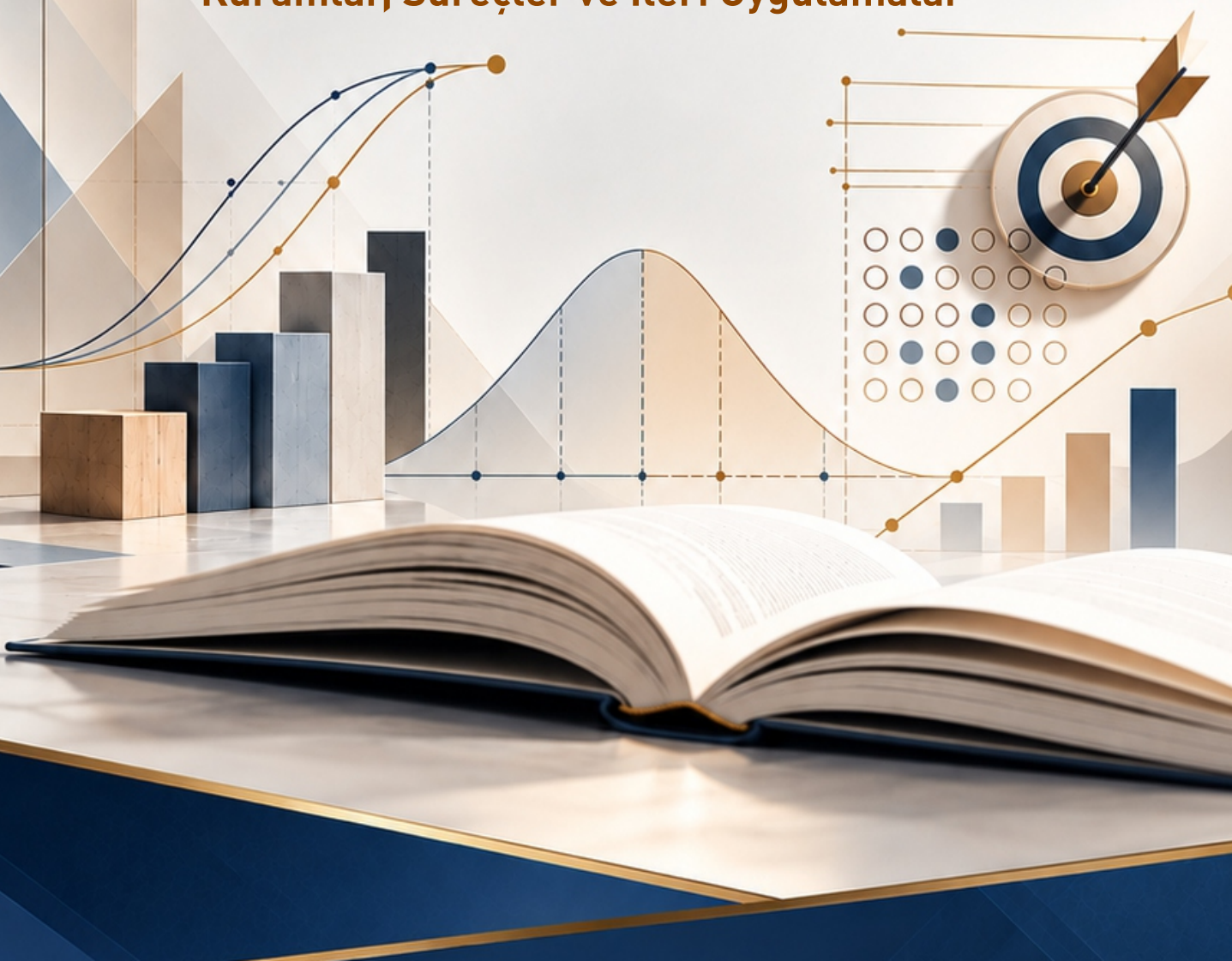
TÜBA
Türkiye Bilimler Akademisi
Turkish Academy of Sciences

ÖSYM

ÖLÇME VE DEĞERLENDİRME

Kuramlar, Süreçler ve İleri Uygulamalar



TÜBA
TÜRKİYE BİLİMLER AKADEMİSİ
TURKISH ACADEMY OF SCIENCES

ÖSYM

ÖLÇME
VE
DEĞERLENDİRME

Kuramlar, Süreçler ve İleri Uygulamalar

Ankara, 2026

Ölçme ve Değerlendirme: Kuramlar, Süreçler ve İleri Uygulamalar
Measurement and Assessment: Theories, Processes, and Advanced Applications

© Türkiye Bilimler Akademisi / © Turkish Academy of Sciences Publications

Bilim ve Düşünce Serisi: 74

ISBN: 978-625-6110-89-2

DOI: 10.53478/TUBA.978-625-6110-89-2

Editörler/Editors
Prof. Dr. Bayram Ali ERSOY
Prof. Dr. Mustafa DOĞAN
Doç. Dr. Yıldız YILDIRIM
Dr. Tuba GÜNDÜZ

Yayın Danışmanı / Publication Advisor
Doç. Dr. Mürsel DOĞRUL

Sekreterya/Secretariat
Büşra AYGÜN

Grafik-Tasarım/Graphic Design
Mustafa ALTINTEPE

Bu kitapta yer alan tüm yazıların dil, bilim, etik ve hukuk açısından bilimsel sorumluluğu yazarlarına aittir. Editörlerin ve Türkiye Bilimler Akademisi'nin sorumluluğu bulunmamaktadır.

Full responsibility for the language, scientific content, ethical standards and legal aspects of all chapters in this book lies with the respective authors. The editors and the Turkish Academy of Sciences do not assume any responsibility in this regard.

Turkish Academy of Sciences
Vedat Dalokay Cad. No: 112 06670 Çankaya - Ankara
Tel: +90 312 442 29 03 • www.tuba.gov.tr

Baskı Yeri, Tarihi / *Publication Place and Date*
Tek Ses Ofset Matbaacılık San. Ltd. Şti. Sertifka No: 44186 • Ankara • Ekim 2026

Ölçme ve Değerlendirme: Kuramlar, Süreçler ve İleri Uygulamalar/ ed. Bayram Ali Ersoy, Mustafa Doğan, Yıldız Yıldırım ve Tuba Gündüz. -- Ankara : Türkiye Bilimler Akademisi, 2026.
712 s. : tabl. ; 19 x 27 sm. – (Bilim ve Düşün Serisi No; 74)
Bibliyografya ve dizin içerir
ISBN 978-625-6110-89-2
1.Eğitimde ölçme ve değerlendirme. 2. Educational measurement and assessment. 3. Ölçme araçları. 4. Measurement instruments. 5. Eğitimde değerlendirme. 6. Educational assessment. 7.Korelasyon. 8. Correlation.

Bu çalışma, Creative Commons Atıf-GayriTicari-Türetilemez (BY-NC-ND) 4.0 Uluslararası Lisansı kapsamında lisanslanmıştır.

This work is licensed under a Creative Commons Attribution-NonCommercial-No Derivative (BY-NC-ND) 4.0 International License.



TÜRKİYE BİLİMLER AKADEMİSİ
TURKISH ACADEMY OF SCIENCES



MEASUREMENT AND ASSESSMENT

Theories, Processes, and Advanced Applications

Ankara, 2026

İÇİNDEKİLER
Table of Contents

Takdim / Foreword
Prof. Dr. Muzaffer ŞEKER
vii

Ön Söz / Preface
Prof. Dr. Bayram Ali Ersoy
xi

Editör Notu / Editorial Note
Editörler/Editors
xv

Ölçme ve Değerlendirmede Temel Kavramlar
Fundamental Concepts in Measurement and Assessment
Prof. Dr. Bayram Çetin
1

Test Kuramları
Test Theories
Prof. Dr. Tahsin Oğuz Başokçu, Prof. Dr. Nuri Doğan
29

Ölçme Araçlarının ve Sonuçlarının Sahip Olması Gereken Nitelikler
Essential Qualities of Measurement Instruments, Scores, and Score Interpretations
Prof. Dr. Mustafa İlhan
75

Ölçmeye Konu Olan Psikolojik Özelliklerin Ölçülmesi
Measurement of Psychological Attributes
Dr. Öğr. Üyesi Ömer Kutlu, Doç. Dr. Özge Altıntaş
133

Test Geliştirme Süreci
Test Development Process
Prof. Dr. Melek Gülşah Şahin
199

Bilişsel Alana Özgü Madde Türleri ve Madde Yazım İlkeleri
Cognitive Domain Item Types and Principles of Item Writing
Dr. Tuba Gündüz, Doç. Dr. Yıldız Yıldırım
237

Performans Değerlendirme Araç ve Yöntemleri
Performance Assessment Tools and Methods
Prof. Dr. Mehtap Çakan
277

Ölçme ve Değerlendirmede Çeşitli Öğrenme Yaklaşımlarının Kullanılması
Using Various Learning Approaches in Educational Measurement and Evaluation
Doç. Dr. Erkan Hasan Atalmış, Prof. Dr. Gökhan Aksu
327

Sınav Çeşitleri ve Sınav Planı
Types of Examinations and Test Planning
Prof. Dr. Seher Yalçın
365

Puanlama Yöntemleri
Scoring Methods
Doç. Dr. Murat Doğan Şahin
391

Ölçme Sonuçlarına Uygulanan İstatistiksel İşlemler
Statistical Analysis of Measurement Results
Doç. Dr. Eren Can Aybek
419

Değerlendirme ve Yerleştirme
Assessment and Placement
Dr. Öğr. Üyesi Sungur Gürel, Dr. Öğr. Üyesi Muhammet Yorulmaz
465

Türkiye Yüzyılı Maarif Modeli'nde Tanımlanan Becerilerin Ölçülmesi
Measurement of the Skills Defined in the Century of Türkiye Education Model
Doç. Dr. Sami Pektaş, Dr. Öğr. Üyesi Semih Aşiret
489

Etik, Adalet ve Yasal Yükümlülükler
Ethics, Justice and Legal Obligations
Doç. Dr. Asiye Şengül Avşar
531

Ölçme ve Değerlendirmede İleri Yaklaşımlar
Advanced Approaches to Measurement and Assessment
Prof. Dr. Ayfer Sayın, Doç. Dr. Selma Şenel
573

Yeterliklerin ve Normların Belirlenmesi
Determining Proficiency Levels and Norms
Doç. Dr. Ahmet Salih Şimşek
655

Dizin / Index
687

TAKDİM

Türkiye Bilimler Akademisi (TÜBA), Osmanlı'dan Cumhuriyet'e uzanan bilim ve kültür mirasımızın üzerine inşa edilen, Türkiye'nin ulusal bilimler akademisidir. TÜBA, bilimsel düşüncenin ve akademik üretimin geliştirilmesi, bilim insanlarının desteklenmesi ve ülkemizin bilimsel kapasitesinin güçlendirilmesi amacıyla farklı disiplinlerden bilim insanlarını ve kurumları bir araya getiren çalışmalar yürütmektedir. Akademimiz, bilimin evrensel niteliğini ve toplumların gelişimindeki belirleyici rolünü esas alarak Türkiye'nin bilim ekosistemine katkı sunmayı sürdürmektedir.

Bilimsel ilerleme, nitelikli insan kaynağının yetiştirilmesi ve bu insan kaynağının bilgi, beceri ve yetkinliklerinin doğru biçimde ortaya konulmasıyla yakından ilişkilidir. Bu bakımdan ölçme ve değerlendirme, eğitim ve öğretimin her aşamasında başvurulan, nicel ve nitel veriler üzerinden performans ve kaliteye ilişkin sağlıklı analizler yapılmasına, doğru kararlar alınmasına ve süreçlerin geliştirilmesine imkân sağlayan temel bir alandır. Geçerli, güvenilir, adil ve şeffaf ölçme sonuçları; eğitimde kalite ve fırsat eşitliğinin yanı sıra akademik liyakat, rekabet ve dürüstlük bakımından da önemli bir zemin oluşturmaktadır. Nitekim çağdaş ölçme ve değerlendirme anlayışı, psikometri ve veri analitiğinin yanında etik, adalet, kapsayıcılık ve kalite süreçlerini de kapsayan disiplinler arası bir niteliğe ulaşmıştır.

Bu anlayış doğrultusunda, TÜBA ile Ölçme, Seçme ve Yerleştirme Merkezi Başkanlığı (ÖSYM) iş birliğiyle hazırlanan bu eserin, ülkemizin ölçme ve değerlendirme alanındaki birikimini bütüncül biçimde ortaya koyması bakımından değerli bir çalışma olduğunu düşünüyorum. 16 bölüm ve 23 yazarın katkısıyla hazırlanan eserde, ölçme ve değerlendirmenin temel kavramlarından test kuramlarına, geçerlik ve güvenirlikten test geliştirmeye, madde yazımından performans değerlendirmesine, sınav planlamasından puanlama ve istatistiksel analizlere, değerlendirme ve yerleştirmeden etik, adalet ve yasal yükümlülüklere kadar alanın temel konuları bir arada ele alınmıştır. Bunun yanında yapay zekâ destekli uygulamalar, bilgisayar ortamında bireye uyarlanmış testler, çok aşamalı testler, test eşitleme, yeterlik ve norm çalışmaları gibi güncel yaklaşımlara da yer verilmiştir.

Bu kapsamlı çalışma, yalnızca mevcut uygulamaları değerlendiren bir kaynak olmanın ötesinde, ölçme ve değerlendirme alanındaki gelişmeleri, yeni yöntemleri ve geleceğin ihtiyaçlarını birlikte düşünme imkânı sunmaktadır. TÜBA olarak bu tür çalışmalarla ülkemizin bilimsel birikiminin görünür kılınmasına, kurumlar arasındaki iş birliğinin güçlendirilmesine ve özellikle genç bilim insanlarının güncel bilimsel yaklaşımlarla buluşmasına katkı sağlamayı önemsiyoruz. Bu doğrultuda akademimizin, bilimsel niteliği ve toplumsal faydayı esas alan çalışmaların desteklenmesine yönelik katkıları devam edecektir.

Türkiye'nin bilimsel kapasitesinin gelişmesi, uluslararası bilim camiasıyla kurduğumuz ilişkiler bakımından da önem taşımaktadır. TÜBA, bilim diplomasisi faaliyetleri kapsamında ülkemizin bilimsel birikimini, karşılaştığı sorunları ve geliştirdiği iyi uygulamaları uluslararası platformlarda paylaşırken, farklı ülkelerin ve bilim akademilerinin deneyimlerinden de yararlanmaktadır. Ölçme ve değerlendirme alanındaki bilimsel birikimin de bu karşılıklı bilgi ve tecrübe paylaşımının önemli başlıklarından biri olduğuna inanıyorum.

Bu değerli eserin hazırlanmasına katkı sunan başta ÖSYM Başkanı Prof. Dr. Bayram Ali Ersoy olmak üzere editörlerimize, 23 yazarımıza, ÖSYM'nin değerli uzmanlarına ve emeği geçen tüm çalışma arkadaşlarımıza teşekkür ediyorum. Eserin ülkemizde ölçme ve değerlendirme alanının gelişmesine, bilimsel kalite anlayışının güçlenmesine ve bu alanda yetişecek yeni nesil araştırmacılara katkı sunmasını temenni ediyorum.

Prof. Dr. Muzaffer ŞEKER
Türkiye Bilimler Akademisi Başkanı

FOREWORD

The Turkish Academy of Sciences (TÜBA) is Türkiye's national academy of sciences. It was founded on a rich scientific and cultural heritage that extends from the Ottoman era to the Republic. TÜBA carries out initiatives that bring together scientists and institutions from a variety of disciplines, with the aim of promoting scientific thought and academic research, supporting scientists and enhancing our country's scientific capabilities. Recognising the universal nature of science and its pivotal role in societal development, our academy continues to contribute to Türkiye's scientific ecosystem.

Scientific progress is closely linked to cultivating highly qualified human resources and accurately assessing their knowledge, skills and competencies. Assessment and evaluation are fundamental to every stage of education and training, providing a basis for analysing performance and quality, making informed decisions, and continuously improving processes. Valid, reliable, fair and transparent assessment results are essential for ensuring quality and equality of opportunity in education, as well as academic merit, competition and integrity. The contemporary understanding of measurement and evaluation has indeed evolved into an interdisciplinary field, encompassing not only psychometrics and data analytics, but also ethics, fairness, inclusivity, and quality assurance processes.

In accordance with this comprehension, I deem this book, which was meticulously assembled through the collaborative efforts of the TÜBA and the Measurement, Selection and Placement Center of Türkiye (ÖSYM), to be a noteworthy contribution that exemplifies Türkiye's collective knowledge and proficiency in the domain of measurement and evaluation from a comprehensive vantage point. Comprising 16 chapters with contributions from 23 authors, it addresses core topics ranging from fundamental measurement and evaluation concepts to test theories, validity and reliability, test development, item writing, performance assessment, examination planning, scoring and statistical analysis, evaluation and placement, and ethics, fairness and legal obligations. Contemporary approaches covered include artificial intelligence-assisted applications, computerised adaptive testing, multistage testing, test equating, and studies on proficiency standards and norms.

As well as serving as a resource for evaluating current practices, this comprehensive work provides an opportunity to consider developments in measurement and evaluation, emerging methodologies and future requirements all together. At TÜBA, we place great importance on raising the profile of our country's scientific knowledge and expertise through initiatives such as this one, strengthening cooperation between institutions and, in particular, enabling young scientists to engage with contemporary scientific approaches. In this regard, our Academy will continue to support endeavours that uphold scientific excellence and generate societal benefit.

Advancing Türkiye's scientific capacity is also important for our relations with the international scientific community. Through its science diplomacy activities, TÜBA shares Türkiye's scientific expertise, the challenges it faces, and its successful practices on international platforms. In turn, TÜBA benefits from the experiences of other countries and academies of science. I believe that the field of measurement and evaluation constitutes an important area for the mutual exchange of knowledge and experience.

I would like to express my sincere gratitude to Prof. Dr. Bayram Ali Ersoy, President of ÖSYM, the editors, the 23 contributing authors, ÖSYM's esteemed experts, and all the other colleagues who contributed to this valuable book. I hope that this work will contribute to advancing the field of measurement and evaluation in Türkiye, strengthening the culture of scientific quality and serving as a valuable resource for the next generation of researchers in this field.

Prof. Dr. Muzaffer ŞEKER
President of the Turkish Academy of Sciences

ÖN SÖZ

Ölçme ve değerlendirme; sadece eğitim sistemlerinin sonuçlarını ortaya koyan bir süreç olmayıp aynı zamanda öğretimin niteliğini geliştiren, öğrenmeyi yönlendiren ve eğitim politikalarına veri sağlayan bir alandır. Bu alan, stratejik bir öneme sahip olup eğitimde alınan kararların temelinde geçerliği ve güvenirliği yüksek ölçme sonuçları bulunmaktadır. Bu nedenle ölçme ve değerlendirme öğretmenlik mesleğinin vazgeçilmez yeterliklerinden biri olmanın yanı sıra eğitim bilimleri, psikometri, veri analitiği ve karar verme süreçlerinin ortak noktasında yer alan disiplinler arası bir çalışma alanıdır.

Son yıllarda eğitimde yaşanan hızlı değişimler, dijitalleşme, yapay zekâ uygulamaları, bilgisayar tabanlı sınav sistemleri ve geniş ölçekli değerlendirme programlarının yaygınlaşması, ölçme ve değerlendirme alanının kapsamını önemli ölçüde genişletmiştir. Artık sadece "iyi soru yazmak" yeterli görülmeyip ölçülmek istenen yapının doğru tanımlanması, uygun kuramsal modelin seçilmesi, testlerin psikometrik özelliklerinin incelenmesi, puanların anlamlı biçimde yorumlanması, etik ilkelerin gözetilmesi ve elde edilen sonuçların eğitim süreçlerine doğru biçimde yansıtılması beklenmektedir. Dolayısıyla çağdaş ölçme ve değerlendirme anlayışı, kuramsal bilgi ile uygulama deneyimini birlikte işe koşmayı gerektiren bütüncül bir bakış açısını zorunlu kılmaktadır.

Bu kitap, bahsi geçen gereksinimlerden hareketle hazırlanmış ve kitabın ÖSYM uzmanlarından alandaki tüm paydaşlara kadar geniş bir kesime hitap eden; gerektiğinde örnek uygulamalara da yer veren, temel düzeyden ileri düzeye uzanan başat bir kaynak olması hedeflenmiştir. Kitapta ölçme ve değerlendirmenin temel kavramlarından başlayarak sınıf içi ölçme ve değerlendirme yaklaşımlarından geniş ölçekli standart test uygulamalarına; klasik test kuramı, madde tepki kuramı ve genellenebilirlik kuramı gibi psikometrik kuramlara; geçerlik ve güvenirlik çalışmalarından test geliştirme süreçlerine; madde yazım ilkelerinden performans değerlendirmeye; puanlama yöntemlerinden istatistiksel veri analizlerine; sınav planlamasından değerlendirme ve yerleştirme süreçlerine kadar alanın temel bilgi ve uygulamaları bir bütün olarak ele alınmaktadır.

Bununla birlikte kitap, günümüzde giderek önem kazanan pek çok güncel konuya da yer vermektedir. Yapay zekâ destekli otomatik madde üretimi, otomatik puanlama sistemleri, bilgisayar ortamında bireye uyarlanmış ve çok aşamalı testler, test eşitleme, standart belirleme yöntemleri, yeterlilik ve norm çalışmaları, etik, adalet, kapsayıcılık ve kalite süreçleri gibi başlıklar, kitabın sadece bugünün gereksinimlerine değil, geleceğin ölçme ve değerlendirme uygulamalarına da ışık tutmaktadır. Ayrıca “Türkiye Yüzyılı Maarif Modeli” kapsamında öne çıkan beceri temelli değerlendirme yaklaşımına yer verilmesi, kitabın güncel eğitim politikalarıyla uyumlu ve çağın beklentilerine cevap veren bir kaynak olmasını sağlamaktadır.

Kitabın hazırlanmasındaki temel amaç; eğitim fakültelerinde öğrenim gören lisans öğrencileri için temel bir ders kitabı, lisansüstü öğrenciler ve araştırmacılar için kapsamlı bir başvuru kaynağı, öğretmenler için uygulamaya dönük bir rehber, ölçme ve değerlendirme uzmanları ile soru geliştiricileri için ise güncel gelişmeleri de kapsayan nitelikli bir kaynak ortaya koymaktır. Özellikle geniş ölçekli sınavların planlanması, uygulanması ve değerlendirilmesi süreçlerinde görev alan uzmanlar açısından kitabın kuram ile uygulamayı bütünleştiren yapısının önemli katkılar sağlayacağına inanıyorum.

Bu kitabın, ölçme ve değerlendirmenin kuramsal temellerini sağlam bir biçimde sunarken; uygulamaya yönelik süreçleri, güncel yaklaşımları ve ileri psikometrik yöntemleri bir bütünlük içerisinde ele alarak okuyucuya kapsamlı bir bakış açısı kazandıracağını düşünüyorum. Böylece okuyucuların sadece temel kavramları öğrenmekle kalmayıp aynı zamanda nitelikli ölçme araçları geliştirebilmeleri, elde edilen verileri bilimsel yöntemlerle analiz edebilmeleri ve bu sonuçları eğitim süreçlerinde etkili biçimde kullanabilmeleri amaçlanmıştır.

Başkanlığımızın kuruluşunun 50. yılı vesilesiyle hazırlanan bu eserin ortaya çıkmasında katkı sağlayan Prof. Dr. Mustafa Doğan'a, Doç. Dr. Yıldız Yıldırım'a ve Dr. Tuba Gündüz'e, bölüm yazarlarımıza, son nüsha düzeltmeleri için Prof. Dr. Mustafa İlhan'a, eserin basımında her türlü desteği veren TÜBA Başkanı Prof. Dr. Muzaffer Şeker'e ve Başkan Yardımcısı Doç. Dr. Mürsel Doğrul ile ekibine teşekkür ediyorum.

Prof. Dr. Bayram Ali ERSOY
Ölçme, Seçme ve Yerleştirme Merkezi Başkanlığı (ÖSYM) Başkanı

PREFACE

Measurement and assessment is a discipline that not only produces the results of education systems but also one that improves the quality of education, guides educational practice and provides data for educational policies. This field is strategically important, with reliable and valid assessment results providing the basis of educational decision-making. Therefore, measurement and assessment is not only one of the indispensable competencies of the teaching profession but also an interdisciplinary field positioned at the intersection of educational sciences, psychometrics, data analytics and decision-making processes.

Rapid changes experienced in education in recent years, such as digitalization, artificial intelligence applications, computer-based testing systems and the proliferation of wide-scale assessment programs, have significantly broadened the scope of the field of testing and assessment. It is no longer enough to "develop high-quality test items;" expectations now extend to accurately defining the constructs to be measured, selecting an appropriate theoretical model, evaluating the psychometric properties of tests, interpreting test scores meaningfully, adhering to ethical principles and correctly incorporating the assessment results into educational processes. Therefore, the contemporary approach to measurement and assessment requires a holistic perspective that integrates theoretical knowledge with practical expertise.

This book has been prepared with these requirements in mind, and aims to serve as a leading resource – covering topics from introductory to advanced levels and supported by illustrative examples where appropriate – that appeals to a broad audience, including ÖSYM experts and other stakeholders in the field. The book comprehensively addresses the fundamental concepts and practices of the field, ranging from basic concepts of measurement and assessment to classroom-based measurement and large-scale standardized testing; from psychometric theories—such as classical test theory, item response theory, and generalizability theory—to validity and reliability studies and test development processes; and from item-writing principles and performance assessment to scoring methods, statistical data analysis, and the planning, administration, assessment and placement processes associated with large-scale examinations.

Additionally, the book also addresses numerous contemporary topics that are becoming increasingly important. The inclusion of topics such as AI-assisted automated item generation, automated scoring systems, computerized adaptive and multistage testing, test equating, standard-setting methods, proficiency and norming studies, as well as ethics, justice, inclusivity, and quality management, ensures that the book responds not only to current requirements but also prepares readers for future developments in

assessment and evaluation. Furthermore, the inclusion of the skill-based approach to assessment, a key feature of the “Century of Türkiye Education Model”, ensures that the book aligns with current educational policies while addressing the evolving demands of contemporary education.

The primary aim of this book is to serve as a foundational textbook for undergraduate students studying in the faculties of education, a comprehensive reference book for graduate students and researchers, a practical guide for teachers, and a high-quality resource for measurement and assessment specialists and item developers that incorporates the latest developments in the field. I believe that the book's structure that integrates theory and practice will make a valuable contribution to the field, particularly for professionals involved in the planning, administration, and evaluation of large-scale examinations.

I believe that this book will provide readers with a comprehensive perspective by presenting a solid foundation in the theoretical principles of measurement and assessment while addressing practical applications, contemporary approaches, and advanced psychometric methods within a unified framework. Therefore, the aim is to enable the readers not only to learn the fundamental concepts but also to develop high-quality measurement tools, to analyse the collected data using scientific methods, and to effectively use these results during the process of education.

I would like to extend my gratitude to Prof. Dr. Mustafa Doğan, Assoc. Prof. Dr. Yıldız Yıldırım, and Dr. Tuba Gündüz for their contributions to the preparation of this work, which was prepared on the occasion of the 50th anniversary of our Presidency; to Prof. Dr. Mustafa İlhan for the proof reading of the book; and to TÜBA President Prof. Dr. Muzaffer Şeker, TÜBA Vice President Assoc. Prof. Dr. Mürsel Doğrul, and his team for their invaluable support throughout the publication process.

Prof. Dr. Bayram Ali ERSOY
Center for Assessment, Selection and Placement (ÖSYM), President

EDİTÖR NOTU

Bu kitap, ölçme ve değerlendirme alanının kuramsal temelleri ile güncel uygulamalarını, bilimsel araştırma ve psikometrik yöntemlerdeki çağdaş gelişmeler ışığında bütüncül bir yaklaşımla ele alan kapsamlı bir başvuru kaynağı niteliğindedir. Kitapta, alanın temel kavram, kuram ve ilkelerinden başlayarak ölçme araçlarının geliştirilmesi, uygulanması ve değerlendirilmesi süreçleri ile daha ileri düzey psikometrik yaklaşımlar bir bütünlük içerisinde sunulmaktadır. Kitabın bölümleri, ölçme ve değerlendirme sürecinin farklı aşamalarını ve bu aşamalar arasındaki yöntemsel ilişkileri ortaya koyacak biçimde birbirini tamamlayan bir yapı içerisinde kurgulanmıştır. Bu yönüyle eser, temel bilgi ve kavramların kazandırılmasının ötesine geçerek okuyucunun ölçme ve değerlendirme süreçlerini bilimsel bir bakış açısıyla bütüncül olarak kavramasını, elde edilen ölçme sonuçlarının doğru bir biçimde yorumlanmasını ve bu sonuçların eğitimsel karar süreçlerine etkili biçimde yansıtılmasını amaçlamaktadır. Eserin, eğitim fakültelerinde öğrenim gören lisans öğrencilerinin alanın temel bilgi ve becerilerini edinmelerine, lisansüstü öğrenciler ile araştırmacıların kuramsal ve yöntemsel bilgilerini derinleştirmelerine, öğretmenlerin ölçme ve değerlendirme uygulamalarını geliştirmelerine ve ölçme-değerlendirme uzmanları ile soru geliştiricilerin güncel bilimsel yaklaşımları ve uygulamalarına katkı sağlayacak nitelikte olduğu değerlendirilmektedir. Özellikle geniş ölçekli sınavların planlanması, uygulanması ve değerlendirilmesi gibi yüksek düzeyde teknik uzmanlık gerektiren süreçler bakımından eser, kuram ile uygulama arasındaki ilişkiyi görünür kılan ve farklı düzeylerdeki okuyucuların gereksinimlerine yanıt verebilecek kapsamlı bir çerçeve sunmaktadır.

Başkanlığımızın yarım asırlık kuruluş yıldönümünü anmak amacıyla yayımlanan bu kitap, 16 bölümden oluşmaktadır. Kitabın ilk bölümü olan "Ölçme ve Değerlendirmede Temel Kavramlar", Prof. Dr. Bayram Çetin tarafından hazırlanmıştır. Ölçme ve değerlendirmenin temelini oluşturan kavramlar, mantıksal bir bütünlük içinde ele alınmış; okuyucunun ilerleyen bölümlerde karşılaşacağı kuramsal ve uygulamalı konular için sağlam bir altyapı oluşturulmuştur. Bu yönüyle ilk bölüm, kitabın kavramsal omurgasını oluşturmaktadır.

İkinci bölüm "Test Kuramları", alanın üç temel yaklaşımını aynı çatı altında buluşturmaktadır. Prof. Dr. Nuri Doğan, klasik test kuramını yalın ve anlaşılır bir dille okuyucuya aktarırken; Prof. Dr. Tahsin Oğuz Başokçu günümüzde psikometrik araştırmaların odağında yer alan madde tepki kuramı ve bilişsel tanı modellerini kuramsal temelleri ve uygulama boyutlarıyla ele almıştır. Böylece bölüm, ölçme kuramlarının tarihsel gelişimini ve günümüzde ulaştığı noktayı bütüncül bir bakış açısıyla sunmaktadır.

Kitabın üçüncü bölümü "Ölçme Araçlarının ve Sonuçlarının Sahip Olması Gereken Nitelikler", Prof. Dr. Mustafa İlhan tarafından hazırlanmıştır. Geçerlik ve güvenirlik başta olmak üzere ölçme araçlarının sahip olması gereken temel psikometrik özellikler, son yıllarda geçerlik anlayışında yaşanan dönüşümler ve APA Standartları doğrultusunda güncel bir yaklaşımla ele alınmış; klasik bilgi birikimi ile çağdaş değerlendirme anlayışı başarılı biçimde bütünleştirilmiştir.

Dördüncü bölüm olan "Ölçmeye Konu Olan Psikolojik Özelliklerin Ölçülmesi", Dr. Öğr. Üyesi Ömer Kutlu ve Doç. Dr. Özge Altıntaş tarafından kaleme alınmıştır. Psikolojik yapıların ölçülmesine ilişkin kuramsal temeller, çağdaş psikometrik yaklaşımlar çerçevesinde açıklanmış; soyut kavramlar çok sayıda örnek ve uygulama ile somutlaştırılarak okuyucuların konuyu derinlemesine kavramaları hedeflenmiştir.

Kitabın beşinci bölümü "Test Geliştirme", Prof. Dr. Melek Gülşah Şahin tarafından hazırlanmıştır. Test geliştirmenin yalnızca belirli adımlardan oluşan doğrusal bir süreç olmadığı; sürekli gözden geçirilen, geliştirilen ve güncellenen dinamik bir yapı olduğu vurgulanarak test geliştirme sürecinin bütün aşamaları güncel yaklaşımlar ışığında ayrıntılı biçimde ele alınmıştır.

Altıncı bölüm "Bilişsel Alana Özgü Madde Türleri ve Madde Yazım İlkeleri", Dr. Tuba Gündüz ile Doç. Dr. Yıldız Yıldırım tarafından hazırlanmıştır. Bölümde bilişsel alanda yaygın olarak kullanılan madde türleri ve madde yazım ilkeleri yalnızca kuramsal yönleriyle değil, son yıllarda gerçekleştirilen bilimsel araştırmaların bulguları ve uygulama örnekleriyle birlikte değerlendirilmiştir. Böylece okuyucular, madde yazımına ilişkin temel ilkelerin yanı sıra güncel araştırmaların ortaya koyduğu yeni yaklaşımları da takip edebilme olanağı bulacaktır.

Kitabın yedinci bölümü "Performans Değerlendirme Araç ve Yöntemleri", Prof. Dr. Mehtap Çakan tarafından kaleme alınmıştır. Son yıllarda giderek önem kazanan performans değerlendirme yaklaşımları, geleneksel yöntemlerden çağdaş değerlendirme uygulamalarına uzanan geniş bir perspektifle ele alınmış; performansın ölçülmesinde kullanılan araçlar, puanlama yöntemleri ve güncel eğilimler sunulmuştur.

Sekizinci bölüm olan "Ölçme ve Değerlendirmede Çeşitli Öğrenme Yaklaşımlarının Kullanılması", Prof. Dr. Gökhan Aksu ile Doç. Dr. Erkan Hasan Atalmış tarafından yazılmıştır. Bölümde çağdaş öğrenme yaklaşımlarına dayalı ölçme ve değerlendirme uygulamaları ile farklı ölçme araçlarının eğitim ortamlarında nasıl kullanılabileceği çok sayıda örnek üzerinden açıklanmış; teori ile uygulama arasındaki bağ güçlü biçimde kurulmuştur.

Kitabın dokuzuncu bölümü “Sınav Çeşitleri ve Sınav Planı” Prof. Dr. Seher Yalçın tarafından yazılmıştır. Bölümde sınavlar standart olup olmaması durumuna, katılımcı/aday sayısına, risk düzeyine, amacına ve uygulama türüne göre sınıflandırılarak detaylandırılmıştır. Sınav planı başlığında ise sınav güvenliği ve itirazlar başlıkları dikkat çekmektedir.

Onuncu bölüm olan “Puanlama Yöntemleri” Doç. Dr. Murat Doğan Şahin tarafından kaleme alınmıştır. Ortamına göre ve puanın veriliş biçimine göre puanlama yaklaşımları sınıflandırılarak sunulmuştur. Dahası bölümde otomatik puanlama ve karşılaştırma yargılarıyla puanlama gibi yenilikçi yaklaşımlara da yer verilmiştir.

Kitabın on birinci bölümü “Ölçme Sonuçlarına Uygulanan İstatistiksel İşlemler”, Doç. Dr. Eren Can Aybek tarafından hazırlanmıştır. Bölüm, az sayıda kişinin seçilmesinin amaçlandığı sınavlar için yeni bir ayırt edicilik indeksi sunmaktadır. Aynı zamanda tamamıyla ücretsiz bir uygulama olan Shiny Item Analysis yazılımını örneklerle tanıtmasıyla öne çıkmaktadır.

Kitabın on ikinci bölümü olan “Değerlendirme ve Yerleştirme”, Dr. Öğr. Üyesi Sungur Gürel ile Dr. Öğr. Üyesi Muhammet Yorulmaz tarafından kaleme alınmıştır. Bölümde biçimlendirici değerlendirme başlığında geri bildirim kavramı da ele alınmış olup geniş ölçekli değerlendirmelere de detaylı değinilmiştir. Bölüm, alanyazından farklı olarak yerleştirme algoritmasının ayrıntılı biçimde sunulmasıyla öne çıkmaktadır.

Kitabın on üçüncü bölümü olan “Türkiye Yüzyılı Maarif Modeli’nde Tanımlanan Becerilerin Ölçülmesi”, Doç. Dr. Sami Pektaş ile Dr. Öğr. Üyesi Semih Aşiret tarafından yazılmıştır. Bu bölümde öncelikle Türkiye Yüzyılı Maarif Modeli kapsamındaki mevcut öğretim programlarının yapısı tanıtılmıştır. Dahası program kapsamında ele alınan becerilerin ölçülmesinde kullanılabilecek bağlam temelli sorulara yönelik örnekler sunulmuştur.

On dördüncü bölüm “Etik, Adalet ve Yasal Yükümlülükler”, Doç. Dr. Asiye Şengül Avşar tarafından kaleme alınmıştır. Bölümde ölçme ve değerlendirme sürecinin tüm paydaşlarına ilişkin sorumluluklar tek tek ele alınmıştır. Ayrıca ölçmede adalet, yansızlık, kapsayıcılık, hesap verebilirlik ve şeffaflık gibi kavramlar APA standartlarından da alıntı yapılarak yakın mercek altında incelenmiştir.

On beşinci bölüm olan “Ölçme ve Değerlendirmede İleri Yaklaşımlar”, Prof. Dr. Ayfer Sayın ile Doç. Dr. Selma Şenel tarafından hazırlanmıştır. Bölümde bilgisayar ortamında bireye uyarlanmış testler ve çok aşamalı testler tüm süreciyle ele alınmış olup uluslararası örnek uygulamalar ile beraber sunulmuştur. Otomatik madde üretimi konusu ise yöntemleri, aşamaları ve uygulama örnekleriyle detaylandırılmıştır. Test eşitleme konusunu da ele alan bölümde test eşitlemenin temel kavramları, desenleri ve yöntemleri uygulama örnekleriyle ayrıntılı biçimde sunulmuştur.

Kitabın on altıncı ve son bölümü olan “Yeterliklerin ve Normların Belirlenmesi” ise Doç. Dr. Ahmet Salih Şimşek tarafından kaleme alınmıştır. Bölümde yeterlik ve norm kavramları detaylarıyla yer almaktadır. Standart belirleme yöntemleri kapsamında ise alanyazında bulunan hem test-merkezli yaklaşımlar hem de aday-merkezli yaklaşımlar sunulmuştur.

Bu kitabın eğitim fakültelerinde, lisansüstü programlarda, okullarda, ölçme ve değerlendirme merkezlerinde ve sınav geliştirme birimlerinde görev yapan tüm okuyucular için yararlı ve güvenilir bir başvuru kaynağı olmasını; ölçme ve değerlendirme alanındaki kuramsal bilgiye ve uygulama süreçlerine büyük katkılar sunmasını diliyoruz.

Prof. Dr. Bayram Ali ERSOY
ÖSYM
ORCID: 0000-0002-8307-9644

Prof. Dr. Mustafa DOĞAN
ÖSYM
ORCID: 0000-0002-7380-9869

Doç. Dr. Yıldız YILDIRIM
ÖSYM
ORCID: 0000-0001-8434-5062

Dr. Tuba GÜNDÜZ
ÖSYM
ORCID: 0000-0002-0921-9290

EDITORIAL NOTE

This book constitutes a comprehensive reference work that addresses the theoretical foundations and current practices of measurement and assessment from an integrated perspective, in light of contemporary developments in scientific research and psychometric methods. Beginning with the fundamental concepts, theories, and principles of the field are presented alongside the processes of developing, administering, and evaluating measurement instruments, as well as more advanced psychometric approaches, within a coherent and integrated framework. The chapters are structured in a complementary manner, with the aim of demonstrating the different stages of the measurement and assessment process and the methodological relationships among these stages. In this respect, the work goes beyond the acquisition of basic knowledge and concepts, aiming to enable readers to develop a comprehensive understanding of assessment and evaluation processes from a scientific perspective, to interpret measurement results accurately, and to effectively incorporate these results into educational decision-making processes. The work is considered to be of value in supporting undergraduate students in faculties of education in acquiring the fundamental knowledge and skills of the field; enabling graduate students and researchers to deepen their theoretical and methodological knowledge; helping teachers improve their assessment and evaluation practices; and contributing to the application of current scientific approaches by assessment and evaluation specialists and item developers. In particular, with regard to processes requiring a high level of technical expertise, such as the planning, administration, and evaluation of large scale assessments, the work provides a comprehensive framework that makes the relationship between theory and practice explicit and responds to the needs of readers at different levels.

This book, published to commemorate the half-century anniversary of our Presidency, comprises 16 chapters. The first chapter, "Fundamental Concepts in Measurement and Assessment," was written by Prof. Dr. Bayram Çetin. The foundational concepts of measurement and assessment are presented in a coherent and systematic manner, providing a solid foundation for the theoretical and practical topics that readers will encounter in subsequent chapters. In this respect, the first chapter serves as the conceptual backbone of the book.

The second section, "Test Theories," brings together the three fundamental approaches in the field. Prof. Dr. Nuri Doğan presents Classical Test Theory – a subject to which he has devoted many years of research – in clear, accessible language, while Prof. Dr.

Tahsin Oğuz Başokçu discusses Item Response Theory and Cognitive Diagnosis Models – currently at the forefront of psychometric research – covering both their theoretical foundations and practical applications. Thus, the chapter provides a holistic perspective on the historical development of test theories and their current status.

The third chapter of the book, "Essential Qualities of Measurement Instruments, Scores and Score Interpretations," was prepared by Prof. Dr. Mustafa İlhan. It addresses the fundamental psychometric properties required of measurement instruments – particularly validity and reliability – through a contemporary approach that reflects recent shifts in the conceptualization of validity and aligns with the Standards for APA; it effectively integrates established knowledge with modern assessment perspectives.

The fourth chapter, “Measurement of Psychological Attributes” was authored by Asst. Prof. Dr. Ömer Kutlu and Assoc. Prof. Dr. Özge Altıntaş. The theoretical foundations underlying the measurement of psychological constructs are explained within the framework of contemporary psychometric approaches; abstract concepts are made concrete through numerous examples and applications, with the aim of enabling readers to grasp the subject matter in depth.

The fifth chapter of the book, "Test Development Process," was authored by Prof. Dr. Melek Gülşah Şahin. Emphasizing that test development is not merely a linear sequence of predefined steps – but rather a dynamic process subject to continuous review, improvement, and updating – the chapter examines all stages of the test development process in detail, in light of contemporary approaches.

The sixth chapter, "Cognitive Domain Item Types and Principles of Item Writing," was authored by Dr. Tuba Gündüz and Assoc. Prof. Dr. Yıldız Yıldırım. In this chapter, commonly used item types and item-writing principles are examined not only in terms of their theoretical foundations but also through recent scientific research and practical examples. Thus, readers will be able to follow the fundamental principles regarding item writing as well as the emerging approaches from current research.

The seventh chapter of the book, "Performance Assessment Tools and Methods," was authored by Prof. Dr. Mehtap Çakan. Performance assessment approaches – which have become increasingly important in recent years – are examined from a broad perspective ranging from traditional methods to contemporary assessment practices; the chapter also presents the tools used to assess performance, scoring methods, and current trends.

The eighth chapter, "Using Various Learning Approaches in Educational Measurement and Evaluation," was authored by Prof. Dr. Gökhan Aksu and Assoc. Prof. Dr. Erkan Hasan Atalmış. In this chapter, measurement and evaluation practices grounded in contemporary learning approaches and the application of various assessment tools in educational settings are explained through numerous examples, thereby strengthening the connection between theory and practice.

The ninth chapter of the book, “Types of Examinations and Test Planning,” was authored by Prof. Dr. Seher Yalçın. In this chapter, exams are classified and described in detail based on criteria such as standardization, the number of examinees, stakes level, purpose, and the mode of administration. The section on exam planning particularly highlights issues related to exam security and appeals.

The tenth chapter, “Scoring Methods,” was authored by Assoc. Prof. Dr. Murat Doğan Şahin. It presents a classification of scoring approaches based on different assessment contexts and scoring procedures. Furthermore, the chapter covers innovative approaches such as automated scoring and comparative judgment-based scoring.

The eleventh chapter of the book, “Statistical Analysis of Measurement Results,” was authored by Assoc. Prof. Dr. Eren Can Aybek. The chapter introduces a new discrimination index for exams designed to select a limited number of individuals. It also stands out for presenting – through illustrative examples – the Shiny Item Analysis software, which is a free software application.

The twelfth chapter of the book, titled “Assessment and Placement,” was authored by Asst. Prof. Dr. Sungur Gürel and Asst. Prof. Dr. Muhammet Yorulmaz. The chapter discusses the concept of effective feedback within the context of formative assessment and provides a detailed discussion of large-scale assessments. It stands out from existing literature by presenting the placement algorithm in considerable detail.

The thirteenth chapter of the book, titled “Measurement of the Skills Defined in the Century of Türkiye Education Model,” was authored by Assoc. Prof. Dr. Sami Pektaş and Asst. Prof. Dr. Semih Aşiret. This chapter first outlines the structure of the current curricula within the framework of the Century of Türkiye Education Model. Furthermore, it presents examples of context-based questions that can be used to assess the skills defined in the curriculum.

The fourteenth chapter, “Ethics, Justice, and Legal Obligations,” was authored by Assoc. Prof. Dr. Asiye Şengül Avşar. This chapter systematically addresses the responsibilities of all stakeholders involved in the measurement and assessment process. Furthermore, concepts such as fairness, impartiality, inclusivity, accountability, and transparency in measurement are examined in depth, supported by references to APA standards.

The fifteenth chapter, titled "Advanced Approaches in Measurement and Assessment," was authored by Prof. Dr. Ayfer Sayın and Assoc. Prof. Dr. Selma Şenel. This chapter examines the entire process of computerized adaptive testing and multistage testing, presenting them alongside international examples of their implementation. The chapter provides a detailed discussion of automatic item generation, including its methods, stages, and practical examples. The chapter also addresses the subject of test equating; it presents the fundamental concepts, designs, and methods of test equating in detail, accompanied by practical examples.

The sixteenth and final chapter of the book, titled "Determining Proficiency Levels and Norms," was authored by Assoc. Prof. Dr. Ahmet Salih Şimşek. This chapter examines the concepts of competencies and norms in detail. In discussing standard-setting methods, it presents both test-centred and candidate-centred approaches found in the literature.

We hope that this book serves as a useful and reliable reference for all readers from faculties of education, graduate programs, schools, measurement and evaluation centres, and test development units, and that it makes significant contributions to both theoretical knowledge and practical processes in the field of testing and assessment.

Prof. Dr. Bayram Ali ERSOY
ÖSYM
ORCID: 0000-0002-8307-9644

Prof. Dr. Mustafa DOĞAN
ÖSYM
ORCID: 0000-0002-7380-9869

Assoc. Prof. Dr. Yıldız YILDIRIM
ÖSYM
ORCID: 0000-0001-8434-5062

Dr. Tuba GÜNDÜZ
ÖSYM
ORCID: 0000-0002-0921-9290

1. BÖLÜM

ÖLÇME VE DEĞERLENDİRMEDE TEMEL KAVRAMLAR

FUNDAMENTAL CONCEPTS IN MEASUREMENT AND ASSESSMENT

Prof. Dr. Bayram ÇETİN

ÖLÇME VE DEĞERLENDİRMEDE TEMEL KAVRAMLAR

Prof. Dr. Bayram ÇETİN
Gaziantep Üniversitesi

Öz

Bu bölümde, ölçme ve değerlendirme alanına ilişkin temel kavramlar ele alınmaktadır. İlk olarak bilimsel araştırmaların ve eğitimde ölçme uygulamalarının temelini oluşturan değişken kavramına yer verilmekte; değişkenler ifade ediliş şekillerine (nitel ve nicel), alabileceği değerlerin sürekliliğine (sürekli ve süreksiz) ve araştırmadaki rollerine (bağımlı, bağımsız, kontrol, aracı ve düzenleyici) göre sınıflandırılmaktadır. Ardından ölçme kavramı tanımlanarak doğrudan, dolaylı ve türetilmiş ölçme üzerinde durulmaktadır. Devamında ölçme kuralı ve ölçüm kavramlarından söz edilmekte, ölçmede sıfır kavramından bahsedilip mutlak ve tanımlanmış sıfır örneklerle açıklanmaktadır. Ölçmede birim başlığı altında birim türleri (doğal birim ve yapay birim) ile birimin sahip olması gereken özelliklere (eşitlik, genellik ve kullanışlılık) değinilmektedir. Ölçek türleri başlığında sınıflama, sıralama, eşit aralıklı ve oranlı ölçek türlerinin özelliklerine, birbirlerinden ayrışan yönlerine ve istatistiksel işlemler açısından sağladıkları olanaklara ilişkin bilgiler sunulmaktadır. Bunu takiben korelasyon konusu işlenmekte; bu kapsamda korelasyon katsayısına ilişkin temel özellikler, saçılım grafikleri, korelasyon katsayısının hesaplanması, korelasyon teknikleri ve korelasyon matrisi alt başlıkları aktarılmaktadır. Bölümde ayrıca ölçüt terimi ve ölçüt türleri (mutlak ve bağıl) açıklığa kavuşturulmaktadır. Değerlendirmenin tanımı yapıldıktan sonra ölçütüne (bağıl ve mutlak) ve amacına (tanıma-yerleştirmeye, biçimlendirme-yetiştirmeye ve değer biçmeye) göre değerlendirme türleri başlıkları ile bölüm sonlandırılmaktadır. Böylece bölüm, ölçme ve değerlendirme alanına ait temel kavramları açıklayıp kuramsal bilgileri örneklerle destekleyerek ilerleyen konular için gereken altyapıyı oluşturmayı amaçlamaktadır.

Anahtar Kelimeler

Ölçme, Değerlendirme, Değişken, Ölçüt, Ölçek, Korelasyon

FUNDAMENTAL CONCEPTS IN MEASUREMENT AND ASSESSMENT

Prof. Dr. Bayram ÇETİN
Gaziantep University

Abstract

This chapter introduces the fundamental concepts in measurement and assessment. It begins by examining the concept of variables, which forms the basis of scientific research and educational measurement. Variables are classified according to their mode of expression (qualitative or quantitative), whether their values are continuous or discrete, and their roles in research (dependent, independent, control, mediating, and moderating). The chapter then defines measurement and discusses direct, indirect, and derived measurement. It proceeds to explain measurement rules and measurement results, followed by a discussion of the role of zero in measurement, distinguishing between absolute and arbitrary (defined) zero through illustrative examples. The chapter also introduces natural and artificial measurement units, together with the essential characteristics of measurement units—equality, generality, and practicality. It further examines the four levels of measurement (nominal, ordinal, interval, and ratio), highlighting their defining characteristics, key distinctions, and implications for statistical analyses. The concept of correlation is then presented, covering the properties and calculation of correlation coefficients, scatterplots, methods of correlation analysis, and correlation matrices. In addition, the chapter explains the concept of a criterion and distinguishes between absolute and relative criteria. Finally, after defining assessment, it discusses types of assessment based on the criterion used (criterion-referenced and norm-referenced) and their primary purposes (placement, formative, and summative). Overall, the chapter lays the conceptual and theoretical foundation for subsequent chapters by integrating key concepts with illustrative examples.

Keywords

Measurement, Assessment, Variable, Criterion, Scale, Correlation

1. GİRİŞ

Ölçme ve değerlendirme eğitim sürecinin vazgeçilmez bir unsurudur. Öğrencilerin öğrenme düzeyine ilişkin çıkarım yapmayı mümkün kılan, ölçme ve değerlendirme uygulamalarıdır. Bunun yanında eğitimde ve psikolojide yapılan ölçmeler pek çok farklı uygulamanın etkililiğini görebilmeyi sağlar. Ölçme ve değerlendirme olmasaydı herhangi bir müdahalenin etkisini tespit etmeye yönelik deneysel çalışmaları yapmak mümkün olmazdı. Örneğin, bir öğretim yönteminin başarı üzerindeki etkisini sınamaya yönelik deneysel bir çalışmada, gerçekleştirilen müdahalenin etkisi ancak ölçme ve değerlendirme uygulamalarıyla (sınavlar, gözlem vb.) tespit edilebilir. Benzer şekilde psikolojik iyi oluş düzeyini artırmaya yönelik bir programın etkililiğini araştırmak için depresyon ölçeği ve psikolojik iyi oluş ölçeği gibi ölçme araçlarından yararlanılır.

Bir öğrenci, öğrenim hayatı boyunca sürekli ölçme ve değerlendirme uygulamalarından geçer. Bu uygulamaların amacı değişkenlik gösterebilir. Okul öncesi eğitimde öğrencinin okula hazırlık durumunu saptamak için yapılan ölçmelerden, bilişsel, duyuşsal, psikomotor gelişimini izlemek için yapılan ölçmelere, ders geçme durumlarına karar vermek için uygulanan testlerden kademeler arası geçişte yapılan sınavlara kadar farklı amaçlarla birçok ölçme işlemi yürütülmektedir. Eğitim dışındaki alanlarda da değişik amaçlarla ölçme ve değerlendirme uygulamaları gerçekleştirilmektedir. Bir işyerine eleman seçiminde, kliniklerde ruh sağlığının değerlendirilmesinde, mesleki yönlendirme için kişilerin ilgi alanlarının belirlenmesinde yapılan ölçmeleri, ölçme ve değerlendirmeye ilişkin farklı uygulamalar olarak düşünebiliriz.

Bu bölümde kitabın ilerleyen bölümlerinde kullanılacak çeşitli kavramların tanıtılması amaçlanmıştır. Bunun okuyucular için kitabı daha anlaşılır kılacağı düşünülmektedir. Bu kapsamda aşağıda önce değişken kavramı ile türleri açıklanmıştır. Devamında sırasıyla ölçme ve ölçme türleri, ölçme kuralı, ölçüm, ölçmede sıfır noktası, ölçmede birim ve birimin sahip olması gereken özellikler ile ölçek ve ölçek türleri başlıklarına yer verilmiştir. Sıralanan başlıkları takiben korelasyon konusu ele alınıp korelasyon katsayısı, saçılım grafikleri, korelasyonun hesaplanması ve korelasyon teknikleri ile korelasyon matrisi üzerinde durulmuştur. Ardından ölçüt ve ölçüt türleri ile değerlendirme ve değerlendirme türlerinden bahsedilerek bölüm tamamlanmıştır.

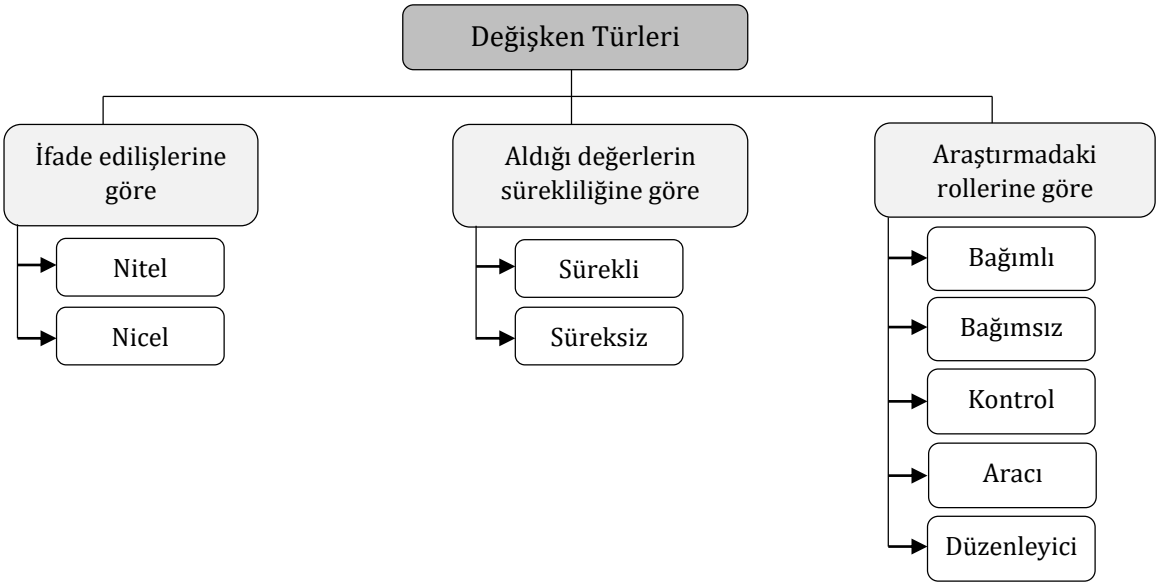
2. DEĞİŞKEN VE TÜRLERİ

Değişken; birden fazla değer alabilen, diğer bir deyişle sabit olmayan özelliklerdir. Daha açık bir anlatımla değişkenler bir araştırmada veya analizde ölçülen, gözlemlenen ya da manipüle edilen aynı zamanda en az iki farklı değeri bulunan nitelikler/özelliklerdir. Değişkenler ifade edilişine, aldığı değerin sürekliliğine ve araştırmadaki rolüne göre farklı şekilde sınıflandırılabilir. Bu sınıflamalara göre değişken türleri Şekil 1'de özetlenmiştir.

Yapısı itibariyle değişken olan nitelikler, kimi durumlarda sabit halini alabilir. Örneğin, Ankara'da yer alan ve toplamda 20 şubesi bulunan bir okulda yapılan araştırmada toplam şube sayısı değişkenlik göstermeyeceği için sabit olarak değerlendirilir. Fakat aynı araştırmayı tek bir okulda değil de Ankara'nın Çankaya ilçesindeki ilkokulları kapsayacak şekilde yürütürsek okullardaki şube sayıları farklılık gösterecek, haliyle şube sayısı araştırmada değişken konumunda olacaktır (Boztunç Öztürk & Şahin, 2021).

Şekil 1

Değişkenlerin Sınıflandırılması



2.1. İfade Edilişlerine Göre Değişkenler

Değişkenler ifade edilişlerine göre nitel ve nicel değişken olarak ikiye ayrılır.

2.1.1. Nicel Değişken

Aldığı değerler sayısal olarak ifade edilebilen ve bu sayıların miktar (büyüklük/küçüklük) bildirdiği değişkenlerdir. Boy uzunluğu (birim ile), mesleki kıdem, sınıftaki öğrenci sayısı, yarışmacıların bir koşuyu tamamlama sıraları, hava sıcaklığı vb. nicel değişkenlere örnek olarak verilebilir.

2.1.2. Nitel Değişken

Nicel ölçüm birimleri ile ifade edilemeyen, kategorik verilere dayanan değişkenlerdir. Nitel değişkenlere verilen değerler, sayısal olarak büyüklük-küçüklük anlamı içermez. Cinsiyet, medeni durum, etnik köken veya doğum yeri nitel değişkenlere örnek olarak verilebilir.

2.2. Alabileceği Değerlerin Sürekliliğine Göre Değişkenler

Değişkenler aldıkları değerin sürekliliğine göre sürekli ve süreksiz olarak sınıflandırılır.

2.2.1. Sürekli Değişken

Belli bir aralıkta sonsuz sayıda değer alabilen değişkenlerdir. Örnek olarak ortam sıcaklığı, ağırlık, zekâ düzeyi veya boy uzunluğu verilebilir. Mesela 165 cm ile 167 cm arasındaki tek değer 166 olmayıp bu aralıkta sonsuz sayıda değer söz konusudur. Bir kişinin boy uzunluğu 165 ya da 166 cm olabileceği gibi 165,1 cm, 165,5 cm, 165,9 vb. ondalık değerler de alabilir. Yani sürekli değişkenler yalnızca tam sayılı değerler almaz, ondalık-yüzdelik-bindelik vb. değerleri de içerebilir.

2.2.2. Süreksiz Değişken

Diğer adı kesikli değişken olup belli bir aralıkta sınırlı sayıda değer alabilir. Kardeş sayısı, cinsiyet, eğitim durumu veya göz rengi süreksiz değişkene örnek olarak gösterilebilir. Söz gelimi 9-C sınıfındaki öğrenci sayısı 20 ile 25 arasındadır dediğimizde bu sınıftaki öğrenci sayısının alabileceği değerler 21, 22, 23 veya 24 olur. Fakat öğrenci sayısının 20,5 gibi ondalık içeren bir değer alması mümkün değildir. Bu nedenle, sınıf mevcudu süreksiz bir değişkendir.



Nitel değişkenlerin tamamı süreksizdir. Diğer taraftan nicel bir değişken sürekli veya süreksiz olabilir. Örnek vermek gerekirse cinsiyet, medeni durum, göz rengi nitel ve süreksiz değişkenlerdir. Ağırlık nicel sürekli bir değişken iken nüfus nicel süreksiz bir değişkendir.

2.3. Araştırmadaki Rollerine Göre Değişkenler

Değişkenlerin araştırmadaki rollerine göre kategorilere ayrılması araştırma deseninin doğru kurgulanması, değişkenler arasındaki ilişkilerin tespiti ve elde edilen bulguların doğru yorumlanması açısından oldukça önemlidir. Bilimsel araştırmalarda değişkenler, araştırma sürecinde üstlendikleri rollere göre bağımlı, bağımsız, kontrol, aracı ve düzenleyici değişkenler olarak sınıflandırılmaktadır.

2.3.1. Bağımsız Değişken

Bir araştırmada neden durumunda olan, bir başka deyişle bağımlı değişken üzerindeki etkisi incelenen değişkendir. Deneysel araştırmalarda bağımsız değişken araştırmacı tarafından manipüle edilirken betimsel ve ilişkisel araştırmalarda bağımsız değişken manipüle edilmeden mevcut durumu temsil eder.

2.3.2. Bağımlı Değişken

Bir araştırmada sonuç durumunda yer alan, diğer bir ifadeyle bağımsız değişkenden etkilenip etkilenmediği sınanan değişkendir. Bu bakımdan alabileceği değerlerin başka değişkenlere göre farklılaşıp farklılaşmadığı sorgulanan değişkenler olarak da tanımlanabilir.

2.3.3. Kontrol Değişkeni

Bağımlı değişken üzerindeki olası etkisini kontrol etmek için araştırma kapsamında sabit tutulan değişkendir. Tıpkı bağımsız değişken gibi kontrol değişkeninin de bağımlı değişkeni etkileme potansiyeli olabilir. Ancak bağımsız değişkenden faklı olarak araştırmacı kontrol değişkenin etkisini test etme amacı gütmez. Aksine bağımsız değişkenin bağımlı değişken üzerindeki etkisine dair daha doğru sonuçlara ulaşabilmek için kontrol değişkenin muhtemel etkilerini elimine etmeye çalışır. Bağımlı ve bağımsız değişken ile kontrol değişkeni aşağıda bir araştırma problemi üzerinden örneklendirilmiştir.

Örnek: Bir araştırmacı ölçme ve değerlendirme dersinde farklı öğretim yöntemlerinin (yüz-yüze öğretim ve uzaktan öğretim) öğrenci başarısı üzerindeki etkisini incelemeye dönük deneysel bir çalışma gerçekleştirmiştir.

Bu araştırmada;

- Bağımsız değişken öğretim yöntemidir. Araştırmacı, bir grupta yüz yüze öğretim uygularken diğer grupta uzaktan öğretime yer vermek suretiyle öğretim yöntemi değişkenini manipüle eder.
- Bağımlı değişken öğrencilerin başarı düzeyidir. Araştırmacı her iki gruptaki öğrencilerin sınavdan aldığı puanları ölçerek başarı değişkenine ilişkin verileri elde eder. Daha sonra başarının kullanılan öğretim yöntemine bağlı olarak değişip değişmediğini uygun istatistiksel analizler aracığıyla test eder.
- Öğrenci başarısını etkileyebilecek diğer değişkenler kontrol değişkenleri olarak alınabilir. Öğrencilerin sosyoekonomik düzeyi, yaşı, destek eğitimi alma durumu vb. kontrol değişkeni olarak düşünülebilir. Yukarıdaki araştırmada öğreticiden kaynaklı olası farklılıkları kontrol altına almak için iki sınıfta derslerin aynı öğretim üyesi tarafından yürütülmesi sağlanabilir. Böylece "öğretici" kontrol değişkeni olur. Deneysel araştırmalarda kontrol değişkenlerin sayısını artırmak bazı zorluklar içerse de ulaşılan sonuçların iç geçerliğini artırabilmek için kontrol süreçlerinin işletilmesi oldukça önemlidir.

2.3.4. Aracı Değişken

A, B ve C şeklinde üç değişken üzerinden yürütülen bir araştırmada A değişkeni B'yi, B değişkeni de C'yi etkiliyorsa B değişkeni A ile C arasındaki ilişkide aracı değişken olarak ifade edilir (Mackinnon, 2008). Yani aracı değişken bağımsız değişkenin bağımlı değişkeni nasıl/neden etkilediğinin anlaşılmasına katkıda bulunur. Örneğin; A değişkeni öğretim yöntemini, C değişkeni ise test puanlarını temsil etsin. Araştırmalar A'nın C'yi etkilediğini, bir başka deyişle test puanlarının kullanılan öğretim yöntemine bağlı olarak farklılaştığını gösteriyor olsun. Burada öğrenci motivasyonu, öğretim yöntemi ile başarı arasında aracı bir rol üstelenebilir. Kullanılan öğretim yönteminin öğrencilerin motivasyonunu, motivasyonun da başarıyı etkilediği zincirleme bir ilişki söz konusu olabilir (Johnson & Christensen, 2017). İşte öğrenci motivasyonu, bu ilişki zincirindeki aracı değişkendir.

2.3.5. Düzenleyici Değişken

Bağımlı ve bağımsız değişken arasındaki ilişkinin gücünü ve/veya yönünü etkileyen nitel ya da nicel özellikteki değişkendir (Baron & Kenny, 1986). Örneğin, herhangi bir sağlık sorunu olmayan kişilerde para ile mutluluk arasında pozitif yönlü güçlü bir ilişki gözlenirken kronik sağlık sorunu olan bireylerde para ile mutluluk arasında bir ilişki çıkmayabilir. Bu da sağlığın, para ile mutluluk arasındaki ilişkide düzenleyici role sahip olduğu anlamına gelir.



Sürekli ve süreksiz değişken ile nitel ve nicel değişken kavramları değişkenin doğasıyla ilgili özelliklerdir. Öte yandan bağımlı, bağımsız, aracı, düzenleyici değişken ile kontrol değişkeni kavramları değişkenin araştırmada nasıl konumlandırıldığıyla ilgilidir. Buna göre bir değişken bir araştırmada bağımsız değişken durumunda iken bir başka çalışmada bağımlı, bir diğer araştırmada ise kontrol değişkeni durumunda olabilir.

3. ÖLÇME VE ÖLÇME TÜRLERİ

Ölçme kavramı, eğitim alanında genel olarak sınavları çağrıştırır. Ancak ölçme yalnızca eğitime özgü bir kavram değildir. Eğitimde ölçme sadece sınavlarla yapılmadığı gibi diğer pek çok farklı alanda ve günlük hayatta oldukça farklı ölçme işlemleri gerçekleştirilebilmektedir (Mehrens & Lehman, 1991). Turgut (1997) ölçmeyi bir niteliğin gözlenip sonuçların sayı veya sembollerle ifade edilmesi şeklinde tanımlamaktadır. Bu tanımda nitelik yerine değişken sözcüğü konulursa ölçme "değişkenlere değer verme işlemi" olarak da ifade edilebilir. Miller ve arkadaşları (2009) ölçmeyi, bir sınava veya başka bir özelliğe belli kurallara göre sayılar atama şeklinde açıklamaktadır. Ölçülen özelliğin doğasına bağlı olarak doğrudan (temel), dolaylı ve türetilmiş olmak üzere üç ölçme türü vardır (Turgut & Baykul, 2010).

3.1. Doğrudan Ölçme

Doğrudan gözlemlerle, doğal birimler kullanılarak veya değişkenin kendi cinsinden birimi olan ölçme araçları yardımıyla gerçekleştirilen ölçme işlemidir. Bu anlamda doğrudan ölçmeler daha çok fiziksel özelliklerin ölçülmesinde veya sayma işlemlerine dayalı ölçmelerde karşımıza çıkar. Örneğin, sınıftaki öğrenci sayısı doğrudan ölçülür ve burada "kişi" gibi doğal bir birim vardır. Boy uzunluğu "cm" ile birimlenmiş ölçme aracıyla ölçülür ve burada ölçme aracının birimi ölçülen özelliğin birimi ile aynıdır. Bir nesnenin ağırlığının eşit kollu terazi ile belirlenmesi doğrudan ölçmeye yönelik bir başka örnektir. Sıvıların hacminin bir litre kabı yardımıyla ölçülmesi de doğrudan ölçmeye örnek olarak verilebilir. Doğrudan ölçme sonuçları genellikle daha az hata içerdiğinden daha geçerli ve güvenilir sonuçlar verir.

3.2. Dolaylı Ölçme

Dolaylı ölçmede ölçülen özellik, göstergesi olarak kullanılabilecek başka değişken ile ilişkilendirilir. Yani dolaylı ölçmede bir nitelik; doğrudan gözlemlerle veya birimi direkt olarak ölçmeye konu özellikle ilişkili ölçme araçlarıyla değil başka bir özelliğe dayalı olarak ölçülür. Hava sıcaklığının termometreler aracılığıyla ölçülmesi dolaylı ölçmeye yönelik tipik bir örnektir. Aslında ölçülen termometre içerisindeki civanın yüksekliğidir. Hava sıcaklığı artıkça civa genleşir ve termometre içerisindeki yüksekliği artar. Hava sıcaklığı düştüğünde ise civanın termometre içindeki yüksekliği düşer. Sıcaklık başka bir fiziksel özellikten, civanın hacmindeki değişimden yola çıkılarak ölçüldüğü için termometreyle sıcaklık ölçümü dolaylı bir ölçme işlemidir. Eğitimde ve psikolojide yapılan ölçmelerin çoğu doğrudan gözlenemez, dolaylı olarak ölçülebilir. Örneğin, zekâyı manyetik rezonans görüntüleme (MR) veya bilgisayarlı tomografi (BT) gibi yöntemler ile doğrudan gözlemek olanaksızdır. Zekâ ancak bireyin belli uyaranlar karşısında gösterdiği tepkilerden yola çıkılarak ölçülebilir. Bu anlamda zekânın çeşitli testler ve problemler kullanılarak ölçülmesi dolaylı bir ölçmedir. Benzer durum akademik başarı, kişilik, tutum, motivasyon, mutluluk vb. özellikler için de geçerlidir. Dolaylı ölçme doğrudan ölçmeye kıyasla hataya biraz daha açıktır.

3.3. Türetilmiş Ölçme

Doğrudan ölçülemeyen bir değişkenin birden fazla başka değişken yardımıyla belirli bir formüle dayalı olarak hesaplanmasıdır. Araçların süratinin ölçülmesi, aldığı yol ve bu yolu ne kadar sürede kat ettiğine dayalı olarak ölçülür. İki ölçme sonucu arasında matematiksel bir işlem yapılarak elde edildiği için süratin ölçümü türetilmiş bir ölçmedir. Benzer şekilde vücut kitle indeksinin ölçülmesi türetilmiş ölçmeye karşılık gelir. Çünkü vücut kitle indeksi, ağırlık (kilogram) ile boy uzunluğunun (metre) bir formüle yerleştirilmesiyle elde edilir. Üçgeninin alanın bulunması, yoğunluğun hesaplanması, ivmenin belirlenmesi, basıncın saptanması türetilmiş ölçmeye örnek teşkil eden durumlardır.

4. ÖLÇME KURALI

Ölçme sürecinde hangi büyüklüğe ne kadar değer atanacağına, ölçme işlemi yapılırken hangi koşulların sağlanması gerektiğine, değer verirken hangi sayı/sembol kümesinin kullanılacağına vb. ilkelere ölçme kuralı denir. "Her soru 5 puandır" veya "Dört yanlış bir doğru cevabı götürmektedir" sıkça karşılaştığımız ölçme kurallarıdır. Ölçme yapabilmek için ölçme kurallarını belirlemiş olmak gerekir.

5. ÖLÇÜM

Ölçme işleminden elde edilen sayısal değerler ya da semboller ölçüm olarak tanımlanır. Örneğin, "Mustafa sınavdan 60 puan aldı", "Neriman deneme sınavında birinci oldu", "Bugün hava güneşli" cümlelerinde geçen "60 puan", "birinci" ve "güneşli" ölçüm ifadeleridir.

6. ÖLÇMEDE SIFIR (BAŞLANGIÇ) NOKTASI

Ölçme sonuçlarını ifade ederken genellikle sayılardan yararlanılır. Büyüklüğüne dayanarak ölçülen özelliğe değer ataması yaparken kullanılan sayılardan biri de sıfırdır. Yokluk ifade edip etmemesine göre ölçmede iki ayrı sıfır noktası vardır. Bunlardan ilki mutlak, diğeri ise tanımlı sıfır değeridir.

6.1. Mutlak Sıfır

Gerçek sıfır olarak da ifade edilir ve yokluğu, yani hiçliği belirtir. Mutlak sıfır olan ölçeklerde sıfırın altında değerler bulunmaz. Uzunluk ve ağırlık ölçümlerinde ölçek mutlak sıfır ile başlar. Örneğin; bebek sıfır yaşındaysa henüz dünyada yaş almamış, doğmamış demektir. Aynı şekilde banka hesap bakiyemiz sıfır ise hesabımızda hiç para olmadığı anlaşılır.

6.2. Tanımlanmış Sıfır

Keyfi, bağıl, izafi veya göreceli sıfır olarak da adlandırılır. Bu tür ölçeklerde sıfır yokluk bildirmez, tanımlanmış bir değeri belirtir ve sıfırın altındaki değerler de kullanılır. Sıfır değerinin tanımı ölçeğin birimlerini oluşturmada belirleyicidir. Örneğin, santigrat derece (°C) birimiyle sıcaklık ölçümünde sıcaklık sıfırın altında da değerler alabilir. Yani derece santigratta sıfır yokluğu değil uzmanlar tarafından tanımlanmış bir noktayı (suyun donma noktasını) gösterir.

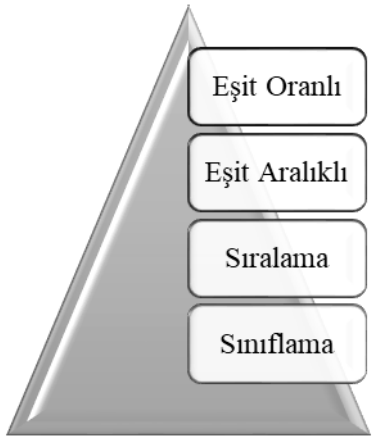
Kelvin termometresindeki sıfır tanımlanmış değil mutlak sıfırdır. Çünkü Kelvin cinsinden sıfır değeri enerjinin olmadığını ifade eder.

7. ÖLÇMEDE BİRİM

Ölçmede birim, ölçme aracının puanlanabilen en küçük parçası ya da ölçme sırasında işe koşulan en küçük standart miktar olarak tanımlanabilir. Özer-Özkan (2022) ise birimi ölçme aracının ölçme sonucuna anlamlı katkı sağlayabilen en küçük parçası şeklinde kavramsallaştırmaktadır. Söz gelimi; Ahmet'in boy uzunluğu 160 santimetre dediğimizde birim *santimetre*, masanın genişliği 10 karış dediğimizde birim *karış*, sepette 12 tane yumurta var dediğimizde birim *tane* ya da karpuz 5,3 kilogram dediğimizde birim *kilogram*dır. Ölçmede kullanılan birimler doğal ve yapay birim olmak üzere iki başlıkta incelenirken birimin sahip olması gereken özellikler genellik, eşitlik ve kullanışlılık olarak sıralanmaktadır.

Birim Türleri	Doğal Birim	Doğada kendiliğinden var olan uzman eliyle belirlenmemiş birimlerdir. Tane, adet, karış, adım, kulaç doğal birime verilebilecek örneklerdir.
	Yapay Birim	Bilimsel veya pratik ihtiyaçları karşılaşmak üzere uzmanlar tarafından tanımlanmış birimlerdir. Santimetre, kilogram, dakika, litre, amper, yapay birime ilişkin örneklerdir.
Birim Bulunması Gereken Özellikler	Genellik	Birimin ne kadar geniş gruplar için anlaşılır olduğuyla ilgilidir. Örneğin Euro, İsviçre Frank'ına kıyasla genellik özelliği daha yüksek bir birimdir. Uluslararası Birim Sistemi (SI), birimde evrenselliği sağlamak, daha açık ifade etmek gerekirse dünyada herkesin aynı özellik için aynı birimi kullanmasını teşvik etmek, böylece ölçmede olası karışıklıkların önüne geçmek amacıyla kurulmuştur.
	Eşitlik	Birimin gerek ölçme işlemi boyunca gerekse bir ölçmeden diğerine değişmeyip aynı kalmasıdır. Örneğin iki kişi bir masanın uzunluğunu karışlayarak ölçtüğünde farklı sonuçlar elde edilir. Çünkü karış eşitlik özelliği olmayan bir birimdir, karışın belirttiği uzunluk her zaman aynı değildir. Öte yandan santimetrenin bildirdiği uzunluk daima aynıdır. Bu anlamda, santimetre eşitlik özelliği olan bir birimdir.
	Kullanışlılık	Birimin ölçülen büyüklüğe uygun olmasıyla ilgilidir. Örneğin, Ankara ile Eskişehir arasındaki uzaklığı ifade ederken birim olarak kilometre kullanılırken bireylerin boy uzunluklarının metre ya da santimetre cinsinden ifade edilmesi birimin kullanışlılık özelliği ile ilgilidir.

8. ÖLÇEK VE ÖLÇEK TÜRLERİ



Ölçek, ölçme ve değerlendirme literatüründe farklı anlamlar yüklenen bir sözcüktür. Ölçeğin ilk anlamı "birim" olarak ifade edilebilir. İkinci anlamı bir tür ölçme aracını belirtir ve bir psikolojik yapıyı ölçmek amacıyla hazırlanan, hedef yapının davranışsal göstergeleri olabilecek maddeler içeren ve içerdiği maddelerden toplam ya da ortalama bir puan alınabilen ölçme araçları ölçek olarak tanımlanır (İlhan ve ark., 2020). Ölçeğin üçüncü anlamı ise ölçme sonuçlarının düzeyiyle (levels of measurement) ilgilidir. Ölçek kavramı bu yönüyle ölçme sonuçlarının ölçülen özellik hakkında ne kadar yüzeysel ya da derin bilgi verebildiğini yansıtır. Ölçme sonuçları üzerinde yapılabilecek matematiksel ve istatistiksel işlemler ölçümlerin hangi ölçek düzeyinde elde edildiğine göre farklılaşır.

Stanley Smith Stevens 1946 tarihinde yayınladığı *Ölçme Düzeyleri Teorisi Üzerine (On the Theory of Scales of Measurement)* başlıklı makalesinde sınıflama (nominal), sıralama (ordinal), eşit aralıklı (interval) ve oranlı (ratio) olmak üzere dört ölçme düzeyi tanımlamıştır. Bu düzeyler hiyerarşik bir yapıya sahip olup daha üst düzey ölçekler alt seviyedeki ölçeklere ait tüm özellikleri içerir (Bandalos, 2018). Sıralanan ölçek düzeyleri aşağıda başlıklar halinde açıklanmıştır.

8.1. Sınıflama Ölçeği

Varlıkların ölçülen özellik açısından benzerlik ve farklılıklarına göre kategorilere ayrılmasıdır. Örneğin, cinsiyeti ölçtüğümüzde kadın-erkek şeklinde bireyleri gruplara ayırırız, dolayısıyla sınıflama düzeyinde bir ölçme işlemi gerçekleştirmiş oluruz. Benzer şekilde medeni durum, meslek, doğum yeri, saç rengi ya da kan grubu ölçülürken kişiler ölçmeye konu özellik açısından gruplara ayrılır ve yapılan ölçme işlemi sınıflama düzeyine karşılık gelir. Bu tür ölçmeler sonucunda oluşan kategoriler birbirini dışlayan bir yapıya sahiptir. Daha açık bir anlatımla, kişinin/nesnenin aynı anda birden fazla kategoride yer almasına imkân yoktur. Mesela bireyin kan grubu Rh (+) ya da Rh (-)'dir. Bireyin aynı anda hem Rh (+) hem de Rh (-) kategorisinde yer alması gibi bir durum söz konusu değildir.

Sınıflama düzeyindeki ölçümlere sayısal değerler atanması da mümkündür. Ancak atanan sayılar sadece farklı kategorileri birbirinden ayırt etmeye hizmet eder, büyüklük-küçüklük bildirmez. Örneğin, kadınlar 1, erkekler 2 olarak kodlandığında kullanılan sayısal değerler miktar ifade etmez; yalnızca kategorileri ayırt etmeye yarar. Bu anlamda medeni durumu ölçerken evlileri 1, bekârları 2 şeklinde kodlamak ile evliler için 10,

bekârlar için 20 kodunu kullanmak arasında bir fark yoktur. Vatandaşlık numaraları ile illerin telefon veya plaka kodları sayısal değerlerin miktar bildirmeyip yalnızca etiket görevi üstlendiği örneklerdir.

Sınıflama düzeyindeki ölçmelerde frekans, yüzde ve mod/tepe değer (veri setinde en yüksek frekans sahip olan, diğer bir deyişle en fazla tekrar eden gözlem) belirlenebilir. Bir beden eğitimi öğretmeni okulun spor salonundaki topları futbol, voleybol, basketbol ve hentbol şeklinde gruplandırmış olsun. Bu sınıflama işleminin ardından öğretmen spor salonunda her bir top türünden kaç tane bulunduğunu saptayabilir (frekans hesabı) ve bunların toplam top sayısı içerisindeki yüzdelerini tespit edebilir. Ayrıca spor salonunda en fazla hangi türde topun bulunduğu ortaya koyabilir yani modu belirleyebilir. Sınıflama düzeyinde ölçmelerde ki-kare testi, McNemar testi ve phi korelasyonu gibi kategorilerdeki frekans değerlerine dayalı istatistikler de hesaplanabilir. Buna karşın büyüklük-küçüklük sıralaması yapılamadığından medyan (ortanca değer) bulunamaz. Sınıflama ölçeğindeki verilerle toplama, çıkarma, çarpma ya da bölme yapılamaz. İlkokul sıralarında sıkça duyduğumuz "elmalar ile armutları toplayamazsınız" yargısı buradan gelir.

8.2. Sıralama Ölçeği

Varlıkların ölçülen özelliğe sahip oluş düzeylerine göre sıralamaya tabi tutulmasıdır. Bir sınıftaki öğrencilerin boy uzunluklarına göre en kısadan en uzuna doğru dizilmesi ya da birinci, ikinci ve üçüncü şeklinde bir koşudaki derecelerin belirlenmesi sıralama düzeyindeki ölçmelere verilebilecek örneklerdir. Sıralama ölçeğinde kullanılan sayılar büyüklük-küçüklük karşılaştırması yapmak için kullanılabilir. Fakat ardışık ölçek noktaları arasındaki uzaklıkların eşit olduğu gibi bir varsayım yoktur. Bir yarışa ait sonuçları ele alalım. Yarışı birinci tamamlayan kişinin ikinciden daha hızlı olduğunu biliriz ama ne kadar hızlı olduğuna dair fikrimiz olmaz. Çünkü sıralı ölçümler, ölçek noktaları arasındaki farkın ne kadar olduğu hakkında bir şey söylemez. Aynı sebeple yarışı birinci ile ikinci tamamlayanlar arasındaki zaman farkının, ikinci ve üçüncü bitirenler arasındaki zaman farkına eşit olup olmadığını bilmemiz mümkün değildir (Bandalos, 2018).

Sıralama düzeyindeki ölçmelerde genel olarak sayma sayıları kullanılır, sıfır değerinin bir anlamı yoktur. Tabi her zaman sayısal değerlere yer verme zorunluluğu olmadığını da belirtmek gerekir. Mesela kıyafetler ile içecekler için kullanılan *small* (küçük), *medium* (orta), *large* (büyük) ve *x-large* (çok büyük) etiketleri herhangi bir sayısal değer içermese de büyüklük-küçüklük sıralaması bildirir. Dolayısıyla sıralama düzeyine karşılık gelen ölçmelerdir. Yine ilkokul, ortaokul, lise ve üniversite şeklinde bireylerin eğitim düzeylerinin belirlenmesi ya da A, B, C, D vb. tarzdaki harf notları sıralama düzeyinde olan ölçme sonuçlarıdır. Sıralama düzeyindeki ölçümlerde bir önceki ölçek düzeyinde yapılabilen frekans, yüzde ve mod hesabının yanı sıra medyan belirlenebilmekte, ayrıca Spearman sıra farkları korelasyonu hesaplanabilmektedir. Ama sınıflama ölçeğinde olduğu gibi bu ölçek düzeyinde de dört işlem (toplama, çıkarma, çarpma ve bölme) gerçekleştirilememektedir.

8.3. Eşit Aralık Ölçeği

Eşit aralık düzeyindeki ölçmelerde ardışık ölçek noktaları arasındaki uzaklıklar eşittir ve tanımlanmış bir sıfır noktası vardır. Termometre ile sıcaklık ölçümü eşit aralık düzeyindeki ölçmelere ilişkin tipik bir örnektir. Burada sıfır tanımlanmış olup Celcius ölçeği için başka, Fahrenheit ölçeği için daha başka bir noktadır. Hava sıcaklığının 0 °C olarak ölçülmesi sıcaklığın olmadığını bildirmez uzmanlar tarafından tanımlanmış bir noktayı (1 atmosfer basınç altında suyun donma noktasını) ifade eder. Dahası 10 °C ile 11 °C arasındaki fark 25 °C ile 26 °C arasındaki farka eşittir, yani birbirini izleyen ölçek noktaları arası mesafeler aynıdır. Herhangi bir yerleşim yerinin rakımının belirlenmesi eşit aralık düzeyindeki ölçmelere verebilecek bir diğer örnektir. Rakım ölçülürken deniz seviyesi sıfır olarak tanımlanır. Benzer şekilde takvim ile zamanı ölçerken eşit aralık düzeyinde bir ölçme işlemi gerçekleştirilir ve sıfır noktası kullanılan takvime göre farklılaşır. Hicri takvimde Hz. Muhammed'in Mekke'den Medine'ye hicreti sıfır noktası olarak kabul edilirken miladi takvimde Hz. İsa'nın doğumu sıfır noktası olarak alınır.

Herhangi bir özellik için birimleri farklı olan eşit aralık düzeyindeki iki ölçme sonucu, y = ax + b biçimindeki doğrusal denklemlerle birbirine dönüştürülebilir (Bandalos, 2018). Örneğin Celcius termometresiyle elde edilen sıcaklık değeri Fahrenheit termometresine dönüştürülebilir. Benzer şekilde kitabın ilerleyen bölümlerinde karşımıza çıkacak *z* ve *T* standart puanları arasında dönüşümler yapılabilir ($T = 10z + 50$).

Eğitimde ve psikolojide yapılan ölçmelerin büyük kısmı eşit aralık düzeyindedir. Çoktan seçmeli bir test aracılığıyla öğrencilerin matematik başarılarını ölçtüğümüzde ya da bir zekâ testi uygulayarak bireylerin IQ puanlarını belirlediğimizde elde ettiğimiz veriler eşit aralıklı ölçek düzeyine karşılık gelir. Likert tipi ölçeklerle toplanan veriler de teorik olarak sıralama düzeyinde bulunmasına rağmen uygulamada eşit aralık düzeyindeymiş gibi işlem görür. Şöyle ki Likert tipi bir ölçekte *Hiç katılmıyorum*, *Katılmıyorum*, *Biraz katılıyorum*, *Katılıyorum* ve *Tamamen Katılıyorum* cevap kategorileri arasındaki uzaklıklar gerçekte eşit değildir. "Katılıyorum" ifadesi "Biraz Katılıyorum" seçeneğine kıyasla; "Tamamen Katılıyorum" ifadesi de "Katılıyorum" seçeneğine kıyasla katılma eğiliminin daha fazla olduğunu bildirir. Fakat *Tamamen Katılıyorum* ile *Katılıyorum* noktaları arasındaki fark, *Katılıyorum* ile *Biraz Katılıyorum* kategorileri arasındaki farka eşit olmayabilir. Ölçek kategorileri için kullanılan etiketlere bağlı olarak ardışık noktalar arasındaki uzaklıklar farklı yorumlanabileceği gibi kategoriler arasındaki mesafelerin farklı cevaplayıcılar tarafından farklı şekillerde algılanması da ihtimal dâhilindedir. Ancak uygulamada bu ihtimal göz ardı edilir ve ardışık cevap kategorileri arasındaki uzaklıkların aynı olduğu varsayılır. Başarı, zekâ, tutum vb. özellikler eşit aralık düzeyinde ölçüldüğünden bu özellikler için elde edilen sıfır değeri yokluğu ifade etmez. Yani bir öğrencinin sayısal yetenek testinden sıfır puan alması bu öğrencinin sayısal yeteneğinin hiç olmadığı anlamına gelmez.

Eşit aralık düzeyindeki ölçmelerde sıralı verilerde yapılabilen frekans, yüzde, mod ve medyan hesaplamalarına ek olarak toplama ve çıkarma işlemleri de gerçekleştirilebilir. Bununla birlikte mutlak bir sıfır noktası bulunmadığından eşit aralık düzeyindeki verilerde çarpma ve bölme işlemi uygulanamaz. Katıldıkları sınavdan sırasıyla 80 ve 40 alan Hale ve Uğur isimli iki öğrenci düşünelim. Sınav puanları eşit aralık düzeyinde olduğundan Hale'nin notu Uğur'unkinden 40 puan yüksektir diyebiliriz. Ancak bu puanları oranlayıp Uğur Hale'nin yarısı kadar başarılı olmuştur demek yanlış olur.

8.4. Oran Ölçeği (Eşit Oranlı Ölçek)

Daha çok fiziksel özellikleri ölçerken karşımıza çıkan bu ölçek düzeyinde (Stevens, 1946) eşit aralık ölçeğinde olduğu gibi birimler arası uzaklıklar eşittir. Bununla beraber, eşit aralıklı ölçekteki tanımlanmış sıfırın yerini eşit oranlı ölçekte gerçek sıfır (mutlak sıfır) alır. Sınıf mevcudunu ölçtüğümüzü farz edelim. Bu ölçme işleminde sıfır sınıfta hiç öğrenci bulunmadığını ifade eder, dolayısıyla mutlak bir sıfır söz konusudur. Ayrıca birimler arası uzaklıklar aynıdır, diğer bir deyişle 17 ile 18 arasındaki uzaklık, 34 ile 35 arasındaki farka eşdeğerdir. Buna bağlı olarak sınıf mevcudunu ölçtüğümüzde elde ettiğimiz sonuç eşit oranlı ölçek düzeyinde olur. Bireylerin aylık gelirini hesapladığımızda, mezura aracılığıyla boy uzunluklarını belirlediğimizde, baskül aracılığıyla ağırlıklarını tespit ettiğimizde ya da bir kompozisyondaki yazım hatalarının saydığımızda ulaştığımız veriler eşit oranlı ölçek düzeyine karşılık gelir. Oran ölçeğindeki verilerde her türlü matematiksel ve istatistiksel işlem gerçekleştirilebilir.



- Teorik olarak eşit aralıklı ve oranlı ölçek arasında ayrım bulunmakla birlikte uygulamada bu iki ölçek düzeyindeki ölçme sonuçları sürekli veri kabul edilir ve aynı istatistiksel işlemlere tabi tutulur. Nitekim istatistik programlarını incelediğimizde bu iki kategorideki ölçümler için ayrı tanımlamalar yapmadıkları, aynı etiketleri kullandıkları görülür. Söz gelimi en sık başvurulan istatistik programlarından olan SPSS ile günümüzde kullanımı giderek yaygınlaşan JASP eşit aralıklı ve oranlı ölçek düzeyinde olan verileri "Scale" şeklinde etiketler. Giderek popülerleşen bir başka istatistiksel yazılım olan JAMOVI ise aralık ve oran ölçeğindeki verileri "Continuous" olarak ele alır.
- Ölçek düzeyleri hiyerarşik bir yapıya sahip olduğundan daha üst düzey ölçekler alt düzey ölçeklere kolaylıkla dönüştürebilir. Dört kişinin ağırlığını 100, 70, 80 ve 50 kg olarak ölçtüğümüzü varsayalım. Burada yokluğu bildiren gerçek bir sıfır noktası vardır ve veriler oranlı ölçek düzeyindedir. Diyelim ki 70'i sıfır noktası olarak tanımladık. Artık sıfır keyfi olarak tanımlanmış bir değerdir ve yokluğu ifade etmez. Dolayısıyla veriler 30, 0, 10 ve -20 olarak ifade edilir ve eşit aralıklı ölçek düzeyine iner. Bu kişileri ağırlıklarına göre sıraladığımızda 1, 3, 2 ve 4 şeklinde dizilir ve veriler sıralama düzeyine gelir. Ölçeklerin alt düzeyden üst düzeye getirilmesi ise bazen zor, çoğu zaman da olanaksızdır. Örneğin, sıralama düzeyindeki verilerin bir kısmı çeşitli istatistiksel süreçlerden geçirilerek eşit aralıklı ölçek düzeyine çıkarılabilir. Ancak eşit aralık düzeyindeki verileri oranlı ölçek düzeyine getirmek mümkün değildir.

9. DEĞİŞKENLER ARASI İLİŞKİ: KORELASYON

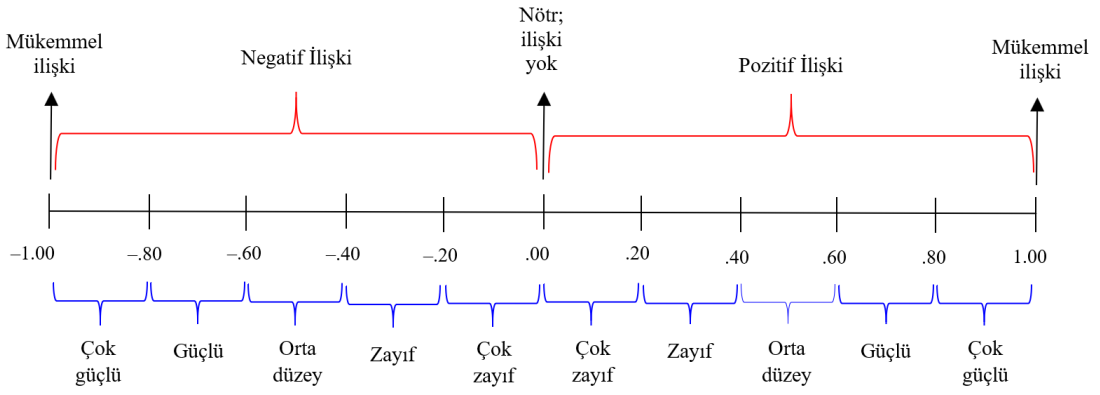
Korelasyon iki ya da daha fazla değişken arasındaki ilişkinin derecesi ile yönünü tanımlamak amacıyla kullanılan istatistiksel bir yöntemdir. Korelasyon katsayısına bakılarak değişkenlerin birlikte değişimi hakkında yargıya varılabilmektedir. Daha açık bir anlatımla korelasyon bir değişkenin değeri artarken veya azalırken diğer değişkenin nasıl değiştiği hakkında bilgi vermektedir. Eğitimde ve psikolojide yapılan pek çok araştırmada değişkenler arası ilişkiler incelenir. Örneğin, sınav kaygısı ile akademik başarı arasındaki ilişkiyi incelemek için korelasyon analizine başvurabilir, böylece kaygı artıkça veya azaldıkça başarıda nasıl bir değişim meydana geldiği ortaya konulabilir. Benzer şekilde öğrencinin herhangi bir dersteki başarısının; zekâsı, günlük ders çalışma süresi, okula yönelik tutumu, ebeveynlerinin eğitim düzeyi, ailesinin sosyoekonomik durumu vb. değişkenler ile ilişkisini incelemek için korelasyon analizi yapılabilir. Korelasyon analizi sonucunda korelasyon katsayısı olarak adlandırılan bir değer elde edilir. Bu katsayıya ilişkin temel özellikler aşağıda sıralanmıştır.

- ⇨ Korelasyon katsayısı r sembolü ile gösterilir ve –1 ile +1 arasında değişen değer alabilir ($-1 \leq r \leq 1$).
- ⇨ +1 korelasyon değeri değişkenlerin birinde bir birimlik artış olduğunda diğerinde oransal bir artış gözleneceği anlamına gelir. Aynı şekilde birinci değişkendeki bir birimlik azalmaya ikinci değişkende oransal bir azalış eşlik eder. Bu korelasyon katsayısı, daha çok fen bilimlerinde iki değişken arasındaki ilişkiyi etkileyebilecek diğer değişkenler sabit tutulduğunda gözlenir. Örneğin, kapalı bir kaptaki gazın hacmi ve mol sayısı sabit tutulduğunda basıncı ile sıcaklığı arasındaki korelasyonun değeri +1 olur. Bu korelasyon katsayısı, mükemmel bir ilişkiyi temsil eder.
- ⇨ –1 korelasyon değeri, ilk değişkendeki bir birimlik artışa diğer değişkende oransal bir azalışın eşlik edeceğini ima eder. Yukarıdaki örnek için konuşacak olursak kapalı bir kaptaki gazın mol sayısı ve sıcaklığı sabit tutulduğunda hacmi ile basıncı arasındaki korelasyon katsayısı –1 çıkar. Bu korelasyon, mükemmel bir ilişkiyi temsil eder.
- ⇨ 0 korelasyon değeri değişkenler arasında ilişki olmadığı anlamına gelir. Öğrencilerin ayakkabı numarası ile günlük ders çalışma süresi arasında sıfır ya da sıfıra yakın korelasyon katsayısı bulunması beklenir.
- ⇨ Değişkenlerden biri artarken diğeri de artma ve biri azalırken diğeri de azalma eğilimi gösteriyorsa bu iki değişken arasındaki ilişkiye ait korelasyon katsayısı pozitif çıkar ($0 < r \leq +1$). Örneğin, boy uzunluğu ile ayakkabı numarası arasında pozitif korelasyon olması beklenir.
- ⇨ Değişkenlerden birinin değeri artarken diğerinde azalma gözleniyorsa bu iki değişken arasında negatif bir korelasyon var demektir ($-1 \leq r < 0$). Örneğin, öğrencilerin gün içinde sosyal medyada geçirdikleri süre ile ders çalışmaya ayırdıkları süre arasında negatif korelasyon olması beklenir.

⇨ Korelasyonun gücü yorumlanırken elde edilen katsayının mutlak değeri esas alınır. Mutlak değerce 1'e yakın katsayılar daha güçlü ilişkiyi ifade ederken 0'a yakın değerler ilişkinin zayıf olduğunu bildirir. Salkind (2010) mutlak değerce 0 ile 0,20 arasındaki korelasyon katsayılarının çok zayıf, 0,20 ile 0,40 arasındaki değerlerin zayıf, 0,40 ile 0,60 arasındaki değerlerin orta düzey, 0,60 ile 0,80 arasındaki değerlerin güçlü ve 0,80 ile 1 arasındaki değerlerin çok güçlü ilişkiye işaret ettiğini belirtmektedir. İlhan ve arkadaşları (2022), sıralanan aralıklardan yararlanarak korelasyon katsayısının yorumlanmasına ilişkin ilkeleri Şekil 2'deki gibi görselleştirmiştir.

Şekil 2

Korelasyon Katsayısının Yorumlanması





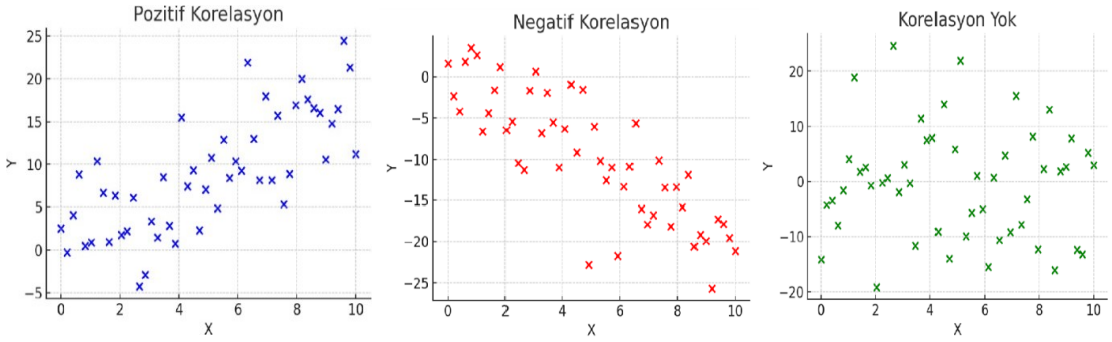
Korelasyon katsayısı iki değişken arasındaki ilişkinin nedenselliğine kanıt oluşturmaz. Örneğin, ilköğretime devam eden çocukların ayakkabı numaraları ile zekâ düzeyleri arasında pozitif yönlü güçlü bir korelasyon hesaplanabilir. Fakat bu korelasyondan hareketle bahse konu iki değişken arasında sebep–sonuç ilişkisi bulunduğu çıkarımına varılamaz. Hesaplanan korelasyona kaynaklık eden hem ayakkabı numarasının hem de zekâ düzeyinin yaş değişkeninden etkilenmesidir (Tan, 2015).

9.1. Saçılım Grafikleri

Değişkenlerden biri yatay (X), diğeri dikey (Y) eksene yerleştirilerek iki değişken arasındaki ilişki koordinat düzleminde üzerinde görsel olarak sunulabilir. Bu şekilde oluşturulan grafikler saçılım grafiği olarak isimlendirilir. Saçılım grafiğindeki noktalar eğimi 45 derece olan doğruya ne kadar yakın kümelenirse ilişki o kadar güçlü demektir. Korelasyonun işaretine göre saçılma grafikleri şekildeki gibi yorumlanır.

Şekil 3

Korelasyonun İşareti ile X ve Y Değişkenlerinin Saçılma Grafiği



9.2. Korelasyon Katsayısının Hesaplanması ve Korelasyon Teknikleri

Değişkenlerin ölçek düzeyine ve dağılımına göre iki değişken arasındaki ilişkiyi incelerken başvurulabilecek farklı korelasyon teknikleri bulunmaktadır. Korelasyon katsayısını hesaplarken sıkça kullanılan teknikler bir Tablo 1'de sunulmuştur.

Tablo 1

Değişkenlerin Ölçek Düzeyine Göre Kullanılan Korelasyon Teknikleri

1. Değişken	2. Değişken	Korelasyon Tekniği
Eşit Aralık/Oran	Eşit Aralık/Oran	Pearson Momentler Çarpımı
Eşit Aralık/Oran	İki Kategorili Sınıflama (Doğal)	Nokta Çift Serili
Eşit Aralık/Oran	İki Kategorili Sınıflama (Yapay)	Çift Serili
Sıralama	Sıralama	Kendall Tau/Spearman Sıra Farkları
İki Kategorili Sınıflama (Yapay)	İki Kategorili Sınıflama (Yapay)	Tetrakorik
İki Kategorili Sınıflama (Doğal)	İki Kategorili Sınıflama (Doğal)	Fi (Dörtlü)
Çok kategorili	Çok kategorili	Cramer V

[*] *İki değişken de en az eşit aralıklı ölçek düzeyinde yer almasına rağmen bu değişkenlere ait veriler normal dağılım göstermiyorsa Pearson korelasyonu yerine Kendall tau ya da Spearman sıra farkları korelasyonu kullanılır.*

Ölçek düzeyi üst kategoride olan bir değişkenin ölçek düzeyi alt kategoriye indirilerek yukarıdaki korelasyon tekniklerinden uygun olanı kullanılabilir. Örneğin, değişkenlerden biri eşit aralık/oran ölçeğinde iken diğer değişken sıralama düzeyinde ise aralık/oran ölçeğindeki değişken sıralama düzeyine indirilerek Sperman sıra farkları korelasyonu hesaplanabilir. En temel ve en sık kullanılan korelasyon tekniği şüphesiz Pearson momentler çarpımı korelasyonudur. Pearson korelasyonu her iki değişken de en az eşit aralıklı ölçek düzeyinde ise kullanılır ve aşağıdaki eşitlik yardımıyla hesaplanır.

$$r_{xy} = \frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum(x_i-\bar{x})^2 \sum(y_i-\bar{y})^2}}$$

r_{xy} = Korelasyon katsayısı

x_i = Bireylerin birinci değişken ölçümü

y_i = Bireylerin ikinci değişken ölçümü

$\sum$ = Toplam sembolü

$\bar{x}$ = Birinci değişkenin aritmetik ortalaması

$\bar{y}$= İkinci değişkenin aritmetik ortalaması

Eşitlik 1

Tablo 2'de 10 kişilik bir öğrenci grubunun Türkçe ve matematik dersi başarılarının 10 üzerinden ölçümleri verilmiştir. Bu öğrenci grubunun Türkçe ve matematik dersi başarıları arasındaki ilişkiyi Pearson momentler çarpımı korelasyonu ile hesaplayalım.

Tablo 2

Türkçe ve Matematik Dersi Başarı Puanları Arasındaki İlişkiye Ait Pearson Momentler Çarpımı Korelasyonunun Hesaplanması

Bireyler	X (Türkçe Ders Başarısı)	Y (Matematik Ders Başarısı)	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})^2$	$(y_i - \bar{y})^2$	$(x_i - \bar{x})(y_i - \bar{y})$
Ahmet	4	8	-1	1	1	1	-1
Musa	7	9	2	2	4	4	4
Hatice	3	5	-2	-2	4	4	4
Salih	8	10	3	3	9	9	9
Sena	2	6	-3	-1	9	1	3
İrfan	6	8	1	1	1	1	1
Lale	4	5	-1	-2	1	4	2
Feyza	8	6	3	-1	9	1	-3
Gülden	7	8	2	1	4	1	2
İbrahim	1	5	-4	-2	16	4	8
Ortalama	5	7	**Toplam**		58	30	29

$$r_{xy} = \frac{29}{\sqrt{(58\times(30)}} = \frac{29}{41,71} = 0,70$$

olarak elde edilir. Bu da öğrencilerin Türkçe ve Matematik deri başarıları arasında pozitif yönlü güçlü bir ilişki olduğunu yansıtır.

9.3. Korelasyon Matrisi

Araştırmacılar sadece iki değişken arasındaki ilişkiyi incelediklerinde hesapladıkları korelasyon katsayısını tablo biçiminde sunmak yerine genelde metin içerisinde verirler. Diğer taraftan çok sayıda değişken arasındaki ikili ilişkilere odaklandıklarında elde ettikleri korelasyon katsayılarını korelasyon matrisi şeklinde isimlendirilen tablolar aracılığıyla sunarlar (Huck, 2012). Bu durumda *n* sayıda değişken arasındaki ikili ilişkilere bakılıyorsa korelasyon matrisinde rapor edilecek korelasyon katsayılarının

sayısı $n(n-1)/2$ kadar olur. Örneğin, dört değişken arasındaki ikili ilişkilerin test edilmesine yönelik bir araştırma yapıyorsak korelasyon matrisinde vereceğimiz korelasyon katsayılarının sayısı 6'ya karşılık gelir (4×3/2=6, bkz. Tablo 3). Tablo 3 algılanan akademik destek, akademik öz-yeterlik algısı, sınav kaygısı ve akademik başarı arasındaki ilişkilerin incelendiği bir çalışmaya ait korelasyon matrisini içermektedir.

Tablo 3

Korelasyon Matrisi Örneği

Değişkenler	1	2	3	4
1. Algılanan akademik destek	1,00			
2. Akademik öz-yeterlik algısı	0,241	1,00		
3. Sınav kaygısı	–0,234	–0,373	1,00	
4. Akademik başarı	0,254	0,515	–0,340	1,00

Tablo 3'teki korelasyon matrisinde görüldüğü üzere, tablonun satır ve sütunlarında aynı değişkenler yer almaktadır. Ayrıca, köşegen üzerindeki hücrelerde 1.00 değerinin bulunduğu dikkat çekmektedir. Bu değerler bir değişkenin kendisiyle olan korelasyonunu ifade etmekte, bu sebeple tablo yorumlanırken bu değerler hesaba katılmamaktadır. Tabloda göze çarpan bir diğer husus, köşegenin üst tarafında kalan hücrelerin herhangi bir sayısal değer içermemesidir. Buna kaynaklık eden faktör, korelasyon matrisinin esas köşegene göre simetrik olmasıdır. Yani köşegenin alt ve üst tarafındaki değerler bir cismin aynadaki yansıması gibidir. Buna bağlı olarak matriste tüm hücreler verilmemekte; köşegenin ya alt ya da üst tarafındaki değerler sunulmaktadır. Tablo 3'ün herhangi bir hücresindeki korelasyon katsayısı, bulunduğu satır ve sütunda yer alan iki değişken arasındaki ilişkiyi temsil etmektedir. Örneğin, r= –0,234 değerinin yer aldığı hücrenin bulunduğu satırda sınav kaygısı, sütunda ise 1 numaralı değişken olan algılanan akademik destek vardır. Buna göre "–0,234" değeri sınav kaygısı ile algılanan akademik destek arasındaki korelasyonu ifade etmekte ve bu değişkenler arasında negatif yönlü zayıf bir ilişki olduğu şeklinde yorumlanmaktadır.

10. ÖLÇÜT

Ölçme sonuçlarıyla karşılaştırılarak yargıya varmak için kullanılan sınır değer ölçüt olarak tanımlanır. Bundan dolayı ölçüt; ölçme işleminin bir sonuca bağlanmasında, yorum yapılmasında veya karar verilmesinde temel alınacak kriter olarak da ifade edilebilir. Ölçüt mutlak ve bağıl ölçüt olmak üzere iki başlıkta ele alınır.

10.1. Mutlak Ölçüt

Ölçülen gruptan etkilenmeyen ve kesin/net değeri ölçme işlemi tamamlanmadan önce belirlenebilen ölçüt türüdür. Örneğin, "Sınavdan 60'ın üstüne not alan dersi geçer." ifadesi mutlak bir ölçüttür. Sınıf ortalaması kaç olursa olsun öğrenci 60 veya üzerinde bir not alırsa dersten geçer, 60'ın altında bir not alması halinde ise dersten kalır. Mutlak ölçütte "Her koyun kendi bacağından asılır." anlayışı vardır. Yeterlik sınavlarında ve standart belirleme çalışmalarında mutlak ölçüt tercih edilmelidir.

10.2. Bağıl Ölçüt

Ölçülen gruba göre şekillenen ve net değeri ancak ölçme işlemi tamamlandıktan sonra belli olan ölçüt türüdür. Örneğin, "Sınavdan sınıf ortalamasının üstünde not alanlar dersi geçer." ifadesi bağıl bir ölçüttür. Böyle bir durumda sınıf ortalamasına bağlı olarak öğrenci 40 alıp dersi geçebilir ya da 80 alıp dersten kalabilir. Çok sayıda başvuran aday arasından az sayıda kişinin seçileceği durumlarda bağıl ölçütün kullanılması uygun olur.



- Kimi durumlarda mutlak ve bağıl ölçüt bir arada kullanılabilir. Örneğin, "Yazılı sınavda 250 baraj puanını aşan ve en yüksek puanı alan ilk 50 kişi mülakata katılmaya hak kazanır" ifadesi hem mutlak hem de bağıl ölçüt içerir.
- Ölçüt, ölçme için değil değerlendirme yapabilmek için gereklidir. Diğer bir deyişle ölçüt belirlenmeden ölçme yapılabilir ama ölçüt olmaksızın değerlendirme yapılabilmesi mümkün değildir.

11. DEĞERLENDİRME

Eğitimde ve diğer pek çok alanda değerlendirme önemli bir yer tutar. Hangi kurum olursa olsun verimliliği sağlamanın yolu belli aralıklarla değerlendirme yapmaktan geçer. Değerlendirme yapılmadığı takdirde yürütülen işte verimlilik beklemek gerçekçi olmaz. Eğitim programlarının ya da gerçekleştirilen öğretim faaliyetlerinin verimliliği ancak değerlendirme süreçleriyle ortaya konulabilir. Öğrencilerin öğretim programında yer verilen öğrenme çıktılarına ne düzeyde eriştiğini ve kazandırmayı hedeflediğimiz becerileri ne düzeyde edindiğini anlamanın yolu değerlendirme yapmaktır. Değerlendirme ölçme sonuçlarını bir ölçütle ya da ölçüt takımıyla karşılaştırarak karar vermeyi ifade eder. Bu tanımdan anlaşılacağı üzere değerlendirmenin "ölçme sonucu, ölçüt ve karar" olmak üzere üç öğesi vardır.



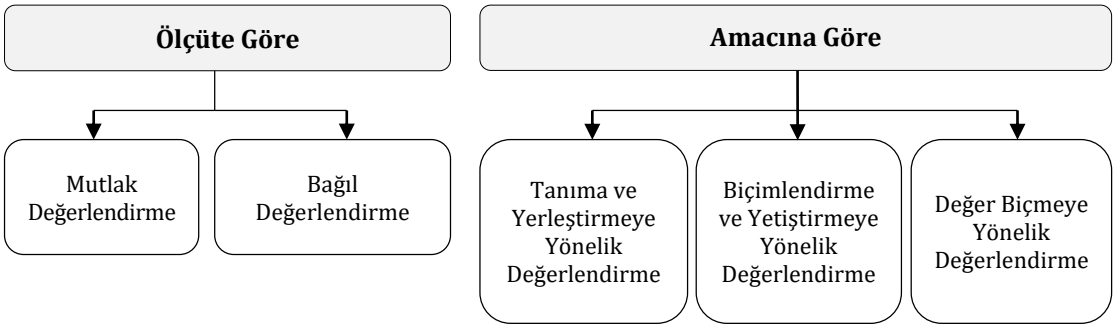
Örneğin; bir doktor, işyeri hekimliği sertifikası alabilmek için Ölçme, Seçme ve Yerleştirme Merkezi Başkanlığı (ÖSYM) tarafından yapılan İş Yeri Hekimliği ve İş Güvenliği Uzmanlığı Sınavı'na (İSG) katılmış olsun. Bu doktorun sınavdan 75 puan aldığını düşünelim. İSG kılavuzundaki bilgilere göre iş yeri hekimliği sertifikası alanından sınava giren adayların başarılı sayılabilmeleri için sınavdan 70 ve üzerinde bir puan almaları gerekir. Buna göre adayın puanı, ölçüt olarak tanımlanan değeri aşmaktadır. Dolaysıyla bu adayın başarılı olduğu ve işyeri hekimliği sertifikası almaya hak kazandığı sonucuna varılır. Değerlendirme ve değerlendirme türleri 12. bölümde ayrıntılı biçimde ele alınmıştır. Ancak burada da değerlendirme türlerine kısaca değinilmiştir.

11.1. Değerlendirme Türleri

Değerlendirmeye ilişkin sınıflama değerlendirmede kullanılan ölçüt ile değerlendirmenin amacına göre yapılır. Bu iki başlık dikkate alındığında değerlendirme Şekil 4'teki gibi sınıflandırılır.

Şekil 4

Değerlendirme Türleri



11.1.1. Ölçütüne Göre Değerlendirme

Ölçütüne göre iki değerlendirme türünden söz edebiliriz. Bunlar mutlak ve bağıl değerlendirmedir.

11.1.1.1. Mutlak Değerlendirme

Mutlak ölçüt kullanılarak yapılan değerlendirmedir. Bir üniversitenin eğitim-öğretim ve sınav yönetmeliğinde geçme puanını 50 olarak belirlediğini varsayalım (mutlak ölçüt). Bu üniversitede öğrenim gören bir öğrencinin dönem sonu notu 45 olsun (ölçme). Bu durumda, "Öğrencinin notu 50'nin altında olduğundan bu öğrenci dersten kalır" (mutlak değerlendirme).

11.1.1.2. Bağıl Değerlendirme

Bağıl ölçüt kullanılarak yapılan değerlendirmedir. Sınav yönetmeliğinde "Sınavdan sınıf ortalamasının üstünde puan alan öğrenciler dersi geçer" (bağıl ölçüt) bilgisine yer veren başka bir üniversite düşünelim. Bu üniversitede okuyan Cengiz isimli bir öğrenci dönem sonu sınavından 45 almış olsun (ölçme). Diyelim ki bahsi geçen sınavda sınıf ortalaması 40 olarak hesaplandı. Buna göre Cengiz sınıf ortalamasının üstüne bir not aldığından dersten geçer (bağıl değerlendirme). Bağıl değerlendirmede öğrencinin bilgi ve becerileri mutlak bir yeterlik standardına göre değerlendirilemediğinden ölçülen özellik bakımından bilgi kaybı yaşanabilmekte; yani gruba göre başarılı bulunan kişilerin gerçekten istenen yetkinlik/başarı standardına sahip olup olmadığı belirsiz kalmaktadır (Oyar & Gündüz, 2023).



Seçme sınavlarda genellikle bağıl değerlendirme yapılır. Örneğin, Yükseköğretim Kurumları Sınavı (YKS) sonuçlarına dayalı program yerleştirmelerinde hangi programa kaç puanla yerleşileceği belli değildir. Ancak YKS yerleştirmelerinde kaç kişinin bir programa yerleşeceği bellidir. En son yerleşecek kişinin puanı yerleştirme sırasında belli olur; önceden net bir puan tanımlanamaz. Ölçüt ve kontenjan değişmediği halde aynı programa yerleşecek en düşük puan yıldan yıla farklılaşır. Dolayısıyla yapılan değerlendirme bağıl olur.

11.1.2. Amacına Göre Değerlendirme

Değerlendirmeler yapılış amacına göre tanıma ve yerleştirmeye (placement and diagnostic assessment), biçimlendirme ve yetiştirmeye (formative), değer biçmeye/düzey belirlemeye (summative) yönelik değerlendirme olmak üzere üçe ayrılır.

11.1.2.1. Tanıma ve Yerleştirmeye Yönelik Değerlendirme

Tanıma ve yerleştirmeye yönelik değerlendirmede öğrenci/aday giriş özellikleri değerlendirilir (Miller ve arkadaşları, 2009). Bu değerlendirme türünde "Öğrenci, planlanan öğretime başlamak için gereken bilgi ve becerilere sahip mi?", "Öğrencinin yeterlilikleri bu eğitim için uygun mu yoksa başka bir grupla mı eğitimine devam etmesi gerekir?" gibi sorulara cevap aranır. Muafiyet ve hazır bulunuşluk sınavları tanıma-yerleştirme yönelik değerlendirme yapabilmek için uygulanan sınavlardır. Örneğin, yabancı dil ile eğitim yapan bir programa yerleşen üniversite öğrencileri genellikle İngilizce muafiyet sınavına tabi tutulurlar. Uygulanan sınavın sonuçlarına göre öğrencilerden hangilerinin birinci sınıftan başlayabileceğine ve hangilerinin yabancı dil hazırlık eğitimi alması gerektiğine karar verilir. Bu bağlamda İngilizce muafiyet sınavları tanıma-yerleştirmeye yönelik değerlendirme amacı taşır. Yine öğretim yılının başında seviye grupları oluşturmak amacıyla gerçekleştirilen sınavlar tanıma-yerleştirmeye yöneliktir. Millî Eğitim Bakanlığı tarafından uygulanan Liselere Giriş Sınavı (LGS) ile ÖSYM tarafından uygulanan YKS, Tıpta Uzmanlık Sınavı (TUS), Kamu Personeli Seçme Sınavı (KPSS), Dikey Geçiş Sınavı (DGS) vb. de tanıma-yerleştirme amacıyla gerçekleştirilir.

11.1.2.2. Biçimlendirme ve Yetiştirmeye Yönelik Değerlendirme

Öğretmenlerin; öğrencilerin eksik, yanlış ya da hatalı öğrenmelerini belirlemek ve elde ettikleri kanıtlar doğrultusunda öğretim uygulamalarını düzenlemek amacıyla yaptıkları değerlendirmelerdir (Popham, 2025). Yani amaç puan vermek değil öğrencileri izlemek ve öğrenmelerini artırmaktır. Dolayısıyla biçimlendirme ve yetiştirmeye yönelik değerlendirme öğrenme için değerlendirme (assessment for learning) olarak da isimlendirilir. Öğrencilerin öğrenmesini desteklemek, konu hakkında kavram eksikleri ya da yanılgıları varsa bunları gidermek için öğrencilere ayrıntılı ve bireyselleştirilmiş geri bildirimler verilmesi son derece önemlidir. Konu ve ünite testleri ile İngilizce sınıflarında sıkça uygulanan kısa sınavlar (quizler), sınıf-içi tartışmalar ile gözlemler biçimlendirme ve yetiştirmeye yönelik değerlendirme amacıyla yapılır.

11.1.2.3. Değer Biçmeye Yönelik Değerlendirme

Değer biçmeye dönük değerlendirmede amaç öğrencilere not vermektir. Bazen dönem sınavları olarak isimlendirilen sınavlarla, bazen de diğer ölçme araçları kullanılarak öğrencilerin geçme-kalma notlarının belirlendiği değerlendirme türüdür. Ara sınavlar ile dönem sonu ve telafi sınavları (vize, final ve bütünleme) değer biçmeye yönelik değerlendirme amacıyla uygulanır.

12. SONUÇ

Ölçme ve değerlendirme hem teknik uygulamalar içeren hem de belirli kavramsal temellere dayanan sistematik bir süreçtir. Bu anlamda, alanda kullanılan temel kavramların kapsamının, özelliklerinin ve birbirleriyle ilişkilerinin bilinmesi ölçme ve değerlendirme süreçlerinin doğru şekilde anlaşılabilmesi için oldukça önemlidir. Buna bağlı olarak bu bölümde ölçme ve değerlendirme alanının teorik çerçevesini oluşturan kavramlar bir arada sunulmuştur. Üzerinde durulan kavramların kitabın ilerleyen bölümlerinde ayrıntılı olarak ele alınacak ölçme araçlarının geliştirilmesi, uygulanması, analiz edilmesi ve yorumlanmasına yönelik konulara kuramsal bir alt yapı oluşturması beklenmektedir.

KAYNAKÇA | REFERENCES

Bandalos, D. L. (2018). *Measurement theory and applications for the social sciences*. The Guilford.

Baron, R. M. & Kenny, D. A. (1986). The moderator-mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology*, *51*(6), 1173–1182. https://doi.org/10.1037//0022-3514.51.6.1173

Boztunç Öztürk, N. & Şahin, M. G. (2021). Bilimsel araştırmaya temel oluşturan ölçme kavramları. B. Çetin, M. İlhan & M. G. Şahin (Ed.), *Eğitimde araştırma yöntemleri: Temel kavramlar, ilkeler ve süreçler* (s. 1–34) içinde. Pegem Akademi.

Huck, S. W. (2012). *Reading statistics and research* (6. baskı.). Pearson.

İlhan, M., Güler, N. & Taşdelen Teker, G. (2020). Nicel veri toplama araçları. B. Oral & A. Çoban (Ed.), *Kuramdan uygulamaya eğitimde bilimsel araştırma yöntemleri* (s. 77–111) içinde. Pegem Akademi.

İlhan, M., Taşdelen Teker, G. & Güler, N. (2022). Korelasyon analizi. S. Göçer Şahin & M. Buluş (Ed.), *Adım adım uygulamalı istatistik* (s. 57–104) içinde. Pegem Akademi.

Johnson, R. B. & Christensen, L. (2017). *Educational research: Quantitative, qualitative, and mixed approaches* (6. baskı) Sage.

Mackinnon, D. P. (2008). *Introduction to statistical mediation analysis*. Lawrance Erlbaum.

Mehrens, W. A. & Lehmann, I.J. (1991). *Measurement and evaluation in education and psychology*. Harcourt Brace Jonanovich: Forth Worth.

Miller, M. D, Linn, R. L. & Gronlund, N. E. (2009). *Measurement and assessment in teaching,* (10. baskı.), Prentice Hall Inc.

Mukaka, M. M. (2012). A guide to appropriate use of correlation coefficient in medical research. *Malawi Medical Journal, 24*(3), 69–71.

Oyar, E. & Gündüz, T. (2023). Öğrenme çıktılarının değerlendirilmesi: not verme. Tabak, H. & Özdemir, V. (Ed.). *Sınıf içi öğrenmeleri ölçme ve değerlendirme* (s. 279-308) içinde. Pegem Akademi.

Özer-Özkan, Y. (2022). Ölçme ve değerlendirme ile ilgili temel kavramlar. B. Çetin (Ed.). *Eğitimde ölçme ve değerlendirme* (2. baskı, s. 1–22) içinde. Anı.

Popham, W. J. (2025). *Classroom assessment: What teachers need to know*. Pearson.

Salkind, N. J. (2010). *Statistics for people who (think they) hate statistics* (3. baskı.). Sage.

Stevens, S. S. (1946). On the theory of scales of measurement. *Science*, *103*(2684), 677–680. https://doi.org/10.1126/science.103.2684.677

Tan, Ş. (2015). *Öğretimde ölçme ve değerlendirme KPSS el kitabı*. Pegem Akademi.

Turgut, M. F. (1997). *Eğitimde ölçme ve değerlendirme metotları.* Yargıcı.

Turgut, M. F. & Baykul, Y. (2010). *Eğitimde ölçme ve değerlendirme*. Pegem Akademi.

YAZAR HAKKINDA | ABOUT AUTHOR

PROF. DR. BAYRAM ÇETİN
Gaziantep Üniversitesi | bcetin27[at]gmail.com
ORCID: 0000-0001-5321-8028

Prof. Dr. Bayram Çetin, lisans eğitimini 1993 yılında Çukurova Üniversitesi, Eğitim Fakültesi, Sınıf Öğretmenliği Bölümünde tamamlamıştır. Yüksek lisans derecesini 2002 yılında, doktora derecesini ise 2005 yılında Hacettepe Üniversitesi, Sosyal Bilimler Enstitüsü, Eğitim Bilimleri Anabilim Dalı, Eğitimde Ölçme ve Değerlendirme Bilim Dalında almıştır. 2011 yılında Doçent olmuş, 2014–2019 yılları arasında Millî Eğitim Bakanlığı'nda Ölçme, Değerlendirme ve Sınav Hizmetleri Genel Müdürü olarak görev yapmış ve 2019 yılında profesör unvanını almıştır. Daha önce Hacettepe, Sakarya ve Gazi Üniversitesi'nde çalışmış olup şu anda Gaziantep Üniversitesi, Gaziantep Eğitim Fakültesi, Eğitim Bilimleri Bölümü, Eğitimde Ölçme ve Değerlendirme Anabilim Dalı'nda görev yapmakta, rektör yardımcılığı ile birlikte Gaziantep eğitim fakültesi dekanlığı görevini yürütmektedir.

PROF. DR. BAYRAM ÇETİN
Gaziantep University | bcetin27[at]gmail.com
ORCID: 0000-0001-5321-8028

Bayram Çetin is Professor of Educational Measurement and Evaluation. He received his bachelor's degree in Primary Education from the Faculty of Education at Çukurova University in 1993. He earned his master's degree in 2002 and his Ph.D. in 2005 from the Department of Educational Sciences (Division of Educational Measurement and Evaluation), Institute of Social Sciences, Hacettepe University. He was promoted to Associate Professor in 2011. From 2014 to 2019, he served as Director General for Measurement, Evaluation and Examination Services at the Turkish Ministry of National Education before being promoted to Professor in 2019. Throughout his academic career, Professor Çetin has held faculty positions at Hacettepe University, Sakarya University, and Gazi University. He is currently a faculty member in the Division of Educational Measurement and Evaluation, Department of Educational Sciences, Faculty of Education, Gaziantep University, where he also serves as Vice Rector and Dean of the Faculty of Education.

2. BÖLÜM

TEST KURAMLARI

TEST THEORIES

Prof. Dr. Tahsin Oğuz BAŞOKÇU

Prof. Dr. Nuri DOĞAN

TEST KURAMLARI

Prof. Dr. Tahsin Oğuz BAŞOKÇU
Ege Üniversitesi

Prof. Dr. Nuri DOĞAN
Hacettepe Üniversitesi

Öz

Bu bölüm ile eğitimde ve psikolojide ölçme ve değerlendirmede kullanılan başlıca test kuramları olan klasik test kuramı, madde tepki kuramı ve bilişsel tanı modellerinin tanıtılması amaçlanmıştır. Testlerin, ölçme ve değerlendirme sürecinde yaygın olarak kullanılmaya başlanması test kuramlarına duyulan ihtiyacı artırmıştır. 20. yüzyılın başlarında, öğrenme güçlüğü yaşayan bireylerin belirlenmesi amacıyla geliştirilen zekâ testleri ile Birinci Dünya Savaşı sırasında askerlerin becerilerine uygun görevlere yerleştirilmelerini sağlamak amacıyla ABD Ordusunda kullanılan testlerin psikometrik özelliklerinin değerlendirilmesi sürecinde, test kuramlarına duyulan ihtiyaç belirgin hâle gelmiştir. Bu dönemde Galton'ın bireysel farklılıklar ve bireysel farklılıkları etkileyen genetik ve çevresel faktörler üzerine yaptığı çalışmalar; Pearson tarafından korelayon kavramı üzerinde yapılan kuramsal ve hesaplamaya dönük yayınlar ve Spearman tarafından ilk kez söz edilen ölçme hataları ve faktör analizine ilişkin açıklamalar klasik test kuramının temellerinin atılmasını sağlamıştır. Klasik test kuramı 1950'li yıllara kadar olgunlaşmıştır ve tüm test kuramlarının cevabını aradığı sorulara cevap vermeyi kendi sınırlılıkları içinde başarmıştır. Ancak klasik test kuramının temel denklemlerinden ve varsayımlarından gelen zayıflıkları 1940'lı yıllardan itibaren yeni test kuramlarının gelişimini hızlandırmıştır. Temeli 1940'lı yıllardaki araştırmalara dayanan "Örtük özellik kuramı" daha sonra kendi içinde farklı kuramların ve/veya ölçme modellerinin gelişimine temel sağlamıştır. Özellikle madde tepki kuramı, Rasch modeli, örtük sınıf analizi, bilişsel tanı modelleri, yapısal eşitlik modelleri gibi birçok ölçme modeli konusunda ilk çalışmalara 1950-1970 yılları arasında yayınlanan dergilerde rastlanmaktadır. Ancak bu dönemde temel denklemleri oluşturulan bazı modellerin uygulama alanı bulması ve gelişimlerinin hızlanması bilgisayar teknolojisindeki gelişimi beklemek zorunda kalmıştır. Klasik test kuramına alternatif olarak ortaya çıkan madde tepki kuramı, yapısal eşitlik modelleri, örtük sınıf analizi, bilişsel tanı modelleri gibi ölçme modellerinin analizinde başvurulan iteratif ve/veya karmaşık süreçler, bu kuramların yaygınlaşmalarını geciktirmiş; geniş kapsamlı ve karşılaştırmalı çalışmaların ertelenmesine neden olmuştur. İnternetin ve bilgisayarların bireysel kullanıma sunulduğu 1990'lı yıllardan itibaren hem klasik test kuramı hem de alternatif test kuramları üzerinde yapılan karşılaştırmalı yayınların her geçen yıl katlayarak arttığı; alternatif test kuramlarına ilişkin hesaplamalı çalışmaların kolaylaştığı görülmektedir. Diğer yandan 2000'li yılların başından itibaren büyük veri ve veri madenciliği tekniklerinin psikometrik analizlerde kullanılmaya başlandığı ve özellikle son 10 yıl içinde yapay zekâ destekli psikometrik analizlerin yaygınlaşmaya başlandığı söylenebilir. Bu gelişmeler farklı test kuramlarına dayalı sonuçların aynı koşullar altında karşılaştırılmasını kolaylaştırırken psikometrik analizlerden daha geniş kesimlerin yararlanmasını da sağlamaktadır. Bu bölümde öne çıkan test kuramlarından klasik test kuramı madde tepki kuramı ile bilişsel tanı modelleri sunulmuştur.

Anahtar Kelimler

Test Kuramları, Klasik Test Kuramı, Madde Tepki Kuramı, Örtük Sınıf Analizi, Bilişsel Tanı Modelleri

TEST THEORIES

Prof. Dr. Tahsin Oğuz BAŞOKÇU
Ege University

Prof. Dr. Nuri DOĞAN
Hacettepe University

Abstract

The purpose of this chapter is to introduce the main test theories used in measurement and assessment in education and psychology: classical test theory, item response theory, and cognitive diagnostic models. The widespread use of tests in the measurement and assessment process has increased the need for test theories. In the early 20th century, the need for test theories became apparent during the evaluation of the psychometric properties of intelligence tests—developed to identify individuals with learning difficulties—and of tests used by the U.S. Army during World War I to ensure that soldiers were assigned to tasks suited to their abilities. During this period, Galton's studies on individual differences and the genetic and environmental factors influencing them; Pearson's theoretical and computational publications on the concept of correlation; and Spearman's explanations regarding measurement errors and factor analysis—which he first introduced—laid the foundations for classical test theory. Classical test theory matured by the 1950s and, within its own limitations, succeeded in answering the questions that all test theories sought to address. However, weaknesses stemming from the fundamental equations and assumptions of classical test theory accelerated the development of new test theories beginning in the 1940s. The "Latent trait theory," which is rooted in research from the 1940s, later served as the foundation for the development of various theories and/or measurement models within its own framework. Early studies on numerous measurement models—such as item response theory, the Rasch model, latent class analysis, cognitive diagnostic models, and structural equation models—can be found in journals published between 1950 and 1970. However, the practical application and accelerated development of some models, whose fundamental equations were established during this period, had to wait for advances in computer technology. The iterative and/or complex processes involved in analyzing measurement models—such as item response theory, structural equation modeling, latent class analysis, and cognitive diagnostic models—which emerged as alternatives to classical test theory, delayed the widespread adoption of these theories and led to the postponement of large-scale and comparative studies. Since the 1990s, when the internet and computers became available for individual use, the number of comparative publications on both classical test theory and alternative test theories has increased exponentially each year; computational studies related to alternative test theories have also become easier to conduct. On the other hand, it can be said that big data and data mining techniques began to be used in psychometric analyses starting in the early 2000s, and that AI-supported psychometric analyses have become increasingly widespread, particularly over the past 10 years. While these developments facilitate the comparison of results based on different test theories under the same conditions, they also ensure that a broader audience can benefit from psychometric analyses. This section presents the classical test theory, item response theory, and cognitive diagnostic models as prominent test theories.

Keywords

Test Theories, Classical Test Theory, Item Response Theory, Latent Class Analysis, Cognitive Diagnostic Models

1. GİRİŞ

Bireylerin bilişsel, duyuşsal ve çeşitli psikomotor davranışlarını ölçmek için testlerden yararlanılır. Dolayısıyla testler bireylerin sahip olduğu özellikleri saptamak amacıyla tüm dünyada yaygın şekilde kullanılır. Okullara yönetici ve öğretmen atamalarında, iş yerlerine çalışan seçiminde, bilimsel araştırmalarda, derslerde başarıyı belirlemede, kişileri belirli mesleklere yönlendirmede, bireylerin psikolojik sorunlarına ilişkin tanı koymada, öğrencileri üst eğitim kademelerine yerleştirmede, kurum ve kuruluşların kamuoyu araştırmalarında vb. pek çok alanda testlere başvurulur. Testlerin kullanımı o kadar yaygındır ki hayatı boyunca herhangi bir test uygulamasına katılmamış bir bireye rastlama olasılığı oldukça düşüktür.

Elbette geliştirilen her test amaca hizmet etmeyebilir ve bireylerin ölçmeye konu olan özellikleri hakkında doğru sonuçlar ortaya koymayabilir. Testlerin beklenen faydayı sağlaması, hedeflenen özelliği doğru şekilde ölçebilmesine bağlıdır. Ancak kimi zaman testler yetkin olmayan bireyler tarafından oluşturulmakta veya uygulanmakta, testten elde edilen sonuçlar yetkin olmayan bireyler tarafından analiz edilmekte ve analiz sonuçları yine yeterli uzmanlığa sahip olmayan kişiler tarafından yorumlanmaktadır. Oysa testlerin geliştirilmesinde, uygulanmasında ve sonuçlarının yorumlanmasında gerekli teknik becerilere sahip (yetkin) bireylerin görev alması gerekir. Ölçme süreçlerinde görev alan kişilerin sahip olması beklenen nitelikler olduğu gibi kullanılan testler ile bu testlere dayalı yorumların da taşıması gereken psikometrik özellikler bulunmaktadır. İşte, bireylerin ölçülen yapıya ilişkin puanlarının belirlenmesi, bu puanların ne kadar doğru olduğunun tespiti ve elde edilen ölçümlerin psikometrik kalitesinin saptanması test kuramlarının konusunu oluşturmaktadır.

Test kuramları (ölçme kuramları); psikolojik ölçmenin temel ilkelerini açıklama, test puanlarını yorumlama ve karşılaştırma, testlerden maksimum düzeyde yararlanma, bireylerin ihtiyaçlarına uygun test geliştirme vb. teknik konuları bütünsel olarak ele alıp değerlendirir. Bu çerçevede test kuramları; ölçme sonuçlarının kalitesini tanımlamak, sınıflandırmak ve değerlendirmek için kurallar belirler. Tüm bunlar dikkate alındığında test kuramlarının ölçümlerin kullanışlılığı, doğruluğu ile anlamlılığını iyileştirmeyi ve daha nitelikli ölçme

Genellenebirlik (G) kuramı KTK'nin bir uzantısı olup varyans analizi yoluyla ölçmeye karışan hataları bileşenlerine ayırabilmektedir. Ancak bireylerin yetenek düzeylerinin kestirimine ilişkin KTK'nin ötesinde ayrı bir model ileri sürmemiştir. Bundan dolayı G kuramı bu bölüm kapsamına alınmamıştır. G kuramı hakkında bilgi edinmek isteyen okuyucular Shavelson & Webb (1991), Brennan (2001), Güler ve arkadaşları (2012), Atılgan (2019), İlhan ve arkadaşları (2024) kaynaklarına başvurulabilirler.

araçları geliştirmeye hizmet edecek yöntemler ortaya koymayı amaçlayan bir çalışma alanı olduğu söylenebilir. Kitabın bu bölümünde öncelikle test kuramlarının tarihsel gelişimi üzerinde durulmuştur. Ardından geleneksel test kuramlarından Klasik Test Kuramı'na (KTK) yer verilmiş, devamında çağdaş test kuramlarından Madde Tepki Kuramı (MTK) ile Bilişsel Tanı Modelleri'ne (BTM) değinilmiştir.

2. TEST KURAMLARININ GELİŞİM SÜRECİ

On dokuzuncu yüzyılda test kuramlarının gelişimine yön veren iki ana çizgi vardır. Bunlardan ilki psikolojinin gelişimi ve deneysel bir bilim olarak ön plana çıkmasıdır. Diğeri ise istatistik yöntem ve tekniklerdeki gelişmelerdir. Bu iki alandaki gelişmelerin paralel şekilde ilerlediği, birbirini etkilediği ve desteklediği, psikolojik testlerin ve test kuramlarının daha hızlı gelişmesine zemin hazırladığı söylenebilir. Bahse konu iki ana çizgiye yirminci yüzyılın ikinci yarısından itibaren bilgisayar teknolojilerindeki gelişimeler ile internetin bireysel kullanım için sunulması da eklenmiştir. Özellikle MTK gibi görece yeni test kuramlarının uygulanması ve büyük örneklemlerde karmaşık analizlerin gerçekleştirilmesi bilgisayar teknolojisindeki gelişmelerle mümkün olmuştur.

Du Bois (1970) psikolojik testlerin tarihini memurluk sınavları, okul içi sınavlar ve bireysel farklılıkların belirlenmesi için yapılan sınavlar olmak üzere üç ana başlıkta inceler. Memurluk sınavlarının tarihi yaklaşık 3000 yıl öncesine, Çin'de imparatorluğun memurların yeterliliğini değerlendirmek için yaptığı sınavlara kadar uzanmaktadır. Çin'de memurları işe almak veya değerlendirmek için kullanılan sınavlar "zihinsel test" uygulamalarının ilk örnekleri olarak görülmektedir. Bu sınavlar müzik, binicilik, medeni hukuk, yazı, din bilgisi, kamu ve özel tören bilgisi gibi konuları kapsayan sorulardan oluşmuştur. Avrupa'daki üniversiteler, 13. yüzyılda adaylara ilk kez resmî sözlü sınavlar uygulamıştır. Kâğıt kullanımının yaygınlaşmasıyla Cizvitler 16. yüzyılda öğrencileri seçme ve okul içinde değerlendirme süreçlerinde basılı testler kullanmaya başlamıştır.

Sosyal bilimlerde özellikle psikoloji alanındaki gelişmeler çerçevesinde 19. yüzyılın başlarında bireysel farklılıklara ilgi artmıştır. Darwin'in görüşlerinden fazlasıyla etkilenen kuzeni Francis Galton, bireysel farklılıklara odaklanarak psikometri alanını destekleyen önemli fikirler ortaya koymuştur. İlgili dönemde birçok fiziksel özelliğin (örneğin; boy, ağırlık) normal dağılım eğrisine benzer şekilde dağıldığı bilinmekteydi. Galton bunu bir adım ileri taşımış ve psikolojik özelliklerin de normal dağılacağını öne sürmüştür. Bu dönemde kalıtım çalışmalarına da yaygın bir ilgi söz konusu olmuştur. Kalıtıma duyulan merak zihnin ölçülmesine ilgiyi artırmış ve psikometrik testlerin gelişimini hızlandırmıştır. Galton özellikle zihinsel işlevlere, zekâ ve genetik ilişkisine odaklanmıştır. 1880'lerin ortalarına gelindiğinde fiziksel güç testi, hız testleri, görme keskinliği ve hafıza testleri gibi özelliklere ilişkin ölçümler geliştirmiştir. Galton'ın Büyük Britanya'da kurduğu Antropometrik (Antropoloji) Laboratuvarı'nda duyusal ve motor

becerileri ölçmeye yönelik kullandığı araçlar bireysel farklılıklara ilişkin sistematik ölçümlere temel teşkil etmiştir. Galton bu ölçümlerin birbirleriyle ilişkisi, erken yaşta alınan puanların daha sonraki başarıyı kestirip kestirmeyeceği ve başarının bir nesilden diğerine aktarılıp aktarılmadığı gibi konularla da ilgilenmiştir.

Galton, insanların yeteneklerini ve çeşitli duyusal özelliklerini ölçerek incelemiştir. Bu yaklaşım devrim niteliğinde kabul edilmektedir. Galton ölçme yöntemlerini ve ölçme yoluyla veri toplamak için gereken mekanik aletleri, bunun yanı sıra ölçme sonuçlarıyla başa çıkmak için gereken istatistiksel yöntemleri kendisi geliştirmiştir. Doğrudan sayısal ölçümler elde edemediği durumlarda bile değişkenlik ve dağılımı hesaba katarak sayısal olarak akıl yürütmüştür. Zekâ testleri, antropometri ve biyometri Galton'ın bu çabaları sonucunda ortaya çıkmıştır. Galton iki değişken arasındaki olasılıksal ilişkiyi kalibre etmek için bir endekse ihtiyaç duymuş ve devamında Gillman dağılım grafiklerini kullanarak bir değişkene ait ölçümlerin diğer bir değişkenin ölçümlerini tahmin etmedeki işlevini fark edip iki puan kümesinin ilişkisini hesaplamaya yarayan bir yöntem geliştirmiştir. Meslektaşı, istatistikçi Karl Pearson (1857-1936), bu çalışmayı resmileştirmiş ve genişletmiştir. Bazı bilim insanları tarafından "İstatistiğin kurucusu" olarak anılan Pearson, "normal eğri" ve "standart sapma" gibi yeni terimleri kullanarak modern psikometri için istatistiksel yapı taşları haline gelecek yöntemleri inşa etmiştir. Pearson ayrıca, ölçme teorisinin özünü oluşturan ve psikometrik çalışmalarda uzun yıllar boyunca temele alınan Pearson moment çarpımları korelasyon katsayısını ve ardından nokta-çift serili korelasyon gibi bazı teknikleri geliştirmiştir (Gilham, 2001).

Alfred Binet'e (1857-1911) kadar zihinsel özelliklerle, genellikle de zekâyla ilgilenenler zihinsel yapıları değerlendirmede kullanabilecekleri etkili araçlardan yoksundu. Binet Fransa'da meslektaşı Theophile Simon ile birlikte öğrenme güçlüğü yaşayan çocukları saptamak için çalışmalar yürütmüştür. 1905 yılında bireysel farklılıklar üzerine yaptığı çalışmaların bir parçası olarak bazı zihinsel özellikleri ölçtüğünü iddia ettiği bir ölçek yayımlamıştır. Öğrencilerin bir dizi görevdeki performansının yeteneklerinin sıralaması için gösterge olabileceğini, diğer bir deyişle öğrencilerin gerçekleştirmeleri beklenen çeşitli görevlerdeki performanslarının yeteneklerine ilişkin yeterli ve anlamlı bir temel oluşturacağını savunmuştur. Binet bunu gerçekleştirmek için öğrencilerin okulda karşılaştığı görevlere benzeyen ama bire bir aynısı olmayan, karmaşık zihinsel işlevleri içeren bir soru örnekleminin yeterli olacağını kabul etmiş, böylece bireyler arası farkları ortaya çıkaracak bir puan dağılımına ulaşılacağını ifade edip ölçeğinin bir zekâ testi olmadığını, yalnızca özel gereksinimli çocuklar için bir tarama aracı olduğunu vurgulamıştır. Binet'in ölçeğinin yayınlanmasından kısa bir süre sonra Alman psikolog ve filozof William Stern 1912 yılında günümüzdeki Zekâ Katsayısı (IQ) tanımından oldukça farklı olsa da zihinsel (ölçülen) yaşın kronolojik (gerçek) yaşa oranı olarak tanımladığı bir katsayı (IQ) geliştirmiştir.

Binet'in vurgusunun tersine Goddard gibi bazı Amerikalı psikologlar, Binet'in testinin zekâyı ölçtüğünü, testteki soruların zekâdaki genel bir faktörü yansıttığını ve bu genel faktörün büyük ölçüde kalıtsal özellik olduğunu öne sürmüştür. Dahası Terman II. Dünya Savaşı'ndan önce Binet'in testini hem Amerikan kültürüne uyarlamış hem de revize etmiştir. Terman'ın uyarlamasının revize geçirmiş bir formu günümüzde de kullanılmakta ve Stanford-Binet testi olarak bilinmektedir. Binet testinden farklı olarak Stanford-Binet testi öğrencilerin yaşına uygun güçlük düzeylerine göre gruplanmış problemler veya "maddeler" içermekte ve farklı zekâ alanlarına zekâ puanı (IQ) vermektedir. Birinci Dünya Savaşı sırasında ABD ordusunda Binet tipi testlerin yaygın kullanımı, test süreçlerini kolaylaştırmış ve prosedürlerini standartlaştırmaya yardımcı olmuştur. Bu uygulamalar, standartlaştırılmış testlerle grup testlerinin ve çoktan seçmeli yanıt formatlarının ilk büyük ölçekli uygulaması olarak kabul görmektedir (Sokal, 1987; Danziger, 1990; Rose, 1979).

20. yüzyılın başlarında, James Cattell ve meslektaşları, tepki süresi ve duyusal keskinlik üzerine daha ayrıntılı bir dizi antropometrik ölçüm geliştirmiştir. Ancak Cattell'in ölçümleri, bireylerin daha sonraki okul ve mesleki başarılarıyla tutarlı sonuçlar vermemiştir. Pearson gibi bir istatistikçi olan Charles Spearman (1863-1945), söz konusu tutarsızlığın Cattell ve meslektaşlarının elde ettiği ölçümlerdeki hatalardan kaynaklanmış olabileceğine dikkat çekmiş; herhangi bir testten elde edilen sonuçların kaçınılmaz olarak ölçme hatası içereceğini ve hataya açık başka ölçme sonuçlarıyla korelasyonunun gerçek puanlarla (hata içermeyen puanlarla) korelasyonundan daha küçük çıkacağını savunmuştur. Spearman'a göre, belli bir testle elde edilen ölçme sonuçlarındaki ölçme hatasını tahmin etmenin bir yolu, tekrarlı uygulamalardan elde edilen sonuçları ilişkilendirmektir. Spearman "tesadüfi hata" nedeniyle oluşan korelasyonlardaki bu "zayıflamayı" belirlemeyi mümkün kulan bir formül tanımlamıştır. Brown da Spearman'dan bağımsız olarak aynı formülü geliştirmiş ve her ikisi de Yule'a dayandırdıkları bazı istatistiksel kanıtlar sunmuştur. Galton ve Pearson'ın sadık bir takipçisi olan Spearman'ın çalışmaları, modern test puanı güvenirliği ile faktör analizi kavramlarının ortaya çıkışına öncülük etmiştir. Spearman'ın iki bileşik ölçümün ilişkisini (korelasyonunu) varyans ve kovaryans cinsinden ifade etmesi daha sonra "güvenilirlik endeksi" olarak bilinen istatistiğin formülasyonu ile sonuçlanmıştır (Levy, 1995).

1800'lü yılların sonu ve 1900'lü yılların başında uygulanan testlerin çoğu, bireysel test şeklinde tasarlanmıştır. Yirminci yüzyılın başlarında, okullardaki öğrencilere grup testlerinin uygulandığına nadiren rastlanmaktadır. Grup uygulamaları I. Dünya Savaşı sürecinde ABD'de askerlere uygulanan testlerle yaygınlaşmıştır. Savaş döneminde uygulanan zekâ testleri sonucunda ABD nüfusunun geniş bir örneklemini meydana getiren askerlerin ortalama zekâ puanının "genel ortalamanın altında" kaldığı tespit edilmiştir. Bu durum test puanlarını yorumlamada esas alınacak, test puanlarına anlam kazandıracak uygun bir "norm" grubu seçimi sorununa işaret etmiştir. Daha sonra geliştirilen zekâ testlerinde, örneğin Wechsler zekâ ölçeklerinde, bu durum göz önüne alınarak ölçme sonuçlarının karşılaştırılacağı norm grupları belirlenirken daha dikkatli seçimler yapılmıştır.

Birinci ve ikinci dünya savaşı arasında "geçerlik" kavramı, yani testlerin ölçülmek istenen değişkeni ölçüp ölçmediği ya da ne derece ölçtüğü sorunu ortaya çıkmıştır. Binet ve o dönemdeki bazı zekâ testi uygulayıcıları, "Zeka, zekâ testlerinin ölçtüğü şeydir" ifadesini kullanarak ortaya çıkan geçerlik sorunundan bir totolojiyle kurtulmayı başarmıştır (Rogers 1995; Michell, 1999). Geçerliğe ilişin bu yaklaşım, soruna geçici ve bir ölçüde yaratıcı bir çözüm geliştirerek işlemselcilik (operationalizm) başlığı altında bir dönem için kabul görmüştür. Örneğin 1930'ların ortalarında Stevens psikolojik bir yapının (kavramın/özelliğin) kuramsal anlamının, o psikolojik yapıyı ölçmek için kullanılan işlemlerle tanımlanabileceğini savunmuştur. Psikolojik bir kavramı ölçmek için kullanılan işlemler genellikle ölçülen özelliğin gözlenmesini ve gözlem süreci sonunda ilgili özelliğe sistematik biçimde sayısal değerler atamayı içermektedir. Bu dönemde ölçme ile ilgilenen farklı alanlardaki birçok uzman, bir psikolojik kavramı ölçülebilir hale getirmek için gerekli olan işlemleri uygulamanın psikolojik kavramı üretmek için yeterli olacağı görüşünü kabul etmiştir (Rogers, 1995; Michell, 1999; Wright, 1997).

Zekâ testleri üzerinde yapılan çalışmaların pratikteki başarısından dolayı psikologlar, eğitimciler ve test geliştiriciler işlemselci anlayışı ilgi, tutum ve kişilik gibi duyuşsal alan ölçümlerine de genişletmiştir. I. Dünya Savaşı'ndan sonra test geliştirme sürecinde zekâ testlerinin standardizasyonunu yapmaya önem verilmiş ve grup karşılaştırma teknikleri devreye sokulmuştur. Bu dönemde geliştirilen testler, belli bir içeriğini örnekleyen ve sabit yanıt formatına sahip kısa maddeler içermiştir. Maddelere verilen yanıtlara ilişkin puanlar toplanmış, doğrusal ölçek oluşturmak için birleştirilmiştir. Test toplam puanları, ölçülen psikolojik kavramın sayısal bir göstergesi olarak yorumlanmış ve norm gruplarının puan dağılımlarıyla karşılaştırmak için kalibre edilmiştir. Zekâ, yetenek ve başarı testlerinin aksine ilgi, tutum veya kişilik gibi duyuşsal alanı ölçen testlerde durum kısmen farklılaşmıştır. Çünkü bu testlerde kullanılan maddelerin doğruluğu ya da yanlışlığı söz konusu değildir. Duyuşsal alan testlerindeki belirsiz madde içeriği ve diğer ölçümlerle zayıf ilişkiler; ilgi, tutum, kişilik gibi yapılara ilişkin ölçme araçlarının geliştirilmesinde ölçülecek psikolojik özelliğe ilişkin tanımların dış ölçütlere göre belirlendiği yöntemlerin yaygınlaşmasına yol açmıştır (Buchanan, 2002). II. Dünya Savaşı sonrası dönemde psikologlar ve test geliştiriciler, işlemselci ölçme yaklaşımlarını yavaş yavaş terk etmiştir. İşlemselci ölçme yaklaşımlarına yenileri eklenerek kuramsal temellendirme ve gerekçelendirme ön plana çıkmıştır. Spearman'ın temelini attığı, ardından Thurstone ve diğerlerinin geliştirdiği faktör analizi tekniklerinin, gizil (örtük) psikolojik kavramları türetmek için matematiksel bir model olarak kullanılması yaygınlaşmaya başlamıştır (Lovie & Lovie, 1993; Bartholomew, 1995). O dönemden günümüze değin faktör analizi ölçme araçlarının geçerliğine ilişkin kanıtlar sunmada önemli bir rol oynamıştır. Örneğin Cattell'in ve Eysenck'in kişilik envanterlerini geliştirme sürecinde ve başka birçok ölçme aracının yapısal geçerliğinin araştırılmasında ya da kanıtlanmasında faktör analitik teknikler başvurulan temel istatistiksel yöntem olarak karşımıza çıkmaktadır.

Spearman'ın başlattığı temel çalışmalar ve birçok psikometristin katkısıyla oluşan sonuçlar Gulliksen tarafından kapsamlı bir şekilde birleştirilmiştir. Gulliksen (1950) "Zihin Testleri Kuramı" (Theory of Mental Tests)" kitabında "gerçek puan modelini" birçok yönüyle açıklamıştır. Lord ve Novick 1960'ların sonunda KTK olarak adlandırılan bu modele, olasılık teorisine dayanan katkılar sunmuştur (Traub, 1997). Lord ve Novick aynı zamanda genelleştirilmiş yanıt modelleri oluşturmak için madde düzeyindeki analizleri genişletmiş ve MTK'yi geliştirmiştir. MTK savunucuları, bu analizlerin KTK'nin önemli sınırlılıklarını (madde istatistiklerinin uygulama yapılan örnekleme, örneklemdeki bireylerin yetenek ölçülerinin de uygulanan testin güçlük düzeyine bağlı olması) aştığını iddia etmektedir. Dolayısıyla MTK test güçlüğü ve örneklem yetenek dağılımı kısıtlamalarından bağımsız olarak gizil özelliklerin kestirilmesine ilişkin analiz yöntemleri ve modeller geliştirmeyi temele almıştır (Bock, 1997; Embretson, 1996).

Du Bois'e göre (1970), bir disiplin olarak test kuramı 1930'larda gelişmeye başlamıştır. Test kuramlarından ilki olan, en çok kullanılan ve sınıf içi ölçmelerde uygulanan testleri geliştirirken hala kaçınılmaz olarak tercih edilen KTK'nin temelleri bu tarihsel süreç içerisinde atılmış ve büyük oranda 1950'lerde tamamlanmıştır. Diğer yandan eğitimsel ve psikolojik ölçme üzerinde veya psikometrik yöntemler üzerine araştırmalar artarak sürmektedir. KTK'nin uzantısı olarak görülen G Kuramı, MTK ve Rasch modeli gibi "modern test kuramları" mevcut ölçme uygulamalarını desteklemek, eksiklerini gidermek ve gerektiği durumlarda değiştirmek için alternatif modeller ve teknikler üzerinde çalışmaktadır. Aşağıda ilk olarak KTK açıklanmış, daha sonra MTK ile BTM başlıklarına yer verilmiştir.

3. KLASİK TEST KURAMI

Ne kadar dikkat edersek edelim, ölçme araçlarımızın hassasiyeti ne kadar yüksek olursa olsun hiçbir ölçme aracı mükemmel düzeyde güvenilir veya hatasız ölçümler vermez. Örneğin günümüzde zamanı ölçmek için kullandığımız araçlar oldukça gelişmiş, mikro-saniyelerle ölçümler yapabilecek duruma gelmiş olsa da saatler farklı koşullarda yavaş ya da hızlı çalışabilmektedir. İnsanların zihinsel özelliklerine ilişkin ölçümler söz konusu olduğunda ise ölçme sonuçlarının hassasiyeti daha da azalmaktadır. Örneğin psikologlar bireylerin zekâ, yetenek, başarı, tutum, ilgi, kişilik, kaygı gibi psikolojik niteliklerini; otomobil üreten mühendislerin hızı veya coğrafyacıların iki bölge arasındaki uzaklığı ölçmede sağladıkları hassasiyetle ölçememektedir. Bunun tipik örneğini aynı testi aynı gruba farklı zamanlarda uygulayarak test-tekrar test güvenirlik katsayısını hesaplayan araştırmacılar net biçimde gözleyebilmektedir.

Psikolojik ölçümler için kolay uygulanabilecek bir örnekle bu durumu açıklamak mümkündür. Eğer bir grup öğrenciye bir matematik testini uygulasaydık, test puanlarının bir öğrenciden diğerine değiştiğini, yani tüm öğrencilerin aynı puanı almadığını saptayabilirdik. Bunun birçok nedeni vardır. Bunlardan en önemlisi testi uyguladığımız öğrencilerin matematik bilgi ve/veya becerilerindeki farklılıktır. Test

puanlarının, öğrencilerin matematik bilgilerindeki/ becerilerindeki farklılığı yansıtması istenen bir durumdur. Ancak şimdi, aynı matematik testini aynı öğrencilere bir gün sonra (uygulamada bu süre daha uzundur ve genellikle 2-6 hafta arasında değişir) ikinci bir kez uyguladığımızı varsayalım. Acaba her bir öğrenci testin ikinci uygulamasından ilk uygulamada elde ettiği puanının aynısını alabilir mi? Bu konuda yapılan çalışmalarda ortaya çıkan sonuç cevabın neredeyse her zaman "hayır" olduğudur. Yani öğrencilerin testin birinci ve ikinci uygulamasından elde ettikleri puanlarının değişmesi neredeyse kesindir. Öğrencilerin matematik testinin farklı zamanlardaki iki uygulamasından elde ettiği puanlar birbirine yakın olabilir ancak çoğunlukla aynı kalmaz. Bazı öğrenciler testin ikinci uygulamasında daha yüksek puan alırken bazıları daha düşük puan alabilir. Matematik becerisinin her öğrencide aynı kaldığını varsayarsak (sonuçta sadece 1 gün sonra aynı testi tekrar uyguluyoruz), ikinci uygulamada elde edilen puanlarda farklılığa yol açan nedir?

Geleneksel ölçme modeli bu duruma "gerçek puan modeli" adı verilen matematiksel bir model ile açıklama getirir. KTK bu model üzerine temellendirilmiştir. Geleneksel ölçme modeline göre bireyin test puanı gözlenen puan olarak adlandırılır ve iki bağımsız bileşenden oluşur. İlk bileşen, testin ölçmek üzere tasarlandığı niteliğin bireyde bulunan gerçek miktarının bir ölçüsüdür. Dolayısıyla bu bileşene gerçek puan (True Score) adı verilir. İkinci bileşen ise testin uygulandığı ortam, ölçme araç veya yöntemi, ölçülen birey, ölçmeyi/puanlamayı yapan kişi gibi farklı kaynaklardan gelebilen, yönü ve/veya miktarı belirlenemeyen tesadüfi hata puanıdır. KTK'nin temelini oluşturan gerçek puan denklemi Eşitlik 1'de sunulmuştur.

$X_i = T_i + E_i$

X: Gözlenen puan (test puanı, ölçme sonucu)
T: Gerçek puan (bireyin ölçülen özelliğe ilişkin gerçek puan değeri)
E: Tesadüfi (random = seçkisiz/rastgele/gelişigüzel) hata puanı

Eşitlik 1

Gerçek puan denklemi iki bilinmeyenli bir denklemdir. Denklemdeki terimlerden sadece gözlenen puanı doğrudan bilinmektedir. Pratikte gerçek puan ve hata puanını hipotetiktir; doğrudan belirlenememekte, birtakım varsayımlar altında kestirilebilmektedir. Daha açık bir anlatımla gerçek puan, bir birey bir testi sonsuz kez aldığında elde edeceği puanların ortalamasının hesaplanmasıyla ulaşılabilecek hipotetik bir puandır. Herhangi bir test uygulamasında meydana gelebilecek tesadüfi hatalar, bir test sonsuz kez uygulandığında birbirlerini sıfırlayacak, yani negatif ve pozitif yönlü hatalar birbirini götürecektir. Bu nedenle, sonsuz tekrardan elde edilen tüm puanların ortalaması alındığında sonuç artık tesadüfi hata içermeyen bir puanı temsil edecektir. Bu ortalama, bireyin testin ölçtüğü özelliğe ilişkin gerçek puanı olarak kabul edilir. Yukarıdaki eşitlikten görülebileceği üzere ölçmeye karışan hata, gözlenen puan ile gerçek puan arasındaki farka karşılık gelir. Gerçek puan denkleminden hareketle ölçme hatasına ilişkin formül Eşitlik 2'deki gibi yazılabilir.

$E_i = X_i - T_i$ *Eşitlik 2*

Yukarıdaki örnek çerçevesinde bir öğrencinin matematik testi puanının bir uygulamadan diğerine değişmesine neden olan şey (öğrencinin gerçek matematik puanının değişmeyeceği varsayımı altında) tesadüfi hatalardır. Bu hatalar gerçek puandan bazen daha yüksek bazen de daha düşük bir gözlenen puan elde edilmesine sebebiyet verir. Dolayısıyla, öğrencinin gerçek matematik puanının ne olduğunu belirlemek için sonsuz sayıda uygulama yapıp elde edilen puanların ortalamasını almak gerekir. Sonsuz sayıda uygulama yapmak pratikte mümkün olmadığından eşitliği çözümleyebilmek için hata ya da gerçek puana ilişkin birtakım varsayımlarda bulunmak zorunluluk halini alır. Tesadüfi hataların normal dağılım göstermesi (artı ve eksi yönde dağılması) ve başka değişkelerle ilişkisinin bilinebilir olması nedeniyle gerçek puan denklemi tesadüfi ölçme hatalarına ilişkin varsayımlardan hareketle çözümlenebilir. KTK'nin hata için öngördüğü varsayımlar aşağıda açıklanmıştır.

3.1. Hatalara İlişkin Varsayımlar

KTK'nin temelini oluşturan gerçek puan modelinde ölçme sonuçlarına karışan hatalar tesadüfi olduğundan bu hatalar evrende (ya da yeterince büyük örneklemlerden elde edilen ölçümlerde) normal dağılıma sahiptir. Tesadüfi hatalar artı ve eksi yönde değerler alacağından yeterince gözlem yapıldığı takdirde ortalamaları sıfır olur. KTK'nin tesadüfi hatalara ilişkin üç varsayımı şöyle özetlenir (Baykul, 2000).

- $\mu_E = 0$ (Hataların ortalaması sıfırdır.)
- $r_{TE} = 0$ (Hatalarla gerçek puan korelasyonu sıfırdır)
- $r_{E1E2} = 0$ (İki paralel ölçme işleminde tesadüfi hatalar arasındaki korelasyon sıfırdır)

Gerçek puan denklemine göre şu yorumları yapmak mümkündür: Ölçme sonucunda ulaşılan gözlenen puan gerçek puana ne kadar yakınsa tesadüfi hata o kadar düşük, güvenirlik de o kadar yüksek demektir. Haliyle güvenirliği yükseltmenin yolu ölçme sonuçlarına karışan tesadüfi hataları önlemekten geçer.

Gerçek puanı daha iyi açıklamak için daha önce değindiğimiz matematik testine ilişin örneği genişletebiliriz. Öğrencilerin matematik bilgi ve becerilerini belirlemek için tek bir sorudan oluşan bir test uygulamış olalım. Bir öğrencinin matematik becerisini sadece bir soruyla belirlemek ne derece mümkündür? Öğrencinin bir soruluk testten alacağı puan gerçek matematik becerisini ne düzeyde temsil edebilir? Bir maddeden elde edilen puana çok sayıda farklı faktörden gelen tesadüfi hatalar karışacak; buna bağlı olarak hata puanı yüksek, güvenirlik ise düşük olacaktır. Örneğin öğrenci ilgili maddeyi şansla doğru cevaplamış veya başka bir öğrenci söz konusu maddeyi cevaplarken kopya çekmiş olabilir. Bu durumda gözlenen puan, gerçek puandan oldukça farklı bulunacak, tesadüfi hata puanı yüksek olacaktır. Dolayısıyla bireyin gerçek matematik yeteneğini tahmin etmek için daha fazla maddeye ihtiyaç duyulur. Böylece tesadüfi hatalar azalırken

gözlenen puan ile gerçek puanın birbirine yaklaşması beklenir. Yani matematik becerisini daha fazla madde ile ölçtüğümüzde bireyin matematik performansı üzerindeki rastgele etkiler dengelenecektir. Tesadüfi hatanın doğası tam olarak bu şekilde işler. Bu anlamda, herhangi bir psikolojik özelliği ölçerken bireye sonsuz sayıda maddeden oluşan bir test uygulanır ve tüm madde puanlarının ortalaması alınırsa tesadüfi hata tamamen ortadan kalkar ve öğrencinin gerçek puanı elde edilebilir.

3.2. KTK'de Güvenirliğin Hesaplanması

KTK'de güvenirliğe ilişkin tüm kavramsallaştırmaların temelinde, aynı özelliğin tekrarlı ölçümler yoluyla değerlendirilmesi anlayışı yatar. Bu anlayış doğrultusunda KTK güvenirliği açıklarken ve kestirirken *paralel testler* varsayımını temel alır (Crocker & Algina, 1986). İki testin ortalamasının, varyansının, dağılımsal özelliklerinin ve dışsal değişkenler ile korelasyonunun eşit olması halinde bu testler paralel test olarak nitelendirilir ve ancak bu şartlar altında gerçek puan ile hata puanı ilişkisiz kabul edilebilir. Bu da KTK'nin temel denkleminde her iki tarafın varyansını alarak Eşitlik 3'e ulaşmayı mümkün kılar (John & Benet-Martinez, 2000).

$$\text{Varyans}(X) = \text{Varyans}(T + E)$$
$$\text{Varyans}(X) = \text{Varyans}(T) + \text{Varyans}(E) \qquad \textit{Eşitlik 3}$$

Güvenilirlik konusundaki herhangi bir tartışma, ölçmenin standart hatası ele alınmadan eksik kalır. Bu bakımdan KTK'de güvenirliğin nasıl belirleneceğine geçmeden önce güvenirlikle doğrudan ilişkili olan "ölçmenin standart hatası" kavramından da bahsedilmelidir (Kurpis & Stafford, 2006). Tesadüfi hataların gerçek puan etrafındaki dağılımının standart sapması, ölçmenin standart hatası (Standard Error of Measurement, SEM) olarak tanımlanır (Hogan, 2019). Ölçmenin standart hatası ne kadar düşükse, tesadüfi hatalar gerçek puan etrafında o kadar dar bir aralıkta kümelenir. Dolayısıyla ölçümlerin güvenirliğini belirlemede başvurulabilecek göstergelerden biri, ölçmenin standart hatasıdır. KTK'de güvenirliğin nasıl kavramsallaştırıldığını anlayabilmek için üzerinde durulması gereken bir nokta da şudur: Daha önce değinildiği gibi KTK'ye göre gerçek puana ulaşmanın yolu bir testi aynı bireye sonsuz sayıda uygulayıp bu uygulamalardan elde edilen puanların ortalamasını almaktır. Peki, bir birey aynı ölçme aracını neden tekrar tekrar cevaplamak istesin ki? Burada sözü edilen tekrarlı uygulamalar KTK'de güvenirliğin nasıl hesaplandığını simüle etmek için son derece kullanışlı olmakla beraber pratik açıdan imkânsızdır (Kline, 2025). Aynı test aynı öğrencilere tekrar tekrar uygulandığında öğrenciler isyan eder veya aşırı derecede itaatkâr olsalar bile kısa sürede yorgunluktan bayılırlar (Popham, 2025). Bu koşullar altında güvenirlik hesaplamalarına nasıl devam edilebilir?

KTK'de güvenirlik formülasyonunda son derece önemli bir adım, psikometristlerin aynı kişiden birden fazla örneklem alarak (örneğin ölçme aracını aynı kişiye 1000 kez uygulayarak) geliştirilen gerçek ve hata puanları teorisinin, bir ölçme aracının birden fazla kişiye tek bir defa uygulanması durumunda da (yani, ölçme aracını 1000 farklı kişiden oluşan bir gruba bir defa uygulayarak) geçerli olduğunu göstermesiyle

gerçekleşmiştir. Böylece KTK'de güvenirliğin hesaplanması önemli ölçüde kolaylaşmıştır. Buna göre ölçme hatasını ve güvenirliği kestirebilmek için aynı ölçme aracını bir kişiye belli bir zaman dilimi boyunca tekrar tekrar uygulamak yerine, bu ölçme aracını yeterli büyüklükte bir örnekleme sadece bir kez uygulayıp ölçmenin standart hatasına ilişkin aynı değer elde edilebilir ve bu değer evrene genellenebilir (Kline, 2025). Bu da hata varyansını tahmin edilebilir hale getirdiğinden artık güvenirlik hesaplanabilir ve KTK'de bunun için Eşitlik 4'teki formül kullanılır (Shultz ve arkadaşları, 2021).

$$R = \frac{Var(T)}{Var(X)} = \frac{Var(X) - Var(E)}{Var(X)} = 1 - \frac{Var(E)}{Var(X)}$$

R = Güvenirlik katsayısı
$Var(X)$ = Gözlenen puana ilişkin varyans
$Var(T)$ = Gerçek puana ilişkin varyans
$Var(E)$ = Hata puanına ilişkin varyans

Eşitlik 4

Bu formüle göre güvenirlik katsayısının alabileceği değerler 0 ile 1 arasında uzanır. Ölçme sonucuna karışan hata azaldıkça güvenirlik katsayısı yükselir ve 1'e yakın değerler alır. Öte yandan ölçümlere karışan hata arttıkça güvenirlik katsayısı düşer ve sıfıra yaklaşır.

3.3. KTK'nin Sınırlılıkları

KTK, ölçme ve değerlendirmenin temel ilkelerinin anlaşılmasına önemli katkılar sunmuş olsa da çeşitli sınırlılıklar içermektedir (Thakur, 2025). Birey ve test özelliklerinin birbirinden bağımsız olarak ele alınamaması, daha açık bir ifadeyle bireye ve teste ilişkin özelliklerin yalnızca birbirleri bağlamında anlamlı biçimde yorumlanabilmesi KTK'nin en önemli sınırlılığıdır (Juve, 2008). Buna göre KTK'de bireyin yetenek düzeyi uygulanan testteki maddelerin parametrelerine bağımlı iken testteki maddeler için kestirilen parametreler testin uygulandığı örneklemdeki bireylerin yetenek düzeylerine göre değişebilmektedir. Aynı birey için kolay maddelerden oluşan bir testte yüksek, zor maddelerden oluşan bir testte ise düşük yetenek kestirimi yapılabilmektedir. Benzer şekilde bir test yüksek yetenek düzeyindeki bireylere uygulandığında maddelerin kolay olduğunu gösteren güçlük indeksleri elde edilirken bu test düşük yetenek düzeyindeki bireylere uygulandığında maddelerin zor olduğuna işaret eden güçlük indekslerine ulaşılabilmektedir.

KTK'ye ilişkin ikinci bir sınırlılık hataların tesadüfi olduğunun kabul edilmesi ve bireylerin gerçek puanına bağlı olmadığının varsayılmasıdır. Ancak gerçekte durum her zaman böyle değildir. Örneğin bireylerin bazı maddelerde hata yapma olasılığı daha fazla olabilmekte ya da bazı maddeler bir gruba avantaj sağlayarak bireyin yetenek puanının yanlı kestirimine yol açabilmektedir (Mir ve ark., 2025). Üçüncü olarak KTK ölçme hatasını ve güvenirliği testin uygulandığı tüm bireyler için aynı kabul eder. Ancak gerçekte durum farklıdır. Yüksek ve düşük yetenek düzeyindeki bireylerin gerçek

puanlarına ilişkin kestirimler, yetenek dağılımının orta bölgesinde bulunan bireylerin gerçek puanlarına ilişkin kestirimlere kıyasla genellikle daha düşük bir kesinlik düzeyine sahiptir. Dördüncü olarak KTK korelasyon temellidir ve gözlenen puanlar ile gerçek puan arasında doğrusal bir ilişki olduğunu varsayar. Ancak araştırmalar bu varsayımın daima karşılanmadığını göstermektedir (Shultz ve ark., 2021). KTK'ye ilişkin bu sınırlılıklar araştırmacıları söz konusu sınırlılıkların aşılmasına hizmet edebilecek yöntem arayışına sevk etmiştir. Mevcut ölçme uygulamalarını desteklemek, eksiklerini gidererek iyileştirmek ve gerektiği durumlarda değiştirmek için alternatif modeller üzerinde çalışan araştırmacılar, 1930'lu yılların sonucunda KTK'nin sınırlılıklarının üstesinden geleceği iddiasıyla örtük özellikler kuramı olarak da anılan MTK'yi ileri sürmüştür (Doğan & Tezbaşaran, 2003).

4. MADDE TEPKİ KURAMI

KTK için yukarıda açıklanan sınırlılıkların önüne geçebilmek, daha doğru yetenek kestirimleri ile daha geçerli ve güvenilir test istatistiklerine ulaşabilmek amacıyla yetenek-madde etkileşimini ön plana çıkaran MTK geliştirilmiştir.

4.1. KTK ile MTK'yi Ayıran Özellikler

KTK ölçülen bireyin toplam puanı üzerinden bir yetenek kestirimi yapmaktadır. Bu yaklaşım işlevsel ve tutarlı olsa da madde güçlüğünü ölçümlerde doğrudan bir kriter olarak değil bir varsayım olarak kabul etmektedir. KTK'de testin her bir maddesinden alınan puanların toplamı bireyin ham puanını vermektedir. MTK'de ise bireyin testten aldığı ham puan yerine her bir maddeye verdiği cevabın doğru veya yanlış olmasına odaklanılır. Çünkü MTK testten alınan toplam puana değil testin her bir maddesine dayalıdır (Baker, 2016). Ülkemizde, özellikle ÖSYM ya da MEB tarafından yapılan sınav uygulamaları göz önünde bulundurulduğunda bir testte daha zor ve ayırt ediciliği daha yüksek maddeleri doğru cevaplamanın testten alınan puanı etkilediğine ilişkin yanlış bir inanış vardır. Örneğin, 20 soruluk bir testin en zor ve en ayırt edici 5 sorusunu doğru yanıtlayan bir kişinin aynı testin en kolay ve düşük ayırt edicilikteki 5 maddesini doğru yanıtlayan kişiden daha yüksek puan alacağı düşüncesi tamamen yanlıştır. İki öğrenci de toplamda 5 soruya doğru cevap verdikleri için tam olarak aynı puanı alacaklardır. Çünkü KTK, zor maddeleri doğru cevaplayabilecek yetenek düzeyindeki öğrencinin kolay soruları da doğru cevaplayacağını kabul eder (Lord & Novick, 2008). Bu varsayıma örnek olarak bir kişi eğer 100 kg kütleli bir cismi kaldırdıysa bu kişinin 50, 60, 70 kg kütleli cisimleri de kaldırması beklenir. Bu durumda eğer farklı kütlelere sahip 10 cismin kaldırılması gereken bir test yapılırsa 100 kg kütleli bir cismi kaldıran bir kişinin en fazla 70 kg kütleli bir cismi kaldırabilen kişiden daha fazla başarılı kaldırış gerçekleştireceği varsayılır. Bu varsayım aslında oldukça mantıklı ve tutarlıdır. Bu nedenle KTK, standart test puanlarını hesaplarken ya da bir testin standardizasyonunu yaparken toplam puan üzerinden analizler gerçekleştirir. Bir diğer ifade ile öğrencinin toplam puanını, o öğrencinin gerçek yetenek düzeyi olarak belirler.

MTK ise yetenek kestirimi yaparken öğrencinin yeteneğini o öğrenciye en uygun olan madde ile eşleştirir. Önceki örnek üzerinden konuşacak olursak şans eseri olmadan ve tekrarlanabilecek şekilde kişinin kaldırabileceği en uygun kütlenin belirlenmesinin odağa konulduğu söylenebilir. Sonuç olarak bireyin 100 kg civarındaki bir kütleyi kaldırabilme olasılığının yüksek olması, onun bu alandaki kapasitesinin göstergesi olup yetenek düzeyi bu değer üzerinden tanımlanır. Birçok üstün yönü bulunan bir yaklaşım olmasına rağmen MTK'nin de birtakım sınırlılıklarının bulunduğuna bölümün ilerleyen kısımlarında değinilecektir.

Önceki örnekte, madde güçlüğünü temsil ettiği düşünülen kütlelerin aslında kesin olarak bilinen ve doğru ölçümler olduğu aşikârdır. Ancak eğitimde ve psikolojide ölçmeye konu olan bir özellik ölçülürken madde güçlüğü, o maddeyi doğru cevaplayanların oranını vermektedir. Bir diğer ifade ile madde güçlüğü, kesin değil gruba göre değişen bir istatistiktir. KTK bir maddenin güçlük düzeyini gruba bağlı belirlediği için aynı maddenin başarılı gruplarda daha kolay, başarısız gruplarda ise daha zor belirlenmesine yol açmaktadır. Bahse konu problemin MTK tarafından çözümü aslında bu modelin bir diğer özelliğine işaret etmektedir. MTK'de maddelere ilişkin güçlük ve ayırt edicilik değerleri, veriden hesaplanan bir istatistik değil, modelin yapısını tanımlayan parametreler olarak kabul edilir (Hambleton & Swaminathan, 2013; Lord, 1980). Bunun anlamı, madde parametrelerinin gruptan bağımsız matematiksel değerler olmasıdır. Yani yeterli büyüklükte bir örneklem verilerine dayalı olarak hesaplanması koşuluyla kestirilen güçlük ve ayırt edicilik parametreleri, test farklı gruplara uygulansa da örneklem bağımsızlığı özelliği nedeniyle değişmeyecektir. Bu sebeple MTK analizlerinde çok büyük bir örnekleme ihtiyaç duyulduğu göz ardı edilmemelidir. Diğer bir deyişle MTK açıklanırken kuramın dayandığı varsayımlar mutlaka göz önünde bulundurulmalıdır.

4.2. MTK Varsayımları

4.2.1. Monotonik Artış

Bu varsayımına göre bireyin sahip olduğu gizil özellik/yetenek düzeyi yükseldikçe testteki herhangi bir maddeyi doğru cevaplama olasılığı artar (DeMars, 2010; Hambleton & Swaminathan, 2013). Bu varsayım aslında KTK'de de karşımıza çıkar. Bir konuda daha fazla bilgisi olan ya da belli bir özelliğe/yeteneğe daha fazla sahip olan bir kişinin bu konu ile ilgili soruları doğru yanıtlama olasılığının daha yüksek olduğu kabul edilir.

4.2.2. Tek Boyutluluk

Model, ölçmede kullanılan aracın baskın tek bir gizil özelliği (yetenek ya da beceri) ölçtüğünü kabul eder. Bir başka deyişle, testte yer alan tüm maddelerin aynı örtük özelliği ölçtüğü ve aynı bilişsel süreçle çözüldüğü varsayılır (Embretson & Reise, 2013). İçinde hem fizik hem de biyoloji konularının olduğu bir test düşünelim. Bu testi alan bir öğrencinin fizik sorularını çözerken kullandığı bilişsel beceri ile biyoloji sorularında kullandığı beceri arasında belirgin farklar bulunmaktadır. Bu durumda bu testin tek boyutlu olduğunu söylemek doğru olmayacaktır. Bu varsayım monotonik artış varsayımı gibi KTK'de de kabul edilmektedir.

4.2.3. Yerel Bağımsızlık

Bireyin yetenek düzeyi (θ) sabit tutulduğunda bir testteki farklı maddelere verilen yanıtların birbirinden bağımsız olduğu varsayılır (Embretson & Reise, 2013; Hambleton & Swaminathan, 2013; Lord, 1980). Bu varsayım KTK için de geçerlidir. Şöyle ki KTK'de de bir maddeye verilen yanıtın bir başka maddeyi olumlu ya da olumsuz yönde etkilememesi esastır. Yani testi alan kişi her maddede birbirinden bağımsız bir ölçümle karşılaşmalıdır. Bu durum ortak köklü maddeler için kafa karıştırıcı olabilir. Aynı köke dayalı birden fazla soru mevcut ve bu sorulardan birinin doğru cevaplanması bir diğerinin doğru cevaplanmasını kolaylaştırıyorsa yerel bağımsızlık varsayımı ihlal edilmiş demektir. Tabi ortak köke dayalı tüm sorularda böyle bir sorun olmayabilir, birbirinden bağımsız cevaplanabilecek ortak köklü sorular hazırlamak pek tabi mümkündür.

4.2.4. Değişmezlik

Bir maddenin parametreleri, kişinin yetenek düzeyi ile maddeye doğru cevap verme olasılığı arasındaki ilişkiyi gösteren Madde-Yanıt Eğrisi'nin herhangi bir noktasından tahmin edilebilir. Başka bir deyişle madde parametreleri yalnızca belli bir örnekleme özgü değildir; maddeyi yanıtlayan farklı gruplardan elde edilen veriler üzerinden de aynı parametreler tahmin edilebilir (Embretson & Reise, 2013). Daha önce belirtildiği gibi madde parametrelerinin gruptan bağımsız olması MTK'nin temel varsayımlarından biridir.

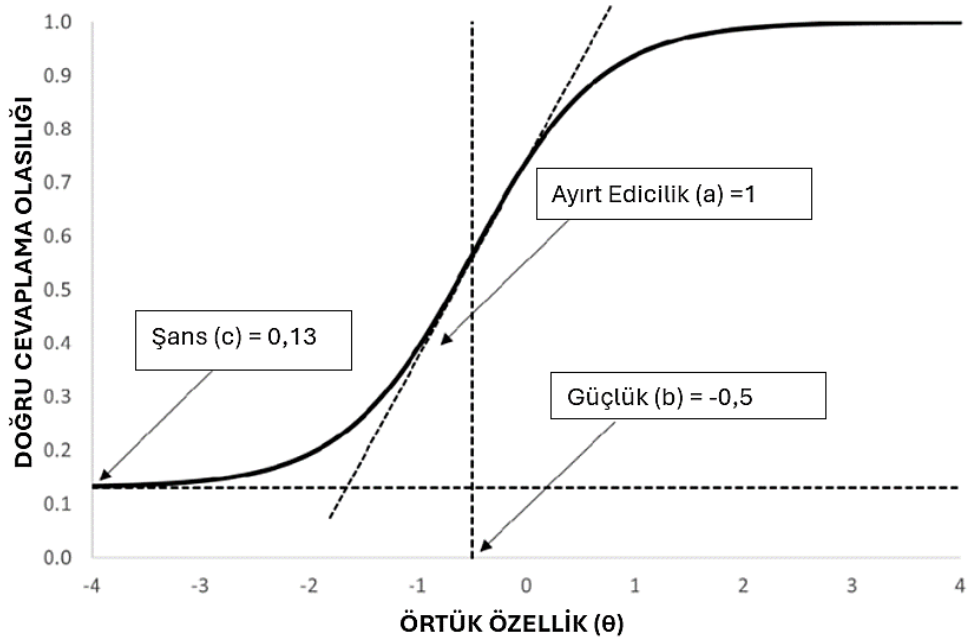
Hambleton ve arkadaşları (1991), MTK'de bir kişinin bir test maddesindeki ediminin-eyleminin, örtük özellik ya da yetenek olarak tanımlanabilecek bir faktör aracılığıyla tahmin edilebileceğini ve aynı zamanda bir kişinin bir maddedeki edimi-eylemi ile o maddeye cevap vermesini sağlayan özellik arasındaki ilişkinin Madde Karakteristik Eğrisi (MKE) olarak tanımlanan bir fonksiyonla gösterilebileceğini belirtmiştir. MKE'nin mantığını ve doğasını anlamak MTK'nin yapısını anlamanın öncelikli koşuludur.

4.3. Madde Karakteristik Eğrisi

MTK modelleri katılımcıların bir ölçeğe verdikleri yanıtları belli bir örtük özelliğe sahip oluş düzeyleri ve madde parametrelerine göre tahmin eder. Bu ilişki matematiksel olarak Madde-Tepki Fonksiyonu ile gösterilir ve bir ölçeğin her maddesine verilen yanıtların ölçülen örtük özellik hakkında fikir verdiği varsayımına dayanır. Kişinin yetenek düzeyi (θ), doğrudan bir olasılık değildir; fakat yetenek arttıkça bir maddeyi doğru yanıtlama olasılığı monoton olarak artar. Yetenek ile maddeyi doğru yanıtlama olasılığı arasındaki bu ilişki Madde Karakteristik Eğrisi (MKE) adı verilen grafikle gösterilir (Duncan & MacEachern, 2008). Teorik olarak bir bireyin yetenek düzeyi (θ) -∞ ila +∞ aralığında yer alır ancak uygulamada genellikle -3 ila + 3 aralığında değişir. Şekil 1'de bir madde için oluşturulmuş MKE örneği verilmiştir.

Şekil 1

Örnek MKE

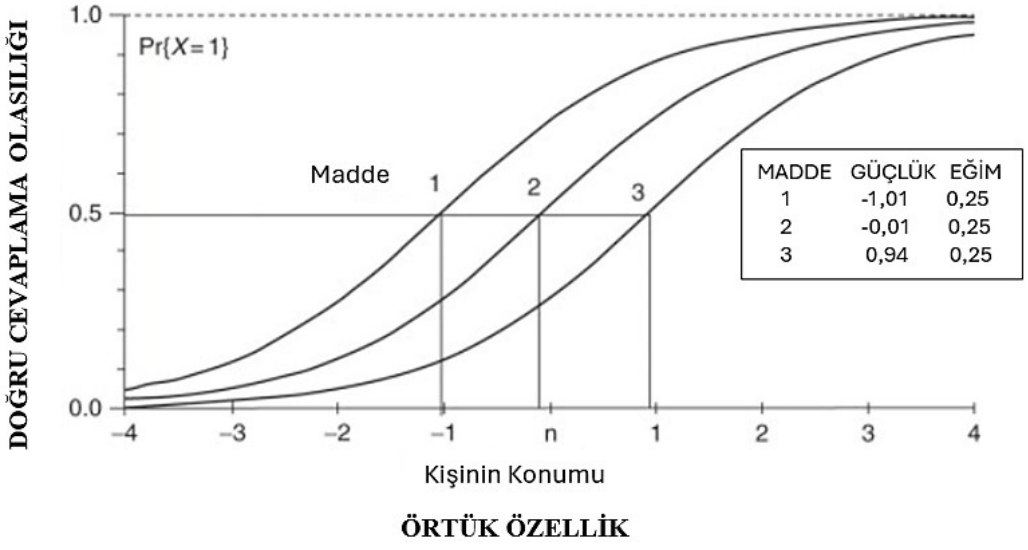


Şekil 1'deki grafik incelendiğinde madde güçlük parametresinin (*b* parametresi) –0.5 düzeyinde olduğu görülmektedir. Bu değer, fonksiyonda doğru cevaplama olasılığının 0.5'e (%50'ye) karşılık gelen örtük özellik (θ) düzeyinin değeridir.

MTK modelinde madde güçlüğü (b_i), maddenin yetenek skalası boyunca nasıl çalıştığını belirleyen bir parametredir. Olasılığın 0.5 olduğu, yani yanıtlayıcıların doğru cevap verme olasılığının %50 olduğu nokta, referans noktası olarak alınır. MKE'de zor maddeler, eğrinin skalanın sağına kaymasıyla temsil edilir. Bu durum, ilgili maddeleri doğru yanıtlayan kişilerin daha yüksek yetenek düzeyine sahip olduklarını gösterir. Buna karşılık kolay maddeler, eğrinin sola kaymasıyla gösterilir. Şekil 2 farklı yetenek düzeyleri için madde fonksiyonunun nasıl değiştiğini göstermektedir.

Şekil 2

Farklı Güçlük Düzeyleri İçin MKE

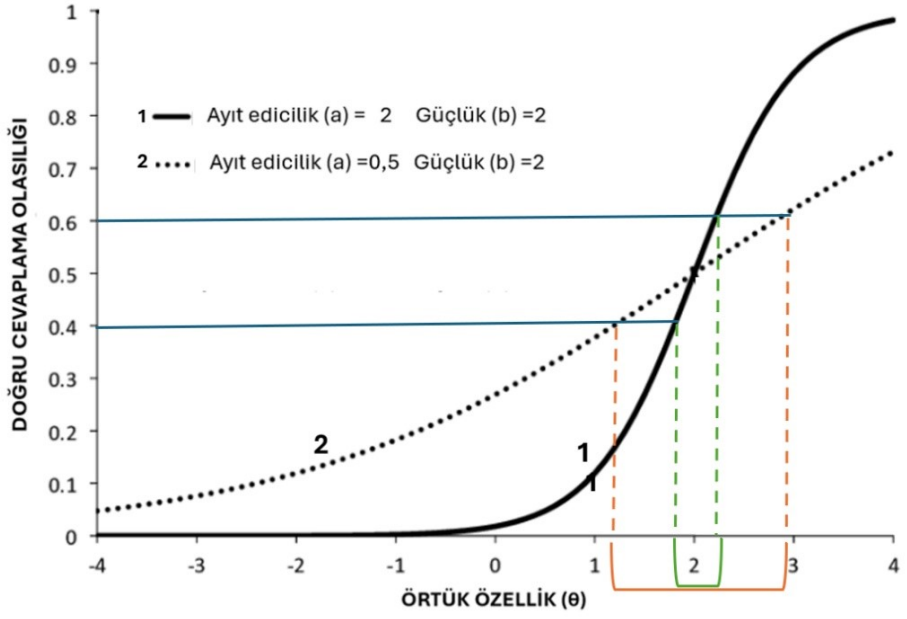


Şekil 2'deki MKE'ler hem maddelere ait farklı güçlük düzeylerini hem de farklı yetenek düzeylerindeki öğrencilerin konumlarını göstermektedir. Bu durum, MTK'de madde-yetenek etkileşiminin bir göstergesidir. Öte yandan 2. madde orta düzeyde güçlüğe sahip olup maddeler arasında en zor olanı 3., en kolay olanı ise 1. maddedir.

Madde ayırt edicilik parametresi "*a* parametresi" olarak tanımlanır ve MKE'nin eğimini ifade eder. Şekil 1'de MKE'nin en dik olduğu yerdeki eğimi 1.00 olup bu maddenin *a* parametresi de 1.00 olarak hesaplanır. Eğim arttıkça madde, sadece belli seviyedeki öğrencileri daha iyi ayırt eder; maddeyi doğru cevaplama olasılığı da bu küçük aralıkta değişir. KTK'deki madde ayırt edicilik indeksinde olduğu gibi MTK'deki *a* parametresinin değerindeki artış madde ayırt ediciliğinin arttığı anlamına gelir. *a* parametresi teorik olarak -∞ ile +∞ arasında değer alabilirken pratikte genellikle 0 ila +2 aralığında değer alır. Farklı ayırt edicilik değerlerine sahip iki madde için MKE'ler Şekil 3'te verilmiştir.

Şekil 3

Farklı Ayırt Edicilik Düzeyleri İçin MKE



Şekil 3 incelendiğinde 1. maddenin yüksek düzeyde ayırt edici olduğu görülmektedir. Yeşil çizgi ile gösterilen bölümden de görülebileceği üzere bu madde, belirli bir yetenek düzeyi için kestirim yapmaktadır. Başka bir deyişle, bu madde ile eşleşen yetenek düzeyi +2 civarında dar bir aralıkta yoğunlaşmaktadır. 2. maddeye bakıldığında ise turuncu renkle gösterilen bölümden bu maddeyi doğru cevaplayan bir öğrencinin yetenek düzeyinin yaklaşık +1 ile +3 arasında herhangi bir yerde olabileceği anlaşılmaktadır. Bu durum, maddenin ayırt ediciliğinin düşük olduğuna işaret eder. MTK'de bir maddenin yüksek ayırt edicilikte olması, MKE'nin eğiminin yüksek olduğunu bildirir. Bu kitabın 11. bölümünde MTK'ye dayalı parametre kestirimlerine yönelik detaylı açıklamalar vardır.

4.4. MTK'de Güvenirlik ve Ölçme Hatası

MTK'nin en önemli katkılarından biri, ölçümlerin hangi yetenek düzeylerinde daha güvenilir olduğunu gösterebilmesidir. KTK'de güvenirlik, testin tümüne ait tek bir katsayıyla (örneğin, Cronbach alfa) ifade edilir. Bu değer, testin farklı yetenek düzeylerinde ne kadar hassas ölçüm yaptığını göstermez. Oysa MTK'de her bir madde ve test için bilgi fonksiyonları hesaplanabilir.

Test Bilgi Fonksiyonu (Test Information Function – TBF), bir testin farklı yetenek düzeylerinde ne kadar bilgi sağladığını gösterir. Yani test, hangi yetenek aralığında daha hassas ölçüm yapıyor sorusunun cevabını verir. Matematiksel olarak bilgi fonksiyonu, yanıtların varyansı ile doğrudan ilişkilidir; varyans arttıkça test daha fazla bilgi üretir (Embretson & Reise, 2013; Hambleton & Swaminathan, 2013).

Bilgi fonksiyonu ile ilişkili en önemli kavramlardan biri de ölçmenin standart hatasıdır (ÖSH). ÖSH, yetenek kestiriminin ne kadar isabetli olduğunun göstergesidir ve bilgi fonksiyonunun tersi ile hesaplanır. Yani testin sağladığı bilgi arttıkça, ölçüm hatası azalır (Baker & Kim, 2004). Bu nedenle bilgi fonksiyonunun yüksek olduğu yetenek düzeylerinde test daha güvenilir, düşük olduğu bölgelerde ise daha az güvenilirdir. Grafiklerle gösterildiğinde şu yorumu yapmak mümkündür: Eğer bir testin bilgi fonksiyonu -1 ile +1 arası yetenek düzeyinde yüksek çıkıyorsa, bu test özellikle orta düzey yeteneklerde daha hassas ölçüm yapıyor demektir. Benzer şekilde, ileri düzey öğrencileri seçmek isteyen bir sınavın bilgi fonksiyonunun +2 ve üzerindeki bölgelerde yüksek olması beklenir. Sonuç olarak TBF ve ÖSH birlikte incelendiğinde her yetenek düzeyinde ölçmenin ne kadar güvenilir olduğu anlaşılır. Bu yaklaşım, özellikle sınavların amaçlarına uygunluğunu değerlendirmek için önemli bir araçtır. Nitekim uluslararası sınavlarda (PISA, TIMSS, GRE vb.) ve Bilgisayar Ortamında Bireye Uyarlanmış Testlerde (BOBUT), TBF ve ÖSH analizleri standart bir uygulamadır (OECD, 2017; van der Linden & van der Linden, 2016).

4.5. MTK Modelleri

MTK özünde tek bir model değil bir modeller ailesidir. Ölçülen özelliğin yapısı ve kullanılan testin türüne göre MTK çatısı altında yer alan çeşitli yaklaşımlar vardır. Daha açık bir anlatımla ölçülen özelliğin boyutluluğuna (tek boyutlu mu, çok boyutlu mu), madde cevaplarının iki kategorili (dikotomik [1-0]) ya da çoklu cevap formatlarında olmasına ya da ölçme sonuçlarında farklılığa yol açabilecek değişkenlik kaynaklarının sayısına göre geliştirilmiş farklı MTK modelleri bulunmaktadır (İlhan & Güler, 2018). Alanyazında MTK modellerini sınıflandırırken esas alınan değişkenlerden biri de maddeler için kestirilen parametrelerin sayısıdır (Birnbaum, 1968; Lord & Novick, 2008). Bölümün devamında bu modeller tanıtılacaktır.

4.5.1. 1 Parametreli Lojistik Model

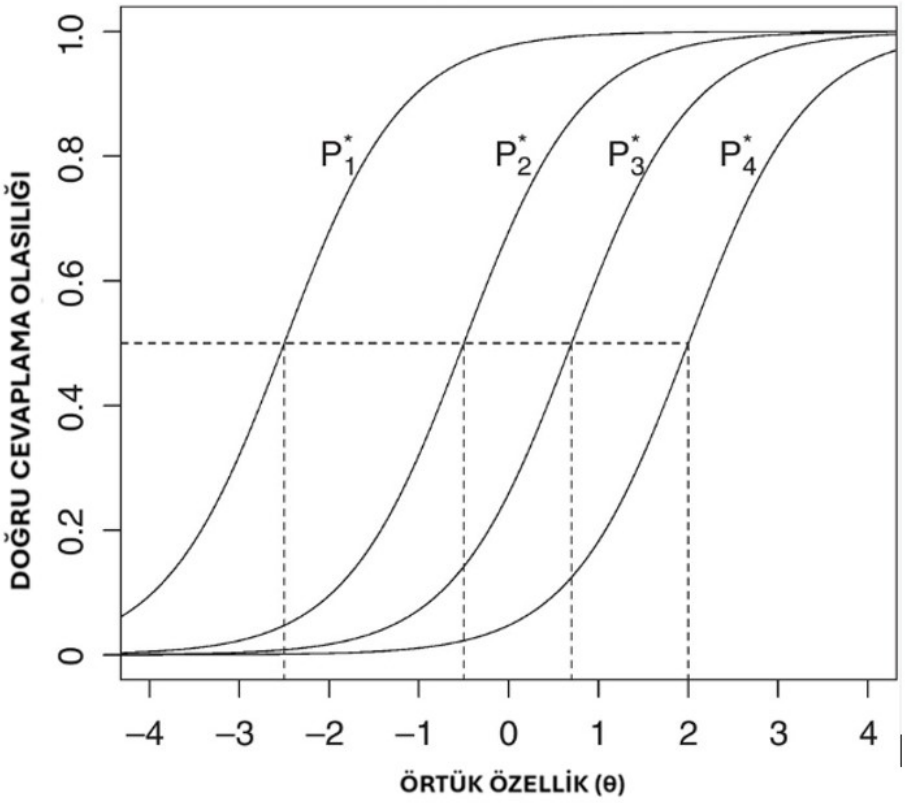
MTK modellerinin en basiti olan 1 Parametreli Lojistik Model (1PLM) maddelere yanıt veren kişinin örtük özelliğini (yetenek – θ) tanımlayan parametre ile maddenin zorluk düzeyini gösteren b (güçlük) parametresi üzerine kurulmuştur. Aşağıda 1PLM için kullanılan eşitlik verilmiştir (Baker & Kim, 2004).

$$P_i(\theta) = \frac{e^{1,7(\theta - b_i)}}{1 + e^{1,7(\theta - b_i)}} \qquad \textit{Eşitlik 5}$$

Bu denklem θ yetenek düzeyine sahip herhangi bir öğrencinin i maddesini doğru cevaplama olasılığını ifade etmektedir. Eşitlikteki b_i, i maddesine ait güçlük parametresini göstermektedir. Görüldüğü gibi bu model sadece madde güçlüğünü bir değişken olarak almaktadır. Modelde ayırt edicilik parametresi bütün maddeler için sabittir ve kümülatif normal ogive dağılımının yaklaşık eğimi (1,7) olarak kabul edilmektedir. Bu nedenle her madde için MKE'ler paralel olarak çizilmektedir. Şekil 4, 1PLM için dört ayrı maddenin MKE'sini göstermektedir.

Şekil 4

Farklı Madde Güçlükleri İçin 1PLM'de MKE'ler



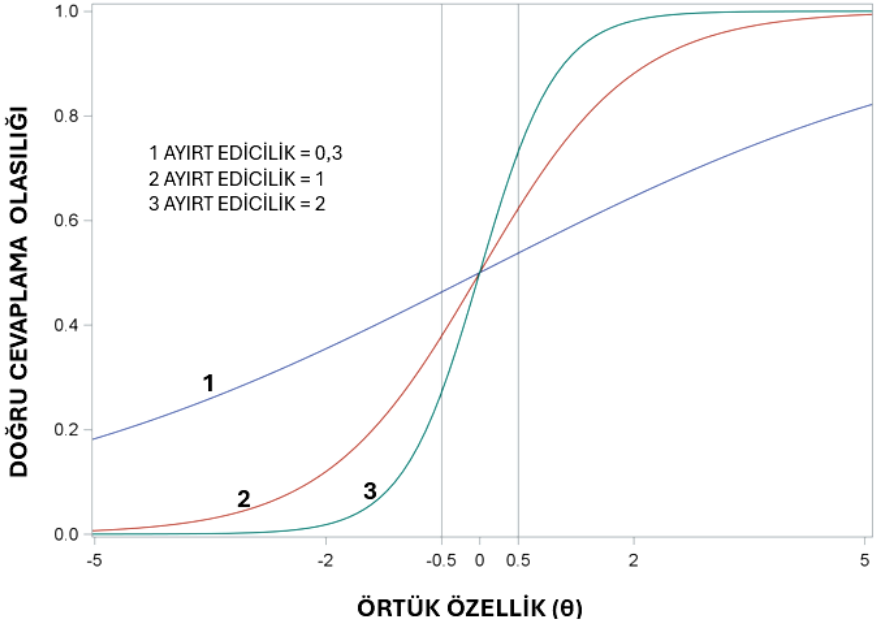
Şekil 4'te dört farklı yetenek düzeyini temsil eden MKE'ler verilmiştir. Eğimleri yani ayırt edicilikleri aynı olan bu maddelerin değişen tek özelliği maddeyi doğru yanıtlama olasılığının 0.5'ye karşılık geldiği noktadır. Bu noktalar ise farklı yetenek düzeylerine ya da o yetenek düzeylerine sahip öğrencileri temsil eden maddelere işaret etmektedir.

4.5.2. 2 Parametreli Lojistik Model

2 Parametreli Lojistik Model (2PLM), bir öğrencinin doğru maddeye cevap verme olasılığını hesaplarken maddenin hem güçlüğünü hem de ayırt ediciliğini kullanır (Baker & Kim, 2004). Şekil 5'te aynı güçlük düzeyi için farklı ayırt edicilikte üç maddenin MKE'leri çizilmiştir.

Şekil 5

2PLM'de Farklı Ayırt Edicilikteki Maddeler İçin MKE



Şekil 5'te görüldüğü gibi aynı güçlük düzeyinde olan 1, 2 ve 3. maddelere ait MKE'lerin eğimleri farklıdır. Dolayısıyla bu maddelerin ayırt edicilik değerleri de farklıdır. Bu durum, (-0,5, 0,5) aralığındaki yetenek düzeyini belirlemede 3. maddenin diğer maddelere kıyasla daha etkili olduğunu göstermektedir. Bu model θ yetenek düzeyindeki bir bireyin maddeyi doğru cevaplama olasılığını belirlemek için aşağıdaki eşitliği kullanır.

$$P_i(\theta) = \frac{e^{a_i(\theta - b_i)}}{1 + e^{a_i(\theta - b_i)}} \qquad \textit{Eşitlik 6}$$

Eşitlik 6'da görüldüğü üzere modelde doğru cevaplama olasılığı kestirilirken maddenin hem güçlük (b_i) hem de ayırt edicilik (a_i) parametreleri kullanılmaktadır.

4.5.3. 3 Parametreli Lojistik Model

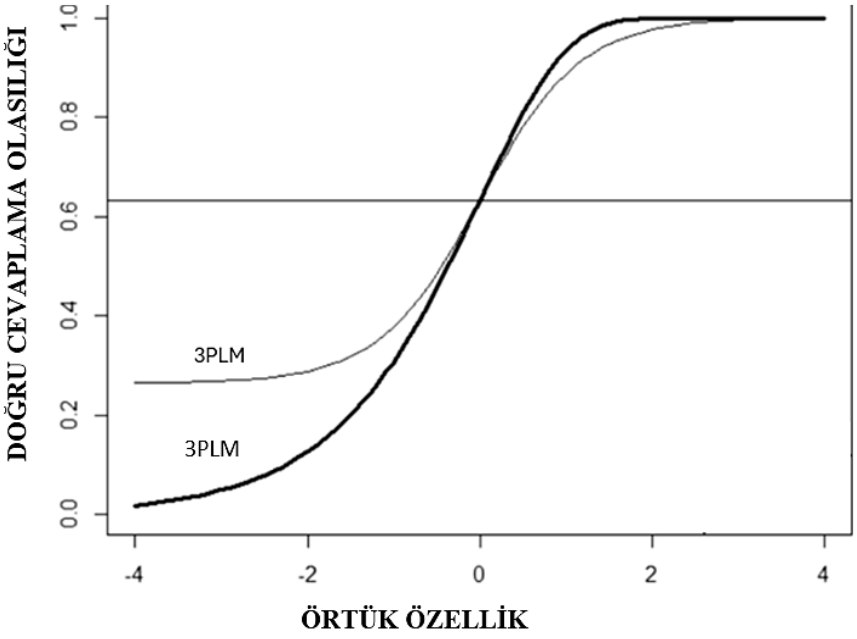
Bu model, 2PLM'ye şans (c) parametresinin eklenmesi ile oluşturulmuştur. Yani 3 Parametreli Lojistik Model'in (3PLM) hesaplamalarında güçlük (b), ayırt edicilik (a) ve şans (c) parametresi kullanılmaktadır (Baker & Kim, 2004). Modelin eşitliği aşağıdaki gibidir.

$$P_i(\theta) = c_i + (1 - c_i)\frac{e^{-a_i(\theta - b_i)}}{1 + e^{-a_i(\theta - b_i)}} \qquad \textit{Eşitlik 7}$$

Eşitlik 7'de 1 ve 2 parametreli modellerden farklı olarak *c* parametresi karşımıza çıkmaktadır. Bu parametre hesaplanırken öncelikle maddenin güçlüğüne karşılık gelen yetenek düzeyinin altında kalan öğrencilerin maddeyi doğru cevaplaması referans alınmaktadır. 3 PLM'de MKE'nin nasıl değiştiği Şekil 6'da gösterilmiştir.

Şekil 6

3PLM İçin MKE



Şekil 6'da görüldüğü üzere şans parametresinin modele eklenmesi, eğrinin şeklini etkilemektedir. Özellikle şans parametresiyle birlikte ayırt edicilik parametresinin değerinin düşeceği görülmektedir. Bu modelin şans başarısını kontrol ederek öğrencinin yetenek kestirimini daha doğru şekilde yapacağını iddia etmek mümkündür. Ancak yapılan araştırmalar ve genel olarak tercih edilen uygulamalara bakıldığında 1PLM ve 2PLM'nin daha yaygın olarak kullanıldığı gözlenmektedir.

4.5.4. 4 Parametreli Lojistik Model

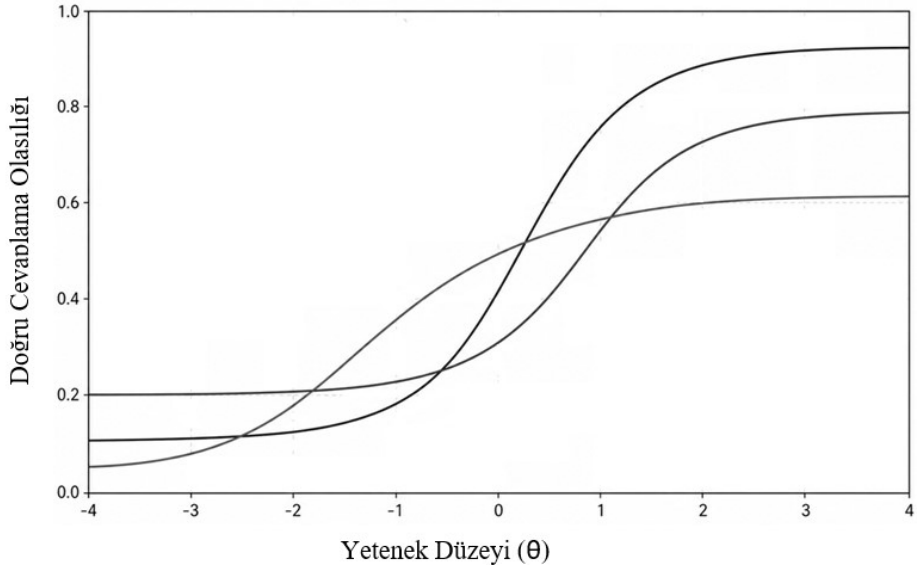
Dört parametreli lojistik model (4PLM) güçlük, ayırt edicilik ve şans parametrelerinin yanı sıra madde karakteristik eğrisinin üst asimptotuna karşılık gelen dikkatsizlik parametresini (*d*) de içermektedir. Yani bu model yüksek yetenek düzeyine sahip bireylerin dikkatsizlik ya da benzeri nedenlerle doğru yanıtlamaları beklenen maddeleri hatalı yanıtlama olasılığını modele dâhil etmek amacıyla geliştirilmiştir (Paek & Cole, 2020). Modele ilişkin eşitlik (Cheng & Liu, 2015) aşağıda verilmiştir.

$$P_i(\theta) = c_i + (d_i - c_i)\frac{e^{-a_i(\theta - b_i)}}{1 + e^{-a_i(\theta - b_i)}} \qquad \textit{Eşitlik 8}$$

Eşitlikte görüldüğü üzere 3PLM'deki $(1 - c_i)$ terimi 4PLM'de $(d_i - c_i)$ olarak değişmiştir. Böylece yetenek (θ) ölçeği boyunca maddeyi doğru yanıtlama olasılığının ulaşabileceği en yüksek artık değer 1'e değil d parametresine karşılık gelmektedir. Modelin geçerli olabilmesi için $c < d \leq 1$ olmalıdır. Diğer bir deyişle d parametresinin alabileceği değerler (c, 1] aralığında uzanır (Meng ve ark., 2019). 4PLM için MKE örnekleri Şekil 7'de resmedilmiştir.

Şekil 7

4PLM İçin MKE



Şekil 7'de görüldüğü üzere üç maddenin üst asimptotları yani maddeyi doğru yanıtlama olasılığına ilişkin en yüksek değerler birbirinden farklıdır. Yine maddelerin güçlük (b), ayırt edicilik (a) ve şans (c) parametrelerinin de birbirinden farklı olduğu görülmektedir.

4.6. MTK ile KTK'nin Karşılaştırılması

MTK, KTK'nin içerdiği bazı sınırlılıkları matematiksel bir modelleme yöntemiyle çözen tutarlı bir test kuramı olarak değerlendirilebilir. Daha önce belirtildiği gibi KTK'de bireylerin yetenek kestiriminde madde güçlük ve ayırt edicilik değerleri doğrudan bir rol oynamaz (Birnbaum, 1968; Hambleton ve arkadaşları, 1991; Lord & Novick, 2008). Bu durum, özellikle karmaşık yapıları ölçen testlerde bazı soru işaretlerini doğurmaktadır. Bunu bir örnekle açıklamak gerekirse bir sınavda matematik başarısını ölçen tüm maddelerin eşdeğer olduğunu varsaymak ve bu maddelerin ölçtüğü özellik hakkında verilen kararların bütünüyle gerçeği yansıttığını iddia etmek çoğu durumda geçerli değildir.

Bununla birlikte, KTK'de madde güçlüğü ve ayırt ediciliği testin genel güvenirliği ile geçerliğine dolaylı bir kanıt olarak kullanılır. Halbuki MTK'de bu parametreler, her bir yetenek düzeyi için ayrı ayrı değerlendirilebilir. Böylece yalnızca testin genel güvenirliği

ve geçerliği değil, aynı zamanda her bir maddenin belirli yetenek düzeylerinde ne kadar güvenilir ölçüm yaptığı da ortaya konur (Hambleton & Swaminathan, 2013; Van der Linden & van der Linden, 2016). Buna bağlı olarak MTK, özellikle psikometrik çalışmalarda, belli yetenek grupları için belli çıkarımları yapabilme konusunda farklı yaklaşımlarda bulunmaya olanak tanır.

MTK'nin bütün bu güçlü özellikleriyle birlikte kullanım alanlarının KTK kadar yaygın olmamasının önemli sebeplerinden biri, parametrelerin hesaplanması için büyük örnekleme ihtiyaç duyulmasıdır. Küçük örneklemlerde MTK'nin uygulanabilirliğini inceleyen birçok araştırma bulunmakla birlikte, sınıf içi ölçmelerde ve küçük gruplarda güvenilir sonuçlar verdiğini söylemek henüz mümkün değildir (Fotaris ve ark., 2011; Houts ve ark., 2018; Lawson, 2006; LeBeau & McVay, 2017). Bununla beraber büyük gruplara uygulanan geniş ölçekli testlerde (PISA, TIMSS, PIRLS, GRE vb.) MTK'nin sıkça kullanıldığı görülmektedir (Educational Testing Association, 2014; Mullis ve ark., 2012; OECD, 2017).

Bunun yanı sıra MTK günümüzde BOBUT için zorunlu bir altyapı sağlamaktadır. BOBUT uygulamalarında her bireye kendi yetenek düzeyine uygun sorular yöneltilir. Bu süreçte MTK'nin sağladığı madde parametreleri (güçlük, ayırt edicilik, şans) ve yetenek kestirimleri kritik rol oynamaktadır (Van Der Linden & Glas, 2010). Bu sayede hem daha kısa testlerle güvenilir ölçmeler yapılabilmekte hem de öğrencilerin test deneyimleri iyileştirilmektedir.

MTK, tıpkı KTK gibi bilimsel olarak hipotezleri kanıtlanmış sağlam matematiksel, istatistiksel ve psikometrik özellikler barındıran bir test kuramıdır. Bu anlamda MTK ile ilgili olan bazı sınırlılıkların bu kuramın bilimsel tutarlılığını bozmadığını belirtmekte yarar vardır. MTK çoğu açıdan bir testin, özellikle de maddelerin yapısını anlama, psikometrik özelliklerini belirleme ve bu sayede o testle ilgili çıkarımları daha tutarlı yapabilme olanağı sağlamaktadır.

5. BİLİŞSEL TANI MODELLERİ

KTK ve MTK akademik ve psikolojik testlerde bireylerin performanslarını mümkün olduğu kadar güvenilir ve geçerli şekilde ölçmek amacıyla ortaya çıkmış, süreç içinde farklı yaklaşımlarla birlikte geliştirilmiştir. Testler ile testlere duyulan ihtiyacın artması, ölçülmesi hedeflenen özelliklerin giderek karmaşıklaşması ve bu modellerin temel varsayımlarıyla çelişen durumların ortaya çıkması, araştırmacıları yeni yöntemler üzerinde çalışmaya yöneltmiştir. Özellikle tek boyutluluk varsayımı ve test maddelerinin tek bir özelliği ölçmesi kabulü bu iki varsayıma da ihtiyaç duymayan Bilişsel Tanı Modellerine (BTM) ilgiyi arttırmıştır.

BTM öğrencide belli bir bilginin yapısını ya da bir becerinin gelişimini, öğrencinin güçlü ve zayıf yönlerini dikkate alarak hesaplamak amacıyla ileri sürülmüş (Leighton & Gierl, 2007), 2001 yılında Amerika'da başlayan "No Child Left Behind" hareketinden sonra yaygınlık kazanmıştır. Bu yaklaşımın amacı öğrencilere, öğretmenlere ve ailelere bilişsel dönüt verebilmektir. Geleneksel yaklaşımlara dayalı testlerin sunduğu geri bildirmeler testin geneline ya da alt testlerine ilişkin toplam puan düzeyinde kalmaktadır. BTM ise toplam puanların ötesine geçip her bir öğrencinin profiline, hangi özelliklere sahip olduğuna ve hangi özellikler konusunda öğrenme eksikliklerinin olduğuna dair detaylı dönütler sunabilmektedir (Cheng & Chang, 2007).

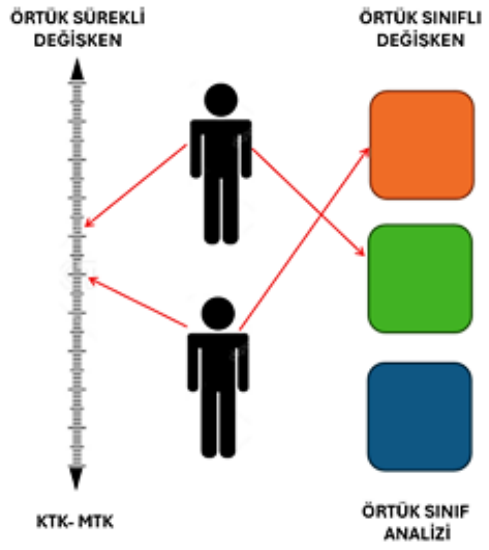
Bu noktada BTM'nin iki önemli özelliğine dikkat edilmelidir. Öncelikle BTM öğrenciyi sahip olduğu özellik bakımından sürekli bir değişken olarak raporlamayı, yani öğrencinin ölçülen özelliğine ait bir puan hesaplamayı öncelemez. Öğrencinin ilgili özelliğe sahip olup olmadığına odaklanır. Alanyazında öğrencinin hedef özelliğe sahip olma düzeyini üç ya da daha fazla sınıfla belirtmeye yönelik güncel araştırmalar bulunmakla birlikte sonuçlar yine çoğunlukla kategoriktir. İkinci olarak BTM henüz ne KTK ne de MTK gibi hipotezleri doğrulanmış, olası durumlar için bilimsel olarak sınanmış bir teori değildir. BTM hakkındaki birçok konu hala araştırmacılar tarafından deneysel olarak çalışılmakta ve sınanmaktadır. Bütün bunlarla birlikte, BTM'nin madde ve ölçülen özellik arasındaki ilişkiye yönelttiği yaratıcı bakış, test ve madde ilişkisine dair sunduğu yenilikçi yaklaşım ve KTK ile MTK'nin sınırlı kaldığı bazı konularda öne sürdüğü istatistiksel yaklaşımlar, psikometri alanına farklı bir perspektif kazandırmıştır. BTM'nin temel özelliklerini daha anlaşılır kılacağı düşüncesiyle aşağıda önce bu modellerin istatistiksel alt yapısını oluşturan "örtük sınıf analizi" yaklaşımı açıklanmıştır.

5.1. Örtük Sınıf Analizi

Örtük sınıf analizinin altında yatan düşünce, gözlenen alt gruplar arasındaki farkı belirleyen varsayımsal bir istatistik modelin parametrelerdir (Vermunt & Magidson, 2002). Bu alt gruplar örtük özelliğe ait kategorilerdir. Örtük sınıf kavramı 1950 yılında Lazarsfeld tarafından ortaya atılmıştır. Lazarsfeld iki kategorili gözlenen değişkenlerin tipolojisini (veya kümelemesini) kurmak amacıyla bu tekniğe başvurmuştur. Lazarfeld özellikle tutuma ilişkin ölçek verilerini içeren araştırmalarda örtük özelliği belirleyebilmek amacıyla bu yöntemi geliştirmiştir (McCutcheon, 1987). Bu yaklaşımda, gözlenen sürekli değişkenler temele alınarak faktör analizini de içeren bir örtük yapı yöntemi kullanılmıştır. Böylece örtük sürekli değişkenler (faktörler) tanımlanmaya çalışılmıştır. Ancak daha önce de belirtildiği gibi örtük sınıf analizi bireyleri sürekli bir değişken üzerine yerleştirmez, kategorik olarak belli sınıflara atar. Benzer şekilde BTM ölçme sonuçlarında kişiler için bir puan kestirmek yerine bireylerin önceden belirlenmiş sınıflardan hangisine ait olduklarını en doğru biçimde kestirmeye çalışır. Şekil 8, sürekli ve kategorik ölçme yaklaşımının anlatımını basit bir diyagramla göstermektedir.

Şekil 8

Sürekli ve Kategorik Ölçme Yaklaşımı



Bir özellikle ilgili ölçüm yapmanın temel mantığı bireyler arasındaki farkları belirlemektir. Sürekli değişkenlere ilişkin ölçümler incelendiğinde bireylerin bir değişkende hangi konumda olduklarını yani sürekli artan puanlar içinde hangi puana karşılık geldiklerini saptamanın en güvenilir ve geçerli yöntemleri araştırılmaktadır. Örneğin, bir matematik testinde matematik başarısı sürekli bir değişken olarak kabul edilir ve en tutarlı ölçüm yöntemiyle bireylerin bu puan skalasındaki yeri tespit edilmeye çalışılır. Örtük sınıf analizi ise matematik başarısını sürekli değil kategorik bir değişken olarak ele alır ve analiz eder.

Tabi şu noktanın altını çizmek gerekir: Örtük sınıf analizi matematik başarısının doğası itibariyle kategorik olduğunu iddia etmez. Ancak sosyal bilimlerde değişkenler genellikle sürekli halleri ile kullanılmaz. Yukarıdaki örneği ele alalım. Matematik başarısının kullanıldığı durumların çoğu öğrencinin sürekli puanını değil bu puana ilişkin kategorik kararları içerir. Eğitim süreçlerine bakıldığında yerleştirme ve sıralama amaçlı sınavlar dışında matematik başarısı; öğrencinin konuyu anlayıp anlamadığı, dersten geçip geçmediği veya hangi harf kategorisine dâhil olduğu gibi kararları almak için kullanılır. Örtük sınıf analizi ile bu mantığı test uygulamalarında işe koşan BTM, tam da bu yaklaşım üzerine temellendirilmiştir. Bahse konu yaklaşım, iki öğrenci arasındaki farkı keskin bir biçimde belirlemektense birbirinden farklı oldukları bilinen iki öğrenciyi hangi sınıfa atamak gerektiğini saptamaya çalışır.

5.2. BTM'nin Genel Özellikleri

BTM'nin amacı öğrencileri kategorik niteliklerin sıralanışına dayalı örtük kategorilere göre sınıflandırmaktır. Öğrenciye yönelik teşhisin temelinde yer alan beceri seti, örtük değişkenleri oluşturan nitelik aracılığıyla tanımlanır. Burada "nitelik" olarak adlandırılan örtük değişkenler; özellik, vasıf, görev, alt görev, bilişsel süreç veya beceri olarak tanımlanabilir (Tatsuoka, 1985). Görüldüğü gibi BTM'de özellik olarak tanımlanan değişken, aslında bir testte ölçülmek istenen alt becerilerin tamamını ifade eder. Örneğin dört işlem becerisini ölçen bir testte her bir işlem bir özellik olarak tanımlanabilir. Aynı şekilde yalnızca toplama işlemi özelinde 'tek basamaklı sayılarda toplama', 'iki basamaklı sayılarda toplama' veya 'eldeli toplama' gibi alt boyutlar özellik olarak nitelendirilebilir. Üst düzey becerileri ölçmeye yönelik testlerde ise bu özellikler daha genel tanımlarla belirlenebilir.

BTM, bir testin birbiriyle ilişkili birden fazla yapıyı ölçtüğü durumlarda daha uygundur. Testteki her madde bu yapıları ya da bilişsel bileşenlerini ölçecek biçimde düzenlenir. BTM'de her madde testte yer alan birden fazla boyutu aynı anda ölçebilir. BTM, öğrencilerin örtük ölçekteki yetenek düzeylerine odaklanmaz; bunun yerine her bir bilişsel öge üzerindeki performanslarını belirler. Elde edilen olasılıklar, öğrencinin uzmanlaştığı becerilerin profilini ortaya koyacak şekilde dönüştürülebilir. BTM'de maddelerin ölçülecek özelliklerle eşleştirilmesi için Q-matris kullanılır. Bilişsel tanı yaklaşımını kavrayabilmek için Q-matrisi anlamak ve bu matrisin nasıl oluşturulduğunu bilmek kritik öneme sahiptir.

5.3. BTM Analizlerinde Q-Matris

Q-matriste her bir sütun bir nitelik ya da becerinin vektörünü, her satır ise bir maddeyi temsil eder. Nitelikler; alan uzmanları tarafından belirlenen özellikler, prosedürler, buluş yöntemleri, stratejiler, beceriler ve diğer bilgi bileşenleri olabilir. Q-matris, 1 ve 0 şeklindeki ikili kodlamalarla bir özelliğin maddede bulunup bulunmadığını gösterir. Bu numaralandırma ilk kez Fisher (1973) tarafından 'ağırlıklandırma' olarak tanımlanmıştır. Buna göre, bir k niteliği bir i maddesinde bulunduğunda 1, bulunmadığında ise 0 ile kodlanır. Q-matrisinin nasıl oluşturulduğu aşağıda bir örnek üzerinden açıklanmıştır.

Tablo 1

Örnek Q-matris

Maddeler	α_1	α_2	α_3
1	1	0	0
2	0	1	0
3	1	1	0
4	0	0	1
5	0	1	1
6	1	0	1
7	1	1	1

Tablo 1'de basit bir Q-matris örneği verilmiştir. Yedi maddelik bu test ile üç özellik ölçülmektedir. Özellikler α1, α2 ve α3 şeklinde gösterilmiştir. Birinci madde incelendiğinde bu maddenin sadece birinci özelliği ölçtüğü görülmektedir. İkinci madde sadece ikinci özelliği ölçmektedir. Üçüncü maddenin birinci ve ikinci özelliği birlikte ölçtüğü anlaşılmaktadır. Yedinci maddeye bakıldığında ise teste ölçülmek istenen üç özelliğin de bu maddede yer aldığı görülmektedir.

Q-matrisin yapısı aslında BTM için çok önemli bir özelliği daha vurgular. KTK ve MTK'de (tek boyutlu modelleri), bir maddenin birden fazla özelliği ölçmesi hata olarak kabul edilir. Bu kuramlar, bir testteki bir uyarının (maddenin) sadece bir davranışı tetikleyebileceği, yani bir özelliği ölçüleceği varsayımına dayanır. BTM'de ise bir maddenin birden fazla özelliği ölçebileceği kabul edilir ve bu doğrultuda bir istatistiksel yaklaşım kullanılır. Bu yönü BTM'yi önceki yaklaşımlara göre ayrıcalıklı kılmakta ve test geliştiricilere geniş imkânlar sunmaktadır.

Ölçülen niteliklerle birlikte düzenlenen Q-matrisi, madde güçlüğünün psikometrik model ve bilişsel süreçle arasındaki bağlantıyı kurar. Modelin seçimi ve Q-matris birlikte, örtük sınıf cevaplama olasılıklarını belirlemeyi sağlar. Son dönemde yapılan araştırmalarda, Q matrisinin doğruluğunu ve etkisini ölçmeye yönelik güçlü kanıtlar sunan yöntemler geliştirilmiştir. Templin ve Henson (2006), Q-matrisin kalitesinin, öğrencilerin beceri profilinin kestirimini doğrudan etkilediğini ifade etmektedir. Dolayısıyla söz konusu çalışmalar 'tanısal analiz' açısından büyük öneme sahiptir (de la Torre, 2008).

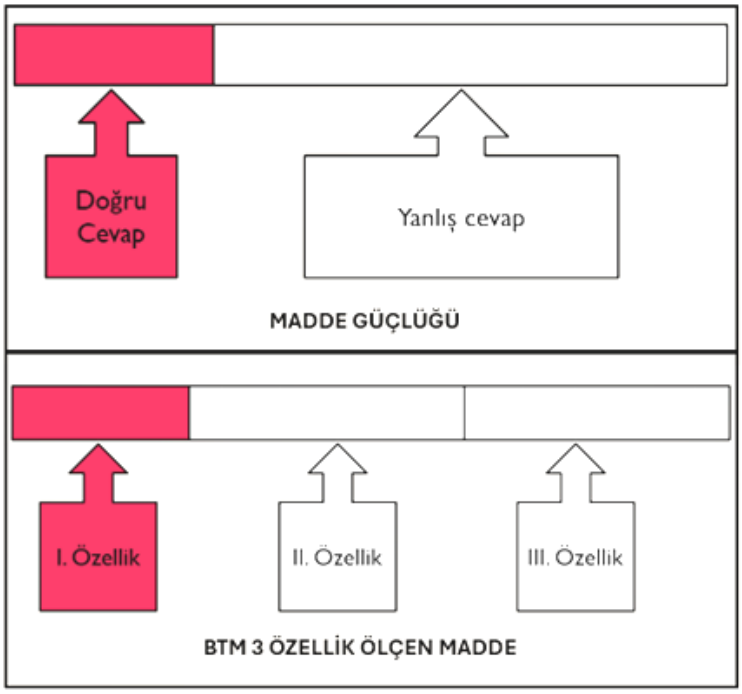
Maddelerin oluşturulması BTM analizleri için kritik önem taşımaktadır. Burada en önemli konu, Q-matrisini oluşturan özelliklerin doğru şekilde tanımlanmasıdır. Bir testte hangi becerilerin ölçüleceği, testin hangi öğrenci grubunun hangi özelliklerini ayırmak için kullanılacağı önceden belirlenmelidir. Ayrıca maddeler de bu özelliklere göre tanımlanmalı ve oluşturulmadır. BTM analizlerine yöneltilen temel eleştirilerden biri, Q-matrisini oluşturma sürecinin uzman kanaatlerine dayanmasıdır. Bu eleştirilerin haklılık payı bulunmakla birlikte benzer endişeler KTK ile MTK için de söz konusudur. Gerek KTK gerekse MTK'ye göre geliştirilen testlerde maddenin hangi özelliği ölçtüğü konusunda uzman kanaati dışında kesin bir kanıt bulunmamaktadır. Özellikle çok boyutlu testlerde, maddeler ile boyutlar arasındaki ilişki yine uzman kararıyla belirlenmektedir. Bütün test geliştirme süreçlerinde uzman görüşüyle belirlenen kazanım, davranış, özellik, öğrenme çıktısı veya beceri daha sonra istatistiksel yöntemlerle ilişkilendirilerek kanıtlanmaktadır. BTM modellerinde de parametreler, Q-matrisinin doğruluğu hakkında bilgi verecek şekilde geliştirilmiştir. Burada eleştirilen nokta, sosyal bilimlerde karmaşık yapıları ölçen testlerde hangi becerilerin ölçüldüğünün açık ve kesin biçimde tanımlanmasının güç olmasından kaynaklanmaktadır. BTM, test geliştirmeye yönelik yaklaşımıyla bu sürecin ölçülen özellikler bağlamında daha titiz yürütülmesi gerektiğine dikkat çekmektedir.

5.4. BTM'de Madde Özellik İlişkisi

Bilişsel tanıya dayalı yaklaşımın test geliştirme süreçlerine en önemli katkılarından biri, ölçülen özellik ile soru arasındaki ilişkiye getirdiği bakıştır. Test geliştirme sürecinde soru yazarlarının genel eğilimi, bir soruda ölçülen baskın özelliği sorunun ölçtüğü özellik olarak kabul etmektir. Sorunun baskın kazanımın yanında çözüm için ihtiyaç duyulan diğer özelliklerin de ortaya konulmasını gerektirmesi, test maddesinin birden fazla özelliği ölçtüğünü yansıtır. Bu durumda aslında sorunun neyi ölçtüğünden çok, soruyu çözerken hangi özelliklerin kullanıldığını tespit etmek daha önemlidir. Şekil 9'daki diyagramda, maddeye ilişkin geleneksel yaklaşım ile bilişsel tanılama yaklaşımına arasındaki farklılık resmedilmektedir.

Şekil 9

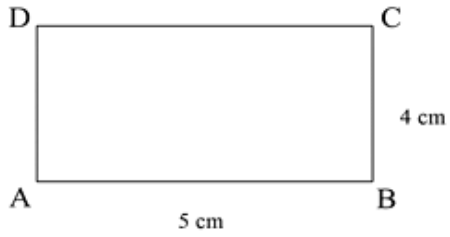
Maddeye İlişkin Geleneksel Yaklaşım İle BTM Yaklaşımına Ait Farklılık

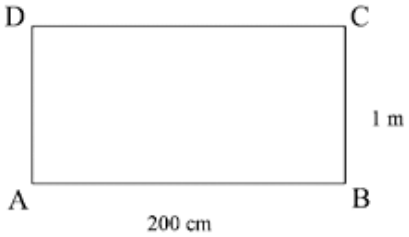


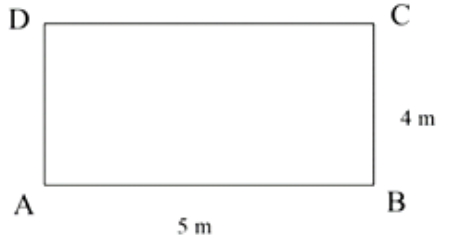
Şekil 9 incelendiğinde genel olarak kullanılan madde güçlüğü anlayışını yansıtmaktadır. KTK'de madde güçlüğü bir maddeyi doğru cevaplayanların sayısının, teste katılan toplam kişi sayısına oranı olarak tanımlanır. MTK'de ise madde güçlüğü, maddeyi doğru yanıtlama olasılığının %50 olduğu yetenek düzeyine karşılık gelir. Yani her iki yaklaşımda da madde güçlüğü hesaplanırken maddeyi doğru cevaplayan kişilerin ne bildiğine odaklanılmaktadır. Diyagramın ikinci kısmında BTM için üç özelliği ölçen bir maddenin yapısı verilmiştir. Burada maddeyi doğru cevaplamak için gereken özelliklerin yanı sıra, maddeyi yanlış cevaplayanların hangi özelliklere sahip olmadıklarını da belirlemeye ihtiyaç vardır. Tam da bu sebeple BTM için soru geliştirilirken soruyu doğru cevaplama prosedürü ayrıntılı olarak açıklanmalı, soruyu doğru cevaplamak için gereken özellikler net biçimde belirlenmeli ve sonuç olarak öğrencinin soruyu nasıl cevapladığı açıkça anlaşılabilmelidir. Modelin bu yapısı, aynı kazanımı ölçtüğü düşünülen ancak farklı güçlük düzeylerinde bulunan soruların tek bir ana başlık altında yazılmasının ölçme sürecinde yarattığı karmaşaya da dikkat çekmektedir. Şekil 10'da açıklayıcı bir örnek yer almaktadır.

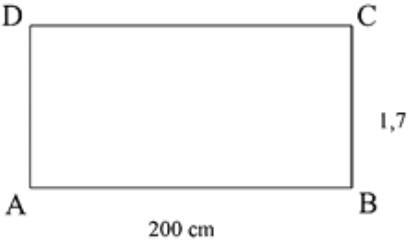
Şekil 10

Aynı Soru İfadesiyle Ölçülen Faklı Özelliklere İlişkin Örnek Maddeler

<table>
<tr>
<td>

Yukarıdaki dikdörtgenin alanı kaç cm^2 dir?</td>
<td>

Yukarıdaki dikdörtgenin alanı kaç m^2 dir?</td>
</tr>
<tr>
<td>• Dikdörtgenin alanının kenarları çarpımına eşit olduğu.</td>
<td>• Dikdörtgenin alanının kenarları çarpımına eşit olduğu.</td>
</tr>
<tr>
<td>

Yukarıdaki dikdörtgenin alanı kaç cm^2 dir?</td>
<td>

Yukarıdaki dikdörtgenin alanı kaç cm^2 dir?</td>
</tr>
<tr>
<td>• Dikdörtgenin alanının kenarları çarpımına eşit olduğu.
• Birimler arası dönüşüm.</td>
<td>• Dikdörtgenin alanının kenarları çarpımına eşit olduğu.
• Birimler arası dönüşüm.
• Ondalık sayılarda çarpma işlemi yapabilmesi</td>
</tr>
</table>

Şekil 10'daki dört madde, aynı soru ifadesiyle öğrenciye yöneltilmiştir. Maddeler incelendiğinde birinci ve ikinci maddenin aynı özelliği ölçtüğü anlaşılmaktadır. Buna karşılık, üçüncü ve dördüncü soruları doğru yanıtlayabilmek için gerekli özellikler birinci ve ikinci sorudan farklıdır. Her sorunun altında bu soruyu doğru cevaplamak için gerekli olan özellikler tanımlanmıştır. Geleneksel ölçme yaklaşımında bu soruyu yanlış cevaplayan bir öğrencinin dikdörtgen alanını hesaplama bilgisinin yetersiz olduğu kabul edilmektedir. Ancak BTM ile yapılan analizler sayesinde öğrencinin üçüncü veya dördüncü soruyu yanlış cevaplamasının hangi bilgi eksikliğinden kaynaklandığına dair istatistiksel bir dönüt sunulabilmektedir.

5.5. BTM'de Madde Parametreleri

BTM modelleri sürekli yeni yaklaşımlarla gelişmekte, bu noktada madde ve teste ait farklı istatistikler kullanılmaktadır. Bu kitabın içeriği göz önüne alındığında, bugüne kadar ortaya konmuş ve çalışmalara konu olmuş bütün madde yaklaşımlarını incelemek yerine bilişsel tanı yaklaşımının mantığını ortaya koyan maddelere ilişkin bazı parametreler açıklanacaktır.

BTM, temel olarak örtük sınıf analizine dayandığından ölçülen bireylerin atandıkları sınıfların doğruluğu hakkında kestirim yapmaktadır. Bu nedenle modeller, genel olarak örtük özellik ile gözlenen özellik arasındaki ilişkiyi olasılıkla temellendirmekte, her madde için çoğunlukla iki madde parametresini hesaplamaktadır. Bunlar *s* "kaydırma" (slip) ve *g* "tahmin" (guess) parametreleridir.

$$s_j = P\left[Y_{ij} = 0 | \eta_{ij} = 1\right]$$

$$g_j = P\left[Y_{ij} = 1 | \eta_{ij} = 0\right],$$

s_j, örtük özelliğe sahip bireyin *j* maddesine yanlış cevap verme olasılığını (yanlış pozitif olasılık), g_j ise örtük özelliğe sahip olmayan bireyin doğru cevap verme olasılığını (doğru pozitif olasılık) ifade etmektedir. s_j parametresi düştükçe, aranan özelliklere sahip bireylerin doğru cevap verme olasılığı artmaktadır.

g_j parametresine tahmin parametresi de denir. Maris (1999), g_j parametresini, alternatif olarak zihinsel tahmin yeteneğine başarıyla bel bağlama durumu olarak açıklar. Tahmin parametresi bu anlamda MTK'deki şans parametresinden farklı bir yapıdadır. Modelde *g* parametresi, yalnızca gerekli özelliğe sahip olmayan bireyin maddeyi doğru cevaplaması anlamına gelmez. Ayrıca cevaplayıcının gerekli olduğu düşünülen özellikler dışında başka özellikleri kullanarak maddeyi doğru cevaplaması anlamına da gelir. Bu nedenle, bir maddenin doğru cevaplanması için Q matris tarafından tanımlanmayan farklı özellikler de yeterli olabilir. Bir madde için *g* parametresi '1' değerine yakın çıktığında yapılabilecek tek yorum o maddeyi ilgili özelliğe sahip olmayan bireylerin doğru cevapladığı olmaz. 1'e yakın g parametresi maddeyi doğru cevaplamak için gerekli bazı özelliklerin belirlenemediğini de işaret edebilir.

5.6. BTM'de Yetenek Kestirimi

Bilişsel tanıda farklı puanlama türleri ve daha fazla kategoriye yönelik sınıflamalar üzerine çok sayıda araştırma yapılmıştır. Ancak temel yaklaşım, öğrenci yeteneğinin iki kategorili bir değişken olarak belirlenmesi fikrine dayanmaktadır. Bu anlamda testi alan kişinin yetenek kestiriminin iki kategorili bir yaklaşımla nasıl belirlendiği ve sonuçların nasıl raporlandığına dair açıklamalar yapmak, konunun anlaşılmasına yardımcı olacaktır.

Bilişsel tanı modeline uygun şekilde geliştirilen ya da uygulanan bir testte, belirlenen *k* sayıdaki özelliğe dayalı olarak 2^k sayıda örtük sınıf tanımlanır. Örneğin bir testte üç özelliğin ölçüldüğü varsayılırsa testi alan bireyler toplam $2^3 = 8$ örtük sınıfa ayrılır. Üç özellik için olası sınıflar "000", "100", "010", "001", "110", "011", "101" ve "111" şeklinde sıralanır. Hiçbir özelliğe sahip olmayanlar ilk sınıfa dâhil olurken sadece birinci ve üçüncü özelliğe sahip olan bireyler yedinci sınıfa yerleşir.

Modelin, bir cevaplayıcının belirli bir özelliğe sahip olup olmadığını belirleme süreci, farklı aşamaların birleşimidir. Öğrencinin özellik bakımından 0 sınıfında mı yoksa 1 sınıfında mı yer alacağı, bir olasılık değeriyle belirlenir. Bu değer değişebileceği gibi genellikle kullanımda 0,5 eşiği temel alınır. Dolayısıyla öğrencinin özelliğe sahip olma olasılığı %50'nin altında kalırsa 0 sınıfına, %50 olur ya da bu değeri aşarsa 1 sınıfına dâhil olur. DINA model öğrenciye ilişkin α kestirimini yaparken o özelliği temsil eden maddelerin doğru cevaplanma oranını temel almaz. Öğrencinin bir özelliğe sahip olma olasılığı, o özelliği temsil eden maddelerin güçleriyle ilişkilidir (de la Torre, 2008c). Maddenin olasılık belirleyen temel parametresi ise *s* dir. Bu durumu örneklendirmek daha açıklayıcı olacaktır.

Tablo 2'de 40 maddelik 8 özellik için Q matrisi belirlenmiş bir testte ilişkin A ve B cevaplayıcılarının belli sorulara ilişkin cevap örüntüsü verilmiştir. İki cevaplayıcı da toplamda 20 doğru yanıt vermiştir. Buna rağmen A cevaplayıcısının atandığı örtük sınıf "11101010" iken B cevaplayıcısı "00000000" sınıfına dâhil olmuştur. Görüldüğü gibi DINA model tarafından A cevaplayıcısı 5 özelliğe sahipken B cevaplayıcısının hiçbir özelliğe sahip olmadığı kararı verilmiştir. Tablo 2'de iki cevaplayıcının birinci ve ikinci özellikle ilişkili cevap örüntüsü ve madde parametreleri verilmiştir. Bu analizlerin, geniş örneklemli bir uygulamanın uç iki puanlamasına örnek teşkil ettiğini belirtmek gerekir.

Tablo 2

İki Özellik İçin Parametre Listesi

Madde	Özellik	A	B	*g*	*S*
2	1	0	1	.37	.45
5	1	0	0	.24	.18
6	1	1	0	.13	.02
7	1	0	1	.38	.35
17	1	0	0	.27	.20
36	1	0	1	.16	.38
12	2	0	1	.45	.28
23	2	0	1	.35	.47
32	2	1	0	.14	.00
33	2	1	0	.17	.09

Tablo 2 incelendiğinde 2 öğrencinin cevap verdiği maddelerin parametrelerinin önemli düzeyde farklılaştığı gözlenmektedir. A öğrencisi, birinci özelliği ölçen tek bir soruya doğru cevap vermesine rağmen, bu maddeye ait hem *g* hem de *s* parametrelerinin çok düşük olduğu görülmektedir. Buna karşılık, B öğrencisinin aynı özelliğe ait doğru cevap

verdiği üç soruda bu parametrelerin yüksek olduğu gözlenmektedir. Aynı durum ikinci özelliği ölçen sorularda da geçerlidir. Bu durum, BTM'nin yetenek kestiriminde toplam puandan bağımsız, özelliğe sahip olma düzeyiyle ilişkili bir yaklaşım kullandığını göstermektedir. Bununla birlikte, yukarıdaki örneğin uç bir örnek olduğu ve konunun anlaşılması için özellikle verildiğini unutmamak gerekir. Yapılan birçok araştırma modellerin toplam puanla ilişki düzeyinin yüksek olduğunu göstermektedir.

De la Torre (2009) örtük sınıfların belirlenmesi ve α parametresi kestirimi konusunda maddenin gücünün önemli bir etkisi olduğunu söylemiştir. Yukarıdaki örnek bu durumu desteklemektedir. Bununla birlikte modelin çok boyutlu yapısı da öğrencinin örtük sınıfını belirlemeyi etkilemektedir. DINA modelde bir madde birden çok özellikle ilişkilendirilebildiği için her madde her öğrenci için aynı değeri taşımamaktadır. Bazı durumlarda, aynı maddelere doğru cevap vermelerine rağmen, bir öğrencinin 0 sınıfına, diğerinin ise 1 sınıfına atanması mümkün olmaktadır.

5.7. BTM Modelleri ve Türleri

Bilişsel tanı yaklaşımı giderek yaygınlaşan bir araştırma alanı olduğundan dolayı literatürde alana ilişkin yeni modeller sürekli güncellenmektedir. Bütün modellerin tek tek anlatılması ve farklarının açıklanması yerine, modelleri benzerlik ve farklılıklarına göre listelemek, okuyucular açısından daha yararlı olacaktır. Tablo 3'teki listede testin yapısı, ölçülen özelliğin türü ve modellerin istatistiksel yaklaşımı dikkate alınarak bir sınıflandırma yapılmıştır.

Tablo 3

Bilişsel Tanı Modellerinde Kullanılan Teknikler

Gözlenen Tepki Değişkenleri	Örtük Yordayıcı Değişkenler		Model Tipi
	DİKOTOMİK	ÇOKLU	
İkili Puanlanmış	1. RSM 3. AHM 5. DINA 7. HO-DINA 8. MS-DINA 10. NIDA 11. RERUM 12. BIN 13. MCLCM 14. F NC-RUM 15. RE NC-RUM	2. BIN 4. MCLCM 6. F NC-RUM 9. RE NC-RUM	Tamamlayıcı Olmayan Model
	16. DINO 1. NIDO 3. BIN 5. MCLCM 7. C-RUM 9. GDM 11. H-GDM 13. LCDM 14. G-DINA	17. BIN 2. MCLCM 4. C-RUM 6. GDM 8. H-GDM 10. LCDM 12. G-DINA	Tamamlayıcı Model
Çoklu Puanlanmış	1. RSM 1. AHM 3. BIN 5. MCLCM 7. F NC-RUM 8. RE NC-RUM	2. BIN 2. MCLCM 4. F NC-RUM 6. RE NC-RUM	Tamamlayıcı Olmayan Model
	1. BIN 1. MCLCM 3. C-RUM 5. GDM 7. H-GDM 9. LCDM 11. G-DINA	2. BIN 2. MCLCM 4. C-RUM 6. GDM 8. H-GDM 10. LCDM 12. G-DINA	Tamamlayıcı Model

Tablo 3'teki listede *gözlenen tepki değişkeni*, bireyin sorudan aldığı puan türünü ifade etmektedir. Örneğin çoktan seçmeli, doğru yanlış, eşleştirmeli, kısa cevaplı sorular için sorunun cevabı ya doğru ya yanlış şeklinde puanlanıyorsa dikotomik (1-0), çoklu puanlanıyorsa çoklu şeklinde tanımlanır. Çoklu puanlamaya kısmi puan verilen sorular (Likert tipi ölçekler, dereceleme ölçekleri vb.) örnek olarak verilebilir. Listede *örtük yordayıcı değişken* olarak tanımlanan bölüm, yetenek kestiriminin iki veya daha fazla kategoriden oluşmasını ifade etmektedir. Bu bölümde *model tipi* olarak tanımlanan sütundaki modeller, bireylerin özelliğe sahip olma düzeylerini 1-0 şeklinde raporlar. Aynı sütundaki çoklu modeller ise bireylerin özelliğe sahip olma düzeylerini ikiden fazla sınıfla tanımlayabilmektedir. Listenin son sütununda yer alan model türü, modellerin istatistiksel yapısının telafi edici, denkleştirici ya da burada olduğu gibi tamamlayıcı olup olmadığını göstermektedir.

5.8. Bilişsel Tanı Modellerinin Genel Değerlendirilmesi

Bilişsel tanı yaklaşımının ilgi çekici bir alan olmakla birlikte oldukça yeni bir araştırma konusu olduğunu belirtmek gerekir. BTM ile ilgili üç temel özelliğin ön plana çıktığı, bunun da hem sosyal bilim araştırmalarında hem de eğitim alanında yeni fikir ve yaklaşımlara yol açtığı söylenebilir.

Öncelikle belirtilmesi gereken, BTM modellerinde bir maddenin birden fazla özellikle ilişkilendirilebilmesidir. Bu durum, bir madde oluşturulurken tek bir öğrenme çıktısına veya kazanıma bağlı kalma sınırlılığını ortadan kaldırmaktadır. Bu sayede, özellikle beceri ölçümlerinde birden fazla konu alanını ve bilişsel süreci bir arada kullanmayı gerektiren maddelerin analizini yapılabilmektedir.

İkinci olarak, öğrencilerin bilgi, beceri, bilişsel süreç ve yetenek gibi değişkenlerdeki farklılık ve benzerliklerine göre gruplanabilmesine olanak tanımaktadır. Bu gruplar, özellikle eğitim ve son çalışmalarda gözlenen psikolojik özellikler açısından müdahale, düzenleme ve eğitim ihtiyaçlarını karşılama konusunda tutarlı bilgiler sunmaktadır.

Son olarak, test geliştirme açısından bir maddeye farklı bir gözle bakma; maddenin ölçülen özellikle ilişkisini, o maddeyi cevaplayan kişinin deneyimiyle tanımlama gerekliliğini BTM ortaya koymaktadır. Bu yaklaşım, iyi test maddesi oluşturmak için soru yazarının bakış açısını yalnızca maddenin kurgusuna değil maddenin nasıl cevaplanacağına; maddeyi doğru cevaplamak için gereken özelliklerin yanında yanlış cevaplayanların sahip olmadığı özelliklere de yöneltmektedir.

Yukarıda açıklanan üç hususla birlikte, BTM'de testi alan bireyler yalnızca kategorik olarak tanımlandığından bu modeller bireyleri karşılaştırmak, sıralamak veya seçmek gibi durumlarda kullanılamaz. Başta da belirtildiği gibi bu modellerin amacı öğrenci profillerini oluşturmak ve öğrencilerin yeterli ya da eksik yönlerini belirlemektir. Bireylerin aralarındaki farklara göre sıralamayı ya da ayırmayı gereken ölçümlerde, en azından şu an için, BTM'nin kullanılması uygun görülmemektedir.

6. SONUÇ

Eğitimsel ve psikolojik süreçlerde bireysel farklılıkların, yeteneklerin, akademik başarıların ve duyuşsal özelliklerin objektif kriterlerle saptanması modern toplumların en temel ihtiyaçlarından biridir. Bu doğrultuda geliştirilen ölçme ve değerlendirme kuramları, 19. yüzyıldan itibaren deneysel psikolojideki ilerlemeler ve istatistik alanındaki metodolojik gelişmelerle beslenerek sistematik bir yapıya kavuşmuştur. İlk dönem zihinsel tarama ve grup testlerinden günümüze uzanan bu tarihsel süreç; ölçme araçlarının güvenirliği, geçerliği, norm gruplarının tespiti ve ölçme hatasının en aza indirilmesi gibi temel psikometrik arayışlar etrafında şekillenmiştir.

Bu kapsamda, psikometrik ölçmelerin teorik zeminini oluşturan en köklü yaklaşım Klasik Test Kuramı'dır (KTK). KTK, bireylerin bir testten elde ettiği gözlenen puanı, hipotetik "gerçek puan" ile yönü ve miktarı kestirilemeyen "tesadüfi hata" bileşenlerinin bir toplamı olarak ele alan matematiksel bir modele dayanır. Kuram, testlerin geliştirilmesi ve sınıf içi ölçme uygulamalarında pratikliği ve güçlü mantıksal yapısı sayesinde günümüzde de vazgeçilmezliğini korumaktadır. Ancak KTK'nin madde istatistiklerini örnekleme, bireylerin yetenek kestirimlerini ise testin güçlük düzeyine bağımlı kılan temel sınırlılıkları, modern ölçme kuramlarının doğuşuna zemin hazırlamıştır.

Yirminci yüzyılın ikinci yarısından itibaren bilgisayar ve internet teknolojilerinin sunduğu işlem gücü sayesinde Madde Tepki Kuramı (MTK) ve Rasch Modeli gibi çağdaş yaklaşımlar ön plana çıkmıştır. Modern ölçme kuramları, KTK'nin örneklem ve test bağımlılığı kısıtlamalarını aşarak, test güçlüğünden ve örneklem dağılımından bağımsız, gizil özelliklerin daha hassas kestirilebildiği modeller sunmayı başarmıştır. Aynı şekilde KTK'nin bir uzantısı olan Genellenebirlik (G) kuramı da varyans analizi yoluyla ölçmeye karışan hataları çok boyutlu bileşenlerine ayırarak hata kaynaklarının daha detaylı incelenmesine olanak tanımıştır.

Tüm bu gelişmelerin yanı sıra, ölçme sonuçlarının sadece düzey belirleme amacıyla değil, öğrenme süreçlerine doğrudan müdahale ve dönüt sağlama amacıyla kullanılması ihtiyacı, Bilişsel Tanı Modellerini (BTM) ön plana çıkarmıştır. BTM, bireylerin toplam test puanlarını ya da genel yetenek düzeylerinin belirlenmesi yerine testin ölçtüğü alt becerilere veya niteliklere (öğrenme bileşenlerine) ne derece sahip olduklarını mikro düzeyde ortaya koyan güçlü bir sınıflandırma aracıdır. Teorik yapısı ve sunduğu potansiyel açısından oldukça güçlü olan bu modeller, günümüzde halen gelişim ve olgunlaşma aşamasındadır. Öğrencilerin güçlü yönlerini ve öğrenme güçlüğü çektikleri spesifik alanları (kavram yanılgılarını) tespit ederek biçimlendirici değerlendirme süreçlerine ve bireyselleştirilmiş öğretim tasarımlarına doğrudan rehberlik etme vaadi, kuramsal ve uygulamalı araştırmalarla sürekli olarak geliştirilmeye devam etmektedir.

Nihai olarak, ölçme ve değerlendirme kuramlarının evrimi, işlemselci ve basit doğrusal yaklaşımlardan daha esnek, kuramsal temelleri güçlü ve olasılık teorilerine dayalı gizil modellemelere doğru bir gelişim göstermiştir. Günümüzde eğitimde ve psikolojide kullanılan testlerin bireyler için hayati kararlara dayanak oluşturduğu göz önüne olduğunda; testlerin geliştirilmesi, uygulanması ve analiz edilmesinde kuramsal esaslara bağlı kalınması ve psikometrik kalitenin gözetilmesi teknik bir zorunluluktur. Klasik ve modern test kuramlarının sunduğu metodolojik araçlar doğru analiz edilip sentezlendiğinde, bireysel potansiyellerin en adil ve en az hatayla ölçülmesi mümkün olacaktır.

KAYNAKÇA | REFERENCES

Atılgan, H. (2019). *Genellenebilirlik kuramı ve uygulaması*. Anı.

Baker, F. B. & Kim, S.-H. (2004). *Item response theory: Parameter estimation techniques*. CRC.

Baker, F. B. (2016). *Madde tepki kuramının temelleri* (N. Güler, Çev. Ed.). Pegem Akademi. https://doi.org/10.14527/9786053185673

Baykul, Y. (2000). *Eğitimde ve psikolojide ölçme: Klasik test teorisi ve uygulaması*. ÖSYM.

Birnbaum, A. L. (1968). Some latent trait models and their use in inferring an examinee's ability. F. M. Lord & M. R. Novick (Ed.), *Statistical theories of mental test scores* (s. 395–479) içinde. Addison-Wesley.

Brennan, R. L. (2001). *Generalizability theory*. Springer.

Bock, R. D. (1997). A brief history of item response theory. *Educational Measurement: Issues and Practice*, *16*, 21-33.

Buchanan, R.D. (2002). On not 'Giving Psychology Away': The MMPI and Public Controversy over Testing in the 1960s. *History of Psychology*, *5*, 284-309.

Cheng, Y. & Chang, H. (2007). *The modified maximum global discrimination index method for cognitive diagnostic computerized adaptive testing*. Proceedings of the 2007 GMAC Computerized Adaptive Testing Conference içinde.

Cheng, Y. & Liu, C. (2015). The effect of upper and lower asymptotes of IRT models on computerized adaptive testing. *Applied Psychological Measurement*, *39*(7), 551–565. https://doi.org/10.1177/0146621615585850

Crocker, L. & Algina, J. (1986). *Introduction to classical and modern test theory*. Holt. Rinehart and Winston

de la Torre, J. (2008). An empirically based method of Q-matrix validation for the DINA model: Development and applications. *Journal of Educational Measurement*, *45*(4), 343–362. https://doi.org/10.1111/j.1745-3984.2008.00069.x

de la Torre, J. (2009). A cognitive diagnosis model for cognitively based multiple-choice options. *Applied Psychological Measurement*, *33*(3), 163–183. https://doi.org/10.1177/0146621608320523

Danziger, K. (1990). *Constructing the subject: Historical origins of psychological research*.

Cambridge University Press.

DeMars, C. (2010). *Item response theory*. Oxford University.

De Gurijter, D. N. M. & van der Kamp, L. J. T. (2008*). Statistical test theory for the behavioral science*. Chapman & Hall/CRC.

DeVellis, R. F. (2003). *Scale development: theory and application* (2. Baskı). SAGE Publication.

Doğan, N. & Tezbaşaran, A. (2003). Klasik test kuramı ve örtük özellikler kuramının örneklemler bağlamında karşılaştırılması. *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi*, *25*, 58–67.

DuBois, P. (1970). *A history of psychological testing.* Allyn and Bacon.

Duncan, K. A. & MacEachern, S. N. (2008). Nonparametric bayesian modelling for item response. *Statistical Modelling*, *8*(1), 41–66. https://doi.org/10.1177/1471082X0700800104

Educational Testing Association. (2014). *A snapshot of the individuals who took the GRE revised general test*. https://www.ets.org/research/policy_research_reports/publications/report/2013/jork.html

Embretson, S. E. (1996). The new rules of measurement. *Psychological Assessment*, *8*, 341-349.

Embretson, S. E. & Reise, S. P. (2013). *Item response theory for psychologists*. Psychology.

Erkuş, A. (2003). *Psikometri üzerine yazılar*. Türk Psikologlar Derneği.

Fisher, R. A. (1973). *Statistical methods and scientific inference* (3. Baskı). Hafner Press.

Fotaris, P., Mastoras, T., Mavridis, I. & Manitsaris, A. (2011). Identifying potentially flawed items in the context of small sample IRT analysis. *International Journal on Advances in Intelligent Systems*, *4*(1 & 2), 31-42.

Gillham, N. W. (2001). *A life of Sir Francis Galton: From African exploration to the birth of eugenics.* Oxford University Press.

Gulliksen, H. (1950). *Theory of mental tests.* John Wiley & Sons.

Güler, N., Uyanık, G. & Taşdelen Teker, G. (2012). *Genellenebilirlik kuramı*. Pegem.

Hambleton, R. K. & Swaminathan, H. (2013). *Item response theory: Principles and applications*. Springer Science & Business Media.

Hambleton, R. K., Swaminathan, H. & Rogers, H. J. (1991). *Fundamentals of item response theory*. Sage.

Hogan, T. P. (2019). *Psychological testing: A practical introduction* (4. baskı.). Wiley.

Horst, P. (1966). *Psyhological measurement and prediction*. Wadsworth.

Houts, C. R., Morlock, R., Blum, S. I., Edwards, M. C. & Wirth, R. (2018). Scale development with small samples: A new application of longitudinal item response theory. *Quality of Life Research*, *27*(7), 1721-1734. https://doi.org/10.1007/s11136-018-1801-z

İlhan, M. & Güler, N. (2018). A Comparison of difficulty ındices calculated for open-ended items according to classical test theory and many facet Rasch model. *Eurasian Journal of Educational Research*, *75*, 99–114.

İlhan, M., Güler, N. & Taşdelen Teker, G. (2024). *Puanlayıcı yargılarına dayalı ölçmelerde geçerlik ve güvenirlik*. Pegem.

John, O. P. & Benet-Martinez, V. (2000). Measurement: Reliability, validity, construct validation and scale construction. H. T. Reis & C. M. Judd (Ed.), *Handbook of research methods in social and personality psychology* (339–369) içinde. Cambridge University.

Juve, J. (2008). Testing and assessment. S. F. Davis & W. Buskist (Ed.) *Testing and assessment* (Vol. 2, pp. I-383-I-391) içinde. Sage. https://doi.org/10.4135/9781412956321.n41

Linacre, J. M. (2001). Generalizability theory and Rasch measurement. Rasch Measurement Transactions, *15*(1), 806–807.

Keeves, P.John. (1988) Research methodology, and measurement: An international handbook (2. baskı). Emeral

Kline, T. J. B. (2005). *Psychological testing: A practical aprroach to design and evaluation*. Sage.

Kurpis, S. E. R. & Stafford, M. E. (2006). *Testing and measurement: A user-friendly guide*. Sage.

Lawson, D. M. (2006). Applying the item response theory to classroom examinations. *Journal of Manipulative and Physiological Therapeutics*, *29*(5), 393-397. https://doi.org/10.1016/j.jmpt.2006.04.006

Leighton, J. & Gierl, M. (2007). Cognitive diagnostic assessment for education: Theory and applications. Cambridge University. http://books.google.com.tr/books?id=E13E9lOFUMYC

Levy, P. (1995). Charles Spearman's contributions to test theory. *British Journal of Mathematical and Statistical Psychology*, *48*, 221-235.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems* (Repr). Routledge.

Lord, F. M. & Novick, M. R. (2008). *Statistical theories of mental test scores*. IAP.

Lovie, A.D. & Lovie, P. (1993). Charles Spearman, Cyril Burt, and the origins of factor analysis. *Journal of the History of the Behavioral Sciences*, *29*, 308-321.

Maris, E. (1999). Estimating multiple classification latent class models. *Psychometrika*, *64*(2), 187–212. https://doi.org/10.1007/bf02294535

McCutcheon, A. L. (1987). Latent class analysis (Issue 64. no.). Sage. https://doi.org/10.4135/9781412984713

Meng, X., Xu, G., Zhang, J. & Tao, J. (2020), Marginalized maximum a posteriori estimation for the four-parameter logistic model under a mixture modelling framework. *British Journal of Mathematical and Statistical Psychology*, *73*(S1), 51-82. https://doi.org/10.1111/bmsp.12185

Michell, J. (1999). *Measurement in psychology: Critical history of a methodological concept*. Cambridge University Press.

Mir, U. A., Joldan, S., Kumar, V. & Saini, S. (2024). *International handbook of research methods and statistics: Volume-1*. The Readers Paradise.

Mullis, I. V., Martin, M. O., Foy, P. & Arora, A. (2012). TIMSS 2011 international results in mathematics. https://eric.ed.gov/?id=ED544554

Mullis, I. V., Martin, M. O., Foy, P. & Drucker, K. T. (2012). *PIRLS 2011 international results in reading*. https://eric.ed.gov/?id=ED544362

OECD. (2017). *PISA 2015 assessment and analytical framework: Science, reading, mathematic, financial literacy and collaborative problem solving*. OECD. https://doi.org/10.1787/9789264281820-en

Paek, I. & Cole, K. (2020). *Using R for item response theory model applications*. Routledge.

Popham, W. J. (2025). *Classroom assessment: What teachers need to know* (10. baskı). Pearson.

Rogers, T. B. (1995). *The Psychological Testing Enterprise: An Introduction*. Brooks/Cole, Pacific Grove.

Rose, N. (1979). The psychological complex: Mental measurement and social administration. *Ideology and Consciousness*, *5*, 5-68.

Shavelson, J. R. & Webb N. M. (1991). *Generalizability theory: A primer*. Sage.

Shultz, K. S., Whitney, D. J. & Zickar, D. J. (2021). *Measurement theory in action: Case studies and exercises* (3. baskı). Routledge.

Sokal, M. M. (1987). Psychological testing and American Society, 1890-1930, Rutgers University Press.

Tatsuoka, K. K. (1985). A probabilistic model for diagnosing misconceptions by the pattern classification approach. *Journal of Educational and Behavioral Statistics, 10*(1), 55-73. https://doi.org/10.3102/10769986010001055

Templin, J. L. & Henson, R. A. (2006). Measurement of psychological disorders using cognitive diagnosis models. *Psychological Methods*, *11*(3), 287–305. https://doi.org/10.1037/1082-989X.11.3.287

Thakur, T. (2025). *Rasch measurement theory: A complete course*. Educohack.

Traub, R. E. (1997). Classical Test Theory in Historical Perspective. *Educational Measurement: Issues and Practice*, *16*, 8-14.

Van Der Linden, W. J. & Glas, C. A. W. (Ed.). (2010). *Elements of adaptive testing*. Springer. https://doi.org/10.1007/978-0-387-85461-8

Van der Linden, W. J. & van der Linden, W. (2016). *Handbook of item response theory*. CRC.

Vermunt, J. K. & Magidson, J. (2002). Latent class cluster analysis. J. A. Hagenaars & A. L. McCutcheon (Ed.), *Applied latent class analysis* (s. 89–106) içinde. Cambridge University Press.

YAZARLAR HAKKINDA | ABOUT AUTHORS

PROF. DR. TAHSİN OĞUZ BAŞOKÇU

Ege Üniversitesi | tahsin.oguz.basokcu[at]ege.edu.tr

ORCID: 0000-0002-4821-0045

Prof. Dr. Tahsin Oğuz Başokçu, Ege Üniversitesi Eğitim Bilimleri Bölümü Eğitimde Ölçme ve Değerlendirme Anabilim Dalı'nda öğretim üyesi olarak görev yapmaktadır. Lisans eğitimini Gazi Üniversitesi'nde tamamlayan Başokçu, yüksek lisans derecesini Ankara Üniversitesi'nden, doktora derecesini ise Hacettepe Üniversitesi Eğitimde Ölçme ve Değerlendirme Anabilim Dalı'ndan almıştır. Akademik çalışmalarında psikometri, test geliştirme, örtük sınıf analizi, bilişsel tanı modellerinin yanı sıra üst düzey düşünme becerilerinin ölçülmesi ve üstbiliş ölçümleri alanlarına odaklanmaktadır. Ulusal ve uluslararası düzeyde birçok bilimsel projede aktif roller üstlenen Başokçu, özellikle TÜBİTAK destekli araştırmalarda görev yapmıştır. Başokçu, akademik çalışmalarının yanında çok sayıda bilimsel dergide hakemlik ve istatistik editörlüğü faaliyetlerini de sürdürmektedir.

PROF. DR. TAHSİN OĞUZ BAŞOKÇU

Ege University | tahsin.oguz.basokcu[at]ege.edu.tr

ORCID: 0000-0002-4821-0045

Prof. Dr. Tahsin Oguz Basokcu works as a faculty member in the Division of Measurement and Evaluation within the Department of Educational Sciences at Ege University. He completed his undergraduate studies at Gazi University, earned his master's degree from Ankara University, and received his PhD in Educational Measurement and Evaluation from Hacettepe University. His academic research primarily focuses on psychometrics, test development, latent class analysis, cognitive diagnostic models, as well as the assessment of higher-order thinking skills and metacognition measurements. Basokcu has taken active roles in numerous national and international research projects, contributing extensively to studies funded by TÜBİTAK. Alongside his research, he also actively serves as a statistics editor and reviewer for various scientific journals.

PROF. DR. NURİ DOĞAN
Hacettepe Üniversitesi | nurid[at]hacettepe.edu.tr
ORCID: 0000-0001-6274-2016

Prof. Dr. Nuri DOĞAN, Hacettepe Üniversitesi, Eğitim Fakültesi, Ölçme ve Değerlendirme Anabilim Dalında öğretim üyesidir. Lisansüstü eğitimini Hacettepe Üniversitesinde Eğitimde Ölçme ve Değerlendirme alanında tamamlayan Doğan, 1998'dan bu yana akademik çalışmalarını sürdürmektedir. 2014 -2015 yılarında Massachusets Amherst Üniversitesi Araştırma, Değerlendirme, Ölçme ve Psikometri Bölümünde ziyaretçi araştırmacı olarak bulunmuş ve madde tepki kuramı, değişen madde fonksiyonu, faktör analizi ve şans başarısının test boyutluluğuna etkisi konularında çalışmalar yapmıştır. 2010 yılında doçent, 2016 yılında profesör unvanını almıştır. Uzmanlık alanları arasında test geliştirme ve uyarlama, klasik test kuramı, Rasch modeli, bilgisayarda bireyselleştirilmiş testler, çok aşamalı testler, değişen madde fonksiyonu ve geniş ölçekli değerlendirmeler yer almaktadır. Uluslararası saygın dergilerde çok sayıda çalışması yayımlanmış; ulusal ve uluslararası birçok projede araştırmacı olarak görev almıştır. Ayrıca eğitime giriş, ölçmenin psikolojik temelleri, klasik test kuramı, Rasch ölçme modeli, eğitimde ölçme ve değerlendirme, araştırma yöntemleri, istatistik, sözel alanlarda soru hazırlama vb. başlıklarında lisans ve lisansüstü düzeyde dersler vermektedir.

Prof. Dr. Nuri DOĞAN
Hacettepe University | nurid[at]hacettepe.edu.tr
ORCID: 0000-0001-6274-2016

Prof. Dr. Nuri DOĞAN is a faculty member in The Department of Measurement and Evaluation at Faculty of Education, Hacettepe University. He completed his postgraduate studies in Measurement and Evaluation in Education at Hacettepe University and has been pursuing academic work since 1998. From September 2014 to August 2015, he was a visiting scholar at the Department of Research, Evaluation, Measurement and Psychometrics at the University of Massachusetts Amherst, where he conducted research on item response theory, differential item functioning, factor analysis, and the effect of chance success on test dimensionality. He received the title of Associate Professor in 2010 and Professor in 2016. His areas of expertise include test development and adaptation, classical test theory, Rasch model, computer-adaptive tests, multi-stage tests, differential item functioning, and large-scale assessments. He has published numerous papers in internationally reputable journals and has served as a researcher in many projects. He also teaches undergraduate and graduate courses on topics such as introduction to education, psychological foundations of measurement, classical test theory, Rasch model, measurement and evaluation in education, research methods, statistics, question writing in verbal fields, etc.

3. BÖLÜM

ÖLÇME ARAÇLARININ VE SONUÇLARININ SAHİP OLMASI GEREKEN NİTELİKLER

ESSENTIAL QUALITIES OF MEASUREMENT INSTRUMENTS, SCORES, AND SCORE INTERPRETATIONS

Prof. Dr. Mustafa İLHAN

ÖLÇME ARAÇLARININ VE SONUÇLARININ SAHİP OLMASI GEREKEN NİTELİKLER

Prof. Dr. Mustafa İLHAN
Dicle Üniversitesi

Öz

Ölçme sonucunda elde edilen puanların doğruluğunu ve bu puanlara dayalı olarak bireyler hakkında yapılan değerlendirmelerin isabet derecesini belirlemeye yönelik çalışmalar, ölçmede geçerlik ile güvenirliğin konusunu oluşturur. Ölçme sonuçlarına dayanarak herhangi bir karar vermeden önce ölçümlerin güvenilir ve ölçümlerden hareketle yapılacak yorumların geçerli olduğuna ilişkin kanıtları ortaya koymak gerekir. Bununla beraber, bir ölçme aracının taşıması gereken nitelikler geçerlik ve güvenirlikten ibaret değildir. Ölçme aracının uygulanabilirliğini, kullanım kolaylığını ve farklı koşullarda ne ölçüde tercih edileceğini etkileyen bir diğer kritik özellik kullanışlılıktır. Bu bölümde sıralanan bu kavramlar üzerinde durulmuştur. Önce ölçmede hatadan bahsedilmiş; hata kavramı, hata kaynakları ile türleri başlıklarına yer verilmiştir. Daha sonra güvenirlik başlığı ele alınmış, bu başlık altında klasik test kuramında kullanılan güvenirlik belirleme yöntemleri (test-tekrar test güvenirliği, paralel formlar güvenirliği, iç tutarlılık güvenirliği ve puanlayıcı güvenirliği), güvenirliği raporlarken dikkat edilmesi gereken hususlar, güvenirliği artırmak için alınabilecek önlemler ile ölçmenin standart hatası açıklanmıştır. Devamında geçerlik konusu işlenmiş; kapsama, içyapıya, ölçme aracından alınan puanların diğer değişkenlerle ilişkisine, cevaplama süreçlerine ve ölçme sonuçlarına dayalı geçerlik kanıtları başlıklar halinde sunulmuş, sıralanan başlıkları takiben geçerliği artırmak için alınabilecek önlemlere değinilmiştir. Ardından kullanışlılık başlığı altında kullanışlılığın ne anlama geldiği tanımlanmış; geçerlik, güvenirlik ile kullanışlılık dengesinden söz edilerek bölüm sonlandırılmıştır.

Anahtar Kelimeler

Psikometrik Özellikler, Ölçme Hatası, Güvenirlik, Geçerlik, Kullanışlılık

ESSENTIAL QUALITIES OF MEASUREMENT INSTRUMENTS, SCORES, AND SCORE INTERPRETATIONS

Prof. Dr. Mustafa İLHAN
Dicle University

Abstract

The accuracy of measurement scores and the extent to which decisions and inferences based on these scores appropriately reflect individuals are central concerns of reliability and validity in educational and psychological measurement. Before measurement results are used to support any decision-making process, sufficient evidence should be provided to demonstrate that the instrument produces reliable scores and supports valid interpretations of those scores. Nevertheless, the qualities expected of a measurement tool are not limited to reliability and validity alone. Another essential attribute that influences the applicability of an instrument, its ease of use, and its suitability across different contexts is practicality. This chapter examines these fundamental concepts in detail. It begins with a discussion of measurement error, including the concept of error, its sources, and its types. The chapter then addresses reliability by presenting the reliability estimation methods commonly employed within classical test theory, including test–retest reliability, parallel-forms reliability, internal consistency reliability, and rater reliability. In addition, issues that should be considered when reporting reliability, strategies for improving reliability, and the standard error of measurement are discussed. Next, the chapter turns to validity. Validity evidence based on test content, internal structure, relationships between test scores and other variables, response processes, and the consequences of testing is presented under separate headings. This discussion is followed by practical recommendations for enhancing validity. Subsequently, under the heading of practicality, the concept of practicality is defined, and the chapter concludes with a discussion of the balance among validity, reliability, and practicality.

Keywords

Psychometric Properties, Measurement Error, Reliability, Validity, Practicality

1. GİRİŞ

Tüm ölçme işlemlerinde amaç; ölçülen özelliğin gerçek değerini belirlemek, hatasız sonuçlara ulaşmaktır. Ancak gerçek puan platonik bir kavramdır, ne kadar uğraşırsak uğraşalım ölçülen özelliğin gerçek değerini tespit etmek olanaksızdır (Lord & Novick, 2008). Özellikle sosyal bilimlerde zekâ, başarı, yetenek, tutum, kaygı, kişilik gibi soyut kavramlarla ilgilenildiğinden ölçmeye bir miktar hata karışması kaçınılmazdır. Bundan dolayı ölçme sonucunda elde ettiğimiz puan, ölçülen özelliğin gerçek değeri ve hata olmak üzere iki bileşenden oluşur. Klasik test kuramı bunu $X = T + E$ eşitliği ile gösterir. Bu eşitlikte X gözlenen puanı, T (true) ölçülen özelliğin gerçek değerini, E (error) ise hata bileşenini yansıtır (Kline, 2005). Hatalı ölçmeler bireyler hakkında yanlış kararlar almamıza neden olacağından ölçmeye hata karışmasına yol açan faktörleri bilmemiz ve hatayı asgari düzeye çekecek tedbirleri hayata geçirmemiz gerekir. Ayrıca hatayı tamamıyla ortadan kaldırmak mümkün olmadığından yaptığımız ölçmelere ne kadar hata karıştığını ve ulaştığımız sonuçlara ne düzeyde itimat edebileceğimizi saptamamız zorunluluk arz eder. Ölçümlerin ve bu ölçümlerden hareketle yapılan çıkarımların isabet derecesini kestirmek amacıyla gerçekleştirilen işlemler ölçmede güvenirlik ile geçerliğin konusunu oluşturur. Ölçme sonuçlarını bireyler hakkında kararlar almak için kullanmadan önce bu iki psikometrik nitelik mutlaka test edilmelidir. Geçerlik ve güvenirlik dışında; ölçme aracının işlevselliğini etkileyen, özellikle hangi durumda ne sıklıkta işe koşulduğu üzerinde belirleyici rol oynayan üçüncü faktör ise kullanışlılıktır. Bu bölümde sıralanan bu kavramlar üzerinde durulacaktır. Önce ölçmede hatadan bahsedilecek; daha sonra güvenirlik, geçerlik ve kullanışlılık başlıklarına yer verilecektir.

2. ÖLÇMEDE HATA

Bir özelliğin ölçülen değeri ile gerçek değeri arasındaki fark hata olarak ifade edilir. Örneğin, bir kişinin ağırlığı gerçekte 70 kg iken baskülde görülen değerin 72 kg olması ölçme işlemine 2 kg'lık bir hata karıştığını gösterir. Baskülün aynı kişinin ağırlığını 67 kg olarak belirlemesi ise ölçme işlemine –3 kg'lık bir hata karıştığı anlamına gelir. Yani hatanın bir miktarı, bir de yönü vardır. Ölçme sonuçlarına farklı miktarlarda ve yönde hata karışabilmektedir.

2.1. Hata Kaynakları

Ölçme sonuçlarına ölçme aracından, ölçmenin yapıldığı ortamdan, ölçmeyi yapan kişiden ve ölçülen bireylerden kaynaklanan hatalar karışabilmektedir. Hata, bu kaynakların sadece birinden veya aynı anda birkaçından dolayı oluşabileceği gibi sözü edilen değişkenler arasındaki etkileşimler de hataya kaynaklık edebilmektedir. Bu hata kaynakları aşağıda örnekler üzerinden açıklanmıştır.

2.1.1. Ölçme Aracından Kaynaklanan Hatalar

Ölçme aracının yeterli duyarlılıkta olmaması hatalı ölçümlere yol açan faktörlerden ilkidir. Duyarlılık ölçme aracının ölçülen özellikteki farklılıkları ortaya koyabilme derecesini ifade eder. Ölçme aracının birimi küçüldükçe ölçülen özellikteki farklılıkları ortaya çıkarma gücü artar, buna bağlı olarak ölçmeye karışan hata azalır. Örneğin, gerçek ağırlığı 503 gram olan bir cismi hassas tartı ile tarttığımızda 503 değerine ulaşabiliriz. Ancak aynı cismi eşit kollu terazi ile tarttığımızda ulaşacağımız sonuç 500 gram olur. Eşit kollu terazinin duyarlılığının düşük olması ölçme işlemine 3 gramlık bir hata karıştırır. Benzer durum eğitimde kullanılan ölçme araçları için de geçerlidir. Bir başarı testinde yeterli sayıda maddeye yer verilmezse testin duyarlılığı düşer ve ölçme işlemine karışan hata artar. Yine Likert tipi ölçeklerde cevap kategorilerinin sayısı yeteriz olursa duyarlılık azalır ve hata artar. Yanlış madde türünün tercih edilmesi, ölçme aracının kapsaması gereken konuların bir kısmıyla ilgili maddeler içermemesi, hedef kitleye uygun olmayan bir yazı puntosunun kullanılması, baskının kalitesizliği, maddelerin silik çıkması, madde yanlılıkları (sorunun bir gruba haksız avantaj sunması), yönergenin ve/veya maddelerin net olmaması ölçme aracından kaynaklanan hatalara verilebilecek diğer örneklerdir.

2.1.2. Ölçmenin Yapıldığı Ortamdan Kaynaklanan Hatalar

Baskülün eğri bir zeminde bulunması ağırlığın hatalı ölçülmesine neden olacağı gibi sınavların uygulandığı ortamlar da hataya sebebiyet verebilmektedir. Ortamdaki havalandırmanın ya da aydınlatmanın yetersizliği; ortamının çok sıcak, çok soğuk veya gürültülü olması, sınava girdiği sıranın boyutunun adayın fiziksel özelliklerine uygun olmaması, çevrimiçi yapılan sınavlarda internetin yavaş çalışması ve/veya bilgisayarda teknik problemlerin bulunması ölçme hatalarını tetikleyebilecek ortamdan kaynaklı özelliklerdir.

2.1.3. Ölçülen Kişilerden Kaynaklanan Hatalar

Bireyin sınav günü hasta olması, sınav anında yüksek kaygı/heyecan duyması, uykusuz, yorgun veya karnı aç bir halde sınava girmesi, sınava yakın tarihte travmatik bir olay yaşaması, kopya çekmesi ya da cevaplarını optik forma geçirirken kaydırma yapması ölçme sonuçlarında hataya yol açabilen ölçülen kişiden kaynaklı faktörlerdir.

2.1.4. Ölçmeyi Yapan Kişiden Kaynaklanan Hatalar

Uygulanması uzmanlık gerektiren bir sınavın bu konuda yeterli bilgisi/deneyimi bulunmayan bir kişi tarafından yapılması, öğretmenin puanlamayı yaparken taraflı davranması ya da dikkatsizlik/yorgunluk sebebiyle bazı öğrencilere gereğinden fazla ya da eksik puan vermesi ölçmeyi yapan kişiden kaynaklı hatalara örnek olarak verilebilir.

2.2. Hata Türleri

Ölçmeye karışan hatalar sabit, sistematik ve tesadüfi olmak üzere üçe ayrılır. Bu hata türleri Tablo 1'de başlıklar halinde açıklanmıştır.

Tablo 1

Sabit, Sistematik ve Tesadüfi Hatalar

Hata Türü	Örnekler
Sabit Hata Kaynağı bilinen, tüm ölçümlere aynı miktarda ve yönde karışan, puanların standart sapması ile varyansını değiştirmeyen hatadır (Başol, 2018).	• Öğretmenin sınav kâğıtlarını puanlarken tüm öğrencilere 5 puan fazla vermesi ya da her öğrencinin notuna sınıf ortalamasının %10'u kadar puan eklemesi, • Bir madde silik çıktığı ya da net olmadığı için sınava giren hiç kimse tarafından doğru cevaplanamaması, • Bir baskülün her nesneyi 100 gram eksik tartması, • Bir saatin sürekli olarak 5 dakika geri göstermesi, • İlk *x* santimi kırık bir cetvelin bir defa kullanıldığı uzunluk ölçümleri
Sistematik Hata Ölçülen büyüklüğe, ölçmeyi yapan kişiye ya da ölçme koşullarına göre miktarı ve/veya yönü değişen; kaynağı belli olan hatalardır (Turgut, 1983).	• Öğretmenin yazısı güzel olanlara 5 puan fazla vermesi ya da her öğrenciye sınavdan aldığı notun %10'u kadar puan eklemesi, • Bir maddenin içeriği itibariyle aynı yetenek düzeyindeki farklı alt gruplardan birine avantaj sağlar nitelikte olması, • Çoktan seçmeli bir maddede yanlış seçeneğin doğru cevap olarak belirlenmesi, • Bir saatin her 60 dakikada bir 5 dakika geri kalması, • İlk *x* santimi kırık bir cetvelin birden fazla kez kullanıldığı uzunluk ölçümleri
Tesadüfi Hata Kaynağı bilinmeyen; ölçme sonuçlarına nerede, nasıl ve hangi miktarda karıştığı belli olmayan hatalardır (Başokçu, 2009).	• Öğretmenin farkında olmadan bazı öğrencilere fazla ya da eksik puan vermesi, • Öğrencinin cevaplarını optik forma geçirirken kaydırma yapması, • Şans başarısı (maddenin kör tahminle doğru cevaplanması), • Maddenin cevaplayıcılar tarafından farklı şekillerde anlaşılması, • Sınav ortamının sıcak, soğuk, gürültülü olmasından dolayı ya da sınav sırasında kaygılı/hasta olduğu için öğrencinin gerçek performansını ortaya koyamaması • Sınavda kopya çekilmesi

Tablo 1'deki hata türleriyle ilgili vurgulanması gereken diğer noktalar şunlardır: Sistematik hata bir grubun lehine, diğer grubun aleyhine sonuçlar ürettiğinden yanlılık olarak da isimlendirilir. Tesadüfi hata için ise gelişigüzel/rastgele etiketleri de kullanılır. Tesadüfi hata güvenirliği düşürücü etkiye sahipken sabit ve sistematik hata güvenirliği etkilemez. Geçerlik ise her üç hata türünden de etkilenir.

3. GÜVENİRLİK

Güvenirlik, aynı özellik için elde edilen iki ya da daha fazla ölçme sonucu arasındaki tutarlılığı ifade eder. Örneğin, bir basküle arka arkaya birkaç kez çıktığımızda her seferinde aynı sonucu görürsek baskülün tutarlı ölçmeler yaptığını düşünür, bunu güvenirlik kanıtı olarak yorumlarız (Banks, 2012). Ancak basküle her çıkışımızda başka bir değer ile karşılaşırsak baskülün güvenilir ölçmeler yapmadığını anlarız. Benzer şekilde, kan basıncımızı biri analog diğeri dijital iki ayrı tansiyon aleti ile ölçtüğümüzde aynı sonuca ulaşırsak ölçümlerin güvenilir olduğuna kanaat getiririz. Analog ve dijital tansiyon aletinin gösterdiği sonuçlar arasında tutarsızlık olması halinde ise tansiyon aletlerinden en az birinin hatalı ölçme yaptığı çıkarımına varır, elde ettiğimiz sonuçlara güvenemeyiz. Öğrencilerin matematik dersine yönelik tutumlarını, cevap kategorileri "Katılıyorum, Biraz Katılıyorum, Katılmıyorum" şeklinde etiketlenen Likert tipi bir ölçek aracılığıyla ölçtüğümüz bir başka örnek ele alalım. Diyelim ki öğrencilerden biri bu ölçekteki "Matematik dersini seviyorum" ve "Matematik dersinden nefret ediyorum" maddelerinin ikisine de *Katılıyorum* cevabını verdi. Bu durumda öğrencinin iki maddedeki cevapları tutarsızdır ve bir güvenirlik sorunu var demektir.

Ölçme sonuçlarındaki tutarsızlıklara ölçme işlemine karışan tesadüfi hatalar yol açar. Dolayısıyla güvenirlik, ölçme sonuçlarının tesadüfi hatalardan arınıklık derecesi olarak da tanımlanır. Yani güvenirlik ile tesadüfi hata arasında negatif bir korelasyon vardır. Hata arttıkça güvenirlik azalırken hata düştüğünde güvenirlik yükselir. Ölçme sonuçlarına; ölçme aracı, ölçmenin yapıldığı ortam, ölçülen ve ölçmeyi yapan kişi gibi farklı kaynaklardan gelen hatalar karışabildiğinden güvenirliği test ederken olası hata kaynaklarını göz önünde bulundurmak gerekir. Bu süreçte kullanılan yöntemler güvenirlik kestiriminde hangi ölçme kuramının esas alındığına bağlı olarak değişir. Daha açık ifade etmek gerekirse ölçme kuramları bölümünde yer verilen klasik test kuramı (KTK), madde tepki kuramı (MTK) ile genellenebilirlik (G) kuramı güvenirliği farklı şekillerde kavramsallaştırır ve analiz eder. Örneğin, MTK'ye göre aynı testi alan farklı yetenek düzeyindeki kişiler için ölçme sonuçlarına karışan hata miktarı aynı değildir, yetenek düzeyine göre farklılaşır. Haliyle farklı yetenek düzeyindeki bireyler için elde edilen güvenirlik değerleri birbirinden farklı olur. Öte yandan KTK'de hata miktarı ve güvenirlik değeri yetenek düzeyinden bağımsız olup testin uygulandığı gruptaki tüm kişiler için özdeştir. Güvenirlik kestiriminde G kuramı ile KTK'nin birbirinden ayrıştığı temel nokta ise şudur. Genellenebilirlik kuramı, potansiyel hata kaynaklarının tamamını aynı anda analiz edip güvenirliğe ilişkin tek bir katsayı verir. Buna karşın KTK her bir hata kaynağı için ayrı bir güvenirlik çalışması yürütmeyi gerektirir. Bu bölümde güvenirlik konusu KTK çerçevesinde işlenmiş, KTK'de güvenirliği hesaplarken kullanılan yöntemler Tablo 2'de sunulmuştur.

Tablo 2

KTK'de Kullanılan Güvenirlik Hesaplama Yöntemleri[*]

Hata Kaynağı	Güvenirlik Belirleme Yöntemi	Taşıdığı Anlam
Testin uygulandığı bireylerde veya ölçme koşullarında zaman içinde meydana gelen değişiklikler	Test-tekrar test güvenirliği	Kararlılık
Alternatif formlar arasındaki farklılıklar	Paralel formlar güvenirliği	Denklik/Eşdeğerlik
Ölçme aracındaki maddeler	İç tutarlılık güvenirliği	Maddelerin homojenliği / heterojenliği
Puanlayıcılar/gözlemciler/hakemler	Puanlayıcı güvenirliği	Objektiflik

[*] Han'dan (2021) uyarlanmıştır.

Tablo 2'deki güvenirlik belirleme yöntemlerinin her biri tutarlılığın başka bir yönüne odaklanır. Test-tekrar test güvenirliğinde analiz edilen zamana karşı tutarlılıktır. Paralel formlar güvenirliğinde formlar arasındaki tutarlılık incelenir. İç tutarlılık güvenirliğinde maddelerin birbiriyle tutarlılığına bakılır. Puanlayıcı güvenirliği ise aynı özellik için farklı kişilerin ya da aynı kişinin farklı zamanlarda yaptığı puanlamalar arasındaki tutarlılığı irdeler.

3.1. Güvenirlik Belirleme Yöntemleri

Güvenirlik, tutarlılık kavramı ile eşleştirilse de (Popham, 2025) KTK'de güvenirliği test etmek için tutarlılığa; kararlılık, eşdeğerlik, iç tutarlılık ve objektiflik gibi başka anlamlar yükleyen farklı yöntemler ileri sürülmüştür. KTK'de kullanılan güvenirlik kestirim yöntemleri aşağıda başlıklar halinde açıklanmıştır.

3.1.1. Test-Tekrar Test Güvenirliği

Test-tekrar test güvenirliğinde ölçme aracı aynı gruba belli aralıklarla iki defa uygulanır. Daha sonra katılımcıların ölçme aracının birinci ve ikinci uygulamasından aldığı puanlar arasındaki korelasyon hesaplanır. Elde edilen korelasyon katsayısı *kararlılık* anlamında güvenirlik olarak yorumlanır. Peki, ölçümlerin kararlı olduğunu iddia edebilmek için hesaplanan korelasyon katsayısı en az kaç olmalıdır? Bu konuda net bir alt sınır ileri sürmek güçtür. Crocker ve Algina'ya (2008) göre kişilik, ilgi, tutum vb. bilişsel olmayan özellikler ölçülüyorsa 0,80'i aşan test-tekrar test katsayıları, ölçülmesi hedeflenen yapının iyi tanımlandığına ve ölçme aracının iyi yapılandırıldığına işaret eder. Yetenek

Test-tekrar test güvenirliği, birinci ve ikinci uygulama arasında geçen sürede bireylerin ölçülen özelliğe ilişkin gerçek puanlarında bir değişim olmadığı varsayımına dayanır (Furr & Bacharach, 2008). Bu varsayımın ihlal edildiği biliniyorsa, örneğin iki uygulama arasında bir öğretim müdahalesi söz konusuysa, test-tekrar test güvenirliğini hesaplamak yanlış olur.

testlerinde ise daha yüksek bir test-tekrar test güvenirliğine ulaşılması beklenir. Elbette, ölçülen özelliğin yanı sıra iki uygulama arasında geçen süre de alt sınır olarak kabul edilmesi gereken değeri etkiler (Bandalos, 2018).

"Birinci ve ikinci uygulama arasında ne kadar süre bulunmalıdır?" test tekrar güvenirliğine ilişkin bir başka kritik sorudur. Güvenirlik için esas alınması gereken alt sınır gibi, bu sürenin ne kadar olması gerektiğine dair de net bir şey söylemek mümkün değildir. Çünkü ideal süre ölçülen özelliğe ve ölçme aracının uygulandığı gruptaki kişilerin gelişim özelliklerine bağlı olarak farklılaşır (Mitchell ve ark., 2000). Akademik başarı gibi kısa vadede değişebilen bir özellik ölçülüyorsa iki uygulama arasındaki süre kısa tutulmalıdır. Diğer taraftan zekâ, siyasi görüş ya da tutum gibi değişmesi uzun zaman alan bir özellik üzerinde duruluyorsa iki uygulama arasındaki süre daha uzun olabilir. Bu anlamda birinci ve ikinci uygulama arasındaki süreye karar verirken dikkat edilmesi gereken temel nokta şudur: Aradaki zaman dilimi katılımcılarda olgunlaşma ya da öğrenmeye bağlı değişimler yaşanmayacağı kadar kısa ama ilk uygulamada verilen cevapların hatırlanmayacağı kadar da uzun olmalıdır (Crocker & Algina, 2008). Test-tekrar test güvenirliğine ilişkin uygulamalarda gözetilmesi gereken diğer kritik hususlar şöyle sıralanabilir:

- ⇨ İki uygulama arasında ne kadar süre olduğu mutlaka belirtilmelidir.
- ⇨ Ölçme aracının birinci ve ikinci uygulamasındaki koşullar olabildiğince benzer olmalıdır. Mesela test ilk uygulamada kâğıt kalem formatında, ikinci uygulamada çevrimiçi uygulanırsa katılımcıların puanlarında uygulama biçiminden kaynaklı farklar ortaya çıkabilir. Bu da test-tekrar test güvenirliği hakkında yanıltıcı sonuçlar doğurur.
- ⇨ Birinci ve ikinci uygulamaya ait veriler aynı sınıftan geldiğinden, diğer bir deyişle aynı metrikte olduğundan iki uygulamadan elde edilen puanlar arasındaki korelasyon hesaplanırken eğer varsayımları (örneğin, normallik) karşılanıyorsa sınıf-içi korelasyon katsayısına başvurulmalıdır.
- ⇨ Test-tekrar test yöntemi her bir katılımcının birinci ve ikinci uygulamada yanıtladığı ölçme araçlarını eşleştirmeyi gerektirir. Bunun için ölçme aracı isim/rumuz alınarak katılımcılara uygulanır. Bu işlem zorlayıcı olabildiğinden test-tekrar test güvenirliğine yönelik uygulamalarda büyük örneklemlerden veri toplamak zordur. Ancak yine de yeterli büyüklükte bir örneklemden veri toplandığından emin olunmalıdır. Ulaşılması gereken örneklem büyüklüğü hakkında farklı araştırmacılar değişik görüşler ileri sürmüştür. Koçak ve arkadaşları (2022) en az 30 katılımcıdan oluşan bir örneklemden veri toplanmasını önermiştir. Kennedy (2022) yaptığı ampirik araştırmada 20-30 kişilik örneklemlerin test-tekrar test güvenirliğine ilişkin doğru sonuçlar vermediğini bildirmiş, 40 ila 50 katılımcı içeren örneklemleri kısmen kabul edilebilir bulsa da bu örneklem büyüklüklerinde 0,70 eşiğine ulaşmanın zor olduğunu rapor etmiştir. Test-tekrar test güvenirliğine ilişkin daha güçlü kestirimlere ulaşabilmek için en az 100 kişiden oluşan bir örnekleme

ulaşılmasını önermiştir. Charter (1999) ise güvenilirlik tahmininin kesinliğini örneklem büyüklüğünün bir fonksiyonu olarak nitelendirip test-tekrar test güvenilirliği için örneklemin en az 400 kişiden oluşması gerektiğini ifade etmiştir.

Örnek teşkil etmesi açısından Tablo 3'te, test-tekrar test güvenirliğine ilişkin hipotetik olarak üretilmiş küçük bir veri seti sunulmuştur. Verilerin, çoktan seçmeli altı maddeden oluşan bir testin 10 kişilik bir öğrenci grubuna 3 hafta arayla iki kez uygulanarak elde edildiği varsayılmaktadır. Maddelere ait doğru yanıtlar 1 olarak puanlanmış, yanlış yanıtlanan ve boş bırakılan maddelere ise 0 puan verilmiştir. Tablodaki "Toplam" etiketli sütunlar öğrencilerin testten aldıkları toplam puanı bildirmektedir. Tabloya bakıldığında birinci ve ikinci uygulamadaki toplam puanların genel olarak birbirine benzer olduğu görülmektedir. Buna bağlı olarak test-tekrar test güvenirliği için 0,93 gibi yüksek bir korelasyon katsayısı hesaplanmıştır.

Tablo 3

Test-Tekrar Test Güvenirliğine İlişkin Örnek Veri Seti

Öğrenciler	Birinci Uygulama							İkinci Uygulama						
	M1	M2	M3	M4	M5	M6	Toplam	M1	M2	M3	M4	M5	M6	Toplam
Ö1	1	1	1	1	0	1	5	1	1	1	1	1	1	6
Ö2	1	1	1	1	1	1	6	1	1	1	1	1	1	6
Ö3	1	0	0	0	0	0	1	1	0	1	0	0	0	2
Ö4	1	1	0	0	0	1	3	1	1	1	0	0	1	4
Ö5	1	0	0	0	0	1	2	1	1	0	0	0	1	3
Ö6	1	1	0	1	1	1	5	1	1	0	1	0	1	4
Ö7	1	1	1	0	0	0	3	1	0	1	0	0	0	2
Ö8	0	0	0	0	0	0	0	1	0	0	0	0	0	1
Ö9	1	1	1	1	1	1	6	1	1	1	1	0	1	5
Ö10	1	1	0	1	0	1	4	1	1	1	1	0	1	5
İki uygulamadan alınan puanlar arasındaki sınıf-içi korelasyon katsayısı: 0,93 [%95 GA (0,753; 0,984)]														

* *GA, güven aralığını ifade etmektedir. Görüldüğü gibi korelasyon katsayısına ilişkin kestirim, güven aralığının alt ve üst sınırının orta noktasında değildir. Örneklem küçükse, korelasyon katsayısı mutlak değerce 1'e yakınsa (Cumming & Fidler, 2015) ya da verilerde ranj daralması söz konusu ise (Fidler & Cumming, 2019) korelasyon katsayısına ilişkin güven aralığı asimetrik olmaktadır.*

Test-tekrar test güvenirliğinde ilk uygulamanın ardından belli bir süre geçmiş olsa bile katılımcılar ölçme aracındaki maddeleri ve maddelere verdikleri cevapları hatırlayabilmekte, bu da ikinci uygulamadaki test puanlarını etkileyebilmektedir (carry over effect) (Price, 2017). Bu sınırlılık özellikle maksimum performansın (başarı, zekâ, yetenek vb.) ölçüldüğü durumlarda kendini göstermektedir.

3.1.2. Paralel (Eşdeğer) Formlar Güvenirliği

Bu güvenirlik türünde aynı özelliği ölçen ancak farklı maddelerden oluşan iki ölçme aracı aynı gruba uygulanır. Ardından iki formdan alınan puanlar arasındaki korelasyon hesaplanır ve bulunan korelasyon katsayısı *eşdeğerlik* anlamında güvenirlik olarak yorumlanır. Paralel formlar güvenirliğinde en önemli konu iki formun içerik itibariyle gerçekten denk olup olmadığıdır. İki formu denk olarak nitelendirilebilmemiz için bu formlar aynı kapsamı ölçmelerinin yanında madde sayıları ve türleri açısından da özdeş olmalı, benzer güçlük ve ayırt edicilikte maddeler içermelidir. Biri çoktan seçmeli, diğeri açık uçlu maddelerden meydana gelen iki matematik testi düşünelim. Bu testleri paralel form olarak ele almak uygun değildir. Çünkü madde türündeki farklılıklar katılımcıların test puanlarında farklılık yaratacak, böyle bir durumda iki forma ait puanlar arasındaki korelasyon paralel form güvenirliğine ilişkin doğru bir sonuç sunmayacaktır.

Formların katılımcılara nasıl uygulanacağı paralel form güvenirliğindeki bir diğer önemli hususdur. Formlar çok sayıda maddeden oluşuyorsa aynı anda uygulanmaları katılımcılarda yorgunluğa ve dikkatsizliğe yol açabilir. Buna mani olmak için formları farklı oturumlarda uygulama yoluna gidilmelidir. Ancak bu sefer de uygulama sırasından kaynaklı etkiler oluşabileceği unutulmamalı, söz konusu etkiyi kontrol altına alabilmek için karşıt dengelenmiş desen (counter balanced design) takip edilmelidir. Buna göre örneklem iki gruba ayrılmalı, grubun yarısına önce A sonra B formu uygulanırken diğer yarısında tersi bir sıra izlenmelidir. Formlar tek oturumda uygulanabilecek kadar kısa olduğunda ise iki form katılımcılara aynı anda sunulmalı, tabi burada da yine karşıt dengelenmiş desene bağlı kalınmalıdır. Grubun yarısı önce A arkasından B formundaki maddeleri cevaplarken diğer yarısı B formundaki maddeler ile başlayıp sonrasında A formundaki maddelere geçmelidir.

Şimdi paralel form güvenirliğine örnek oluşturacak bir veri seti ele alalım. Elimizde; içerdikleri maddelerin güçlük ve ayırt edicilik indekslerinin benzer olduğu öngörülen, vücudumuzdaki sistemler öğrenme alanıyla ilgili iki biyoloji testi bulunduğunu varsayalım. Testlerin ikisinde de dolaşım, solunum, sindirim, boşaltım, üreme, iskelet ve sinir sistemleriyle ilgili birer madde olmak üzere yedi madde (M1–M7) yer alıyor olsun. Diyelim ki bu testleri 10 öğrenciye uyguladık ve uygulama sonucunda Tablo 4'te sunulan verilere ulaştık.

Tablo 4

Paralel Formlar Güvenirliğine ilişkin Örnek Veri Seti

Öğrenciler	A Formu								B Formu							
	M1	M2	M3	M4	M5	M6	M7	Toplam	M1	M2	M3	M4	M5	M6	M7	Toplam
Ö1	1	1	1	0	1	1	1	6	1	1	0	0	1	1	1	5
Ö2	1	1	0	0	1	1	1	5	1	1	1	0	1	1	1	6
Ö3	1	0	0	0	1	0	1	3	1	1	0	0	1	1	1	5
Ö4	1	0	0	0	0	1	1	3	1	1	0	0	0	1	1	4
Ö5	1	1	1	1	1	1	1	7	1	1	1	1	1	1	1	7
Ö6	1	0	0	0	0	0	0	1	1	2	0	0	0	0	1	4
Ö7	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
Ö8	1	1	1	0	0	0	1	4	1	1	1	0	0	1	1	5
Ö9	0	1	1	1	1	0	1	5	1	1	1	0	1	1	1	6
Ö10	1	1	1	1	1	0	1	6	1	1	1	0	1	1	1	6

A ve B formundan alınan puanlar arasındaki Spearman sıra farkları korelasyonu: 0,864 [%95 GA (0,515; 0,967)]

Tablo 4'te görüldüğü üzere maddeler 1-0 şeklinde puanlanmış, öğrenciler A ve B formu olarak isimlendirilen iki biyoloji testinden oldukça benzer puanlar almıştır. İki formdan elde edilen puanlar arasındaki korelasyon 0,864 bulunmuştur. Popham (2000) standart test geliştiricilerinin genelde 0,80 ila 0,90 civarında bir paralel form güvenirliğine ulaşmayı hedeflediklerini belirtmektedir. Ancak iki form farklı zamanlarda uygulandığında arada geçen süreden kaynaklı etkilerin de ölçme sonuçlarına karışabileceğini ifade etmekte, bu tür durumlarda 0,70 değerinin ölçüt olarak alınmasını önermektedir. Sıralanan referans değerlere göre, Tablo 4'teki korelasyon katsayısının eşdeğerlik anlamında yüksek bir güvenirliğe işaret ettiği anlaşılmaktadır.

Test-tekrar test güvenirliği için geçerli olan cevapların hatırlanması gibi bir risk taşımasa da paralel form güvenirliğinin de önemli bir sınırlılığı vardır. Aynı konu alanını örnekleyen, madde sayıları ile istatistikleri bakımından benzer olan iki form geliştirmenin zor ve zaman alıcı olması bu yöntemin temel sınırlılığıdır. Bu sebeple paralel formlar güvenirliği çok fazla kullanılmamaktadır.

3.1.3. İç Tutarlılık Güvenirliği

Ölçme aracındaki maddelerin birbiriyle ve ölçme aracından alınan toplam puanla ilişkisine dayalı olarak hesaplanan bir güvenirlik türüdür. Ölçme aracındaki tüm maddeler aynı örtük özelliği ölçmeye yönelik olduğunda, diğer bir deyişle maddeler ölçtükleri özellik bakımından homojen bir yapı gösterdiğinde iç tutarlılık güvenirliği yüksek çıkar. Maddeler ölçtükleri özellik açısından heterojen bir yapıya sahip olduğunda ise düşük bir iç tutarlılık katsayısı hesaplanır.

Bir Türkçe testini ele alalım. Testte 15 madde bulunuyor olsun. Bu testteki 14 madde okuduğunu anlama becerisini ölçmeye yönelikken bir maddenin dilbilgisi ile ilgili olduğunu varsayalım. Dilbilgisi içerikli madde testteki diğer maddelerden daha başka bir özelliği ölçtüğünden iç tutarlılık güvenirliğini düşürücü etki yaratacaktır. İç tutarlılık güvenirliğine ilişkin kanıt elde edebilmek için eşdeğer yarılar (test yarılama) yöntemi kullanılabilir ya da verilerin yapısı göz önünde bulundurularak KR-20, KR-21, Cronbach alfa, McDonald omega, tabakalı alfa ve hiyerarşik omega katsayılarından uygun olanı hesaplanır.

> Tek uygulamaya dayandığından ve birden fazla forma ihtiyaç duyulmadığından iç tutarlılık güvenirliği test-tekrar test ve paralel (eşdeğer) formlar güvenirliğine kıyasla çok daha sık başvurulan bir güvenirlik belirleme yöntemidir.

3.1.3.1. Eşdeğer Yarılar (Test Yarılama) Yöntemi

Bu yöntemde test bir bütün olarak katılımcılara uygulanır ama uygulamadan sonra testteki maddeler olabildiğince eş iki yarıya ayrılır. Bu ayırma işlemi genelde tek ve çift numaralı maddeler şeklinde yapılır. Tabi bu noktada "Testi neden ilk ve ikinci yarı şeklinde değil de tek ve çift numaralı maddeler şeklinde ikiye bölüyoruz?" sorusunu cevaplamamız gerekir. Öncelikle katılımcıların sınav anındaki motivasyonlarını olumsuz etkilememek için testteki maddeler genelde kolaydan zora doğru sıralanır. Dolayısıyla testi ilk ve ikinci yarı şeklinde bölmemiz halinde iki yarıdaki maddelerin güçlük düzeyleri birbirinden farklı olur. Bu durumda iki yarı eşdeğer kabul edilemez. Ayrıca bir testte aynı konu alanına yönelik maddeler çoğunlukla bir arada sunulur. Örneğin, bir ölçme ve değerlendirme testinde en başta genellikle temel kavramlarla ilgili maddeler yer alır. Bu maddeleri sırasıyla ölçme araç ve sonuçlarında bulunması gereken özellikler, eğitimde kullanılan geleneksel ve tamamlayıcı ölçme araçları, test geliştirme, madde analizi ile ölçme sonuçları üzerinde istatistiksel işlemler konularına yönelik maddeler takip eder. Durum böyleyken testi ilk ve ikinci yarı şeklinde bölmek eşdeğer olmayan iki yarıyla sonuçlanır. Örneğin, matematiksel hesaplamaların yer aldığı ölçme konularının tamamı ikinci yarıda bulunur. İlk ve ikinci yarı şeklinde yapılacak bir yarılamadan kaçınmayı gerektiren bir başka faktör de öğrencilerin bir kısmının testin sonundaki maddeleri yetiştirememiş olma ihtimalidir.

Testi en uygun şekilde ikiye ayırdıktan sonra her bir öğrencinin testin iki yarısından aldıkları puanlar belirlenir. Daha sonra iki yarıdan alınan puanlar arasındaki korelasyon katsayısı hesaplanır. Bu korelasyon katsayısı testin tamamına değil sadece bir yarısına ait güvenirliği bildirir. Bunu şöyle düşünebiliriz. Test-tekrar test yönteminde 10 maddelik bir testten alınan puanlar ile bu 10 maddelik testin ikinci uygulamasından elde edilen puanlar arasındaki korelasyon hesaplanır ve bulunan korelasyon 10 maddelik teste ait bir güvenirlik indeksi verir. Aynı şekilde eşdeğer formlar yönteminde 10 maddelik bir testten alınan puanlarla 10 madde içeren eşdeğer forma ait puanlar

arasındaki korelasyon hesaplanır ve ulaşılan korelasyon değeri 10 maddelik test için bir güvenirlik kanıtı sunar. Diğer taraftan test yarılama yönteminde test yapay olarak iki yarıya ayrılır ve tek numaralı maddeler (5 madde) ile çift numaralı maddelerden (5 madde) alınan puanlar arasındaki korelasyon hesaplanır ve bu korelasyon katsayısı sadece 5 sorunun güvenirliğini gösterir (Aybek, 2021). Testin tamamının güvenirliğine ulaşabilmek için hesaplanan korelasyon katsayısına Sperman-Brown düzeltmesi uygulanır (bkz. Eşitlik 1).

$$r_T = \frac{2 \times r_Y}{1 + r_Y}$$

r_Y = İki yarı arasındaki korelasyon

r_T = testin tamamına ait güvenirlik

Eşitlik 1

Her bir maddenin 1 ile 5 arasında puanlandığı açık uçlu 8 maddeden oluşan bir testin 10 öğrenciye uygulandığını ve uygulama sonucunda Tablo 5'te sunulan verilere ulaşıldığını farz edelim. Öğrencilerin testin tek numaralı maddelerinden (M1, M3, M5 ve M7) aldığı toplam puan "Tek" etiketli sütunda, çift numaralı maddelerinden (M2, M4, M6 ve M8) aldığı toplam puan ise "Çift" başlıklı sütunda verilmiştir. Tabloda görüldüğü gibi testin tek ve çift numaralı maddelerinden alınan puanlar arasındaki Spearman sıra farkları korelasyonu (r_Y) 0,793 çıkmıştır. Testin yarısına ait güvenirliği ifade eden bu korelasyon katsayısı, Spearman-Brown formülünde yerine konulduğunda testin bütününe ait iç tutarlılık katsayısı 0,885 olarak kestirilmiştir.

Tablo 5

Test Yarılama Yöntemiyle İç Tutarlılık Katsayısının Hesaplanması

Öğrenciler	Maddeler								Toplam	Tek	Çift
	M1	M2	M3	M4	M5	M6	M7	M8			
Ö1	5	4	3	4	4	4	5	4	33	17	16
Ö2	2	4	1	3	2	2	2	3	19	7	12
Ö3	5	5	4	5	5	5	5	4	38	19	19
Ö4	5	5	5	5	5	5	5	5	40	20	20
Ö5	3	2	1	2	4	2	3	2	19	11	8
Ö6	2	3	2	1	2	3	2	2	17	8	9
Ö7	5	4	3	4	3	4	4	5	32	15	17
Ö8	4	4	3	3	4	4	5	4	31	16	15
Ö9	3	2	1	2	3	2	3	2	18	10	8
Ö10	5	4	3	3	5	4	5	4	33	18	15

r_Y = 0,793 olarak elde edilir ve bu korelasyon katsayısına göre r_T = [(2×0,793)/(1+0,793)] = 0,885 bulunur.

Dikkat edilirse testin tamamına ait güvenirlik katsayısı testin bir yarısına ait güvenirlikten yüksektir. Bu sonuç, madde sayısı ile güvenirlik arasındaki ilişkiye kanıt oluşturmaktadır. Şöyle ki madde sayısı artırıldığında testin duyarlılığı artacağından güvenirlik de yükselir. Ancak bu aşamada, teste eklenen her maddenin muhakkak

güvenirlik katsayısında artış sağlayacağı gibi bir yanılgıya düşülmemelidir. Eğer teste eklenen yeni bir madde testteki mevcut maddelerden daha başka bir özelliği yokluyorsa güvenirliği artırmayacak, tam tersi güvenirlik katsayısında düşüşe neden olacaktır. Ayrıca teste eklenen maddelerin testin güvenirliğini belli bir noktaya kadar artıracağı, bu noktadan sonra madde sayısındaki artışın güvenirliği pek fazla etkilemeyeceği hatta madde sayısının çok fazla artırılmasının öğrencilerde yorgunluğa neden olarak güvenirliği düşüreceği gözden kaçırılmamalıdır.

3.1.3.2. KR-20 ve KR-21 Katsayıları

KR-20 ile KR-21 Chicago Üniversitesi'nde görev yapan G. Frederic Kuder ve Marion Richardson (1937) tarafından maddelerin iki kategorili olarak (0–1) puanlandığı ölçme araçları için geliştirilmiş iç tutarlılık katsayılarıdır. Daha açık bir anlatımla çoktan seçmeli, doğru-yanlış ya da eşleştirme türünde maddelerden oluşan testler ile var/yok, evet/hayır gibi iki kategorili puanlanan performans görevleri iç tutarlılığın KR-20 ve KR-21 katsayıları ile kestirilebileceği ölçmelerdir.

Kuder ve Richardson 1937 tarihli "*Test güvenirliğini kestirim teorisi [The theory of the estimation of test reliability]*" başlıklı makalelerinde 22 ayrı eşitlik sıralamışlardır. KR-20 ile KR-21 katsayıları, ilgili makaledeki 20 ve 21 numaralı eşitlikler kullanılarak hesaplanmaktadır. İşte KR-20 ve KR-21 kısaltmasındaki *K* ile *R* harfleri bu formülleri geliştiren araştırmacıların soyadlarından, 20 ve 21 sayıları ise sözü edilen makaledeki eşitlik numaralarından gelmektedir.

KR-20 katsayısı ölçme aracındaki maddelerin güçlük düzeyleri birbirinden farklı olduğunda kullanılırken KR-21 madde güçlüklerinin aynı olduğunu varsayar. Ölçme aracındaki maddelerin güçlük düzeyleri bilinmiyorsa tüm maddelerin güçlük düzeyleri eşit kabul edilir ve iç tutarlılık KR-21 ile kestirilir. Maddelerin güçlük düzeylerinin birbirine yakınlığı ölçüsünde KR-21, KR-20'ye yaklaşır. Ancak KR-21'in KR-20'den büyük olması söz konusu değildir. Yani bu iki katsayı arasında KR-20 ≥ KR-21 şeklinde bir matematiksel ilişki vardır.

Günümüzde ölçümlerin güvenirliğini tespit etmeye yönelik analizler istatistiksel yazılımlar aracılığıyla gerçekleştirildiğinden KR-20 ve KR-21 katsayılarının elle hesaplandığı bir örnek verilmemiştir. Ancak güvenirliği etkileyen faktörleri anlayabilmek için bu katsayıların formüllerini bilmek önemlidir. Bundan dolayı KR-20 ile K-21 katsayılarının hesaplanmasında kullanılan formüller aşağıda sunulmuştur.

$$\text{KR-20} = \left(\frac{k}{k-1}\right)\left(1 - \frac{\sum_{j=1}^{k} p_j q_j}{S^2}\right)$$

Eşitlik 2

k = Testteki madde sayısı

S^2 = Test puanlarının varyansı

p_i = i. maddeyi doğru cevaplayan kişi sayısının tüm cevaplayıcılara oranı (madde güçlük indeksini ifade eder)

q_i = i. maddeyi yanlış cevaplayan kişi sayısının tüm cevaplayıcılara oranı ("*1-p*" değerine eşittir)

$$\text{KR-21} = \left(\frac{k}{k-1}\right)\left(1 - \frac{\bar{X}(k-\bar{X})}{kS^2}\right)$$

k = Testteki madde sayısı
$\bar{X}$ = Test puanlarının ortalaması

Eşitlik 3

Yukarıdaki eşitliklere bakıldığında (k/k-1) oranın iki formüldeki ortak terim olduğu görülmektedir. Bu terim, madde sayısının güvenirlik üzerindeki etkisini azaltmaya dönük bir düzeltme işlemi olarak formüllere dâhil edilmiştir. KR-20 formülündeki $\sum_{j=1}^{k} p_j q_j$ terimi, testteki maddelerin varyansları toplamına karşılık gelmektedir. Test puanlarına ilişkin varyans ise madde varyansları ile maddeler arasındaki kovaryansların toplamıdır. Maddeler arasında ilişki yoksa daha teknik bir dille ifade edecek olursak maddeler arasında kovaryans bulunmuyorsa test puanlarının varyansı madde varyanslarının toplamına eşit, yani $S^2 = \sum_{j=1}^{k} p_j q_j$ olur. Bu durumda Eşitlik 2'nin sağındaki ikinci parantezin değeri ve beraberinde KR-20 katsayısı 0 çıkar. Maddeler birbiriyle ne kadar ilişkiliyse maddeler arasındaki kovaryans o kadar fazla olur. Buna paralel olarak $\sum_{j=1}^{k} p_j q_j / S^2$ oranı küçülür ve güvenirlik katsayısı 1'e yaklaşır (Kaplan & Saccuzzo, 2018).

KR-21 madde güçlük indekslerinin eşitliği varsayımına dayandığından Eşitlik 3, madde güçlükleriyle ilgili bir terim içermemektedir**.** Fakat uygulamada bir testteki tüm maddelerin güçlük indekslerinin eşit olması neredeyse imkânsızdır. Bu nedenle KR-21 katsayısını hesaplamak çoğu zaman yanlış bir tercihtir. Unutmamak gerekir ki KR-21, güvenirlik hesaplamalarının kâğıt-kalem ile yapıldığı dönemlerde KR-20'ye kıyasla daha kolay hesaplanabilen bir katsayı olarak ileri sürülmüştür ve sadece iç tutarlılığın alt sınırına ilişkin bir kestirim sunmaktadır (Crocker & Algina, 2008). Artık matematiksel olarak en karmaşık güvenirlik hesaplamaları bile bilgisayarlar aracılığıyla kolayca yapılabildiğinden KR-20 yerine KR-21'i hesaplamayı haklı kılabilecek koşullar ortadan kalkmıştır.

3.1.3.3. Cronbach Alfa ile McDonald Omega

Bu iki katsayı, KR-20 ile KR-21'den farklı olarak, maddelerin ikili puanlandığı ölçme araçlarında kullanılabildiği gibi çoklu puanlandığı (örneğin; Likert tipi ölçekler, dereceleme ölçeği veya rubrikler aracılığıyla puanlanan açık uçlu maddeler) ölçme araçlarında da kullanılabilmektedir. Bu anlamda Cronbach alfa ile McDonald omega katsayıları, KR-20 ve KR-21'in puanlamadaki kategorisi sayısının ikiden fazla olduğu durumlara genelleştirilmiş hali olarak nitelendirilmektedir (İlhan ve ark., 2024).

Cronbach alfanın McDonald omegadan ayrıldığı nokta ise daha katı varsayımlara (tau denkliği, maddelerin hata varyanslarının ilişkisizliği) sahip olmasıdır. McDonald omega (ω) ile Cronbach alfa (α) arasında "ω ≥ α" şeklinde bir matematiksel ilişki bulunmaktadır (McDonald, 1999). Bu iki iç tutarlılık katsayısının hesaplanmasında kullanılan formüller aşağıda sunulmuştur.

> Maddelerin iki kategorili (1-0) olarak puanlandığı bir test için Cronbach alfa hesaplanacak olursa elde edilecek katsayı bu teste ait KR-20 iç tutarlılık katsayısına eşit çıkar.

$$\alpha = \left(\frac{k}{k-1}\right)\left(1 - \frac{\sum_{j=1}^{k} S_i^2}{S^2}\right)$$

k = Ölçme aracındaki madde sayısı

S_i^2 = i. maddeye ait varyans

S^2 = Ölçme aracından alınan puanlara ilişkin varyans

*Eşitlik 4**

Tüm maddeler ortalaması 0, standart sapması 1 olacak biçimde standartlaştırılırsa α aşağıdaki formülle de hesaplanabilir.

$$\alpha = \frac{k\bar{r}}{1+(k-1)(\bar{r})}$$

k = Ölçme aracındaki madde sayısı

$\bar{r}$ = maddeler arası olası bütün ikili kombinasyonlar için hesaplanan korelasyon değerlerinin ortalaması

*Eşitlik 5**

$$\omega = \frac{\left(\sum_{i=1}^{k} \lambda_i\right)^2}{\left(\sum_{i=1}^{k} \lambda_i\right)^2 + \left(\sum_{i=1}^{k} (\Psi_i)\right)}$$

λi = i. maddeye ait standartlaştırılmış faktör yükü

Ψ_i = i. maddeye ait hata varyansı

*Eşitlik 6***

* Hogan (2019), ** McDonald (1999)

Eşitlik 4'e yakından bakarsak Cronbach alfa katsayının KR-20'nin genel bir formu olduğunu kolayca fark edebiliriz. KR-20 ile Cronbach alfanın formülü arasındaki tek fark, madde varyansının hesaplanma biçimidir. İki kategorili (1-0) puanlanan maddelerde, madde varyansı maddenin güçlük indeksi üzerinden hesaplanır. Buna karşın çoklu puanlanan maddelerde madde varyansını hesaplamak için daha başka bir yol takip edilir ve bu yol iki kategorili puanlanan maddelere de uygulanabilmektedir. Eşitlik 5 incelendiğinde ise Cronbach alfanın hem maddeler arası ilişkilerden hem de ölçme aracındaki madde sayısından etkilendiği açık şekilde görülmektedir. Son olarak, McDonald omeganın formülü bu noktaya kadar değinmediğimiz faktör yükü kavramını içermektedir. Şu aşamada, McDonald omeganın bir iç tutarlılık katsayısı olduğunu ve Cronbach alfanın görece yanlı iç tutarlılık katsayıları ürettiği durumlarda da kullanılabildiğini bilmemiz yeterlidir. Faktör yükü kavramı bu bölümün kapsamı dışında kalmakta, konu hakkında bilgi edinmek isteyen okuyucuların faktör analizi kitaplarına ve/veya faktör analizini bir bölüm olarak içeren kaynaklara başvurmaları önerilmektedir.

Örnek olması amacıyla aşağıda biri dört, diğer sekiz maddelik iki ölçme aracı için hesaplanan maddeler arası korelasyon değerleri ile Cronbach alfa ve McDonald omega katsayıları sunulmuştur. Tablolardaki "madde-kalan korelasyonu (item-rest corelation)" ifadesiyle herhangi bir maddeye ait puanlar ile bu madde hariç ölçme aracındaki diğer maddelerden alınan toplam puan arasındaki ilişki kastedilmektedir. Tablo 6 ve Tablo 7 maddeler arası korelasyonlar açısından karşılaştırıldığında dört maddelik ölçme aracındaki değerlerin daha yüksek olduğu göze çarpmaktadır. Buna rağmen sekiz

maddelik ölçme aracı için hesaplanan Cronbach alfa ve McDonald omega katsayıları daha yüksek çıkmıştır. Çünkü Cronbach alfa ile McDonald omega maddeler arası ilişkilere dayanır ama aynı zamanda madde sayısının bir fonksiyonudur. Ölçme aracına benzer özellikte maddeler eklendikçe iç tutarlılık katsayıları yükselme eğilimi gösterir, bu eğilim madde sayısı belli bir eşiğe ulaşıncaya dek devam eder.

Tablo 6

Dört Maddelik Veri Seti için Maddeler Arası Korelasyonlar ile Cronbach Alfa ve Mcdonald Omega Katsayıları

Maddeler	M1	M2	M3	M4	Madde-kalan korelasyonu
M1	–				0,594
M2	0,457	–			0,580
M3	0,513	0,504	–		0,626
M4	0,474	0,457	0,488	–	0,581
Cronbach alfa: 0,786 [95% GA (0,734; 0,831)]; McDonald omega: 0,787 [95% GA (0,739; 0,836)]					

Tablo 7

Sekiz Maddelik Veri Seti için Maddeler Arası Korelasyonlar ile Cronbach Alfa ve Mcdonald Omega Katsayıları

Maddeler	M1	M2	M3	M4	M5	M6	M7	M8	Madde-kalan korelasyonu
M1	–								0,500
M2	0,308	–							0,512
M3	0,308	0,313	–						0,503
M4	0,358	0,325	0,308	–					0,488
M5	0,348	0,358	0,358	0,300	–				0,520
M6	0,328	0,330	0,370	0,308	0,330	–			0,503
M7	0,328	0,340	0,325	0,305	0,348	0,318	–		0,511
M8	0,318	0,368	0,323	0,340	0,333	0,320	0,375	–	0,521
Cronbach alfa: 0,799 [95% GA (0,753; 0,838)]; McDonald omega: 0,799 [95% GA (0,757; 0,841)]									



- Hız testlerinin güvenirliğini kestirirken iç tutarlılığa dayalı yöntemler kullanılmamalıdır (Kurpis & Stafford, 2006).
- KR-20, KR-21, Cronbach alfa ve McDonald omega tek boyutluluğa dayalı güvenirlik katsayılarıdır. Fakat bu katsayıların yüksek çıkması tek boyutluluğa kanıt oluşturmaz.
- Ölçme aracının iki yarısı aynı uzunlukta ise Cronbach alfa katsayısı, iki yarı yönteminden elde edilebilecek olası tüm güvenirlik değerlerin ortalamasına eşit olur (Cronbach, 1951; Warrens, 2015). Daha açık ifade etmek gerekirse bir ölçme aracı çok sayıda farklı şekilde iki yarıya ayrılabilir. Oluşan yarıların her biri ayrı bir güvenirlik değeri üretir ve bu güvenirlik değerleri yaklaşık olarak normal dağılım sergiler. İlgili güvenirlik değerlerinin ortalaması alınırsa bulunan değer aynı veri setine ait Cronbach alfa katsayısına eşit olur. Bundan dolayı, iki yarı güvenirliğini Cronbach alfa (ya da McDonald omega) ile birlikte sunmak yersizdir. Cronbach alfa verildiyse iki yarı güvenirliğini sunmaya gerek yoktur.
- Alanyazın incelendiğinde Cronbach alfanın McDonald omegaya kıyasla çok daha fazla kullanıldığı görülmektedir. Ancak gerektirdiği varsayımların karşılanması oldukça zayıf bir ihtimalken ve McDonald omega gibi varsayım açısından daha esnek bir alternatifi mevcutken Cronbach alfayı kullanma konusunda ısrarcı davranmak mantıklı değildir (Hayes & Coutts, 2020). Özellikle günümüzde istatistiksel yazılımların hızla geliştiği ve hemen her yazılımın güvenirlik analizi sekmesine Cronbach alfayla birlikte McDonald omeganın da dâhil edildiği dikkate alındığında hesaplama kolaylığı Cronbach alfa için tercih nedeni olmaktan çıkmalıdır.

3.1.3.4. Tabakalı Alfa ve Hiyerarşik Omega

Kimi ölçme araçlarında birden fazla alt boyut bulunur ama bu boyutlar bir araya gelip daha genel bir yapıyı oluşturur. Ölçme aracının alt boyutlar içerdiği ve bu boyutların daha üst bir yapıyı meydana getirdiği durumlarda toplam puanlara ait iç tutarlılığı kestirirken tabakalı alfaya ya da hiyerarşik omegaya başvurulur (Kalkbrenner, 2023; Tan, 2009). Örneğin, bir matematik testinde cebir ve geometri şeklinde iki boyut yer alabilir. Bu durumda cebir ve geometri boyutlarının iç tutarlılığı Cronbach alfa ya da McDonald omega ile hesaplanır. Ancak cebir ve geometri boyutlarının bileşiminden oluşan matematik testine ait iç tutarlılığı belirlerken Cronbach alfa ya da McDonald omegayı kullanmak doğru olmaz, yanlı güvenirlik kestirimleri ile neticelenir. Burada izlenmesi gereken yol toplam puanlara ait iç tutarlılığı tabakalı alfa veya hiyerarşik omega ile hesaplamaktır.

Yukarıda sıralanan iç tutarlılık katsayılarının tamamı, katılımcıların ölçme aracındaki maddelere verdikleri yanıtlar arasındaki tutarlılığı test eder. Maddelerin ölçtükleri özelliğin homojen bir yapı göstermesi, yanıtlar arasında yüksek bir tutarlılık olmasını sağlar ve iç tutarlılık güvenirliğini yükseltir. Maddelere verilen yanıtlar arasında tutarsızlığa yol açacak faktörler ise iç tutarlılığı düşürür. Bu faktörlerden ilki maddelerin birbirinden farklı özellikleri ölçmesi, yani heterojen olmasıdır. Hedef yapıyla ilişkisi düşük maddelerin yanı sıra birden fazla cevabı olan, kafa karıştırıcı ve/veya cevaplayıcıların aşina olmadığı ifadeler içeren maddeler de iç tutarlılığı azaltır. Mesela olumsuz maddeler, bireylerde kafa karışıklığına yol açarak iç tutarlılığı düşürücü etki yaratabilmektedir.

Alanyazında iç tutarlılık için en sık referans alınan ölçütün 0,70 olduğu bilinmektedir. Ancak her koşulda bu ölçüte bağlı kalmak uygun değildir. İç tutarlılığı etkileyen önemli bir değişken ölçme aracındaki madde sayısı olduğundan az sayıda madde içeren ölçme araçlarında 0,70'in altındaki değerler de makul karşılanabilir. Öyle ki 3 ila 5 maddeden oluşan bir ölçme aracında iç tutarlılığın 0,60'ı hatta 0,50'yi aşması yeterli olabilir (İlhan & Çetin, 2021). Bir işe ya da eğitim kurumuna seçilecek kişileri belirlemek amacıyla yapılan bir sınavda ise 0,70 ölçütü düşük kalacaktır. Bireyler için hayati sonuçları olabilen, dolayısıyla çok daha yüksek bir iç tutarlılığa ulaşmayı gerektiren yüksek riskli testlerde 0,90 civarı değerler ölçüt kabul edilmelidir. İç tutarlılık katsayısı 0,95'i aştığında da temkinli olmak gerekir. Çünkü iç tutarlığın bu denli yüksek çıkması ölçme aracındaki artık maddelerden kaynaklanıyor olabilir. Bu bağlamda, 1'e çok yakın iç tutarlılık katsayıları elde edildiğinde ölçme aracında tamamıyla aynı içeriği ölçen ve çıkarıldığında hedef kapsamı daraltmayacak maddelerin olup olmadığıyla ilgili kontroller yapılmalı, varsa tekrar niteliğindeki maddeler ölçme aracından çıkarılmalıdır (İlhan ve ark., 2024).

3.1.4. Puanlayıcı Güvenirliği

Açık uçlu maddelerden oluşan testlerin ya da performans görevlerinin kullanıldığı ölçmelerde puanlayıcı kaynaklı faktörler de ölçme sonuçlarında hataya yol açabilmektedir. Dolayısıyla puanlayıcı yargılarına dayanan bir ölçme işlemi yürüttüğümüzde puanlayıcı güvenirliğini de hesaplamamız gerekir. Puanlayıcı güvenirliği objektiflik anlamında bir güvenirlik ölçüsü olup *puanlayıcılar arası* ve *puanlayıcı içi* güvenirlik şeklinde iki başlıkta incelenir.

Puanlayıcılar arası güvenirlikte aynı özelliği/performansı puanlayan iki ya da daha fazla puanlayıcı arasındaki uyum test edilirken puanlayıcı içi güvenirlikte aynı puanlayıcının aynı özellik/performans için iki farklı zamanda yaptığı puanlamalar arasındaki uyuma bakılır. Burada bahsi geçen uyumun mutlak uyum ve göreli uyum (tutarlılık) olmak üzere iki yönü vardır. Aynı öğrenci grubunu puanlayan iki puanlayıcı düşünelim. İki puanlayıcı öğrencilere birbirinden farklı notlar vermiş olmasına rağmen bu notlara dayalı olarak

öğrenciler arasında sıralama yapıldığında benzer bir tablo ortaya çıkıyorsa iki puanlayıcı arasındaki mutlak uyum düşüktür ama aralarında yüksek bir göreli uyum var demektir. Tablo 8'deki veriler bu durumu örneklendirmektedir.

Tablo 8

Puanlayıcılar Arası Güvenirliğe ilişkin Örnek Veri Seti

Öğrenciler	Puanlayıcı1	Puanlayıcı2	Puanlayıcı3
Ö1	10	10	10
Ö2	6	4	3
Ö3	3	3	1
Ö4	7	7	4
Ö5	8	8	6

Tabloya bakıldığında birinci ve ikinci puanlayıcının beş öğrenciden dördüne aynı puanı verdiği görülmektedir. Yani bu iki puanlayıcı arasında yüksek bir mutlak uyum bulunmaktadır. Üçüncü puanlayıcı ise birinci ve ikinci puanlayıcıdan farklı puanlar vermiştir. Buna göre üçüncü puanlayıcı ile diğer iki puanlayıcı arasındaki mutlak uyumun düşük olduğu anlaşılmaktadır. Bununla birlikte, öğrenciler en başarılıdan en başarısıza doğru sıralandığında her üç puanlayıcı için de Ö1, Ö5, Ö4, Ö2 ve Ö3 şeklinde bir sonuç ortaya çıkmaktadır. Bu da üçüncü puanlayıcı ile birinci ve ikinci puanlayıcı arasında yüksek bir göreli uyum (tutarlılık) olduğunu yansıtmaktadır. Eğer bir sıralama sınavı yapıldıysa ya da bağıl bir değerlendirme uygulanacaksa puanlayıcılar arasında göreli uyumunun sağlanması yeterli kabul edilebilir. Ancak bir yeterlik sınavı yapıldıysa, standart belirleme çalışması yürütülecekse ya da mutlak değerlendirme uygulanacaksa puanlayıcılar arasında göreli uyumun sağlanmış olması yetmez mutlak uyuma odaklanmak gerekir.

Puanlayıcı güvenirliğini incelerken kullanılabilecek farklı teknikler vardır. Bunun için en sık başvurulan tekniklerden biri, iki puanlayıcının aynı kişilere (ya da tek bir puanlayıcının aynı kişilere iki farklı zamanda) verdiği puanlar arasındaki korelasyonu hesaplamaktır. Bu hesaplamada aynı sınıfa ait ölçümler arasındaki ilişki inceleneceğinden Pearson momentler çarpımı ya da Spearman sıra farkları korelasyonu yerine sınıf-içi korelasyon katsayısı tercih edilmelidir (Mathews, 2010). Uyum yüzdesi, Kappa istatistikleri, Kendall uyuşum değeri ile Krippendorf alfa katsayısı da puanlayıcı güvenirliğini test ederken başvurulabilecek teknikler arasında yer almaktadır.

3.2. Güvenirliğin Raporlanması

Güvenirlik ölçme aracına ait sabit bir özellik değildir, gruba bağımlıdır. Aynı ölçme aracı bir grupta güvenirliği yüksek ölçümler verirken bir başka grupta düşük güvenirlikte ölçümler üretebilir. Bu da her ölçme çalışmasının ardından güvenirliğin raporlanmasını zorunlu kılar. Diğer bir deyişle güvenirliği ölçme aracının yalnızca ilk uygulamasında hesaplamak yeterli değildir. Ölçme aracının her kullanımında ilgili uygulamadaki

güvenirliği ortaya koymak gerekir. Ölçümlere ait güvenirliğin raporlanmaması alanyazında *güvenirlik tümevarımı (reliability induction)* olarak adlandırılmaktadır (Vacha-Haase ve ark., 2000). Güvenirlik tümevarımı ölçme aracını uygulayıp elde edilen sonuçları raporlarken güvenirlikten hiç bahsedilmemesi ya da ölçme aracının önceki uygulamalarında ulaşılan güvenirlik değerinin sunulması ama eldeki verilere ait güvenirlik katsayısının bildirilmemesi şeklinde gerçekleşebilmektedir (Sánchez-Meca ve ark., 2021). Hangi biçimde gerçekleşmiş olursa olsun güvenirlik tümevarımı kaçınılması gereken hatalı bir yaklaşımdır.

Güvenirliği raporlarken dikkat edilmesi gereken ikinci bir husus, güvenirlik kestiriminde kullanılan yöntemin ve yazılımın ayrıntılı bir şekilde sunulmasıdır. Örneğin, "Ölçümlere ait güvenirlik katsayısı 0,75 bulunmuştur" demek yerine "Ölçümlere ait test-tekrar test güvenirliği 0,75 bulunmuştur" ifadesi tercih edilmelidir. Dahası test-tekrar test güvenirliğine ait bu katsayısının hangi korelasyon tekniğiyle elde edildiği de açıklanmalıdır. Benzer şekilde iç tutarlılık güvenirliği hesaplandıysa bunun hangi iç tutarlılık katsayısı (Cronbach alfa, McDonald omega vb.) olduğu söylenmelidir. Güvenirlik analizinde kullanılan yazılım da sadece adıyla değil sürümüyle birlikte (örneğin, ITEMAN 4.4) belirtilmelidir.

Hesaplanan güvenirlik katsayısını nokta kestirimi şeklinde sunmayıp güven aralığı ile birlikte vermek güvenirliği raporlarken gözetilmesi gereken ilkelerden bir diğeridir. Ölçme sonuçlarına karışan tesadüfi hata miktarı tam olarak ortaya konulamadığından güvenirliğin değerini kesin olarak bilmek olanaksızdır. Ancak eldeki verilerden hareketle güvenirlik katsayısı hakkında bir tahmin yapılabilmektedir (Köse, 2012). Güvenirliği raporlarken güvenirlik kestirimine ilişkin bu süreç hesaba katılmalıdır. Elde edilen katsayı herhangi bir hata payı içermiyormuşçasına kesin bir değer olarak verilmemeli, aşağıdaki örnekte görüldüğü gibi yanında güven aralığı yer alacak tarzda rapor edilmelidir.

"Ölçümlere ait Cronbach alfa iç tutarlılık katsayısı 0,745 olarak bulunmuştur."	
"Ölçümlere ait Cronbach alfa iç tutarlılık katsayısı 0,745 olarak bulunmuştur [%95 GA (0,710; 0,780)]."	

3.3. Güvenirliği Artırmak için Alınabilecek Önlemler

Daha önce de belirtildiği gibi güvenirlik ile tesadüfi hata arasında negatif bir korelasyon vardır. Dolayısıyla güvenirliği artırmanın yolu ölçme işlemine karışabilecek tesadüfi hataları minimize etmekten geçer. Ölçme sonuçlarına ölçme aracının kendisinden, uygulama koşullarından, ölçmeyi yapan kişiden veya ölçülen gruptan kaynaklanan

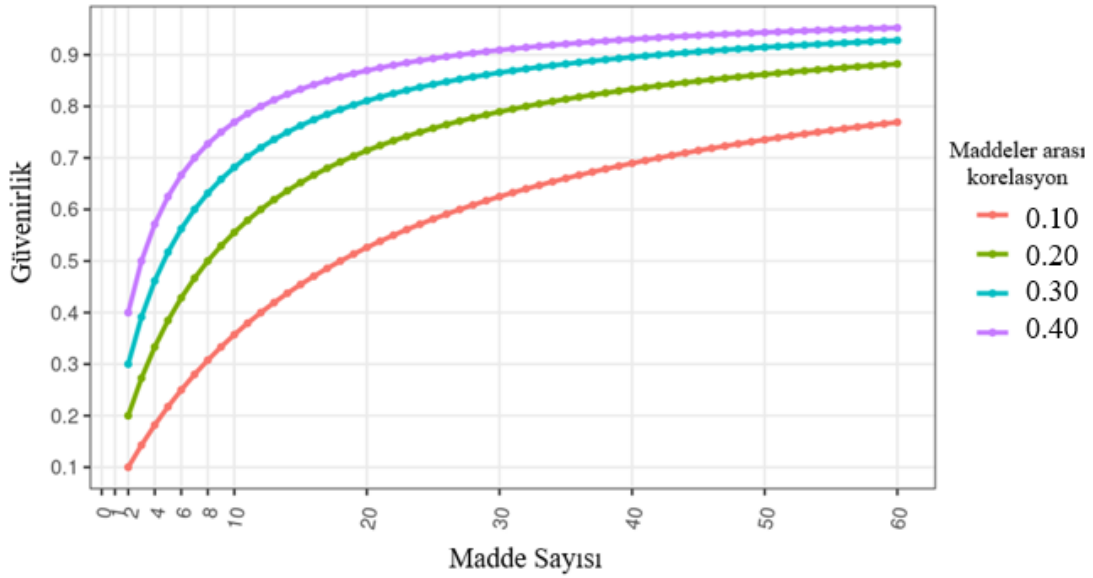
hatalar karışabildiğinden güvenirliği artırmak için sıralanan faktörlerin her biriyle ilgili gözetilmesi gereken belli başlı hususlar vardır. Güvenirliği artırmak için dikkat edilmesi gereken noktalar aşağıdaki gibi sıralanabilir:

- ⇨ Ölçme aracının başında; ölçme aracının uygulama süresi, puanlama şekli ve cevaplamada dikkat edilmesi gereken hususlar gibi temel bilgileri içeren açık bir yönerge sunulmalıdır.
- ⇨ Ölçme aracındaki maddeler net olmalı, tüm cevaplayıcılar tarafından aynı şekilde anlaşılmalı, farklı kişiler tarafından farklı yorumlanabilecek ifadeler içermemelidir. Bunu sağlayabilmek için maddenin anlaşılırlığını negatif yönde etkileyecek olumsuz ifadelerden mümkün olduğunca kaçınılmalı, özellikle de çift olumsuz yargılardan uzak durulmalıdır. Ayrıca maddenin cevap hakkında ipucu verebilecek herhangi bir ifade içermediğinden emin olunmalıdır.
- ⇨ Ölçme aracında yeterli sayıda maddeye yer verilmelidir. Madde sayısı az olduğunda duyarlı ölçmeler yapılamayacak, dolayısıyla katılımcılar birbirinden yeterince iyi ayırt edilemeyecek ve güvenirlik düşük çıkacaktır. Şöyle düşünelim: Elimizdeki testte tek bir madde olduğunda testin uygulandığı gruptaki kişiler "maddeyi doğru cevaplayanlar" ile "yanlış cevaplayanlar/boş bırakanlar" olarak iki gruba ayrılabilecektir. Madde sayısını ikiye çıkardığımızda testin uygulandığı kişiler "her iki maddeyi de doğru cevaplayanlar", "maddelerden yalnızca birini doğru cevaplayanlar" ve "iki maddeyi de yanlış yanıtlayanlar/boş bırakanlar" olmak üzere üç gruba ayrılabilecektir. Eklenen yeni maddelerin testte var olan maddeler ile arasında kabul edilebilir düzeyde korelasyonlar bulunduğu varsayımı altında, madde sayısı artıkça bir noktaya kadar güvenirlik yükselecektir (bkz. Şekil 1).

> "*Ölçme aracında kaç madde olmalıdır?*" sorusuna verilebilecek sihirli bir yanıt yoktur (Aybek, 2021). İdeal madde sayısı ölçülmesi hedeflenen kapsama, ölçme sonuçlarının hangi amaçla kullanılacağına, cevaplayıcıların gelişim özelliklerine vb. faktörlere bağlı olarak değişir.

Şekil 1

Maddeler Arası Belli Korelasyon Değerleri için Madde Sayısı ile Güvenirlik Arasındaki İlişki (Evans, 2021)



Şekil 1'e bakıldığında madde sayısındaki artışa bağlı olarak güvenirlik ilk başta hızlı bir şekilde yükselmekte madde sayısı artmaya devam ettikçe güvenirlikte gözlenen artış giderek daha düşük olmaktadır. Yani eşdeğer yarılar yöntemi başlığında da ifade edildiği üzere belli bir noktadan sonra madde sayısındaki artış güvenirlikte önemli bir yükselme sağlamamakta, dahası madde sayısının gereğinden fazla artırılması cevaplayıcıların yorulmasına/sıkılmasına neden olarak güvenirliği ters yönde etkilemektedir.

- ⇨ Açık uçlu maddeler ya da performans görevleri kullanılarak yapılan ölçmelerde, puanlamada kullanılan anahtarın cevaplardaki olası farklılıkları açığa çıkarılabilecek kadar kategori içerdiğinden emin olunmalıdır. Söz gelimi açık uçlu bir matematik sorusuna verilen cevapları 1-0 şeklinde puanladığımızda öğrencileri sadece iki kategoriye ayırmış oluruz. Oysaki bazı öğrenciler doğru çözüm yolunu kullanmış, sadece işlem hataları sebebiyle yanlış bir sonuç bulmuş olabilir. Ya da uygun şekilde başlangıç yapmış ama çözümün devamını getirememiş öğrenciler bulunabilir. Bu koşullar altında ilgili maddeyi 1 (*tam doğru*), 0,75 (*neredeyse tamamen doğru*), 0,50 (*yarı yarıya doğru*), 0,25 (*kısmen doğru*) ve 0 (*yanlış/boş*) şeklinde puanlamak öğrencileri daha iyi ayırt etmemizi sağlar ve daha yüksek bir güvenirlikle neticelenir.
- ⇨ Ölçme aracındaki maddelerin güçlük düzeyi uygulama yapılacak grubun yetenek düzeyine uygun olmalıdır. Testteki maddeler grup için çok kolay olduğunda gruptaki kişilerin tamamı maddelerin hepsini doğru yanıtlar. Böyle bir durumda katılımcılar arasındaki farklılıklar ortaya konulamaz, haliyle de güvenirlik düşük

çıkar. Aynı şekilde, testteki maddelerin tamamının grup için çok zor olması halinde katılımcılar testteki hiçbir maddeyi doğru yanıtlayamaz. Buna bağlı olarak, katılımcılar arasındaki farklılıklar ortaya çıkarılamaz ve yüksek bir güvenirlik değerine ulaşılamaz. Güvenirliği artırabilmek için katılımcıların yanıtlarında değişkenlik yaratacak maddeler teste dâhil edilmelidir. Bunun için de madde güçlükleri, ölçülen grubun yetenek düzeyleri açısından gösterdiği dağılıma göre ayarlanmalıdır. Buradan şu sonuca da varmak mümkündür: Aynı ölçme aracı heterojen bir grupta yüksek güvenirlikte ölçümler üretirken daha homojen bir grupta düşük güvenirlikte ölçümler verebilir.

- ⇨ Açık uçlu maddeler ya da performans görevleri gibi doğası itibariyle sübjektif nitelik taşıyan ölçme araçları kullanıldığında puanlama sürecini daha nesnel hale getirecek uygulamalara yer verilmelidir. Bu doğrultuda puanlamaların puanlama anahtarına dayalı olarak yapılması sağlanmalı, eğer mümkünse puanlama işlemi öncesinde puanlayıcılara eğitim verilmeli ve puanlamalar birden fazla puanlayıcı tarafından gerçekleştirilmelidir.
- ⇨ Şans başarısını (maddenin kör tahminle doğru yanıtlanma olasılığını) azaltacak uygulamalara yer verilmelidir. Örneğin; çoktan seçmeli maddelerde düzeltme formülü uygulanabilir, yani seçenek sayısının bir eksiği kadar yanlış cevabın bir doğruyu götürdüğü bir puanlama yapılabilir. Yine madde çoktan seçmeli ise çeldiricilerin (doğru cevap dışındaki seçeneklerin) iyi yazılması, diğer bir deyişle maddenin ölçtüğü özelliğe sahip olmayan ya da düşük düzeyde sahip olan öğrencilerin bile kolayca eleyeceği çeldiricilere yer verilmemesi şans başarısını düşürür. Böylece tesadüfi hata olasılığı azalır ve güvenirlik yükselir.
- ⇨ Kopya çekilmesini engelleyecek önlemler alınmalıdır. Alternatif kitapçıkların oluşturulması, sınav salonundaki gözetmen sayısının artırılması ve sınavın kamerayla kayıt altına alınması yüz yüze yapılan sınavlarda kopya ihtimalini azaltacak düzenlemelerdir. Elektronik ortamda yürütülen sınavlarda ise başka sekmeler açmaya, kopyala/yapıştır işlemine ya da ekran görüntüsü almaya mani olacak güvenli tarayıcı sistemlerinin devreye sokulması ve yüz tanıma uygulamalarının kullanılması bu amaçla hayata geçirilebilecek uygulamalardır.
- ⇨ Sınav ortamı iyi ayarlanmalıdır. Ortam çok sıcak, çok soğuk ya da gürültülü olmamalı, yeterli havalandırma ve ışıklandırma sağlanmalı, sınava girdiği sıranın/masanın boyutu öğrenciye uygun olmalıdır. Elektronik sınav yapılıyorsa sistemin uygun hızda ve kesintisiz çalışmasını sağlayabilecek bir ağ alt yapısı ile sunucu kullanılmalıdır.
- ⇨ Ölçme aracını cevaplamaları için katılımcılara tanınacak süre iyi ayarlanmalıdır. Gereğinden az ya da fazla süre verilmesinin güvenirliği düşüreceği unutulmamalıdır. Uygun süreye karar verirken uygulama yapılacak grup göz önünde bulundurulmalı, gruptaki kişilerin ölçülen özellik bakımından ne düzeyde olduğu ve ne kadar heterojen bir yapı gösterdiği hesaba katılmalıdır.

- ⇨ Ölçme aracının uygulandığı kişilere ait değişkenlerin (kaygı, motivasyon, heyecan, hastalık, açlık, susuzluk vb.) güvenirliği önemli ölçüde etkileyeceğinin farkında olunmalı (Boztunç Öztürk & Şahin, 2021) ve yapılabiliyorsa bu değişkenlerle ilgili önlemler alınmalıdır. Örneğin, testten aldığı puanın birey için önemli olduğu sınavlarda kişi zaten yüksek bir motivasyonla maddeleri cevaplar. Öte yandan kimi ölçme araçlarında kaç puan aldığı birey için bir anlam ifade etmez. Bu durumda bireyin motivasyonu düşük olabilir ve rastgele yanıtlanmış ölçme araçlarıyla karşılaşılabilir. Böyle yanıtlar güvenirliği olumsuz etkileyeceğinden katılımcıların motivasyonunu artıracak yollar aranmalıdır. Katılımcılara ölçme aracının neden uygulandığına dair bilgi verilmesi, güvenilir sonuçlara ulaşmanın uygulayıcı için taşıdığı öneme dair ikna edici açıklamalar yapılması ve mümkünse cevaplayıcılara küçük hediyeler verilmesi bireylerin motivasyonlarını artırıp gerçeğe yakın yanıtlar alınmasını sağlar. Bu da güvenirliği yükseltir.

3.4. Ölçmenin Standart Hatası

Güvenilirlik katsayısı ölçümlerin psikometrik özellikleri hakkında son derece önemli bilgiler sunmakla birlikte, ölçme hatasının ne boyutta olduğunu doğrudan ortaya koyamaz. Şöyle ki güvenirlik katsayısı [0-1] aralığında uzanan bir metrikte ifade edilir. Dolayısıyla bir testi bir grup bireye uyguladığımızda güvenirliğin, ölçme sonuçlarında gözlenen ortalama hatanın ne büyüklükte olduğuna dair söyledikleri test puanlarıyla aynı birimde olmaz (Furr & Bacharach, 2008). Oysa bireyin gözlenen puanından yola çıkarak gerçek puanı hakkında tahminde bulanabilmek için ölçme sonuçlarına karışan hatayı test puanlarıyla aynı birimde ifade etmek gerekir. Bu noktada ölçmenin standart hatası kavramı karşımıza çıkar. Ölçmenin standart hatası test puanlarının standart sapması ile güvenirliğine dayalı olarak hesaplanır ve hesaplamada Eşitlik 7'deki formül kullanılır.

$$S_E = (S_x) \times \sqrt{1 - r}$$

S_E = Ölçmenin standart hatası
S_x = Test puanlarının standart sapması
r = Güvenirlik katsayısı

Eşitlik 7

Ölçmenin standart hatası güvenirlik ile ters bir ilişki içerisindedir. Daha yüksek güvenirlik katsayıları ölçmenin standart hatasına ilişkin daha düşük değerler ile sonuçlanır. Ölçmenin standart hatasından yararlanarak bir kişinin ölçülen özelliğe ilişkin gerçek puanının hangi aralıkta yer alacağı belirli bir olasılık dâhilinde tespit edilebilmektedir. Hatırlarsak ölçme sonucunda elde ettiğimiz puanı gözlenen puan olarak nitelendirmiş, ölçme işlemi ne kadar dikkatli yapılırsa yapılsın gözlenen puanlarda bir miktar hata bulunmasının kaçınılmaz olduğunu vurgulamıştık. Bu sebeple bir testi bir bireye tekrar tekrar uyguladığımızda araya öğrenme, unutma gibi gerçek puanlarda farklılık yaratacak başka değişkenler girmese dahi tesadüfi hataların etkisiyle bireyin testten alacağı puanlar bir uygulamadan diğerine değişecektir. Bu değişimin

hangi aralıkta olacağını kestirirken ölçmenin standart hatasına başvurulmaktadır. Ölçmenin standart hatası gözlenen puandan çıkarıldığında gerçek puana ilişkin alt sınır, gözlenen puana eklendiğinde ise gerçek puana ilişkin üst sınır tahmin edilebilmektedir. Bu alt ve üst sınırın oluşturduğu aralık ise güven aralığı adıyla anılmaktadır. Ölçme sonuçları raporlanırken genellikle %68, %95 ya da %99 olasılık güven aralıkları sunulmaktadır. Gözlenen puan, diğer bir deyişle bireyin testten aldığı puan *X* ve ölçmenin standart hatası S_E olmak üzere %68, %95 ve %99 olasılıklı güven aralıkları aşağıdaki gibi hesaplanmaktadır.

	Güven aralığının alt sınırı	**Güven aralığının üst sınırı**
% 68 olasılıklı güven aralığı	$X - S_E$	$X + S_E$
% 95 olasılıklı güven aralığı	$X - 2S_E$	$X + 2S_E$
% 99 olasılıklı güven aralığı	$X - 3S_E$	$X + 3S_E$

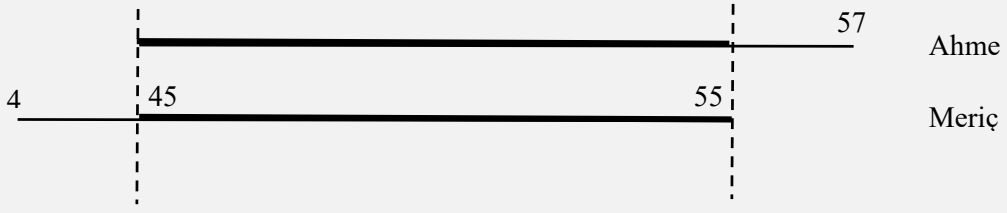
Standart normal dağılım eğrisi altındaki alanları gösteren Z tablosundan yararlanarak %68, %95 ve %99 dışında daha başka olasılıklar için de güven aralıklarını hesaplamak mümkündür.

Anlaşılacağı üzere ölçme sonucunda bireyin gerçek puanına ulaşmak mümkün olmasa bile ölçmenin standart hatası sayesinde bireyin gerçek puanının hangi aralıkta değişeceği tahmin edilebilmektedir. Bir örnek üzerinde sıralanan kavramları biraz daha somutlaştıralım. Diyelim ki Enes isimli bir öğrenci uyguladığımız testten 72 puan aldı. Test puanlarının standart sapması 5 ve güvenirlik katsayısı 0,84'tü. Buna göre ilgili ölçme işlemine ait standart hata 2'dir $[(5)\times\sqrt{1-(0{,}84)}]$. Dolayısıyla Enes'in gerçek puanı %68 olasılıkla 72±2, %95 olasılıkla 72±4 ve %99 olasılıkla 72±6 aralığındadır. Yani Enes bu sınava aynı koşullar altında 100 defa girerse

- 68'inde 70–74 arasında, 32'sinde ise bu aralığın dışında bir puan alır,
- 95'inde 68–76 aralığında bir puan elde eder ama diğer 5 sınavda puanı bu aralığın dışında kalır,
- 99'unda 66–78 aralığında bir puan alır ama kalan 1 sınavda puanı bu aralığın dışında olur.

Ölçmenin standart hatası küçükse aynı test bireye tekrarlı olarak uygulandığında elde edilecek gözlenen puanlar bireyin gerçek puanlarına yakın, dar bir aralıkta kümelenir. Öte yandan ölçmenin standart hatasına ilişkin yüksek değerler, aynı testin bireye tekrarlı olarak uygulanması halinde bireyin gözlenen puanlarında belirgin dalgalanmalar olacağını bildirir. Böyle bir durumda bireyin gerçek puanını tahmin etmek zorlaşır ve standart hatanın büyüklüğü ölçüsünde güven aralığı genişler.

Geçme notunun 50 ve ölçmenin standart hatasının 3'e karşılık geldiği bir sınavda Meriç 49, Ahmet 51 puan almış olsun. Ölçmenin standart hatasını dikkate almazsak Ahmet dersi geçecek Meriç ise dersten kalacaktır. Hâlbuki gözlenen puanlara ilişkin güven aralıklarına göre %95 olasılıkla Ahmet'in gerçek puanı 45-57 ve Meriç'in gerçek puanı 43-55 arasında yer alacaktır. Aşağıdaki şekilde net biçimde görüldüğü gibi iki öğrencinin gerçek puanını içeren aralıklar büyük oranda kesişmektedir. Bu da ölçme sonuçlarına dayalı olarak karar verirken, özellikle de belli bir kesme puanına göre geçti/kaldı, başarılı/başarısız türünde sınıflandırmalar yaparken, ölçmenin standart hatasını hesaba katmanın daha adil/savunulabilir sonuçlara ulaşabilmek için ne denli önemli olduğunu gözler önüne sermektedir.



4. GEÇERLİK

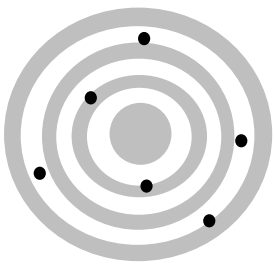
Geçerlik, ölçme ve değerlendirme alanının en yanlış anlaşılan ve en hatalı kullanılan terimlerinden biridir (Frisbie, 2005). Yanlış anlaşılmaya ve yorumlanmaya müsait olması sebebiyle geçerlik için geçmişten günümüze çeşitli tanımlar ileri sürülmüştür. 1940 ile 1950'lerin başında yapılan ilk tanımlarda test puanlarının nasıl kullanıldığı dikkate alınmadan geçerlik ölçme aracına ait mutlak bir özellik olarak karakterize edilmiştir (Oerman & Gaberson, 2021). Ölçme aracının "amaca hizmet etme derecesi" ve "ölçmeye çalıştığı yapıyı başka değişkenler ile karıştırmadan ölçebilmesi" bu bakış açısının ürünü olan tanımlardır. Ancak geçerlik ölçme aracının kendisine veya ölçümlere ait bir nitelik değildir, ölçme aracından elde edilen puanların ne kadar doğru yorumlandığıyla ilgili bir özelliktir. Bu da vurgunun ölçme aracında ya da ölçümlerde olduğu geçerlik tanımlarını gözden geçirme ihtiyacını doğurmuştur. Bu bağlamda *Eğitsel ve Psikolojik Test Standartları*'nda geçerlik "test puanlarına dayalı olarak yapılan yorumların, bu yorumların doğruluğunu test etmek için toplanan kanıtlar ve ölçülen özelliğin teorik çerçevesi ile uyumu" şeklinde kavramsallaştırılmıştır (AERA, APA & NCME, 2014). Mesela bir fizik testindeki maddeleri cevaplayabilmek için ağır matematiksel işlemler yapmak gerekiyorsa öğrencilerin test puanları fizik bilgilerinin yanı sıra matematik bilgilerini de yansıtır. Diğer bir deyişle ölçülmesi hedeflenen özellik fizik başarısı iken matematik başarısı da ölçme sonuçlarına karışmış olur. Bu durumda test puanlarını sadece fizik başarısının göstergesi olarak yorumlamak bir geçerlik sorununu gündeme getirir. Buna karşın tüm cevaplayıcıların matematik yeterliklerinin oldukça yüksek olduğu bir başka grupta ilgili testten alınan puanlar fizik başarısının göstergesi olarak kabul edilebilir. Dolayısıyla aynı test ikinci grupta amaca daha fazla hizmet eder ve çok daha geçerli sonuçlar verir. Benzer şekilde, lise öğrencilerinin kimya

bilgisini yoklamaya yönelik Almanca hazırlanmış bir test düşünelim. Bu test Almanya'daki liselerde uygulandığında test puanları öğrencilerin kimya bilgilerinin göstergesi olabilir. Oysa aynı test Türkiye'deki liselerde uygulandığında test puanlarını kimya bilgisinin bir ölçüsü olarak yorumlamak hatalı olur. Çünkü kullanılan test Türkiye örnekleminde kimya bilgisi kadar Almanca bilgisini de ölçer ve test puanlarını yalnızca kimya bilgisiyle ilişkilendirmek bir geçerlik problemi yaratır (İlhan ve ark., 2024).

Geçerlik, güvenirlik ile yakından ilişkilidir. O kadar ki Salkind ve Frey (2023) geçerlik ile güvenirliği birinci dereceden kuzenler olarak nitelendirir. Ölçümler güvenilir değilse bu ölçmelerden hareketle yapılan yorumların geçerli olması imkânsızdır. Bununla beraber güvenilir ölçümler, ölçümlere dayalı çıkarımların da mutlaka geçerli olacağı anlamına gelmez. Güvenirlik, geçerlik için gereklidir fakat yeterli değildir. Bunun tersi ise doğrudur. Test puanlarına bağlı olarak yapılan yorumlar geçerliyse söz konusu puanlar güvenilir demektir. Domino ve Domino (2006) geçerlik ve güvenirlik arasındaki ilişkiyi şöyle örneklendirmektedir: Bir mezura aracılığıyla başınızın çevresini ölçüp bulduğunuz değeri 6,3 ile çarpın. Daha sonra ulaştığınız sonuca sol elinizdeki parmak sayısının üç katı ile göz sayınızın altı katını ekleyin. Bu işlem sizi IQ puanınıza götürür. Bir sınıftaki öğrencilerden bunları yapmasını ve sıralanan işlemleri birkaç kez tekrarlamalarını istedik diyelim. Öğrencilerin baş ebatları değişmediği ve tekrarlı ölçmeler arasında vücut uzuvlarında bir kayıp yaşamadıkları sürece hesaplanan IQ puanları aynı kalır. Bu da elde edilen puanların son derece güvenilir olduğunu bildirir. Ancak ulaşılan ölçümler gerçekte IQ puanını yansıtmayacağından sonuçlar geçerli olmaz. Güvenirliğin geçerlik için ön koşul niteliği taşıdığı ama geçerliği garanti etmeyeceğini açıklamak için sıklıkla Şekil 2'deki görselden yararlanılır. Şekil 2'ye bakıldığında bir numaralı görselde atışların ne tutarlı ne de isabetli olduğu görülmektedir. Bu sonuç, güvenirliğin ve geçerliğin düşük olduğunu ima etmektedir. İki numaralı görselde atışlar tutarlıdır ama isabetli değildir. Buna göre güvenirlik yüksek ama geçerlik düşüktür. Üç numaralı görselde ise atışların hem tutarlı hem isabetli olması güvenirliğin de geçerliğin de yüksek olduğunu yansıtmaktadır.

Şekil 2

Geçerlik ve Güvenirlik Arasındaki İlişki (Fraenkel ve ark., 2012; Suter, 2012)



(I) Düşük Güvenirlik, Düşük Geçerlik



(II) Yüksek Güvenirlik, Düşük Geçerlik



(III) Yüksek Güvenirlik, Yüksek Geçerlik

Geçerlik bazı yönleriyle güvenirliğe benzerken kimi yönleriyle güvenirlikten ayrılır. Öncelikle geçerlik de güvenirlik gibi doğrudan gözlenemez, dolaylı olarak tespit edilebilir. Geçerliğin güvenirliğe benzeştiği ikinci nokta; derecelerinin olması, var-yok biçimde değerlendirilememesidir. Bu bakımdan "geçerlidir" ya da "geçerli değildir" gibi kalıplar kullanmak doğru değildir. Bunun yerine "geçerliği düşüktür", "geçerliği yüksektir" gibi ifadeler benimsenmelidir. Geçerlik ve güvenirliği birbirinden ayıran başlıca özellikler ise şunlardır: Güvenirlik *tutarlılık* kavramı ile eşleştirilirken geçerlik *doğruluk* kavramı ile eşleştirilebilir. Güvenirlik tesadüfi hatalardan etkilenirken geçerlik sabit, sistematik ya da tesadüfi fark etmeksizin tüm hata türlerinden etkilenir. Ayrıca güvenirliğin kararlılık, iç tutarlılık, eşdeğerlik ve objektiflik gibi farklı anlamları bulunsa da bunlar tek bir istatistiksel değer ile belirlenebilir. Söz gelimi ölçümlerin iç tutarlılığı Cronbach alfa katsayısı ile saptanabilir. Ancak geçerliğe gelince işin rengi değişir, tek bir katsayıya dayanarak geçerlik hakkında yorum yapmak mümkün olmaz. Geçerliğin test edilmesi, ölçme aracından alınan puanlara dayanarak ne tür çıkarımlar yapılabileceğine ya da hangi çıkarımlarda bulunmanın yanlış olacağına dair bir kanıt oluşturma sürecini içerir. Bu süreçte bir suçluyu ortaya çıkarmak için delil peşinde koşan dedektifler gibi hareket etmek, test puanlarının doğru yorumlandığı konusunda güçlü kuramsal ve ampirik kanıtlar sunmak gerekir (Coaley, 2010). Geçerliği değerlendirme işleminin statik değil dinamik bir yapıya sahip olduğu da unutulmalıdır (Sireci, 2007). Buna göre başlangıçta suçsuz bulunan bir sanığın toplanan yeni delillerin ardından suçlu olduğu anlaşılabileceği gibi geçerlik hakkında yeni kanıtlar toplandıkça önceki kanıtların işaret ettiğinden daha farklı bir sonuçla karşılaşılabilir. Elbette, her zaman böyle olmak zorunda değildir. Nasıl ki yeni deliller kimi zaman sanığın suçluluğuna ya da suçsuzluğuna yönelik kanaati güçlendirebiliyorsa yeni kanıtların geçerliğe ilişkin önceki değerlendirmeleri desteklemesi de mümkün olabilir. Sözün özü, "Test puanlarına ilişkin yorumların geçerli olduğu doğrulanmıştır, daha başka geçerlik kanıtları toplamaya gerek yoktur" gibi bir iddiada bulunmak olanaksızdır (Swanson, 2012).

4.1. Geçerlik Kanıtı Türleri

Geçerlik hakkında toplanabilecek kanıtlar oldukça çeşitlidir. Bununla birlikte, ilgili kanıtlar belli başlıklar altında gruplandırılmaktadır. *Eğitsel ve Psikolojik Test Standartları*'nın 1985 yılındaki basımında geçerlik kanıtları; kapsamla, yapıyla ve ölçütle bağlantılı kanıtlar olmak üzere üçe ayrılmıştır. Kitabın 1999 yılında yayınlanan basımıyla birlikte ise geçerliğe ilişkin kanıtlar Şekil 3'te sunulan beş başlıkta kategorize edilmiştir (McDonald, 2018). Aynı kategorizasyon kitabın 2014 tarihli son baskısında da benimsenmiştir. Bu aşamada şunun da altını çizmekte fayda vardır. Şekil 3'te yer alan ve günümüzde kabul gören başlıklar geçerlik türlerini değil, geçerlik kanıtı türlerini ifade eder. Geçerlik tekil (üniter) bir kavram olup farklı geçerlik türleri yoktur, geçerliğe ilişkin farklı kanıtlar mevcuttur. Bu anlamda, 1950'li yılların ilk yarısından 1985'e kadar hâkimiyetini sürdüren geleneksel geçerlik teorisinde var olan "geçerlik türleri" anlayışı (Messick, 1989) tamamıyla terk edilmelidir.

Şekil 3

Geçerlik Kanıtı Türleri



4.1.1. Kapsama Dayalı Kanıtlar

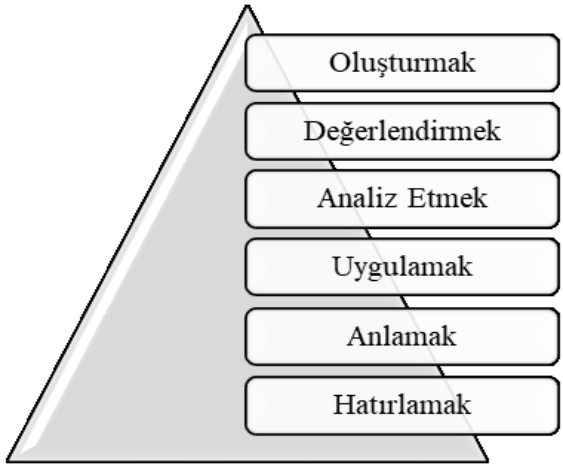
Kapsama dayalı geçerlik kanıtları, ölçme aracının ölçmeye konu özelliği ne kadar iyi temsil edebildiği sorusuna yanıt arar. Teorik olarak bir yapıyı ölçmeye yönelik sonsuz sayıda madde oluşturulabilir. Ancak uygulamada bu mümkün olmadığından belli sayıda madde seçilir ve ölçme aracına dahil edilir (Baumgartner ve ark., 2021). Hedef yapıyı tam anlamıyla temsil eden bir madde örneklemi seçebilmek için ilk yapılması gereken ayrıntılı bir test planı hazırlamaktır. Bu aşama, bir mimarın inşa süreçleri öncesinde kâğıt üzerinde çizimler yapmasına benzer (Violato, 2019). İnşaat sonucunda ortaya çıkacak binanın sağlam ve beklentileri karşılar nitelikte olması mimarın sürecin başında doğru çizimler yapmasına bağlı olduğu gibi ölçme sonucunda geçerli ölçümlere ulaşabilmenin yolu da uygun bir test planıyla işe başlamaktan geçer. Test planı sayesinde ölçülecek kapsama ilişkin analitik bir çözümleme yapılır. Çözümlemede, hedef içeriğin hangi bileşenlerden oluştuğunu betimlemeye yarayan ve ölçme literatüründe "belirtke tablosu (table of specification/test blueprint)" olarak adlandırılan bir matris kullanılır.

Belirtke tablosu satırlarında hedef davranışları, sütunlarında ise bu hedef davranışların bilişsel düzeylerini listeler. Ölçülecek içeriğin böyle bir tabloya yerleştirilmesi; bazı hedef davranışların ölçme aracında temsil edilmemesi, gereğinden az ya da fazla temsil edilmesi gibi riskleri minimuma indirir. Ayrıca tablo hedef davranışların bilişsel düzeylerini içerdiğinden hangi madde türünü/türlerini kullanmak gerektiğine dair de yol gösterici olur. Bu anlamda belirtke tablosu, hem geçerliği sağlamaya hizmet eder hem de geçerliğe kanıt oluşturur. Test geliştiricisi belirtke tablosunu hazırlarken önce hedef davranışları, diğer bir deyişle ölçülecek öğrenme çıktılarını tespit etmeli ve devamında

öğrenme çıktılarının bilişsel düzeylerini saptamalıdır. Alanyazında bu amaçla kullanılabilecek Bloom, SOLO, Fink ve Dettmer gibi farklı taksonomiler bulunmaktadır. Bunlar arasında en fazla kullanılanı ise Bloom taksonomisidir (Arı, 2013).

Bloom taksonomisi doğrudan bu bölümün kapsamına giren bir başlık değildir. Bununla birlikte aşağıda Bloom taksonomisinden ana hatlarıyla söz edilmiş, taksonomi hakkında sunulan bilgiler belirtke tablosunu örneklendirmeye imkân tanıyacak bir içerikle sınırlandırılmıştır. Dolayısıyla yalnızca taksonomiye ait bilişsel süreçler üzerinde durulmuştur.

Bloom taksonomisine göre bilişsel süreçler; *hatırlamak*, *anlamak*, *uygulamak*, *analiz etmek*, *değerlendirmek* ve *oluşturmak* şeklinde basitten karmaşığa doğru ilerleyen altı düzeyden oluşur. Bu düzeyler, alt basamaklar geçilmeden üst basamaklara erişmenin mümkün olmadığı hiyerarşik bir yapıya sahiptir (Anderson & Krathwohl, 2001). Düzeylerin temel özellikleri Şekil 4'te açıklanmıştır.



Şekil 4

Bloom Taksonomine göre Bilişsel Süreç Basamakları [Krathwohl'dan (2002) Yararlanılarak Oluşturulmuştur]

Hatırlamak: Görünce tanıma, ezberden tekrar etme ile ilgilidir. Örneğin; öğrenciden bir tanım yapması istendiğinde kitaptaki tanımı aynen aktarması, benzer şekilde bir kavramı örneklendirmesi talep edildiğinde verdiği örneklerin derste sunulanlardan ibaret kalması hatırlama basamağına karşılık gelir. Yine bir olayın tarihini söyleme, bir antlaşmanın maddelerini sıralama hatırlama düzeyine ilişkin örneklerdir.

Anlamak: Bu basamakta öğrenci öğrendiklerini kendi cümleleriyle ifade edebilir, farklı temsil biçimleri arasında dönüşümler yapabilir, ayrıca sınıflama, yorumlama, açıklama ve özetleme işlemlerini gerçekleştirebilir. Örneğin; öğrenci bir kavramı kitapta geçtiği şekliyle tanımlamak yerine kendi cümleleriyle açıklayabilir, kavrama yönelik kitapta/derste sunulanların ötesinde başka örnekler verebilir, okuduğu bir parçayı özetleyip parçanın ana fikrini belirleyebilir, aktif bir cümleyi edilgen cümleye çevirebilir, sözel ifadeleri grafiğe aktarabilir ya da bir grafiği sözel olarak yorumlayabilir.

Uygulamak: Bireyin öğrendiklerini karşılaştığı problem durumlarında ve/veya gerçek hayatta kullanılabildiği basamaktır. Konu hakkındaki bilgilerinden yola çıkarak bir matematik problemini çözmek veya kompozisyon yazarken noktalama işaretlerini doğru şekilde kullanmak uygulama basamağına örnek olarak sunulabilir.

Analiz etmek: Bir bütünü bileşenlerine ayırma, parçaların birbiriyle ve bütünle ilişkisini saptama, parçalar arasındaki ilişkilerden hareketle bütüne ulaşma analiz etme düzeyinin karakteristik özellikleridir. Söz gelimi, merkezi sınavlarda sıkça yer verilen "Bu parça iki paragrafa ayrılacak olursa ikinci paragraf hangi cümleyle başlar?" tarzındaki sorular analiz etme düzeyine karşılık gelir.

Değerlendirme: Bir ölçütle karşılaştırıp yargıya varma söz konusudur. Mesela bir test yönergesini taşıması gereken özellikler açısından gözden geçirip yönergenin ne kadar iyi hazırlandığına dair yargıya varmak değerlendirme basamağıyla eşleşen bir işlemdir.

Oluşturmak: Yeni, tutarlı ve özgün bir ürünün ortaya konulabildiği basamaktır. Tez yazmak, proje hazırlamak, beste yapmak, bir öğretim materyali tasarlamak oluşturma basamağında gerçekleştirilebilecek eylemlerdir.

Bloom taksonomisinin hatırlama, anlama ve uygulama basamakları alt; analiz, değerlendirme ve oluşturma basamakları ise üst düzey bilişsel süreçler olarak kabul edilir. Alt bilişsel süreçlerden üst bilişsel süreçlere doğru gidildikçe yanıtı yapılandırmayı gerektiren ölçme araçlarına (örneğin; yazılı ve sözlü sınavlar, performans görevleri) duyulan ihtiyaç artar. Özellikle oluşturma basamağına gelindiğinde yanıtı seçmeyi gerektiren ölçme araçları (doğru-yanlış testleri, çoktan seçmeli testler vb.) amaca hizmet edemez, öğrenme çıktıları ancak yanıtı yapılandırmayı gerektiren araçlarla ölçülebilir.

Bloom taksonomisinin temel çerçevesini öğrendiğimize göre artık örnek bir belirtke tablosu inceleyebiliriz. Aşağıda özel dörtgenler konusuna ilişkin bir belirtke tablosu sunulmuştur (bkz. Tablo 9). Tablo 9'a göre ölçülmesi gereken öğrenme çıktılarının sayısı 25'tir ve geliştirilecek test beş ayrı dörtgen türüyle ilgili maddeler içermelidir. En fazla madde yamukla ilgili olmalı (%28), bunu kare (%20) ve dikdörtgene (%20) ilişkin madde sayıları izlemeli, paralelkenar (%16) ve eşkenar dörtgene (%16) ilişkin madde sayıları ise görece daha az olmalıdır. Tablo 9'daki dağılımdan testte anlama, uygulama ile analiz etme düzeylerine karşılık gelen maddelere yer verilmesi ve maddelerin yaklaşık yarısının uygulama düzeyinde olması gerektiği anlaşılmaktadır. Belirtke tablosu tam olarak bunu, öğrenme çıktılarını kuşbakışı görebilmeyi sağlar (Earnest & Gupta, 2022). Böylece hedef içeriğin bazı kritik bileşenlerinin test geliştirme sürecinde gözden kaçması ihtimalini azaltır.



Sınıf içinde uygulanan sınavlar, konu ya da ünite testleri genellikle belirtke tablosunda listelenen her bir öğrenme çıktısıyla ilgili maddeler içerecek tarzda hazırlanır. Diğer taraftan geniş ölçekli sınavlarda bir veya birkaç öğretim yılı boyunca işlenen kapsamlı bir içeriğin ölçülmesi amaçlandığından bunu yapmak olanaksızdır. Bu gibi durumlarda kritik hedef davranışlara odaklanılır ve ölçme aracı bu doğrultuda organize edilir.

Tablo 9

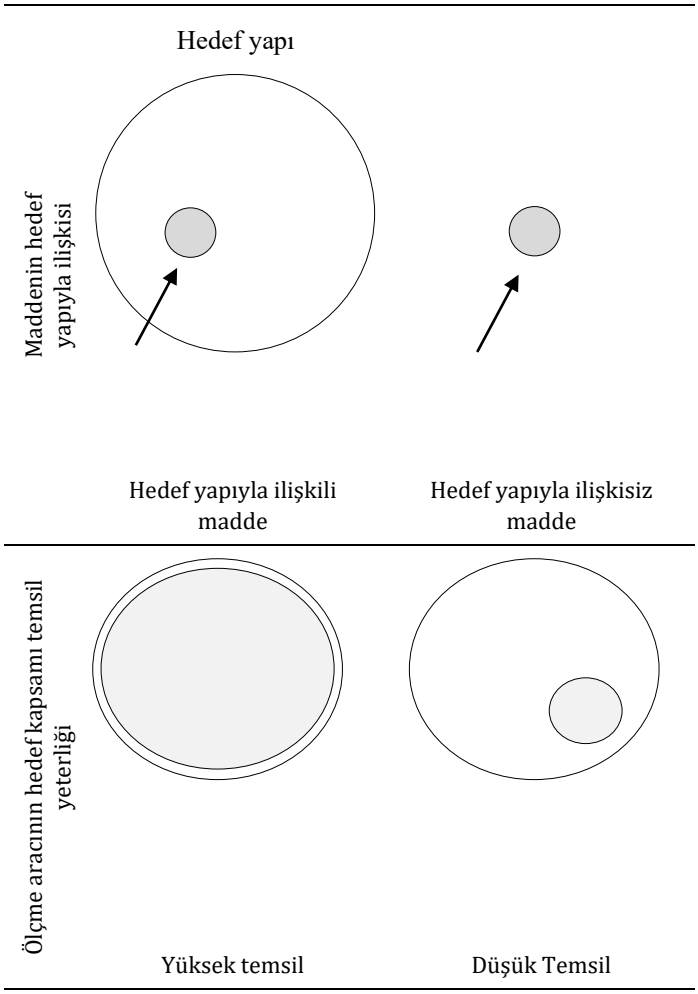
Bloom Taksonomisine göre Hazırlanmış Belirtke Tablosu Örneği

	Bilişsel Düzeyler						
Konular	**Hatırlamak**	**Anlamak**	**Uygulamak**	**Analiz Etmek**	**Değerlendirmek**	**Oluşturmak**	**Toplam**
Dikdörtgen	–	1 (%4)	2 (%8)	2 (%8)	–	–	5 (%20)
Kare	–	1 (%4)	2 (%8)	2 (%8)	–	–	5 (%20)
Paralelkenar	–	1 (%4)	2 (%8)	1 (%4)	–	–	4 (%16)
Eşkenar dörtgen	–	1 (%4)	2 (%8)	1 (%4)	–	–	4 (%16)
Yamuk	–	2 (%8)	3 (%12)	2 (%8)	–	–	7 (%28)
Toplam	–	6 (%24)	11 (%44)	8 (%32)	–	–	25 (%100)

Belirtke tablosunu düzenleyip ölçme aracını bu tablodaki dağılım doğrultusunda hazırlamak kapsama dayalı önemli bir geçerlik kanıtı sağlasa da bu kanıtı uzman görüşleriyle desteklemek son derece önemlidir. Bunun için taslak ölçme aracı alan uzmanlarına sunulmalı, uzmanlar hem her bir maddenin hedef yapıyla ilişkisini (item relevance) değerlendirmeli hem de maddelerin bu yapıyı ne kadar kapsadığına (content coverage) dair görüş bildirmelidir. Çünkü maddelerin tamamı ölçülen yapıyla ilişkili olduğu halde ölçme aracı hedef yapıyı yeterince kapsamıyor olabilir. Örneğin, obsesif-kompulsif semptomları değerlendirmeye yönelik bir ölçek için yazılan maddelerin tamamı ritüel davranışlar (el yıkama alışkanlığının varlığı ve sıklığı gibi) ile ilişkiliyse madde bazındaki incelemeler geçerliğe ilişkin olumlu bir tablo ortaya koyar. Ancak maddelerin obsesif kompülsif bozukluğun ritüel davranışların dışında kalan semptomlarını içermemesi ölçme aracının hedef yapıyı tam anlamıyla temsil edemediğini yansıtır (Reynolds ve arkadaşları, 2021). Bu durumda madde düzeyindeki incelemelerin aksine ölçme aracının geneline ilişkin değerlendirmeler geçerliğe dair olumsuz bir sonuç bildirir. Maddenin ölçülen özellikle ilişkisi ve ölçme aracının hedef kapsamı temsil yeterliği arasındaki bu ayrım Şekil 5'te resmedilmiştir.

Şekil 5

Maddenin Ölçülen Özellikle İlişkisinin ve Ölçme Aracının Hedef İçeriği Temsil Kabiliyetinin Görsel Sunumu (Reynolds ve ark., 2021)



İlhan ve arkadaşları (2024) maddeler hakkında en az iki alan uzmanından görüş alınmasını önermektedir. Ancak uzman sayısının odak nokta haline getirilmemesi gerektiğini savunmakta, yetkinliği yüksek az sayıda uzmanın düşük yetkinlikteki çok sayıda uzmana kıyasla maddeler hakkında daha önemli geri bildirimler sunabileceğini belirtmektedir. Tabi alan uzmanlarının yanında maddelerin madde yazım ilkelerine uygunluğunu değerlendirmesi amacıyla ölçme ve değerlendirme uzmanlarından; yazım, imla ve noktalama kuralları açısından inceleme yapması amacıyla da dil uzmanlarından görüş talep edilmelidir.

Uzmanlardan görüş alırken uzmanların maddelerle ilgili düşüncelerini nicel olarak ifade edebilecekleri bir form kullanılması, kapsama dayalı geçerlik kanıtlarını daha somut hale getirir. Bundan dolayı uzman görüşlerine başvururken Şekil 6'daki tarzda bir form hazırlanması önerilir.

Şekil 6

Örnek Uzman Görüş Formu [Özer Özkan (2022) ile İlhan ve arkadaşlarından (2024) uyarlanmıştır]

Sayın Uzman;

Aşağıda yer alan maddeler …….…….. amacıyla hazırlanmış olup ölçme aracının ……..…….. kişilerden oluşan bir örnekleme uygulanması planlanmaktadır. Sizden taslak ölçme aracındaki her bir maddenin belirtilen amaca ne kadar hizmet ettiğini değerlendirmeniz rica edilmektedir. Madde;

- Hedef yapıyla ilişkili ve doğrudan kullanılabilir nitelikte ise *Gerekli-Aynen Kullanılabilir*
- Hedef yapıyı ölçebilir özellikte ama revize edilmesi gereken kısımları varsa *Gerekli-Düzeltilmeli*
- Hedef yapıyla ilişkili ancak gerekli değilse *Yararlı fakat Gerekli Değil*
- Hedef yapıyı temsil etmiyorsa *Gereksiz*

seçeneklerini işaretleyiniz. Maddelere ilişkin görüşlerinizi en sağ sütunda detaylandırabilirsiniz. Özellikle revize edilmesi gerektiğini düşündüğünüz maddeler için görüşlerinizi ayrıntılı biçimde sunmanız ilgili düzeltmelerin yapılmasını kolaylaştıracaktır. Katkılarınızdan dolayı teşekkür ederim/ederiz.

Ölçme aracını geliştiren kişinin/kişilerin;
Unvanı, Adı ve Soyadı
İletişim bilgileri

Maddeler	Ölçtüğü öngörülen boyut/kazanım	Gerekli		Yararlı fakat Gerekli Değil	Gereksiz	Görüş ve Öneriler
		Aynen Kullanılabilir	Düzeltilmeli			
……	……					
……	……					
……	……					
……	……					
……	……					

Ölçme aracının hedef kapsamı temsil yeterliğini 1 ile 10 arasında puanlayacak olursanız kaç puan verirsiniz? ……. *(puan)*

Eklemek istedikleriniz:

……
……
……
……

Uzmanlardan görüş alındıktan sonra her bir madde için kapsam geçerlik oranı (KGO) hesaplanır. KGO uzman görüşlerini nicel verilere dönüştürerek kapsama dayalı geçerlik kanıtlarına nesnel bir kimlik kazandırır (Yurdugül & Bayrak, 2012). KGO'yu hesaplamak için Lawshe'nin (1975) geliştirdiği Eşitlik 8'deki formülden yararlanılır.

$$KGO = \frac{N_G}{N/2} - 1$$

N_G = Gerekli görüşünü bildiren uzman sayısı
N = Görüşüne başvurulan toplam uzman sayısı

Eşitlik 8

Eşitlik 8'e göre, bir maddeyi görüşüne başvurulan uzmanlardan hiçbiri gerekli bulmazsa KGO -1 çıkar. Madde için tüm uzmanların gerekli şeklinde görüş bildirmesi durumunda ise KGO +1 olarak hesaplanır. Yani KGO'nun alabileceği değerler ±1 aralığında uzanır (Özer Özkan, 2022). Burada akla gelecek temel soru ise taslak ölçme aracındaki bir maddenin ölçme aracının nihai formuna dâhil edilebilmesi için KGO'sunun en az kaç olması gerektiğidir. Bununla ilgili ilk tanımlama Lawshe (1975) tarafından yapılmıştır. Lawshe (1975), *p*=0,05 anlamlılık düzeyinde uzman sayısına göre KGO'nun eşik değerlerini içeren bir tablo oluşturmuş, bu tablodaki kritik değerler daha sonra farklı araştırmacılar (Ayre & Scally, 2014; Wilson ve ark., 2012) tarafından da test edilmiştir. Ayre ve Scally (2014) yaptıkları incelemede Lawshe'nin (1975) belirlediği kritik değerlerin güvenilir olduğunu rapor etmiş, bununla beraber daha kesin olduğunu iddia ettikleri yeni sınır değerler ileri sürmüşlerdir. Tablo 10, KGO için Lawshe (1975) ile Ayre ve Scally (2014) tarafından tanımlanan kritik değerleri içermektedir.

Tablo 10

KGO için Tanımlanan Kritik Değerler

Lawshe (1975)		**Ayre ve Scally (2014)**			
Uzman Sayısı	**Kritik Değer**	**Uzman Sayısı**	**Kritik Değer**	**Uzman Sayısı**	**Kritik Değer**
5	0,99	5	1,00	23	0,391
6	0,99	6	1,00	24	0,417
7	0,99	7	1,00	25	0,440
8	0,75	8	0,750	26	0,385
9	0,78	9	0,778	27	0,407
10	0,62	10	0,800	28	0,357
11	0,59	11	0,636	29	0,379
12	0,56	12	0,667	30	0,333
13	0,54	13	0,538	31	0,355
14	0,51	14	0,571	32	0,375
15	0,49	15	0,600	33	0,333
20	0,42	16	0,500	34	0,353
25	0,37	17	0,529	35	0,314
30	0,33	18	0,444	36	0,333
35	0,31	19	0,474	37	0,297
40	0,29	20	0,500	38	0,316
		21	0,429	39	0,333
		22	0,455	40	0,300

Hangi maddelerin ölçme aracında kalması gerektiğine karar verirken her bir madde için hesaplanan KGO, Tablo 10'daki kritik değer ile karşılaştırılır. Bu karşılaştırma doğrultusunda iki kutuplu bir karar alınır. KGO'su kritik değeri aşıyorsa madde ölçme aracının nihai formuna alınır, yok eğer eşik değerin altında kalıyorsa madde testten

çıkarılır. Son adımda ise nihai forma seçilen maddelere ait KGO değerlerinin ortalaması alınıp ölçme aracının geneline ilişkin kapsam geçerlik indeksi bulunur. Çok boyutlu ölçme araçlarında bu son adım her boyut için ayrı ayrı yapılır (Yurdugül, 2005). Mesela ölçme aracında A ve B şeklinde iki boyutlu bir yapı varsa A boyutuna seçilen maddelerin KGO'larının ortalaması alınarak bu boyutun kapsam geçerlik indeksi elde edilir ve aynı işlem B boyutu için de tekrarlanır. Örnek teşkil etmesi açısından Tablo 11'de 10 maddelik bir test için 15 uzmandan görüş alındığı varsayılmış, uzmanların değerlendirmeleri doğrultusunda maddelere ait KGO değerleri ile ölçme aracının geneline ilişkin kapsam geçerlik indeksinin hesaplanmasına yönelik süreç özetlenmiştir.

Tablo 11

KGO İçin Örnek Uygulama

Maddeler	Gerekli diyen uzman sayısı	KGO	Ölçme Aracının Geneline ait Kapsam Geçerlik İndeksi
M1	14	[14/(7,5)] – 1 = 0,87	15 uzman için kritik KGO değeri Lawshe'ye (1975) göre 0,49 iken Ayre ve Scally'ye (2014) göre 0,60'tır. Bu iki ölçütten hangisi dikkate alınırsa alınsın ölçme aracından çıkarılması gereken maddelerin M5, M7 ve M9 olduğu anlaşılmaktadır. Kalan yedi madde üzerinden nihai test için kapsam geçerlik indeksi hesaplanacak olursa $\frac{(0,87+1,00+0,73+1,00+0,60+0,73+1,00)}{7} = 0,72$ bulunur.
M2	15	[15/(7,5)] – 1 = 1,00	
M3	13	[13/(7,5)] – 1 = 0,73	
M4	15	[15/(7,5)] – 1 = 1,00	
M5	7	[7/(7,5)] – 1 = –0,07	
M6	12	[12/(7,5)] – 1 = 0,60	
M7	10	[10/(7,5)] – 1 = 0,33	
M8	13	[13/(7,5)] – 1 = 0,73	
M9	9	[9/(7,5)] – 1 = 0,20	
M10	15	[15/(7,5)] – 1 = 1,00	

4.1.2. İçyapıya Dayalı Kanıtlar

Bu geçerlik kanıtında ölçme aracını oluşturan maddelerin ve boyutların ölçülmesi amaçlanan gizil yapı ile ne kadar örtüştüğü belirlenmeye çalışılır (Holmboe ve ark., 2018). Kapsama dayalı geçerlik kanıtlarına benzemekle birlikte burada maddeler bireysel olarak değerlendirilmez, maddeler arası ilişkilere odaklanılır (Mertler, 2022). Örneğin, bir test tek bir yetenek ya da özelliği ölçmek amacıyla tasarlandıysa testin her bir maddesinden alınan puanlar ile geneline ait puanlar arasında güçlü korelasyonlar bulunmalıdır. Öte yandan ölçmeye konu özellik çeşitli faktörlerden oluşuyor ve test bu faktörler doğrultusunda hazırlandıysa maddeler, ölçmeye çalıştığı faktörden alınan toplam puan ile yüksek korelasyonlar göstermeli; diğer faktörlerden ve testin genelinden elde edilen puanlar ile görece daha zayıf bir ilişki içerisinde olmalıdır (Morrow ve ark., 2023). Kısacası maddeler ve boyutlar arasında gözlenen ilişkiler, hedef yapının teorik çerçevesi ile paralellik taşımalıdır.

Konuyu biraz daha işevuruk hale getirmeye çalışalım. Kişiliği ölçmek amacıyla beş faktörlü kişilik modeline dayalı bir envanter geliştirdiğimizi varsayalım. Büyük beşli olarak da adlandırılan beş faktörlü modele göre kişilik; "dışadönüklük, uyumluluk, sorumluluk, nevrotiklik ve deneyime açıklık" boyutlarından oluşur (McCrae & Costa, 2003; McCrae & John, 1992). Bu durumda hazırladığımız ölçme aracı sıralanan beş boyutlu yapı ile uyumlu olmalıdır. Bu uyumu test etmek için ölçme aracı yeterli büyüklükte bir gruba uygulanır ve toplanan veriler üzerinden çeşitli analizler gerçekleştirilir. Yapılan analizler maddeler arası ilişkilerin referans alınan beş boyutlu yapıyla eşleştiğini gösterirse bu sonuç geçerlik için içyapıya dayalı kanıt teşkil eder.

İçyapıya dayalı geçerlik kanıtlarına ulaşmak için en çok başvurulan teknik faktör analizidir (Reynolds ve ark., 2021). Faktör analizi ölçme aracındaki maddeleri aralarındaki korelasyonların örüntüsü ile uyumlu, doğrudan gözlenmesi mümkün olmayan bir veya birkaç faktör altında toplamaya dönük bir veri indirgeme tekniğidir (İlhan ve ark., 2024). Kapsamı gereği bu bölümde faktör analizine yönelik ayrı bir başlık açılmamıştır. Bununla beraber, faktör analizinin en azından mantığının kavranmasına yardımcı olacağı düşüncesiyle Tablo 12'de örnek bir korelasyon matrisi sunulmuştur. Hipotetik bir veri seti üzerinden elde edilen bu korelasyon matrisi, 6 maddelik (M1-M6) bir testte katılımcıların madde puanları arasındaki ilişkileri göstermektedir. Testin ilk üç maddesinin sözel, devamındaki üç maddesinin ise sayısal yeteneği ölçtüğü öngörülmektedir.

Tablo 12

İki Boyutlu Olduğu Öngörülen Altı Maddelik Bir Test için Korelasyon Matrisi

Maddeler	M1	M2	M3	M4	M5	M6
M1	–					
M2	0,610	–				
M3	0,427	0,481	–			
M4	0,144	0,159	0,163	–		
M5	0,166	0,177	0,124	0,617	–	
M6	0,183	0,161	0,177	0,414	0,560	–

Tablo 12'deki korelasyon katsayılarına göre M1, M2 ve M3'ün birbiriyle ilişkisi, testin devamındaki maddeler ile olan ilişkilerine kıyasla daha güçlüdür. Benzer şekilde M4, M5 ve M6'nın kendi aralarındaki ilişkiler testin ilk üç maddesiyle olan ilişkilerinden belirgin derecede yüksektir. Yani korelasyon matrisi, maddelerin beklendiği gibi iki boyut altında kümelendiğine işaret etmektedir. İşte faktör analizi, maddeler arası ilişkiler üzerine temellendirilen bir dizi istatistiksel işlemle ölçme sonucunda gözlenen yapının kuramsal olarak öngörülen bileşenleri destekleyip desteklemediğini ortaya koyar. Bu yolla geçerliğe ilişkin içyapıya dayalı kanıtlar sunar.

4.1.3. Ölçme Aracından Alınan Puanların Diğer Değişkenler ile İlişkisine Dayalı Kanıtlar

Ölçülmesi hedeflenen yapıya benzer olduğu bilinen ya da bu yapıdan ayrı bir özelliği temsil ettiğine inanılan dışsal bir değişken ölçüt olarak alınır. Ölçme aracından elde edilen puanlar ile ölçüt değişkene ait puanlar arasındaki ilişki incelenerek geçerlik hakkında değerlendirme yapılır. Ulaşılan sonuçların hedef yapıyla ilgili kuramsal çerçeveyi ve/veya daha önce yapılmış ampirik araştırmaları destekler nitelikte olması geçerliğe kanıt sağlar. Kettler (2019) ölçme aracının diğer değişkenlerle ilişkisine dayalı kanıtları; eş zamanlı ve yordayıcı kanıtlar, yakınsak ve ıraksak kanıtlar ile karşıt gruplara dayalı kanıtlar şeklinde kategorilere ayırmıştır. Bu bölümde Kettler'in (2019) yaptığı sınıflama dikkate alınıp ölçme aracının diğer değişkenlerle ilişkine dayalı kanıtlar üç başlıkta sunulmuştur.

Ölçmeye konu hiçbir özellik diğer yapılardan bütünüyle izole değildir. Aksine her bir yapı, içerisinde başka yapıların da bulunduğu bir sistemin parçasıdır. Bu sistemde birbirine benzer özellikleri ölçen yapılar bulunabildiği gibi farklı özellikleri ölçen yapılar da yer alabilir. Geçerliğe kanıt ararken bunun farkında olmak ve hedef yapının nomolojik ağını incelemek önemlidir (Suen & Rzasa, 2004). Geçerlik çalışmalarında ölçme aracının diğer değişkenlerle ilişkisini analiz etmenin arkasında yatan felsefe budur.

4.1.3.1. Eş Zamanlı ve Yordayıcı Kanıtlar

Geçerliği sorgulanan ölçme aracından elde edilen puanlar ile onunla ilişkisi olması beklenen bir başka değişkene ait veriler arasındaki ilişkiye bakılır. Eş zamanlı ve yordayıcı kanıt ayrımının altında, ölçüt değişkene ait verilerin ne zaman elde edildiği yatar. Eş zamanlı kanıtlarda ölçüt değişkene ait veriler, hedef yapıya ait ölçümler ile aynı anda toplanır. Buna karşılık yordayıcı kanıtlarda ölçüt değişkenin verileri daha ileri bir zaman dilimine aittir. Akademik erteleme eğilimi ölçeğinin geçerliğine ilişkin bir çalışma yürütüyor olalım. Bu kapsamda üniversite son sınıf öğrencilerine akademik erteleme eğilimi ölçeği ile zaman yönetimi becerileri ölçeğini aynı anda uyguladık diyelim. Ayrıca öğretim yılı sonunda, öğrenci işleri daire başkanlığından çalışmaya katılan öğrencilerden kaçının mezun olmaya hak kazandığı ve kaçının dönem uzattığı ile ilgili veri elde ettiğimizi kabul edelim. Bu çalışmada akademik erteleme eğilimi ölçeği ile zaman yönetimi becerileri ölçeğinden alınan puanlar arasında negatif yönlü anlamlı korelasyon saptanması geçerliğe yönelik eş zamanlı kanıt sağlar. Akademik erteme eğilimi ölçeğinden alınan puanların mezun olan ve dönem uzatan öğrencileri büyük ölçüde doğru sınıflaması, yani ölçekten düşük puan alan öğrenciler genelde mezun olurken dönem uzatanların çoğunlukla ölçekten yüksek puan alanlar arasından olması geçerlik hakkında yordayıcı kanıt sunar.



Yordayıcı kanıtlara ulaşmak eş zamanlı geçerlik kanıtlarına kıyasla daha fazla vakit alır ve daha çok çaba gerektirir (Johnson & Christensen, 2017). Bu sebeple literatür incelendiğinde geçerliğe ilişkin yordayıcı kanıtların daha az raporlandığı görülür. Ancak unutmamak gerekir ki özellikle seçme ve yerleştirmeye yönelik sınavlarda, yordayıcı kanıtların incelenmesi sınavın amaca ne kadar hizmet ettiğine dair son derece önemli bilgiler verir.

4.1.3.2. Yakınsak ve Iraksak Kanıtlar

Geçerliği test edilen ölçme aracından alınan puanlar ile aynı özellik için farklı yöntemlerden ve/veya daha önce geçerliği kanıtlanmış başka bir ölçme aracından elde edilen puanlar arasında pozitif yönlü güçlü korelasyonlar bulunması yakınsak geçerlik kanıtı sağlar (Dougherty, 2019; Whitley & Kite, 2013). Iraksak geçerlik kanıtlarında ise hedef yapıdan farklı bir özelliği ölçtüğü bilinen dışsal bir değişken vardır. Geçerli sonuçlar verdiği daha önce ispatlanmış olan bu dışsal değişkene ait ölçümler ile hedef yapı için elde edilen puanlar arasında ilişki bulunmaması ya da zayıf ilişkiler olması geçerliğe dair ıraksak kanıtlar sunar (Aiken, 1996). Yakınsak ve ıraksak geçerlik kanıtları genellikle Campbell ve Fiske (1959) tarafından ileri sürülen çoklu özellik çoklu yöntem yaklaşımı ile incelenir. Yakınsak ve ıraksak kanıtları aynı anda değerlendirmeye imkân veren bu yaklaşımda birden fazla özellik (örneğin, genel zihinsel yetenek ve motivasyon) birden fazla yöntem (örneğin, öz bildirim ve akran değerlendirme) kullanılarak ölçülür. Daha sonra ölçümler arasındaki korelasyonlar hesaplanır ve bir korelasyon matrisine yerleştirilir (Cascio & Aguinis, 2019). Aynı özellik için farklı yöntemlerden, farklı özellikler için aynı yöntemden ve farklı özellikler için farklı yöntemlerden elde edilen korelasyon katsayılarını içeren bu matrisin esas köşegeni üzerine ise ölçümlere ait güvenilirlik katsayıları yazılır (bkz. Tablo 13).

Tablo 13

*Çoklu Özellik Çok Yöntem Yaklaşımı için Örnek Korelasyon Matrisi**

		Yöntem1		**Yöntem2**	
		Özellik-A	**Özellik-B**	**Özellik-A**	**Özellik-B**
Yöntem1	Özellik-A	r_{A1}			
	Özellik-B	IG_1	r_{B1}		
Yöntem2	Özellik-A	YG_1	IG_3	r_{A2}	
	Özellik-B	IG_2	YG_2	IG_3	r_{B2}

Korelasyon matrisinin farklı özellikler için aynı yöntemden elde edilen ölçümler arasındaki ilişkileri ortaya koyması, istenmeyen sistematik bir varyans olan yöntem etkisini de tespit etmeyi mümkün kılar (Dimitrov, 2012).

Rowe ve Mahar'dan (2006) uyarlanmıştır. YG: Yakınsak geçerlik, IG: Iraksak geçerlik

İki ayrı özellik iki farklı yöntem ile ölçüldüğünden Tablo 13'te esas köşegen üzerinde dört ayrı güvenirlik değeri (r_{A1}, r_{B1}, r_{A2} ve r_{B2}) mevcuttur. Tablo yorumlanırken önce güvenirlik kestirimleri kontrol edilmeli, ölçümlerin güvenirliğinden emin olunduktan sonra korelasyon katsayılarına bakılmalıdır. Hesaplanan korelasyonların geçerliğe kanıt oluşturabilmesi için (a) aynı özellik için farklı yöntemlerden elde edilen ölçümler arasında pozitif yönlü güçlü ilişkiler gözlenmeli, (b) farklı özellikler aynı yöntemle ölçülmüş olsa dahi bu özellikler arasında zayıf ilişkiler bulunmalıdır (Sullivan & Feldman, 1979).

Çoklu özellik çoklu yöntem yaklaşımı hakkında yukarıda sıralanan özellikleri bir örnek üzerinde pekiştirelim. İlkokul öğrencilerinde saldırganlık ile dijital bağımlılığı öz bildirim ve davranışsal gözlem şeklinde iki farklı yöntemle ölçtüğümüzü kabul edelim. Bu iki yapı için iki ayrı yöntemden elde ettiğimiz puanlar arasında Tablo 14'teki korelasyon katsayılarını hesapladığımızı varsayalım.

Tablo 14

*Saldırganlık ve Dijital Bağımlılık Verileri için Çoklu Özellik Çoklu Yöntem Yaklaşımı Korelasyon Matrisi**

Yöntem	Özellik	Öz Bildirim		Davranışsal Gözlem	
		Saldırganlık	Dijital Bağ.	Saldırganlık	Dijital Bağ.
Öz Bildirim	Saldırganlık	(0,91)			
	Dijital Bağ.	*0,44*	(0,89)		
Davranışsal Gözlem	Saldırganlık	**0,81**	0,25	(0,88)	
	Dijital Bağ.	0,28	**0,77**	*0,41*	(0,85)

* Örneğin oluşturulmasında ve yorumlanmasında Urbina (2004, s. 172) ile Finch ve French (2019, s. 210) referans alınmıştır.

Tablo 14'te esas köşegen üzerinde parantez içerisinde sunulan katsayılar iç tutarlılık yöntemiyle hesaplanan güvenirlik değerlerini göstermekte, tüm katsayıların 0,80'i aşması ölçümlerin yüksek güvenirlikte olduğunu simgelemektedir. Buna göre ölçümlere ait güvenirliğin değişkenler arasındaki korelasyonda düşüşe yol açması ihtimali zayıftır, tablodaki korelasyon katsayıları güvenle yorumlanabilir. Tablo 14'te daha koyu yazılan korelasyon değerleri, aynı özellik için iki farklı yöntemden elde edilen puanların ilişkisini ifade etmektedir. Bu korelasyonların pozitif yönlü ve güçlü olması beklenmekte, tablodaki katsayılar bu beklentinin karşılandığını yansıtmaktadır. İtalik biçimde sunulan korelasyon katsayıları farklı psikolojik yapılar için aynı yöntemden elde edilen puanlar arasındaki ilişkileri belirtmektedir. Bu korelasyonlar, beklendiği gibi, aynı özelliğin farklı yöntemlerle ölçüldüğü durumdaki katsayılardan (koyu yazılan değerlerden) daha düşüktür. Son olarak düz metin formatındaki korelasyon katsayıları farklı özellikler için farklı yöntemlerden gelen puanların ilişkisini yansıtmakta, matristeki en düşük değerlerin bunlara ait olması beklenmektedir. Tablo 14'teki korelasyonlar bu beklentiyle

de uyumludur. Bütün bulgular bir arada değerlendirildiğinde değişkenler arasındaki ilişkide hedef yapıdan kaynaklanan varyansın yüksek, yönteme atfedilebilecek varyansın ise düşük olduğu anlaşılmaktadır. Dolayısıyla çoklu özellik çoklu yöntem matrisindeki korelasyon katsayılarının geçerliğe ilişkin sunduğu yakınsak ve ıraksak kanıtlar oldukça güçlüdür denilebilir.



- Çoklu özellik çoklu yöntem yaklaşımı geçerlik hakkında oldukça kritik bilgiler sağlayabilmektedir. Bu potansiyeline rağmen veri toplama sürecindeki güçlükler nedeniyle nadiren kullanılmaktadır (Göçer Şahin, 2021).
- Geçerliğe ilişkin yakınsak kanıtlar ile eş zamanlı ve yordayıcı kanıtların hareket noktası benzerdir. Üçü de hedef yapıya ait kavramsal çerçevenin ölçme aracından alınan puanlarla ilişkili olduğuna işaret ettiği, geçerliği önceden ispatlanmış dışsal bir değişkeni ölçüt olarak alır. Yakınsak kanıtlarda bu dışsal değişken geçerliği araştırılan ölçme aracıyla *aynı* özelliği ölçen başka bir araçtan/yöntemden elde edilen puanlardır. Eş zamanlı ve yordayıcı kanıtlarda ise geçerliği sorgulanan ölçme aracıyla aynı değil *benzer* özelliği ölçen bir dışsal değişken vardır (Brain, 2016; Brown, 2023).

4.1.3.3. Karşıt (Bilinen) Gruplara Dayalı Kanıtlar

Bu yöntemde ölçme aracından alınan puanlarla ilişkisi incelenen değişken sınıflama düzeyindedir. Ölçme aracının bu sınıflama değişkeninin farklı kategorilerinde yer alan bireyleri birbirinden ayırt edip edemediği incelenir. Bu incelemelerde genelde bağımsız örneklemler *t*-testi ve varyans analizi gibi grup ortalamalarını karşılaştırmaya yönelik analizler kullanılır. Analiz sonuçlarının grupların puanları arasında istatistiksel açıdan anlamlı fark bulunduğunu ortaya koyması ve bu farkın kuramsal yapının işaret ettiği yönde çıkması geçerliğe kanıt oluşturur. Daha istatistiki terimler ile ifade edecek olursak bu yöntemde yönlü bir alternatif hipotez test edilir. Yapılan analizler yokluk hipotezinin reddedilip yönlü alternatif hipotezin kabulü yönünde bir kararla neticelenirse bu sonuç geçerlik için karşıt gruplara dayalı kanıtlar sağlar. Örneğin; Gezer ve İlhan (2020) iklim değişikliği bilgi testini öğretmen adaylarından oluşan bir örneklem üzerinde Türkçeye uyarladıkları araştırmada, "Coğrafya öğretmenliği programına kayıtlı öğretmen adaylarının test puanları, diğer bölümlerde okuyan öğretmen adaylarınınkine kıyasla anlamlı derece daha yüksektir" hipotezini test etmişlerdir. Bu amaçla tek yönlü varyans analizi uygulamış, analiz çıktılarının hipotezlerini doğrulamasını test puanlarının geçerliğine kanıt olarak göstermişlerdir.

Biçer'in (2024) dikkat eksikliği tanılamasında kullanılmak üzere materyal destekli test bataryası geliştirdiği çalışma, test puanlarının geçerliğinin karşıt gruplara dayalı kanıtlarla desteklendiği bir başka araştırmadır. Biçer (2024) çalışmasında üç grubun dikkat eksikliği envanterinin alt testlerinden aldıkları puanları karşılaştırmıştır. İlk

grupta çocuk psikiyatristleri tarafından dikkat eksikliği tanısı konmuş çocuklar yer almıştır. İkinci grup henüz dikkat eksikliği tanısı konmamış fakat sınıf öğretmenlerinin kendilerinde dikkat eksikliği bulunduğu izlenimini edindiği öğrencilerden oluşmuştur. Üçüncü grup ise dikkat eksikliği ile ilgili herhangi bir tanı almamış öğrencilerden meydana gelmiştir. Araştırma sonucunda beklendiği gibi dikkat eksikliği tanısı konmuş ya da tanı konmamış olmasına rağmen öğretmenlerinde bu yönde bir kanaat oluşmuş öğrenciler, dikkat eksikliği tanısı almamış öğrencilere kıyasla alt testlerin çoğunda daha düşük performans göstermiştir. Araştırmacı bu sonucu geliştirilen test bataryasının geçerliğine kanıt olarak yorumlamıştır.

4.1.4. Cevaplama Süreçlerine Dayalı Kanıtlar

Ölçme aracındaki maddelerin ölçmeyi hedeflediği yapı ile bireyin bu maddeleri cevaplarken kullandığı bilişsel süreçler arasındaki uyumdan hareketle incelenen bir geçerlik kanıtıdır (Engelhard & Winel, 2018). Birey cevaplama sırasında maddeyi yanıtlayabilmek için işe koşulması gerektiği öngörülen bilişsel süreçleri kullanıyorsa bu durum geçerliğe kanıt sağlar. Öte yandan madde, cevaplanabilmesi için gerekli olduğu düşünülen zihinsel süreçlere başvurulmadan, daha başka yollar kullanılarak da doğru yanıtlanabiliyorsa geçerliğin sağlanmadığı anlaşılır. Dolayısıyla cevaplama süreçlerini incelemeden maddeyi doğru yanıtlayan öğrencilerin cevaba ulaşabilmek için gereken bilişsel süreçleri kullandığını düşünmek bir varsayımın ötesine geçemez (Cizek, 2020).

Matematiksel akıl yürütme becerisini ölçmeye yönelik bir testi öğrencilere uyguladığımızı farz edelim. Öğrencilerin cevaplarını incelediğimizde akıl yürütme süreçlerinin işletildiğini görürsek testin amacına hizmet ettiğini, test puanlarının öğrencilerin akıl yürütme becerisini yansıttığını söyleyebiliriz. Ancak maddelerin ezber bilgiler, formüller ve bilindik algoritmalar takip edilerek cevaplandığını saptarsak test puanlarını akıl yürütme becerisinin göstergesi olarak yorumlayamayız ve bir geçerlik sorunu olduğu kanaatine varırız (Banks, 2012). Başka bir örnek üzerinden konuyu biraz daha somutlaştıralım. Çoktan seçmeli bir matematik testinde "4x – 3 = 2x + 17 ise x kaçtır?" maddesine yer vermiş olalım. Bu madde "Birinci dereceden bir bilinmeyenli denklemleri çözer" hedefini ölçmeye yöneliktir. Gelen cevapları incelediğimizde çoğu öğrencinin maddeyi denklem çözme adımlarını takip ederek değil seçeneklerdeki değerleri denklemde yerine koyarak yanıtladığını görürsek geçerliğin düşük olduğu çıkarımına varırız. Bu da ilgili maddeyi açık uçlu formatta ya da cevabı gizlenmiş çoktan seçmeli madde biçiminde (soru kökünde "*x kaçtır?*" yerine "*x'in rakamları toplamı kaçtır?*" demek gibi) sunduğumuzda geçerliği daha yüksek ölçümler elde edebileceğimizi gösterir.

Cevaplama süreçlerine dayalı geçerlik kanıtlarını toplamak için öğrencilerin cevaplarını oluştururken kullandıkları stratejileri açık biçimde görebilmek gerekir. Bu kapsamda bireylerden maddeleri yanıtlarken sesli düşünmeleri istenebilir (protokol analizi), cevaplayıcılar ile retrospektif görüşmeler gerçekleştirilebilir (Messick, 1989), göz izleme yöntemi kullanılabilir ve maddelerin cevaplanma süresi kontrol edilebilir (Bandalos,

2018). Ek olarak bireylerin sınav kâğıdı üzerinde yaptıkları işlemler ile tuttukları notlar analiz edilebilir ve açık uçlu maddelerden meydana gelen bir sınav uygulandıysa puanlamayı yapan kişilerden de görüş alınabilir (İlhan ve ark., 2024).

4.1.5. Ölçme Sonuçlarına Dayalı Kanıtlar

Testler çeşitli faydalar sağlayacağı düşüncesiyle geliştirilir ve bireylere uygulanır. Örneğin, bir işe alınacak kişilere karar vermek amacıyla yapılan bir sınavın kalifiye elemanların seçilmesini sağlayacağı ve vasıfsız kişilerin işe girmesini engelleyeceği varsayılır (Reynolds ve ark., 2021). Ancak uygulamada durum her zaman böyle olmaz, süreç yeterliliği düşük kişilerin işe alımıyla da sonuçlanabilir. Daha açık bir anlatımla, ölçme işlemleri istendik sonuçlar ortaya çıkarabileceği gibi beklenmedik bir takım çıktılar da doğurabilir. Bunlar geçerlikle ilgili önemli geri bildirimler sağlar (Miller ve ark., 2011). Ölçme sonuçlarına dayalı geçerlik kanıtları bu geri bildirimleri analiz edip ölçme işleminin kendisinden beklenen faydaları ne ölçüde sağladığını ve istenmeyen sonuçlara yol açıp açmadığını saptamaya çalışır (Sandilands & Zumbo, 2014). Geçerlik için ölçme sonuçlarına dayalı kanıtlar toplanırken aşağıdaki sorular üzerinden ilerlenir (Goodwin, 2002, akt: McDonald, 2018):

- Testten beklenen faydalar ne ölçüde gerçekleşiyor?
- Test sonucunda beklenmedik çıktılar (pozitif veya negatif) ne ölçüde meydana geliyor?
- Testi alan farklı alt gruplar için ne düzeyde farklı sonuçlar gözleniyor?

İlhan ve arkadaşları (2024) sıralanan soruları referans alarak ölçme sonuçlarına dayalı geçerlik kanıtlarını tıp alanından bir örnekle açıklamıştır. Buna göre nasıl ki bir aşının etkili olduğunu söyleyebilmek için hastalığı büyük oranda engellemesi ve yan etkilerinin olumlu etkilerinden ağır basmaması gerekiyorsa bir ölçme aracının geçerli ölçümler ürettiğini iddia edebilmek için kendisinden beklenen faydaları sağlaması, sebep olduğu herhangi bir olumsuz sonucun bulunmaması ya da olumsuz sonuçlarının minimumda kalması gerekir. Ayrıca aşının ar-ge süreçlerinde farklı gruplar (bebekler, hamileler, emziren kadınlar vb.) üzerindeki etkisini tanımlamak gerektiği gibi bir test uygulandığında da bu testin sınava giren farklı gruplarda ne kadar benzer çalıştığı sorgulanmalıdır.

Ölçme sonuçlarına dayalı geçerlik kanıtları tüm sınavlara uygulanabilir. Bununla birlikte özellikle seçme, yerleştirme, terfi edecekleri belirleme gibi amaçlarla yapılan sınavlarda bu geçerlik kanıtıyla ilgili incelemelerde bulunmak daha önemlidir (Reynolds ve ark., 2021). Çünkü ölçme sonuçlarına dayalı geçerlik kanıtları, sınavın geçerliği ile mevcut sistemin geçerliği arasında bir köprü vazifesi görür (Sireci, 2016).

Buraya kadar ele alınan diğer geçerlik kanıtları tüm ölçme uzmanları tarafından kabul görmektedir. Eğitsel ve psikolojik test standartlarına 1999 yılında giren bu geçerlik kanıtı (Gliner ve ark., 2017) hakkında ise psikometrisyenler farklı görüşler ileri sürmüştür. Uzmanların bir kısmına göre kapsama, içyapıya ve ölçme aracının diğer değişkenlerle ilişkisine dayanan geleneksel kanıtlar sağlanıyorsa geçerlik hakkında yeterince güçlü argümanlar mevcuttur ve daha farklı kanıtlar incelemeye gerek yoktur. Kimi uzmanlar ise buna karşı çıkmakta, diğer geçerlik kanıtları sağlansa bile bazı geçerlik problemlerinin hala ihtimal dâhilinde olabileceğini belirtmekte, ölçme sonuçlarının birey ve toplum üzerindeki etkileri ile sosyal ve politik sonuçlarına da bakmak gerektiğini savunmaktadır (Lovett, 2023).

Gerçekten de geçerlik hakkında değerlendirme yaparken ölçme sonuçlarına dayalı kanıtları dikkate almak son derece elzemdir. Şöyle bir örnek düşünelim. Öğretimi iyileştirmek, eksik öğrenmeleri gidermek ve hatalı öğrenmeleri düzeltmek amacıyla biçimlendirici değerlendirmeler yaptık, bu kapsamda öğrencilere belli aralıklarla konu ve tarama testleri uyguladık diyelim. Bu testlerden elde edilen sonuçların öğrenmeyi desteklemesi ve öğretimde iyileşme sağlaması biçimlendirici değerlendirme amacına hizmet eder ve geçerliğe kanıt oluşturur. Fakat yaptığımız testler öğretimi iyileştirmekten çok öğrencilerde strese neden oluyorsa ne kadar geçerli olduklarını sorgulamamız gerekir. Çünkü bu testler ölçmeye konu içeriği iyi temsil etse (kapsama dayalı yanıtlar), istatistiksel özellikleri bakımından gayet iyi olsa (içyapıya dayalı kanıtlar) ve ilişkili çıkması beklenen dışsal değişkenlerle anlamlı korelasyonlar gösterse (diğer değişkenlerle ilişkiye dayalı kanıtlar) dahi amacımıza hizmet etmemiştir. Bu da ortada bir geçerlik sorunu var demektir.



Cizek ve arkadaşları (2010) bilimsel kaynaklarda ölçme sonuçlarına dayalı geçerlik kanıtlarının ne sıklıkta sunulduğunu belirlemek amacıyla 2408 makale ile farklı kurumlar (AERA, APA ve NCME) tarafından yayınlanan 326 raporu incelemiştir. Çalışmanın sonuçlarına göre ne makalelerde ne de kuramsal raporlarda bu geçerlik kanıtına yer verilmiştir. Cizek ve arkadaşları (2010) bunu, ölçme sonuçlarına dayalı geçerlik kanıtlarını toplamanın zor olması ve araştırmacıların daha kolay elde edilebilecekleri geçerlik kanıtlarına yönelmesiyle açıklamıştır. Elbette tek neden bu değildir. Bir ölçme işleminin ya da testin toplumsal etkileri ile sosyal ve politik sonuçlarının görünür hale gelmesinin genellikle uzun sürmesi ve test geliştiricilerin bu kadar uzun süre bekleyecek vakitlerinin bulunmaması, test raporlarında ölçme sonuçlarına dayalı geçerlik kanıtlarından bahsedilmemesine yol açan bir başka faktördür (Lovett, 2023). Diğer kanıtlara kıyasla ölçme sonuçlarına dayalı geçerlik kanıtlarının nasıl toplanabileceğine dair kılavuzların sınırlı oluşu da (McIntire & Miller, 2007) bu geçerlik kanıtının göz ardı edilmesine neden olabilmektedir.

4.2. Geçerliği Artırmak için Alınabilecek Önlemler

Güvenirlik geçerliğin ön koşulu olduğundan güvenirliği destekleyen uygulamaların tamamı geçerliği de artırır. Ancak geçerlik, güvenirliğin aksine sabit ve sistematik hatadan da etkilendiğinden güvenirliği artıran uygulamalar arasında doğrudan yer almayan bazı düzenlemeler, ölçme aracının hedeflenen yapıyı daha doğru temsil etmesini sağlayarak geçerliği destekleyebilir. Aşağıda güvenirliği artırmaya yönelik önlemler arasında listelenmeyen ama geçerliği güçlendirecek uygulamalar maddeler halinde sıralanmıştır.

- ⇨ Hedef yapıyı ölçmeye hizmet edebilecek ölçme araçları kullanılmalıdır. Söz gelimi yazma becerisi çoktan seçmeli bir test ile yoklanırsa ulaşılan sonuçların geçerliği düşük olur.
- ⇨ Madde sayısı ve beraberinde sınav süresi ölçülecek kapsam dikkate alınarak belirlenmelidir. Mesela bir dersin vize sınavında 20-25 hatta bazen daha az sayıda madde ile geçerli sonuçlara ulaşılabilir. Ancak yükseköğretim kurumlarına yerleştirilecek öğrencileri belirlemek amacıyla yapılan bir sınavda oldukça geniş bir kapsam yoklandığından öğrenciler hakkında doğru kararlar alabilmek için sınavda daha fazla sayıda maddeye yer vermek gerekir.
- ⇨ Ölçme aracında herhangi bir grubun lehine çalışabilecek yanlı maddeler bulunmadığından emin olunmalıdır. Bir matematik sınavında olasılık konusuyla ilgili "*Bir satranç takımındaki taşlardan rastgele bir tane seçilmiştir. Bu taşın piyon olma olasılığı nedir?*" maddesine yer verildiğini varsayalım. Bu soru satrançta toplam kaç taş bulunduğunu, kaç tane piyon olduğunu bilmeyen öğrencileri dezavantajlı duruma getirir ve geçerliği düşürür. Dolayısıyla aynı yetenek düzeyindeki farklı alt gruplardan birine avantaj sağlayabilecek maddelerden uzak durulmalıdır.
- ⇨ Ölçülmesi amaçlanan özellik dışındaki faktörler puanlama işlemine karıştırılmamalıdır. Örneğin, açık uçlu maddelerden oluşan bir tarih sınavında yazı güzelliği/kâğıt düzeni gibi faktörler puanlamayı etkilememelidir.
- ⇨ Değerlendirmede kullanılan ölçüt doğru belirlenmelidir. Örneğin, bir yeterlik sınavında bağıl ölçüt kullanılması geçerliği düşük çıkarımlar ile sonuçlanır. Yine ölçme aracının diğer değişkenlerle ilişkisinin araştırıldığı geçerlik çalışmalarında ölçüt değişken doğru belirlenmiş ve yeterli güvenirlikte ölçülmüş olmalıdır. Ayrıca ölçülen özellikte ya da ölçüt değişkende ranj daralması sorunu bulunmamalı, diğer bir deyişle verilerin toplandığı örneklem hem ölçmeye konu özellik hem de ölçüt değişken açısından heterojen bir yapı göstermelidir.
- ⇨ Bilgisayar ortamında yapılan sınavlarda adayların bilişim teknolojilerine ilişkin okuryazarlıklarının test puanlarını etkilemeyeceği bir sistem oluşturulmalıdır. Örneğin, bir matematik sınavı çevrimiçi yapılıyor ve öğrencilerin matematiksel ifadelerin girildiği ara yüzü kullanma becerileri test puanlarını etkiliyorsa bu durum geçerliği düşürür (Wu ve arkadaşları, 2016). Bu anlamda çevrimiçi sınavlarda kullanımı özel bilgi ya da yeterlikler gerektirmeyen ara yüzler tercih edilmelidir.

5. KULLANIŞLILIK

Kullanışlılık ölçme aracının hazırlanması, çoğaltılması, uygulanması, puanlanması ve elde edilen puanların yorumlanması bakımından pratik ve ekonomik olması demektir (Tekin, 2009). Kullanışlılık üzerinde etkili olan bir dizi değişken bulunduğundan bir ölçme aracı bir yönüyle kullanışlıyken bir başka yönüyle kullanışlı olmayabilir. Bunu çoktan seçmeli testler ve yazılı sınavlar üzerinden örneklendirelim. Çoktan seçmeli maddelerden oluşan bir test geliştirirken çok sayıda madde yazmak, her bir madde için madde kökünü oluşturup seçeneklere karar vermek ve çeldiricileri (doğru cevap dışındaki seçenekleri) ölçülen özelliğe düşük düzeyde sahip olan kişileri kendisine çekebilecek şekilde yapılandırmak gerekir. Bu işlemler uzun zaman almakta ve çoktan seçmeli soru yazma deneyimi az olan madde yazarlarını zorlayabilmektedir. Yazılı sınavlarda ise daha az sayıda maddeye yer verilmekte ve bu maddeler için çeldirici yazma gibi bir iş yükü bulunmamaktadır. Tüm bunlar hazırlanışı itibariyle yazılı sınavları çoktan seçmeli testlere kıyasla daha kullanışlı bir araç yapmaktadır. Bu iki ölçme aracı puanlanışı açısından karşılaştırıldığında ise tersi bir durum söz konusudur. Çoktan seçmeli testler optik okuyucular tarafından otomatik olarak oldukça kısa sürede puanlanabilmektedir. Diğer taraftan yazılı sınavlarda çoktan seçmeli testlerdeki kadar gelişmiş bir otomatik puanlama sistemi bulunmamakta, puanlama işlemi daha uzun sürmekte ve eğer puanlayıcılara ücret ödeniyorsa maliyet de yükselmektedir. Bu sebeple çoktan seçmeli maddeler puanlanışı açısından yazılı sınavlara göre çok daha kullanışlıdır.

Uygulama koşulları kullanışlılık üzerinde etkili olan bir başka önemli unsurdur. Ölçme aracının aynı anda çok sayıda kişiye hızlıca uygulanabilmesi, uygulama süreçlerinin büyük ölçüde standart olması ve uzmanlık gerektirmemesi kullanışlılığını pozitif yönde etkiler. Öte yandan cevaplanmasının uzun sürmesi, bir seferde tek bir kişiye uygulanabilmesi ve uzman uygulayıcılara ihtiyaç duyulması ölçme aracının kullanışlılığını düşürür. Bu anlamda, grup testleri bireysel testlere kıyasla daha kullanışlıdır. Ayrıca uygulama süresi dikkate alındığında özellikle kalabalık gruplarda, çoktan seçmeli testler yazılı sınavlara, yazılı sınavlar da sözlü sınavlara göre daha kullanışlı olmaktadır.

Gerektirdiği bütçe de ölçme aracının kullanışlılığını etkilemektedir. Örneğin; üç sayfalık bir sınav kâğıdı yazı puntosu küçülterek, satır aralığı azaltılarak, şekil/grafik/tablo boyutları küçültülerek veya başka işlemlerle okunurluğu zedelenmeden önlü arkalı basılabilecek şekilde iki sayfaya sığdırılabiliyorsa basım/fotokopi maliyeti azalacak, bu durum ölçme aracını daha kullanışlı hale getirecektir. Yine kimi sınavlar elektronik ortamda uygulandığında maliyet düşebilmekte, dolayısıyla bazen elektronik uygulama basılı sınava kıyasla daha kullanışlı bir seçenek olabilmektedir. Ölçme aracında renkli basılması gereken maddeler/görseller yoksa gri tonlamalı basımın tercih edilmesi de maliyeti düşürüp kullanışlılığı artırılabilmektedir.

Madde sayısı kullanışlılık üzerinde kilit bir role sahip değişkenlerden biridir. Madde sayısını azaltmak uygulama süresini ve basım maliyetini aşağı çekerek kullanışlılığı artırır. Ancak kullanışlılık ile ilgili diğer tüm işlemler gibi madde sayısında da temel ilke, ölçme aracından çıkarılan maddelerin geçerlikte ve güvenirlikte kayba yol açmamasıdır. Evet, kullanışlılık ölçme araç ve sonuçlarının sahip olması gereken bir niteliktir ama geçerlik ve güvenirliğin önüne konulacak bir özellik değildir. Öncelikli hedef geçerli ve güvenilir ölçümler sağlayacak bir araca ulaşmak olmalı; kullanışlılık, geçerlik ile güvenirlikten taviz verilmeden artırılmaya çalışılmalıdır. İç tutarlılık katsayısı 0,92 olan ve 25 madde içeren bir ölçme aracı düşünelim. Bu ölçme aracındaki madde sayısını 20'ye düşürdüğümüzde hedef yapıyı yine tüm yönleriyle yoklayabiliyor ve iç tutarlılık güvenirliğinde belirgin bir düşüş gözlemiyorsak kullanışlılığı artırmak amacıyla madde sayısını azaltabiliriz. Fakat çıkardığımız maddelerin ardından bazı hedef davranışlar ölçme aracında temsil edilmiyorsa ve/veya güvenirlikte ciddi bir kayıp yaşanıyorsa madde sayısını azaltmamız yanlış bir karar olur. Kısacası işin odağında geçerli ve güvenilir sonuçlara ulaşmak bulunmalı, geçerlik ve güvenirlik üzerinde negatif bir etki yaratmadığı sürece daha az emek ve maliyet gerektiren ölçme araçları tercih edilmelidir (Akbaş, 2022).

Son olarak kullanışlılıkla ilgili özelliklerin teknolojik yeniliklerden etkilenmeye açık olduğuna değinmekte fayda vardır. Örneklendirmek gerekirse optik okuyucular geliştirilmeden önce çoktan seçmeli testler puanlanışı yönüyle bugünkü kadar kullanışlı bir araç değildi. Günümüzde, yapay zekâ teknolojileriyle birlikte açık uçlu maddelerin otomatik puanlanmasına yönelik önemli gelişmeler olduğu (bkz. Aydın ve ark., 2025; Demir, 2023; Sayın & Melanlıoğlu, 2024; Tate ve ark., 2024; Yiğiter & Boduroğlu, 2025) dikkate alındığında yakın tarihte açık uçlu maddelerden oluşan ölçme araçlarının puanlanışı itibariyle daha kullanışlı hale gelmesi muhtemel görünmektedir. Benzer şekilde yapay zekâ araçlarının madde üretme ve çeldirici yazma konusunda hızla gelişme kaydettiği (bkz. Barış Pekmezci & Şengül Avşar, 2023; Kıyak ve ark., 2024; Lin & Chen, 2024) göz önünde bulundurulduğunda teknolojideki ilerlemelerin hazırlanışı boyutuyla da ölçme araçlarının kullanışlılığını etkilemesi beklenmektedir.

6. SONUÇ

Ölçme araçları ile sonuçlarının niteliği hakkında değerlendirme yaparken güvenirlik, geçerlik ve kullanışlılık bütüncül olarak ele alınmalıdır. Güvenirlik ölçümlerin tutarlılığını ve hatalardan arınıklığını ifade ederken geçerlik ölçme sonuçları dikkate alınarak bireyler, kurumlar, öğretim programları ve/veya eğitim sistemleri hakkında verilen kararların ne kadar doğru olduğunu yansıtır. Kullanışlılık ise ölçme araçlarının ve sonuçlarının sahip olması gereken bu özelliklerin uygulanabilir bir zemine taşınmasıyla ilgilidir. Bu anlamda psikometrik açıdan güçlü ölçme sonuçları ile uygulamada tercih edilen ölçme araçlarına ulaşmanın yolu güvenirlik, geçerlik ve kullanışlılığın dengeli biçimde bir araya gelmesinden geçer. Böylece hem bilimsel araştırmalarda hem de uygulama ortamlarında sağlam temellere dayanan, daha doğru ve savunulabilir kararlar ile neticelenen ölçme ve değerlendirme süreçleri yürütülebilir.

KAYNAKÇA | REFERENCES

Aiken, L. R. (1996). *Assessment of intellectual functioning* (2. baskı). Springer.

Akbaş, U. (2022). Ölçme araç ve sonuçlarında bulunması gereken nitelikler. B. Çetin (Ed.), *Eğitimde ölçme ve değerlendirme* (2. baskı, s. 45–84) içinde. Anı.

American Educational Research Association, American Psychological Association & National Council on Measurement in Education (Ed.). (2014). *Standards for educational and psychological testing*. American Educational Research Association.

Anderson, L.W. & Krathwohl, D.R. (2001). *A taxonomy for learning teaching and assessing: A revision of Bloom's taxonomy of educational objectives*. Longman.

Arı, A. (2013). Bilişsel alan sınıflamasında yenilenmiş Bloom, SOLO, Fink, Dettmer taksonomileri ve uluslararası alanda tanınma durumları. *Uşak Üniversitesi Sosyal Bilimler Dergisi*, *6*(2), 259–290. https://doi.org/10.12780/UUSBD164

Aybek, E. C. (2021). Ölçme araçlarının temel özellikleri: Güvenirlik. M. D. Şahin (Ed.), *Eğitimde ölçme ve değerlendirme* (s. 75–98) içinde. Nobel.

Aydın, B., Kışla, T., Tan Elmas, N. & Bulut, O. (2025). Automated scoring in the era of artificial intelligence: An empirical study with Turkish essays. *System*, *133*, 103784. https://doi.org/10.1016/j.system.2025.103784

Ayre, C. & Scally, A. J. (2014). Critical values for Lawshe's content validity ratio: Revisiting the original methods of calculation. *Measurement and Evaluation in Counseling and Development*, *47*(1), 79–86. https://doi.org/10.1177/0748175613513808

Bandalos, D. L. (2018). *Measurement theory and applications for the social sciences*. The Guilford.

Banks, J. A. (Ed.) (2012). *Encyclopedia of diversity in education (Vol. 4).* Sage. https://doi.org/10.4135/9781452218533

Banks, S. R. (2012). *Classroom assessment: Issues and practices* (2. baskı.). Waveland.

Barış Pekmezci, F. & Şengül Avşar, A. (2023, 8–11 Haziran). *Madde yazımında yeni bir yol: Yapay zekâ teknolojisinin kullanımıyla otomatik madde üretimi* [Kongre sunum özeti]. 10. Uluslararası Eğitim Araştırmaları Kongresi, TED Üniversitesi, Ankara, Türkiye.

Başokçu, O. T. (2009). Ölçme sonuçlarının nitelikleri: Ölçme hatası, güvenirlik, geçerlik ve kullanışlılık. N. Doğan (Ed.), *Eğitimde ölçme ve değerlendirme* (s. 31–74) içinde. Pegem Akademi.

Başol, G. (2018). *Eğitimde ölçme ve değerlendirme* (5. baskı). Pegem Akademi.

Baumgartner, T. A., Hensley, L. D., Zhu, W. & Kulinna, P. H. (2021). *Conducting and reading research in kinesiology* (6. baskı). Jones & Bartlett Learning.

Biçer, M. (2024). *Dikkat eksikliği tanılamasında yeni yöntem: Materyal destekli test bataryası geliştirme* (Tez No. 890786). [Yüksek Lisans Tezi, Gaziantep Üniversitesi]. YÖK Ulusal Merkezi.

Brain, C. (2016). *Edexcel psychology for A level book 2*. Hodder Education.

Boztunç Öztürk, N. & Şahin, M. G. (2021). Bilimsel araştırmaya temel oluşturan ölçme kavramları. B. Çetin, M. İlhan & M. G. Şahin (Ed.), *Eğitimde araştırma yöntemleri: Temel kavramlar, ilkeler ve süreçler* (s. 1–34) içinde. Pegem Akademi.

Brown, C. (2023). *The evidence-based practitioner: Applying research to meet client needs* (2. baskı). F. A. Davis Company.

Campbell, D. T. & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, *56*(2), 81–105. https://doi.org/10.1037/h0046016

Cascio, W. F. & Aguinis, H. (2019). *Applied psychology in talent management* (8. baskı). Sage.

Charter, R. A. (1999) Sample size requirements for precise estimates of reliability, generalizability, and validity coefficients. *Journal of Clinical and Experimental Neuropsychology*, *21*(4), 559–566. https://doi.org/10.1076/jcen.21.4.559.889

Cizek, G. J. (2020). *Validity: An integrated approach to test score meaning and use*. Routledge.

Cizek, G. J., Bowen, D. & Church, K. (2010). Sources of validity evidence for educational and psychological tests: A follow-up study. *Educational and Psychological Measurement*, *70*(5), 732–743. https://doi.org/10.1177/0013164410379323

Coaley, K. (2010). *An introduction to psychological assessment and psychometrics*. Sage.

Crocker, L. & Algina, J. (2008). *Introduction to classical and modern test theory*. Cengage Learning.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*, 297–334. https://doi.org/10.1007/BF02310555

Cumming, G. & Fidler, F. (2015). Confidence intervals better answers to better questions. *Journal of Psychology*, *217*(1), 15–26. https://doi.org/10.1027/0044-3409.217.1.15

Demir, S. (2023). Investigation of ChatGPT and real raters in scoring open-ended items in terms of inter-rater reliability. *International Journal of Turkish Educational Sciences*, *11*(21), 1072–1099. https://doi.org/10.46778/goputeb.1345752

Dimitrov, D. M. (2012). *Statistical methods for validation of assessment scale data in counseling and related fields*. American Counseling Association.

Domino, G. & Domino M. L. (2006). *Psychological testing: An introduction* (2. baskı) Cambridge University.

Dougherty, J. (2019). Collecting evidence. N. A. Schmidt & J. M. Brown (Ed.), *Evidence-based practice for nurses: Appraisal and application of research* (4. Baskı, s. 253–283) içinde. Jones & Bartlett Learning.

Earnest, J. & Gupta, S. K. (2022). *Outcome-based curriculum in engineering education*. Phi.

Engelhard, G., Jr. & Wind, S. A. (2018). *Invariant measurement with raters and rating scales: Rasch models for rater-mediated assessments*. Routledge/Taylor & Francis Group.

Evans, C. (2021, April 9). *Chris (Evans) R SAFAQ: Spearman-Brown formula*. https://www.psyctc.org/R_blog/posts/2021-04-09-spearman-brown-formula/

Fidler, F. & Cumming, G. (2019). Effect sizes and confidence intervals. G. R. Hancock, L. M. Stapleton & R. O. Mueller (Ed.), *The reviewer's guide to quantitative methods in the social sciences* (2. Baskı, s. 72–85) içinde. Routledge.

Finch, W. H. & French, B. H. (2019). *Educational and psychological measurement*. Routledge.

Fraenkel, J. R., Wallen, N. E. & Hyun, H. H. (2012). *How to design and evaluate research in education* (8. baskı.). McGraw-Hill.

Frisbie, D. A. (2005). Measurement 101: Some fundamentals revisited. *Educational Measurement: Issues and Practice*, *24*(3), 21–28. https://doi.org/10.1111/j.1745-3992.2005.00016.x

Furr, R. M. & Bacharach, V. R. (2008). *Psychometrics: An introduction*. Sage.

Gezer, M. & İlhan, M. (2018, 18–22 Nisan). *İklim değişikliği bilgi testinin Türkçeye uyarlanması: Geçerlik ve güvenirlik çalışması*. 27. Uluslararası Eğitim Bilimleri Kongresi [Sözlü sunum], Antalya, Türkiye.

Gliner, J. A., Morgan, G. A. & Leech, N. C. (2017). *Research methods in applied settings: An integrated approach to design and analysis* (3. baskı.). Routledge.

Göçer Şahin, S. (2021). Ölçme araçlarının temel özellikleri: Geçerlik ve kullanışlılık. M. D. Şahin (Ed.), *Eğitimde ölçme ve değerlendirme* içinde (s. 99–121). Nobel.

Hayes, A. F. & Coutts, J. J. (2020). Use omega rather than Cronbach's alpha for estimating reliability. but…. *Communication Methods and Measures*, *14*(1), 1–24. https://doi.org/10.1080/19312458.2020.1718629

Hogan, T. P. (2019). *Psychological testing: A practical introduction* (4. Baskı). Wiley.

Holmboe, E. S., Durning, S. J. & Hawkins, R. E. (2018). *Practical guide to the evaluation of clinical competence* (2. baskı). Elsevier.

İlhan, M. & Çetin, B. (2021). Nicel verilerin toplanması ve analizi. B. Çetin, M. İlhan & M. G. Şahin (Eds.), *Eğitimde araştırma yöntemleri: Temel kavramlar, ilkeler ve süreçler* (s. 161–200) içinde. Pegem Akademi.

İlhan, M., Güler, N. & Taşdelen Teker, G. (2024). *Puanlayıcı yargılarına dayalı ölçmelerde geçerlik ve güvenirlik*. Pegem Akademi.

Johnson, R. B. & Christensen, L. (2017). *Educational research: Quantitative, qualitative, and mixed approaches* (6. baskı). Sage.

Kalkbrenner, M. T. (2023). Alpha, omega, and H internal consistency reliability estimates: Reviewing these options and when to use them. *Counseling Outcome Research and Evaluation*, *14*(1), 77–88. https://doi.org/10.1080/21501378.2021.1940118

Kaplan, R. M. & Saccuzzo, D. P. (2018). *Psychological testing: Principles, applications & issues* (9. baskı). Cengage Learning.

Kennedy, I. (2022). Sample size determination in test-retest and Cronbach alpha reliability estimates. British *Journal of Contemporary Education*, *2*(1), 17–29. https://doi.org/10.52589/BJCE-FY266HK9

Kettler, R. J. (2019). *Research methodologies of school psychology: Critical skills (Foundations of school psychology research and practice).* Routledge.

Kıyak, Y. S., Coşkun, Ö., Budakoğlu, I. İ. & Uluoğlu, C. (2924). ChatGPT for generating multiple-choice questions: Evidence on the use of artificial intelligence in automatic item generation for a rational pharmacotherapy exam. *European Journal of Clinical Pharmacology*, *80*, 729–735. https://doi.org/10.1007/s00228-024-03649-x

Kline, T. J. B. (2005). *Psychological testing: A practical aproach to design and evaluation*. Sage.

Koçak, D., Gülleroğlu, H. D. & Çokluk Bökeoğlu, Ö. (2022). *Ölçme aracı uyarlama el kitabı*. Pegem Akademi.

Köse, İ. A. (2012). Ölçmede güvenirlik. N. Çıkrıkçı Demirtaşlı (Ed.), *Eğitimde ölçme ve değerlendirme* (s. 91–115) içinde. Edge Akademi.

Krathwohl, D. R. (2002). A revision of Bloom's taxonomy: an overview. *Theory Into Practice*, *41*(4), 212–218. https://doi.org/10.1207/s15430421tip4104_2

Kuder, G. F. & Richardson, M. W. (1937). The theory of the estimation of test reliability. *Psychometrika, 2*, 151–160. https://doi.org/10.1007/BF02288391

Kurpis, S. E. R. & Stafford, M. E. (2006). *Testing and measurement: A user-friendly guide*. Sage.

Lawshe, C. H. (1975). A quantitative approach to content validity. *Personnel Psychology*, *28*(4), 563–575. http://dx.doi.org/10.1111/j.1744-6570.1975.tb01393.x

Lin, Z. & Chen, H. (2024). Investigating the capability of ChatGPT for generating multiple-choice reading comprehension items. *System*, *123*, 103344. https://doi.org/10.1016/j.system.2024.103344

Lord, F. M. & Novick, M. R. (2008). *Statistical theories of mental test scores*. Information Age.

Lovett, B. L. (2023). *Practical psychometrics: A guide for test users*. Guilford.

Mathews, P. (2010). *Sample size calculations: Practical methods for engineers and scientists*. Mathews Malnar & Bailey.

McCrae, R. R. & Costa, P. T. (2003). *Personality in adulthood: A five-factor theory perspective* (2. baskı). The Guilford.

McCrae, R. R. & John, O. P. (1992). An introduction to the five-factor model and its applications. *Journal of Personality*, *60*(2), 175–215. https://doi.org/10.1111/j.1467-6494.1992.tb00970.x

McDonald, M. E. (2018). *The nurse educator's guide to assessing learning outcomes* (4. Baskı). Jones and Bartlett.

McDonald, R. P. (1999). *Test theory: A unified treatment*. Lawrence Erlbaum.

McIntire, S. A. & Miller, L. A. (2007). *Foundations of psychological testing: A practical approach* (2. baskı). Sage.

Mertler, C. A. (2022). *Introduction to educational research*. Sage.

Messick, S. (1989). Validity. R. L. Linn (Ed.), *Educational measurement* (3. baskı, s. 13–103) içinde. American Council on Education and Macmillan.

Miller, L. A., McIntire, S. A. & Lovler, R. L. (2011). *Foundations of psychological testing: A practical approach* (3. baskı). Sage.

Mitchell, J. E., Crosby, R. D., Wonderlich, S. & Adson, D. E. (2000). *Elements of clinical research in psychiatry*. American Psychiatric

Morrow Jr, J. R., Mood, D. P., Zhu, W. & Kang, M. (2023). *Measurement and evaluation in human performance* (6. baskı.). Human kinetics.

Oerman, M. H. & Gaberson, K. B. (2021). *Evaluation and testing in nursing education* (6. baskı). Springer.

Özer Özkan, Y. (2022). Ölçek geliştirme sürecinde ön inceleme maddeleri gözden geçirme ve düzeltme. M. Acar Güvendir & Y. Özer Özkan (Ed.), *Tüm yönleriyle ölçek geliştirme süreci* (s. 27–36) içinde. Pegem Akademi.

Popham, W. J. (2000). *Modern educational measurement: Practical guidelines for educational leaders* (6. baskı.). Allyn & Bacon.

Popham, W. J. (2025). *Classroom assessment: What teachers need to know* (10. Baskı). Pearson.

Price, L. R. (2017). *Psychometric methods: Theory into practice*. The Guilford.

Reynolds, C. R., Altmann, R. A. & Allen, D. N. (2021). *Mastering modern psychological testing: Theory and methods*. Springer. https://doi.org/10.1007/978-3-030-59455-8

Rowe, D. A. & Mahar, M. T. (2006). Validity. T. M. Wood & W. Zhu (Ed.), *Measurement theory and pratice in kinesiology* (s. 9–26) içinde. Human Kinetics.

Salkind, N. J. & Frey, B. B. (2023). *Tests & measurement for people who (think they) hate tests & measurement*. Sage.

Sánchez-Meca, J., Marín-Martínez, F., López-López, J. A., Núñez-Núñez, R. M., Rubio-Aparicio, M., López-García, J. J., López-Pina, J. A., Hidalgo-Montesinos, M. D., Rosa-Alcázar, Á. I., Gómez-Conesa, A., López-Soler, C. & Gómez-García, F. (2021). Improving the reporting quality of reliability generalization meta-analyses: The REGEMA checklist. *Research Synthesis Methods*, 12, 407–570. https://doi.org/10.1002/jrsm.1487

Sandilands, D. & Zumbo, B. D. (2014). (Mis)alignment of medical education validation research with contemporary validity theory: The Mini-CEX as an example. Zumbo, B. & Chan, E. (Ed), *Validity and validation in social, behavioral, and health sciences* (s. 289–310) içinde. Springer. https://doi.org/10.1007/978-3-319-07794-9_17

Sayın, A. & Melanlıoğlu, D. (2024). Üniversite öğrencilerinin yazılı üretimlerinin OpenAI GPT ile değerlendirilmesi. *Yükseköğretim Dergisi*, *14*(3), 121–134. https://doi.org/10.53478/yuksekogretim.1418870

Sireci, S. G. (2007). On validity theory and test validation. *Educational Researcher*, *36*(8), 477–481. https://doi.org/10.3102/0013189X07311609

Sireci, S. G. (2016). A theory of action for test validation. H. Jiao & R. W. Lissitz (Eds.), *The next generation of testing: Common core standards, smarter-balanced, PARCC, and the nationwide testing movement* (s. 253–271) içinde. Information Age.

Suen, H. K. & Rzasa, S. E. (2004). Psychometricfoundations of behavioral assessment. S. N. Haynes & E. M. Heiby (Ed.), *Comprehensive handbook of psychological assessment* (s. 37–56) içinde. John Wiley & Sons.

Sullivan, J. L. & Feldman, S. (1979). *Multiple indicators: An introduction*. Sage.

Suter, W. N. (2012). *Introduction to educational research: A critical thinking approach* (2. baskı). Sage.

Swanson, J. L. (2012). Measurement and assessment in counseling psychology. E. M. Althamier & J.-I. C. Hansen (Ed.), *The Oxford handbook of counseling* (s. 208–236) içinde. Oxford University.

Tan, Ş. (2009). Misuses of KR-20 and Cronbach's alpha reliability coefficients. *Education and Science*, *34*(152), 101–112.

Tate, T. P., Steiss, J., Bailey, D., Graham, S., Moon, Y., Ritchie, D., Tseng, W. & Warschauer, M. (2024). Can AI provide useful holistic essay scoring? *Computers and Education: Artificial Intelligence, 7*, 100255. https://doi.org/10.1016/j.caeai.2024.100255

Tekin, H. (2009). *Eğitimde ölçme ve değerlendirme*. Yargı.

Turgut, M. F. (1983). *Eğitimde ölçme ve değerlendirme metotları* (2. baskı). Saydam.

Urbina, S. (2004). *Essentials of psychological testing*. John Wiley & Sons.

Vacha-Haase, T., Kogan, L. R. & Thompson, B. (2000). Sample compositions and variabilities in published studies versus those in test manuals: Validity of score reliability inductions. *Educational and Psychological Measurement*, *60*(4), 509–522. https://doi.org/10.1177/00131640021970682

Violato, C. (2019). *Assessing competence in medicine and other health professions*. CRC.

Warrens, M. J. (2015). On Cronbach's alpha as the mean of all split-half reliabilities. *Quantitative Psychology Research*, *89*, 293–300. https://doi.org/10.1007/978-3-319-07503-7_18

Whitley, Jr, B. E. & Kite, M. E. (2013). *Principle of research in behavioral science* (3. baskı) Routledge. https://doi.org/10.4324/9780203085219

Wilson, F. R., Pan, W. & Schumsky, D. A. (2012). Recalculation of the critical values for Lawshe's content validity ratio. *Measurement and Evaluation in Counseling and Development*, *45*(3), 197–210. https://doi.org/10.1177/0748175612440286

Wu, M., Tam, H. P. & Jen, T. H. (2016). *Educational measurement for applied researchers: Theory into practice*. Springer. https://doi.org/10.1007/978-981-10-3302-5

Yiğiter, M. S. & Boduroğlu, E. (2025). Öğrencilerin el yazısıyla yanıtladığı açık uçlu maddelerin puanlanmasında yapay zekâ performansının incelenmesi. *Eğitim ve Bilim*, *50*(Ek Sayı 1), 1–18, https://doi.org/10.15390/EB.2025.14119

Yurdugül, H. (2005, 28–30 Eylül). *Ölçek geliştirme çalışmalarında kapsam geçerliği için kapsam geçerlik indekslerinin kullanılması* [Sözlü sunum]. XIV. Ulusal Eğitim Bilimleri Kongresi, Denizli, Türkiye.

Yurdugül, H. & Bayrak, (2012). Ölçek geliştirme çalışmalarında kapsam geçerlik ölçüleri. Kapsam geçerlik indeksi ve Kappa istatistiğinin karşılaştırılması. *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi*, *Özel Sayı 2*, 264–271.

YAZAR HAKKINDA | ABOUT AUTHOR

PROF. DR. MUSTAFA İLHAN

Dicle Üniversitesi | mustafailhan21[at]gmail.com

ORCID: 0000-0003-1804-002X

Prof. Dr. Mustafa İlhan, lisans eğitimini 2009 yılında Dicle Üniversitesi, Ziya Gökalp Eğitim Fakültesi, İlköğretim Bölümü, İlköğretim Matematik Öğretmenliği programında tamamlamıştır. 2011 yılında Dicle Üniversitesi, Sosyal Bilimler Enstitüsü, Eğitim Bilimleri Anabilim Dalı'nda yüksek lisans derecesini; 2015 yılında Gaziantep Üniversitesi, Eğitim Bilimleri Enstitüsü, Eğitim Bilimleri Anabilim Dalı'nda doktora derecesini ve 2020 yılında Eğitimde Ölçme ve Değerlendirme alanında doçent unvanını almıştır. 2010 ile 2016 yılları arasında Dicle Üniversitesi, Ziya Gökalp Eğitim Fakültesi, Matematik ve Fen Bilimleri Eğitimi Bölümü, Matematik Eğitimi Anabilim Dalı'nda araştırma görevlisi olarak çalışmıştır, 2016 ile 2025 yıllarında arasında aynı bölümde öğretim üyesi olarak görevini sürdürmüştür. 2025 yılı Aralık ayından bu yana Dicle Üniversitesi, Ziya Gökalp Eğitim Fakültesi, Eğitim Bilimleri Bölümü, Eğitimde Ölçme ve Değerlendirme Anabilim Dalı'nda profesör olarak görev yapmakta ve Eğitimde Ölçme ve Değerlendirme Anabilim Dalı Başkanlığı görevini yürütmektedir.

PROF. DR. MUSTAFA İLHAN

Dicle University | mustafailhan21[at]gmail.com

ORCID: 0000-0003-1804-002X

Prof. Dr. Mustafa İlhan received his bachelor's degree in Elementary Mathematics Education from Dicle University in 2009. He earned his master's degree in Educational Sciences from Dicle University in 2011 and his Ph.D. in Educational Sciences from Gaziantep University in 2015. He was awarded the title of Associate Professor in Educational Measurement and Evaluation in 2020. He served as a research assistant in the Department of Mathematics and Science Education, Division of Mathematics Education, Ziya Gökalp Faculty of Education, Dicle University, from 2010 to 2016. From 2016 to 2025, he served as a faculty member in the same department. Since December 2025, he has been a professor in the Department of Educational Sciences, Division of Educational Measurement and Evaluation, Ziya Gökalp Faculty of Education, Dicle University. He also serves as the Head of the Division of Educational Measurement and Evaluation.

4. BÖLÜM

ÖLÇMEYE KONU OLAN PSİKOLOJİK ÖZELLİKLERİN ÖLÇÜLMESİ

MEASUREMENT OF PSYCHOLOGICAL ATTRIBUTES

Dr. Öğr. Üyesi Ömer KUTLU

Doç. Dr. Özge ALTINTAŞ

ÖLÇMEYE KONU OLAN PSİKOLOJİK ÖZELLİKLERİN ÖLÇÜLMESİ

Dr. Öğr. Üyesi Ömer KUTLU
Ankara Üniversitesi

Doç. Dr. Özge ALTINTAŞ
Ankara Üniversitesi

Öz

Bu bölümde ölçmeye konu olan psikolojik özelliklerin kuramsal temelleri, ölçme ve değerlendirme süreçlerindeki yeri çağdaş psikometrik yaklaşımlar çerçevesinde ele alınmaktadır. Bölümde öncelikle psikolojik özelliklerin doğrudan gözlenemeyen, ancak davranışsal göstergeler aracılığıyla dolaylı olarak ölçülebilen örtük yapılar olarak ele alındığı açıklanmakta; ardından bilişsel, duyuşsal ve psikomotor özelliklerin ölçülmesine ilişkin temel yaklaşımlar tartışılmaktadır. Bilişsel özellikler bağlamında eğitim hedefleri sınıflamaları, bilgi türleri, beceri ve yetkinlik kavramları; duyuşsal özellikler bağlamında tutum, ilgi, güdü, özyeterlik, kişilik ve kaygı; psikomotor özellikler bağlamında ise performansa dayalı gözlem ve beceri değerlendirme süreçleri incelenmektedir. Ayrıca zekâ ve yetenek ölçmelerinin tarihsel gelişimi, okul öğrenmelerinin ölçülmesi, kültürel bağlamın ölçme sonuçlarına etkisi ve dijitalleşme ile yapay zekâ destekli ölçme uygulamalarının sunduğu yeni olanaklar değerlendirilmektedir. Bölüm, psikolojik özelliklerin ölçülmesinde geçerlik, güvenirlik, adalet, etik ve yorumlanabilirlik ilkelerinin önemini vurgulayarak ölçme sonuçlarının bireysel ve toplumsal karar süreçlerinde sorumlu biçimde kullanılmasına ilişkin kavramsal bir çerçeve sunmaktadır.

Anahtar Kelimeler

Psikolojik Özellikler, Psikometri, Bilişsel Özellikler, Duyuşsal Özellikler, Psikomotor Özellikler

MEASUREMENT OF PSYCHOLOGICAL ATTRIBUTES

Dr. Öğr. Üyesi Ömer KUTLU
Ankara University

Doç. Dr. Özge ALTINTAŞ
Ankara University

Abstract

This chapter examines the theoretical foundations of psychological attributes targeted for measurement, their role in educational measurement and assessment, and the ways in which they are conceptualized within contemporary psychometric approaches. It begins by explaining that psychological attributes are conceptualized in psychometrics as latent constructs that cannot be observed directly but can be measured indirectly through behavioral indicators using scores obtained from measurement instruments supported by strong validity and reliability evidence. The chapter then discusses the principal approaches to measuring cognitive, affective, and psychomotor attributes. Within the cognitive domain, it addresses educational taxonomies, types of knowledge, and the concepts of skills and competencies. Within the affective domain, it examines attitudes, interests, motivation, self-efficacy, personality, and anxiety. Within the psychomotor domain, it focuses on performance-based observation and skill assessment. The chapter further reviews the historical development of intelligence and ability assessment, the assessment of school learning, the influence of cultural context on the interpretation of assessment results, and the opportunities created by digitalization and artificial intelligence-supported assessment practices. Finally, it presents a conceptual framework for the responsible use of assessment results in individual and societal decision-making by emphasizing the importance of validity, reliability, fairness, ethics, and interpretability in the measurement of psychological attributes.

Keywords

Psychological Attributes, Psychometrics, Cognitive Attributes, Affective Attributes, Psychomotor Attributes

1. GİRİŞ

İnsan davranışlarının anlaşılması ve açıklanması sosyal bilimlerin temel amaçlarından biridir. Bu amaca ulaşabilmenin ön koşulu ise bireyin doğrudan gözlenemeyen psikolojik özelliklerinin bilimsel yöntemlerle ölçülebilmesidir. Bu noktada testler, bireyin bilişsel, duyuşsal ve psikomotor özelliklerine ilişkin geçerli ve güvenilir bilgiler elde edilmesini sağlayan temel ölçme araçları olarak önemli bir işleve sahiptir. Ölçme ve değerlendirme ile psikometri alanlarında yürütülen bilimsel çalışmalar sayesinde zekâ, başarı, tutum, ilgi, güdü, kişilik, kaygı ve benzeri psikolojik özellikler geçerliği ve güvenirliği yüksek ölçme araçları aracılığıyla sistematik olarak değerlendirilebilmekte; bireylerin eğitimsel, mesleki ve sosyal yaşantılarına ilişkin bilimsel temelli kararların alınmasına katkı sağlanmaktadır.

İnsan doğası hâlâ birçok bilinmeyeni içinde barındırmaktadır. Bu bilinmeyenleri anlama çabası yalnızca insanın bilgi arayışını beslemekle kalmamış, aynı zamanda toplumların eğitim, çalışma ve sosyal yaşamı düzenleme biçimlerini de derinden etkilemiştir. Psikolojik özelliklerin ölçülmesi, aslında insanlığın kendine bir "ayna tutması" ve bu aynayı sürekli inşa etme çabasıdır. Bu çaba bireyin yeteneklerindeki potansiyelin keşfedilmesine ve onun toplu yaşam içinde "birey" haline gelmesine katkıda bulunmaktadır. Bu anlamda, psikolojik ölçmelerin en büyük katkısı bireyin henüz farkına varmadığı potansiyelini gün yüzüne çıkarmasıdır. Bireyin yeteneklerini ve ilgilerini keşfetmesi, kendi başarısını yönetmesine, ilgilerine uygun bir mesleği seçmesine, toplumsal yaşamda nasıl davranması gerektiğine, farklı yeteneklerine odaklanmasına ve hatta kim olduğuna kadar birçok işlevi yerine getirmesini sağlayacaktır. Bu yönüyle testlere dayalı ölçmeler, bireysel farklılıkların rastlantısal değerlendirilmesini önleyen sessiz bir devrim niteliğindedir. Zihni ölçmek, yalnızca sayısal verileri toplamak ve onları çözümlemek işi değildir. Bu, insanın kendi doğasını anlama, hatalarını en aza indirme ve geleceğini daha sağlam temeller üzerine kurma çabasıdır.

Psikolojik özelliklerin ölçülmesine yönelik çalışmalar, antik dönemlerde bireyin davranışlarını sistematik biçimde gözlemlemeye yönelik ilk girişimlerden başlayarak günümüzde ileri psikometrik modellere uzanan uzun bir gelişim sürecinden geçmiştir. Özellikle son yıllarda, büyük veri analitiği, bilgisayar ortamında bireye uyarlanmış testler, bilişsel tanı modelleri ve yapay zekâ destekli ölçme uygulamaları; yalnızca bireyin neyi bildiğini değil, nasıl düşündüğünü, nasıl öğrendiğini ve öğrenme süreçlerinin hangi koşullar altında şekillendiğini anlamaya yönelik yeni olanaklar sunmaktadır. Bu gelişmeler, ölçme ve değerlendirme alanının kapsamını genişletmekte; psikometrik modellerin daha güçlü kuramsal temellere dayanmasına ve bireysel farklılıkların daha duyarlı biçimde değerlendirilmesine katkı sağlamaktadır.

Ölçme ve değerlendirme alanı, bireylerin psikolojik özelliklerini ölçerek bilişsel, içsel, kişilerarası ve devimsel (psikomotor) yönlerini bilimsel yöntemlerle anlamayı amaçlar. 20. yüzyıl boyunca psikolojik özelliklerin ölçülmesi; başarı, zekâ ve yetenek gibi ***bilişsel*** temelli özelliklerin yanı sıra tutum, ilgi, kişilik, kaygı ve güdü gibi ***duyuşsal***, el ya da bedenin kullanımını gerektiren ***psikomotor*** beceriler gibi çok boyutlu alanlara yönelmiştir (Bloom ve ark., 1956; Cronbach, 1990; Domino & Domino, 2006; Schultz & Schultz, 2011).

Özellikle 21. yüzyılın başlarında yapılan araştırmalar insanların psikolojik özellikleri üstünde yalnızca anlama, problem çözme, eleştirel düşünme gibi ***bilişsel*** özelliklerin değil, aynı zamanda özgüven, özyeterlik, güdülenme, yılmazlık, kararlılık, ilgi, tutum gibi ***içsel***; iş birliği yapma, iletişim kurma, sorumluluk alma, ikna etme, üretkenlik gibi ***kişilerarası*** özelliklerinin de etkili olduğunu ortaya koymuştur. Tıp ve biyoloji biliminin beyin üstüne yaptığı araştırmalarla desteklenen bulgular ölçme araçlarının kapsamını genişletmiş ve farklı psikolojik özellikleri belirlemeye dönük yöntemler geliştirilmeye başlanmıştır (Anastasi & Urbina, 1997; Kutlu & Altıntaş, 2021; Liu ve ark., 2023; Marzano & Heflebower, 2012).

Bu gelişmeler ölçme ve değerlendirme (assessment) anlayışının da değişmeye başladığını göstermektedir. Ölçme; bireyleri sınıflandırma ve karşılaştırma amacının ötesine taşımış, yaşamın herhangi bir noktasından edindikleri bilgi ve becerileri etkin kullanacakları ve kendilerini psikolojik özelliklerinden de yararlanarak etkin biçimde yönlendirecekleri bir sürece odaklanmıştır (Collins, 2014; Kutlu & Kula-Kartal, 2018; Kyllonen, 2012; Payne & Kyllonen, 2012). Bu nedenle psikolojik ölçmeler hem bireysel gelişim hem de toplumsal karar süreçlerinde temel bir araç haline gelmiştir.

2. BİLİŞSEL ÖZELLİKLERİN ÖLÇÜLMESİ

Eğitim bilimlerinin en temel uğraşlarından biri, öğrenmenin doğasını açıklamak ve öğretimi bu anlayışa göre biçimlendirmektir. Öğrencilerin derslere ait bilgileri edinmelerinin yanında, bu bilgileri diğer bilgilerle nasıl ilişkilendirdikleri, gerçek yaşam durumlarında nasıl ve ne düzeyde kullandıkları, öğretim sürecinin başarısını doğrudan etkileyen faktörler arasında yer almaktadır. Bu nedenle eğitimciler, öğrenme çıktılarını sistemli biçimde sınıflandırmak ve ölçülebilir duruma getirmek amacıyla çeşitli modeller geliştirmişlerdir. Bu sınıflamalar (taksonomi), özellikle bilişsel alan üzerine odaklanmış ve öğretim tasarımının merkezine yerleşmiştir.

2.1. Bilişsel Alan Sınıflamaları

Bu kısımda, bilişsel özelliklerin sınıflanmasına yönelik geliştirilmiş dört etkili yaklaşım tarihsel akış içinde ele alınmıştır: Bloom ve arkadaşlarının (1956) eğitim hedefleri sınıflaması, Anderson ve Krathwohl'un (2001) Bloom'un güncelleştirilmiş sınıflaması, Haladyna'nın (1997) ölçme temelli sınıflaması ile Marzano ve Kendall'ın (2007)

sınıflama modeli. Her biri farklı kuramsal altyapılara dayansa da ortak amaçları, öğrencilerin düşünme becerilerini anlamak, desteklemek, zenginleştirmek ve değerlendirmektir. Bu bağlamda bilişsel sınıflamalar; başta öğretim yöntemleri olmak üzere ölçme araçlarının yapısını, puanlamayı, geribildirimi ve değerlendirme süreçlerini de etkilemiştir.

Sınıflamaların oluşturulması bazı eğitsel tartışmaların sonunda ortaya çıkmıştır. 20. yüzyılın ortalarına gelindiğinde, özellikle ABD'de yükseköğretim kurumlarında sınavların içeriği ve amacı konusunda görüş ayrılıkları bulunmaktaydı. Örneğin, bir öğretmenin "Öğrencinin anlamasını istiyorum." derken ne demek istediği; başka bir öğretmenin "Öğrencinin yorumlaması gerekir." derken beklentisinin ne olduğu açık değildi. Bu belirsizlik, ortak sınav soruları geliştirme çabalarını da sekteye uğratıyordu. Bu dönemde Benjamin Bloom, öğrenme hedeflerini sınıflandırmak ve bu sınıflandırmayı öğretim süreçleriyle bütünleştirmek için Columbia Üniversitesi'ndeki meslektaşlarıyla çalıştaylar düzenlemiş ve 1956 yılında *Taxonomy of Educational Objectives: Handbook I - Cognitive Domain* adlı eseri yayımlamışlardır. Bloom ve arkadaşları (1956), sınıflama (taksonomi) sözcüğünü biyolojiden alarak eğitsel hedefleri sınıflandırmış ve öğretmenler arasında ortak bir dil oluşturmaya çalışmışlardır. Bunun yanında, eğitim hedeflerinin sınıflandırılmasının; öğretmenleri, program geliştiricileri, ölçme ve değerlendirme uzmanlarını ortak bir noktada buluşturacağını vurgulamışlardır.

2.1.1. Bloom Sınıflaması

Bloom ve ekibi tarafından 1956 yılında yayımlanan *Taxonomy of Educational Objectives: Handbook I - Cognitive Domain* yedi yıl süren bir çalışmanın ürünüdür. Bu eserde, bilişsel alan altı temel düzeyde aşamalı olarak düzenlenmiştir (Bloom ve ark., 1956). Tablo 1'de bilişsel düzeyler ve bu düzeylerin özellikleri verilmiştir. Devamında yer alan açıklamalarda her bir bilişsel düzey ve bu düzeylerin özellikleri madde örnekleriyle[1] desteklenmiştir.

[1] Kitabın bu bölümünde yer alan tüm örnek maddeler, ulusal düzeyde gerçekleştirilen ortaöğretime ve yükseköğretime geçiş testlerinde kullanılmış maddelerden seçilerek uyarlanmıştır. Bu testlerde yer alan maddeler çoktan seçmeli yapıdadır ancak örnek olması açısından kısa yanıtlı ya da açık uçlu madde yapısına dönüştürülmüştür.

Tablo 1

Bloom Sınıflamasının Bilişsel Düzeyleri ve Özellikleri

Bilişsel Düzeyler	Özellikleri
Bilgi (Knowledge)	Terimlerin, olguların, alışıların, araç ve gereçlerin, ilke ve yöntemlerin tanınması ya da hatırlanmasını içerir.
Kavrama (Comprehension)	Çevirme, yorumlama, öteleme olmak üzere üç alt beceriden oluşur. Öğrencinin bir bilgiyi başka anlatıma çevirmesini, yorumlamasını, geçmişte ya da gelecekte ne olacağını söylemesini içerir.
Uygulama (Application)	Daha önce öğrenilen bilgileri, soyutlamaları, genel ilkeleri ya da yöntemleri belirli, yeni ve somut durumlarda kullanmayı içerir.
Analiz (Analysis)	Karmaşık bir bütünü; parçalarına ayırmayı ve parçalar arasındaki ilişkileri anlamayı ve ilişkiler arasında ilkeler ya da kurallar oluşturmayı içerir.
Sentez (Synthesis)	Dağınık parçalardan, karmaşık düşüncelerden yeni, anlamlı ve özgün bir ürün ortaya koymaktır. Süreç ve ürün önemlidir. Özgün ürün; bir iletişim ögesini, bir plan önerisini, soyut bir modeli ya da sistemi içerebilir.
Değerlendirme (Evaluation)	Ürünleri, yöntemleri, düşünceleri ve eserleri, geçerli, nesnel dış ya da iç ölçütlere dayanarak yargılara ulaşmayı içerir. Yargılar; kanıtlarla ve gerekçelerle desteklenmelidir. Birey bazen kendi seçtiği kabul edilebilir ölçütleri de kullanarak yargıya varabilir.

Örnek 4.1. Bilgi Düzeyinin Ölçülmesine Yönelik Kısa Yanıtlı Maddeler

- Ölçme nedir? Tanımını yazınız.
- Osmanlı Devleti'nde toprağını boş bırakan köylüden alınan vergiye ne ad verilir? Yazınız.
- Divan Edebiyatının nazım türlerinden ikisini yazınız.
- İnsan derisindeki bir kesikten mikroorganizmalar girdikten sonra, ilk olarak hangi olay meydana gelir? Yazınız.

Örnek 4.2. Kavrama Düzeyinin Ölçülmesine Yönelik Kısa Yanıtlı Maddeler

- Bu kadar düzenli ve disiplinli çalışmadan sonra, yarışmada birinci olması hiç de sürpriz değildi.

Bu cümledeki *"neden-sonuç"* ilişkisi dikkate alındığında, *"neden"* olan durumu yazınız.

- Özgür insan kitapların ürünüdür. İnsanlar kitapları nasıl yaratıyorsa, kitaplar da insanları öyle yaratır. Onun içindir ki özgürlükten korkan, kitaplardan da korkar. Beğenmedikleri, kişisel nedenlerle hoşlanmadıkları düşünceleri ortadan kaldırmaya çalışırlar.

Bu paragrafın anlamına uygun bir son cümle yazınız.

Örnek 4.3. Uygulama Düzeyinin Ölçülmesine Yönelik Açık Uçlu Maddeler

- Bir öğrenci şişirdiği balonu önce sıcak bir ortamda tutarak genleşmesini izlemekte, daha sonra ise soğuk ortama taşıyarak küçüldüğünü gözlemektedir.

Bu öğrenci yaptığı deneyde, aşağıdaki sorulardan hangisine yanıt vermeye çalışmaktadır? Bu sonuca nasıl ulaştınız yazınız.

Buğday, yetişme döneminde bol su, olgunlaşma döneminde ise çok az su isteyen bir üründür. Buğday ekimi yapılan bir bölgede, nisan ayı az yağışlı, temmuz ayı ise yağışlı geçmiştir.

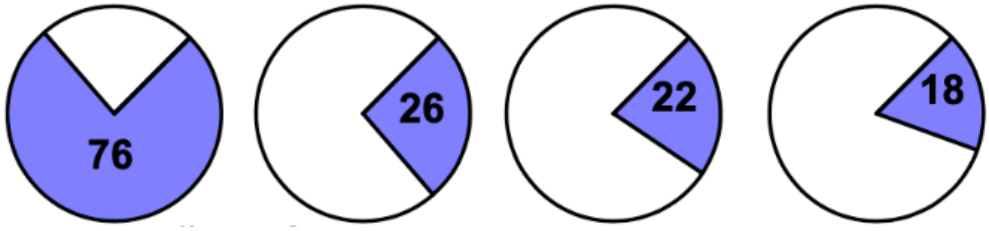
Bu olayda, gözlenecek olası sonuç nedir? Bu sonuca nasıl ulaştınız yazınız.

Örnek 4.4. Analiz Düzeyinin Ölçülmesine Yönelik Açık Uçlu Maddeler

- Bazı insanları memnun etmek pek kolay değildir. (2) Kırk yıl sırtında taşısan ve bir kerecik "yoruldum" desen, kırk yıllık emekler unutulur. (3) Ancak herkes de böyle olmayabilir. (4) Bir fincan kahvenin kırk yıl hatırını sayanlar da vardır. (5) Bu nedenle bir insanın davranışını herkese genellemek doğru değildir. (6) Demek ki, insanlar üç aşağı beş yukarı, benzer anlayıştadır.

Bu paragraftaki hangi cümle, anlam yönünden diğer cümlelerle çelişmektedir? Neden belirterek yazınız.

- Aşağıdaki grafiklerde bir coğrafi bölgenin başlıca tarım ürünlerinin Türkiye üretimindeki payları yüzde olarak verilmiştir:



Buna göre bu coğrafi bölge için;

1. Yağ endüstrisi gelişmiştir.
2. Tarım alanlarının çoğu seracılığa ayrılmıştır.
3. Akdeniz ikliminin etkileri görülür.

yargılarından hangilerine ulaşılır? Ulaşılan durumların her biri için neden belirterek yazınız.

Örnek 4.5. Sentez Düzeyinin Ölçülmesine Yönelik Görev

- Okulunuzda okuyan 5. sınıf kız ve erkek öğrencilere uygulanmak üzere bir görüşme formu hazırlayın. Bu formla onlara öğlen ve akşam öğünlerinde en çok neler yediklerini sorun ve ayrıntılı notlar alın. Beslenme açısından uygun olan ve uygun olmayan durumları raporlayın. Raporda öğretmenlere ve okul yöneticilerine neler yapmaları gerektiğine dair gerekçeli önerilerde bulunun.

Örnek 4.6. Değerlendirme Düzeyinin Ölçülmesine Yönelik Görev

- Göreviniz beğendiğiniz bir şarkıcının söylediği bir şarkı hakkında yazmakla ilgilidir. Bunun için dört farklı değerlendirme ölçütü (örneğin sözler, söz müzik uyumu, solistin ses kalitesi vb.) belirleyiniz. Bu ölçütlere göre, şarkı hakkında bir müzik dergisi için 500-600 sözcük arası görüş yazısı yazınız.

Haladyna (1997), Bloom sınıflamasının özellikle test maddesi yazımında, öğretmenlere yol gösterici bir araç olduğunu, her bilişsel düzeyin farklı türde madde biçimleriyle ilişkilendirilebileceğini ortaya koyduğunu belirtmektedir. Bu fark, öğretmenin yalnızca "öğrenci ne biliyor?" sorusuna değil, "hangi düzeyde biliyor ve ne yapabiliyor?" sorusuna da odaklanmasını sağlamıştır.

Bloom sınıflaması; hedeflerin ölçülebilir ve gözlenebilir şekilde yazılmasını zorlaştırması, "sentez" ve "değerlendirme" düzeylerinin yerinin doğru olmayışı, kavrama ve analiz basamağındaki binişiklikler gibi nedenlerle sorgulanmıştır (Anderson & Krathwohl, 2001). Bu eleştiriler, sınıflamanın güncellenmesi gerekliliğini doğurmuştur.

2.1.2. Anderson ve Krathwohl Sınıflaması (Bloom Sınıflamasının Güncellenmesi)

Bloom ve arkadaşlarının 1956 yılında yayımladığı eğitim hedeflerinin sınıflandırılması çalışması onlarca yıl eğitimciler tarafından benimsenmiştir. Ancak 1990'lı yılların ortalarına gelindiğinde, öğrenme psikolojisindeki gelişmeler ve eğitim uygulamalarındaki değişimler, bu modelin yeniden ele alınmasını zorunlu duruma getirmiştir. Bloom'un öğrencisi Lorin W. Anderson ile 1956 sınıflamasının yazarlarından David R. Krathwohl önderliğinde yeni bir çalışma grubu kurulmuş, bu grup, 1995-2000 yılları arasında gerçekleştirdikleri *A Taxonomy for Learning, Teaching, and Assessing* çalışmalarını 2001 yılında yayımlamışlardır (Anderson & Krathwohl, 2001). İyileştirilmiş (revize edilmiş) model, öğretim hedeflerinin daha açıklayıcı, davranışa dönüştürülebilir ve değerlendirmeye uygun olmasını temel amaç edinmiştir. Krathwohl (2002), yeni sınıflamanın öğretmenlerin uygulamada kullanabileceği bir planlama aracına dönüşmesini istediklerini belirtmiştir.

İyileştirilmiş sınıflama, 1956 sınıflamasının hiyerarşik yapısını korumakla birlikte onu hem iki boyutlu hem de yapılandırmacı öğrenme kuramıyla uyumlu bir duruma getirmiştir. Bu yeni yapı, *bilişsel süreç boyutu* ve *bilgi birikimi boyutu* olmak üzere iki ana eksen üzerine oturtulmuştur (Anderson & Krathwohl, 2001).

Bilişsel süreç boyutu (cognitive process dimension), Bloom ve arkadaşları (1956) tarafından önerilen sınıflamada olduğu gibi altı aşamalı bir yapıya sahiptir. Bilişsel süreç boyutları ve bu boyutların özellikleri Tablo 2'de sunulmuştur.

Tablo 2

Anderson ve Krathwohl Sınıflamasının Bilişsel Süreç Boyutları ve Özellikleri

Bilişsel Süreç Boyutları	**Özellikleri**
Hatırlama (Remember)	Bu boyut uzun süreli bellekte yer alan bilgilere erişilmesiyle ilgilidir. Öğrenciden bilgileri *tanıması* ya da *hatırlaması* istenir.
Anlama (Understand)	Öğrenciden sözlü, yazılı, şema, tablo, grafik ya da resim gibi ögelerden öğretimle ilgili anlamlar oluşturması beklenir. Bu boyutun becerileri şunlardır: *Yorumlama, örneklendirme, sınıflama, özetleme, sonuç çıkarma, karşılaştırma, açıklama*.
Uygulama (Apply)	Öğrenciden işlem yolunu; verilen yeni bir durumda kullanması ya da yeni duruma uygulaması beklenir. Bu boyutun becerileri şunlardır: *Yapma, yararlanma.*
Çözümleme (Analyze)	Öğrenciden materyali anlamlı parçalarına ayırması, parçaların birbiriyle ve materyalin bütünüyle ilişkilerini belirlemesi beklenir. Bu boyutun becerileri şunlardır: *Ayrıştırma, örgütleme, irdeleme.*
Değerlendirme (Evaluate)	Öğrenciden ölçütler ve standartlar kullanarak yargılara ulaşması beklenir. Bu boyutun becerileri şunlardır: *Denetleme, eleştirme.*
Yaratma (Create)	Öğrenciden parçaları bir araya getirerek, yeni, özgün bir ürün oluşturması beklenir. Bu boyutun becerileri şunlardır: *Oluşturma, planlama, üretme.*

Anderson ve Krathwohl (2001) sınıflamasının getirdiği önemli bir yenilik ise, bilgi türlerinin de sınıflandırılmış olmasıdır. Buna göre bilgi türleri; *olgusal bilgi, kavramsal bilgi, işlemsel bilgi ve üstbilişsel* bilgi olmak üzere dört başlıkta ele alınmıştır [Kullanılan Türkçe terimler Anderson ve Krathwohl'un (2001) Prof. Dr. Durmuş Ali Özçelik tarafından çevrilen kitabından alınmıştır]. Bu sınıflandırma, özellikle 1980'lerden itibaren öğrenmenin çok katmanlı bir süreç olduğunu savunan bilişsel psikoloji kuramlarıyla paralellik göstermektedir.

Bilgi birikimi boyutu, Bloom ve arkadaşlarının (1956) sınıflamasında olduğu gibi altı aşamalı bir yapıya sahiptir. Bilgi birikimi boyutları ve bu boyutların özellikleri Tablo 3'te verilmiştir.

Tablo 3

Anderson ve Krathwohl Sınıflamasının Bilgi Birikimi Boyutları ve Özellikleri

Bilgi Birikimi Boyutu	**Özellikleri**
Olgusal Bilgi (Factual Knowledge)	Bir konu alanına ait bilgileri bilen, o alana ait problemleri çözebilen bir öğrencinin bilmesi zorunlu ögeleri içerir. Bu bilgi boyutu; *terimlerin bilgisi, özel ayrıntı ve ögelerin bilgisi* olmak üzere iki kısımdan oluşur.
Kavramsal Bilgi (Conceptual Knowledge)	Geniş bir yapının temel ögeleri arasında bulunan ve bu yapıyı oluşturan ögelerin birlikte hareket etmesini sağlayan ilişkilerin bilgisi olarak tanımlanmaktadır. Bu bilgi boyutu; *sınıflamalar ve sınıfların bilgisi, ilkeler ve genellemelerin bilgisi, kuram, model ve yapıların bilgisi* olmak üzere üç kısımdan oluşur.
İşlemsel Bilgi (Procedural Knowledge)	Bir şeyin nasıl yapılacağı konusundaki bilgileri tanımlanmaktadır. Bu bilgi boyutu; *alana özel beceri ve algoritmaların bilgisi, alana özel teknik ve yöntemlerin bilgisi, uygun yöntemlerin hangi durumlarda kullanılacağının belirlenmesine ilişkin ölçütlerin bilgisi* olmak üzere üç kısımdan oluşur.
Üstbilişsel Bilgi (Metacognitive Knowledge)	Üstbilişsel bilgi, bireyin kendi bilişinin farkında ve bilgi sahibi olması olarak tanımlanmaktadır. Bu bilgi boyutu; *öğrenme ve düşünmeyle ilgili genel stratejiler bilgisi*, *bilişsel görevlerde farklı stratejilerin ne zaman ve nasıl kullanılacağıyla ilgili bilgisi ve bireyin kendisi hakkındaki bilgisi* olmak üzere üç kısımdan oluşur.

Tablo 2 ve 3 birlikte incelendiğinde; bilişsel süreç boyutlarının basitten karmaşığa, bilgi birikimi boyutunun ise somuttan soyuta sıralandığı görülecektir. Anderson ve Krathwohl (2001), öğretmenlerin bu iki tablonun çaprazlanarak oluşturulduğu iki boyutlu belirtke tablosunu kullanarak hem öğrenme çıktılarını hem de ölçme ve değerlendirme etkinliklerini kolaylıkla planlayabileceklerini belirtmektedirler.

Tablo 4'te bilgi birikimi ve bilişsel süreç boyutunun çaprazlanarak oluşturulduğu örnek bir belirtke tablosu sunulmuştur. Okuyucu bu örnek belirtke tablosunu bir dersin bir ya da birkaç ünitesinde yer alan kazanımlara dayanarak hazırlayacağı sınav için düşünebilir.

Tablo 4

. . . . . Dersinin 1. Yazılı Sınavı Belirtke Tablosu

Bilgi Birikimi Boyutu	**Bilişsel Süreç Boyutu**					**Toplam**
	Hatırlama	**Anlama**	**Uygulama**	**Çözümleme**	**Değerlendirme**	
Olgusal Bilgi	1	1	1	1		**4**
Kavramsal Bilgi	2	1				**3**
İşlemsel Bilgi		1	1	1	1	**4**
Üstbilişsel Bilgi		2	1		1	**4**
Toplam	**3**	**5**	**3**	**2**	**2**	**15**

Örnek belirtke tablosunda ünitelerin ders saatine göre ağırlıkları dikkate alınmış ve ünitelere ait soru sayıları buna göre belirlenmiştir. Bu dersin sınavını hazırlayacak bir öğretmen, kısa yanıtlı ve açık uçlu olmak üzere toplamda 15 madde kullanacaktır. Bunun yanında öğretmenler, bu tabloya bakarak hem kazandıracakları bilgi türünü hem de bu bilgiyle öğrenciden hangi bilişsel düzeyde bir işlem beklediklerini belirleyebileceklerdir. Bu yapı, bu özelliğiyle öğretim hedeflerinin formüle edilmesinde hem açıklık sağlamakta hem de ölçülebilirliğini kolaylaştırmaktadır. Anderson ve Krathwohl sınıflaması, kuramsal bir çerçeve sunmaktan öte, öğretmenlerin eğitimde doğrudan kullanabilecekleri bir araç olarak tasarlanmıştır.

Yeni sınıflama uygulayıcılara şu katkıları sunmaktadır:

1. Kazanımların, "Öğrenci şunu bilmelidir." ifadesinden çok, "Öğrenci, X konusundaki kavramsal bilgisini kullanarak Y problemini çözümler" gibi daha işlevsel (işevuruk ve ölçülebilir) duruma gelmesini sağlamıştır.
2. Test maddelerinin ve performans görevlerinin, hangi bilişsel düzeye ve hangi bilgi türüne hitap ettiğini belirlemek mümkün duruma gelmiştir.
3. MEB ve benzeri kurumlar, öğretim programlarını yapılandırırken bu iki boyutlu matrisi kullanarak kazanımları daha sistematik duruma getirebilmişlerdir.

Yeni sınıflama tüm katkılarının yanında yaygın biçimde kabul görse de bazı yönlerden eleştirilmiştir. Bu eleştirilerden bazıları şöyledir:

1. Özellikle deneyimsiz öğretmenler iki boyutlu matrisi günlük ders planları için kullanırken karmaşık bulabilirler.
2. Bilişsel süreçler, sınıflamada belirtildiği gibi her zaman birbirinden net biçimde ayrılmayabilir.
3. Sınıflama, Batı merkezli bir bilişsel anlayışa dayandığından, farklı kültürlerdeki öğrenme dinamiklerini yeterince yansıtmayabilir.

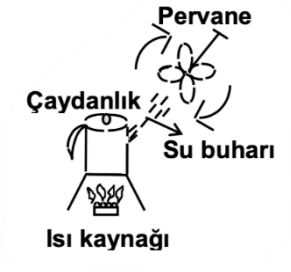
Bu eleştiriler, modelin geçerliği açısından olumsuz görünse de farklı öğrenme modelleriyle etkileşime açık bir yapıya sahip olduğunu göstermektedir.

Bilişsel süreç ve bilgi birikimi boyutlarının ilişkilendirildiği bazı madde örnekleri aşağıda verilmiştir:

Örnek 4.7. Olgusal Bilginin Ölçülmesine Yönelik Açık Uçlu Madde

Bilgi birikimi boyutu: Olgusal bilgi
Bilişsel süreç boyutu: Çözümleme
Sınıf Aralığı: Ortaokul 6-8. sınıf
Aşağıdaki görselde bazı kuşların ayak ve gaga özellikleri verilmiştir: **Şekil 1.** *Kuşların Ayak ve Gaga Özellikleri*  **Şekil 1(a)** **Şekil 1(b)** **Şekil 1(c)** • Taneli besinlerle beslenme • Ağaç dallarına tutunabilme • Yüzememe **Bu şekillerin hangisinde verilen hayvanlar bu özellikler sahiptir, hangileri değildir? Yanıtınızı şekillerle ilişkilendirerek yazınız.**

Örnek 4.8. Kavramsal Bilginin Ölçülmesine Yönelik Açık Uçlu Madde

Bilgi birikimi boyutu: Kavramsal bilgi
Bilişsel süreç boyutu: Uygulama
Sınıf Aralığı: Ortaokul 7-8. sınıf
Aşağıda "kaldırma kuvveti**"nin etkisiyle ilişkili üç farklı görsel verilmiştir:** **Şekil 2** *Kaldırma Kuvvetinin Olası Etkileri*  **Şekil 2(a)**  **Şekil 2(b)**  **Şekil 2(c)** **Bu şekillerin hangisinde kaldırma kuvvetinin etkisi vardır, hangisinde yoktur? Gerekçenizi kaldırma kuvvetiyle ilişkilendirerek yazınız.**

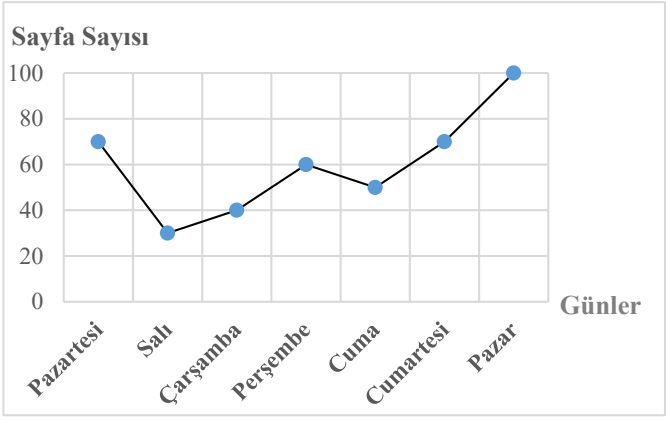
Örnek 4.9. Kavramsal Bilginin Ölçülmesine Yönelik Açık Uçlu Madde

Bilgi birikimi boyutu: Kavramsal bilgi
Bilişsel süreç boyutu: Anlama
Sınıf Aralığı: Ortaokul 5-8. sınıf
(1) Uygar ve kültürlü bir insan olmak için çok okumalı, ders kitapları ile yetinmemeliyiz. (2) Uygarlık, bizden önceki kuşakların biriktirdiği bilgi ve anıların bir toplamıdır. (3) Biz, uygarlığa o kuşakların kitaplarını okumakla katılabiliriz. (4) Uygar bir insan olmanın biricik yolu okumaktır. **Yukarıdaki cümlelerin hangisinde, bir kavram ve açıklaması birlikte verilmiştir? Yanıtınızı yorumlayarak yazınız.**

Örnek 4.10. İşlemsel Bilginin Ölçülmesine Yönelik Açık Uçlu Madde

Bilgi birikimi boyutu: İşlemsel bilgi
Bilişsel süreç boyutu: Anlama
Sınıf Aralığı: Ortaokul 6-8. sınıf
(1) Sanatta başarı durağan değildir. (2) Bunun içindir ki gerçek sanatçı, sanatta doruk diye bir noktayı kabul etmez. (3) Onun amacı kendinden öncekileri aşarak daha ileriye gitmektir. Çünkü güzelliğin sınırı yoktur. (4) Bugün dünyanın beğendiği sanatçılar, alanlarında her şeyi yapmış sanatçılardır. **Yukarıdaki cümlelerden hangisi, paragrafta işlenen düşünceyle tutarlı <u>değildir</u>? Yanıtınızı dil kuralları açısından neden belirterek yazınız.**

Örnek 4.11. İşlemsel Bilginin Ölçülmesine Yönelik Açık Uçlu Madde

Bilgi birikimi boyutu: İşlemsel bilgi
Bilişsel süreç boyutu: Analiz etme/Çözümleme
Sınıf Aralığı: Ortaokul 8, Lise 9. sınıf
Aşağıda bir öğrencinin haftalık okuduğu kitap sayfa sayısını gösteren grafik verilmiştir:  **Buna göre, bu öğrenci günde ortalama kaç sayfa kitap okumaktadır? Çözümünüzü yazarak anlatınız.**

Örnek 4.12. Üstbilişsel Bilginin Ölçülmesine Yönelik Açık Uçlu Madde

Bilgi birikimi boyutu: Üstbilişsel bilgi
Bilişsel süreç boyutu: Çözümleme
Sınıf Aralığı: Lise 9-12. sınıf
Yaşamın getirdiği bazı güçlükler ve sıkıntılar altında bunalır; dertleşecek, sohbet edecek bir arkadaş ararız. Bu nedenle, içten ve samimi arkadaşlık ilişkileri mutlu olmamıza katkı sağlar. **Bu görüşün, sizin bilgi ve deneyimlerinizle ne kadar uyumlu olduğunu düşünüyorsunuz? Kendi yaşamınızdan örnekler vererek ve yukarıdaki görüşle ilişkilendirerek yazınız.**

2.1.3. Thomas Haladyna Sınıflaması

Thomas Haladyna, test maddesi yazımına dair sistematik yaklaşımıyla öne çıkmaktadır. *Writing Test Items to Evaluate Higher Order Thinking* adlı çalışmasında, öğretmenlerin yalnızca bilgi (hatırlama) düzeyinde sorular yazmakla yetinmemesi gerektiğini vurgular. Bunun yerine, öğrencinin eleştirel düşünmesini ve akıl yürütmesini sağlayacak sorulara ağırlık verilmesini önerir. Üst düzey düşünme becerileri (higher order thinking skills), hatırlamanın ötesine geçer; analiz etme, sentez yapma, değerlendirme, transfer etme, yaratma gibi bilişsel etkinlikleri kapsar. Haladyna'ya göre bu tür düşünme, birden fazla bilgi kaynağını kullanarak karar verme, problem çözme ve yeni fikirler üretme sürecini içerir.

Haladyna (1997) bilişsel alan davranışlarının entelektüel davranışları kapsadığını ve bu alanın *erişi* (achievement) ve *yetenek* (ability) olmak üzere iki ana davranış kategorisini içerdiğini vurgulamaktadır. Erişi bir ders, ünite ya da dönem süresi gibi kısa zamanda gelişen öğrenci davranışlarıdır. Erişi aynı zamanda, kolaylıkla değişebilen ve öğretmenlerin öğrencileri kolaylıkla etkileme eğiliminde olduğu bilişsel davranışlara karşılık gelmektedir. Erişi; bilgi (knowledge) ve beceri (skill) olmak üzere iki bileşene sahiptir.

Bilgiler kısa sürede edinilen, ders içeriğinin anlaşılmasını ifade eden daha tekil öğrenmelerle ilgiliyken, *beceriler* anlaşılan birden fazla bilginin uygulamaya yönelik kullanılmasını içerir ve edinilmesi bilgilere göre daha uzun zaman alır. *Yetenek* ise bilişsel birden fazla becerinin bir araya gelmesinden oluşan, geliştirilmesi daha uzun zaman alan ve erişiye göre değişmesi daha zor bilişsel davranışlardır. Yetenekler kısaca üst düzey zihinsel becerilerdir.

Beceri ve yetenek boyutları için, şöyle bir örnek verilebilir: Öğrencinin öğretim sürecinde açıkça verilmiş adımları izleyerek ikinci derece bir denklemi çözebilmesi "erişi"yi (öğrenilmiş bilgi ve becerinin uygulanması) temsil ederken, öğrenciden gerçek yaşamdan bir problem durumu oluşturup bu durumu ikinci derece bir denklem formuna dönüştürmesi, farklı çözüm yolları sunması ve çözümü gerekçelendirmesi "yetenek" boyutunu temsil eder. İkinci durumda öğrenciden işlemin çözüm yolunu tekrarlaması değil, bilgiyi yeni bir bağlama transfer etmesi, problem kurması, çözüm stratejileri sunması ve çözüm için akıl yürütmeyi temellendirmesi beklenmektedir. Erişiler, öğrenilmiş bilgi ve becerilerden oluşurken; yetenekler, bilgi ve beceriye dayanmanın yanında daha karmaşık ve öğrenilmesi daha zor üst düzey bilişsel davranışlardan oluşurlar. Haladyna'ya (1997) göre okulların asıl hedefi öğrencilere yetenek adı verilen üst düzey düşünme becerilerinin kazandırılması olmalıdır. Buradan hareketle, öğrencilerden beklenen üst düzey zihinsel yetenekleri; *anlama, problem çözme, eleştirel düşünme ve yaratıcı düşünme* olmak üzere dört grupta sınıflandırmıştır. Bu sınıflamanın düzeyleri ve özellikleri Tablo 5'te verilmiştir.

Tablo 5

Haladyna Sınıflamasının Düzeyleri ve Özellikleri

Düzeyler	**Özellikleri**
Anlama (Understanding)	Bu düzey, Bloom sınıflamasındaki kavrama basamağına benzer biçimde, olgular, kavramlar, ilkeler ve süreçler için geçerlidir. Anlama, üç belirli öğrenci davranışı türüyle ilişkilidir: 1. Öğrenci, bir tanımı anladığını göstermek için başka sözcüklerle ifade etmeli, doğru bir tanım üretmelidir. 2. Öğrenci; bir olguyu, kavramı, ilkeyi ya da sürece ilişkin özelliği anladığını göstermek için ezberlenmemiş bir dizi özellik üretmelidir. 3. Öğrenci; bir olguyu, kavramı, ilkeyi ya da süreci anladığını göstermek için daha önce karşılaşılmamış örnekler üretmelidir.
Problem Çözme (Problem Solving)	Problem çözme, genellikle bir yanıta, bir hedefe ulaşmaya yol açan bir dizi zihinsel süreçtir. Problem çözme; ezberleme, anlama ve eleştirel düşünme, yaratıcılık gibi diğer zihinsel davranış türlerini içerebilir. Bazı problemlerin çözümünde birden fazla doğru yanıt bulunabilir. Problem çözme şu aşamaları içerir: *Problemi belirleme, problemi tanımlama, çözümleme, çözüm önerme, deney (test etme, deneme), sonuçlandırma.*
Eleştirel Düşünme (Critical Thinking)	Eleştirel düşünme; düşünmeyi, karşılaştırmayı, değerlendirmeyi ve ardından bir yargıya varmayı sağlar. Eleştirel düşünmeyi tanımlamak kolay değildir; genellikle öğretilebilen bir dizi zihinsel beceriler bütünü olarak düşünülmektedir. Bu düzeyde eleştirel düşünme; *değerlendirme ve tahmin etme, tümdengelim ve tümevarım* akıl yürütme olarak ele alınmaktadır.
Yaratıcı Düşünme (Creative Thinking)	Yaratıcılığın *yaratıcı düşünme* ve *yaratıcı üretim* olmak üzere iki yönü vardır. Bunlardan yaratıcı üretim *bilimsel* ve *sanatsal* olmak üzere ikiye ayrılır. Yaratıcılık öğretim programının resmi bir parçası değil, insan olmanın doğal bir yönü olarak düşünülmelidir. Bu yönüyle okullar yaratıcılığın gelişimini desteklemelidir. "Yenilikçi, özgün, yeni ve benzersiz" sözcükleri yaratıcılıkla ilişkilendirilmektedir. Yaratıcılık; tanıdık parçaların ve ögelerin yeni bir birleşimini içerir. Yaratıcı düşüncenin çoğalması, yoğun yaratma güdüsü, yüksek düzeyde yaratıcı eylem, enerji, bağlılık, disiplin, özveri, sıkı çalışma ve güdü gerektirir.

Haladyna'ya (1997) göre, üst düzey düşünme becerilerini ölçen sorular aşağıdaki özelliklere sahip olmalıdır:

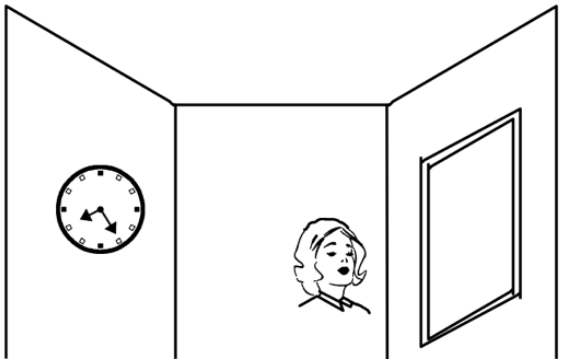
- Öğrenciye yeni bir durum sunmalıdır.
- Yalnızca hatırlamaya değil, ilişkilendirmeye ve çıkarıma dayanmalıdır.
- Alternatif yanıtları düşündürmeli, yanıt tek doğruya indirgenmemelidir.
- Gerekçelendirme istemelidir.

Bunun için yalnızca çoktan seçmeli maddeler değil, senaryo temelli, görsel içerikli, problem çözmeye yönelik madde türleri kullanılmalıdır. Haladyna sınıflamasına dayalı bazı madde örnekleri aşağıda verilmiştir:

Örnek 4.13. Anlama Düzeyinin Ölçülmesine Yönelik Açık Uçlu Madde

Ders: Türkçe
Beceri: Örneklendirme
Sınıf Aralığı: Ortaokul 6-8. sınıf
Sık sık dil konularına "eğilen" yazarlarımız var. **Bu cümlede kullanılan "**eğilmek**" sözcüğüyle aynı anlama gelen farklı bir cümle yazınız.**

Örnek 4.14. Anlama Düzeyinin Ölçülmesine Yönelik Açık Uçlu Madde

Ders: Fen Bilimleri
Beceri: Karşılaştırma
Sınıf Aralığı: Ortaokul 6-8. sınıf
Aşağıdaki şekilde bir birey karşısındaki düz aynadan arkasındaki duvarda bulunan saate bakmaktadır.  **Birey bu saati aynada nasıl görür? Çizerek gösteriniz ve bu görüntüyü nasıl belirlediğinizi yazınız.**

Örnek 4.15. Problem Çözme Düzeyinin Ölçülmesine Yönelik Açık Uçlu Madde

Ders: Fen Bilimleri
Beceri: Deney yapma
Sınıf Aralığı: Ortaokul 7-8, Lise 9. sınıf
Bir öğrenci gübre miktarının bitkilerin büyümesine olan etkisini incelemek istemektedir. **Bu öğrenci nasıl bir deney yaparsa araştırdığı soruya yanıt verebilir? Deneyi ayrıntılı olarak yazınız.**

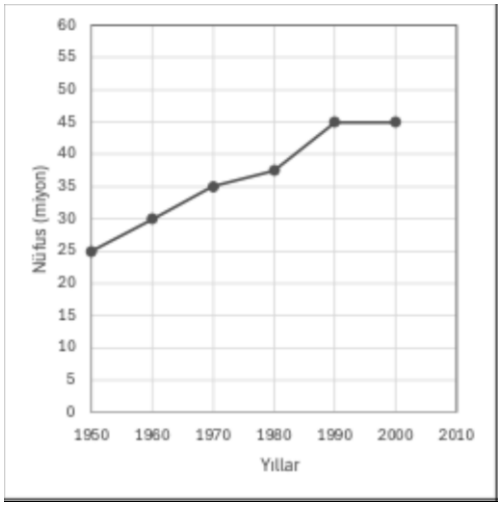
Örnek 4.16. Problem Çözme Düzeyinin Ölçülmesine Yönelik Açık Uçlu Madde

Ders: Sosyal Bilgiler
Beceri: Sorunu tanımlama ve çözüm önerme
Sınıf Aralığı: Ortaokul 7-8, Lise 9-10. sınıf
Geçmişte yaşanan tarihi olaylar günümüze önemli bilgiler sunmaktadır. Aşağıdaki soruyu duruma göre yanıtlayınız. Hun Türkleriyle Çinliler arasındaki çatışmalar Mete Han'ın ölümünden sonra da devam etti. Çinliler, Türkleri savaşlarla yenemeyeceklerini anlayınca Türkleri birbirine düşürmek için propaganda yaptılar. Sonunda Büyük Hun Devleti ikiye ayrıldı. Bir süre sonra da varlıkları sona erdi. **Bu olayda bir devletin devamlılığıyla ilgili hangi sorun üstünde durulmuştur? Sorunu tanımlayınız ve bir çözüm öneriniz.**

Örnek 4.17. Eleştirel Düşünme Düzeyinin Ölçülmesine Yönelik Açık Uçlu Madde

Ders: Sosyal Bilgiler
Beceri: Değerlendirme
Sınıf Aralığı: Ortaokul 8, Lise 9-10. sınıf
Mondros Ateşkes Antlaşması'nın bazı maddeleri şunlardır: • Osmanlı ordusu terhis edilecek. • Orduya ait silahlar, İtilaf Devletlerinin emrine verilecek. • Haberleşme ve ulaşım araçları İtilaf Devletlerinin denetimine verilecek. • İtilaf Devletleri güvenliklerini tehdit edecek bir durumda herhangi bir stratejik noktayı işgal edebilecek. **Antlaşmanın bu maddelerine göre, Osmanlı Devleti'nin içine düştüğü durum nasıl tanımlanır? Bu durumu, "**toplumsal, ekonomik, siyasal, kültürel ya da askeri**" etkenlerden birini ölçüt alarak ve antlaşma maddeleriyle ilişkilendirerek yazınız.**

Örnek 4.18. Eleştirel Düşünme Düzeyinin Ölçülmesine Yönelik Açık Uçlu Madde

Ders: Sosyal Bilgiler
Beceri: Yordama
Sınıf Aralığı: Ortaokul 8, Lise 9-10. sınıf
Aşağıdaki grafikte bir ülke nüfusunun yıllara göre değişimi verilmiştir:  Bu ülkenin 2000-2010 yılları arasındaki nüfus artış miktarının, 1960-1970 yılları arasındaki artış miktarı ile aynı olacağı söyleniyor **Buna göre, 2010 yılındaki nüfus kaç milyon olacaktır? Tahmininizi hem matematiksel işlemle hem de matematik diliyle yazarak anlatınız.**

2.1.4. Marzano ve Kendall'ın Sınıflaması

Robert J. Marzano ve John S. Kendall, 2000'li yılların başında, Bloom ile Anderson ve Krathwohl'un modellerinin öğretmenler için yol gösterici olduğunu ancak bilişsel süreçleri açıklamada yetersiz kaldığını ileri sürmüşlerdir. Onlara göre bu modeller, öğrencinin öğrenmeyi nasıl yapılandırdığını ve zihinsel süreçlerin etkileşimini tam olarak açıklayamamaktadır. Özellikle öğrenme yalnızca bilişsel değil; duyuşsal, üstbilişsel ve sosyal boyutlara da sahiptir. Bu gerekçeyle geliştirilen *New Taxonomy of Educational Objectives* hem bilişsel süreçleri hem de bireyin öğrenme güdüsünü ve zihinsel kontrol mekanizmalarını kapsayan bir model sunmaktadır. Yeni sınıflama, üç geniş sistem altında ve altı düzeyde yapılandırılmıştır. Bu yapı, öğrencinin öğrenmeyi nasıl yönettiğini (üstbiliş), hangi süreçler aracılığıyla bilgiyle etkileşime girdiğini (biliş) ve öğrenmeye ne kadar istekli olduğunu (duyuşsal) birlikte ele alır (Marzano & Kendall, 2007);

A. Bilişsel Sistem (Cognitive System)

1. Geri Getirme
2. Anlama
3. Çözümleme
4. Bilgiyi Kullanma

B. Üstbilişsel Sistem (Metacognitive System)

5. Üstbiliş

C. Öz-Sistem (Self-System)

6. Öz-Sistem Düşüncesi

Bu sistemler ve içerikleri en üst düzeyden alt düzeye doğru Tablo 6'da verilmiştir.

Tablo 6

Marzano ve Kendall Sınıflaması

Sistem	Düzey	İşlem	Hedeflerin Genel Biçimi
Öz-Sistem	**6. Öz-Sistem Düşüncesi** (Self-System Thinking)	Önemi İnceleme	Öğrenci; bilgilerin, zihinsel ya da psikomotor süreçlerin kendisi için ne kadar önemli olduğunu ve bu algının altında yatan mantığı belirleyebilecektir.
		Etkililiği İnceleme	Öğrenci; bilgiler, zihinsel ya da psikomotor süreçlerle ilgili olarak yetkinliğini ya da anlayışını geliştirme yeteneği hakkındaki inançlarını ve bu algının altında yatan mantığı belirleyebilecektir.
		Duygusal Tepkiyi İnceleme	Öğrenci; bilgiye, zihinsel ya da psikomotor işleme verdiği duygusal tepkileri ve bu tepkilerin nedenlerini belirleyebilecektir.
		Güdüyü İnceleme	Öğrenci; bilgiler, zihinsel ya da psikomotor süreçlerle ilgili yeterlik ya da anlayışı geliştirmeye yönelik genel güdü düzeyini ve bu güdü düzeyinin nedenlerini belirleyebilecektir.
Üstbilişsel Sistem	**5. Üstbiliş** (Metacognition)	Hedef Belirleme	Öğrenci; bilgilere, zihinsel ya da psikomotor süreçlere göre bir hedef ve bu hedefe ulaşmak için bir plan oluşturabilecektir.
		Süreç İzleme	Öğrenci belirli bir hedefe ulaşma yolundaki ilerlemeyi; bilgi, zihinsel ya da psikomotor süreç açısından izleyebilecektir.
		İzleme Netliği	Öğrenci; bilgiler, zihinsel ya da psikomotor süreçler hakkında ne ölçüde netliğe sahip olduğunu belirleyebilecektir.
		İzleme Doğruluğu	Öğrenci; bilgiler, zihinsel ya da psikomotor süreçler hakkında ne ölçüde doğru olduğunu belirleyebilecektir.

Tablo 6 (Devamı)

Marzano ve Kendall Sınıflaması

Sistem	Düzey	İşlem	Hedeflerin Genel Biçimi
Bilissel Sistem	**4. Bilgiyi Kullanma** (Knowledge Utilization)	Karar Verme	Öğrenci, genel olarak karar vermek; bilginin, zihinsel ya da psikomotor süreçlerin kullanımı hakkında kararlar almak için bilgileri, zihinsel ya da psikomotor süreçleri kullanabilecektir.
		Problem Çözme	Öğrenci, genel olarak sorunları çözmek; bilgi, zihinsel ya da psikomotor süreçlerle ilgili sorunları çözmek için bilgileri, zihinsel ya da psikomotor süreçleri kullanabilecektir.
		Deney Yapma	Öğrenci, genel olarak hipotezler üretmek ve test etmek; bilgiler, zihinsel ya da psikomotor süreçler hakkında hipotezler üretmek ve test etmek için bilgileri, zihinsel ya da psikomotor süreçleri kullanabilecektir.
		Araştırma	Öğrenci, genel olarak araştırmalar yapmak; bilgiler, zihinsel ya da psikomotor süreçler hakkında araştırmalar yapmak için bilgileri, zihinsel ya da psikomotor süreçleri kullanabilecektir.
	3. Çözümleme (Analysis)	Eşleştirme	Öğrenci; bilgilerle, zihinsel ya da psikomotor süreçlerle ilgili önemli benzerlikleri ve farklılıkları belirleyebilecektir.
		Sınıflandırma	Öğrenci; bilgilerle, zihinsel ya da psikomotor süreçlerle ilgili üst ve alt kategorileri belirleyebilecektir.
		Çözümleme Hataları	Öğrenci; bilgilerin, zihinsel ya da psikomotor süreçlerin sunumunda ya da kullanımında hataları belirleyebilecektir.
		Genelleme	Öğrenci; bilgilere, zihinsel ya da psikomotor süreçlere dayalı olarak yeni genellemeler ya da ilkeler oluşturabilecektir.
		Belirtme	Öğrenci; bilgilerin, zihinsel ya da psikomotor süreçlerin mantıksal sonuçlarını belirleyebilecektir.

Tablo 6 (Devamı)

Marzano ve Kendall Sınıflaması

Sistem	Düzey	İşlem	Hedeflerin Genel Biçimi
Bilişsel Sistem	**2. Anlama** (Comprehension)	Bütünleştirme	Öğrenci; bilgilerin, zihinsel ya da psikomotor süreçlerin temel yapısını, eleştirel ve eleştirel olmayan özellikleri belirleyebilecektir.
		Sembolleştirme	Öğrenci; kritik ve kritik olmayan ögeleri ayırt ederek bilginin, zihinsel ya da psikomotor sürecin doğru bir sembolik temsilini oluşturabilecektir.
	1. Bulup/Geri Getirme (Retrieval)	Tanıma	Öğrenci; bilginin özellikleriyle ilgili doğru ifadeleri doğrulayabilir, ancak bilginin yapısını anlayamayabilir ya da kritik ve kritik olmayan bileşenleri ayırt edemeyebilir.
		Hatırlama	Öğrenci; bilginin özelliklerini üretebilecektir, ancak bilginin yapısını anlayamayabilir ya da kritik ve kritik olmayan bileşenleri ayırt edemeyebilir.
		Uygulama	Öğrenci; önemli bir hata yapmadan bir işlemi gerçekleştirebilecek, ancak işlemin nasıl ve ne işe yaradığını tam olarak anlayamayacaktır.

Tablo 6'da gösterildiği gibi yeni sınıflama; üç bilgi alanını (bilgi, zihinsel işlemler ve psikomotor işlemler) ve üç süreç kategorisini (bilişsel, üstbilişsel ve öz-sistem) kapsayan geniş bir hedef yelpazesi oluştururken ölçme ve değerlendirme açısından da zengin bir bakış açısı sunmaktadır. Düzey 5 ve 6 (üstbilişsel ve öz-sistem düşüncesi), daha akademik nitelikte kabul edilen ilk dört düzeyin tamamlayıcısı ve destekleyici olarak görülmekte ve derslerin öğretim programlarında değerlendirilmesi önerilmektedir (Marzano & Kendall, 2007). Bu modele göre "çözümleme" ve "bilgiyi kullanma" üst düzey zihinsel özelliklerinin gelişiminde önemli rol oynamaktadır. Özellikle "bilgiyi kullanma" düzeyi, üst düzey bilişsel işlemi temsil eder ve gerçek yaşam problemlerinde bilgileri kullanmayı ve uygulamayı içerir. Bloom sınıflamasındaki "uygulama", "sentez" ve "değerlendirme" boyutlarından daha kapsamlıdır.

Aşağıda bu düzeylere ilişkin madde (performans görevi) örneklerine yer verilmiştir. Örnek 4.20'de yer alan görevin puanlanması için Tablo 7'deki dereceli puanlama anahtarı eklenmiştir (Marzano ve ark., 1993):

Örnek 4.19. Bilişsel Sistemin Çözümleme Düzeyinin Ölçülmesine Yönelik Performans Görevi

Sistem: Bilişsel Sistem
Düzey 3: Çözümleme
İşlem: Çözümleme Hataları Görevi
Sınıf Aralığı: Ortaokul - Lise
Görev: Yerel gazete ve televizyon kanallarının yakın zamanda haber yaptığı bir felaketi belirleyin. Göreviniz, medyanın felaketle ilgili bilgileri halka nasıl ilettiğini ve yanlış bilginin belirli örneklerini belirlemektir. Bu görevde, çeşitli gazetelerden, televizyon yayınlarından, görgü tanığı ifadelerinden ya da röportajlardan örnekler eklemeniz beklenmektedir. Ayrıca, bu yanlışlıkların bilgiyi alan kişileri nasıl etkilediğine dair belirli örnekler sunmalı ve bu yanlışlıkların halka iletilmesini önlemek için neler yapılabileceğini de açıklamalısınız. Bulgularınızı bir medya mensubuna sunmanın bir yolunu belirleyin, o kişinin yanıtını kaydedin ve sınıftaki diğer üyelerle paylaşın. *İçerik Standardı* Sosyal Bilgiler / Dil Sanatları 1. Yerel medyayı ve bir felaket sırasında bilginin halka nasıl iletildiğini anlama beceriniz. *Yaşam Boyu Öğrenme Standartları* Karmaşık Düşünme: Çözümleme Hatalarını Bulma 1. Bilgi ya da süreçlerdeki önemli hataları belirleme ve ifade etme beceriniz. 2. Hataların, kaynaklandıklarını bilgi ya da süreç üzerindeki etkilerini doğru biçimde tanımlama beceriniz. Etkili İletişim 1. Çeşitli bilgi toplama tekniklerini ve bilgi kaynaklarını kullanma beceriniz.

Örnek 4.20. Bilişsel Sistemin Bilgiyi Kullanma Düzeyinin Ölçülmesine Yönelik Performans Görevi

Sistem: Bilişsel Sistem
Düzey 4: Bilgiyi Kullanma
İşlem: Tümevarım Görevi
Sınıf Aralığı: Ortaokul - Lise
Görev: İnsan davranışlarını gözlemlemek ilginç ve genellikle çok açıklayıcı olabilir. İnsanları ve davranışlarını gözlemleyebileceğiniz birkaç farklı durum seçin. Örneğin, bir alışveriş merkezinde, bir spor etkinliğinde, bir öğle yemeğinde ya da bir film başlamadan önceki on beş dakika içinde insanları gözlemlemeyi seçebilirsiniz. İnsanları belirli bir durumda; birkaç kez, tercihen her seferinde farklı kişilerle gözlemlemelisiniz. Gözlemlerinizi yazın; her bir gözlemin tarihini ve saatini kaydedin. Bu gözlem günlüğü, incelediğiniz çeşitli insan davranışı kategorileri hakkındaki bilginizi yansıtmalıdır. Gözlemlerinize dayanarak hangi önemli sonuçlara varabilirsiniz? Bu sonuçlara ulaşmak için attığınız adımları açıklayın. Sonuçlarınızı bir üniversitenin öğrenci panelinde sunacak ve kanıtlarınızı savunacaksınız. Ayrıca Tablo 7'deki dereceli puanlama anahtarında verilen ölçütlere göre değerlendirmeye tabi tutulacaksınız. *İçerik Standardı* Sosyal Bilimler 1. İnsan davranışı kategorileri hakkındaki anlayışınız. 2. İnsan davranışı kategorilerini anlama düzeyiniz. *Yaşam Boyu Öğrenme Standartları* Karmaşık Düşünme: Destekli Tümevarım 1. Tümevarım yapmak için ögeleri (bilgi parçaları ya da gözlemler) belirleme yeteneğiniz. 2. Seçilen bilgi ya da gözlemlerden doğru sonuçlar (tümevarımlar) çıkarma ve bunları ifade etme yeteneğiniz. Etkili İletişim 1. Çeşitli kitlelerle etkili bir şekilde iletişim kurma yeteneğiniz.

Tablo 7

Analitik Dereceli Puanlama Anahtarı

Ölçütler	Düzeyler			
	1	**2**	**3**	**4**
Tümevarım yapmak için belirli parçaları belirler/gözler.	Önemsiz parçalar seçer.	Önemli sayılan bazı parçaları seçer.	Konu için önemli olan parçaların çoğunu seçer.	Konuyu yaratıcı biçimde yansıtacak tüm parçaları doğru olarak seçer.
Tümevarım yapılan gözlemleri yorumlar.	Konuyla ilgisi olmayan mantıksız yorumlar yapar.	Konuyla ilgili az sayıda doğru yorumlar yapar.	Birkaç istisna dışında konu hakkında önemli yorumlar yapar.	Konunun ayrıntılı incelendiğini gösteren doğru yorumlar yapar.
Seçilen gözlemlerden doğru sonuçlar çıkarır ve sonuçları doğru yorumlar.	Gözlemlerden hatalı sonuç çıkarır ve sonuçları yorumlayamaz.	Gözlemlerden az sayıda yorum yansıtan sonuçlar sunar.	Gözlemlerden, birkaç istisna dışında, mantıksal sonuçlar sunar.	Gözlemler ile yorumlar arasında net ve mantıklı bağlantılar kuran sonuçlar çıkarır.

Örnek 4.21. Bilişsel Sistemin Bilgiyi Kullanma Düzeyinin Ölçülmesine Yönelik Performans Görevi

Sistem: Bilişsel Sistem
Düzey 4: Bilgiyi Kullanma
İşlem: Karar Verme Görevi
Sınıf Aralığı: Ortaokul - Lise
Görev: 15. ya da 16. yüzyılda yaşamış ve eylemleri önemli sonuçlar doğurmuş etkili bir kişi seçin. Bu kişinin bu önemli eylemleri gerçekleştirmeden önce dikkate almış olması gereken etkenleri belirleyin. Bu kişi için başka hangi alternatifler mevcuttu? Kişiyi eylemde bulunmaya güdüleyen hedefleri ve eylem kararı verirken uygulamış olabileceği ölçütleri belirleyin. Bir alternatifi diğerine tercih etmenin olası dezavantajları nelerdi? Riskler ve ödüller nelerdi ve bunlar nasıl ölçülebilirdi? Geriye dönüp bakma fırsatı olsaydı, aynı seçimi yapar mıydı? Nedenini açıklayınız. Bulgularınızı sözlü ya da yazılı olarak sunabilirsiniz. Değerlendirme ölçütleri aşağıda verilmiştir: *İçerik Standardı* Sosyal Bilgiler ve Tarih 1. Tarihteki etkili insanların değer ve inanç sistemlerinden ne kadar etkilendiğiniz ve onları ne kadar iyi anladığınız. *Yaşam Boyu Öğrenme Standartları* Karmaşık Düşünme: Karar Verme 1. Dikkate alınması gereken uygun ve önemli alternatifleri belirleme yeteneğiniz. 2. Alternatifleri değerlendirmek için önemli ve uygun kriterleri belirleme yeteneğiniz. Bilgi İşleme 1. Bilgileri yorumlama ve sentezleme yeteneğiniz. Etkili İletişim 1. İletişim yoluyla çeşitli amaçlara hizmet etme yeteneğiniz.

2.1.5. Bilişsel Sınıflamalara Eleştirel Bakış

Bilişsel sınıflamalar; öğrencilerin düşünme, öğrenme, edindikleri bilgileri kullanma anlayışlarına katkıda bulunmak için geliştirilmiş sistemlerdir. Ancak her modelin altında yatan kuramsal varsayımlar ve öğretimsel beklentiler farklıdır. Bu farklar, doğrudan eğitim kararlarına ve öğretmenlerin günlük uygulamalarına yansımaktadır.

Yukarıda yapılan açıklamalardan da görüleceği gibi, Marzano ve Kendall'ın sınıflaması, bilgi işleme dışında, öğrencinin öğrenme güdüsünü ve stratejik karar süreçlerini de kapsamına alan daha bütünsel bir yapı önermektedir. Buna karşılık Bloom'un, Anderson ve Krathwohl'un ve de Haladyna'nın modelleri, daha çok öğretim hedeflerinin davranışa dönüştürülmesiyle ilgilenmiştir. Bu yönüyle sınıflamaların birbirlerini tamamladıkları ve giderek daha fazla bilgiyi eğitim ortamlarına sunduklarını söyleyebiliriz.

Bilişsel sınıflamaların uygulamadaki üstünlükleri ve güçlükleri açısından kısaca şunlar söylenebilir:

Bloom Sınıflaması*:* Öğretmenlerin ilk kez sistematik olarak öğrenme hedefi belirlemelerine yardımcı olmuş; eğitim ve öğretim süreçlerine önemli katkılar sunmuştur. Ancak sınıflamada yer alan kavramların soyutluğu, sınıflamanın anlaşılmasında bazı belirsizlikler yaratmıştır. Çoğunlukla davranışçı kuramdan etkilenmiş olmasına rağmen bu sınıflamada bilişsel psikolojinin izleri de bulunmaktadır.

Anderson ve Krathwohl Sınıflaması*:* Göreli olarak Bloom sınıflamasına bağlı kalan bir yapısı bulunmaktadır. Bu durum eğitim kararlarındaki süreklilik açısından önemlidir. Sınıflamada yer alan hedefler; açık, davranışa dönüştürülebilir ve ölçülebilir biçimde yazılması açısından güçlüdür. İki boyutlu matris, öğretmenlerin ders içeriğini hem bilgi türü hem de süreç açısından yapılandırmasına ve uygun belirtke tablosu hazırlamalarına olanak tanımaktadır. Ancak bu sınıflamada, sosyal ve duyuşsal özelliklerin azlığı dikkat çekmektedir. Sınıflama, bilişsel psikolojiyi ve yapılandırmacı yaklaşımı temsil etmektedir.

Haladyna Sınıflaması*:* Birçok bilişsel sınıflamanın ortak özelliklerini dikkate alarak hazırlanmıştır. Sınıflamanın özellikle üst düzey düşünme becerilerine odaklanması, içinde bulunduğumuz 21. yüzyılın okuldan beklentilerini karşılaması açısından önemlidir. Ayrıca sınıflama, öğretmenlerin düşünme düzeyine uygun maddeler yazabilmelerini kolaylaştıran pratik kurallar sunmaktadır. Ancak sınıflamadaki bilgi türleri çok açık değildir.

Marzano ve Kendall Sınıflaması*:* Sınıflamada yer alan hedefler yalnızca bilişsel değil, üstbilişsel ve duyuşsal özellikleri de içermektedir. Bu nedenle bireyselleştirilmiş öğrenme tasarımları açısından daha işlevseldir, çünkü öğrenci açısından bireysel güdüyü ve kontrol süreçlerini de hesaba katmaktadır. Bu sınıflama henüz yaygın biçimde kurumsallaşmamış olmakla birlikte, öğrenci merkezli öğretim uygulamalarında önemli bir potansiyel sunmaktadır. Ancak sınıflamanın okul ortamında uygulanması diğer sınıflamalara göre daha karmaşıktır.

Genel olarak bakıldığında, Benjamin Bloom'un orijinal modeli, öğrenme hedeflerinin sınıflandırılmasında devrimsel bir başlangıç noktası sunmuştur. Anderson ve Krathwohl ise bu modeli güncelleyerek davranışsal ve bilişsel uyumu güçlendirmiştir. Haladyna, bu yapıların ölçülebilirliğini ve üst düzey zihinsel süreçlerle ilişkisini somut örnekler ve soru yazımı ilkeleriyle desteklemiştir. Marzano ve Kendall ise bilişsel süreçlerin psikolojik boyutlarına dikkat çekerek bütünsel bir öğrenme modeline öncülük etmiştir.

Tüm bunların yanında sınıflamalar aşağıda sunulan açılardan da tartışılmakta ve yorumlanmaktadır:

Aşamalı (hiyerarşik) mı değil mi?

Tüm sınıflamalar belirli bir sıralama ve aşama önerse de eğitimciler arasında bu sıralamaların mutlak olup olmadığına dair farklı görüşler bulunmaktadır. Örneğin, bazı uzmanlara göre "değerlendirme" her zaman "yaratma"dan daha zorlayıcı olabilmekte (Krathwohl, 2002) ya da problem çözme diğer basamakları içerebilmektedir (Haladyna, 1997).

Kültürel açıdan uygun mu?

Marzano ve Kendall tarafından önerilen sınıflama dışındaki tüm sınıflamaların çoğunlukla Batı merkezli bilişsel normlara dayandığı görülmektedir. Öğrenci profilinin kültürden etkilenen yapısı göz önüne alındığında, bu sınıflamaların öğrenme ve öğretme süreçlerindeki kullanımında sorunlara neden olabilecektir.

Eğitim teknolojileriyle bütünleşebilir mi?

Bilgisayar teknolojilerinin bilişsel sınıflamalardan daha hızla değiştiği göz önüne alınacak olursa, sınıflamaların bu teknolojilerle bütünleşmesini sağlayacak önlemlerin alınması da gerekmektedir. Bu açıdan bakıldığında; Anderson ve Krathwohl'un güncelleştirilmiş Bloom sınıflaması ile Marzano ve Kendall tarafından önerilen modelin, dijital öğrenme ortamlarıyla daha kolay eşleştirilebileceği söylenebilir.

2.2. Bilişsel Özellikler ve Ölçülmesi

2.2.1. Zekâ ve Ölçülmesi

Zekâ, bireylerin öğrenme, problem çözme, soyut düşünme, çevreye uyum sağlama ve karşılaşılan yeni durumlarla başa çıkma kapasitesi olarak tanımlanmaktadır. Yıllar boyunca toplumlar, bireylere ait statü ve rolleri düzenlemek ve üretim sürecinde görev alacak işgücünü verimli kullanmak amacıyla bireylerin öğrenme kapasitesini belirlemek istemişler, bu nedenle de zekânın ölçülmesine yönelik girişimlerde bulunmuşlardır. Sanayi devriminden sonra, özellikle eğitim alanında öğrencilerin öğrenme potansiyellerinin belirlenmesi, askeri alanda görev dağılımlarının yapılması ve çalışma yaşamında yeteneğin yönetilmesi gibi gereksinimler zekâ ölçümlerinin gelişimini hızlandırmıştır (Neisser ve ark., 1996).

Günümüzde kullanılan anlamıyla "zekâ" kavramının temelleri, Sir Francis Galton tarafından atılmıştır. Galton, insanların beceri düzeylerinin nasıl ve neden farklılık gösterdiğini merak etmiştir. 1869 yılında yayınladığı Kalıtsal Deha (Hereditary Genius) adlı kitabında zekâyı kalıtımsal açıdan ele almıştır. Galton'ın, zekâyla ilgili ortaya koyduğu dört temel nokta bugün hâlâ geçerliğini korumaktadır (Cohen & Swerdlik, 2009; Toker ve ark., 1968):

1. Zekânın bir düzeyi vardır ve bu ölçülerek gösterilebilir.
2. Zekâ düzeyi farklılıkları normal dağılım oluşturur.
3. Zekâ, tek doğru yanıtı olan nesnel testlerle ölçülebilir.
4. İki test sonucunun birbiriyle ilişkisi istatistiksel olarak belirlenebilir.

Farklı bilim insanları zekâya çeşitli bakış açılarıyla yaklaşmış ve günümüzdeki zekâ anlayışına katkı sağlamışlardır. Günümüzde hâlâ üzerinde durulan zekâ kuramlarından bazıları şöyledir: Galton'ın Zekâ Kuramı, Spearman'ın Genel Zekâ Kuramı, Thorndike'ın Çok Faktör Kuramı, Thurstone'un Grup Faktör Kuramı, Cattel'in Zekâ Kuramı, Horn'un Zekâ Kuramı, Carroll'ın Üç Katman Kuramı, Guilford'un Üç Boyutlu Zekâ Kuramı, Sternberg'in Üçlü Sac Ayağı Kuramı, Gardner'ın Çoklu Zekâ Modeli, Goleman'ın Duygusal Zekâ Yaklaşımı.

Zekâ testlerinin gelişimindeki dönüm noktası, Alfred Binet ve Théodore Simon'un 1905'te geliştirdiği zekâ testidir. Fransa'da ilköğretimde öğrenme güçlüğü yaşayan çocukları belirlemek amacıyla tasarlanan bu test, zihinsel gelişim yaşını belirlemeye dayalıdır (Binet & Simon, 1916). Daha sonra Lewis Terman, Binet-Simon testini uyarlayarak Stanford-Binet Zekâ Testi'ni geliştirmiş ve Amerika'da zekâ ölçümlerinin standart biçimde yapılmasını sağlamıştır (Terman, 1916). Zekâ çalışmaları Charles Edward Spearman'ın iki faktörlü modelinden daha ileriye taşınmış, Edward Thorndike, Louis Leon Thurstone ve Joy Paul Guilford tarafından zekânın çok boyutlu yapısı ortaya konmuştur (Toker ve ark., 1968). 20. yüzyılın ortalarında David Wechsler, zekâyı çok boyutlu bir kavram olarak ele almış ve bireylerin sözel, performans ve genel yeteneklerini ölçmeyi amaçlayan ölçekler geliştirmiştir. Wechsler Zekâ Ölçekleri, günümüzde hâlâ en yaygın kullanılan ölçme araçları arasında yer almaktadır.

Yirminci yüzyılda zekâ testleri, bireylerin sınıflandırılması, eğitimde seçme ve yerleştirme kararlarının verilmesi, askeri görev dağılımlarının yapılması ve işgücü planlamasının oluşturulması gibi amaçlarla kullanılmıştır (Carson, 2007). Ancak zekâ testleri; bireyleri yalnızca bir puanla etiketlediği, toplumsal eşitsizlikleri pekiştirdiği ve bireyin potansiyelini tam olarak yansıtmadığı eleştirilerine de maruz kalmıştır (Gould, 1996). Zekâ testlerinin en çok tartışılan yönlerinden biri de kültür boyutuyla ilgilidir. Zekâ testlerinin belirli kültürel değerlere göre geliştirilmesi, farklı sosyoekonomik ve kültürel özelliklere sahip grupların dezavantajlı duruma düşmesine yol açmıştır. Bu

nedenle kültürden bağımsız testler ya da Raven'in İlerleyen Matrisleri gibi dile dayalı olmayan görsel-uzamsal testler, zekânın ölçülmesinde daha evrensel ölçme araçları olarak öne çıkmıştır (Raven, 2000). Günümüzde zekâ ölçümleri, daha adil, geçerliği yüksek ve kültüre duyarlı araçlar aracılığıyla elde edilmektedir.

Zekâ testleri aynı zamanda eğitimde bireysel farklılıkların dikkate alınmasına da katkı sağlamıştır. Öğrencilerin güçlü ve zayıf yönlerinin belirlenerek öğretim yöntemlerinin bu farklılıklara göre uyarlanması, özel eğitime gereksinim duyan öğrencilerin erken dönemde belirlenmesi ve üstün yetenekli bireylerin desteklenmesi için zekâ testleri önemli bir araç olmuştur (Özgüven, 2012).

2.2.2. Yetenek ve Ölçülmesi

Yapılmış çalışmalar, zekâ ve yetenek kavramlarının bazı araştırmacılar tarafından ayrı anlamlarda (Ebel & Frisbie, 1991; Cohen & Swerdlik, 2009), bazıları tarafından ise aynı anlamda kullanıldığını (Thorndike & Thorndike-Christ, 2017) göstermektedir. Bu durum, zekâ ile yetenek arasında önemli bir örtüşme alanı bulunduğunu ortaya koymaktadır. Son yıllarda okul başarısının öğrencinin; anlama, problem çözme, akıl yürütme, eleştirel düşünme, yaratıcılık gibi üst düzey düşünme becerileriyle ilişkilendirilmesi, zekâ ve yetenek kavramlarının yanına başarı kavramının da eklenmesine neden olmuştur. Haladyna (1997) yeteneğin uzun vadeli öğrenmeyi gerektirdiğini vurgulamakta ve yeteneği bilgi ve beceriden ayırarak üst düzey düşünme alanı olarak tanımlamaktadır. Kutlu ve arkadaşları (2024) bu görüşü destekleyerek yeteneğin bilişsel özelliklerin yanında duyuşsal ve psikomotor özelliklerin birleşiminden oluşan zihinsel bir alan olduğuna vurgu yapmakta ve üstbilişsel (metacognition) süreçlerden destek aldığını da belirtmektedir.

Türk Dil Kurumu "yetenek" kavramını, bireyin bir şeyi anlama ya da yapabilme niteliği, istidat, kabiliyet; kalıtıma dayanan ve öğrenmeyi çerçeveleyen sınır; dışarıdan gelen etkiyi alabilme gücü gibi anlatımlarla tanımlamaktadır. Yetenek, bireylerin belirli alanlarda öğrenme, problem çözme, yeni beceriler geliştirme ve performans sergileme potansiyeli olarak ifade edilmektedir. Eğitimden çalışma yaşamına kadar geniş bir alanda kullanılan yetenek testleri, bireylerin hangi alanlarda başarılı olabileceği konusunda kestirim yapmaya yardımcı olmaktadır. Yetenek testleri, özellikle meslek seçimi ve işe alım süreçlerinde bireylerin hangi görevlerde daha verimli çalışabileceğini belirlemede önemli bir araç olarak kullanılmaktadır (Anastasi & Urbina, 1997).

Yeteneğin ölçülmesine yönelik ilk çalışmalar, zekâ testlerinin ötesinde daha özel becerilere odaklanmıştır. Spearman, geliştirdiği "genel yetenek" (g faktörü) kavramıyla, tüm bilişsel performansların ortak bir zihinsel kapasiteye dayandığını ileri sürmüştür (Spearman, 1927). Buna karşın Thurstone, yeteneklerin bağımsız boyutlara ayrılabileceğini savunmuş ve sözel, sayısal, mekânsal gibi farklı alanlarda yeteneklerin ölçülmesini önermiştir (Thurstone, 1938). Bu kuramsal ayrışma, modern yetenek

testlerinin çeşitlenmesine de temel oluşturmuştur. Örneğin, Differential Aptitude Tests (DAT), bireylerin farklı alanlardaki potansiyellerini ölçmek için geliştirilmiş en kapsamlı araçlardan biri olmuştur. DAT; sözel akıl yürütme, sayısal akıl yürütme, soyut düşünme, mekanik akıl yürütme, dil yetenekleri ve mekânsal görselleştirme gibi alt ölçekler içermektedir (Bennett ve ark., 1974). Bu tür ölçme araçları, özellikle lise ve üniversite öğrencilerinin mesleki yönlendirilmelerinde ve kariyer planlamalarında önemli katkılar sağlamıştır. Eğitimde; bireylerin farklı yeteneklerine göre yönlendirilmesine, toplumsal hareketliliğin artırılmasına ve bireylerin potansiyellerine uygun kariyer yolları seçmelerine katkıda bulunmuştur.

Yetenek ölçümleri, çalışma yaşamında da önemli bir rol oynamaktadır. Özellikle sanayi ve hizmet sektörlerindeki işe alım süreçlerinde işin özelliklerine uygun adayların belirlenmesi kritik bir öneme sahiptir. İşverenler, adayların yalnızca işe ilişkin akademik donanımlarını değil, aynı zamanda problem çözme, karar verme, sayısal çözümleme ya da mekânsal düşünme gibi alanlardaki yeteneklerini de değerlendirmek istemektedirler. Schmidt ve Hunter (1998), genel bilişsel yeteneğin iş performansını güçlü biçimde yordadığını ortaya koymuştur. Bu durum, yetenek ölçümlerinin işe alım süreçlerinde güvenilir bir karar aracı olarak kullanılmasına önemli bir dayanak oluşturmuştur.

Yetenek testleri, fırsat eşitliği açısından tartışılmaya devam etmektedir. Testlerin kültüre dayalı içerikleri ve kültüre olan bağımlılıkları, farklı sosyoekonomik gruplardan gelen bireylerin eşit koşullarda değerlendirilememesine neden olabilmektedir (Helms, 1992). Bu nedenle kültürel yanlılıkların azaltılması, modern psikometrinin öncelikli hedefleri arasında yer almıştır.

Türkiye'de, akademik yetenek ölçümleri eğitim sisteminde uzun süredir kullanılmaktadır. Ortaöğretime, yükseköğretime ve lisansüstü eğitime geçiş testlerinde ve bazı kamu personeli seçme sınavlarında; okuduğunu anlama, problem çözme, analitik düşünme, akıl yürütme, karar verme gibi düşünme süreçlerini ölçen yetenek maddelerinden yararlanılmaktadır. Tüm bu gelişmelere rağmen, günümüzde bireylerin düşünme potansiyellerini çok yönlü ölçebilecek kapsamlı ölçme araçlarına gereksinim duyulmaktadır. Bunun yanında yetenek testlerinin, bireylerin güçlü yönlerini belirlemenin yanı sıra toplumsal fırsat eşitliğini destekleme ve adil ölçmeler yapma özellikleriyle birlikte değerlendirilmesi gerekmektedir.

3. DUYUŞSAL ÖZELLİKLERİN ÖLÇÜLMESİ

Duyuşsal özellikler bireylerin; ilgi, tutum, güdü ve değer gibi duygusal tepkilerini kapsayan geniş bir alanı ifade etmektedir. Yapılan araştırmalar, öğrenci başarısında bilgi ve becerilerin yanı sıra; ilgi, tutum ve güdü gibi duyuşsal tepkilerin de kritik rol oynadığını göstermektedir (Krathwohl ve ark., 1964). Duyuşsal özelliklerin ölçülmesi, bireylerin öğrenme süreçlerine katılımını artırmak açısından zorunludur. John Dewey, eğitimin yalnızca bilişsel bilgi aktarımıyla sınırlı olamayacağını, sosyal ve duyuşsal

gelişimin öğrenmenin ayrılmaz bir parçası olduğunu belirtmiştir (Dewey, 1930). Carl Rogers ise, bireyin duyuşsal dünyasının eğitsel süreçlerde belirleyici olduğunu ve öğrenmenin ancak güvenli ve destekleyici bir ortamda duyuşsal katılım yoluyla gerçekleşebileceğini savunmuştur (Rogers, 1969). Bu yaklaşımlar, duyuşsal özelliklerin düzenli olarak ölçülmesine olan gereksinimi artırmıştır.

2000'li yılların başından itibaren Ekonomik Kalkınma ve İşbirliği Örgütü (Organisation for Economic Cooperation and Development -OECD) tarafından öğrencilerin temel yaşam becerilerini belirleme amacı taşıyan Uluslararası Öğrenci Değerlendirme Programı (Programme for International Student Assessment -PISA) uygulamaları okuma, fen (bilim) ve matematik okuryazarlık becerileri üzerinde duyuşsal özelliklerin etkili olduğunu ortaya koyan ölçme uygulamaları gerçekleştirmektedir (OECD, 2023a; 2023b). Bunun yanında bu çalışmalarda, okulların yaşam için gerekli bilişsel, duyuşsal ve sosyal becerileri yeterince kazandıramadığı da ele alınmıştır (Haladyna, 1997; Kutlu ve ark., 2024; Marzano, 1992; NCLB, 2001; Popham, 2004).

Duyuşsal özelliklerin ölçülmesinde kullanılan ilk araçlar genellikle öz bildirim anketleri, gözlem formları ve derecelendirme ölçekleridir. Özellikle Likert ölçekleri, bireylerin bir konuya yönelik tutumlarını derecelendirme yoluyla belirlenmesini ve nicel veriye dayanarak incelenmesini olanaklı kılmıştır (Likert, 1974). Günümüzde bu ölçekler eğitim alanında, öğrencilerin derslere, öğretmenlere ya da öğrenme süreçlerine yönelik tutumlarını belirlemede yaygın olarak kullanılmaktadır. Çalışma yaşamında ise duyuşsal özelliklerin ölçülmesi, çalışan bağlılığı, iş doyumu, güdü ve örgütsel iklim gibi özelliklerin değerlendirilmesinde önemli duruma gelmiş; bu ölçümler insan kaynakları politikalarının şekillendirilmesine, çalışan verimliliğini artırılmasına ve iş doyumunun geliştirilmesine katkı sağlamıştır.

3.1. Duyuşsal Alan Hedefleri

Duyuşsal özellikler hakkında ilk ve kapsamlı çalışmalardan biri Krathwohl ve arkadaşları (1964) tarafından yapılmıştır. Duyuşsal alan "duygular, değerler, ilgiler, değer verme, coşkular, güdüler ve tutumlar" gibi psikolojik özelliklerle duyuşsal olarak başa çıkma biçimimizi içermektedir. Krathwohl ve arkadaşları tarafından önerilen duyuşsal alan öğretim hedefleri Tablo 8'de verilmiştir (Türkçe terimler Doç. Dr. Halil Tekin'in (1979) orijinal kaynağı Türkçe'ye genişleterek çevirdiği kitabından alınmıştır). Öğretim hedeflerine ilişkin özellikler sunulurken günümüz okullarına ve uygulayıcılara bazı önerilerde de bulunulmuştur.

Tablo 8

Duyuşsal Alan Öğretim Hedeflerinin Düzeyleri, Kategorileri ve Özellikleri

Düzeyler	Kategoriler	Özellikleri
1. Alma (Receiving)	1.1 Farkında olma 1.2 Almaya açıklık 1.3 Seçici dikkat	Bu düzeydeki davranışlar, öğrencinin belli olaylara ya da uyaranlara dikkat etme isteğiyle ilgilidir. Öğrencinin öğretmenin anlattıklarını dikkatle dinlemesi bu düzeye bir örnektir. *Okullara uygulama önerisi:* Dijital dikkat dağınıklığına karşı odaklanma becerisi geliştirmek.
2. Tepkide Bulunma (Responding)	2.1 Tepkide uysallık 2.2 Tepkiye isteklilik 2.3 Tepkide doyum	Bu düzeydeki davranışlar, öğrencilerin etkin katılımıyla ilgilidir. Öğrenci belirli bir olguya dikkat eder ve tepki verir. Öğrenme çıktıları; yanıt vermede uyum sağlamayı, yanıt verme isteğini ya da yanıt vermede doyum sağlamayı vurgular. Öğrencinin sınıf tartışmalarına katılması, sorular sorması bu düzeye bir örnektir. *Okullara uygulama önerisi:* Online öğrenme platformlarına etkin katılım ve görüşlere yorum yapmak
3. Değer Verme (Valuing)	3.1 Bir değeri kabullenme 3.2 Bir değeri yeğleme 3.3 Bir değere adanmışlık	Bu düzeydeki davranışlar, bir bireyin belirli bir nesneye, olguya ya da davranışa yüklediği değerle ilgilidir. Bu özellikler, basit bir kabullenmeden karmaşık bir bağlılık durumuna kadar uzanır. Değer verme, belirli bir değerler kümesinin içselleştirilmesine dayanırken, bu değerlere dair ipuçları öğrencinin açık davranışlarında ifade edilir ve genellikle tanımlanabilir. Öğrencinin çevreyi koruma fikrini benimseyip savunması bu düzeye bir örnektir. *Okullara uygulama önerisi:* Sosyal sorumluluk projelerine katılım ve sürdürülebilirlik bilincine sahip olmak.

Tablo 8 (Devamı)

Duyuşsal Alan Öğretim Hedeflerinin Düzeyleri, Kategorileri ve Özellikleri

Düzeyler	Kategoriler	Özellikleri
4. Örgütleme (Organization)	4.1 Bir değerin kavramsallaştırılması 4.2 Bir değer sistemi örgütleme	Bu düzeydeki davranışlar, farklı değerleri karşılaştırma, aralarındaki çatışmaları çözme ve benzersiz bir değer sistemi oluşturarak değerleri önceliklerine göre düzenlemeyle ilgilidir. Vurgu, değerleri karşılaştırmak, ilişkilendirmek ve sentezlemek üzerinedir. Öğrencinin akademik başarıyı, etik davranışlarla birlikte değerlendirmesi bu düzeye bir örnektir. *Okullara uygulama önerisi:* Kişisel gelişim planı oluşturma, kariyer hedefleriyle değerleri uyumlaştırma çalışmaları yapmak.
5. Bir Değerler Bütünüyle Nitelenmişlik (Characterization by a Value or Value Complex)	5.1 Genelleşmiş örüntü 5.2 Nitelenme	Bu düzey, davranışlarını kontrol eden bir değer sistemine sahip olmayı içerir. Öğrencinin davranışları; yaygın, tutarlı, içselleştirilmiş, öngörülebilir ve en önemlisi uzun vadeli bir değer sistemi içerir. Öğrencinin dürüstlük ilkesini her ortamda uygulaması bu düzeye bir örnektir. *Okullara uygulama önerisi:* Dijital ve çevrim içi ortamlardaki davranışlarda etik kuralları içselleştirmek ve tutarlı olmak.

Bu konuda yapılan çalışmalar duyuşsal alan davranışlarının ölçülmesinin güçlüğünden söz etmektedir. İlgi, tutum, güdü, değer gibi duyuşsal yapıdaki özellikler neredeyse yüz yıla yaklaşan bir süredir ölçeklerle ölçülmekte ve bireylerin bu tür özellikleri hakkında kararlar verilmesine zemin oluşturmaktadır. Bunun yanında duyuşsal özellikler; gözlem formları, öz değerlendirme formları ve öğrenci günlükleri gibi tipik tepkilere ilişkin bilgi veren yöntemlerle de ölçülmektedir. Kutlu ve Kula Kartal (2018) ile Kutlu ve Altıntaş (2021) duyuşsal yönü bulunan ve *içsel* olarak adlandırılan becerilerin maksimum performans maddeleriyle de ölçülebilirliğini göstermişlerdir.

3.2. Duyuşsal Özellikler ve Ölçülmesi

3.2.1. Tutumlar ve Ölçülmesi

Tutumlar, bireylerin belirli bir psikolojik nesneye, kişiye, gruba, olaya ya da değere karşı geliştirdiği öğrenilmiş eğilimleridir. Araştırmalar, tutumların bireylerin davranışlarını yönlendirmede ve karar verme süreçlerinde kritik rol oynadığını göstermektedir (Kağıtçıbaşı, 1977). Toplumlar, bireylerin sosyal olaylara, kurumlara ve değerlere

yönelik tutumlarını anlamak amacıyla eğitim, siyaset, ticaret, propaganda gibi alanlardaki tutumlarını sistematik ve nicel yöntemlerle belirlemeye çalışmıştır. Bu gereksinim, bu alanda kullanılacak ölçme araçlarının geliştirilmesini zorunlu kılmıştır.

Tutumların bilimsel olarak ölçülebilir yapılar olduğu düşüncesini savunan bilim insanları arasında Gordon Allport önemli bir yere sahiptir. Tutum terimi ilk olarak Allport (1935) tarafından sanat alanında, "eylem için zihinsel olarak hazırlanma" anlamında ve yeteneklerle ilintili olarak kullanılmıştır. Allport, bireylerin kişiliklerinin ve toplumsal rollerinin anlaşılabilmesi için tutumların bilimsel yöntemlerle incelenmesi gerektiğini vurgulamıştır. Kendinden önce yapılmış on altı tutum tanımını inceleyen Allport tutumu "Yaşantı ya da deneyimler yoluyla elde edilen, bireyin ilgili olduğu tüm durumlara, nesnelere verdiği tepkiler üzerinde yönlendirici, dinamik bir etkiye sahip zihinsel ya da sinirsel hazırlık hali" olarak tanımlamıştır (Allport, 1935). Onun bu tanımı, sosyal psikolojide tutum araştırmalarının kuramsal temelini oluşturmuştur.

Tutumların ölçülmesiyle ilgili çabalar ve gelişmeler incelendiğinde bazı önemli yaklaşımlar ön plana çıkmaktadır. Bunlar arasında; Bogardus'un Toplumsal Uzaklık Ölçeği, Thurstone'un Eşit Görünümlü Aralıklar Ölçeği, Likert'in Dereceleme Toplamlarıyla Ölçekleme Tekniği, Guttman'ın Yığışımlı (Birikimli) Ölçekleme Tekniği ve Osgood'un Duygusal Anlam Ölçeği standardize edilmiş ölçme teknikleri olarak anılmaktadır. Bu ölçeklerden Bogardus, Thurstone, Likert ve Guttman ölçekleri cümlelerden/maddelerden, Osgood Duygusal Anlam Ölçeği ise sıfatlardan oluşmaktadır (Tavşancıl, 2002). Aşağıda bu ölçeklerden üçü üzerinde durulmuştur.

Yirminci yüzyılın ilk yarısında Louis Leon Thurstone, tutumların ölçülmesine sistematik bir yaklaşım kazandırmıştır. Tutumun çok boyutlu bir yapı olduğunu savunmasına rağmen, ölçekleme amaçlarına uygun olarak tutumu; "psikolojik bir nesneye yönelen olumlu ya da olumsuz bir yoğunluk sıralaması ve derecelemesi" olarak tek boyutlu tanımlamıştır (Thurstone, 1928; 1931). Thurstone tutumların ölçülmesinde ölçeğin yapısal özellikleri üzerinde çalışmış ve tutumu; "Karşılaştırmalı Yargılar Yasası"na göre ölçeklemiştir. Ancak bu yöntem gözlemciden çok fazla sayıda yargıda bulunmasını gerektirdiğinden (örneğin 130 tutum cümlesi için 8385 karşılaştırma) pek işlevsel bulmamıştır. Bu nedenle, "Ardışık Aralıklar Yöntemi" ile tutum ölçmek için en elverişli olan "Mutlak Yargılarla Ölçekleme" yönteminden yola çıkarak "Eşit Görünen Aralıklar Yöntemi"ni geliştirmiştir (Turgut & Baykul, 1992). Thurstone bu yöntemde (1) *en olumsuz*, (6) *nötr* ve (11) *en olumlu* olmak üzere 11 eşit aralık tanımlamış, bireylerden verilen cümleleri bu aralıklara göre sıralamasını istemiştir.

Emery S. Bogardus 1925 yılında, ulusal ve ırkçı tutumları ölçmek için Bogardus Toplumsal Uzaklık Ölçeği (Bogardus Social Distance Scale) olarak adlandırılan ölçeği geliştirmiştir. Ölçekte yer alan maddeler; çok yakın bir toplumsal ilişkiyi benimseme eğiliminden, uzak bir toplumsal ilişkiden kaçınma eğilimine doğru belirli bir düzende

sıralanmış yedi seçenekten oluşmaktadır (Bogardus, 1925). Bogardus ölçeğinde, birbirini izleyen ölçek maddeleri arasındaki uzaklığın eşit olduğu gibi bir iddia yoktur, bu nedenle ölçek sıralama düzeyindedir.

Rensis Likert ise, beşli derecelendirmeye dayanan ve bireylerin ifadelere katılma derecelerini "*kesinlikle katılmıyorum, katılmıyorum, kararsızım, katılıyorum ve tamamen katılıyorum*" olarak beş düzeyde ölçen "Dereceleme Toplamları ile Ölçekleme" aracını geliştirmiştir (Likert, 1932). Likert tipi ölçekler; yanıtlama kolaylığı, çok sayıda bireye aynı anda uygulanabilirliği ve puanlama yolunun kolaylığı gibi özellikler açısından diğer ölçeklere göre bazı üstünlüklere sahiptir. Bunların yanında Likert'in geliştirdiği dereceleme yöntemi tutum dışında başka psikolojik özelliklerin ölçülmesinde kullanılabilmektedir. Bireyin bir tutum maddesine verdiği tepki "Tamamen Katılıyorum" ise o bireyin aşırı olumlu tutum içinde olduğunu ve maddeden en yüksek puanı alacağını göstermektedir. Eğer birey "Kesinlikle Katılmıyorum" tepkisi verdiyse bu tepki onun olumsuz bir tutum içinde olduğunu ve o maddeden en düşük puanı alacağını göstermektedir. Tezbaşaran (1997) tutumun bir tepkide bulunma eğilimi olduğunu, karar bildirmeyen kararsızım derecesinin ise bir eğilim bildirmediğinden, tutumun ölçülmesinde kullanılamayacağını belirtmektedir.

Tutumlara ilişkin ölçümlerin gelişiminde sosyolojik etkenler belirleyici olmuştur. Bireylerin grup içindeki davranışlarını anlamak ve toplumsal ilişkilerdeki benzerlik ve farklılıkları açıklamak; sosyal psikoloji araştırmalarının temel amaçları arasında yer almaktadır (Ajzen, 2005). Eğitimde ise öğrencilerin öğrenmeye yönelik tutumlarının, başarıları üzerinde etkili olduğu varsayıldığı için öğretmenler bu tutumları anlamak için tutum ölçeklerinden yararlanmaktadır.

Tutumların ölçülmesinde kültürel farklılıklar da dikkate alınmalıdır. Toplumun kültürü, bireylerin yaşam biçimini, olaylara, değerlere ve kurumlara bakışını biçimlendirir. Bu nedenle, bir kültürde geliştirilen tutum ölçeğinin başka bir kültürde doğrudan kullanılması her zaman geçerli sonuçlar vermez. Bu sorun, kültüre duyarlı tutum ölçeklerinin geliştirilmesini zorunlu kılmıştır. (Tavşancıl, 2002; Tezbaşaran, 1997)

Tutumların ölçülmesi yalnızca bireysel özelliklerin belirlenmesinde değil, toplumsal politikaların yönlendirilmesinde de kullanılmaktadır. Propaganda araştırmalarında kitlelerin siyasal ve ideolojik tutumları, eğitim araştırmalarında öğrencilerin derslere yönelik yaklaşımları, sosyal politika çalışmalarında ise bireylerin toplumsal kurumlara güven düzeyi tutum ölçekleriyle ölçülmektedir (Fishbein & Ajzen, 1975). Böylece tutum ölçümleri, bireylerin düşünce, duygu ve davranışsal eğilimlerini bilimsel temelde anlamayı olanaklı kılarak toplumbilimlerin önemli ölçme konularından biri haline gelmiştir.

3.2.2. İlgiler ve Ölçülmesi

Bireylerin ilgi alanlarının belirlenmesi, eğitimde ve mesleki yönlendirme sürecinde kritik bir rol oynamaktadır. Türk Dil Kurumu ilgi kavramını; "bir olay ya da etkinliğe yakınlık duyma, ondan hoşlanma ve ona öncelik tanıma durumu" olarak tanımlamaktadır. Psikolojide ise ilgi, bireyin hoşlandığı, yöneldiği ve tekrar etmekten zevk aldığı etkinliklere karşı duyduğu eğilim olarak açıklanmaktadır. Ayrıca ilgi, bireyin belirli etkinliklere ya da konulara karşı geliştirdiği sürekli tercih ve eğilimler olarak da tanımlanmaktadır. Psikolojik bir özellik olarak ilgi; bir şeye zaman ayırma, o konuda daha fazla bilgi edinmeye istek duyma ve ilgili etkinlikleri yaparken zevk alma durumlarıyla ifade edilmektedir. Eğitimde öğrencilerin ilgi alanlarının belirlenmesi, onların uygun öğrenme ortamlarına yönlendirilmesini ve öğrenmeye yönelik güdülerinin artırılmasını sağlar. Meslek seçimi açısından ise bireylerin ilgi alanlarına göre uygun iş kollarına yönlendirilmesi, iş doyumunun ve toplumsal anlamda üretkenliğin artmasına katkı sağlar.

Yirminci yüzyılın başlarında, ilgilerin ölçülmesine yönelik ilk önemli çaba, Edward Kellog Strong tarafından geliştirilen Strong Mesleki İlgi Envanteri (Strong Vocational Interest Bank -SVIB) olmuştur. 1927 yılında Strong tarafından, özellikle askerlik görevini tamamlayan bireylerin uygun işlere yönlendirilmesi amacıyla geliştirilmiştir. Bu envanter, bireylerin farklı meslek gruplarına yönelik ilgilerini değerlendirmeyi amaçlayarak modern anlamda ilgiye ilişkin ölçümlerinin öncüsü kabul edilmiştir. Bireyler bu envantere yanıt verirken çeşitli mesleklere, etkinliklere, hobilere ve okul derslerine ne kadar ilgi duyduklarını belirtirler. Bu yanıtlar, belirli meslek gruplarıyla karşılaştırılarak bireyin ilgi profili çıkarılarak yönlendirilmesi amaçlanır (Strong, 1945). Strong daha sonra, Strong İlgi Envanteri (Strong Interest Inventory -SII) adını verdiği aracı geliştirmiştir. Bu araç mesleki yönlendirme süreçlerinde yaygın olarak kullanılmıştır (Donnay ve ark., 2005).

Strong ile aynı dönemde G. Frederic Kuder de ilgilerin ölçülmesine yönelik önemli katkılarda bulunmuştur. 1938 yılında geliştirdiği Kuder Tercih Kaydı (Kuder Preference Record) ile, lise ve üniversite öğrencilerinin belirli meslek alanlarına *(açık hava işleri, mekanik, hesaplama, bilimsel, ikna edici, sanatsal, edebi, müzikal, sosyal hizmet, büro işleri)* yönelik genel eğilimlerini belirleyerek eğitim ve kariyer planlamasına rehberlik etmeyi amaçlamıştır. Bu araç, Strong envanterinden farklı olarak özel meslek tercihlerinden çok, daha geniş mesleki ilgi alanlarına odaklanmış ve bireylerin tercihlerini üçlü seçenekler arasından seçmelerini istemiştir. Envanterde bireye, 504 etkinliği içeren 168 adet üçlü grup sunulmakta ve her grupta katılımcıdan hem en çok hoşlandığı etkinliği hem de en az hoşlandığı etkinliği seçmesi istenmektedir. Kuder bu yolla, yanlılığı azaltmayı ve daha güvenilir sonuçlar elde etmeyi hedeflemiştir (Donnay ve ark., 2005).

İlgilerin ölçülmesi konusunda bir diğer önemli katkı da John Holland'ın RIASEC [Realistic (Gerçekçi), Investigative (Araştırmacı), Artistic (Sanatçı), Social (Sosyal), Enterprising (Girişimci), Conventional (Geleneksel)] modelidir. Holland (1997), bireylerin kişilik özellikleri [gerçekçi (yapanlar), araştırmacı (düşünenler), sanatçı (yaratıcılar), sosyal (yardımcılar), girişimci (ikna ediciler) ve geleneksel (organizatörler)] ile meslek seçimleri arasında güçlü bir ilişki bulunduğunu belirtmiştir. Holland'ın RIASEC modeli, mesleki yönlendirme alanında bir dönüm noktası olmuştur ve günümüzde pek çok kariyer danışmanlığı uygulamasında kullanılmaktadır.

İlgilerin ölçülmesi, yalnızca bireysel gelişim açısından değil, ekonomik anlamda verimlilik açısından da büyük önem taşımaktadır. Günümüz iş dünyasında bireylerin uygun meslek alanlarına yönlendirilmesi işgücü verimliliğini ve bireylerin iş doyumunu artırmaktadır (Savickas, 1999). İşe alım süreçlerinde yapılan ilgi ölçümleri, yalnızca bireylerin ilgi ve yeteneklerine uygun bir işe yönlendirilmelerini değil, uzun vadede işe bağlılıklarını öngörmek için de kullanılabilmektedir. Bireylerin ilgi alanları toplumun değerleriyle, sosyoekonomik koşullarıyla ve eğitim politikalarıyla biçimlendiği için kültürel etkenler ilgi ölçümlerinde belirleyici rol oynamaktadır. Bu nedenle farklı kültürlerde geliştirilen ölçeklerin doğrudan başka bir kültürde kullanılması her zaman geçerli sonuçlar vermeyecektir.

3.2.3. Kişilik ve Ölçülmesi

Kişilik, bireylerin çevreleriyle etkileşimlerinde ortaya çıkan, düşünce, duygu ve davranış örüntülerinin görece kalıcı özelliklerini ifade etmektedir. Psikolojide kişilik, bireyin sosyal ilişkilerdeki uyumunu, stresle başa çıkma stratejilerini ve karar verme biçimlerini anlamada temel bir kavramdır (Pervin & Cervone, 2019). Türk Dil Kurumu kişilik kavramını, bir kimseye özgü belirgin özellik, manevi ve ruhsal niteliklerinin bütünü; şahsiyet olarak ve bireyin toplumsal hayatı içinde edindiği alışkanlıkların ve davranışların bütünü olarak tanımlamaktadır. Kişilik ölçümleri, bireylerin güçlü ve zayıf yönlerini anlamada, klinik tanı süreçlerinde ve çalışma yaşamında uygun adayların seçilmesinde kritik bir rol oynamaktadır. Bu açıklamalar doğrultusunda kişilik, bireyin içsel süreçleriyle ilişkili, bireye özgü ve tutarlı davranış kalıplarını içeren bir yapıdır. Bir bireyin kişiliği, onu diğer bireylerden ayıran ve benzersiz kılan özelliklerinin bir bütünüdür.

Kişilik ölçümlerinin bilimsel temelleri, 20. yüzyılın başında Carl Gustav Jung'un kişilik tipolojilerine dayanmaktadır. Jung, içe dönüklük ve dışa dönüklük kavramlarını geliştirerek bireylerin toplumla etkileşim biçimlerini tanımlamıştır (Jung, 1971). Daha sonra Raymond Bernard Cattell, faktör analizi yöntemini kullanarak kişiliğin ölçülebilir boyutlarını ortaya koymuş ve 16 Kişilik Faktörü Envanteri'ni (Sixteen Personality Factor Questionnaire -16PF) geliştirmiştir (Cattell ve ark., 1970). Bu envanter, kişilik ölçümünde çok boyutlu bir yapıyı sistematik biçimde ölçen önemli araçlardan biri olmuştur.

Kişilik ölçümleri alanında en bilinen araçlardan bir diğeri de Minnesota Çok Yönlü Kişilik Envanteri'dir (Minnesota Multiphasic Personality Inventory -MMPI). İlk olarak 1940'larda Starke R. Hathaway ve J. Charnley McKinley tarafından geliştirilen MMPI, klinik psikolojide psikopatolojilerin değerlendirilmesi amacıyla kullanılmıştır (Hathaway & McKinley, 1943). Daha sonra geliştirilen MMPI-2 ve MMPI-2-RF, kişiliğin normal ve anormal boyutlarını da inceleyen kapsamlı ölçme araçları haline gelmiştir. Bu envanter klinik psikoloji alanında; tanı koyma, tedavi planlama ve araştırma süreçlerinde hâlâ yaygın olarak kullanılmaktadır.

21. yüzyılda kişilik ölçümleri büyük ölçüde Beş Faktör Modeli (Five-Factor Model/Big Five Personality Trait Model) çerçevesinde biçimlenmiştir. Bu model; dışadönüklük, uyumluluk, sorumluluk, duygusal denge ve deneyime açıklık boyutları üzerinden kişilik yapısını değerlendirmektedir (McCrae & Costa, 1999). Modelin en önemli katkısı, evrensel ve kültürler arası karşılaştırmalara elverişli bir biçimde kişiliğin ölçülmesine olanak tanımasıdır. Günümüzde birçok kişilik envanteri, Beş Faktör Modeli temelinde geliştirilmektedir.

Kişilik envanterleri ayrıca çalışma yaşamının insan kaynakları süreçlerinde; çalışanların örgütsel uyumlarını, liderlik potansiyellerini ve takım çalışmasına yatkınlıklarını değerlendirmek amacıyla kullanılmaktadır (Ones ve ark., 2007). Bu ölçümler, işe alım kararlarını destekleyerek işe uygun adayların seçilmesine ve örgütsel verimliliğin artırılmasına katkı sağlamaktadır.

Kişilik, sosyokültürel özelliklere bağlı olarak değişkenlik gösterdiği için kişiliğin ölçülmesinde ve değerlendirilmesinde bireylerin toplumla uyum düzeyleri, toplumsal kurallara bağlılıkları ve kültürel değerlerle ilişkileri dikkate alınmaktadır. Örneğin, toplumcu anlayışın ön planda olduğu kültürlerde uyumluluk ve toplumsal ilişkiler öne çıkarken, bireyci anlayışın yaygın olduğu kültürlerde özerklik ve deneyime açıklık daha fazla önem kazanmaktadır (Triandis & Suh, 2002). Bu nedenle bir kültürde geliştirilen kişilik testinin, farklı bir kültürde güvenirliği ve geçerliği yüksek düzeyde uygulanabilmesi uyarlama çalışmalarını zorunlu hale getirmektedir. Sosyokültürel bağlam, kişilik testlerinin yanıtlanma ve yorumlanma biçimini de etkilemekte, bireylerin kendini ifade ediş biçimleri farklılık gösterebilmektedir. Örneğin, Batı kültüründe yetişmiş bireyler kişiliklerine ilişkin özellikleri daha açık ifade edebilmekte, Doğu kültürlerinde ise daha ölçülü yanıtlar verme eğilimi göstermektedirler. Bu durum, testlerden elde edilen sonuçların kültürel farklılıklar ışığında yorumlanmasını zorunlu kılmaktadır (Cheung ve ark., 2011).

3.2.4. Güdünün ve Özyeterliğin Ölçülmesi

Psikolojik ölçme alanında güdü ve özyeterlik kavramları, bireylerin öğrenme süreçlerini ve başarılarını açıklamada önemli bir rol oynamaktadır. Güdü, bireyin davranışlarını başlatan, yönlendiren ve sürdüren içsel ya da dışsal dinamikleri ifade ederken; özyeterlik, bireyin belirli bir görevi başarıyla yerine getirebilme konusundaki inancını

yansıtmaktadır (Bandura, 1997). Eğitimde, bu iki psikolojik özellik hakkında bilgi edinilmesi hem bireyin başarısının altında yatan etkenlerin anlaşılmasında hem de bireyin gelişim süreçlerinin desteklenmesinde önemlidir.

Okul öğrenmelerinde öğrencilerin derslerine neden ve nasıl çalıştıkları, hangi hedeflere yöneldikleri ve öğrenme süreçlerinde ne kadar çaba gösterdikleri güdüsel yapılarla ilişkilidir. Deci ve Ryan'ın (1985) öne sürdüğü öz belirleme kuramı, içsel güdü etkenlerinin dışsal güdü etkenlerine göre öğrenme başarısı üzerinde daha fazla olumlu etkileri olduğunu ortaya koymuştur. Bu doğrultuda geliştirilen ölçme araçlarından elde edilen bulgular da bireylerin öğrenme davranışlarının merak, ilgi ve öz doyum gibi içsel etkenlerle daha çok biçimlendiğini göstermektedir. Güdünün ölçülmesinde sıklıkla kullanılan ölçme araçları arasında Akademik Güdülenme Ölçeği ve Okula Yönelik Güdü Envanteri yer almaktadır.

Özyeterlik, bireylerin performanslarını kestirmede en güçlü değişkenlerden biri olarak kabul edilmektedir. İnsan davranışının önemli bir belirleyicisi olan özyeterlik kavramı, Sosyal-Bilişsel Kuram kapsamında Bandura tarafından tanımlanmıştır. Özyeterlik, bireyin çevresinde olup bitenler üzerinde etkili olabilecek biçimde bir davranışı başlatıp sonuç alıncaya kadar sürdürebileceğine olan inancını ifade eder. Bu anlamda özyeterlik, bireylerin bir göreve yönelik güdülerini, kararlılık düzeylerini ve zorluklarla başa çıkma stratejilerini doğrudan etkilemektedir (Bandura, 1997). Özyeterlik, akademik alanın yanı sıra, sağlık, spor, çalışma yaşamı gibi pek çok alanda birey performansının ve bireydeki davranış değişikliğinin belirleyicisi olarak ifade edilmektedir. Bu alanda yaygın olarak kullanılan araçlardan biri, Schwarzer ve Jerusalem (1995) tarafından geliştirilen Genel Özyeterlik Ölçeği'dir. Eğitimde özyeterlik, öğretmenlerin sınıf yönetimi ve öğretim becerilerinde, öğrencilerin ise öğrenme stratejilerinde belirleyici bir rol üstlenmektedir.

Güdü ve özyeterlik özelliklerinin ölçülmesi, sosyokültürel bağlama göre farklılık göstermektedir. Bireyin ya da bireyselliğin ön planda olduğu kültürlerde bireysel başarı ve bireysel kararlar öne çıkarken, toplumcu kültürlerde grup başarısı ve topluluk desteği bu yapıları biçimlendirmektedir. Bu nedenle güdü ve özyeterlik ölçeklerinin farklı kültürlere uyarlanarak güvenirlik ve geçerlik çalışmalarının yapılması gerekmektedir. Ayrıca, sosyoekonomik farklılıklar da güdü düzeyleri üzerinde etkilidir; düşük gelirli öğrencilerde dışsal güdünün, yüksek gelirli öğrencilerde ise içsel güdünün daha belirgin olduğu gözlenmektedir.

Bilgisayar teknolojilerindeki gelişmeler, öğrencilerin öğrenme ortamlarındaki davranışlarını izleyerek güdüye ve özyeterliğe ilişkin eğilimleri hakkında önemli ipuçları sunmaktadır. Örneğin, öğrencinin ders materyalleriyle etkileşim sıklığı, çevrimiçi tartışmalara katılım düzeyi ve kendisine verilen görevleri tamamlama oranı, güdünün dolaylı göstergeleri olarak değerlendirilmektedir (Siemens & Long, 2011). Benzer biçimde, öğrencilerin özyeterlik algısının düzenli izlenmesi, gerekli desteklerin zamanında belirlenmesine olanak tanımaktadır.

3.2.5. Kaygı ve Ölçülmesi

Kaygı, bireylerin geleceğe yönelik tehdit algıları karşısında yaşadığı duygusal gerilim durumu olarak tanımlanmaktadır. Türk Dil Kurumu kaygıyı; üzücü ya da kötü bir şey olacak korkusundan doğan tedirgin edici duygu; düşünce, tasa, küşüm, endişe, gam olarak ve genellikle kötü bir şey olacakmış düşüncesiyle ortaya çıkan ve nedeni bilinmeyen gerginlik duygusu olarak tanımlamaktadır. Günlük yaşamın doğal bir parçası olan kaygı, belli bir düzeyde bireyleri belli bir işi başarma konusunda güdülerken, aşırı düzeylere ulaştığında bireyde huzursuzluğa, yorgunluğa ve korkuya neden olabilmekte ve başarının gerçekleşmesi önünde engel oluşturmaktadır. Eğitimde sınav kaygısı, çalışma yaşamında performans kaygısı ve toplumsal yaşamda sosyoekonomik kaygılar, bireylerin başarılarını ve yaşam doyumlarını doğrudan etkileyen ögelerdir (Spielberger, 1983). Bu nedenle kaygının ölçülmesi, bireylerin psikolojik durumlarını anlamak, müdahale süreçlerini planlamak ve başarılarını artırmak açısından önemlidir.

Yirminci yüzyılın ortalarında Charles Spielberger, kaygıyı "durumluk" (state) ve "sürekli" (trait) olmak üzere iki boyutta ele almış ve bu temelde Durumluk-Sürekli Kaygı Envanteri'ni (State-Trait Anxiety Inventory -STAI) geliştirmiştir (Spielberger ve ark., 1970). STAI hem klinik psikolojide hem de eğitim araştırmalarında geçerliği ve güvenirliği yüksek bir biçimde kaygının ölçülmesini sağlayarak standart bir araç haline gelmiştir.

Günümüzde kaygı ölçümlerinde öz bildirim yöntemlerinin yanında bireye ilişkin nesnel veriler sundukları için; kalp atış hızı, deri iletkenliği ve kortizol (stres hormonu) düzeyi gibi fizyolojik ve biyolojik ölçütler de kullanılmaktadır. Ancak bu tür ölçümler daha çok araştırma laboratuvarlarında yapılabilmektedir (Beck & Clark, 1997).

Eğitim alanında kaygı ölçümleri, özellikle sınav kaygısının anlaşılmasında kullanılmaktadır. Araştırmalar, yüksek sınav kaygısına sahip öğrencilerin dikkatlerini yoğunlaştırmakta zorlandığını, sınav başarılarının düştüğünü ve güdülerinin olumsuz etkilendiğini göstermektedir (Zeidner, 1998). Bu nedenle rehber öğretmenler ve okul psikologları, öğrencilerin kaygı düzeylerini belirleyerek onlara uygun rehberlik ve destek stratejileri geliştirmektedir.

Çalışma yaşamında iş performansı kaygısı, çalışan verimliliğini doğrudan etkilemektedir. Özellikle yüksek kaygı içeren mesleklerde kaygı ölçümleri, çalışanların tükenmişlik düzeylerini, iş doyumlarını ve verimliliklerini izlemek için kullanılmaktadır. Örneğin, örgütsel psikoloji araştırmaları, kaygının işten ayrılma eğilimiyle ve görev performansıyla yakından ilişkili olduğunu göstermektedir. Bu nedenle iş dünyasında kaygı ölçümleri, insan kaynakları stratejilerinin önemli bir bileşeni haline gelmiştir (LePine ve ark., 2005).

Psikolojik bir özellik olarak kaygının ölçülmesinde sosyokültürel boyut dikkate alınmalıdır. Çünkü bazı toplumlar kaygıyı daha çok somatik belirtilerle açıklarken bazıları bilişsel ya da duyuşsal belirtileri ön plana çıkarmaktadır (Draguns & Tanaka-Matsumi, 2003). Bu nedenle bir kaygı ölçeği farklı bir kültürde kullanılacaksa uyarlama çalışmasının yapılarak ölçeğin güvenirlik ve geçerlik değerlerinin istendik düzeyde olmasının sağlanması gerekmektedir. Türkiye'de STAI'nin yanı sıra sınav kaygısına yönelik özel ölçekler geliştirilmiş ve öğrencilerin performans kaygıları üzerine kapsamlı araştırmalar yapılmıştır (Öner & LeCompte, 1998). Eğitimde sınav kaygısı, iş dünyasında performans kaygısı ve klinik psikolojide anksiyete bozukluklarının değerlendirilmesi, kaygı ölçümlerinin çok boyutlu kullanım alanlarını ortaya koymaktadır.

4. PSİKOMOTOR ÖZELLİKLERİN ÖLÇÜLMESİ

Psikomotor özellikler, bireylerin zihinsel süreçleri ile motor becerilerinin birleşiminden oluşan performanslarını ifade etmektedir. Başka bir anlatımla bu özellikler, bireyin fiziksel hareketlerini, el-göz koordinasyonunu, hızını, çevikliğini ve teknik becerilerini belirleyen özellikleridir. Sanayi devrimiyle birlikte işgücünün niteliğinin artırılması ve özellikle askeri alanda görev dağılımlarının yeteneklere uygun biçimde yapılmasının gerekmesi, psikomotor becerilerin ölçülmesini zorunlu kılmıştır (Fleishman, 1964). Psikomotor özelliklerin ölçülmesi eğitim, sağlık, müzik ve spor alanlarında da yaygın olarak uygulanmaktadır.

Psikomotor özelliklerin değerlendirilmesine yönelik ilk sistematik yaklaşımlardan biri Edward L. Thorndike tarafından geliştirilmiştir. Thorndike (1914), öğrenmenin yalnızca bilişsel değil, psikomotor bir boyutu olduğunu vurgulamış ve bireylerin öğrenme süreçlerinde motor becerilerinin önemine dikkat çekmiştir. Bu yaklaşım, psikomotor ölçümlerin eğitimde ve öğrenme süreçlerinde kullanılmasına zemin hazırlamıştır.

Psikomotor becerileri belirlemeye yönelik ilk ölçme araçları arasında; zamanlama testleri, koordinasyon testleri ve uygulamalı beceri değerlendirmeleri yer almaktadır. Örneğin, Purdue Pegboard Testi, bireylerin el-göz koordinasyonunu ve ince motor becerilerini ölçmede yaygın olarak kullanılan ölçme araçlarından biridir. Bu tür testler, özellikle endüstriyel işlerde çalışanların uygunluğunu belirlemek için geliştirilmiş; daha sonra klinik psikoloji ve rehabilitasyon alanında da kullanılmaya başlanmıştır (Tiffin & Asher, 1948).

Sanayi toplumlarında üretkenlik, yalnızca zihinsel kapasiteye değil; aynı zamanda hız ve fiziksel uyum gerektiren becerilere de bağlıdır. Bu nedenle iş güvenliği, verimlilik ve mesleki uyum açısından psikomotor beceriler kritik görülmüştür (Anastasi & Urbina, 1997). Günümüzde havacılık (pilotaj), tıp (cerrahi) ve spor gibi alanlarda yüksek düzeyde psikomotor beceriler gereklidir ve bu becerilerin değerlendirilmesi özel testlerle yapılmaktadır.

Psikomotor becerilerin ölçülmesinde sosyokültürel etkenler önemli bir rol oynamaktadır. Bazı toplumlarda el becerilerine ve zanaatkârlığa verilen değer, bu alanda yapılan ölçmelerin daha erken gelişmesine yol açmıştır. Buna karşılık akademik başarıya öncelik verilen kültürlerde psikomotor beceriler daha az önemsenmiş, bu da bu türden testlerin gelişimini yavaşlatmıştır. Bu nedenle psikomotor testlerin farklı kültürlerde uygulanabilmesi için norm çalışmaları yapılması gerekmektedir (Özgüven, 2012).

4.1. Psikomotor Alan Hedefleri

Eğitimde psikomotor ölçmeler, özellikle mesleki ve teknik eğitim alanlarında daha sık kullanılmaktadır. Öğrencilerin uygulamalı becerilerini değerlendirmek amacıyla performans görevlerinden, gözlem formlarından ve dereceli puanlama anahtarlarından yararlanılmaktadır. Sağlık alanında psikomotor beceriler, rehabilitasyon programlarında bireylerin ilerlemelerini değerlendirmek için kullanılmaktadır. Spor bilimlerinde ise motor beceri testleri, sporcuların yeteneklerini ve performans düzeylerini belirlemek amacıyla uygulanmaktadır (Magill & Anderson, 2017).

Psikomotor alanla ilgili becerilerin kazanılması, bilişsel ve duyuşsal alanlarda olduğu gibi bir süreç işidir. Psikomotor beceriler de kendi içinde basitten karmaşığa, somuttan soyuta bir sıra izleyerek sınıflandırılmaktadır. Günümüze kadar birçok eğitim bilimci psikomotor alanı sınıflandırmaya çalışmıştır (Harrow, 1972; Linn & Gronlund, 1995; Simpson, 1966). Elizabeth J. Simpson, Benjamin Bloom'un ekolünü devam ettirmiş ve bu alana ait sınıflamayı 1972 yılında tamamlamıştır. Simpson (1966) psikomotor alanı; Bloom ve arkadaşlarının (1956) bilişsel alan sınıflaması ile Krathwohl ve arkadaşlarının (1964) duyuşsal alan sınıflamasını göz önüne alarak davranışların nasıl ve ne yolla başladığına, birbirinin ön koşulu oluş özelliklerine, basitten karmaşıklığına, kolaydan zora gidişine ve bireye kazandırılırken izlenen sıraya bakarak önce beş daha sonra yedi ana düzeye ayırmıştır (Simpson, 1972). Psikomotor alan hedeflerine ilişkin düzeyler ve bu düzeylerin özellikleri Tablo 9'da verilmiştir.

Tablo 9

Psikomotor Alan Hedeflerinin Düzeyleri ve Özellikleri

Düzeyler	**Özellikleri**
Uyarılma (Perception)	Motor etkinliği yönlendirmek için duyusal ipuçlarını kullanma yeteneğidir. Örneğin telleri akort etmeden önce, gitar tellerinin çıkardığı sesleri dinler.
Hazırlık (Set)	Harekete geçmeye hazır olma durumu. Zihinsel, fiziksel ve duygusal hazırlıkları içerir. Bu üç hazırlık, bir kişinin farklı durumlara verdiği tepkiyi önceden belirleyen eğilimlerdir. Beceriyle ilgili alanın model ya da örnekleri belirli süre ve yeterlikte öğrenciye sunulur, öğrenci nasıl yapıldığını izler. Örneğin müzikal bir performansın başlangıcında şefi izlemek ve bir enstrümanı çalmaya hazır duruma gelme işidir.
Yönlendirilmiş Tepki (Guided Response)	Taklit ve deneme-yanılma içeren karmaşık bir beceriyi öğrenmenin erken aşamalarını içerir. Performansın yeterliği çokça uygulama yaparak elde edilir. Beceri haline getirme sürecidir. Beceriyi oluşturan basamakların öğretmen gözetiminde yapılmasını sağlar. Örneğin bir müzik aletini usta gözetiminde çalmaktır.
Mekanik Hale Getirme (Mechanism)	Bu, karmaşık bir beceriyi öğrenmenin ara aşamasıdır. Öğrenilen tepkiler alışkanlık haline gelmiştir ve hareketler bir miktar güvenle ve yeterlilikle gerçekleştirilebilir. Birey beceriyi kendi başına, kimseden yardım almadan, gereken nitelikte ve belirtilen sürede yapar. Örneğin bir müzik aletiyle bir eseri gereken niteliklerde çalmaktır.
Karmaşık ya da Açık Tepki (Complex or Overt Response)	Karmaşık hareket kalıpları içeren motor eylemlerin becerikli biçimde gerçekleştirilmesini içerir. Yeterlik; en az enerjiyle, hızlı, doğru ve son derece koordineli bir performansla gösterilir. Birey bu düzeyde tereddüt etmeden performans sergiler. Örneğin bir müzik aleti çalan sanatçılar bir eseri çalarken farklı yüz ve beden hareketleri gösterirler, çünkü müziğin bütün duygularını içlerinde hissederler, ses kaçırmadan, gerekli duyguyu vererek müziğin akışına kendilerini kaptırırlar.
Uyum (Adaptation)	Beceriler yetkinlik (ustalık) düzeyinde gelişmiştir, becerilerini kimseden yardım almadan kullanır ve karşılaştığı özel durumlara uyacak biçimde hareket kalıplarını değiştirebilir. Örneğin bir grup önünde müzik aleti çalan müzisyenler o an dışsal (gürültü) ya da içsel (heyecan) etkenleri kontrol altında tutarlar.
Yaratabilme (Origination)	Belirli bir duruma ya da özel bir ortama uyacak yeni hareket kalıpları oluşturma işidir. Hem süreçteki işler hem de sonuçtaki çıktılar yüksek düzeyde geliştirilmiş becerilere dayanan yaratıcılığı içerir. Birey bu aşamada yeni ve özgün beceriler geliştirir. Örneğin bir müzisyen uzun yıllar dinlenebilecek yeni bir beste yapabilir ya da başka sanatçılar tarafından da kullanılabilecek yeni bir çalma tekniği geliştirebilir.

Psikomotor becerilerin ölçülmesinde genellikle aşağıdaki üç yaklaşımdan biri kullanılmaktadır. Bunlar sürecin, ürünün ya da sürecin ve ürünün birlikte ele alınarak değerlendirilmesini içerir (Kutlu & Altıntaş, 2015; Micheels & Karnes, 1950, Sönmez, 1985). Bu üç yaklaşımın özellikleri şöyledir:

1. İşin Yapılması Sırasında İzlenen Yolun (Sürecin) Değerlendirilmesi

Öğrenci kendisinden beklenen beceriyi yerine getirirken, öğretmen bu becerinin nasıl yerine getirildiğini adım adım izler ve önceden tanımlanmış olan kendi içinde aşamalı her adıma birer puan verir ya da o becerinin gözlenip gözlenemediğini gözlem formu üzerinde ayrılan yere çarpı (X) işareti koyarak işaretler. Öğrenciden şu davranışları yapması beklenir:

Öğrenci;

- belirlenen yolu ayrıntılarıyla izliyor mu?
- işi yaparken uygun malzemeyi seçiyor mu?
- işi yaparken malzemeyi doğru kullanıyor mu?
- iş ortamını ve malzemeleri temiz tutuyor mu?

2. İşin Bitiminden Sonra Ortaya Çıkan Ürün Kalitesinin ya da Sonucun Değerlendirilmesi

Öğrenci kendisinden istenen beceriyi yerine getirir ve bir ürün ortaya koyar. Öğretmen bu ürünü tanımlanmış ölçütler (doğruluk standartları) açısından inceler ve bu ölçütlere göre ürüne puan verir. Buna göre ürün şu açılardan değerlendirilmelidir:

Ürün;

- istenen özelliklere sahip mi?
- belirlenen ölçülere ya da standartlara sahip mi?
- sahipse ne derece kusursuz olarak sahip?

Ayrıca ürün; güvenirlik ve geçerlik değerleri önceden bilinen, aynı özelliklere sahip bir başka ürünle karşılaştırılarak da değerlendirilebilir.

3. Sürecin ve Ürünün Birlikte Değerlendirilmesi

Öğrenci kendisinden beklenen bir görevi yaparak bir ürün ortaya çıkaracaksa, böyle bir durumda öğretmen hem ürünün ortaya çıkış sürecindeki öğrenci davranışlarını hem de sonunda ortaya çıkan ürünü değerlendirebilir. Bu durumda, süreci ayrı ürünü ayrı puanlaması yerinde olacaktır. Daha sonra isterse bu iki puandan bir toplam puan elde edilebilir.

4.2. Psikomotor Özelliklerin Ölçülmesinde ve Değerlendirilmesinde Kullanılan Araçlar

Psikomotor beceriler, özellikle mesleki eğitim, teknik öğretim ve çalışma yaşamında önemlidir. Eğitimde uygulamalı becerilerin ölçülmesi, öğrencilerin iş başındaki performanslarını, problem çözme yetilerini ve görevleri başarıyla yerine getirme düzeylerini anlamada önemli bir rol oynamaktadır.

Günümüzde okul eğitiminde psikomotor becerilerin ölçülmesinde kullanılan araçlar arasında gözlem formlarının ve kontrol listelerinin yanı sıra, performans görevleri, projeler ve portfolyolar da yer almaktadır. Performans görevleri ve projeler bireylerden gerçek yaşam durumlarını yansıtan bir görevi araştırmasını ya da doğal, toplumsal ve teknik bir sorunu, problem çözme sürecinin aşamalarını kullanarak çözmesini beklerken (Kutlu ve ark., 2024), dereceli puanlama anahtarları bu görev sonunda ortaya konan ürünün belirli ölçütler çerçevesinde nesnel biçimde değerlendirilmesini ve etkili bir geribildirim verilmesini sağlamaktadır (Brookhart, 2013). Gözlem formları, öğretmenlerin ya da değerlendiricilerin bireylerin uygulamalı süreçlerdeki davranışlarını sistematik biçimde kaydetmelerine olanak tanırken (Altıntaş ve ark., 2023) portfolyolar bireylerin çalışmalarına dayalı ürünlerini zaman içinde bir araya getirerek gelişimlerini izleme olanağı sunmaktadır (Kutlu ve ark., 2024). Bu anlamda Türkiye'de, mesleki ve teknik eğitimde psikomotor becerilerin değerlendirilmesinde, performans görevlerinin ve dereceli puanlama anahtarlarının giderek daha yaygın biçimde kullanıldığı görülmektedir.

Psikomotor becerilerin ölçülmesinin ekonomik boyutu da oldukça önemlidir. İşgücünün niteliğini artırmayı hedefleyen toplumlar, çalışanların yalnızca teorik bilgi düzeylerini değil, pratik yeterliklerini de ölçmeye yönelmişlerdir. Özellikle mesleki yeterlik belgeleri ve sertifikasyon süreçleri, bireylerin iş başında gerekli becerilere sahip olduklarını kanıtlamalarını gerektirir (Gonczi, 1994). Bu tür ölçmelere dayalı olarak, yeterlik belgesi almak isteyen bireyler ilgili beceriler açısından ulusal ve uluslararası standartlara uygun biçimde değerlendirilmektedir. Psikomotor becerilerin ölçülmesinde de kültürel özellikler belirleyici bir rol oynamaktadır. Daha önce de vurgulandığı gibi, bazı toplumlarda teorik bilgiye daha fazla önem verilirken, bazılarında uygulamalı beceriler öncelik kazanmaktadır. Örneğin, teknik ve mesleki eğitimin güçlü olduğu ülkelerde performansa dayalı ölçmeler daha önemliyken, sınav odaklı eğitim sistemlerinde uygulamalı beceriler genellikle göz ardı edilmektedir (Darling-Hammond & Adamson, 2014).

Son yıllarda psikomotor becerilerin ölçülmesi, dijital araçların da desteğiyle yeni bir boyut kazanmıştır. Simülasyonlar, sanal laboratuvarlar ve *e*-portfolyolar, bireylerin becerilerini gerçekçi ve etkileşimli ortamlarda göstermelerine olanak tanımaktadır. Özellikle sağlık, mühendislik ve pilotaj gibi yüksek riskli meslek alanlarında simülasyon tabanlı uygulamaya dönük ölçmeler, bireylerin kritik görevlerdeki yeterliklerini güvenli bir ortamda değerlendirme olanağı sunmaktadır (Issenberg ve ark., 2005).

5. OKUL ÖĞRENMELERİNİN ÖLÇÜLMESİ

Okul öğrenmelerinin ölçülmesi, modern eğitim sistemlerinin temel hedeflerinden biridir. Öğrenme süreci yalnızca bilgi aktarımını değil, aynı zamanda beceri ve yetenek/yetkinliklerin geliştirilmesini de kapsamaktadır. Bu nedenle, bu bileşenlerin güvenirliği ve geçerliği yüksek araçlarla ölçülmesi bireysel ve toplumsal düzeyde önemlidir. Eğitim psikolojisi, öğrenme psikolojisi, ölçme ve değerlendirme çalışmaları öğrenme çıktılarının yalnızca akademik bilgi ile sınırlı olmadığını; problem çözme, analitik düşünme, akıl yürütme, eleştirel düşünme ve yaratıcılık gibi üst düzey düşünme becerilerinin de değerlendirme sürecinin kapsamına alınması gerektiğini göstermektedir (Bloom ve ark., 1956; Haladyna, 1997; Kutlu ve ark., 2024; Marzano & Kendall, 2007; Nitko & Brookhart, 2014; Popham 2000).

Yirminci yüzyılın başlarında kısa yanıtlı, çoktan seçmeli, doğru-yanlış maddeleri ölçmelerin standartlaşmasına öncülük etmiştir. Frederick J. Kelly'nin test formatındaki bu yeniliklere olan katkısı, merkezi sınav sistemlerinin yaygınlaşmasına zeminini hazırlamıştır (Kelly, 2001). Ancak bu test maddelerinin yalnızca hatırlama düzeyindeki ölçmelere odaklanması, eğitimde daha karmaşık düşünme süreçlerini gerektiren beceri, yetenek ve yetkinliklerin göz ardı edilmesine yol açmıştır. Bu eksiklik, Benjamin Bloom'un bilişsel alan sınıflaması ve daha sonra gelişen performansa dayalı değerlendirme (performance based assessment) yaklaşımlarıyla aşılmaya çalışılmıştır (Anderson & Krathwohl, 2001; Haladyna, 1997).

Günümüzde okul öğrenmeleri kapsamında yer alan bilgi, beceri ve yetenek/yetkinliklerin; bilişsel olduğu kadar içsel ve kişilerarası becerilerle bağlantısal bütünsellik içinde ilişkilendirilerek ölçülmesi ve belirlenmesi gerekmektedir (Kılıç, 2020; Kutlu & Kula-Kartal, 2018; Kutlu & Altıntaş, 2021; Marzano & Kendall 2007; Marzano & Heflebower 2012). Öğrencilerin yalnızca bilgileri hatırlayıp hatırlamadıklarını değil, o bilgiyi anlama, problem çözme, eleştirel düşünme vb. düzeylerde kullanıp kullanamadıklarını da ölçmek eğitimde kaliteyi artırmanın temel yoludur. Uluslararası öğrenci başarısını belirleme uygulamalarında (PISA, PIRLS, TIMSS vb.) kullanılan maddelerin önemli kısmı okul öğrenmelerinden elde edilen bilgilerin yeni ve gerçek yaşam durumlarında kullanılma gücünü temel becerilerle ilişkilendirerek ölçmektedir (OECD, 2019a). Bu ölçme ve değerlendirme uygulamaları eğitim sistemlerine oldukça önemli geribildirimler sunmaktadır.

Türkiye'deki okullarda öğretmenler, öğrencilerin derslere ilişkin bilgileri öğrenip öğrenmediklerine büyük önem vermekte ve bu nedenle çoğunlukla hatırlama düzeyindeki öğrenme eksikliklerini belirlemeye dönük ölçme uygulamalarına ağırlık vermektedirler. Bu uygulamalardan elde edilen verilere dayalı ders tekrarları yapılmakta, ilgili konuda test maddeleri çözdürülmekte ve tekrar gerektiren ev ödevleri verilmektedir. Ancak bir bilginin kalıcılığını sağlamak, sürekli tekrar edilmesine değil,

gerçek yaşam durumlarında kullanımının sağlanmasına bağlıdır. Öğrenciden bir bilginin derste öğretildiği biçimiyle ya da ders kitabında yazdığı biçimiyle olduğu gibi hatırlamasını beklemek onda ezberleme davranışının pekişmesine neden olacaktır. Öğrenciler bilgilerin kullanım yollarını öğrendikçe ve uzun süreli bellekte diğer bilgilerle bir bağlantısallık içinde ilişkilendirildikçe bilgilerin kalıcılığı söz konusu olabilecektir. Bu nedenle sınıf içi eğitim uygulamalarında açık uçlu maddelerin ve performans görevlerinin kullanılması önemlidir.

Autore ve arkadaşları (2025) tarafından yapılan araştırma; bilgilerin uzun süreli bellekte saklanmasının, bu bilgilerin kullanılabilmesi ve yeni durumlara uyarlanabilmesi açısından yaşamsal önem taşıdığına vurgu yapmaktadır. Ayrıca çalışmada; "unutma"nın bir yok olma ya da silinme olmadığı, bilginin geçici olarak gizli durumda kaldığı belirtilmektedir. Araştırmacılar, sürekli değişen bir ortamda sabit ve değişmez bir belleğin organizma için uyumsuz olacağını, sürekli değişen bir gerçekliğe uygun biçimde yanıt verebilmek için, belirli bilgilerin güncellenmesinin ya da unutulmasının doğal olacağına vurgu yapmaktadır. Bu açıdan bakıldığında, öğrenme eksiklikleri bir unutma olmaktan çok beyin fonksiyonunun yararlı ve uyarlanabilir bir özelliği olarak görülebilecektir. Bu bulguların yakın zamanda bilgilerin ölçülmesine değişik bakış açıları getireceği de kaçınılmazdır.

5.1. Bilgilerin Ölçülmesi

Bilgi, okul öğrenmelerinin en temel yapı taşlarından biridir ve yıllar boyunca eğitimde ölçme ve değerlendirme sürecinin odağında yer almıştır. Sümer, Eski Mısır, Antik Yunan, Antik Roma ve Eski Çin gibi medeniyetler, toplumlarının devamlılığını sağlayacak bireyleri yetiştirmek için okullar kurmuş ve öğrencilerinin başarısını belirleyecek sınav sistemleri geliştirmişlerdir (Kutlu, 2021). İlk sınav sisteminin Eski Çin'de başladığı bilinmektedir ve bu sınavlarda aritmetik, tarım, coğrafya, üretim, hukuk ve askeri konularda bilgilerin ölçüldüğüne vurgu yapılmaktadır (Cohen & Swerdlik, 2009). 1800'lü yıllarda gerçekleşen sanayi devrimi bilgi birikimini hızla değiştirmiştir; ancak okullarda ders kapsamına dayalı bilgi öğretimi ve ölçme anlayışı büyük ölçüde aynı kalmıştır (Saygılı, 2013).

Sanayileşmeyle birlikte sosyoekonomik ve kültürel alanlarda gözlenen değişim bilgilere dayalı ölçmelerin yaygınlaşmasını hızlandırmıştır. 1750'li yıllarda 5-13 yaş grubu çocuklar için eğitimin zorunlu hale getirilmesi, kitlesel eğitim sistemlerinin oluşumuna zemin hazırlamıştır (Melton, 2003). Öğrenci başarısının belirlenmesi ve bu başarıların karşılaştırılması ile işgücü niteliğinin değerlendirilmesi merkezi sınavların geliştirilmesini zorunlu kılmıştır. Bu gelişmeler, eğitimde standartlara ve yeterliklere dayanan yönetim anlayışının temelini oluşturmuştur.

Bilgilere dayalı ölçmelerin kurumsallaşmasında öncü isimlerden biri Frederick J. Kelly'dir. 1914 yılında geliştirdiği çoktan seçmeli test formatı, bilgilerin ölçülmesinde devrim niteliğinde bir yenilik olarak kabul edilmiştir (Kelly, 2001). Çoktan seçmeli testlerin kısa sürede çok sayıda bireye uygulanabilmesi, nesnel biçimde puanlanabilmesi ve geniş kapsamlı içerikleri ölçebilmesi gibi özellikleri, bu testlerin hızla yaygınlaşmasını sağlamıştır. Aynı dönemde Ralph Tyler, eğitimde hedefe dayalı ölçme anlayışını sistemleştirerek ölçme araçlarının öğretim hedefleriyle doğrudan ilişkilendirilmesi gerektiğini vurgulamıştır (Tyler, 1949). Tyler'ın bu yaklaşımı, öğretim programı geliştirme ve değerlendirme süreçlerinde önemli bir anlayış değişikliği yaratmıştır.

Çoktan seçmeli testlerin yanı sıra, bilgiyi ölçmede kullanılan araçlar arasında doğru-yanlış ve kısa yanıtlı maddeler de yer almıştır. Bu test türleri, öğretmenlere geniş öğrenci gruplarının bilgi düzeylerini pratik bir biçimde ölçme olanağı sağlamıştır (Haladyna ve ark., 2002). Bu nedenle 20. yüzyılın ortalarından itibaren eğitim sistemlerinde bu madde türleri yaygın olarak kullanılmaya başlamıştır. Haladyna (1997) bilginin yalnızca gerçekleri, kavramları, ilkeleri ve süreçleri ezberlemeyi içerdiği düşüncesinin bir yanılgı olduğunu belirtmektedir. Bu nedenle erişinin (achievement) "bilgi" boyutu Haladyna sınıflamasındaki bilişsel düzeylerden "anlama" düzeyini oluşturmaktadır.

1950'li yıllar okul öğrenmelerinde bilişsel, duyuşsal ve psikomotor yapıların araştırıldığı önemli bir dönem olmuştur. Bloom ve arkadaşları (1956) tarafından geliştirilen bilişsel sınıflama hem öğretim programlarına hem de ölçme ve değerlendirme süreçlerine bakışı değiştirmiştir. Bloom sınıflamasında yer alan kavrama, uygulama, analiz, sentez ve değerlendirme basamakları, çoktan seçmeli, kısa yanıtlı ve doğru-yanlış maddelerin yapısı nedeniyle büyük ölçüde ölçülememiştir. Bu nedenle bu türden test ve madde formatlarına dayalı ölçme uygulamaları çoğunlukla bilgi (hatırlama) düzeyinde kalmıştır (Kutlu & Altıntaş, 2021).

Yenilenmiş Bloom sınıflamasında; "bilgi" yine en alt basamağı oluşturan "hatırlama" düzeyine karşılık gelmektedir. Hatırlama düzeyindeki öğrenmeler, uzun süreli bellekten bilgilerin öğrenildiği şekliyle geri çağrılmayı gerektirir. Yeni sınıflamada hatırlanacak bilgi olgusal, kavramsal, işlemsel ya da üstbilişsel olabilmektedir (Anderson & Krathwohl, 2001).

Bilgilere dayalı ölçme uygulamalarının kültürel etkileri de dikkat çekicidir. Batı toplumlarındaki yeterlik anlayışı, bireylerin bilgiye dayalı sınavlarla seçilmesi ve sınıflandırılması yönündedir. Buna karşılık bazı Doğu toplumlarında bilgilerin ölçülmesi, daha çok ezbere dayanan sınavlar biçiminde uygulanmış ve öğrencilerin anlama, problem çözme, eleştirel düşünme becerileri ikinci planda kalmıştır (Biggs, 1996). Bu durum, eğitim sistemlerinin kültürel bağlamlara göre farklı ölçme anlayışları geliştirdiğini göstermektedir.

Bilgilerin ölçülmesinde önemli bir gelişme madde bankalarının ortaya çıkmasıdır. 20. yüzyılın ortalarında geliştirilen bu sistemler, öğretmenlere farklı güçlük düzeylerinde ve çeşitli içerik alanlarında madde seçme olanağı sunmuştur. Madde bankaları, ölçme

sürecini daha esnek, veri temelli ve güvenilir duruma getirmiştir (Hambleton & Swaminathan, 1985). Ayrıca MTK'nın gelişimiyle birlikte test maddelerinin psikometrik özellikleri daha hassas biçimde belirlenmiş ve bilgilerin ölçülmesinde bilimsel standartlar güçlendirilmiştir (Embretson & Reise, 2013). Günümüzde öğrenme ve öğretme sürecinde bilgilerin ölçülüyor olması hâlâ vazgeçilmezdir; ancak öğrencilerin öğrenme düzeyinin tam olarak belirlenebilmesi beceri ve yetenek/yetkinlik boyutundaki özelliklerin birlikte ölçülmesiyle olanaklı olacaktır.

5.2. Becerilerin Ölçülmesi

Eğitim sistemlerinde uzun yıllar bilgilerin ölçülmesi ön planda tutulmuş olsa da 20. yüzyılın ortalarından itibaren uygulama ve beceri temelli ölçmelerin önemi artmıştır. Beceriler, bireyin bilgiyi gerçek yaşamda kullanabilme, problem çözme, çözümleme yapma ve eleştirel düşünme kapasitesini ifade etmektedir. Beceriler, bilgilerin daha önce karşılaşılan benzer durumlara aktarılarak uygulanması olarak ifade edilebilir. Ayrıca, sıradan işlem yapma ya da sıradan sorun çözme davranışının üstüne çıkarak bilinen ancak az rastlanılan problem durumlarının çözülmesini de içerirler. Bunun için bir beceri birden fazla bilginin aynı anda ilişkilendirilerek kullanımını içermektedir (Haladyna, 1997; Kutlu ve ark., 2024). Günümüzde sağlık, mühendislik, fen, sosyal ve sanat gibi tüm alanlarda, bireylerin bilme gücünden çok bilgileri kullanarak bir şeyler yapabilme gücü ön plana çıkmıştır (Shavelson ve ark., 1992).

Becerilerin ölçülmesine ilişkin ilk çalışmalardan birini Abraham Flexner gerçekleştirmiştir. Flexner, tıp eğitiminde uygulamalı becerilerin ölçme ve değerlendirme süreçlerinde merkezi bir rol oynaması gerektiğini vurgulamış ve sağlık alanında yaygın olarak kullanılan Nesnel Yapılandırılmış Klinik Sınav (Objective Structured Clinical Examination -OSCE) gibi klinik beceri sınavların gelişmesine öncülük etmiştir (Flexner, 1910). Zamanla gözlem formları, kontrol listeleri, açık uçlu maddeler, göreve dayalı ölçmeler ve laboratuvar uygulamaları beceri ölçümünde kullanılan araçlar olarak geliştirilmiştir. Bu araçlar, öğrencilerin uygulama, çözümleme, değerlendirme ve yaratıcılık gibi becerilerini ölçmeyi hedeflemiştir (Gronlund & Brookhart, 2009). Bunun yanında, günümüz okul eğitiminde, öğrencilerin araştırma yapma, deney yapma, proje hazırlama, sunu yapma, veri toplama, veri çözümleme, yorumlama ve rapor yazma gibi becerilerinin geliştirilmesi anlayışı, eğitimde üst düzey öğrenme çıktılarına odaklanılmasını da sağlamıştır.

Son yıllarda karar verme becerilerinin ölçülmesinde Durumsal Yargı Testleri (Situational Judgment Tests -SJTs) öne çıkmıştır. Gerçek yaşam senaryoları kullanan bu testler, özellikle liderlik, etik karar verme, öğretmen yeterlikleri ve sosyal becerileri belirlemek amacıyla kullanılmaya başlamıştır (McDaniel ve ark., 2001). Ayrıca bu testler, bireylerin yalnızca beceri düzeylerini değil, aynı zamanda değerler ve bağlamsal farkındalık temelinde nasıl karar verdiklerini de değerlendirmektedir.

Liu ve arkadaşları (2023), hızla değişen işgücü bağlamında hangi becerilerin öne çıktığını göstermek ve bu becerilere giden öğrenme yollarını görünür kılmak amacıyla farklı bağlamlarda geliştirilmiş beceri sınıflamalarını karşılaştırmalı olarak sundukları ETS Beceriler Sınıflaması 2025 (ETS Skills Taxonomy 2025) başlıklı çalışmalarını yayımlamışlardır. Bu çalışmalarında; *temel bilişsel*, *ileri bilişsel, içsel, kişilerarası*, *dijital* ve *yaşam boyu öğrenme* olmak üzere geniş bir beceri setinin tanımına yer vermişlerdir. Bunun yanında her bir beceri seti için ölçme ve değerlendirme açısından dikkat edilmesi gereken noktaları belirmişlerdir. Çalışmada ele alınan bu becerilere ek olarak; *uzaktan çalışma*, *eğitilebilir olma (coachability)* ve *bilgisel (sciential) beceriler* gibi güncel gereksinimleri yansıtan yeni beceri alanları da vurgulanmaktadır. Bu geniş beceri seti birbirini içeren özellikler taşısa da günümüzde bireylerden hangi becerilerin geliştirmesinin beklendiğine dikkat çekmesi açısından kapsamlı bir çerçeve sunmaktadır.

Yukarıda sözü edilen bu becerilerin belirlenmesi ve geliştirilmesi kısa yanıtlı, çoktan seçmeli gibi klasik madde türleriyle değil açık uçlu madde, performans görevi ve proje temelli ölçme ve değerlendirme yolları kullanılarak yapılabilecektir. Öğrencilerin uzun süreli bir problem üzerinde çalışmaları, araştırma yapmaları ve özgün ürünler ortaya koymaları, kendilerinden beklenen becerilerin kapsamlı biçimde ölçülmesini sağlayabilecektir (Wiggins, 1998). Bu tür ölçme yaklaşımları, öğrenme sürecine ve bu sürecin sonunda ortaya çıkan ürüne odaklanarak öğrencilerin öğrenme yollarının anlaşılmasına ve etkili geribildirim verilmesine ve becerilerin gelişimine katkı sunacaktır.

Özellikle 20. yüzyılın son çeyreğinden itibaren toplumlar, sosyoekonomik ve kültürel ileri gidişlerinde becerilerin geliştirilmesinin önemi üzerinde durmuşlar, geleceği inşa edecek bireylerin bilişsel, sosyal ve duygusal, davranışsal becerilere sahip olmasına vurgu yapmışlardır (OECD, 2019b). Kutlu ve Altıntaş (2021), 21. yüzyılda öğrencilerin kendi öğrenmelerini yönetebilmelerinin önemine dikkat çekmekte ve bireylerin bu öğrenmeleri bütünsel bir anlayışla yönetebilmeleri için problem çözme, analitik düşünme, akıl yürütme, eleştirel düşünme, yaratıcılık gibi ***bilişsel***; özgüven, özyeterlik, güdülenme, yılmazlık, problem çözmeye açıklık, kararlılık, esneklik, duyguları anlama gibi ***içsel***; iş birliği yapma, iletişim kurma, uyum, bireysel farklılıklara değer verme, sorumluluk alma, ikna etme, hesap verebilir olma, üretkenlik, zaman yönetimi gibi ***kişilerarası*** becerilere sahip olmalarının günlük yaşam ve çalışma yaşamı için önemine dikkat çekmiştir. Soto ve arkadaşları (2021), yaşamdaki başarının bilişsel yeteneklerden daha fazlasına bağlı olduğunu, başarıyı sosyal katılım, iş birliği, özyönetim, duygusal dayanıklılık ve yenilik gibi sosyal, duygusal ve davranışsal becerilerden de etkilediğini belirtmektedir.

Gelecek zaman diliminde birden fazla yeterlik alanını kapsayan becerilerin bağlantısal bütünsellik içeren bir anlayışla ölçülmesi, bireyler hakkında daha gerçekçi ve kapsamlı bilgiler elde edilmesini sağlayacaktır.

5.3. Yetkinliklerin (Üst Düzey Düşünme Becerilerinin) Ölçülmesi

Yirmi birinci yüzyılın eğitim anlayışı, öğrencilerin çağın gereksinimlerine uygun olarak problem çözme, eleştirel düşünme, yaratıcılık, iş birliği ve iletişim gibi üst düzey düşünme gerektiren becerilere sahip olmasını öngörmektedir (Trilling & Fadel, 2009). Bu doğrultuda, eğitim sistemlerinde klasik bilgilere dayalı ölçmelerden farklı olarak, öğrencilerin bu yetkinlikleri ne ölçüde kazanabildiklerini belirlemeye yönelik ölçme araçları geliştirilmiştir.

Yetkinliklerin ölçülmesine ilişkin kuramsal temellerin en önemlilerinden biri Benjamin Bloom tarafından 1956 yılında ortaya konulmuştur. Bloom'un sınıflamasında bilişsel alan; bilgi, kavrama ve uygulamanın ötesine taşınarak analiz, sentez ve değerlendirme gibi üst düzey becerileri içermektedir (Bloom ve ark., 1956). İlerleyen yıllarda Anderson ve Krathwohl'un editörlüğünde yenilenen bilişsel alan sınıflamasında sentez ile değerlendirme basamağı yer değiştirmiş, sentez yerine "yaratma" basamağı eklenmiş ve öğrenmenin en üst düzeyi olan bu basamakta bireylerden özgün ürünler ortaya koymaları beklenmiştir. Ayrıca her bir basamak için beceriler daha net tanımlanmıştır (Anderson & Krathwohl, 2001).

Yetkinliklerin ölçülmesinde öne çıkan ölçme ve değerlendirme yaklaşımlarından biri "performans görevi"dir. Performans görevleri sırasında öğrenciler; görevin belirli bir kalitede ve standartta tamamlanmasını sağlayacak birden fazla beceriyi birbiriyle ilişkilendirirler, öğrendiklerini günlük yaşama uygularlar, ezberden uzak biçimde yaratıcı düşünürler, çalışmaya etkin katılım gösterirler ve bu yolla kendi öğrenmelerini yönetirler. Bu süreçte öğrenciler; özgün bir konuda araştırma yaparak, doğal ya da toplumsal bağlamda yer alan gerçek yaşam problemlerini çözmeye çalışarak yeni bir görüş, plan ya da alternatif çözüm önerileri geliştirirler. Ayrıca sürecin sonunda öğrencilerden somut bir ürün ortaya koymaları da beklenir (Kutlu ve ark., 2024). Performans görevlerinin en önemli üstünlüklerinden biri öğrencinin süreç boyunca gösterdiği performansın farklı beceriler açısından izlenmesine olanak tanımasıdır (Wiggins, 1998). Bu nedenle puanlama sürecinde öğreticiler, geleneksel puanlamanın ötesinde birden fazla ölçütü ve göstergeyi dikkate alan dereceli puanlama anahtarlarını kullanırlar. Ayrıca bu süreçte, öğrencilerin kendilerini değerlendirebilecekleri öz değerlendirme, arkadaşları tarafından değerlendirilebilecekleri akran değerlendirme ve grup çalışmalarında ise grup değerlendirme formları kullanılmaktadır. Bu çoklu değerlendirmeler sayesinde öğrencilere; etkili, yapıcı ve geliştirici bir geribildirim verme zemini de oluşmuş olur.

Yetkinliklerin ölçülmesinde dikkat çeken bir diğer performansa dayalı değerlendirme yaklaşımı ise "gerçek yaşama dayalı durum belirleme (authentic assessment)"dir. Grant Wiggins tarafından geliştirilen bu anlayışta, performans görevlerinin öğrencilerin okul

dışında karşılaşabilecekleri durumlarla tam uyumlu olması gerektiği savunularak öğrenmenin yaşamla bütünleşmesi hedeflenmektedir (Wiggins, 1998). Bu kapsamda öğrencilere verilen görevler, onların gerçek yaşamda karşılaşabilecekleri türden olmalıdır. Örneğin, öğrenci laboratuvar ortamında deney yapmalı, bir gruba müze gezdirmeli, pazarda alışveriş yapmalı, bir müzik grubuyla müzik yapmalı ya da bir tamirat işini gerçekleştirmelidir. Bu yaklaşım, öğrencilerin gerçek yaşam sorunları karşısındaki performansını belirlemeye katkı sağlamaktadır.

Bir diğer önemli ölçme ve değerlendirme yaklaşımı "proje"lerdir. Projelerin temel amacı, öğrencinin kendisi tarafından araştırılarak tanımlanmış yeni bir sorunu çözmesini ya da bir konuyu derinlemesine incelemesi sağlamaktır. Bu süreçte öğrenci, proje konusuyla ilgili yayınları tarar, veri toplar, veriyi sınıflandırır, veriyi çözümler ve elde ettiği sonuçları raporlaştırır. Bu yönüyle proje çalışmaları, bilimsel araştırma yöntemlerinin eğitim ortamına uyarlanmış biçimidir. Proje çalışmalarının en önemli yönlerinden biri özgünlük ilkesidir. Bu süreçte öğrencinin özgünlük ilkesine uygun olarak -bir bilim insanı gibi- kendi bakış açısını, yorumunu ve çözüm önerisini özgürce ortaya koyması beklenir. Bu yaklaşım, ezbere dayalı yöntemlerin ötesine geçerek üretken düşünmeyi özendirir. Disiplinler arası bağlantılar kurarak öğrencinin farklı alanlardaki bilgilerini bütünleştirmesine olanak tanır. Projeler ayrıca; Fen (Bilim), Teknoloji, Mühendislik ve Matematik (Science, Technology, Engineering, Mathematics -STEM) eğitiminde, disiplinler arası öğretimde ve üst düzey becerilerin ölçülmesinde kullanılmaktadır (Thomas, 2000). Bu özellikleriyle ne performans görevleri ne de proje çalışmaları bir ev ödevi değildir; öğrencinin düşünsel gelişimini, akademik derinliğini ve kişisel sorumluluk bilincini artıran çok yönlü birer öğrenme aracıdır.

Yetkinliklerin ölçülmesinde kullanılan önemli araçlardan bir diğeri de "portfolyo"lardır. Öğrencilerin kendi çalışmalarını bir araya getirdiği portfolyolar, onların öğrenme sürecine ilişkin öz değerlendirme yapmalarına, öğretmenlerden geribildirim almalarına ve gelişimlerini izlemelerine olanak tanır (Paulson ve ark., 1991). Portfolyo uygulamaları, öğrencilerin sonuçta ortaya koydukları ürünlerin yanında, süreç boyunca edindikleri deneyimleri de değerlendirmeyi olanaklı kılar. Portfolyolar öğrencilerin tutum, ilgi, güdü, bilgi, beceri ve yetkinlik gibi özelliklerindeki gelişimi ortaya koyan ve bu gelişime katkı sağlayan bir öğrenme ve durum belirleme aracıdır. Bu sayede öğrenci hakkında elde edilen bilgilerin kullanımı sınıf öğretmenleriyle sınırlı kalmaz, rehber öğretmenler tarafından da gelişimsel rehberlik hizmetleri kapsamında etkili biçimde kullanılabilmektedir (Kutlu ve ark., 2024). Tüm bu özellikleriyle yetkinlik temelli ölçmeler, 21. yüzyıl öğrenme anlayışının merkezinde yer almaktadır. Kitabın 7. Bölümünde adı geçen bu uygulamaların detaylarına ayrıca yer verilmiştir.

6. SONUÇ

Psikolojik özelliklerin ölçülmesi, bireyin doğrudan gözlenemeyen bilişsel, duyuşsal ve psikomotor yönlerinin bilimsel yöntemlerle görünür duruma getirilmesini amaçlayan çok boyutlu bir süreçtir. Bu süreçte ölçmeye konu olan psikolojik özelliğin ölçülebilir, gözlenebilir ve işevuruk biçimde tanımlanması, uygun ölçme araçlarıyla karşılaştırılması ve elde edilen puanların geçerlik ve güvenirlik kanıtlarının sunularak etik standartlara uygun biçimde yorumlanması önemlidir. Bilişsel özelliklerin ölçülmesinde öğrenme hedefleri, bilgi türleri ve beceri düzeyleri; duyuşsal özelliklerin ölçülmesinde tutum, ilgi, güdü, kişilik ve kaygı gibi yapıların kültürel bağlamla ilişkisi; psikomotor özelliklerin ölçülmesinde ise performansa dayalı gözlem formları ve uygulamalı değerlendirme yaklaşımları öne çıkmaktadır. Günümüzde bilgisayar teknolojilerine dayalı ölçme ortamları, yapay zekâ destekli sistemler ve öğrenme süreçlerinden elde edilen veriler psikolojik özelliklerin daha ayrıntılı, dinamik ve içeriğe duyarlı biçimde değerlendirilmesine olanak sağlamaktadır. Bununla birlikte, teknolojik olanakların artması ölçme sonuçlarına ait geçerlik ve güvenirlik değerlerinin daha fazla yükseleceği ya da sonuçların daha adil olacağı anlamına gelmemektedir. Bu nedenle psikolojik ölçmelerin geleceği, yalnızca yeni araçların geliştirilmesine değil, bu araçların kuramsal temellere, psikometrik kanıtlara, etik ilkelere ve bireyin yararını önceleyen sorumlu bir kullanım anlayışına dayandırılmasına bağlıdır.

KAYNAKÇA | REFERENCES

Ajzen, I. (2005). *Attitudes, personality, and behavior* (2. baskı). Open University Press.

Al-Azawei, A., Serenelli, F. & Lundqvist, K. (2016). Universal design for learning (UDL): A content analysis of peer-reviewed journal papers from 2012 to 2015. *Journal of the Scholarship of Teaching and Learning, 16*(3), 39-56. https://doi.org/10.14434/josotl.v16i3.19295

Allport, G. W. (1935). *Attitudes*. https://eclass.uowm.gr/modules/document/file.php/NURED262/ΠΡΩΤΟΤΥΠΑ%20ΑΡΘΡΑ/Allport%20GW%20attitudes%201935%20Murchison%20chapter.pdf

Altıntaş, Ö., Akaltan, K. F. & Kutlu, Ö. (2023). Diş hekimliği öğretiminde öğrenci başarısının belirlenmesi. Ö. Kutlu, Ö. Altıntaş & K. F. Akaltan (Ed.), *Diş hekimliğinde klinik öncesi ve klinik becerilerin ölçülmesi ve değerlendirilmesi* (s. 1-30) içinde. Ankara Üniversitesi Basımevi. ISBN: 978-605-136-733-0

Anastasi, A. & Urbina, S. (1997). *Psychological testing* (7. baskı). Prentice Hall.

Anderson, L. W. & Krathwohl, D. R. (Ed.). (2001). *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives*. Longman.

Anderson, L. W. & Krathwohl, D. R. (Ed.). (2001). *Öğrenme, öğretim ve değerlendirme ile ilgili bir sınıflama* (D. A. Özçelik, Çev.). Pegem Akademi. (Orijinal eser 2001 yılında yayımlanmıştır).

Autore L., Drew, M. R. & Ryan, T. J. (2025). The cost of remembering: engram competition as a flexible mechanism of forgetting. *Trends in Neurosciences, 48*(19), 728-738. https://www.cell.com/trends/ neurosciences/fulltext/S0166-2236(25)00153-5

Bandura, A. (1997). *Self-efficacy: The exercise of control*. W. H. Freeman.

Beck, A. T. & Clark, D. A. (1997). An information processing model of anxiety: Automatic and strategic processes. *Behaviour Research and Therapy, 35*(1), 49-58. https://doi.org/10.1016/S0005-7967(96)00069-1

Bennett, G. K., Seashore, H. G. & Wesman, A. G. (1974). *Differential aptitude tests: Manual*. The Psychological Corporation.

Biggs, J. (1996). Western misperceptions of the Confucian-heritage learning culture. D. Watkins & J. Biggs (Ed.), *The Chinese learner: Cultural, psychological, and contextual influences* (s. 45-67) içinde. CERC & ACER.

Binet, A. & Simon, T. (1916). *The development of intelligence in children* (E. S. Kite, Çev.). Williams & Wilkins Company. https://ia801607.us.archive.org/27/items/developmentofint00bineuoft/developmentofint00bineuoft.pdf

Bloom, B. S., Engelhart, M. D., Furst, E. J., Hill, W. H. & Krathwohl, D. R. (1956). *Taxonomy of educational objectives: The classification of educational goals. Handbook I: Cognitive domain*. David McKay Company.

Bogardus, E. S. (1925). Measuring social distances. *Journal of Applied Sociology 9*(1925), 299-308. https://brocku.ca/MeadProject/Bogardus/Bogardus_1925c.html

Brookhart, S. M. (2010). *How to assess higher-order thinking skills in your classroom*. Alexandria, VA: ASCD.

Brookhart, S. M. (2013). *How to create and use rubrics for formative assessment and grading*. ASCD.

Carson, J. (2007). *The measure of merit: Talents, intelligence, and inequality in the French and American republics, 1750-1940*. Princeton University Press.

Cattell, R. B., Eber, H. W. & Tatsuoka, M. M. (1970). *Handbook for the sixteen personality factor questionnaire (16 PF): In clinical, educational, industrial, and research psychology, for use with all forms of the test*. Institute for Personality and Ability Testing.

Cheung, F. M., van de Vijver, F. J. R. & Leong, F. T. L. (2011). Toward a new approach to the study of personality in culture. *American Psychologist, 66*(7), 593-603. https://doi.org/10.1037/a0022389

Cohen, R. J. & Swerdlik, M. E. (2009). *Psychological testing and assessment: An introduction to tests and measurement* (7. baskı). McGraw-Hill.

Collins, R. (2014). Skills for the 21st Century: Teaching higher-order thinking. *Curriculum & Leadership Journal, 12*(14), front page. http://www.curriculum.edu.au/leader/teaching_higher_order_thinking,37431.html?issueID=12910

Cronbach, L. J. (1990). *Essentials of psychological testing*. Harper & Row.

Darling-Hammond, L. & Adamson, F. (2014). *Beyond the bubble test: How performance assessments support 21st century learning*. Wiley.

Deci, E. L. & Ryan, R. M. (1985). *Intrinsic motivation and self-determination in human behavior*. Springer.

Dewey, J. (1930). *Democracy and education: An introduction to the philosophy of education*. The Macmillan Company. https://dn790004.ca.archive.org/0/items/democracyandeduc00deweuoft/democracyandeduc00deweuoft.pdf

Domino, G. & Domino, M. L. (2006). *Psychological testing: An introduction* (2. baskı). Cambridge University Press.

Donnay, D. A. C., Morris, M., Schaubhut, N. & Thompson, R. (2005). Strong interest inventory manual: Research, development, and strategies for interpretation. Mountain View, CA: CPP.

Draguns, J. G. & Tanaka-Matsumi, J. (2003). Assessment of psychopathology across and within cultures: Issues and findings. *Behaviour Research and Therapy, 41*(7), 755-776. https://doi.org/ 10.1016/S0005-7967(02)00190-0

Ebel, R. L. & Frisbie, D. A. (1991). Essentials of educational measurement (5. baskı). *Standardized intelligence and aptitude tests*. Prentice-Hall of India.

Embretson, S. E. & Reise, S. P. (2013). *Item response theory for psychologists*. Psychology Press.

Fishbein, M. & Ajzen, I. (1975). *Belief, attitude, intention, and behaviour: An introduction to theory and research*. Addison-Wesley.

Fleishman, E. A. (1964). *The structure and measurement of physical fitness*. Prentice Hall.

Flexner, A. (1910). *Medical education in the United States and Canada: A report to the Carnegie Foundation for the Advancement of Teaching*. Carnegie Foundation. https://archive.org/details/medicaleducation00flexuoft

Goleman, D. (1995). *Emotional intelligence*. Bantam Books.

Gonczi, A. (1994). Competency based assessment in the professions in Australia. *Assessment in Education: Principles, Policy & Practice, 1*(1), 27-44. https://doi.org/10.1080/0969594940010103

Gould, S. J. (1996). *The mismeasure of man* (Rev. and expanded ed.). W. W. Norton & Company.

Gronlund, N. E. & Brookhart, S. M. (2009). *Writing instructional objectives* (8. baskı). Pearson.

Haladyna, T. M. (1997). *Writing test items to evaluate higher order thinking*. Viacom Company.

Haladyna, T. M., Downing, S. M. & Rodriguez, M. C. (2002). A review of multiple-choice item-writing guidelines for classroom assessment. *Applied Measurement in Education, 15*(3), 309-333. https://doi.org/10.1207/S15324818AME1503_5

Haladyna, T. M. (2004). *Developing and validating multiple-choice test items* (3. baskı). Lawrence Erlbaum Associates.

Hambleton, R. K. & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Springer.

Harrow, A. J. (1972). *A taxonomy of the psychomotor domain: A guide for developing behavioral objectives.* David McKay Co.

Hathaway, S. R. & McKinley, J. C. (1943). *Minnesota multiphasic personality inventory manual*. University of Minnesota Press.

Helms, J. E. (1992). Why is there no study of cultural equivalence in standardized cognitive ability testing? *American Psychologist, 47*(9), 1083-1101. https://doi.org/10.1037/0003-066X.47.9.1083

Holland, J. L. (1997). *Making vocational choices: A theory of vocational personalities and work environments* (3. baskı). Psychological Assessment Resources.

Holmes, W., Porayska-Pomsta, K. & Holstein, K. (2022). Ethics of AI in education: Towards a community-wide framework. *British Journal of Educational Technology, 53*(4), 816-835. https://doi.org/10.1111/bjet.13174

Issenberg, S. B., McGaghie, W. C., Petrusa, E. R., Lee Gordon, D. & Scalese, R. J. (2005). Features and uses of high-fidelity medical simulations that lead to effective learning: A BEME systematic review. *Medical Teacher, 27*(1), 10-28. https://doi.org/10.1080/01421590500046924

Jung, C. G. (1971). *Psychological types* (H. G. Baynes, Çev., Rev. ed.). Princeton University Press. (Original work published 1923). https://www.jungiananalysts.org.uk/wp-content/uploads/2018/07/C.-G.-Jung-Collected-Works-Volume-6_-Psychological-Types.pdf

Kağıtçıbaşı, Ç. (1977). *İnsan ve insanlar: Sosyal psikolojiye giriş* (2. Baskı). Duran Ofset Matbaacılık.

Kane, M. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50*(1), 1-73. https://doi.org/10.1111/jedm.12000

Kelly, F. J. (2001). The Kansas Silent Reading Test. M. J. Feuer, A. I. Behuniak & R. F. Madigan (Eds.), *Education and the law* (s. 19-25) içinde. University Press of America. (Original work published 1914).

Kılıç, T. (2020). *Yeni bilim: Bağlantısallık yeni kültür: Yaşamdaşlık* (8. Basım). Ayrıntı Yayınları.

Krathwohl, D. R., Bloom, B. S. & Masia, B. B. (1964). *Taxonomy of educational objectives: The classification of educational goals. Handbook II: Affective domain*. David McKay.

Krathwohl, D. R. (2002). A revision of Bloom's taxonomy: An overview. *Theory Into Practice*, *41*(4), 212-218. https://doi.org/10.1207/s15430421tip4104_2

Kutlu Ö. & Kula-Kartal, S. (2018). The prominent student competences of the 21st century education and the transformation of classroom assessment. *International Journal of Progressive Education 14*(6). 70-82. https://doi.org/10.29329/ijpe.2018.179.6

Kutlu, Ö. & Altıntaş, Ö. (2015). *Diş hekimliğinde klinik becerilerin ölçülmesi ve değerlendirilmesi.* Ankara Üniversitesi Basımevi.

Kutlu, Ö. & Altıntaş, Ö. (2021). Psikolojik ölçmelerin kısa tarihi ve 21. yüzyılda sınıf içi durum belirleme anlayışı. *Trakya Eğitim Dergisi - Trakya Journal of Educational, 11*(3), 1599-1620. https://doi.org/10.24315/tred.896121

Kutlu, Ö. (2021). *Geçmişten günümüze testler: Standart başarı testlerinden geniş ölçekli testlere*. *İstanbul Ölçme Değerlendirme Dergisi, 2*, 36-49. İstanbul İl Millî Eğitim Müdürlüğü. https://istanbul.meb.gov.tr/dergi/istanbul_olcme_degerlendirme_2/mobile/ index.html

Kutlu, Ö., Doğan, C. D. & Karakaya, İ. (2024). *Ölçme ve değerlendirme: Performansa ve portfolyoya dayalı durum belirleme* (7. Baskı). Pegem Akademi Yayıncılık.

Kyllonen, P. C. (2012, May 7-8). *Measurement of 21st century skills within the common core state standards* [Paper presentation]. Invitational Research Symposium on Technology Enhanced Assessment. K12 Centre at ETS. https://www.ets.org/Media/Research/pdf/ session5-kyllonen-paper-tea2012.pdf

LePine, J. A., Podsakoff, N. P. & LePine, M. A. (2005). A meta-analytic test of the challenge stressor-hindrance stressor framework: An explanation for inconsistent relationships among stressors and performance. *Academy of Management Journal, 48*(5), 764-775. https://doi.org/10.5465/amj.2005.18803921

Likert, R. (1932). *A technique for the measurement of attitudes*. Archives of Psychology, No. 140. https://archive.org/details/likert-1932/mode/1up

Likert, R. (1974). A method of constructing an attitude scale. Maranell, G.M. (Ed.), *Scaling: A Sourcebook for Behavioral Scientists* (s. 233-243) içinde. Aldine Publishing.

Linn, R. L. & Gronlund, N. E. (1995). *Measurement and assessment in teaching*. Prentice-Hall Inc.

Liu, O. L., Kell, H., Williams, K., Ling, G. & Sanders, M. (2023). ETS skills taxonomy 2025. *Chinese/English Journal of Educational Measurement and Evaluation, 4*(4), 1. https://doi.org/10.59863/NMIE9603

Lemann, N. (1999). *The big test: The secret history of the American meritocracy*. Farrar, Straus and Giroux.

Magill, R. A. & Anderson, D. (2017). *Motor learning and control: Concepts and applications* (11. baskı). McGraw-Hill.

Marzano, R. J. (1992). *A different kind of classroom: Teaching with dimensions of learning*. Alexandria: Association for Supervision and Curriculum Development.

Marzano, R. J., Pickering, D. & McTighe, J. (1993). *Assessing student outcomes: Performance assessment using the dimensions of learning model*. https://files.eric.ed.gov/fulltext/ ED461665.pdf

Marzano, R. J. & Kendall, J. S. (2007). *The new taxonomy of educational objectives* (2. baskı). Corwin Press.

Marzano, R. J. & Heflebower, T. (2012). *Teaching and assessing 21st century skills*. Marzano Research.

McCrae, R. R. & Costa, P. T., Jr. (1999). A five-factor theory of personality. L. A. Pervin & O. P. John (Ed.), *Handbook of personality: Theory and research* (2. baskı, s. 139-153) içinde. Guilford Press.

McDaniel, M. A., Morgeson, F. P., Finnegan, E. B., Campion, M. A. & Braverman, E. P. (2001). Use of situational judgment tests to predict job performance: A clarification of the literature. *Journal of Applied Psychology, 86*(4), 730-740. https://doi.org/10.1037/0021-9010.86.4.730

Melton, J. V. H. (2023). *Absolutism and the eighteenth-century origins of compulsory schooling in Prussia and Austria*. Cambridge University Press. https://books.google.com.tr/books/about/Absolutism_and_the_Eighteenth_Century_Or.html?id=2f69mKhTDNQC&redir_esc=y

Meyer, M. N. (2018). Practical tips for ethical data sharing. *Advances in Methods and Practices in Psychological Science, 1*(1), 131-144. https://doi.org/10.1177/2515245917747656

Micheels, W. J. & Karnes, M. R. (1950). *Measuring educational achievement*. McGraw-Hill.

Mislevy, R. J., Steinberg, L. S., Almond, R. G. & Lukas, J. F. (2014). Concepts, terminology, and basic models of evidence-centered design. J. I. Bennett (Ed.), *Educational measurement* (4. baskı, s. 139-188). ACE/Praeger.

NCLB. (2001). *The no child left behind act of 2001*. https://www2.ed.gov/nclb/overview/intro/execsumm.pdf

Neisser, U., Boodoo, G., Bouchard, T. J., Jr., Boykin, A. W., Brody, N., Ceci, S. J., Halpern, D. F., Loehlin, J. C., Perloff, R., Sternberg, R. J. & Urbina, S. (1996). Intelligence: Knowns and unknowns. *American Psychologist, 51*(2), 77-101. https://doi.org/10.1037/0003-066X.51.2.77

Nitko, A. J. & Brookhart, S. M. (2014). *Educational assessment of students* (7. baskı). Pearson Higher Education.

OECD. (2019a). *PISA 2018 assessment and analytical framework*. OECD Publishing. https://doi.org/10.1787/b25efab8-en

OECD. (2019b). *OECD skills strategy 2019: Skills to shape a better future*. OECD Publishing. https://doi.org/10.1787/9789264313835-en

OECD. (2023a). *PISA 2022 results (Volume I): The state of learning and equity in education*. OECD Publishing. https://doi.org/10.1787/53f23881-en

OECD. (2023b). *PISA 2022 results (Volume II). Learning during - and from - disruption*. OECD Publishing. https://doi.org/10.1787/a97db61c-en

Ones, D. S., Dilchert, S., Viswesvaran, C. & Judge, T. A. (2007). In support of personality assessment in organizational settings. *Personnel Psychology, 60*(4), 995-1027. https://doi.org/10.1111/j.1744-6570.2007.00099.x

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science, 349*(6251), aac4716. https://doi.org/10.1126/science.aac4716

Öner, N. & LeCompte, A. (1998). *Durumluk-sürekli kaygı envanteri el kitabı*. Boğaziçi Üniversitesi Yayınları.

Özgüven, İ. E. (2012). *Psikolojik testler.* Nobel Yayıncılık.

Paulson, F. L., Paulson, P. R. & Meyer, C. A. (1991). What makes a portfolio a portfolio? *Educational Leadership, 48*(5), 60-63. https://files.ascd.org/staticfiles/ascd/pdf/journals/ ed_lead/el_199102_ paulson.pdf

Payne, D. & Kyllonen, P. C. (2012, January 11-13). *The role of noncognitive skills in academic success*. [Paper presentation]. 21st Century Knowledge and Skills: The New Curriculum and the Future of Assessment Center for Enrollment Research, Policy, and Practice (CERPP). Marriott Downtown, Los Angeles, United States. https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.418.8555&rep=rep1&type=pdf

Pervin, L. A. & Cervone, D. (2019). *Personality: Theory and research* (14. baskı). Wiley.

Popham, J. W. (2000). *Modern educational measurement: Practical guidelines for educational leaders* (3. baskı). Allyn and Bacon.

Popham J. W. (2004). *America's "failing" schools: How parents and teachers can cope with no child left behind*. Routledge.

Raven, J. (2000). The Raven's Progressive Matrices: Change and stability over culture and time. *Cognitive Psychology, 41*(1), 1-48. https://doi.org/10.1006/cogp.1999.0735

Rogers, C. R. (1969). *Freedom to learn*. Charles E. Merrill.

Savickas, M. L. (1999). The transition from school to work: A developmental perspective. *The Career Development Quarterly, 47*(4), 326-336. https://doi.org/10.1002/j.2161-0045.1999.tb00741.x

Saygılı, S. (2013). Sanayi toplumundan bilgi toplumuna geçiş sürecinde eğitimde dönüştürücü bir entelektüel olarak öğretmenler. *Uşak Üniversitesi Sosyal Bilimler Dergisi*, Özel Sayı, 263-274. https://doi.org/10.12780/UUSBD186

Schmidt, F. L. & Hunter, J. E. (1998). The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psychological Bulletin, 124*(2), 262-274. https://doi.org/10.1037/0033-2909.124.2.262

Schultz, D. P. & Schultz, S. E. (2011). *A history of modern psychology* (10. baskı). Wadsworth, Cengage Learning.

Schwarzer, R. & Jerusalem, M. (1995). Generalized self-efficacy scale. J. Weinman, S. Wright & M. Johnston (Eds.), *Measures in health psychology: A user's portfolio. Causal and control beliefs* (s. 35-37) içinde. NFER-Nelson.

Shavelson, R. J., Baxter, G. P. & Pine, J. (1992). Performance assessments: Political rhetoric and measurement reality. *Educational Researcher, 21*(4), 22-27. https://doi.org/10.3102/0013189X021004022

Siemens, G. & Long, P. (2011). Penetrating the fog: Analytics in learning and education. *EDUCAUSE Review, 46*(5), 30-40. https://er.educause.edu/articles/2011/9/penetrating-the-fog-analytics-in-learning-and-education

Simpson, E. J. (1966). *The classification of educational objectives, psychomotor domain* (Contract No. OE-5-85-104). University of Illinois, College of Education. https://files.eric.ed.gov/fulltext/ED010368.pdf

Simpson, E. J. (1972). *The classification of educational objectives in the psychomotor domain.* Gryphon House.

Soto, C. J., Napolitano, C. M. & Roberts, B. W. (2021). Taking skills seriously: Toward an integrative model and agenda for social, emotional, and behavioral skills. *Current Directions in Psychological Science, 30*(1), 26-33. https://doi.org/10.1177/0963721420978

Sönmez, V. (1985). *Program geliştirmede öğretmen el kitabı*. Öğretmen Yayınları.

Spearman, C. (1927). *The abilities of man: Their nature and measurement*. Macmillan.

Spielberger, C. D. (1983). *Manual for the state-trait anxiety inventory (STAI)*. Consulting Psychologists Press.

Spielberger, C. D., Gorsuch, R. L. & Lushene, R. E. (1970). *Manual for the state-trait anxiety inventory*. Consulting Psychologists Press.

Strong, E. K. (1945). *Vocational interests of men and women* (2. baskı). Stanford University Press. https://archive.org/details/in.ernet.dli.2015.190642/mode/2up

Tam, V., Patel, N., Turcotte, M., Bossé, Y., Paré, G. & Meyre, D. (2019). Benefits and limitations of genome-wide association studies. *Nature Reviews Genetics, 20*(8), 467-484. https://doi.org/10.1038/s41576-019-0127-1

Tavşancıl, E. (2002). *Tutumların ölçülmesi ve SPSS ile veri analizi*. Nobel Yayıncılık.

Tekin, H. (1979). *Eğitimde ölçme ve değerlendirme* (2. Baskı). Yargı Yayınları.

Terman, L. M. (1916). *The measurement of intelligence*. Houghton Mifflin.

Tezbaşaran, A. (1997). *Likert tipi ölçek geliştirme kılavuzu*. Türk Psikologlar Derneği Yayınları.

Thomas, J. W. (2000). *A review of research on project-based learning*. The Autodesk Foundation.

Thorndike, E. L. (1914). *Educational psychology.* Colombia University Press. (Original work published 1904). https://archive.org/details/educationalpsych01thor/page/n7/mode/2up

Thorndike, R. L. (1982). *Applied psychometrics*. Houghton Mifflin Company.

Thorndike, R. M. & Thorndike-Christ, T. (2017). Psikolojide ve eğitimde ölçme ve değerlendirme (8. Baskı). Çev. Ed. M. Otrar. Nobel Yayın Dağıtım.

Thurstone, L. L. (1928). Attitudes can be measured. *American Journal of Sociology, 33*(4), 529-554. https://doi.org/10.1086/214483

Thurstone, L. L. (1931). The measurement of social attitudes. *The Journal of Abnormal and Social Psychology, 26*(3), 249-269. https://doi.org/10.1037/h0070363

Thurstone, L. L. (1938). *Primary mental abilities*. University of Chicago Press.

Tiffin, J. & Asher, E. J. (1948). The Purdue Pegboard: Norms and studies of reliability and validity. *Journal of Applied Psychology, 32*(3), 234-247. https://doi.org/10.1037/h0061266

Toker, F., Kuzgun, Y., Cebe, N. & Uçkunkaya, B. (1968). *Zekâ kuramları*. MEB Talim ve Terbiye Dairesi Araştırma ve Değerlendirme Bürosu.

Triandis, H. C. & Suh, E. M. (2002). Cultural influences on personality. *Annual Review of Psychology, 53*(1), 133-160. https://doi.org/10.1146/annurev.psych.53.100901.135200

Trilling, B. & Fadel, C. (2009). *21st century skills: Learning for life in our times*. Jossey-Bass.

Turgut, M. F. & Baykul, Y. (1992). *Ölçekleme teknikleri*. ÖSYM Yayınları.

Tyler, R. W. (1949). *Basic principles of curriculum and instruction*. University of Chicago Press.

UNESCO. (2021). *Global report on inclusion in education: All means all*. UNESCO Publishing.

von Davier, M. & Halpin, P. F. (2013). Item response theory. In K. F. Geisinger (Ed.), *APA handbook of testing and assessment in psychology*, *1*, 143-164. American Psychological Association.

Wiggins, G. (1998). *Educative assessment: Designing assessments to inform and improve student performance*. Jossey-Bass.

Zeidner, M. (1998). *Test anxiety: The state of the art*. Plenum Press.

YAZARLAR HAKKINDA | ABOUT AUTHORS

DR. ÖĞR. ÜYESİ ÖMER KUTLU

Ankara Üniversitesi (Emekli) | omerkutlu[at]ankara.edu.tr

ORCID: 0000-0003-4364-5629

Hacettepe Üniversitesi Sosyoloji Bölümü mezuniyeti sonrası Eğitimde Ölçme ve Değerlendirme alanında yüksek lisans ve doktora derecesini tamamlayan akademisyen, kariyeri boyunca test geliştirme ve ulusal eğitim politikaları alanlarında öncü çalışmalar yürütmüştür. ÖSYM ve Ankara Üniversitesi'ndeki akademik geçmişinin yanı sıra Millî Eğitim Bakanlığı bünyesinde OKS ile ABİDE gibi geniş ölçekli testlerin ve illerde kurulan Ölçme Değerlendirme Merkezlerinin yapılanmasına katkı sağlamıştır. Madde yazma, üst düzey düşünme süreçleri ile okuduğunu anlama becerilerinin ölçülmesi ve durum belirleme konularında uzmanlaşmış ve birçok yayın yapmıştır. Kurucusu olduğu Ankara Üniversitesi Ölçme ve Değerlendirme Uygulama ve Araştırma Merkezi bünyesinde uluslararası sınav modelleri geliştirmiştir. Ayrıca sağlık bilimleri alanında klinik ölçme altyapıları kurarak disiplinler arası ve özgün bir yaklaşım sergilemiştir. Ankara Üniversitesi'nin kurumsal eğitim reformları biçimlenmesinde etkin rol oynamıştır. Yetiştirdiği 43 yüksek lisans ve 21 doktora öğrencisiyle akademiye katkı sağlayan isim olmuştur. 2018 yılında öğrencisi Doç. Dr. Selma Şenel'le "YÖK Üstün Başarı Ödülü" ile onurlandırılmıştır. 2024'te emekliye ayrılan akademisyen, günümüzde MEV Koleji Okulları bünyesinde akademik danışmanlık yapmaktadır.

ASST. PROF. DR. ÖMER KUTLU

Ankara University (Retired) | omerkutlu[at]ankara.edu.tr

ORCID: 0000-0003-4364-5629

Following his graduation from the Department of Sociology at Hacettepe University, Asst. Prof. Dr. Ömer Kutlu completed his master's and doctoral studies in Educational Measurement and Evaluation. Throughout his academic career, he has conducted pioneering work in test development and national educational policies. In addition to his academic positions at the OSYM and Ankara University, he contributed to the development of large-scale national assessment programs, including the Secondary Education Institutions Selection Examination (OKS) and the Academic Skills Monitoring and Assessment (ABİDE) project under the Ministry of National Education, as well as to the establishment of Provincial Measurement and Evaluation Centers across Türkiye. His research has focused on item writing, the assessment of higher-order thinking skills and reading comprehension, and large-scale educational assessment, resulting in numerous scholarly publications. As the founding director of the Ankara University Center for Measurement and Evaluation Research and Application, he led the development of international assessment models. He also established clinical assessment infrastructures in the health sciences, demonstrating an original interdisciplinary approach to educational measurement. Furthermore, he played a significant role in shaping Ankara University's institutional educational reforms. Having supervised 43 master's theses and 21 doctoral dissertations, he has made substantial contributions to the development of the field. In 2018, he received the Council of Higher Education (YÖK) Outstanding Achievement Award together with his former doctoral student, Assoc. Prof. Dr. Selma Şenel. After retiring in 2024, he continues to serve as an academic advisor to MEV Schools.

DOÇ. DR. ÖZGE ALTINTAŞ
Ankara Üniversitesi | oaltintas[at]ankara.edu.tr
ORCID: 0000-0001-5779-855X

Doç. Dr. Özge Altıntaş, Ankara Üniversitesi Eğitim Bilimleri Fakültesi Eğitimde Ölçme ve Değerlendirme Anabilim Dalı öğretim üyesidir. Lisans eğitimini Marmara Üniversitesi'nde; yüksek lisans ve doktora eğitimini Ankara Üniversitesi'nde tamamlamıştır. Doktora sonrası çalışmalarını Amerika Birleşik Devletleri'nde bulunan Purdue Üniversitesi'nde Bilgisayar Ortamında Bireye Uyarlanmış Testler üzerine gerçekleştirmiştir. 2015-2020 yılları arasında Ankara Üniversitesi Ölçme ve Değerlendirme Uygulama ve Araştırma Merkezi'nin müdür yardımcılığı görevini yürütmüş, burada uluslararası test çalışmalarına öncülük etmiştir. 2024 yılından beri Ankara Üniversitesi Eğitim Programları Koordinatör Yardımcılığı görevini yürütmektedir. Çalışma konuları arasında; örtük sınıf modelleri, bilgisayar ortamında bireye uyarlanmış testler, geniş ölçekli testler, test eşitleme ve klinik becerilerin ölçülmesi yer almaktadır. Ulusal ve uluslararası saygın dergilerce yayınlanmış makaleleri; ulusal ve uluslararası yayınevlerince basılmış ve editörlüğünü yaptığı kitapları ve kitap bölüm yazarlıkları bulunmaktadır. Ayrıca ulusal ve uluslararası projelerde yürütücü, araştırmacı ve danışman olarak görev almıştır. Çalışmalarıyla Üniversitesinin bilim ve akademi ödülüne layık görülmüştür.

ASSOC. PROF. DR. ÖZGE ALTINTAŞ
Ankara University | oaltintas[at]ankara.edu.tr
ORCID: 0000-0001-5779-855X

Assoc. Prof. Dr. Özge Altıntaş is a faculty member in the Department of Educational Measurement and Evaluation at the Faculty of Educational Sciences, Ankara University. She received her bachelor's degree from Marmara University and completed both her master's and doctoral studies at Ankara University. She conducted her postdoctoral research on Computerized Adaptive Testing at Purdue University, USA. Between 2015 and 2020, she served as Vice Director of the Ankara University Center for Measurement and Evaluation Research and Application, where she led several international assessment projects. Since 2024, she has been serving as Assistant Coordinator of Educational Programs at Ankara University. Her research interests include latent class models, computerized adaptive testing, large-scale assessments, test equating, and the assessment of clinical skills. She has published extensively in prestigious national and international journals and has authored, edited, and contributed chapters to numerous books published by national and international publishing houses. She has also served as a principal investigator, researcher, and consultant in various national and international research projects. Her research has been recognized with Ankara University's Science and Academy Award for her outstanding scholarly contributions.

5. BÖLÜM

TEST GELİŞTİRME SÜRECİ

TEST DEVELOPMENT PROCESS

Prof. Dr. Melek Gülşah ŞAHİN

TEST GELİŞTİRME SÜRECİ

Prof. Dr. Melek Gülşah ŞAHİN
Gazi Üniversitesi

Öz

Eğitimin pek çok aşamasında testler kritik öneme sahiptir. Testlere dayalı olarak alınan kararların geçerliğinin sağlanabilmesi her şeyden önce test geliştirme sürecinin uygun şekilde yürütülmesine bağlıdır. Testler geliştirilirken esnek ve dinamik yapıya sahip bir süreç takip edilir. Bu bölümde geniş ölçekli testler dikkate alınarak test geliştirme aşamaları birbiri ile ilişkili sekiz başlıkta sunulmuştur. Testin amacının belirlenmesiyle başlayan test geliştirme sürecinin ikinci aşamasında testin planlaması gerçekleştirilmekte; bu doğrultuda testin içerik, biçimsel, psikometrik, puanlama, uygulama ve zaman özelliklerine dair planlamalar yapılmaktadır. Plana dayalı olarak maddelerin yazılması, puanlama anahtarlarının oluşturulması ve uzman görüşlerinin alınması ise sürecin üçüncü aşamasını oluşturmaktadır. Uzman görüşlerinden hareketle gerekli düzenlemelerin gerçekleştirilmesinden sonra deneme formu oluşturulup deneme uygulaması yürütülmektedir. Bölüm kapsamında deneme uygulaması yapılırken uyulması gereken kurallar ile olası hata kaynaklarına karşı alınabilecek önlemlere de değinilmiştir. Bu aşamada sınav güvenliğine yönelik alınacak tedbirler de önem arz etmektedir. Dolayısıyla bölümde bu tedbirlerden de bahsedilmiştir. Deneme uygulamasının ardından madde ayırt ediciliği, madde güçlüğü, madde karakteristik eğrileri ve değişen madde fonksiyonu gibi madde istatistiklerinin incelenmesi aşaması gelmektedir. Madde istatistiklerine dayalı olarak seçilen maddeler bir sonraki aşamada belli bilgiler ile birlikte madde bankasına seçilerek testlerde kullanılmak üzere saklanabilmektedir. Nihai formun oluşturulmasıyla test geliştirme süreci sonlandırılmaktadır. Bununla beraber bölümde testin geçme puanlarının ve standartlarının belirlenmesi, testin uygulanması, test sonuçlarının ilanı ve testin teknik raporunun hazırlanması gibi test geliştirildikten sonra gerçekleştirilebilecek uygulamalara da yer verilmiştir.

Anahtar Kelimeler

Test Geliştirme Süreci, Test Özellikleri, Test Şeması, Testin Uygulanması, Test Teknik Raporu

TEST DEVELOPMENT PROCESS

Prof. Dr. Melek Gülşah ŞAHİN
Gazi University

Abstract

Tests are of critical importance at many stages of education. Ensuring the validity of decisions derived from test results depends primarily on the rigorous execution of the test development process. A flexible and dynamic process is followed when developing tests. In this section, the stages of test development are presented as eight interrelated phases, with a particular focus on large-scale tests. The test development process begins with defining the purpose of the test. The second phase involves planning the test, including its content, format, psychometric properties, scoring procedures, administration, and timing. The third phase of the process involves writing test items, creating scoring keys, and obtaining expert opinions in line with the plan. the necessary revisions have been made based on expert opinions, a pilot test is created and administered. This section also addresses the rules to be followed during pilot testing and the precautions needed to minimize potential sources of error. At this stage, measures to ensure test security are also of great importance. Accordingly, these measures are discussed in this section. Following the pilot test, the next phase involves analyzing item statistics, including item discrimination, item difficulty, item characteristic curves, and the differential item function (DIF). Based on these analyses, suitable items are added to the item bank together with their associated metadata for future test development. The test development process culminates in the creation of the final version. In addition, the chapter describes post-development procedures, including establishing passing scores and performance standards, administering the test, reporting the results, and preparing the technical report.

Keywords

Test Development Process, Test Specifications, Test Blueprint, Test Administration, Test Technical Report

1. GİRİŞ

Bireylerin hayatında alınan kararlarda önemli yeri olan testlerin varlığı çok eski zamanlara dayanmaktadır. Bugün olduğu gibi geçmişte de toplumlarda önemli kabul edilen bilgi ve becerilerin değerlendirilmesinde testler kullanılıyordu. Örneğin, Antik Yunanlılar spor ve edebiyat gibi alanlarda en iyi kişileri belirlemek için testlerden yararlanıyordu. Benzer şekilde Çin'de yaklaşık M.Ö 2. yüzyıla kadar dayanan testler memuriyete geçiş için kullanılmakta idi. 1800'lü yıllara gelindiğinde ise testler kişilik özelliklerini ve diğer gizil (örtük) yapıları belirlemek amacıyla kullanılmaya başlanmıştır (Cohen & Swerdlik, 2018; Kane, 2006). Matbaanın icadı, bilgisayarların ortaya çıkışı, internetin kullanımı, yapay zekâ teknolojilerindeki gelişmeler gibi insanlık tarihinde önemli rol oynayan faktörler; testlerin türünü, uygulanma ve değerlendirilme biçimini farklılaştırsa da testlere olan ihtiyaç yüzyıllardır değişmemiştir. Günümüzde sınıf-içi öğrenmelerin değerlendirilmesinde, kademeler arası geçiş sınavlarında, kamu personeli alım süreçlerinde, sertifika sınavlarında ve öğretim programlarının etkililiğinin tespit edilmesi gibi pek çok alanda testlerden yararlanılmaktadır.

Haladyna ve Rodriguez (2013), testleri standart koşullar altında bir özelliğin niteliğini veya derecesini saptamak için kullanılan araçlar olarak tanımlayarak standartlaştırma kavramının test puanlarının yorumlanması açısından önemli olduğunu vurgulamıştır. Standart koşullar altında uygulanması, testlerin geçerliği ve güvenirliği yüksek ölçümler üretebilmesi için kritik öneme sahip olmakla birlikte, tek başına yeterli değildir. Geçerliği ve güvenirliği yüksek ölçümlerin elde edilmesi, her şeyden önce test geliştirme sürecinin titizlikle yürütülmesine bağlıdır.

Test geliştirme süreci, bilimsel bir araştırma projesi hazırlamaya benzemektedir. Baykul (2000) test geliştirmeyi, özellikleri önceden kestirilmesi gereken bir süreç olarak tanımlamaktadır. Bu tanımda geçen "özelliklerini önceden kestirebilme" ifadesiyle iyi bir test planı hazırlanması kastedilmektedir. Dolayısıyla test geliştirmeye, bu süreçte görev alacak kişileri ve izlenecek adımları içeren ayrıntılı bir plan oluşturarak başlanmalıdır. İşlem adımlarının detaylı olarak planlanması ve her bir aşamada izlenecek hususların belirlenmesi, test geliştiricilerinin bu süreçte daha objektif kararlar almasını ve amacına uygun ölçme araçları geliştirmesini sağlamaktadır. Bu nedenle test geliştirme aşamaları ve her bir aşamada dikkat edilmesi gereken hususları bilmek son derece önemlidir.

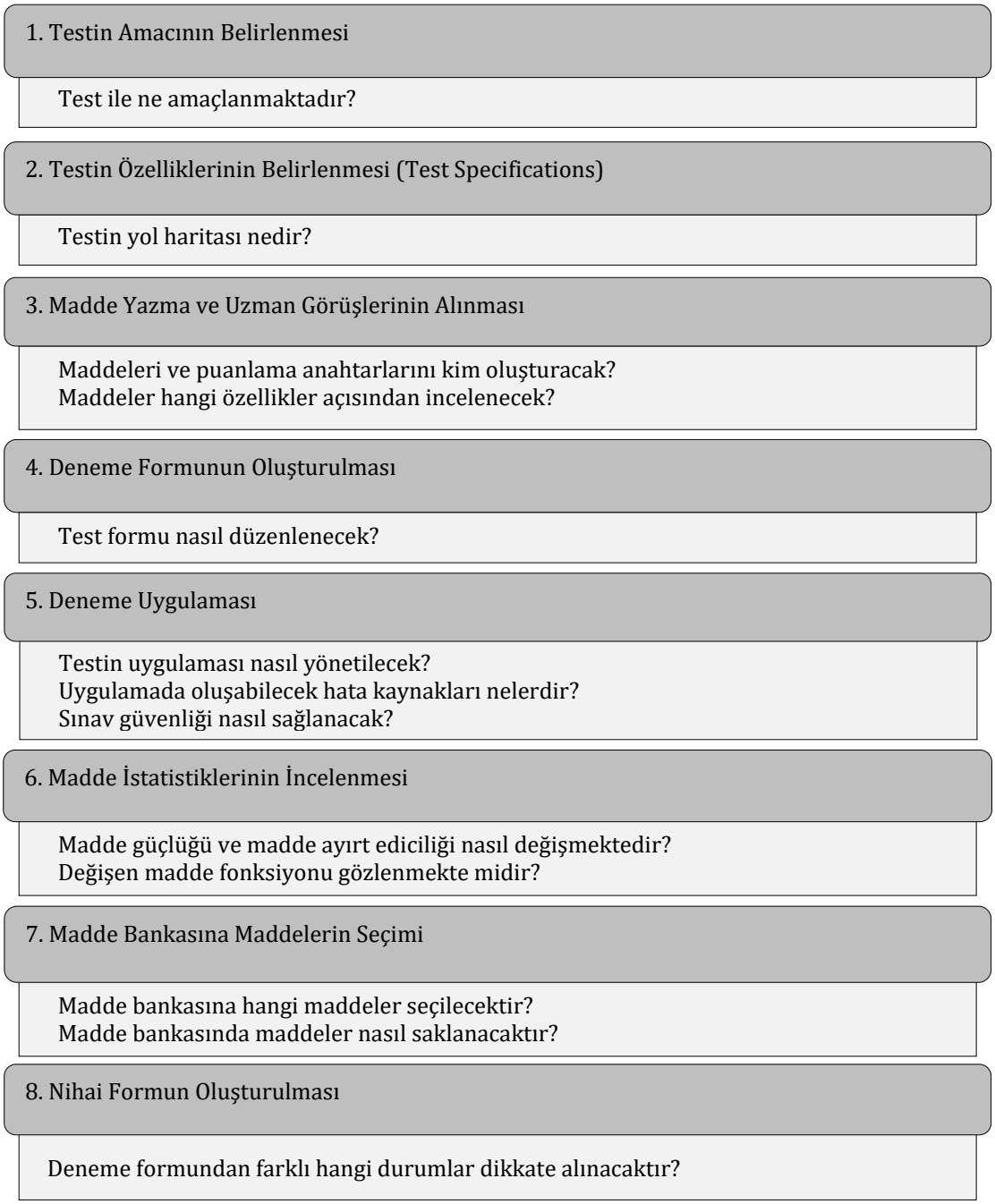
Kitabın bu bölümünde test geliştirme sürecinin esnek ve dinamik bir yapıya sahip olduğu göz önünde bulundurularak test geliştirme aşamaları sunulmuştur. Her bir aşama ayrı başlıklarda detaylandırılmıştır. Buna ek olarak; test geliştirme sürecinden sonra gerçekleştirilebilecek uygulamalara da bölümde kısaca değinilmiştir.

2. TEST GELİŞTİRME AŞAMALARI

Test geliştirme süreci birbirleriyle ilişkili aşamalardan oluşur. Alanyazın incelendiğinde (Eğitim Araştırmaları Birliği (American Educational Research Association-AERA), Amerikan Psikologlar Birliği (American Psychological Association-APA) & Eğitimde Ölçme Ulusal Konseyi (National Council on Measurement in Education-NCME), 2014; Baykul, 2000; Cohen & Swerdlik, 2018; Crocker & Algina, 2008; Çetin, 2019; Doğan, 2019; Downing, 2006; Murphy & Davidshofer, 2001; Özçelik, 1992; Turgut, 1992) bu aşamaların farklı araştırmacılar tarafından değişik şekillerde ele alındığı, genel hatlar aynı kalmakla beraber detaylarda birtakım farklılıklar bulunduğu görülmektedir. Bu bölümde ilgili kaynaklardan hareketle test geliştirme aşamaları organize edilmiştir. Sınıf-içi testler ile geniş ölçekli testlerde test geliştirme sürecinin farklılık gösterebileceği göz önünde bulundurulmuş, bölümde test geliştirme aşamaları geniş ölçekli testler temele alınarak yapılandırılmıştır. Test geliştirme süreci Şekil 1'de, sürecin her bir aşamasına ilişkin anahtar ifadeler ile birlikte özetlenmiştir.

Şekil 1

Test Geliştirme Süreci



2.1. Testin Amacının Belirlenmesi

Test geliştirme sürecindeki ilk basamak testin amacının belirlenmesidir. Testin amacına karar verirken; *i)* Test hangi amaç(lar) için kullanılacak, *ii)* Test ile hangi kararlar alınacak, *iii)* Test hangi yorumları getirecek ve *iv)* Test hangi gruba yönelik olarak geliştirilecek? sorularına cevap aranır. Test geliştirmenin diğer aşamaları testin amacına göre şekillenir. Bu anlamda testin amacını net bir biçimde ortaya koymak, süreci sağlam bir zemine oturtmak anlamına gelir. Baykul (2000) test geliştirmenin genel amaçlarını şu şekilde açıklamaktadır:

- **Psikolojik yapıların betimlenmesi, ilişkilerin incelenmesi:** Testler ilgi, kaygı, motivasyon, öz-yeterlik gibi psikolojik yapıları tanımlamak veya bu yapılar arasındaki ilişkileri ortaya koymak için geliştirilebilir.
- **Seçme ve yerleştirme**: Yükseköğretime geçiş, liselere öğrenci seçimi, bir mesleğe giriş veya bireylerin yabancı dil seviyelerinin tespiti amacıyla gerçekleştirilen testler örnek olarak verilebilir.
- **Öğrenciler hakkında eğitim kararlarının verilmesi:** Öğrencilerin öğrenme eksikleri ile güçlüklerini tespit etme, öğrenme düzeyini saptama ve öğrencileri yönlendirme amacıyla kullanılan testler bu gruba girer.
- **Öğretim ve öğretim programı hakkında kararların alınması**: Öğretim programının etkililiğini tespit etmek ve öğretim hizmetinin niteliğini belirlemek amacıyla gerçekleştirilen testler bu kapsamda yer alır.

Baykul'a (2000) benzer şekilde Grounlund da (1977) testlerin tanıma-yerleştirme, biçimlendirme veya düzey belirleme amacıyla hazırlanabileceğini belirtmiştir. Tanıma-yerleştirmeye yönelik testlerde öğrencinin ön bilgi ve hazırbulunuşluk düzeylerinin belirlenmesi hedeflenir. Bu amaca yönelik hazırlanan testte öncelikle öğrencilerde bulunması beklenen ön koşul davranışların temsil edilmesi önemlidir. Biçimlendirici değerlendirmeye yönelik testlerde amaç, öğrenme eksiklerini ve hatalı öğrenmeleri tespit ederek öğretmen ve öğrencilere öğrenmenin izlenmesine ilişkin geri bildirim vermektir. Dolayısıyla mümkünse bu tür testlerde öğrenme çıktılarının tamamına yer verilmelidir. Düzey belirlemeye dönük testlerde ise not veya sertifika verilmesi esastır. Bu testlerde her bir öğrenme çıktısıyla ilgili madde bulunması genelde mümkün olmaz. Bu nedenle, test öğrenme çıktılarını temsil eden bir madde örneklemi alınarak oluşturulur.

Görüldüğü üzere farklı amaçlara yönelik, çeşitli testler geliştirilebilmektedir. Testin tek ve spesifik bir amaca yönelik geliştirilmesi testten elde edilen sonuçların geçerliği için önemli bir husustur. Geçerlik kanıtları testin amacına göre belirleneceğinden amacın açık biçimde ifade edilmesi geçerlik çalışmalarına da yön verir (AERA, APA & NCME, 2014). Örneğin, amaç bir işe/eğitim kurumuna seçilecek kişileri belirlemekse geçerliğe ilişkin yordayıcı kanıtlar ile ölçme sonuçlarına dayalı geçerlik kanıtlarını incelemek gerekir. Geçerlik kanıtlarına dair detaylı bilgiler kitabın 3. bölümünde sunulmuştur.

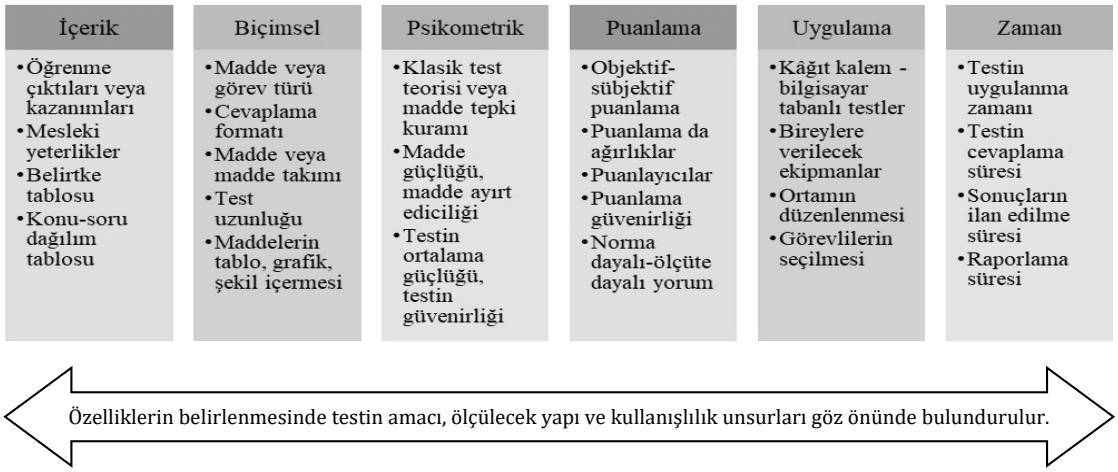
Testin amacı, testte yer alacak maddelerin türünü ve zorluk derecesini de etkiler (Crocker & Algina, 2008; Grounlund, 1977). Örneğin, öğrenme eksiklerinin tespitine yönelik hazırlanmış bir testte açık uçlu maddeler son derece işlevsel olabilirken doğru-yanlış türü maddeler amaca hizmet etmez. Yine çok sayıda başvuran kişi arasından az sayıda kişinin seçileceği bir sınavda, zor maddelerden oluşan bir test amaca hizmet ederken öğrenme güçlüğü yaşayan öğrencileri saptamaya yönelik bir sınavda zor maddeler testin amacı ile bağdaşmaz.

2.2. Testin Özelliklerinin Belirlenmesi (Test Specifications)

Testin özelliklerinin belirlenmesi aşaması, testin inşasına yönelik sistematik bir plan olarak kabul edilebilir. Bu aşamada kararlaştırılan bilgiler test geliştiriciler için kılavuz niteliği taşır. Hedef içeriğin testte nasıl temsil edileceği bu aşamada belirlenir. Test geliştirme süreçlerinin ele alındığı çalışmaların çoğunda bu aşama yer almamakta ya da yüzeysel biçimde sunulmaktadır. Crocker ve Algina (2008), test özelliklerinin belirlenmesini belirtke tablosunun oluşturulması ile sınırlamıştır. Diğer taraftan Downing (2006) test özelliklerinin belirlenmesini, testin özelliklerine yönelik operasyonel (işe vuruk) tanımları içeren daha ayrıntılı bir aşama olarak nitelendirmiştir. AERA, APA ve NCME (2014) tarafından yayınlanan Eğitsel ve Psikolojik Test Standartlarında da testin özelliklerinin belirlenmesi oldukça ayrıntılı tanımlanmış olup bu aşamada pek çok önemli karar alınmasını gerektiğine vurgu yapılmış ve ilgili kararlar amaç, içerik, biçimsel, psikometrik, puanlama ve uygulama (test yönetimi) başlıklarında sunulmuştur. Turgut (1992), testin özelliklerini belirlerken zaman ile ilgili de çeşitli kararlar almak gerektiğine değinmiştir. Test geliştiricisinin sıralanan başlıklarda alması gereken kararlar Şekil 2'de özetlenmiş ve her bir başlık bölümde ele alınmıştır.

Şekil 2

Testin Özelliklerinin Belirlenmesi Aşamasında Üzerinde Durulması Gereken Başlıklar ve Kararlar (AERA, APA & NCME, 2014; Turgut, 1992)



2.2.1. İçeriğin Belirlenmesi

İçerik özelliklerinin belirlenmesinde, yapı veya alanın içerik, beceri, süreç ve tanısal özellikleri kararlaştırılır (AERA, APA & NCME, 2014). Örneğin, yabancı dil seviyesini tespit etmek amacıyla uygulanacak bir sınavın alt testlerinde öğrencilerin İngilizce okuma becerilerinin saptanması amaçlanabilir. Bu amaca ulaşmak için Avrupa Ortak Dil Çerçevesi düzeyleri dikkate alınarak farklı düzeyler için okuma parçaları oluşturulabilir. Başka bir örnek verilecek olursa 6. sınıf öğrencilerinin fen bilimleri dersinde öğrenme eksiklerinin ortaya konulmasının amaçlandığı bir testte içerik öğrenme kazanımlarına göre şekillenmelidir. Kısacası bu aşamada; "Test ile ne ölçülmek istenmektedir, testin kapsamı nedir?" gibi sorulara cevap aranır.

İçerik belirlenirken hedef yapıyla ilgili;

- Görüşler alınabilir ve içerik analizi gerçekleştirilebilir,
- Kaynaklar ve araştırmalar incelenebilir,
- Kritik durumlar belirlenebilir,
- Doğrudan gözlemlere başvurulabilir,
- Öğretim materyalleri, ders kitapları, öğretmen-öğrenci ders notları, öğretim programları ve kazanımları incelenebilir (Baykul, 2000; Crocker & Algina, 2008).

Eğitimde testlerin içeriği öğretim programına veya öğrenme çıktılarına dayalı olarak yapılandırılır. Öğretim programındaki kazanımların / öğrenme çıktılarının dağılımı ile örtüşen bir test geliştirebilmek için genelde ***belirtke tablosu*** veya ***konu-soru dağılım tablosundan*** yararlanılır. Belirtke tablosu, ölçülmesi amaçlanan yapı, davranış veya kazanımların konular ile ilişkilendirildiği bir matristir. Belirtke tablosunda bir boyutta konu/ünite/kazanımlar listelenirken diğer boyutta bilişsel düzeyler bulunur (Turgut, 1992). Tabloda konuların öğretimdeki önemlerine dayalı olarak her bir bilişsel düzeyde yer ve ağırlık verilmelidir (Özçelik, 1992). Bilişsel düzeylerin belirlenmesinde farklı taksonomilerden (Bloom, revize edilmiş Bloom, Barrett, Marzano, Popham, Oosterhof vb.) yararlanılabilir. Hatta öğretim programında o programa özgü bir sınıflandırma söz konusuysa belirtke tablosu hazırlanırken programda tanımlanan bu sınıflandırmadan da yararlanılabilir. Örneğin, Türkiye Yüzyılı Maarif Modeli'nde yer verilen bilişsel düzeyler referans alınarak da bir belirtke tablosu oluşturulabilir. Tablo 1'de ölçme ve değerlendirme dersi ara sınavı için revize edilmiş Bloom taksonomisine göre düzenlenmiş bir belirtke tablosu örneği sunulmuştur.

Tablo 1

Ölçme ve Değerlendirme Dersi Ara Sınav İçin Belirtke Tablosu Örneği

İçerik	Hatırlama	Anlama	Uygulama	Analiz Etme	Toplam
Temel Kavramlar	2 (%8)	3 (%12)	1 (%4)	-	6 (%24)
Ölçme Araçlarının ve Ölçümlerin Psikometrik Nitelikleri	2 (%8)	2 (%8)	2 (%8)	1 (%4)	7 (%28)
Test Geliştirme Süreci	1 (%4)	1 (%4)	-	-	2 (%8)
Geleneksel ve Tamamlayıcı Ölçme Araçları	3 (%12)	4 (%16)	3 (%12)	-	10 (%40)
Toplam	8 (%32)	10 (%40)	6 (%24)	1 (%4)	25 (%100)



- Geniş ölçekli sınavlarda öğretim programındaki tüm kazanımların ölçülmesi zaman ve maliyet açısından pek mümkün olmaz. Bu nedenle test geliştiricinin içerikteki hangi kazanımların daha önemli olduğunu belirlemesi ve içerik dengesini gözeterek maddeleri yapılandırılması gerekir (Crocker & Algina, 2008). Yani kritik davranışlara ve kazanımlara karar verilmelidir.
- Belirtke tablosu hazırlanırken test puanlarına dayalı olarak verilecek kararın önemi de göz önüne alınmalıdır. Test ile alınacak karar ne kadar kritikse belirtke tablosu o kadar ayrıntılı hazırlanmalıdır (Çetin, 2019; Downing, 2006).

Konu-soru dağılım tabloları, ölçmeye konu içeriğin hangi bileşeninin / öğrenme çıktısının testte hangi ağırlıkta temsil edildiğini ortaya koymaya yönelik bir tablodur. Ulusal ve uluslararası sınavlara yönelik kamuoyu ile paylaşılan konu-soru dağılım tabloları bunun tipik bir örneğidir**.** Örneğin Millî Eğitim Bakanlığı, Ölçme ve Değerlendirme Sınav Hizmetleri Genel Müdürlüğü (2025) gerçekleştireceği ortak sınavlar için her bir ders bağlamında oluşturulacak senaryolara ilişkin konu-soru dağılım tabloları yayımlamıştır. Aşağıda 8. sınıf fen bilimleri için hazırlanmış örnek konu-soru dağılım tablosu sunulmuştur.

Tablo 2

*8. Sınıf Fen Bilimleri Dersi Birinci Dönem İkinci Ortak Yazılı Sınavı İçin Örnek Konu-Soru Dağılım Tablosu**

Öğrenme Alanı	Kazanımlar	Madde Sayısı
Dünya ve Evren	F.8.1.1.1. Mevsimlerin oluşumuna yönelik tahminlerde bulunur.	1
Canlılar ve Yaşam	F.8.2.2.2. Tek karakter çaprazlamaları ile ilgili problemler çözerek sonuçlar hakkında yorum yapar.	2
	F.8.2.3.3. Mutasyon ve modifikasyon arasındaki farklar ile ilgili çıkarımda bulunur.	1
Fiziksel Olaylar	F.8.3.1.1. Katı basıncını etkileyen değişkenleri deneyerek keşfeder.	1
	F.8.3.1.2. Sıvı basıncını etkileyen değişkenleri tahmin eder ve tahminlerini test eder.	2
Madde ve Doğa	F.8.4.1.2. Elementleri periyodik tablo üzerinde metal, yarı-metal ve ametal olarak sınıflandırır.	1
	F.8.4.2.1. Fiziksel ve kimyasal değişim arasındaki farkları, çeşitli olayları gözlemleyerek açıklar.	1

**Fen Bilimleri 8. Sınıf Konu Soru Kitapçığı*, Millî Eğitim Bakanlığı (MEB), 2025, (URL-1) adresinden alınmıştır.

Testin içeriğini planlarken belirtke tablosu veya konu-soru dağılım tablosu oluşturmanın dışında daha farklı süreçler de işletilebilir. Örneğin bir test mesleki yeterlik düzeylerini saptamak veya işe alınacak kişileri seçmek amacıyla kullanılacaksa testin içeriğini netleştirirken iş analizine başvurulabilir (AERA, APA & NCME, 2014). Bu kapsamda program geliştirme çalışmalarında ihtiyaç analizi yaparken gerçekleştirilen meslek analizi ile progel - dacum tekniğine benzer adımlar izlenebilir. İş analizinde literatürdeki hazır ölçme ve değerlendirme araçlarından yararlanabileceği gibi her bir mesleğe özgü görev ve faaliyetlerin listesini içeren görev envanterleri de oluşturulabilir. Görev envanterleri meslekteki bireylere uygulanan, görevlerin sıklığı, zorluğu ve önemine yönelik dereceleri içerir (Raymond & Neustel, 2006). Burada başarı testlerinden farklı olarak işin gerektirdiği özellikler net olarak tanımlanmalıdır. Elde edilen veriler üzerinden gerçekleştirilecek içerik analizi mesleğin gereklilikleri ve mesleğe özgü kritik özellikleri saptamaya yardımcı olur.

Testin içeriğinin kim tarafından belirleneceği de önemli bir husustur. Bu aşamada testin türü de etkili olmaktadır. Testin yüksek veya düşük riskli olması önemli bir faktördür. Sınıf içi değerlendirmelerde içerik öğretmen tarafından daha basit ve resmî olmayan biçimde organize edilebilirken, yüksek risk arz eden sınavlarda testin içeriğini daha sistematik biçimde belirlemek gerekir. Bu bakımdan paydaşlardan en az biri için yüksek

risk taşıyan geniş ölçekli testlerde içerik, sınavı uygulayan kurum tarafından oluşturulan komisyonlarca resmî süreçlere dayalı olarak oluşturulmalıdır. Bu tür sınavlarda geçerlik kanıtları için güvenlik ön plana alınarak savunulabilir süreçler işletilmelidir (Çetin, 2019; Downing, 2006).

Belirlenen içeriğin testin amacıyla örtüştüğünü kanıtlayabilmek için geçerlik çalışmaları gerçekleştirilmelidir. Downing (2006) yüksek riskli sınavlarda içeriği oluştururken çok aşamalı bir uzman görüşü alma süreci takip etmeyi önerir. İlk olarak; öğretim programı ve materyalleri ile ders kitapları incelenerek, alan uzmanları ile öğretim üyelerinin, ölçme ve değerlendirme uzmanlarının görüşleri alınarak kapsamlı bir içerik (özellik/kazanım) listesi oluşturulmalıdır. İkinci aşamada ise başka bir uzman grubundan belirlenen test içeriğinin öğrencilerin yeni bilgileri öğrenmesinde veya mesleği icradaki önemini dereceleme ölçekleri aracılığıyla değerlendirmeleri istenebilir. Ardından değerlendirmeyi yapan uzmanların görüşlerinden yola çıkılarak içerik gözden geçirilir. Bu aşamada uzmanlar arasındaki uyumu da dikkate almak gerekir. Uzmanlar arasındaki uyumun hesaplanmasında, uzman sayısı göz önüne alınarak, basit uyum yüzdesinden, Kappa istatistiklerinden (Cohen'in kappası, Fleiss'in kappası, ağırlıklandırılmış kappa) ya da Krippendorf alfa katsayısından faydalanılabilir, genellenebilirlik kuramı ve/veya çok yüzeyli Rasch modeline dayalı analizler gerçekleştirilebilir.

İçeriğin belirlenmesinde geçerliği tehdit edebilecek iki tür sınırlılık söz konusudur: Kapsam eksikliği ve kapsam dışı unsurlar. İlgili yapının/içeriğin eksik temsil edilmesi kapsam eksikliği, yapı ile ilgisiz özelliklerin testin içeriğine karışması ise kapsam dışı unsurlar olarak tanımlanır. Örneğin sözel iletişim becerisini ölçmeye dönük bir test dinleme becerisiyle ilgili maddeler içermiyorsa bu test hedef kapsamı yeterince temsil edemiyor demektir. Bahse konu testte yazma becerisine yönelik maddelerin bulunması ise kapsam dışı özelliklerin teste dâhil edildiği anlamına gelir. Kapsama yönelik bu iki sınırlığın üstesinden gelebilmek için hedef yapının çerçevesi net olarak ortaya konmalı, ölçmeye dâhil edilecek ve hariç tutulacak özelliklerin listesi oluşturulmalıdır (Kline, 2005).

2.2.2. Biçimsel Özelliklerin Belirlenmesi

Biçimsel özelliklerin belirlenmesi aşamasında amaca en uygun madde veya görev türüne karar verilir. Tüm amaca uygun tek bir madde türünden söz etmek olanaksızdır. Alanyazında maddeler veya görevler özelliklerine göre çeşitlilik gösterir ve maddelerin sınıflandırmasına ilişkin birçok farklı yaklaşım vardır (Crocker & Algina, 2008; Osterlind, 2002; Rodriguez, 2002; Turgut, 1992). Genel anlamda maddeler; yanıtı seçmeyi gerektiren ve yanıtı katkı (yapılandırmayı) gerektiren maddeler şeklinde ikiye ayrılabilir. Çoktan seçmeli, doğru yanlış ve eşleştirmeli maddeler yanıtın seçimine dayalı madde türleridir. Şans başarısı içeren ve puanlanışı itibariyle objektif nitelik taşıyan bu maddelerde birey sunulan seçeneklerden birini ya da birkaçını işaretleyerek cevabını verir. Açık uçlu, kısa cevaplı ve boşluk doldurma tipinde maddeler ise yanıtı

yapılandırmayı gerektiren madde türleridir. Bu maddeler bireye cevaplama özgürlüğü tanır ve bu özgürlük yanıtın ne kadar sınırlandırıldığına bağlı olarak değişir. Ayrıca bu maddeler yanıtı seçmeyi gerektiren maddelerin aksine puanlanışı bakımdan subjektif özellik taşır ve şans başarısı içermez. Madde türleriyle ilgili detaylı bilgi edinmek için ilgili kaynaklar ve bu kitabın 6. bölümü incelenebilir.

Testin amacı ve içeriğine yönelik olarak testte tek bir madde türüne yer verilebilir veya test farklı türde maddeler içerecek şekilde hazırlanabilir. Öğrencilerin matematik dersinde problem çözme becerilerini değerlendirmeye yönelik bir test geliştirmek amaçlandığında, test çoktan seçmeli, kısa cevaplı veya açık uçlu maddelerden oluşturulabilir. Madde türü ile sayısına karar verirken hedef kitlenin özellikleri (yaşı, herhangi bir engel durumunun olup olmadığı vb.), grubun büyüklüğü, maliyet ve cevapların ilan edilme süresi gibi değişkenler göz önünde bulundurulmalıdır. Bir mesleğe yerleştirme ya da mesleki yeterlikleri saptama amacıyla yapılan bir sınavda ise mesleki becerileri değerlendirmeye hizmet edecek performans görevleri geliştirilebilir. Bu aşamada sınırlandırılmış ya da sınırlandırılmamış performans görevinin seçimine de karar verilmelidir. Becerinin türü, değerlendirmeye esas ölçütler, zaman ve maliyet unsurları bu seçimde kritik rol oynayacaktır.

Testin amacının veya yapıya ilişkin teorinin başlangıçta net olarak belirlenebildiği durumda, test özelliklerinin her bir aşaması planlanır ve süreçte bu plan ile uyumlu bir ilerleme sağlanabilir. Ancak test geliştirmenin geneli gibi bu süreç de değişmez değildir. Gerekli durumlarda değişikliklere gidilebileceği unutulmamalıdır. Daha açık ifade etmek gerekirse kimi zaman hedef yapının teorik çerçevesi önceden net olmaz, maddelerden edilen verilerden hareketle oluşturulması gerekir. Bu aşamada madde puanlarının ölçülecek yapıyla ilişkili başka özelliklere ait puanlarla korelasyonuna da bakılabilir (AERA, APA & NCME, 2014). Örneğin ölçme ve değerlendirme uzmanlığına yönelik bir mesleki yeterlik sınavında, bazı maddeler bireylerin istatistik özyeterlik ölçeğinden elde ettikleri puanlar ile korelasyonuna dayalı olarak seçilebilir.

2.2.3. Psikometrik Özelliklerin Belirlenmesi

Psikometrik özelliklerin belirlenmesi aşamasında madde ve test istatistiklerine yönelik kararlar alınır. Bu doğrultuda klasik test teorisi bağlamında madde ayırt ediciliği ve madde güçlüğü hesaplanabilir ya da madde tepki kuramı bağlamında madde karakteristik eğrileri incelenebilir. Açık uçlu maddelerden oluşan testlerde madde istatistikleri hesaplanmadan önce puanlamaların güvenirliği de kontrol edilmelidir. Örneğin dört dil becerisinin değerlendirildiği bir yabancı dil sınavında her bir beceride yer alan maddelerin/görevlerin geçerliğini belirlemek amacıyla madde/görev ayırt edicilik değerleri, madde-toplam test korelasyonları hesaplanabilir. Ayrıca konuşma ve yazma becerilerine yönelik puanlar için puanlayıcılar arası güvenirlik katsayıları hesaplanabilir.

2.2.4. Puanlama Özelliklerinin Belirlenmesi

Puanlama özelliklerinin belirlenmesi aşamasında maddelerin nasıl puanlanacağı kararlaştırılır. Elbette bu kararda belirleyici olan madde türüdür. Yanıtı seçmeyi gerektiren maddelerde çoğunlukla 0-1 şeklinde ikili puanlama yapılır. Yanıtı katkı gerektiren maddelerde ise genelde kısmen doğru yanıtların da değerlendirildiği çoklu bir puanlama benimsenir. Bu aşamada madde yazarları, hazırladıkları maddeler için doğru ve kısmen doğru kabul edilecek cevapları belirtmelidir. Becerilerin ölçülmesi söz konusu olduğunda ise performansa dayalı değerlendirme kapsamında puanlama anahtarına (kontrol listesi, dereceleme ölçekleri, rubrik) karar verilmelidir. Örneğin, bir mesleğe yerleştirme ya da mesleki yeterlikleri saptama amacıyla geliştirilen performans görevlerine uygun bütünsel ya da analitik rubrik, genel ya da göreve özel rubriklerden hangisinin kullanılacağına bu aşamada karar verilir. Ayrıca testten alınacak toplam puanda madde ağırlıklarının nasıl yer alacağı bu aşamada kararlaştırılmalıdır. Yine puanlama süreçleri detaylı biçimde planlanmalı; puanlayıcıların nasıl belirleneceği, puanlamanın kaç kişi tarafından, ne zaman ve nasıl gerçekleştirileceği, puanlayıcılar arası uyuşmanın hangi yöntemle hesaplanacağı, olası puanlayıcı yanlılıklarının nasıl önleneceği bu aşamada kararlaştırılmalıdır.

2.2.5. Uygulama Koşullarının Belirlenmesi

Uygulama koşullarının belirlenmesi aşamasında yanıt aranması beklenen sorular Şekil 3'te sunulmuştur.

Şekil 3

Testin Uygulama Koşullarını Belirlerken Yanıt Aranması Gereken Sorular

Testin uygulama formatı nedir?
Test ne kadar süre içerisinde uygulanacaktır?
Bireylere hangi ekipman / ekipmanlar verilecektir?
Sınav yerleşim planı nasıl olacaktır?
Bilgisayar ortamında uygulanacak testler için özel bir yazılım ya da donanım gerekmekte midir?

Uygulama koşullarının belirlenmesi aşamasında ilk olarak testin bireylere hangi şekilde (kâğıt kalem - bilgisayar tabanlı) uygulanacağına karar verilmelidir. Testin uygulamasına kalan süre ve testin uygulamasında bireylere verilecek ekipman / ekipmanlar da uygulama koşulları olarak kararlaştırılması beklenen hususlardır. Bilgisayar tabanlı

testlerin uygulanması söz konusu ise bir diğer husus ise internet bağlantısının gerekliliğinin sorgulanmasıdır. Son olarak sınav yerleşim planı ve görevlilerin tespitine ilişkin karar da bu aşamada alınabilir.

Burada yanıt aranmayan sorular dışında testin formatına dayalı farklı kararlar da söz konusu olabilir. Örneğin, bir yabancı dil seviye tespit sınavının katılımcılara bilgisayar tabanlı uygulanacağını varsayalım. Testin doğrusal mı yoksa bireye uyarlanmış mı olduğu önemlidir. Uyarlanmış testlerde madde seçim algoritmasına da karar vermek gerekir. Maddelerin uygulanmasında göz önünde bulundurulması gereken bir başka kritik husus bireylerin maddeye maruz kalma indeksleridir. Testin uzunluğu doğrusal testlerde sabitken uyarlanmış testlerde sabit veya değişen uzunlukta olabilir. Burada içerik dengelemeye dönük incelemeler de gerçekleştirilebilir. Bilgisayar ortamında bireye uyarlanmış testler ve özellikleri hakkında ayrıntılı bilgi için 15. bölüm incelenebilir.

2.2.6. Zaman Koşullarının Belirlenmesi

Zaman koşullarının belirlenmesi aşamasında testin hazırlanma takvimi, uygulanacağı tarih, uygulamanın kaç saat süreceği, sonuçların ne kadar sürede ilan edileceği ile ilgili planlamalar yapılır. Testin uygulama süresini belirlerken pek çok değişkeni hesaba katmak gerekir. Örneğin testin uygulama yönergeleri söz konusu ise bunun için geçecek zaman test süresine dâhil edilmemeli, ek süre ayarlanmalıdır. Ya da sınavda birden fazla kitapçık olduğu durumda, her bir kitapçık için ayrı bir süre belirlenmelidir. Bir kitapçık toplandıktan sonra diğer kitapçık dağıtılacaksa arada geçen kayıp zaman sınav bitiş saatine eklenmeli ve testin toplam süresinden düşülmemelidir. Özel gereksinimli adaylara ek süre tanınacaksa hangi engel grubundaki adaya fazladan ne kadar süre verileceği de bu aşamada belirlenmelidir.

Testin özelliklerinin belirlenmesi aşamasında, ulusal ve uluslararası test merkezlerinde çoğunlukla testin özelliklerinin bir tablo ile gösterildiği **test şeması (test blueprint)** kullanılır. Testler hazırlanırken test şeması referans alınır. Bazı kaynaklarda test şeması ile belirtke tabloları birbirinin yerine kullanılır. Ancak test şeması, belirtke tablosundan daha fazla bilgi (maddelerin bilişsel düzeyi, türü, sayısı, güçlük düzeyi, puanlamadaki ağırlıkları ile testin uygulanma biçimi [elektronik, basılı vs.], süre gibi) içerir. Test şemaları geniş ölçekli sınavlar yapan test merkezleri için madde ve performans görevlerinin geliştirilmesine, sonuçları raporlanmasına ve araştırmalara da yön verir. (URL-2). Tablo 3'te 20 maddeden oluşan bir test için örnek test şeması yer almaktadır.

Tablo 3

Örnek Test Şeması

İçerik	Öğrenme Hedefi	Bilişsel Düzeyi	Madde Türü	Madde Sayısı	Madde Ağırlığı	Güçlük Düzeyi (1-3)	Puanlama	Süre (dk.)
Konu 1	H1.1	Hatırlama	Kısa cevaplı	4	%20	1	4*3	4*1
	H1.2	Anlama	Çoktan seçmeli	2	%10	1-2	2*5	2*2
	H1.3	Uygulama	Açık uçlu	1	%5	3	1*10	1*5
Konu 2	H2.1	Anlama	Çoktan seçmeli	2	%10	2-3	2*5	2*2
	H2.2	Uygulama	Açık uçlu	2	%10	2-3	2*10	2*5
Konu 3	H3.1	Hatırlama	Doğru-yanlış	4	%20	1	4*3	4*1
	H3.2	Anlama	Kısa cevaplı	2	%10	1-2	2*3	2*1
	H3.3	Uygulama	Çoktan seçmeli	2	%10	1-3	2*5	2*2
	H3.4	Uygulama	Açık uçlu	1	%5	3	1*10	1*5
Toplam				**20**	**%100**		**100**	**42**

2.3. Maddelerin Yazılması ve Uzmanlardan Görüş Alınması

Testler her zaman sorulardan oluşmaz, bazen çözülmesi istenen bir problem veya yargıda bulunulması beklenilen önermelerden meydana gelir. Kitabın 6. bölümünde ifade edildiği üzere madde terimi bu uyarıcıları kapsar ve sorudan daha geniş bir anlama sahiptir (Vale, 2006). Dolayısıyla testin puanlanabilen en küçük birimi için soru yerine madde terimi tercih edilir (Baykul, 2000). Testin özelliklerinin belirlenmesi aşamasında madde türü ve sayısı ile puanlama anahtarı gibi özelliklere karar verilirken bu aşamada aşağıdaki süreçler izlenir.

2.3.1. Madde Yazımında ve Denetiminde Görev Alacak Kişilerin Belirlenmesi

Bir testte yer alan maddelerin kalitesi madde yazarlarının ne kadar nitelikli madde yazabildiği ile ilişkilidir. İyi yazılmamış bir madde için uygun gözden geçirme süreçleri işe koşulsa ve düzeltmeler gerçekleştirilse de nitelikli bir maddeye ulaşma ihtimali zayıftır. Bundan dolayı madde yazarları alanında uzman kişiler arasından titizlikle seçilmelidir. Madde yazarı maddeyi bilimsel hata içermeyecek, herkes tarafından aynı şekilde anlaşılacak ve ölçme ilkeleri ile uyumlu olacak şekilde hazırlamalı, maddenin doğru cevabını belirlemeli ve güçlük indeksine ilişkin tahminde bulunmalı, ayrıca meslektaşları tarafından hazırlanan maddelerin incelemesinde uzman değerlendirici

olarak görev alabilmelidir (Haladyna, 2004). Kısacası madde yazarı ilgili konu alanını veya psikolojik yapıyı iyi tanımalı, ölçme ve değerlendirmenin temel kavramlarını bilmelidir. Baykul (2000) yaratıcılık gerektiren bir iş olmasından dolayı madde yazmanın sanatsal bir yönü olduğunu da ifade etmektedir. Ek olarak madde yazarları; sözel iletişim becerileri gelişmiş, farklı madde türlerine hâkim ve testi alan grubu yeterince tanıyan kişiler olmalıdır. Dahası madde yazarları sabırlı olmalı; madde yazmanın zamanla gelişen bir beceri olduğunu unutulmamalıdır (Mehrens & Lehmann, 1991).

Bu aşamada madde yazarlarının sayısı da belirlenmelidir. Bu sayıyı belirlerken testin içeriği, belirtke tablosuna göre öngörülen madde sayısı, yazarların uzmanlık alanları vb. özellikler göz önünde bulundurulmalıdır. Madde yazarı sayısı kadar, ekipteki kişilerin ne kadar heterojen olduğu da önemlidir. Test geliştirme sürecinde madde yazarlarında olabildiğince çeşitliliğe (kıdem yılı, mezuniyet düzeyi, cinsiyet vb.) gidilmelidir. Böylece yazılan maddelerde çeşitlilik sağlanabilir.

Madde yazarlarının madde yazım kurallarını ne kadar işletebildiği bir diğer önemli hususdur. Madde yazarları madde türleri ve madde yazımında dikkate edilecek hususlarla ilgili hem teorik hem de uygulamalı eğitimlere alınmalıdır. Downing'e (2006) göre ideal olan test geliştirmede görev alan tüm madde yazarlarının madde yazma ve test geliştirme üzerine eğitimlere katılması ve uzmanlığının olmasıdır. Eğitimlerin hazırlanması ve uygulanması, bazı durumlarda zaman ve maliyet açısından ekonomik olmayabilir. Bu gibi durumlarda madde yazarlarına yönelik kılavuz bir kitapçık hazırlanması yoluna gidilebilir. Kitapçıkta; yönergeler, madde formatı, uygun ve uygun olmayan örnekler, madde yazma ve incelemeye ilişkin kurallar yer alabilir. Bu kitapçıklar yeni soru yazarlarının bir eğitimden geçmesine gerek kalmadan madde yazımına ilişkin bilgileri edinmesine hizmet edeceği gibi mevcut yazarlara istedikleri zaman madde yazımına ilişkin ilkeleri tekrar etme imkânı sunar (Haladyna, 2004).

2.3.2. Madde Yazım Kuralları Çerçevesinde Maddelerin Yazılması

Madde türüne göre maddelerin yazımında dikkat edilecek hususlar alanyazında net biçimde tanımlanmış olup konuyla ilgili kaynaklar incelenebilir (Baykul, 2000; Haladyna, 2004, Haladyna ve ark., 2002; Haladyna & Rodriguez, 2013; Mehrens & Lehmann, 1991; Osterlind, 2002; Russell & Airasian, 2012; Turgut, 1992). Ayrıca bu kitabın 6. Bölümüne bakılabilir. Burada testin biçimsel özelliklerini belirlemek amacıyla oluşturulan belirtke tablosu, konu-soru dağılım tablosu veya test şeması madde yazarları tarafından incelenir. Bu doğrultuda maddeler oluşturulur ve hazırlanan maddeler testin madde havuzunu meydana getirir.

Madde havuzuna genellikle test formunda yer alacak madde sayısından daha fazla sayıda madde dâhil edilir (AERA, APA & NCME, 2014). Deneme uygulaması henüz gerçekleştirilmediğinden testin nihai formunda yer alması planlanan madde sayısının en

az üç katı madde içeren bir madde havuzu hazırlanması önerilir. Tabi bu sayının madde yazarlarının ne kadar deneyimli olduğuna ve ne kadar nitelikli madde yazabildiğine göre değişebileceği unutulmamalıdır (Baykul, 2000). Şöyle ki madde yazarları deneyimli ise yazılan maddelerin görece az bir kısmının elenmesi beklenir, dolayısıyla madde havuzu daha az sayıda maddeden oluşabilir. Diğer taraftan madde yazarlarının tecrübesi düşükse yazılan maddelerin elenme ihtimali daha yüksek olur ki bu durumda madde havuzuna dâhil etmek üzere daha fazla sayıda madde yazılması gerekebilir.

Gelişen teknoloji ile madde yazımı bütünüyle insan gücüne bağlı bir süreç olmaktan çıkmıştır. Artık yapay zekâ tabanlı araçlar sayesinde otomatik madde üretimi gerçekleştirilebilmekte, bu da daha kısa sürede, daha az maliyetle daha büyük madde havuzlarının oluşturulmasına olanak tanımaktadır. Otomatik üretilen maddeler için de tıpkı insanlar tarafından yazılan maddelerde geçerli olan uzman görüşü alma süreçleri hayata geçirilmelidir. Otomatik madde üretimi hakkında daha detaylı bilgi edinmek isteyen okuyucular kitabın 15. bölümünü inceleyebilir.

2.3.3. Maddelerin Cevap/Puanlama Anahtarının Hazırlanması

Madde yazarları, hazırladıkları her bir madde için cevap ya da puanlama anahtarı hazırlamalıdır. Yanıtı seçmeyi gerektiren maddeler objektif biçimde ve çoğunlukla 0-1 puanlandığından yanıtları kolaylıkla hazırlanabilir. Yanıtı katkı gerektiren maddelerin puanlanmasında ise puanlama subjektif nitelik taşır ve 0-1 dışında puanlamalar da söz konusudur. Çoktan seçmeli maddeler, doğru-yanlış maddeler, eşleştirmeli maddeler gibi objektif puanlanan ölçme araçları için madde ve yanıtları uzman görüşlerine sunulmalı, görüş birliğinin sağlanması ile cevap anahtarının nihai hali oluşturulmalıdır. Açık uçlu madde ve performans görevi gibi subjektif puanlanan ölçme araçlarında ise ölçme aracının hazırlanması ile eş zamanlı olarak puanlanmada kullanılacak anahtarın da (kontrol, listesi, dereceleme ölçeği veya rubrik) hazırlanması gerekmektedir. Ayrıca puanlama sürecinde geçerliği ve güvenirliği yüksek sonuçlar elde edilmesi için ölçme aracından bağımsız olarak puanlama anahtarının da geçerlik ve güvenirliği incelenmelidir. Bu amaçla uzman görüşü alınabilir ve farklı ölçme kuramlarına göre puanlayıcı güvenirliğine yönelik çeşitli analizler gerçekleştirilebilir. Örneğin, İngilizce yazma becerisinin değerlendirilmesinde kullanılacak bir rubrik, uzman görüşlerine dayalı olarak incelenip revize edilebilir. Revize edilen rubrik küçük bir grupta sınanarak ilgili yapıya uygunluğu, kategori sayısının yeterli olup olmadığı, kategoriler arasındaki farkların belirginliği ve güvenilir puanlamalar üretip üretmediği incelenebilir.

Madde yazarlarının yazdıkları maddeleri cevap ya da puanlama anahtarları ile birlikte bir forma/sisteme aktarması, maddenin künyesini oluşturmak ve geçerlik denetimlerinde uzmanlar tarafından incelenmesini kolaylaştırmak için önemlidir. Formlarda madde yazarının bilgisi, dersin adı, konu ve öğrenme alanları, hedeflenen içerik (kazanım-davranış), tahmini madde güçlüğü, yararlanılan kaynaklar, madde ve

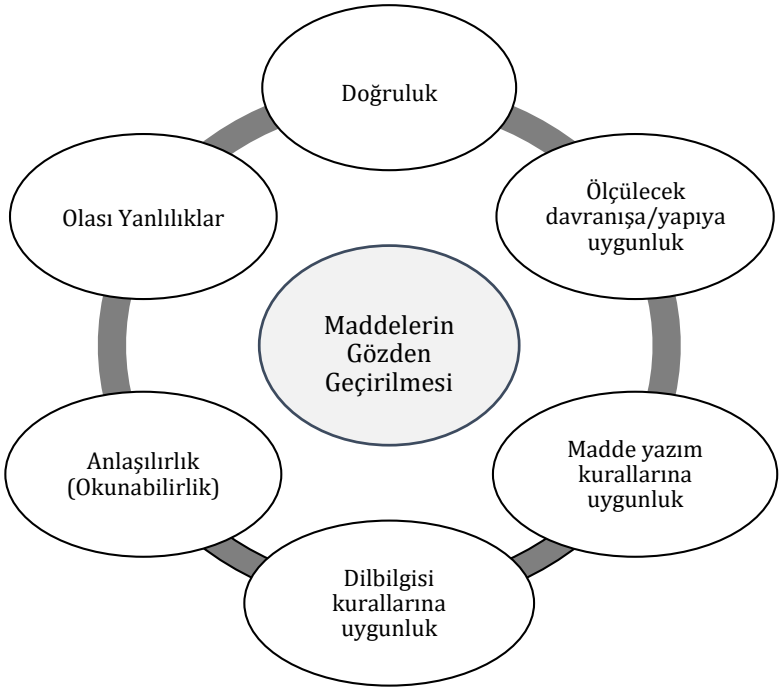
maddenin doğru yanıtı gibi bilgiler yer alabilir (Çetin, 2019). Formlar basılı olarak elde doldurulacağı gibi günümüzde çoğunlukla bilgisayar ortamında madde havuzuna entegre olarak doldurulabilir.

2.3.4. Maddelerin Gözden Geçirilmesi

Madde yazımından sonra maddeler Şekil 4 te yer alan özellikler açısından alan uzmanı, ölçme ve değerlendirme uzmanı ve dil uzmanı tarafından incelenmelidir (AERA, APA & NCME, 2014; Crocker & Algina, 2008).

Şekil 4

Maddeleri Gözden Geçirirken İncelemede Bulunması Gereken Özellikler



Alan uzmanları maddelerin bilimsel doğruluğu, hedef yapıya uygunluğu ve yanlılık içerip içermediği hakkında değerlendirme yapar. Ölçme ve değerlendirme uzmanları her bir maddeyi madde yazım ilkelerine uygunluk açısından denetler ve maddenin yanlılık içerip içermediğine dair görüş bildirir. Dil uzmanları ise maddelerin dilbilgisi, yazım-noktalama kurallarına uygunluğunu ve anlaşılır olup olmadığını kontrol eder. Maddeler uygulanmadan önce uzmanlar tarafından yanlılık incelemesine tabi tutulması sonuçların geçerliği açısından son derece önemlidir. Maddeleri gözden geçirme aşaması bir kalite kontrol süreci olarak düşünülebilir. Bu sayede madde yazarının gözden kaçırdığı noktalar uzmanlar tarafından tespit edilebilir, farklı bir bakış açısıyla maddeler yeniden değerlendirilebilir.

Sınıf içi ölçmelerde öğretmenler hazırladıkları maddeler için bir başka öğretmenden görüş isteyebilirler. Ancak geniş ölçekli sınavlarda maddelerin birden fazla uzmanın görüşüne sunulması gerekir. Uzmanlar inceledikleri maddenin uygun olduğu yönünde görüş bildirebilir, düzeltilerek kullanılmasına veya reddine karar verebilirler. Uzmanların düzeltme önerileri varsa madde, madde yazarları tarafından tekrar gözden geçirilir ve gerekli düzenlemeler gerçekleştirilir. Bu aşamada en az iki uzmandan görüş alınması tavsiye edilir. Böyle bir durumda uzmanlar arasında görüş ayrılığının olması halinde üçüncü bir uzmanın görüşüne başvurulur. Yüksek riskli sınavlarda ise en az üç uzmandan görüş alınması daha uygun bir seçenektir.

Uzmanlardan görüş alırken çoğunlukla her bir uzman diğer uzmanlardan bağımsız olarak maddeleri denetler. Ancak süreci daima bu şekilde ilerletme zorunluluğu yoktur. Duruma göre uzmanlar bir araya gelerek uzman paneli şeklinde de maddeler hakkında inceleme yapabilir.

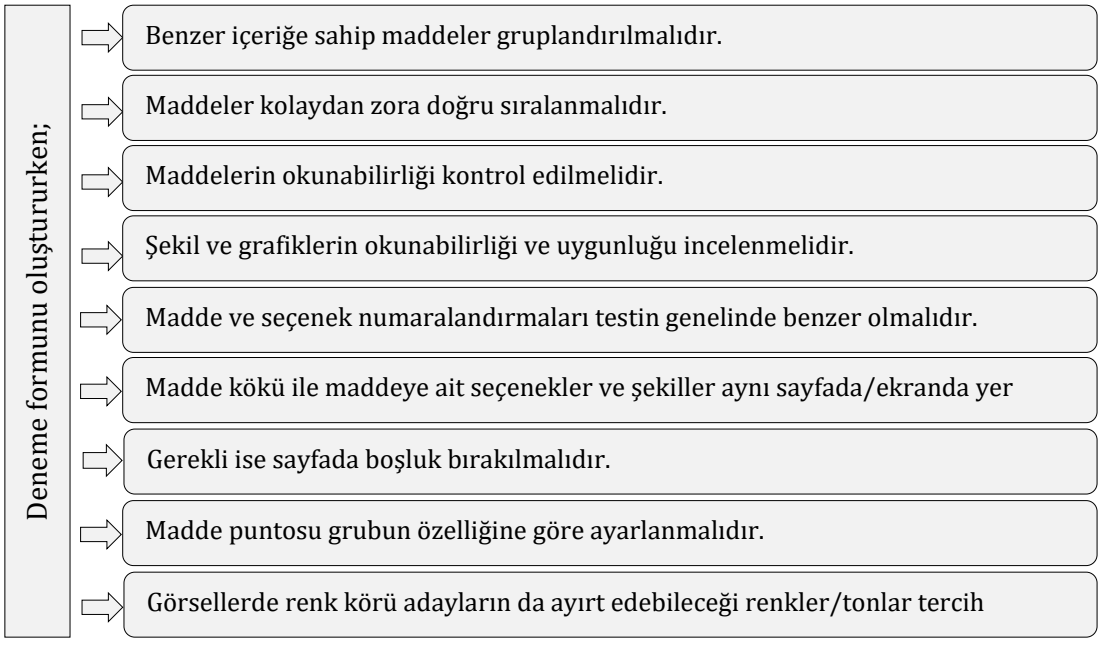
Madde yazarken ve maddelerle ilgili uzmanlardan görüş alırken gözetilmesi gereken bir diğer husus güvenliktir. Tüm sınavlarda madde yazma ve denetleme işlemleri daha önceden belirlenmiş güvenlik prosedürleri dâhilinde gerçekleştirilmelidir. Özellikle yüksek riskli sınavlarda madde yazımında ve denetiminde görev alan uzmanlar belirli gizlilik ve güvenlik protokolüne tabi tutulmalıdır. Örneğin madde yazarları ile gizlilik anlaşmasının imzalanması, madde yazımının ve denetlenmesinin üst düzeyde güvenlik önlemlerinin alındığı ortamlarda yapılması, madde yazarının birinci derece akrabalarının katılım sağlayacağı sınavlara ilişkin madde yazım süreçlerinden muaf tutulması, bir madde yazarı tarafından hazırlanan çok sayıda maddeye aynı sınavda yer verilmemesi gibi kurallar işletilebilir.

2.4. Deneme Formunun Oluşturulması

Bu aşamada test maddeleri denenmek üzere bir testte veya formda birleştirilir. Testin başlangıçta belirlenen amaca ne kadar uygun olduğunu saptamanın yollarından biri madde istatistiklerini incelemektir. Bunun için deneme uygulaması gerçekleştirilir. Deneme uygulamasının ardından bazı maddelerin öngörülen şekilde çalışmadığı tespit edilirse esas uygulama öncesinde bu maddelerin testten çıkarılması gerekir. Çıkarılan maddelerin ardından testin kapsam geçerliğinin korunması için deneme formuna nihai testte yer alacak madde sayısından daha fazla sayıda madde yerleştirilmelidir. Deneme formu, Şekil 5'te sunulan hususlar dikkate alınarak hazırlanmalıdır.

Şekil 5

Deneme Formu Hazırlanırken Dikkat Edilmesi Gereken Hususlar





- Güvenlik riskleri sebebiyle yüksek riskli sınavlarda deneme uygulaması yapılması mümkün olmayabilir, dolayısıyla deneme formunun oluşturulması aşaması atlanır ve doğrudan nihai uygulamaya geçilir. Deneme uygulaması yapılamadığından yüksek riskli sınavlarda maddeleri yazma ve gözden geçirme süreci çok daha titizlikle yürütülmelidir.
- Günümüzde makine öğrenmesi algoritmalarından yararlanılarak deneme uygulamasına gerek kalmadan maddelerin güçlüğü ve yanıt süresi gibi değişkenler kestirilebilmektedir (Fulari & Rusert, 2024; Razavi & Powers, 2025; Yousefpoori-Naeim ve ark., 2024). Bu anlamda deneme uygulamasının yapılamadığı durumlarda makine öğrenme modellerinden yararlanılması önerilebilir.
- Deneme uygulaması gerçekleştirmenin zor olduğu durumlarda başvurulabilecek bir diğer yol; bireylere testi uygularken test formunda belli bir oranda denenmek istenen maddelere yer vermektir. Daha açık bir anlatımla test formu sadece puanlamada dikkate alınacak maddelerden oluşmaz, sonraki uygulamalar için geliştirilen ve nitelikleri kontrol edilmek istenen maddeler de içerir (College Board, 2022). Tabi böyle bir uygulamaya başvurulduğunda denenmek üzere kullanılan maddeler katılımcılarda yorgunluğa neden olmamalı ve katılımcıların asıl maddelerdeki performansını farklılaştırmamalıdır. Bu nedenle puanlamaya dâhil edilmeyip sadece psikometrik niteliklerini belirlemek amacıyla test formuna eklenen maddeler sınırlı sayıda tutulmalıdır.

Yüksek riskli sınavlarda sıkça tercih edilen, çoktan seçmeli maddelerden oluşan testlerde yukarıda belirtilen hususların yanında, doğru cevabın seçeneklere dağılımının dengeli olmalıdır. Örneğin dört seçenekli bir testte her bir seçeneğin doğru yanıt olma olasılığı %25'tir. Maddeler bağımsız hazırlandığından bu adımda gerekli düzenlemeler yapılmazsa doğru cevap hep belli seçeneklere yerleşebilir. Şans başarısına yol açmaması için doğru cevap seçeneklerinin belli bir örüntü oluşturmamasına dikkat edilmelidir (Haladyna, 2004). Ek olarak kesin bir kural olmamakla birlikte arka arkaya 3'ten fazla maddenin doğru cevabının aynı seçenek olmamasına özen gösterilmelidir.

Deneme formu ya da formları oluşturulduktan sonra tüm formlara standart bir yönerge eklenmelidir. Yönergede testin amacı, madde sayısı, testin süresi, cevaplamada dikkate edilecek hususlar (örneğin cevapların optik forma nasıl kodlanması gerektiği) ve puanlamanın nasıl yapılacağı (örneğin düzeltme formülünün uygulanıp uygulanmayacağı) gibi bilgiler yer almalıdır. Unutulmamalıdır ki testi alan bireylerin testi geliştirenin amaçladığı şekilde cevap vermesi yönergelerin açık ve net olmasına bağlıdır. Bu bakımdan standart yönergeler test materyalinin bir parçası olmalıdır (AERA, APA & NCME, 2014).

Impara ve Foster (2006) test uygulamasında kopya çekilmesi, madde havuzundaki maddelerin çokça kullanımından dolayı adaylar tarafından bilinmesi (maddeye maruz kalma oranının artması) ve maddelerin (test veya havuzun) uygulama öncesinde izinsiz ele geçirilmesi gibi geçerliği ve güvenirliği tehdit edebilecek birçok olumsuz durumun, test formunun hazırlanma sürecinde işe koşulacak önlemler ile kontrol altına alınabileceğini belirtmiştir. Impara ve Foster'ın (2006) bu kapsamda önerdiği stratejiler ilgili kaynaktan yararlanılarak aşağıda özetlenmiştir:

- ***Testte yer alan cevap seçeneklerinin rastgele sıralanması:*** Cevapların dizin olarak ezberlenmesini önler. Kâğıt kalem testlerinde bu stratejinin kullanılması halinde birden fazla cevap anahtarı oluşturulmalıdır. Bilgisayar tabanlı sınavlarda yazılım sayesinde bu işlem kolaylaşmaktadır. Seçeneklerde sayılar varsa madde yazım kurallarına göre mantıksal bir sıra izlenmesi gerektiğinden bu strateji önerilmez.
- ***Testin farklı kitapçıklarının ya da formlarının kullanılması***: Testin farklı kitapçıklarının oluşturulması ile testteki maddelerin ve seçeneklerin yerlerinin değiştirildiği testlerin oluşturulması ifade edilmektedir. Testin farklı formları ile test maddelerinin belirli bir oranda (%60 ve %75 aralığında değişen) veya tamamının değiştiği ve ilk forma eşitliğinin sağlandığı formlarının oluşturulması belirtilmektedir. Her iki durumda da adayların soru sıralamasını ezberlemeleri, maddeleri ya da cevapları ezberlemeleri zorlaşacaktır. Kâğıt–kalem testlerinde farklı kitapçıkların oluşturulması maddeler aynı olduğundan dolayı daha ekonomiktir. Testin farklı formlarının geliştirilmesinde ise ekstra madde yazmak gerektiğinden maliyet artmaktadır. Ayrıca test formlarının eşitliğinin sağlanması da

ele alınacak diğer bir süreç olarak karşımıza çıkmaktadır. Bilgisayar tabanlı bireye uyarlanmış testlerde ise bireylerin yetenek düzeyleri göz önünde bulundurularak farklı maddelerden oluşan testler uygulanmaktadır.

- ***Bilgisayar tabanlı testlerin kullanılması***: Bilgisayar tabanlı testler doğrusal biçimde veya bireye uyarlanmış şekilde uygulanabilmektedir. Bireye uyarlanmış testlerde doğrusal testlere kıyasla daha az sayıda madde ile bireylerin yetenek düzeylerine ilişkin benzer kestirimler elde edebilmektedir. Ayrıca bireye uyarlanmış testlerde her aday farklı sorular ve testler alabileceğinden kopya çekilmesi büyük oranda önlenmektedir. Bireye uyarlanmış testlerdeki bazı maddeler, maddeye maruz kalma oranları artıkça adaylar tarafından tanınabilir, yani çok fazla karşılaşılan maddeler söz konusu olabilir. Bu da bir güvenlik riski oluşturup geçerlik ve güvenirliği olumsuz yönde etkileyecektir. Bununla birlikte, yazılımda gerçekleştirilecek ayarlamalar ve büyük madde havuzlarının kullanımı ile bu sınırlılık önemli ölçüde önlenebilir.
- ***Eşdeğer maddelerin kullanılması***: Testlerde madde veya cevap anahtarının ezberlenmesinin önüne geçmenin yollarından biri eşdeğer maddeler kullanmaktır. Bu sayede maddeler ezberlense dahi maddenin eşdeğerinin çözülmesi zorlaşmaktadır. Maddelerin eşdeğerini oluşturmak birçok alanda mümkün olup eşdeğer maddelerin yazılması paralel testlerin oluşturulmasında işlevsel bir rol üstlenmektedir. Sayısal sorularda sayıların değiştirilmesi veya sayıların aynı kalması ancak içeriğin farklılaştırılması ile eşdeğer maddeler oluşturulabilmektedir. Sözel sorularda ise ezberlenme ihtimali olan sözcük veya sözcük grupları değiştirilerek eşdeğer maddeler hazırlanabilmektedir.

2.5. Deneme Uygulaması

Pilot uygulama olarak da isimlendirilen bu aşamada, geliştirilen taslak form testin hedef grubuna benzer fakat daha az sayıda bireyden oluşan bir gruba uygulanır. Bu grubun asıl uygulamanın yapılacağı kitle ile iletişime geçme olasılığının bulunmadığından emin olunmalıdır. Testin deneme formları çoğaltılırken sıkı güvenlik önlemleri alınmalıdır. Buna göre test maddeleri baskı için güvenli olarak sisteme aktarılmalı, fazla veya hatalı basımlar güvenli şekilde imha edilmeli ve test formları güvenli olarak sınav ortamına taşınmalıdır. Son olarak herhangi bir güvenlik açığının oluşmaması için testin belli aşamaları belirli uzmanlara yalnızca sınırlı bir süre açılmalıdır (Downing, 2006). Test formu çoğaltılırken eksik veya hatalı basımların (örneğin sayfanın eksik olması, bazı maddelerin ve/veya görsellerin çıkmaması gibi) kontrolü sağlanmalıdır. Çok sayıda kitapçık olduğunda bu aşamada rastgele seçimler yapılabilir ve seçilen formların farklı uzmanlar tarafından incelemesi sağlanabilir. Ayrıca testin çeşitli formları mevcutsa bu formların denkliği de incelenmelidir. Olası hatalara karşı testi alacak kişi sayısından daha fazla form basılmalıdır.

Deneme uygulamasından önce uzmanların gözden geçirdiği maddeler için cevaplayıcılardan soru hakkında görüşlerini almak ve madde yazarı tarafından öngörülemeyen bir cevap ya da düşünme biçimi olup olmadığını anlamak amacıyla küçük bir grupta ön deneme uygulaması da gerçekleştirilebilir (Baykul, 2000). Ön denemede maddelerde anlaşılmayan noktaların bulunup bulunmadığına ve süre kullanımına ilişkin bilgiler elde edilebilir. Cevaplama sırasında uzun duraksamalar, karalamalar yapma ve cevap değişikliğine gitme durumları maddenin anlaşılırlığının düşük olduğuna dair ipucu sağlayabilir (Crocker & Algina, 2008). Erkuş (2019) ön deneme uygulamasının hedef grubun uç değerlerini (başarılı- başarısız, yüksek yetenek-düşük yetenek vb.) temsil eden 10-15 kişiden oluşan heterojen bir grupta gerçekleştirilmesini, uygulama sırasında öğrencilerin soruları ve çözüm yollarını sesli düşünüp yanıtlamalarının istenmesini ve uygulayıcının süreçte bizzat yer alarak yanıtlayıcının davranışlarını gözlemesini önermiştir. Ön deneme uygulamasına katılan öğrenciler soruları sesli okurken ve yanıtlarken ses/görüntü kaydı alınması tavsiye edilir. Böylece uygulayıcının dışında farklı uzmanların da uygulama süreçlerini incelemesi sağlanabilir ve bu uzmanlardan geribildirim alınarak geçerliğe dair cevaplama süreçlerine dayalı kanıtlara ulaşılabilir. Bilgisayar tabanlı yürütülen ön deneme uygulamalarında ise sesli düşünme protokolüne ilave olarak göz izleme (eye tracking) sistemleri kullanılabilir.

Ön deneme uygulamasından elde edilen bulgular doğrultusunda maddelerde gerekli düzenlemeler gerçekleştirildikten sonra, nihai uygulamaya katılacak gruba benzer nitelikteki bir başka grupta deneme uygulaması gerçekleştirilir. Deneme uygulamasının yapılacağı örneklemin büyüklüğü belirlenirken grubun özelliği, testin türü, madde sayısı, zaman sınırı, maliyet, gerçekleştirilecek analizler vb. değişkenler göz önünde bulundurulmalıdır. Deneme uygulamasında; testi uygulayan kişilerin ve öğrencilerin sınav kurallarına uymaları sağlanmalı, öğrenciler sınava güdülenmeli ve sınav ortamı uygun şekilde düzenlenmelidir. Sıralanan başlıklarla ilgili gerekli düzenlemeler yapılmadığı takdirde ölçümlerin psikometrik özelliklerini zedeleyen hata varyansları oluşur. McCallin (2006) geçerlik ve güvenirliği düşüren hata varyansına kaynaklık edebilecek faktörleri “grubu etkileyen” ve “bireyleri etkileyen” kaynaklar şeklinde ikiye ayırmıştır (bkz. Tablo 4).

Tablo 4

Testin Uygulanmasında Grubu ve Bireyi Etkileyen Hata Kaynakları[*]

	Hata Kaynağı	Örnekler
Grubu Etkileyen	Test uygulama esaslarının değiştirilmesi	Zaman değişimi, test yönergesinde değişikliğe gitmek, farklı uygulama biçimleri kullanmak vb.
	Test uygulama merkezleri veya ortamları arasındaki farklılıklar	Koşulların farklılığı, materyallerin farklılığı vb.
	Testin uygulama aşamasında meydana gelen sorunlar	Adayların yerinin değiştirilmesi, elektrik kesintisi, gürültü vb.
	Testin uygulandığı saatin farklı olması	Adaylar için performanslarının düşük olduğu saatte testin uygulanması
	Puanlama anahtarında değişiklik	Adayların değerlendirme ölçütlerinin testin uygulaması esnasında değişimi
	Teknolojik değişkenlerden kaynaklı ilgisiz varyans kaynakları	Donanımda oluşabilecek sorun, diğer teknik nedenler vb.
Bireyi etkileyen	Test maddelerini önceden alma	Test güvenliğinin ihlal edilmesi
	Adayın performansını etkileyecek test kaygısı	Başarılı olma kaygısı, gözetmenin adayın üzerinde baskı hissi oluşturması vb.
	Gözetmenin kopyaya sebebiyet vermesi	Aday istemeden gözetmenden ipucu alması
	Adayın motivasyon düşüklüğü	Adayın bilerek düşük performans göstermeyi seçmesi
	Adayın yönergeleri yanlış anlaması	Gözetmenin adayın yönergeyi anladığından emin olmaması
	Adayın testte nasıl ilerleyeceğini bilmemesi	Adayın cevabı nasıl kaydedeceğini bilememesi, test sistemini uygun kullanamaması

[*] *McCallin'dan (2006) uyarlanmıştır.*

Tabloda yer alan hata kaynakları incelendiğinde üç önemli kavram göze çarpmaktadır. Bunlar; uygulama koşullarının standartlaştırılması, sınav güvenliği ve gözetmenliktir. Bu kavramları her biri aşağıda açıklanmıştır (Downing, 2006):

- *Standartlaştırma*: Testlerin tüm bireylere mümkün olduğunca aynı koşullarda uygulanması istenmeyen değişkenlik kaynakları önlemeye, yani sonuçları hata varyansından arındırmaya yardımcı olur ve geçerliği artırır.
- *Sınav Güvenliği*: Testin uygulanmasında güvenlik her zaman en temel unsurlardandır. Özellikle yüksek riskli sınavlarda güvenlikle ilgili üst düzey önlemler alınması gerekir. Bu önlemlerden ilki test geliştirme sürecinin birçok

aşamasını kısıtlı sayıda bireyin erişimine açmaktır. Testin üretilmesi ve uygulanması sırasında sürece dışarıdan dâhil olan birey sayısı artacağından güvenlik zinciri genişler. Kâğıt kalem testlerinde test materyallerinin uygulama sahasına gönderilmesi, kuryeler, bina ve salon görevlileri tarafından teslim alınması, testlerin adaylara dağıtılması ve adaylardan toplanması, sınav evraklarının sınav merkezine gönderilmek üzere hazırlanması gibi aşamalarda sınav güvenliğini sağlamaya yönelik kurallar titizlikle uygulanmalıdır. Bilgisayar tabanlı testlerde kitapçıkların basılması, uygulanması ve değerlendirilme aşamaları olmadığından sıralanan başlıklarla ilgili güvenlik ihtiyacı kâğıt kalem testlerine kıyasla nispeten daha azdır. Ancak bilgisayar tabanlı testlerde daha başka güvenlik açıklarının olabileceği ve bunlara yönelik önlemler alınması gerektiğinin altı çizilmelidir. Örneğin, sınavların içeriği ve adayların cevapları şifrelenmeli, adayların kişisel verileri teknik yöntemler ile kontrol altına alınmalıdır (AERA, APA & NCME, 2014). Test bankasına ait dosyalar yerel sınav sunucusuna iletilmeden önce şifrelenmeli, maddelerin tamamı tek bir dosya olarak değil ihtiyaç duyuldukça kısmi olarak sunucudan çekilmeli, sınav merkezindeki sunucu bulunduğu yerde fiziksel olarak korunmalı ve tüm bu güvenlik önlemleri kayıt altına alınmalıdır (Becker & Pomplun, 2006).

- *Gözetmenlik*: Testlerin uygulanmasında sınav kurumları uygulama ve denetlemede yardımcı olması için eğitimli ve deneyimli gözetmenlerden yararlanılmalıdır. Bilgisayar tabanlı testlerde maddeler ve yanıtlar sistem üzerinden iletildiğinden bu sınavlar kâğıt kalem testlerine kıyasla daha az sayıda görevli/gözetmen ile yürütülebilmektedir.

2.6. Madde İstatistiklerinin İncelenmesi

Test geliştirmede temel alınan ölçme kuramına bağlı olarak, deneme formundaki maddelerin madde istatistikleri/özellikleri (madde güçlüğü, madde ayırt ediciliği, madde karakteristik eğrisi, değişen madde fonksiyonu), testin amacına uygun bir nihai form oluşturmak için incelenir. İlgili istatistikler hakkında kısaca bilgi sunulmuştur, daha kapsamlı bilgi edinmek için kitabın 11. bölümü incelenebilir.

2.6.1. Madde Güçlük İndeksi (p_j)

Madde güçlük indeksi, klasik test kuramı'na göre maddenin grupta doğru yanıtlama oranı olup 0 ile 1 arasında değerler alabilir. Madde güçlük indeksinin 0 olması maddeyi grupta kimsenin doğru yanıtlayamadığı anlamına gelirken 1 olması gruptaki tüm kişilerin maddeyi doğru yanıtlandığını belirtir. Anlaşılacağı üzere madde güçlük indeksi 0'a yaklaştıkça madde zorlaşırken 1'e yaklaştıkça madde kolaylaşır. Madde güçlük indeksinin 0,50 olması ise grubun yarısının maddeyi doğru cevapladığını ve maddenin orta güçlükte olduğunu yansıtır. Madde güçlüğü 0 ve 1 değerini aldığında farklı yetenek düzeyindeki bireyler birbirinden ayırt edilemez. Madde orta güçlükte olduğunda (p_j=0,50) ise farklı yetenek düzeyindeki kişiler birbirinden en iyi şekilde ayırt edilebilir ve beraberinde madde varyansı [(p_j) × (1–p_j)] maksimum değerini alır**.** Bu sebeple

testteki maddelerin çoğunun orta güçlükte olması, kolay ve zor madde sayısının daha sınırlı sayıda tutulması önerilir. Madde güçlüklerine ilişkin böyle bir dağılım, bireylerin ölçülen özellik bakımından en iyi şekilde ayırt edilmesini sağlar. Bununla beraber, testin uç gruplar için kullanımı amaçlanıyorsa daha zor veya daha kolay maddelerin ağırlıkta olduğu testler oluşturmak gerekebilir (Allen & Yen, 1979). Yani teste seçilecek maddeler için ideal güçlüğün ne olduğu testin amacına göre şekillenir. Çok sayıda başvuranın olduğu bir test ile az sayıda eleman alınması amaçlanıyor ise zor maddeler teste seçilebilir. Test sadece temel yetenek düzeyindeki bireyleri ayırt etmek amacıyla geliştirilecek ise kolay maddeler teste alınabilir.

2.6.2. Madde Ayırt Ediciliği (r_j)

Madde ayırt ediciliği, maddenin ölçülen gizil özelliğe farklı düzeylerde sahip olan kişileri birbirinden ayırt etme düzeyidir. Klasik Test Kuramı'na göre, madde ayırt ediciliği alt ve üst grup karşılaştırmalarına ve/veya madde ile toplam test puanı arasındaki korelasyona dayalı olarak hesaplanabilir. Madde ayırt edicilik düzeyi -1,00 ile +1,00 arasında değişmekte olup; maddelerin ayırt ediciliğinin 0,40 ve üzerinde olması mükemmel ayırt ediciliği gösterir. Eğer $0{,}30 \leq r_j \leq 0{,}39$ ise madde yeterince ayırt edicidir, herhangi bir düzeltme olmaksızın teste alınabilir. $0{,}20 \leq r_j \leq 0{,}29$ aralığındaki değerler ise maddenin düzeltildikten sonra teste alınabileceğini belirtir. Dolayısıyla bu maddeler uzmanlara sunulmalı, uzman görüşleri doğrultusunda gerekli revizyonlar yapılarak testte kabul edilmelidir (Crocker & Algina, 2008). Ayırt edicilik indeksi 0,20'nin altında kalan maddelerin ise testten çıkarılması gerekir. Maddede ne tür değişikliklerin yapılması gerektiği konusunda çoktan seçmeli testler için çeldirici analizleri de yol gösterici olabilmektedir.

2.6.3. Madde Karakteristik Eğrisi

Madde karakteristik eğrileri, maddenin doğru cevaplama olasılığı ile bireyin yetenek düzeyi arasındaki ilişkinin gösterildiği grafiklerdir. Bu grafikte yatay eksen yetenek düzeyini, dikey eksenin ise maddenin doğru yanıtlanma olasılığını temsil eder. Madde karakteristik eğrisine bakılarak maddeyi doğru cevaplama olasılığının yetenek düzeyine göre nasıl değiştiği görülebilir (Crocker & Algina, 1986), madde güçlük ve madde ayırt edicilik değerleri belirlenebilir. Maddenin doğru yanıtlama olasılığının 0,50 olduğu yetenek düzeyi madde güçlüğünü ifade eder. Madde karakteristik eğrisinin eğimi ise madde ayırt ediciliği hakkında bilgi verir. Daha dik olan eğriler, madde ayırt ediciliğinin yüksek olduğuna işaret ederken daha düz olan eğriler maddenin ayırt ediciliğinin düşük olduğunu yansıtır. Madde karakteristik eğrisi düz olan maddelerde, düşük ve yüksek yetenek düzeylerinde maddenin doğru yanıtlanma olasılığı, yaklaşık aynı olmaktadır (Baker, 2001). İyi bir madde için madde karakteristik eğrisi pozitif eğime ve orta düzey güçlüğe sahip olmalıdır (Allen & Yen, 1979).

2.6.4. Değişen Madde Fonksiyonu

Değişen madde fonksiyonu, maddenin aynı yetenek düzeyindeki farklı alt gruplarda (cinsiyet, yaş grubu, okul türü, etnik köken) farklı şekilde çalışıp çalışmadığını ortaya koyar (Zumbo, 1999). Örneğin, bir maddede cinsiyet değişkeni açısından değişen madde fonksiyonu gözlenmesi, aynı yetenek düzeyindeki bir kadın ile erkeğin ilgili maddeyi doğru cevaplama olasılıkların farklı olduğunu bildirir (Furr, 2022). Bir madde için değişen madde fonksiyonu saptanması o maddenin yanlı olduğu anlamına gelmez, madde üzerinde daha fazla araştırma yapmak gerektiğine ilişkin ikaz niteliği taşır (Skaggs, 2023). Çünkü bu durum maddenin karşılaştırılan gruplardan birinin aleyhine çalışmasından kaynaklanıyor olabileceği gibi gruplardan birinin maddenin ölçtüğü özellikle ilgili ön öğrenmelerinin yetersizliğinden de kaynaklanıyor olabilir (Popham, 2025). Dolayısıyla maddenin bir gruba haksız avantaj sağlayıp sağlamadığını, diğer bir deyişle yanlı olup olmadığını belirlemek için değişen madde fonksiyonu gösteren sorular uzman incelemesine sunulmalıdır. Uzmanların maddede yanlılık bulunduğuna kanaat getirmesi halinde test geliştiriciler maddeyi değiştirmeli ya da düzenlemelidirler (AERA, APA & NCME, 2014). Değişen madde fonksiyonunu incelemek için klasik test kuramına ve madde tepki kuramına dayalı analizler gerçekleştirilebilir (Camili & Shepard, 1994).

2.6.5. Madde Bankasına Maddelerin Seçimi

Madde yazımı zorlu ve yoğun emek gerektiren bir süreçtir. Maddelerin uygulamaya hazır hale gelene kadar geçtiği süreçler düşünüldüğünde maddelerin gelecek testlerde kullanılmak üzere madde bankalarında bekletilmesi test geliştirmede sıkça kullanılan bir prosedürdür. Madde bankaları maddelerin bir sistematik dâhilinde saklandığı sistemlerdir ve bu sistem belirli kurallar çerçevesinde çalışır. Madde bankası olarak günümüzde bilgisayar tabanlı sistemler kullanılmaktadır. Madde bankası oluşturmada bilgisayarların kullanımı 1976'lara uzanmaktadır (Vale, 2006). Madde yazım ilkeleri doğrultusunda hazırlanan maddeler uzman (alan, ölçme ve dil) görüşlerine sunulduktan sonra gerekli düzenlemeler yapılır ve deneme uygulamasına geçilir. Deneme uygulamadan sonra hesaplanan güçlük ve ayırt edicilik indeksleri dikkate alınarak seçilen maddeler madde bankasına alınır. Madde bankası için çeşitli sistemler kullanılabilir. Nitelikli bir madde bankası sistemi esnek ve güvenilir olmalı, maddeler madde bankasına Şekil 6'da yer verilen ilkeler doğrultusunda kaydedilmelidir (Downing, 2006).

Şekil 6

Maddelerin Madde Bankasına Kaydederken Takip Edilmesi Gereken Süreçler

1. Maddeler özgün bir numara ile saklanmalıdır.
2. Maddelerin içeriğine ve bilişsel düzeyine uygun sınıflamalar yapılmalıdır.
3. Maddelerin uygulamaya ilişkin bilgileri (uygulanma sayısı, tarih bilgisi, KTK veya MTK ya göre madde istatistikleri) sisteme girilmelidir.
4. Aynı formda kullanılmaması gereken maddeler varsa bununla ilgili bilgiler sisteme kaydedilmelidir.

Madde bankasının güvenliğinin sağlanması son derece önemlidir. Vale (2006), madde bankasında güvenlik açısından üç temel tehdit belirtmiştir: maddelerin izinsiz ele geçirilmesi, kaybolması ve açığa çıkması. Bu tehditleri önleyebilmek için madde bankasına kontrollü erişim sağlanması ve şifreleme tekniği kullanılması önerilmektedir. Kontrollü erişimde şifre ya da giriş kartı kullanılabileceği gibi ses, parmak okuma, yüz tanıma gibi sistemler devreye sokulabilir. Şifrelemede özel algoritmaların kullanıldığı şifreleme standartları tercih edilmelidir.

2.6.6. Nihai Formun Oluşturulması

Testin amacına ve özelliklerine uygun formları oluşturabilmek için madde bankasından madde istatistiklerine dayalı seçimler gerçekleştirilmelidir. Testin içerik özelliklerinin belirlenmesi aşamasında oluşturulan belirtke tablosu veya konu soru dağılım tablosundan yararlanılıp deneme uygulaması sonucunda elde edilen madde istatistikleri doğrultusunda amaca en uygun maddeler seçilir. Test için öngörülen ortalama güçlük düzeyi hesaba katılarak ilgili konu içeriği ve bilişsel düzeyde yer alan maddeler arasından uygun güçlük indeksine sahip olanlar seçilebilir. Bu aşamada madde güçlük değeri ile birlikte madde ayırt edicilik değeri de dikkate alınmalıdır. Aynı özelliği ölçen, benzer güçlük düzeyine sahip maddelerden madde ayırt edicilik değeri en yüksek olan maddelerin nihai forma seçilmesi önerilebilir.

Bazı maddelerin farklı testin farklı zamanlardaki uygulamalarda birden fazla kez kullanılması söz konusu olabilir. Madde bankasında yer alan maddelerin daha önceki uygulamalarda elde edilmiş madde istatistikleri de bu aşamada ayrıca değerlendirilmelidir. Maddelerin kullanım sıklığının artması maddelerin tanınmasına veya bireylerin maddeye maruz kalma sıklığının artmasına da yol açabileceği göz önünde bulundurulmalıdır. Sonuç olarak, madde bankasından madde istatistikleri yer alan maddelerin teste seçilmesi ile birlikte nihai formun test ve madde istatistikleri de belirli bir hata payı ile kestirilebilir.

Nihai form oluşturulurken, deneme formu hazırlanırken dikkate alınması gereken tüm hususlar göz önünde bulundurulmalıdır. Bu aşamada da maddeler uygun punto seçilerek yazılmalı, okunabilirliği açısından kontrol edilmelidir, grafikler ve şekiller sayfa düzenine uygun olarak düzenlenmelidir. Test formunda öğrencilere işlem yapabilmeleri için uygun boşluklar bırakılmalıdır. Maddelerin test formunda kolaydan zora sıralanmasına ve içerik olarak benzer olanların bir arada verilmesine dikkat edilmelidir.

Nihai formun oluşturulmasında farklı gruplar için gerekli uyarlamalar da yapılmalıdır. Örneğin görme engelli bireyler için büyük puntoyla veya Braille ile yazılmış kitapçıklar ya da ekran okuyucular (AERA, APA & NCME, 2014) kullanılabilir. Ya da okuyucu / okuyucu-işaretleyici isteyen görme engelli adaylar şekil, grafik, tablo içeren maddelerden muaf tutulabilirler.

Testlerin farklı formlarının oluşturulması söz konusu ise formlarda da maddelerin yerleşim düzeninin benzerliği sağlanmalıdır. Maddelerin sadece yerlerinin değiştirilmesi ile oluşturulan farklı kitapçıkları söz konusu ise madde yazımında geçerli olan kurallar yine göz önüne alınarak oluşturulmalıdır. Örneğin matematik testinde maddelerin seçeneklerinde yer alan sayılar büyükten küçüğe ya da küçükten büyüğe sıralanmalıdır. Testin farklı kitapçıklarında maddelerin sırasının değişmesinde, seçeneklerde madde yazım ilkeleri doğrultusunda değişiklik gerçekleştirilmemesi uygun olacaktır. Ayrıca, testteki maddelerin kolaydan zora doğru sıralanması öğrenci motivasyonu açısından önemlidir. Testlerin farklı formlarının oluşturulmasında da bu durum göz önünde bulundurulmalıdır. Aksi takdirde farklı formların alınması öğrencilerin puanlarına hata varyans kaynağı olarak karışabilir.

Nihai test formunun geliştirilmesi ile birlikte test geliştirme süreci sonlanmıştır. Test geliştirme süreci görüldüğü üzere her bir aşaması planlanan ve ilgili planın titizlikle işletildiği süreçlerden oluşmaktadır. Test geliştirme sürecinden bir takım uygulamalar söz konusudur. Bölümde bu uygulamalar da kısaca sunulmuştur.

3. TESTİN GELİŞTİRİLMESİ SONRASINDA GERÇEKLEŞTİRİLECEK UYGULAMALAR

3.1. Geçme Puanlarının / Testin Standartlarının Belirlenmesi

Tüm testler için zorunlu olmamakla beraber testlerde geçme puanlarının veya test puanlarının belirli düzeylerini belirlemeye yönelik kesme puanlarının belirlenmesi süreçleri yürütülebilir. Yüksek riskli sınavlarda savunulabilir kesme puanlarının oluşturulması geçerlik için kritik öneme sahiptir (Downing, 2006). Cizek'e (2006) göre, kesme puanlarının oluşturulma süreci standart belirleme olarak ifade edilmekte olup testin geliştirilmesi, uygulama ve raporlama süreçlerinde düşünülecek önemli bir aşamadır. Hatta test geliştirmede madde veya görevlerin belirlenme süreçlerinden itibaren dikkate alınması uygun kanıtların toplanması açısından da fayda sağlayabilir.

Dört dil becerisine ilişkin yeterlik düzeylerinin tanımlanmak istediği bir İngilizce testi için standart belirleme çalışmaları gerçekleştirilebilir. Burada testin amacına ve uygulanan gruba dayalı olarak düzeyler belirlenebilir. Uygun standart belirleme yöntemine karar verildikten sonra, testin deneme uygulamasının gerçekleştirildiği gruptan farklı ve daha geniş bir örneklemde veri toplanarak uygun analizler gerçekleştirilebilir. Standartların belirlenmesine yönelik olarak ayrıntılı bilgi için kitabın 16. bölümü incelenebilir.

3.2. Testin Uygulanması

Nihai testin uygulanmasında deneme uygulamasında var olan süreçlerin çoğunluğu geçerli olmakla birlikte birtakım farklılar söz konusudur. Deneme uygulaması ile nihai uygulamanın gerçekleştirileceği gruplar özellik bakımından eş değerdir, ancak grup büyüklüğü ile uygulanan madde sayıları farklılaşır. Ayrıca geliştirilen testin özel gereksinimli bireyleri de kapsayacak şekilde planlanması gerekiyorsa, testin bu aşamasında gerekli düzenlemelerin yapılması gerekmektedir. Buna göre nihai uygulamada gözetilmesi gereken hususlar aşağıdaki gibi sıralanabilir:

1. Asıl uygulama deneme uygulamasına kıyasla daha geniş bir grup üzerinde yürütüleceğinden görevli ve gözetmen sayıları güncellenmelidir. Deneme uygulamasında yaşanan sorunlar varsa asıl uygulama öncesinde bu sorunlarla ilgili görevlilere bilgi veya eğitim verilmelidir.
2. Deneme uygulamasında materyal veya ekipmandan kaynaklı sorun yaşandıysa nihai uygulamada bunlara ilişkin düzenlemeler gerçekleştirilmelidir.
3. Gerçek uygulamada sınavın yürütüleceği salon veya merkezlerin sayısı artacağından güvenlik zinciri genişletilip deneme uygulamasına kıyasla daha ciddi işletilmelidir.
4. Sınav sırasında uygulayıcı kurum veya gözetmen tarafından herhangi bir hata yapılması halinde, bu hataları telafi etmek için nelerin yapılacağı önceden planlanmalıdır. Örneğin sınavın adaya eşdeğer bir form ile tekrar uygulanması gibi.
5. Formlarda baskıdan kaynaklı hataların olması ihtimaline karşı sınavların gerçekleştirildiği binalarda yedek kitapçıklar ve cevap kâğıtları bulundurulmalıdır.
6. Sınav esnasında gerçekleşen herhangi bir olumsuzluk karşısında (bilgisayarda yürütülen bir sınavda elektrik kesintisi meydana gelmesi gibi) zorunlu olarak sınava ara verildiyse yaşanan süre kaybı sınav süresine eklenmelidir.

Sınava girecek adaylar ile uygulama öncesinde sınav uygulama kılavuzu/formu basılı veya elektronik ortamda paylaşılabilir. Formda sınav başvuru koşulları, başvuru süreci, sınav uygulamasında dikkat edilmesi gereken durumlar ve sınavın nasıl değerlendirileceğine ilişkin bilgilere ayrıntılı olarak yer verilebilir. Ayrıca adayların uyması gereken kurallar da kılavuzda belirtilmelidir. Sınava girmeden önce, sınav süresi ve sınavda dikkat edilmesi gereken hususlar hakkında adayların bilgilendirilmesi sınav sonuçlarına karışabilecek hata varyanslarının azaltılmasına katkı sağlar.

Sınav uygulama kılavuzunda adayların sınav binasına getirmeleri yasak olan madde/araç/cihaz (hesap makinası, kulaklık, yiyecek-içecek vb.) bilgisi yer almalıdır. Sınav binası ya da sınav salonuna hangi zaman aralığında adayların kabul edileceği bilgisi, sınav esnasında uyulması gereken kurallar (sınav esnasında konuşmama, yardımlaşmama, salon dışına çıkmama vb.) da adaylar ile paylaşılmalıdır. Sınav sonunda adayın sınav salonunu terk etmeden önce uyması gereken kurallarda yine bu kılavuzda sunulmalıdır (bkz. Ölçme, Seçme ve Yerleştirme Merkezi Başkanlığı [ÖSYM], 2024).

3.3. Testin Sonuçlarının İlanı

Testin uygulanmasından sonra testin planında belirlenmiş tarih aralığında sonuçlar birey veya kurumlara sonuç raporu ile birlikte ilan edilir. Test raporu genel olarak beş bileşene sahiptir: gönderici, mesaj, araç, amaç ve hedef kitle (Ryan, 2006). Gönderici testin uygulama ve yönetiminden sorumlu kurum ya da kuruluştur. Mesaj sınav ile iletilmek istenilen bilgilerdir. Bir test raporunda sadece sayısal veriler değil, basit ve anlaşılır olarak testin neyi kapsadığı, puanların ne anlama geldiği, kesinlik derecesi, test puanlarının yorumlanması ile ilgili yanlış anlamalar ve puanların nasıl kullanılacağının da belirtilmesi gerekmektedir (AERA, APA & NCME, 2014). Araç ise mesajın iletilme biçimidir. Mesaj basılı olarak ya da mobil uygulamada iletilebilir. Amaç ise raporun hangi gaye ile (örneğin, bireyleri seviyeleri hakkında bilgilendirmek) paylaşıldığını belirtir. Hedef kitle ise öğrenci, aile, okul yönetimi, kuruluşlar vb. olabilir.

Test sonuçlarının geçerliği hakkında yeterli bilginin sonuç raporunda yer almasına özen gösterilmelidir. Bu nedenle raporda test sonuçlarının nasıl yorumlanacağı belirtilmeli ve test puanlarına dayalı olarak hangi çıkarımlarda bulunmanın uygun olmayacağına ilişkin bilgi verilmelidir. Raporda yer alan bilgilerin ne kadar ayrıntılı olması gerektiği testin türü ile yakından ilişkilidir. Adaylar bir sertifika sınavı aldıklarında sonuçlar daha çok başarı durumunu belirlemeye yönelik olabilir. Testteki doğru ve yanlış sayıları ile sınav puanı öğrenciye raporlanabilir. Adayların bir yabancı dil sınavına katılmaları durumunda ise sonuçlar her bir beceri bağlamında raporlanabilir. Bireyler hakkında geçti-kaldı kararı verilirken tek bir kesme noktasının mı yoksa birden fazla mı kesme noktasının kullanıldığı raporda açıklanmalıdır. Ağırlıklı ortalama ya da alt ölçekler için ayrı ayrı puanlamanın gerçekleştirildiği bilgisi de rapora dâhil edilmelidir (Downing, 2006).

3.4. Testin Teknik Raporunun Hazırlanması

Testin uygulanmasından sonra gerçekleştirilecek son aşama ise test teknik raporunun hazırlanmasıdır. Test teknik raporu test geliştiriciler tarafından hazırlanan ve test geliştirme sürecinde gerçekleştirilen uygulamaların, ulaşılan sonuçların, geçerlik ve güvenirlik kanıtlarıyla beraber açık, net ve objektif olarak ele alındığı bir dokümandır. Teknik rapor uzmanlardan oluşan küçük bir grup tarafından hazırlanabilir. Raporda sınavın kaç merkezde yapıldığı, testi alan grubun büyüklüğü, katılımcıların betimsel

özellikleri, gruplar arası farklılıklara yönelik analizler, test ve madde düzeyinde istatistikler yer almalıdır. Raporda uygulanan test için öngörülen özellikler ile elde edilen bulguların ne derece örtüştüğüne dair bilgiler de sunulmalıdır. Rapor hazırlanırken anlaşılırlığı artırmak için uygun grafik ve şekillerden yararlanılabilir. Testte yer alan madde türüne göre farklılıklar gösterse de raporun aşağıdaki bilgileri içermesi beklenir (AERA, APA & NCME, 2014; Becker & Pomplun, 2006):

- ⇨ Değerlendirmede kullanılan ölçütler,
- ⇨ Norm çalışma süreçleri ve normlar,
- ⇨ Eşitleme ve ölçekleme sonuçları,
- ⇨ Boş veya hatalı girilen maddelerin nasıl değerlendirildiğine ilişkin açıklamalar,
- ⇨ Açık uçlu maddelerin puanlanmasına görev alan puanlayıcı sayısı,
- ⇨ Puanlayıcılar arası güvenirliğe yönelik analizler,
- ⇨ Puanlayıcılar arasında gözlenen farkların nasıl çözümlendiğine yönelik bilgiler,
- ⇨ Puanlama araçlarının geliştirilme süreci,
- ⇨ Ek geçerlik ve güvenirlik kanıtları,
- ⇨ Sonuçların bireylere veya kurumlara raporlanma biçimi,
- ⇨ Sınav güvenliği ile ilgili önlemler.

4. SONUÇ

Test geliştirme süreçleri literatürde farklı aşamalarda ele alınabilse de bu bölümde sekiz aşamada sunulmuştur. Her bir aşama bir sonraki aşama için ön koşul niteliğinde kabul edilebilir. Bölümde test geliştirme sürecinin ikinci aşamasını oluşturan test özelliklerinin belirlenmesi özellikle ayrıntılandırılmıştır. Bu aşamanın testin yol haritası olarak kabul edilebilecek bir plan niteliğinde sürecin ilk aşamalarında detaylı olarak verilmesi test geliştiriciler için önemli görülmektedir. Deneme uygulamasının tüm testlerde güvenlik gibi nedenler ile gerçekleştirilmesinin mümkün olamayacağına içerikte değinilmiştir. Ek olarak, madde istatistiklerinin makine öğrenmesi yöntemi ile kestirilebileceğinden veya testlerde puanlamaya dahil edilmeyen ancak istatistiklerin kestirilmesi amacıyla uygulanan maddelerin kullanılabileceği bilgisine de yer verilmiştir. Bölümde uygulama sürecinde olası hata kaynaklarından ve çözüm önerilerinden söz edilmiştir, bunun test geliştiricilere katkı sağlayacağı düşünülmektedir. Test formunun oluşturulması ile sonlandırılan test geliştirme sürecinden sonra gelen testin uygulaması aşamasında deneme uygulamasından farklı olan gerçekleştirilen işlemler de sunulmuştur. Bölümde ayrıca test geliştirici kurumun hazırlayacağı testin teknik raporuna yer verilmiş ve raporun içermesi beklenen bilgiler psikometrik özellikler ve sınav güvenliği kapsamında ele alınmıştır.

KAYNAKÇA | REFERENCES

Allen, M. J. & Yen, W. M. (1979). *Introduction to measurement theory.* Brooks/Cole.

American Educational Research Association (AERA), American Psychological Association (APA) & National Council on Measurement in Education (NCME) (Ed.). (2014). *Standards for educational and psychological testing*. American Educational Research Association.

Baker, F. B. (2001). *The basics of item response theory*. ERIC Clearinghouse on Assessment and Evaluation, University of Maryland.

Baykul, Y. (2000). *Eğitimde ve psikolojide ölçme: Klasik test teorisi ve uygulaması*. ÖSYM.

Becker, D. F. & Pomplun, M.R. (2006). Technical reporting and documentation. S. M. Downing & T. M. Haladyna (Ed.), *Handbook of test development* (s. 711–723) içinde. Lawrence Erlbaum Associates.

Camili, G. & Shepard, L. A. (1994). *Methods for identifying biased test items*. Sage.

Cizek, G. J. (2006). Standart Setting. S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (s. 225–258) içinde. Lawrence Erlbaum Associates.

Cohen, R. J. & Swerdlik, M. E. (2018). *Psychological testing and assessment: An introduction to tests and measurement* (9. baskı.). McGraw-Hill Education.

College Board. (2022). *Digital SAT suite of assessments specifications overview*. https://satsuite.collegeboard.org/media/pdf/digital-sat-test-spec-overview.pdf.

Crocker, L. & Algina, J. (2008). *Introduction to classical and modern test theory*. Cengage Learning.

Çetin, B. (2019). Test geliştirme. B. Çetin (Ed.), *Eğitimde ölçme ve değerlendirme* (s. 103-126) içinde. Anı.

Doğan, N. (2019). İnsan davranışları ve ölçme araçlarını sınıflama. N. Doğan (Ed.), *Eğitimde ölçme ve değerlendirme* (s. 75-112) içinde. Pegem Akademi.

Downing, S. M. (2006). Twelve steps for effective test development. S. M. Downing & T. M. Haladyna (Ed.), *Handbook of test development* (s. 3–25) içinde. Lawrence Erlbaum Associates.

Fulari, R. & Rusert, J. (2024,). Utilizing machine learning to predict question difficulty and response time for enhanced test construction. *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications* (*BEA 2024)* (s. 528–533) içinde. Association for Computational Linguistics.

Erkuş, A. (2019). *Psikolojide ölçme ve ölçek geliştirme I: Temel kavramlar ve işlemler* (4. baskı). Pegem Akademi.

Furr, R. M. (2022). *Psychometrics: An introduction* (4. baskı). Sage.

Grounlund, N. E. (1977). *Constructing achievement test* (2. baskı). Prentice-Hall.

Haladyna, T. M. (2004). *Developing and validating multiple-choice test items* (3. baskı). Lawrence Erlbaum Associates.

Haladyna, T. M., Downing, S. M. & Rodriguez, M. C. (2002). A review of multiple-choice item-writing guidelines for classroom assessment. *Applied Measurement in Education, 15*(3), 309–334. https://doi.org/10.1207/S15324818AME1503_5

Haladyna, T. M. & Rodriguez, M. C. (2013). *Developing and validating test items*. Routledge.

Hambleton, R. K. (1984). Validating the test scores. R. A. Berk (Ed.), *A guide to criterion- referenced test construction* (s. 199–230) içinde. Johns Hopkins University.

Impara, J. C. & Foster, D. (2006). Item and test development strategies on minimize test fraud. S. M. Downing & T. M. Haladyna (Ed.), *Handbook of test development* (s. 91–114) içinde. Lawrence Erlbaum Associates.

Kane, M. T. (2006). Content-related validity evidence in test development. S. M. Downing & T. M. Haladyna (Ed.), *Handbook of test development* (s. 131–153) içinde. Lawrence Erlbaum Associates.

Kline, T. J. P. (2005). *Psychological testing: A practical approach to design and evaluation* (2. baskı). Sage.

LaDuca, A., Downing, S. M. &. Henzel, T. R. (1995). Test development: Systematic item writing and test construction. J. C. Impara & J. C. Fortune (Ed.), *Licensure examinations: Purposes, procedures, and practices* (s. 117-148) içinde. Buros Institute of Mental Measurements.

McCallin, R. C. (2006). Test administration. S. M. Downing & T. M. Haladyna (Ed.), *Handbook of test development* (s. 625–652) içinde. Lawrence Erlbaum Associates.

Mehrens, W. A. & Lehmann, I. J. (1991). Measurement and evaluation in education and psychology (4. baskı). Wadsworth.

Millî Eğitim Bakanlığı. Ölçme, Değerlendirme ve Sınav Hizmetleri Genel Müdürlüğü. (2025). *Konu soru dağılım tabloları*. https://odsgm.meb.gov.tr/www/1-donem-konu-soru-dagilim-tablolari-2025-2026/icerik/1474

Murphy, K. R. & Davidshofer, C. O. (2001). *Psychological testing: Principles and applications* (6. baskı). Pearson/Prentice Hall.

Osterlind, S. J. (2002). *Constructing test items: Multiple-choice, constructed-response, performance, and other formats* (2. baskı). Kluwer Academic.

Ölçme, Seçme ve Yerleştirme Merkezi Başkanlığı. (2024). *2024-Yükseköğretim Kurumları Sınavı (YKS) Kılavuzu.* https://dokuman.osym.gov.tr/pdfdokuman/2024/YKS/kilavuz_d23052024.pdf

Özçelik, D. (1992). *Ölçme ve değerlendirme*. ÖSYM.

Popham, W. J. (2025). *Classroom assessment: What teachers need to know* (10. baskı). Pearson.

Thompson, S., Johnston, C. J. & Thurlow, M. L. (2002). *Universal design applied to large scale assessments* (Synthesis Report 44). National Center on Educational Outcomes, University of Minnesota.

Turgut, M. F. (1992). *Eğitimde ölçme ve değerlendirme*. Saydam.

Raymond, M. R. & Neustel, S. (2006). Determining the content of credentialing examination. S. M. Downing & T. M. Haladyna (Ed.), *Handbook of test development* (s. 181–223) içinde. Lawrence Erlbaum Associates.

Razavi, P. & Powers, S. J. (2025). *Estimating item difficulty using large language models and tree-based machine learning algorithms*. https://doi.org/10.48550/arXiv.2504.08804

Rodriguez, M. C. (2002). Choosing an item format. G. Tindal & T. M. Haladyna (Eds.), *Large scale assessment programs for all students* (s. 213-231) içinde. Lawrence Erlbaum Associates.

Ryan, J. M. (2006). Practices, issues, and trends in student test score reporting. S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (s. 677–710) içinde. Lawrence Erlbaum Associates.

Russell, M. K. & Airasian, P. W. (2012). *Classroom assessment: Concepts and applications* (7. baskı). McGraw-Hill.

Skaggs, G. (2023). *Test development and validation*. Sage.

Yousefpoori-Naeim, M., Zargari, S. & Hatami, Z. (2024). *Using machine learning to predict item difficulty and response time in medical tests. Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)* (s. 551–560) içinde. Association for Computational Linguistics.

URL-1. Millî Eğitim Bakanlığı (MEB). (2025). *Fen Bilimleri 8. Sınıf Konu Soru Kitapçığı*. https://cdn.eba.gov.tr/yardimcikaynaklar/2025/09/konusoru/fen8.pdf

URL-2. Smarter Content Explorer. *Test Development & Design.* https://contentexplorer.smarterbalanced.org/test-development#blueprints

Vale, C. D. (2006). Computerized item banking. S. M. Downing & T. M. Haladyna (Ed.), *Handbook of test development* (s. 261–285) içinde. Lawrence Erlbaum Associates.

Zumbo, D. B. (1999). *A handbook on the theory and methods of differential item functioning (DIF): Logistic regression modeling as a unitary framework for binary and Likert-type item scores.* Ottowa: Directorate of Human Resources Research and Evaluation, Department of National Defense. https://faculty.educ.ubc.ca/zumbo/DIF/handbook.pdf

YAZAR HAKKINDA | ABOUT AUTHOR

PROF. DR. MELEK GÜLŞAH ŞAHİN

Gazi Üniversitesi | mgulsahsahin[at]gazi.edu.tr

ORCID: 0000-0001-5139-9777

Prof. Dr. Melek Gülşah Şahin, Gazi Üniversitesi Gazi Eğitim Fakültesi Eğitim Bilimleri Bölümü Eğitimde Ölçme ve Değerlendirme alanında öğretim üyesidir. Lisans eğitimini Hacettepe Üniversitesi, Eğitim Fakültesi, Orta Öğretim Fen ve Matematik Alanları Bölümü, Fizik Öğretmenliği Anabilim Dalı'nda 2006 yılında tamamlamıştır. Hacettepe Üniversitesi, Eğitim Bilimleri Enstitüsü, Eğitim Bilimleri Bölümü, Eğitimde Ölçme ve Değerlendirme Anabilim Dalı'nda 2008 yılında yüksek lisans; 2014 yılında ise doktora derecesini almıştır. 2019 yılında Eğitimde Ölçme ve Değerlendirme Alanında doçent ünvanını almıştır. Dr. Şahin'in lisans ve lisansüstü düzeyde yürüttüğü dersler ve tez danışmanlıklarına ek olarak, ulusal ve uluslararası düzeyde çok sayıda yayını ile projelerde görevleri bulunmaktadır. Bilgisayar ortamında bireye uyarlanmış test, çok aşamalı testler, ölçme aracı geliştirme konuları çalışmaları alanları arasında yer alır.

PROF. DR. MELEK GÜLŞAH ŞAHİN

Gazi University| mgulsahsahin[at]gazi.edu.tr

ORCID: 0000-0001-5139-9777

Prof. Dr. Melek Gülşah Şahin is a faculty member in the Division of Educational Measurement and Evaluation, Department of Educational Sciences, Gazi Faculty of Education, Gazi University. She completed, earned her bachelor's degree in Physics Education from the Department of Secondary Science and Mathematics Education, Faculty of Education, Hacettepe University, in 2006. She received her master's degree in 2008 and her Ph.D. in 2014 from the Division of Educational Measurement and Evaluation, Department of Educational Sciences, Graduate School of Educational Sciences, Hacettepe University. In 2019, she was promoted to Associate Professor of Educational Measurement and Evaluation. In addition to teaching undergraduate and graduate courses and supervising graduate theses, Dr. Şahin has numerous national and international publications and has participated in various projects. Her research interests include computerized adaptive testing, computerized multistage testing, and assessment tools development.

6. BÖLÜM

BİLİŞSEL ALANA ÖZGÜ MADDE TÜRLERİ VE MADDE YAZIM İLKELERİ

COGNITIVE DOMAIN ITEM TYPES AND PRINCIPLES OF ITEM WRITING

Dr. Tuba GÜNDÜZ

Doç. Dr. Yıldız YILDIRIM

BİLİŞSEL ALANA ÖZGÜ MADDE TÜRLERİ VE MADDE YAZIM İLKELERİ

Dr. Tuba GÜNDÜZ
Muğla Sıtkı Koçman Üniversitesi

Doç. Dr. Yıldız YILDIRIM
Aydın Adnan Menderes Üniversitesi

Öz

Bu bölümün amacı, bilişsel alana özgü madde türlerini tanıtmak ve bu maddelerin geliştirilmesi sürecinde dikkate alınması gereken temel madde yazım ilkelerini ortaya koymaktır. Testin puanlanabilen en küçük birimi olan madde, testin psikometrik niteliğini doğrudan etkileyen temel yapı taşlarından biridir. Bu nedenle, test geliştirme sürecinde ölçülmek istenen özelliğin doğasına ve ölçmenin amacına en uygun madde türünün seçilmesi, elde edilecek ölçümlerin geçerliği ve güvenirliği üzerinde belirleyici bir role sahiptir. Dolayısıyla madde türünün seçimi, test geliştirme sürecinin kritik aşamalarından biri olarak değerlendirilmektedir. Bu seçim sürecinde, her bir madde türünün üstün ve sınırlı yönlerinin birlikte değerlendirilmesi gerekmektedir. Bu bölümde madde türleri katkı gerektiren ve seçme gerektiren maddeler olmak üzere iki temel grupta ele alınmıştır. Analiz etme, değerlendirme ve yaratma gibi üst düzey bilişsel becerilerin ölçülmesinde katkı gerektiren maddeler; hatırlama, anlama ve uygulama gibi görece alt düzey bilişsel becerilerin ölçülmesinde ise seçme gerektiren maddeler daha yaygın olarak kullanılmaktadır. Bununla birlikte, bu ayrım kesin bir sınıflandırma olarak görülmemeli; uygun biçimde tasarlanan her iki madde türünün de farklı bilişsel düzeylerin ölçülmesine hizmet edebileceği göz önünde bulundurulmalıdır. Madde türünün belirlenmesinden sonra ilgili maddenin madde yazım ilkelerine uygun biçimde yazılması en az madde türünün seçimi kadar önemlidir. Nitekim maddenin niteliği, testten elde edilen puanların geçerlik ve güvenirlik düzeyini önemli ölçüde etkilemektedir. Bu bağlamda, madde yazım ilkelerinin uygulanması, ölçme aracının psikometrik kalitesinin artırılmasına katkı sağlamaktadır. Bu bölümde, madde türlerinin üstün ve sınırlı yönleri dikkate alınarak hem geniş ölçekli testlerde hem de sınıf içi ölçme ve değerlendirme uygulamalarında ölçme amacına uygun madde türünün seçilmesine ve maddelerin madde yazım ilkelerine uygun biçimde hazırlanmasına rehberlik edecek bilgiler, örnek uygulamalarla birlikte sunulmuştur.

Anahtar Kelimeler

Bilişsel Alana Özgü Maddeler, Madde Türleri, Katkı Gerektiren Maddeler, Seçme Gerektiren Maddeler, Açık Uçlu Maddeler, Çoktan Seçmeli Maddeler, Madde Yazım İlkeleri

COGNITIVE DOMAIN ITEM TYPES AND PRINCIPLES OF ITEM WRITING

Dr. Tuba GÜNDÜZ
Muğla Sıtkı Koçman University

Assoc. Prof. Dr. Yıldız YILDIRIM
Aydın Adnan Menderes University

Abstract

This chapter aims to introduce the item types used in the cognitive domain and to present the fundamental principles of item writing for each item type. As the smallest scorable unit of a test, an item is one of the key components that directly influences the psychometric quality of the assessment tools. Therefore, selecting the item type that best aligns with the nature of the construct being measured and the purpose of the assessment plays a decisive role in ensuring the validity and reliability of the measurements. Accordingly, the selection of an appropriate item type is regarded as one of the critical stages of the test development process. This selection should be made by carefully considering both the strengths and the limitations of each item type. Item types used to assess the cognitive domain are generally classified into two broad categories: constructed response items and selected response items. Constructed response items are more commonly used to assess higher-order cognitive skills, such as analyzing, evaluating, and creating, whereas selected response items are more frequently employed to measure relatively lower-order cognitive skills, including remembering, understanding, and applying. Nevertheless, this distinction should not be regarded as absolute, as both item types, when appropriately designed, can be used to assess different levels of cognitive processes. Once the item type has been determined, writing the item in accordance with established item-writing principles is as important as selecting the appropriate item type itself. Indeed, the quality of the item has a substantial influence on the validity and reliability of test scores. Accordingly, adherence to item-writing principles contributes to improving the psychometric quality of the assessment tool. This chapter presents guidance on selecting the most appropriate item type for the intended measurement purpose by considering the strengths and limitations of different item types. It also provides practical recommendations, supported by illustrative examples, for developing items in accordance with item-writing principles for use in both large-scale assessments and classroom assessment practices.

Keywords

Cognitive Domain Items, Item Types, Constructed Response Items, Selected Response Items, Open-ended Items, Multiple Choice Items, Principles of Item Writing

1. GİRİŞ

Ölçmeye konu olan özelliğin ölçülmesinde yararlanılan ölçme araçlarında puanlanabilen en küçük birime madde denir. Duyuşsal alana özgü maddeler çoğunlukla ifade, cümle biçimindeyken bilişsel alana özgü madde türleri genellikle soru ifadesi biçimindedir. Bu sebeptendir ki başarıyı ölçmek için kullanılan testlerde olduğu gibi testi oluşturan maddelere soru ismi de verilir. Soru ifadesi, ölçmeye konu olan özelliğin yoklanmasında bahsedilen her durum için geçerli olmamakla birlikte bilişsel alana özgü maddeler yerine bu ifadenin kullanılması yanlış olmaz. Ancak bölüm boyunca, soru yerine, ölçme ve değerlendirme alanı terminolojisine daha uygun olan madde terimi kullanılmıştır. Kitabın bu bölümünde bilişsel alana özgü madde türleri seçme ve katkı gerektiren maddeler olmak üzere ikiye ayrılarak bu madde türlerinin özellikleri açıklanmıştır. Ardından, her bir madde türüne özgü yazım ilkeleri sunulmuş ve örnekler üzerinden açıklamalar yapılmıştır. Son olarak, madde türlerinin üstün ve sınırlı yönlerine karşılaştırmalı olarak yer verilmiş ve geçerlik ile güvenirlik bağlamında değerlendirilmiştir.

2. MADDE TÜRLERİ

Ölçme aracını oluşturan maddelerin kalitesi ile ölçme aracının kalitesi doğrudan ilişkilidir. Bilimsel doğruluk ve madde yazım ilkelerine uygunluk, maddenin kalitesini belirleyen önemli birer argümandır. Öte yandan her bir madde türünün üstün ve sınırlı yönlerinin bir arada değerlendirilerek amaca uygun kullanılması ise ölçme aracının kalitesi üzerinde önemli bir rol oynamaktadır. Dolayısıyla ölçme ve değerlendirme açısından madde türünün üstün ve sınırlı yönlerinin dikkate alınması ile madde yazım ilkelerinin madde yazım ve denetim süreçlerinde işe koşulması, ölçme aracıyla elde edilecek ölçümlerin geçerlik ve güvenirliğini doğrudan etkilemektedir.

Bilişsel alana özgü madde türleri seçme (selected response / selection items) ve katkı (written response / constructed response / supply response / produce-response items / supply items) gerektirmesine göre iki grupta ele alınabilir (Chappuis ve ark., 2012; Karatoprak-Erşen & Gündüz, 2023; Popham, 2014; Russell & Airasian, 2012; Thorndike & Thorndike-Christ, 2010; Waugh & Gronlund, 2013). Seçme gerektiren madde türleri, adından da anlaşılacağı üzere, bir grup seçenek içinden doğru cevabı işaretleyerek seçme yapmayı gerektirirken, katkı gerektiren madde türleri, bireyin kendi cevabını yapılandırarak yazmayı gerektirir.

2.1. Seçme Gerektiren Madde Türleri

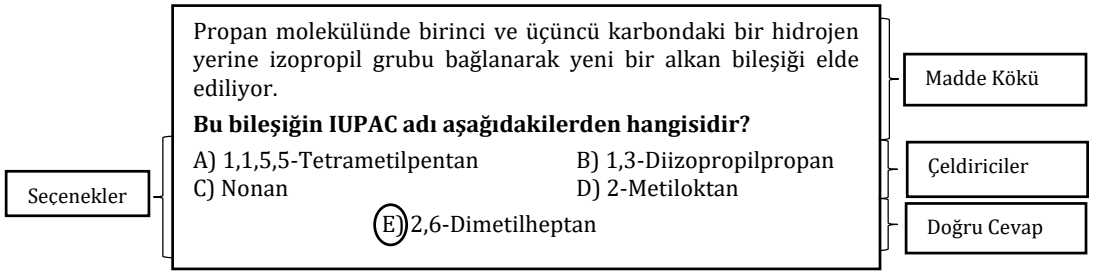
Seçme gerektiren madde türleri çoktan seçmeli, eşleştirmeli ve doğru-yanlış madde türlerinden oluşmaktadır. Bu madde türleri, seçenekler vasıtasıyla basit tanımlara, ezbere ya da sözel çağrışımlara dayalı olarak doğru cevabı verilemeyecek şekilde hazırlanmalıdır (Ebel & Frisbie, 1991).

2.1.1. Çoktan Seçmeli Madde

Çoktan seçmeli madde (multiple-choice item), madde kökü ve bir dizi seçenekten meydana gelir. Seçenek sayısı üç veya üçten fazla olup, seçenekler doğru cevap ve çeldiricilerden oluşur. Madde kökü, tek bir soru ifadesinden oluşabileceği gibi açıklama da içerebilir. Çeldiriciler ise kesinlikle yanlış olmakla birlikte, konuya hâkim olmayan bireylerin seçeceği ve doğru cevapmış gibi görünen seçeneklerdir. Şekil 1'de çoktan seçmeli örnek bir maddenin ögeleri sunulmuştur.

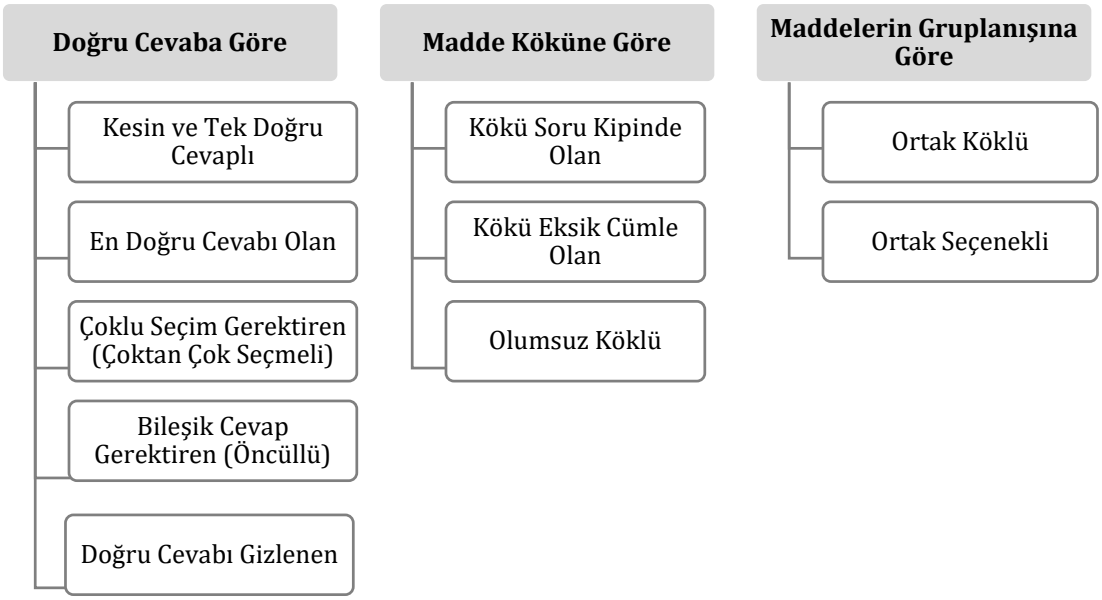
Şekil 1

Çoktan Seçmeli Maddenin Ögeleri



Bu madde türü öncelikli olarak hatırlama ve anlama düzeyindeki öğrenme çıktılarını ölçen başarı testlerinde yaygın olarak kullanılmakla birlikte, iyi tasarlanmış bir çoktan seçmeli madde, daha üst düzey düşünme becerilerini de değerlendirmek için kullanılabilir (Russell & Airasian, 2012). Çoktan seçmeli maddeyi doğru-yanlış türü maddeden ayıran özelliği iki seçenekten fazla seçeneğe sahip olmasıdır. Çoktan seçmeli maddede seçenek sayısına karar vermede alanyazında sıklıkla eğitim düzeyi dikkate alınır. Ancak son zamanlarda yapılan ampirik araştırmalar, üst kademelerdeki eğitim düzeylerinde dahi (örneğin, tıpta uzmanlık gibi) iyi geliştirilmiş üç seçenekli bir çoktan seçmeli maddenin de yeterince iyi çalıştığını kanıtlamıştır (Al-lawama & Kumwenda, 2024; Thanyapa & Currie, 2014). Dahası böyle bir maddeyi beş seçenekli hâle getirmek için doldurma seçenek yazmak maddenin psikometrik ve istatistiksel özelliklerini zedeleyebilir.

Çoktan seçmeli maddeler, farklı açılardan incelenerek sistematik bir sınıflandırmaya tabi tutulabilir. Yapılacak sınıflandırma maddenin farklı unsurlarına ve işlevine göre sınıflara ayrılabilir. Baykul'a (2000) ve Turgut ve Baykul'a (2013) göre çoktan seçmeli madde türleri doğru cevaba göre, madde köküne göre ve maddelerin gruplanışına göre üç ana başlıkta ele alınmaktadır.



2.1.1.1. Doğru Cevaba Göre Çoktan Seçmeli Madde Türleri

Doğru cevaba göre çoktan seçmeli maddeler; kesin ve tek doğru cevaplı, en doğru cevabı olan, cevabı birden fazla olan, bileşik cevap gerektiren ve doğru cevabı gizlenen şeklinde beşe ayrılmaktadır.

Kesin ve Tek Doğru Cevaplı Çoktan Seçmeli Madde

Doğru cevaba göre çoktan seçmeli madde türlerinden en sık kullanılan madde türü kesin ve tek doğru cevaplı çoktan seçmeli maddelerdir. Bu tür maddelerde seçeneklerden yalnızca biri doğru olup diğer seçenekler kesinlikle yanlıştır. Aksi durumda sorunun bilimsel hata içerdiği kanısına varılabilir. Bu nedenle dikkat edilmesi gereken husus, maddenin çeldiricilerinin kesin ve net bir biçimde yanlış olduğunun teyit edilmesidir. Şekil 1'de verilen çoktan seçmeli madde kesin ve tek doğru cevaplı çoktan seçmeli maddelere bir örnektir.

En Doğru Cevabı Olan Çoktan Seçmeli Madde

En doğru cevabı olan çoktan seçmeli maddelerde çeldiriciler kısmen doğrudur. Ancak doğru cevabın doğruluk derecesi çeldiricilere göre daha yüksektir. Bu tür maddelerin cevaplanması kesin ve tek doğru cevaplı maddelere göre daha güçtür. Çünkü çeldiriciler doğru cevaba yakındır ve bireyin seçeneklerin tamamını ele almasını gerektirir. Bu madde türünde dikkat edilmesi gereken husus, çeldiricilerin kesin ve net bir biçimde yanlış olması yerine kısmen doğru olacak şekilde hazırlanmasıdır. Bunun ihlal edildiği bir durum söz konusu olduğunda bu çeldiriciler daha az çalışacak ve madde kolaylaşacaktır. Aynı zamanda bu durum ayırt ediciliğin düşmesine de neden olabilir. Hazırlaması uzmanlık gerektiren bu madde türünde doğruluk derecesi en fazla olan

seçeneğin yani doğru cevabın tartışmaya açık olmaması da dikkat edilmesi gereken başka bir hususdur. Şekil 2'de en doğru cevabı olan çoktan seçmeli bir madde örneği verilmiştir.

Şekil 2

En doğru cevabı olan çoktan seçmeli madde örneği

Bir öğretmen öğrencileriyle açık iletişim kurmaya ve sınıfında olumlu bir iklim oluşturmaya önem vermektedir. Bu doğrultuda öğrencilerinin akademik ve sosyal gelişimlerini desteklemeye çalışmaktadır. Bu öğretmen öğrencilerinden birinin ders sırasında sık sık arkadaşlarının sözünü kestiğini ve alaycı ifadeler kullandığını fark ederek sınıfın iklimini bozmayacak bir tepki vermeye karar vermiştir.

Bu öğretmenin öğrencisine vereceği en uygun tepki aşağıdakilerden hangisidir?

A) Sınıf içinde arkadaşlarından özür dilemesini istemek

*B) Davranışının nedenlerini anlamaya çalışmak

C) Sınıf arkadaşlarını örnek göstermek

D) Teneffüslerde sorumluluk vermek

E) Davranışlarını görmezden gelmek

*: Doğru cevap

Çoklu Seçim Gerektiren Çoktan Seçmeli Madde

Bir madde kökünün birden fazla doğru cevabı olduğunda kullanılabilecek madde türlerinden biri çoklu seçim gerektiren (cevabı birden fazla olan çoktan seçmeli / çoktan çok seçmeli) madde türüdür (Millî Eğitim Bakanlığı, 2025). Bu madde türünde maddenin birden fazla seçeneği doğru olup kalan seçenekler kesinlikle yanlıştır. Bu tür maddelerde seçenek sayısı beş ise; iki veya üç seçeneğin doğru olması beklenir. Çoktan çok seçmeli maddeler geleneksel çoktan seçmeli maddelere göre daha az yaygın olduğundan bu madde türünün kullanıldığı sınavlar için bu durum yönergede mutlaka vurgulanmalıdır. Bu maddenin puanlanmasında kısmî puanlama kullanılması önerilmemektedir (Turgut & Baykul, 2013). Bunun yerine; tam doğru yanıtlara tam puan, kısmî doğru, yanlış ve boş cevaplara 0 puan verilmelidir.

Bileşik Cevap Gerektiren Çoktan Seçmeli Madde

Bileşik cevap gerektiren çoktan seçmeli madde türü, alanyazında öncüllü madde ya da K tipi madde olarak da adlandırılmaktadır. Bileşik cevap gerektiren maddelerde soru kökünde öncüller bulunmaktadır. Bu öncüllerden biri ya da birkaçı doğru cevaptır; diğer öncüller ise çeldirici olarak görev yapar. Bireyden her biri seçeneklerin bir

kombinasyonunu temsil eden rakamlardan / harflerden (I., II, III.; A., B., C. vb.) birini seçerek doğru cevap ya da cevapları belirlemesi istenir (Burton ve ark., 1991). Bu tür maddeler genellikle çeldirici yazmanın zor olduğu durumlarda ve daha fazla çeldirici yazmanın bağlam açısından mümkün olmadığı durumlarda, doldurma seçenek yazmaktan kaçınmak için kullanılmaktadır. Bu nedenle hazırlaması, cevabı kesin ve tek olan ya da cevabı en uygun olan çoktan seçmeli maddelere göre daha kolaydır. Ancak bu madde türünün kullanımı alanyazında önerilmemekle birlikte mümkün olduğunca kullanmaktan kaçınılmalıdır (Burton ve ark., 1991; Butler, 2018; Haladyna ve ark., 2002; Sireci ve ark., 1998). Çünkü cevabı kesin olarak bilmeyen bireyler de bu madde türündeki maddeleri kısmi bilgiyle doğru yanıtlayabilirler. Böylelikle bu maddelerde bilen ve kısmi bilen bireyleri birbirinden ayıramadığı için ayırt edicilik daha düşük olabilir. Bu nedenle bu tür maddelerin güvenirliğinin de düşük olduğu alanyazında bilinmektedir (Rudolph ve ark., 2019). Öte yandan bu tür maddeler, cevabı kesin ve tek olan maddelere göre genellikle daha zordur (Haladyna & Downing, 1989; Rush ve ark., 2016).

Bu tür maddelerin yazımında dikkat edilmesi gereken bazı hususlar aşağıda verilmiştir (Rudolph ve ark., 2019).

- Seçenek yazımında dikkat edilmesi gereken ilkelere öncül yazımında da dikkat edilmelidir.
- Öncüller mümkün olduğunca olumlu ifade edilmelidir.
- Öncüllerin tamamının doğru olmasından kaçınılmalıdır.
- Öncüller seçeneklere mümkün olduğunca eşit oranda dağıtılmalıdır.
- Öncül sayısı dörtten fazla olmamalıdır.

Ayrıca bu madde türüne doldurma seçenek yazmaktan kaçınmak için başvurulduğunda testin genelinde bileşik cevap gerektiren madde sayısı testteki madde sayısının üçte birini geçmemelidir. Ve hatta bu madde türü mümkün olduğunca az kullanılmalıdır. Şekil 3'te bileşik cevap gerektiren çoktan seçmeli bir madde örneği verilmiştir.

Şekil 3

Bileşik Cevap Gerektiren Çoktan Seçmeli Madde Örneği

Fotosentez ile ilgili

I. Fotosentez ışık varlığında gerçekleşir.
II. Fotosentez sırasında oksijen ve su kullanılır.
III. Fotosentez sonucunda karbondioksit açığa çıkar.
ifadelerinden hangileri doğrudur?

*A) Yalnız I B) Yalnız II C) I ve II D) I ve III E) II ve III

*: Doğru cevap

Doğru Cevabı Gizlenen Çoktan Seçmeli Madde

Madde ile yoklanan bilgi veya beceride, sözel alanlarda doğru cevabı bilmeyen bireylerin seçeneklerde verilenlerle sayısal alanlarda ise yerine koyarak deneme-yanılma yöntemiyle maddenin doğru cevaplanmasının önüne geçilmesi için bu madde türünden yararlanılabilir. Böyle durumlarda doğru cevap seçenekler arasında doğrudan yer almayıp gizlenerek sunulur. Bu tür maddelerin kullanımı maddenin test çözme tekniklerine dayalı olarak cevaplanması ihtimalini ve dolayısıyla şans başarısını azaltmaktadır. Bu tür maddelerde bireylerin daha fazla aşina oldukları seçeneği işaretlemelerinin önüne geçilerek madde istatistikleri gerçeğe daha yakın elde edilebilir ve madde istatistiksel açıdan daha doğru yorumlanır. Çünkü aşinalığın madde istatistiklerini bozduğu düşünülmektedir. Özetle bu tür maddelerin kullanımındaki amaç, sadece gerçekten bilen bireyler tarafından doğru cevabın işaretlenerek testin geçerlik ve güvenirliğinin artırılmasıdır. Şekil 4'de doğru cevabı gizlenen çoktan seçmeli maddeye sırasıyla bir sayısal bir sözel örnek verilmiştir.

Şekil 4

Doğru Cevabı Gizlenen Çoktan Seçmeli Madde Örnekleri

$-1<\frac{x-2}{2}<+1$

eşitsizliğini sağlayan x tam sayılarının toplamı kaçtır?

A) 0 B) 2 C) 4 *D) 6 E) 8

İsviçre'nin başkentinin baş harfi aşağıdakilerden hangisidir?

*A) B B) C C) L D) S E) Z

*: Doğru cevap

2.1.1.2. Madde Köküne göre Çoktan Seçmeli Madde Türleri

Madde köküne göre çoktan seçmeli madde türleri; kökü soru kipinde olan, kökü eksik cümle olan ve olumsuz köklü olan şeklinde üçe ayrılmaktadır.

Kökü Soru Kipinde Olan Çoktan Seçmeli Madde

Madde köküne göre çoktan seçmeli madde türlerinden en sık kullanılan madde türü kökü soru kipinde olan çoktan seçmeli maddelerdir. Bu tür maddelerde madde kökü soru ifadesi biçimindedir. Şekil 1'de verilen çoktan seçmeli madde kökü soru kipinde olan çoktan seçmeli maddelere bir örnektir.

Kökü Eksik Cümle Olan Çoktan Seçmeli Madde

Madde kökünün bir soru cümlesi olmayıp seçeneklerle birleştiğinde tam bir cümle oluşturmasının beklendiği çoktan seçmeli madde türüne kökü eksik cümle olan çoktan seçmeli madde denir. Bu tür maddeler, madde kökü soru cümlesi olsa da kökünün herhangi bir kısmında boşluk olacak şekilde de hazırlanabilir. Bu madde türü, bu bölümün devamında anlatılan boşluk doldurma biçimindeki kısa cevaplı madde türünün seçenekli hâli olarak da kullanılabilir. Kökü eksik cümle olan çoktan seçmeli maddelerin kullanımında seçeneklerde gereksiz tekrarlar olup olmadığı ve okuma yükünün azalıp azalmadığı göz önünde bulundurulmalı, azalmıyorsa bu madde türü kullanılmamalıdır (Turgut & Baykul, 2013). Case ve Swanson da (2002) kökü soru kipinde olan çoktan seçmeli madde yerine bu madde türünün kullanımını önermemektedir. Ancak her ne kadar alanyazında önerilmese de dil bilgi ve becerilerini ölçme amaçlı testlerde kullanılabilir. Dil testlerinde madde, cümlenin eksik olan (boş bırakılan) kısmı bir kelime veya söz öbeği olabileceği gibi bazen de cümle tamamlama şeklinde olabilir. Bu tür madde kullanımında madde kökünde ve madde kökü ile seçenek arasında ipucu eklerin kullanılmamasına özen gösterilmelidir ve soru kökü seçeneklerin her biriyle dilbilgisi açısından tamamlanabiliyor olmalıdır. Şekil 5'te kökü eksik cümle olan çoktan seçmeli bir madde örneği verilmiştir.

Şekil 5

Kökü Eksik Olan Çoktan Seçmeli Madde Örnekleri

Sanatçının eserlerinde doğaya yaklaşımı; yüzeyde oldukça sade görünse de aslında derin bir gözlem gücünün ---- ve uzun yıllara yayılan bir birikimin ---- niteliğindedir.

Bu cümlede boş bırakılan yerlere sırasıyla aşağıdakilerden hangisi getirilmelidir?

*A) yansıması – sonucu
B) eleştirisi – karşılığı
C) reddi – başlangıcı
D) taklidi – nedeni
E) sınırlılığı – göstergesi

*: Doğru cevap

Olumsuz Köklü Çoktan Seçmeli Madde

Madde kökünde yapıca veya anlamca olumsuz ifade içeren (örneğin; değildir, yer almaz, yanlıştır gibi ifadelerle biten maddeler ya da hariç, eksikliği gibi anlamca olumsuz ifade içeren maddeler) madde türleridir. Olumsuz ifadeler içeren maddelerden kaçınılması (Chiavaroli, 2017; Harasym ve ark., 1992) alanyazında önerilmekte olup bu tür maddelerin kullanılması durumunda vurgulanması gerekmektedir (Haladyna & Downing, 1989; Haladyna ve ark., 2002). Böyle bir durumda birden fazla olumsuz (çift olumsuz) ifade kullanılmamalıdır. Bunun temelinde olumsuz maddelerin bireylerin dikkatinden kaçabilmesi ya da zihinlerinde karışıklığa yol açabilmesi bulunmaktadır. Dahası bireylerin yanlış cevabı bilmesi doğru cevabı bildiğini göstermediği için de bu tür soruların kullanımı önerilmemektedir (Burton ve ark., 1991). Ayrıca kökü olumsuz maddelerin kökü olumlu ifade edilen maddelere göre çoğunlukla daha zor olduğu bilinmektedir (Harasym ve ark., 1992; Rush ve ark., 2016). Her ne kadar alanyazında olumsuz köklü maddelerin kullanımı önerilmese de hazırlama kolaylığından dolayı sık başvurulmaktadır. Kimi durumda da bağlamın doğası gereği olumsuz maddelerin kullanımı kaçınılmazdır. Ancak önemli olan testin genelinde yer alan olumsuz madde oranıdır ve olumsuz köklü madde sayısı testteki madde sayısının üçte birini geçmemelidir. Ve hatta bu madde türü mümkün olduğunca az kullanılmalıdır.

2.1.1.3. Maddelerin Gruplanışına göre Çoktan Seçmeli Madde Türleri

Maddelerin gruplanışına göre çoktan seçmeli maddeler; ortak köklü ve ortak seçenekli çoktan seçmeli maddeler olmak üzere ikiye ayrılmaktadır.

Ortak Köklü Çoktan Seçmeli Madde

Kökünde ortak bir metin, tablo, grafik ya da şekil olup bu köke bağlı iki veya daha fazla çoktan seçmeli madde bulunan çoktan seçmeli madde türüdür. Bu madde türü alanyazında madde takımı (testlet) olarak da adlandırılır (Kaya-Uyanık & Gelbal, 2018; Yılmaz-Koğar, 2022). Ortak köklü maddelerin kullanımı, ölçümlerin duyarlılığını

artırarak güvenirliği artırma ve test süresini kısaltma gibi önemli üstünlüklere sahiptir (Atalay-Kabasakal & Gören, 2025). Ancak bununla birlikte bazı sınırlı yönleri de bulunmaktadır. Hazırlanması ayrı bir uzmanlık gerektirmekte olup özellikle dikkatli geliştirilmediğinde maddeler birbirleriyle bağımlı olabilmekte ve maddelerin yerel bağımsızlığını ihlal etmektedir. Oysa bu istenen bir durum değildir (Kocaoğlu & Şahin, 2024; Taşdelen-Teker & Doğan, 2015). Bu tür maddelerde ortak köke bağlı bir madde diğer bir maddenin çözümüne ipucu vermemelidir. Ortak köke bağlı maddelerin farklı düzeydeki becerileri ya da farklı öğrenme çıktılarını ölçmesinin kapsama dayalı kanıt oluşturacağı söylenebilir. Bununla birlikte madde güçlüklerinin de birbirinden farklılaşması uygun olacaktır. Örneğin, ortak köke bağlı bir madde hatırlama düzeyinde ve kolayken, diğer bir madde analiz etme düzeyinde ve orta güçlükte olabilir. Diğer bir deyişle madde dağılımının heterojen olması beklenir. Ayrıca ortak köke bağlı dörtten fazla madde yer almamalıdır. Şekil 6'da ortak köklü çoktan seçmeli bir madde örneği verilmiştir.

Şekil 6

Ortak Köklü Çoktan Seçmeli Madde Örneği

1 ve 2. soruları aşağıdaki bilgilere göre birbirinden bağımsız olarak cevaplayınız.

Bir araştırmacı, üç farklı şehirde (A, B ve C) yaşayan üç arkadaşın (Ali, Berke ve Cem) farklı günlerde (pazartesi, salı ve çarşamba) farklı etkinliklere (sinema, tiyatro ve konser) gittiklerini tespit etmiştir. Buna göre aşağıdaki bilgiler bilinmektedir:

Ali, salı günü sinemaya gitmemiştir.
B şehrinde yaşayan kişi, çarşamba günü konsere gitmiştir.
Cem, tiyatroya gitmiştir ve A şehrinde yaşamamaktadır.
Sinemaya giden kişi, pazartesi günü etkinliğe katılmıştır.
Berke, C şehrinde yaşamamaktadır.

1) Berke hangi şehirde ve hangi günde etkinliğe katılmıştır?

A) A - Pazartesi
*B) B - Çarşamba
C) C şehri - Pazartesi
D) A şehri - Çarşamba
E) B şehri - Salı

2) Ali'nin gittiği etkinlik ve Cem'in etkinliğe gittiği gün aşağıdakilerden hangisinde sırasıyla birlikte doğru verilmiştir?

A) Sinema - Pazartesi
*B) Sinema - Salı
C) Sinema - Çarşamba
D) Konser - Pazartesi
E) Konser – Salı

*: Doğru cevap

Ortak Seçenekli Çoktan Seçmeli Madde

Ortak köklü çoktan seçmeli maddenin aksine bu madde türünde seçenekler ortaktır ve ortak seçeneklere bağlı iki veya ikiden fazla madde bulunmaktadır. Bu tür maddelerin üstün yönü uzmanlık gerektiren çeldirici yazma yükünü azaltarak hazırlanması bakımından kullanışlılığı artırmasıdır. Yine de her bir seçeneğin her madde için geçerli olmasına dikkat edilmelidir (Turgut & Baykul, 2013). Ortak seçenekli maddelerde madde sayısını beş seçenek için üç ile sınırlamak önemlidir. Çünkü bireyler, seçeneklerin tekrarlı kullanılmayacağını düşünerek buna göre tepki vereceklerdir ve böylelikle şans başarısı artacaktır. Bu seçeneklerin bir kez kullanılacağı anlamına gelmemelidir. Hatta böyle bir durum varsa madde eşleştirmeli madde türüne dönecektir. Şekil 7'de ortak seçenekli çoktan seçmeli bir madde örneği verilmiştir.

Şekil 7

Ortak Seçenekli Çoktan Seçmeli Madde Örneği

1, 2 ve 3 numaralı soruları aşağıdaki seçeneklere göre cevaplayınız.
A) 1 B) 2 C) 3 D) 4 E) 5

1) Gerçel sayılar kümesi üzerinde tanımlı f fonksiyonu için

$f(x + 1) = f(x) + 2$
$f(0) = 1$

eşitlikleri sağlanıyor.
Buna göre f(2) ifadesinin değeri kaçtır?

(Doğru cevap: C)

2) a ve b gerçel sayılar olmak üzere f fonksiyonu gerçel sayılar kümesi üzerinde tanımlıdır.

$f(x) = ax + b$ doğrusal fonksiyonu için
$f(1) = 3$
$f(2) = 5$

eşitlikleri sağlanıyor.
Buna göre a kaçtır?

(Doğru cevap: B)

3) f fonksiyonu gerçel sayılar kümesi üzerinde tanımlıdır.
Buna göre $f(x) = x^2 - 2x + 1$ fonksiyonu için $f(k) = 0$ olduğuna göre k kaçtır?

(Doğru cevap: A)

2.1.2. Doğru-Yanlış Türü Maddeler

Doğru-yanlış türü maddeler, çoktan seçmeli maddelerin iki seçenekli özel bir hâlidir. Bu sebeple iki seçenekli madde (binary-choice item) olarak da adlandırılmaktadır (Popham, 2014). Bireylere bir önerme sunulur ve bireylerden yanıtlarını doğru-yanlış (true-false / right-wrong / correct-incorrect) veya evet-hayır arasında seçim yaparak vermesi beklenir. Şekil 8'de bir doğru-yanlış maddesi örneği sunulmuştur.

Şekil 8

Doğru-Yanlış Maddesi Örneği

(D) (Y) Araba Sevdası adlı eserin yazarı Recaizade Mahmut Ekrem'dir.

Bu tür maddeler, sınıflama yapmayı gerektiren bilgileri yoklamada en uygun madde türüdür (Thorndike & Thorndike-Christ, 2010). Örneğin, matematik alanında verilen bir sayının rasyonel mi irrasyonel mi olduğunun yoklanmasında doğru-yanlış madde türünün kullanımı oldukça uygundur. Çoğunlukla bilgi ve kavrama düzeyindeki alt düzey düşünme becerileri ölçmek için kullanılan bu madde türü, tıpkı çoktan seçmeli madde türünde olduğu gibi iyi geliştirildiğinde üst düzey düşünme becerilerini de ölçebilmektedir (Frisbie, 1992; Russell & Airasian, 2012).

2.1.3. Eşleştirme Maddeleri

Eşleştirme maddeleri (matching exercises / matching items) bireyin birinci listeden (öncül/önerme listesi) ikinci listeye (seçenekler-cevap listesi) eşleştirme yapmalarını gerektiren iki liste barındıran bir madde türüdür (Nitko & Brookhart, 2014; Thorndike & Thorndike-Christ, 2010). Eşleştirmeli madde türü, hatırlama düzeyindeki bir dizi çoktan seçmeli maddenin birleşimi olarak düşünülebilir. Şekil 9'da eşleştirmeli madde örneği sunulmuştur.

Şekil 9

Eşleştirmeli Madde Örneği

Yönerge: *Şiirler* sütunundaki her bir şiirin solundaki boşluğa ilgili eserin sahibinin *Şairler* sütunundaki harfini yazınız. *Şairler* sütunundaki şairlerin hepsi bir kez kullanılabilir ve her şair kullanılmayabilir.	
Şiirler	Şairler
_____1. Ah Nijad _____2. Münacaat _____3. Sahra _____4. Vaveyla	A. Abdülhak Hamit Tarhan B. Muallim Naci C. Namık Kemal D. Recaizade Mahmut Ekrem E. Şinasi F. Ziya Paşa
Doğru Eşleştirme: 1-D, 2-E, 3-A, 4-C	

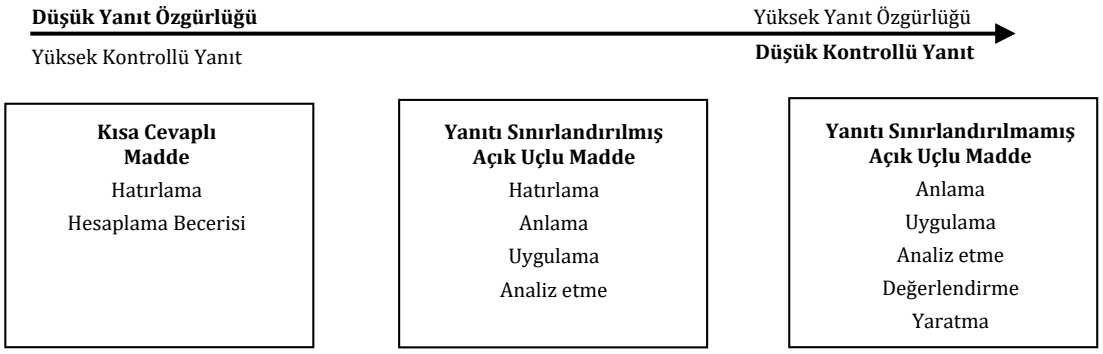
Eşleştirme türü maddeler, alt düzey düşünme becerilerini ölçme ile sınırlıdırlar (Russell & Airasian, 2012). Bireyin ilişkilendirme yapmasına yönelik yoklanacak bir öğrenme çıktısının değerlendirilmesinde, doğru-yanlış ve çoktan seçmeli madde türlerinden daha çok etkilidir.

2.2. Katkı Gerektiren Madde Türleri

Katkı gerektiren madde türleri, kısa cevaplı madde ve açık uçlu madde türlerinden oluşmaktadır. Bu madde türleri ölçtükleri bilişsel alan düzeyi, verilebilecek yanıtın kontrol edilebilirliği ve yanıt özgürlüğü açısından birbirinden ayrışmaktadır. Bu duruma ilişkin diyagram Şekil 10'da verilmiştir.

Şekil 10

Katkı Gerektiren Madde Türlerinin Ölçebildiği Bilişsel Alan Düzeyleri ile Verilebilecek Yanıtın Özelliği (Waugh & Gronlund, 2013'ten uyarlanmıştır.)



Bu madde türlerinin özellikleri karşılaştırmalı olarak katkı gerektiren madde türlerinin üstün ve sınırlı yönleri alt başlığında verilmiştir.

2.2.1. Kısa Cevaplı Madde

Kısa cevaplı madde türünde bireylerin bir soruya veya eksik bir ifadeye cevap olarak bir sayı, sembol, kelime, söz öbeği veya bir cümle yazmaları beklenir. Soru biçiminde olabileceği gibi boşluk doldurma biçiminde de kullanılabilir. Ancak kısa cevaplı madde türünde genelde soru biçiminde olan tercih edilmektedir. Şekil 11'de kısa cevaplı madde örnekleri sunulmuştur.

Şekil 11

Kısa Cevaplı Madde Örnekleri

Osmanlı Devleti _____ yılında kurulmuştur. *(Boşluk doldurma biçiminde)*
Osmanlı Devleti kaç yılında kurulmuştur? *(Soru biçiminde)*

Kısa cevaplı madde türü, hatırlama ve anlama gibi alt düzey düşünme becerilerini ölçmede kullanıldığı gibi, cevaplanması için denklem kurma, problem çözme becerisi gibi beceriler gerektiğinde üst düzey düşünme becerilerini de ölçebilir (Nitko & Brookhart, 2014).

2.2.2. Açık Uçlu Madde

Açık uçlu madde türü (essay item / open-ended item) bireylerin cevaplarını ve düşüncelerini organize ederek kendilerinin yapılandırmasına imkân tanıyan bir madde türüdür (Thorndike & Thorndike-Christ, 2010). Açık uçlu maddeler yanıtı sınırlandırılmış ve yanıtı sınırlandırılmamış olmak üzere ikiye ayrılır (Nitko & Brookhart, 2014; Popham, 2014).

Yanıtı sınırlandırılmış açık uçlu maddelerde bireylerin verebileceği yanıtın içeriği ve biçimi belirlenmiştir. Yanıtı sınırlandırılmamış açık uçlu maddelerde ise bunun aksine, mümkün olduğunca yanıt çerçevesi serbest bırakılmıştır. Şekil 12'de yanıtı sınırlandırılmış ve yanıtı sınırlandırılmamış açık uçlu madde örnekleri sırasıyla sunulmuştur.

Şekil 12

Açık Uçlu Madde Örnekleri (Yanıtı Sınırlandırılmış ve Sınırlandırılmamış)

> Öğretmenlik meslek etiğinin temel ilkelerinden üçünü örnek olayla açıklayınız. (Her bir örnek olay 30 kelimeyi geçmemelidir.) *(Yanıtı sınırlandırılmış)*
>
> Sosyoekonomik düzeyi farklı öğrencilerin bulunduğu bir sınıfta, veliler Öğretmenler Günü'nde sınıf öğretmenine hediye almak istemektedir. Sosyoekonomik düzeyi yüksek olan bazı velilerin bireysel ve daha değerli hediyeler verme eğiliminde olduğu görülmektedir. Bu durumda ortaya çıkabilecek durumları öğretmenlik meslek etiğinin temel ilkelerinden en uygun üçünü ele alarak değerlendiriniz. Bir öğretmen olarak nasıl bir tutum sergilenmesi gerektiğini gerekçeleriyle açıklayınız. Cevabınızı bir sayfayı aşmayacak şekilde hazırlayınız *(Yanıtı sınırlandırılmamış)*

Yazma becerisinin ölçülmesinde kullanılan tek geleneksel madde türü açık uçlu maddedir. Bununla birlikte analiz etme, değerlendirme ve yaratma düzeyindeki becerileri ölçmede kullanılır. Diğer bir deyişle açık uçlu madde türü, bireyin üst düzey düşünme becerilerini ölçmede oldukça etkilidir.

3. MADDE YAZIM İLKELERİ

Seçme gerektiren ve katkı gerektiren maddelerin amacına uygun ve bilimsel açıdan doğru olması kadar ölçme ve değerlendirme açısından da iyi tasarlanmış olması oldukça önemlidir. Bir maddenin ölçme ve değerlendirme açısından iyi tasarlanmış olması madde yazım ilkelerine uyulmasını gerektirir. Madde yazım ilkeleri hem seçme hem katkı gerektiren maddeler için ortak olabileceği gibi sadece tüm madde türlerine özgü

madde yazım ilkeleri de bulunmaktadır. Bu bölümde ortak madde yazım ilkelerine değinilerek akabinde her bir madde türüne özgü madde yazım ilkeleri ayrı ayrı sunulmuştur. Tablo 1'de seçme ve katkı gerektiren madde türlerinin tamamı için uyulması gereken madde yazım ilkeleri verilmiştir.

Tablo 1

Seçme ve Katkı Gerektiren Madde Türlerine Yönelik Ortak Madde Yazım İlkeleri

Ortak Madde Yazım İlkeleri	▪ Madde açık ve net olarak ifade edilmelidir. ▪ Maddede kullanılan terim, sembol, görsel ve kelimeler alanyazına ve hedef kitleye uygun olmalıdır. ▪ Maddede yoklanan bilgi, ölçülmeye değer nitelikte olan bir öğrenme çıktısını veya beceriyi ölçmelidir. ▪ Madde belirli bir konu / durum ile sınırlı olmalıdır. ▪ Kanıya dayalı, uzmanlar arasında birlik bulunmayan, tartışmaya açık bir konu kullanılmamalıdır. ▪ Madde ölçülen özellik dışındaki başka özellikleri ölçmeyecek şekilde hazırlanmalıdır. ▪ Madde; yazım, noktalama ve dil bilgisi açısından doğru yazılmalıdır. ▪ Maddede olumsuz ifade kullanılmasından kaçınılmalıdır. ▪ Madde doğru cevaba ipucu vermeyecek şekilde yapılandırılmalıdır. ▪ Maddenin cevaplanması için gerekli bilgiler eksiksiz verilmelidir. ▪ Maddenin cevaplanması için kullanılmayacak bilgilere / görsellere yer verilmesinden kaçınılmalıdır. ▪ Maddelerin hazırlanmasında madde başına düşen ortalama cevaplama süresi göz önünde bulundurulmalıdır. ▪ Vurgulanmak istenen ifadeler varsa biçimsel olarak vurgulanmalıdır.

Tablo 1'de listelenen ortak madde yazım ilkelerine ek olarak madde eğer geniş ölçekli bir testte kullanılacaksa toplumun tüm kesimleri göz önünde bulundurularak testteki maddelerde kültürel/toplumsal hassasiyet oluşturulabilecek unsurlardan kaçınılması gerekmektedir. Ayrıca toplumun bir kesimini ya da bir grubu rencide edecek ya da övecek ifadelerden de kaçınılması önemlidir. Bu madde yazım ilkelerinin dikkate alınmadan hazırlandığı maddelerin geniş ölçekli bir testte yer alması, testi alan bireylerin dikkatini dağıtmasına sebep olabilir. Bu da testin bütünüyle elde edilen

ölçümler için bir güvenirlik problemi oluşturabilir. Diğer yandan bu tür ifadelerin maddelerde yer alması toplumsal spekülasyonlara sebep olabilir.

Bir madde olumsuz ifade içeriyorsa bu maddenin olumlu ifade içeren maddelere göre ayırt ediciliğinin daha düşük olduğu ve daha zor olduğu alanyazında bilinmektedir (Barker & Ebel, 1982; Dudycha & Carpenter, 1973; Rush ve ark., 2016). Dolayısıyla ortak madde yazım ilkelerinde bahsedildiği üzere maddede olumsuz ifadeler kullanmaktan kaçınılmalıdır. Ancak bazı durumlarda olumsuz madde kullanımı kaçınılmaz olmaktadır. Böyle durumlarda olumsuz ifadeler vurgulanmalı (altı çizilmeli ya da kalın punto kullanılmalı) ve çift olumsuz ifadeler kullanılmamalıdır.

Madde ile ölçülen özellik dışındaki başka özelliklerin yoklanması o maddeyle elde edilen ölçümlerin geçerliğini düşürür. Bu nedenle bireylerin ölçülecek özellik dışındaki özellikleri kullanmalarını gerektirecek maddelerden kaçınılmalıdır. Örneğin, felsefe alanında hazırlanan bir madde yalnızca okuduğunu anlama becerisiyle çözülebiliyorsa bu maddenin amaca hizmet etme derecesinin düşük olduğu söylenebilir. Bununla birlikte bir madde farklı gruplar arasında yansızlığı sağlayacak şekilde hazırlanmalıdır. Diğer bir deyişle iki farklı grupta yer alan bireyler benzer yetenek düzeyinde olmalarına rağmen, mensubu oldukları gruba bağlı olarak maddeyi doğru cevaplama olasılıkları farklılaşmamalıdır. Aksi bir durum yanlılığa neden olabilmektedir. Örneğin, matematik alanında hazırlanan bir maddenin bağlamında futbola ilişkin ögelerin bulunması erkekleri daha avantajlı kılıyorsa bu madde yanlı bir madde olabilir. Bu durumda ölçme sonuçlarına yanlılık karışarak ölçme sonuçlarının geçerliği zedelenir.

Madde yazımında görsel kullanımına özen gösterilmelidir. Kullanılan görseller, madde yazımının temel ilkelerinden biri olan maddenin açık ve net ifade edilmesi ilkesine hizmet edecek biçimde tasarlanmalı ve maddenin anlaşılmasını kolaylaştırmalıdır. Görsellerin, ölçülmek istenen bilgi veya beceriyi destekleyerek sorunun çözümüne anlamlı katkı sağlaması beklenmektedir. Nitekim Aydoğan ve Gündüz (2024), görsel içeren maddelerden oluşan testlerin güvenirlik katsayılarının, aynı maddelerin görsel içermeyen versiyonlarından oluşan testlere göre anlamlı düzeyde daha yüksek olduğunu göstermiştir. Bu bulgu, uygun biçimde kullanılan görsellerin yalnızca maddeyi anlaşılır hâle getirmekle kalmayıp, ölçme sonuçlarının güvenirliğine de olumlu katkı sağlayabileceğine işaret etmektedir. Buna karşın, yalnızca sayfa düzenini zenginleştirmek ya da estetik bir görünüm sağlamak amacıyla kullanılan, sorunun çözümüne herhangi bir katkısı bulunmayan görsellere maddelerde yer verilmemelidir. Dolayısıyla madde yazımında kullanılacak görseller, sorunun çözümüne sağladıkları katkı ve ölçülen yapıyla olan ilişkileri dikkate alınarak seçilmeli ve hazırlanmalıdır.

3.1. Seçme Gerektiren Madde Türlerine İlişkin Madde Yazım İlkeleri

3.1.1. Çoktan Seçmeli Maddelere İlişkin Madde Yazım İlkeleri

Çoktan seçmeli madde yazımında ortak madde yazım ilkeleriyle birlikte bu madde türüne özgü madde yazım ilkeleri de bütüncül olarak ele alınmalıdır. Çoktan seçmeli madde türünde madde köküne ek olarak seçeneklerin (doğru cevap ve çeldiricilerin) ya da öncüllerin de yazımında dikkate alınması gereken hususlar bulunmaktadır. Tablo 2'de çoktan seçmeli bir maddenin tüm ögelerine ilişkin madde yazım ilkeleri verilmiştir.

Tablo 2

Çoktan Seçmeli Madde Türüne Özgü Madde Yazım İlkeleri

Çoktan Seçmeli Maddelere Özgü Madde Yazım İlkeleri	▪ Madde kökü ile seçenekler birbiriyle bütünlük sağlamalıdır. ▪ Seçenekler uzunluk, içerik, ifade ve yapı bakımından birbirleriyle tutarlı ve uyumlu olmalıdır. ▪ Madde kökü ile seçenekler arasında ipucu oluşturacak anahtar kelimelere yer verilmemelidir. ▪ Seçenekler birbirinden bağımsız ve biri diğerini içermeyecek şekilde hazırlanmalıdır. ▪ Seçeneklerden ikisi birbirine zıt olmamalıdır. ▪ Seçeneklerde "hepsi" veya "hiçbiri" kullanılmamalıdır. ▪ Seçeneklerde tekrar eden ifadeler kullanmak yerine tekrarlı ifade madde kökünde kullanılmalıdır. ▪ Seçenekler alfabetik ya da sayısal olarak mantıksal bir sırada verilmelidir. ▪ Seçeneklerde belirsizlik veya kesinlik bildiren ifadelerden kaçınılmalıdır. ▪ Alanyazında bulunmayan kavramlar çeldirici olarak yazılmamalıdır. ▪ Çeldiriciler akla yatkın / makul hazırlanmalıdır. ▪ Maddenin tartışmasız tek bir doğru cevabı olmalıdır. ▪ Doğru cevapta konuya hâkim bireyleri yanıltabilecek ifadelere yer verilmemelidir.

Çoktan seçmeli madde yazımında seçeneklerin hem madde kökü ile hem de kendi arasında tutarlı olmasına özen gösterilmelidir. Madde kökünden farklı bir şekilde sunulmuş seçenekler bireylerin konuyu bilmeden doğru yanıtı bulmalarına neden olacaktır. Diğer yandan seçeneklerin birbiriyle uyumsuz olması durumu da benzer şekilde konuya hâkim olmayan bireylerin seçenekleri elemesine neden olarak şans

başarısını artıracaktır. Ayrıca çoktan seçmeli madde yazımında bir seçeneğin diğer seçeneği içermeyecek şekilde hazırlanması önerilmektedir ve bu husus sayısal değer içeren seçenekler için de geçerlidir. Örneğin, bir aşının etki süresinin yıl olarak sorulduğu bir çoktan seçmeli maddede seçenekler sırasıyla A) 1, B) 2, C) 3, D) 4, E) 5 olsun, bu soruda doğru cevap 3 yıl ise doğru cevap 1 ve 2 seçeneklerini de içerdiği için sorunun hem A hem B hem de C seçenekleri doğrudur ve bu tür bir madde bilimsel olarak hatalı olacaktır. Ancak böyle bir madde için madde kökünde aşının maksimum etki süresi olarak ifade edilseydi seçeneklerin birbirini içermesi engellenmiş olurdu.

Çoktan seçmeli madde yazımında dikkat edilmesi gereken başka bir husus olan *"Seçenekler alfabetik ya da sayısal olarak mantıksal bir sırada verilmelidir."* ilkesine bazı durumlarda uyulmayabilir. Örneğin, ÖSYM'nin gerçekleştirdiği sınavlar gibi geniş ölçekli ve yüksek riskli sınavlarda sınav güvenliği bakımından her adaya özgü farklı soru kitapçıkları oluşturulması adına bu ilkeden vazgeçilebilir. İlkelerden *"Maddenin tartışmasız tek bir doğru cevabı olmalıdır."* ilkesi ise yalnızca tek doğru cevabı olan *geleneksel çoktan seçmeli maddeler* için geçerlidir. Ancak birden fazla doğru seçeneğe sahip bir çoktan seçmeli madde türü olan *çoklu seçim gerektiren çoktan seçmeli maddeler* için geçerli değildir.

Aşağıda çoktan seçmeli madde türünde geliştirilmiş kusurlu bir madde örneği sunulmuştur.

!

Bir testin sahip olması gereken bazı özellikler vardır.

Buna göre, bir teste yeni madde eklendiğinde aşağıdakilerden hangisinin ortaya çıkması beklenir?

A) Testin güvenirliği artar.

B) Testin güvenirliği azalır.

C) Testin objektifliğinin azalması

D) Testin kullanışlılık düzeyinin artırılarak etkin kullanımının sağlanması

E) Testin geçerliği azalır.

*Doğru cevap: A

Bu örnekte ortak madde yazım ilkelerinden *"Maddenin cevaplanması için kullanılmayacak bilgilere / görsellere yer verilmesinden kaçınılmalıdır."* ihlal edilmiştir. Çünkü madde kökünde bulunan "Bir testin sahip olması gereken bazı özellikler vardır." ifadesi sorunun çözümüne katkısı bulunmayan gereksiz bir ifadedir. Diğer yandan *"Madde; yazım, noktalama ve dil bilgisi açısından doğru yazılmalıdır."* ilkesi de ihlal edilerek madde kökünde edattan sonra noktalama işareti kullanılmıştır.

Çoktan seçmeli madde yazım ilkeleri bağlamında incelendiğinde örnek maddede *"Seçeneklerde tekrar eden ifadeler kullanmak yerine tekrarlı ifade madde kökünde kullanılmalıdır."* ilkesi ihlal edilmiştir. Çünkü tüm seçenekler "Testin" kelimesiyle

başlamaktadır. Bunun yerine seçeneklerdeki tekrarlı ifadelerin madde kökünde verilmesi önerilmektedir. Böylelikle okuma yükünün azaltılması sağlanır. Öte yandan örnek maddede *"Seçeneklerden ikisi birbirine zıt olmamalıdır."* ilkesi de ihlal edilmiştir. Çünkü A ve B seçeneklerinden birinde güvenirliğin arttığı ifadesi yer alırken diğerinde azaldığı ifadesi yer almaktadır. Bu tür zıt seçeneklerin kullanılması testi alan bireylerin maddeyi iki seçeneğe düşürmelerine sebep olmakta ve dolayısıyla bireylerin şans başarısını artırmaktadır. Böyle bir durumda madde beş seçenek yerine iki seçenekli çalışmakta ve doğru-yanlış türünde bir maddeden farksız görünmektedir. Ancak kimi durumlarda seçeneklerin zıt kullanılması zorunlu hâle gelebilir. Bu tür durumlar varsa başka iki seçeneğin de birbirinin zıttı olarak tasarlanması önerilmektedir. Örnek maddede ihlal edilen bir diğer ilke ise *"Seçenekler uzunluk, içerik, ifade ve yapı bakımından birbirleriyle tutarlı ve uyumlu olmalıdır."* ilkesidir. Örnek maddede D seçeneği (Testin kullanışlılık düzeyinin artırılarak etkin kullanımının sağlanması) diğer seçeneklerden oldukça uzundur. Dahası seçeneklerin uzunluğunun yanı sıra ifade ve yapı bakımından da farklılıklar söz konusudur. A, B ve E seçeneklerinde çekimli fiiller kullanılmışken C ve D seçeneklerinde fiilimsiler kullanılmıştır. Örnek maddenin bu kusurları giderilmiş hâli aşağıda sunulmuştur.



Bir teste yeni madde eklendiğinde testin hangi psikometrik özelliğinin doğrudan olumlu etkilenmesi beklenir?

*A) Duyarlılık

B) Geçerlik

C) Objektiflik

D) Kullanışlılık

E) Kararlılık

*: Doğru cevap

3.1.2. Doğru-Yanlış Maddelerine İlişkin Madde Yazım İlkeleri

Doğru-yanlış türü madde yazımında ortak madde yazım ilkeleriyle birlikte bu madde türüne özgü madde yazım ilkeleri de bütüncül olarak ele alınmalıdır. Tablo 3'te doğru-yanlış madde türüne özgü madde yazım ilkeleri verilmiştir.

Tablo 3

Doğru-Yanlış Madde Türüne Özgü Madde Yazım İlkeleri

Doğru-Yanlış Maddelerine Özgü Madde Yazım İlkeleri	▪ Verilen önerme kesinlikle yanlış ya da kesinlikle doğru olmalıdır. ▪ Kaynaklardan doğrudan alıntı yapılmamalıdır. ▪ Önermelerde belirsizlik veya kesinlik bildiren ifadelerden kaçınılmalıdır. ▪ Madde, tek bir önerme içermelidir.

Doğru-yanlış madde türünden oluşan bir testte cevabı doğru ve yanlış olan maddelerin yaklaşık aynı uzunlukta olmasına dikkat edilmelidir. Test içinde doğru ve yanlış sayısı yaklaşık olarak birbirine eşit olmalıdır. Cevap anahtarı belirli bir tekrar eden örüntüde oluşturulmamalı ve art arda dörtten fazla doğru veya yanlış seçeneği gelmesinden kaçınılmalıdır.

Aşağıda doğru-yanlış madde türünde geliştirilmiş kusurlu bir madde örneği sunulmuştur.



(D)	(Y)	En büyük gezegen olan Dünya, genellikle Güneş'in etrafında döner.

Bu örnekte hem en büyük gezegenin Dünya olduğuna ilişkin hem de Dünya'nın Güneş'in etrafında dönmesine ilişkin iki önerme bulunmaktadır. Dahası Dünya'nın Güneş'in etrafında genellikle döndüğü önermesi yer almaktadır. Buna göre sırasıyla *"Madde, tek bir önerme içermelidir."* ve *"Önermelerde belirsizlik veya kesinlik bildiren ifadelerden kaçınılmalıdır."* ilkeleri ihlal edilmiştir. Örnek maddenin bu kusurları giderilmiş hâli aşağıda sunulmuştur.



(D)	(Y)	Dünya, Güneş'in etrafında döner.

3.1.3. Eşleştirmeli Maddelere İlişkin Madde Yazım İlkeleri

Eşleştirmeli maddelerin yazımında ortak madde yazım ilkelerine ek olarak bu madde türüne özgü madde yazım ilkeleri de dikkate alınmalıdır. Tablo 4'te eşleştirmeli maddelere özgü madde yazım ilkeleri verilmiştir.

Tablo 4

Eşleştirmeli Madde Türüne Özgü Madde Yazım İlkeleri

Eşleştirmeli Maddelere Özgü Madde Yazım İlkeleri	▪ Listeler kendi içinde homojen olmalıdır. ▪ Cevap listesi öncül-önerme listesinden uzun olmalıdır. ▪ Cevap listesi ve öncül-önerme listesi arasında sistematik ipucu verecek örüntüler bulunmamalıdır. ▪ Öncül-önerme listesi veya cevap listesi alfabetik ya da sayı sırası gibi belirli bir sırada verilmelidir. ▪ Öncül-önerme listesi rakamla, cevap listesi büyük harflerle belirtilmelidir. ▪ Öncül-önerme sayısı ondan fazla olmamalıdır.

Eşleştirmeli madde türünde geleneksel olarak her bir cevap tek bir kez kullanılırken yönergede belirtilmek kaydıyla bir cevap birden fazla önerme için de kullanılabilir (Thorndike & Thorndike-Christ, 2010).

Aşağıda eşleştirmeli madde türünde geliştirilmiş kusurlu bir madde örneği sunulmuştur.



Yönerge: *Kavramlar* sütunundaki her bir kavramın solundaki boşluğa *Tanımlar* sütunundan ilgili kavramın açıklaması olduğunu düşündüğünüz ifadenin harfini yazınız.	
Kavramlar	Tanımlar
_____a. Standart hata _____b. Güvenirlik _____c. Madde güçlüğü _____d. Madde ayırt ediciliği _____e. Geçerlik	1. Ölçme sonuçlarına karışan hata miktarının tahmin edilmesini sağlar. 2. Ölçme sonuçlarının tutarlılığı ve tesadüfi hatalardan arınıklık derecesidir. 3. p = Maddeyi doğru cevaplayan bireylerin sayısı / Maddeyi alan tüm bireylerin sayısı 4. Bir maddenin yüksek ve düşük başarı düzeyindeki bireyleri ayırt etme gücüdür. 5. Ölçme aracının ölçmek istediği özelliği ne derece doğru ölçtüğünü ifade eder.
Doğru Eşleştirme: a-1, b-2, c-3, d-4, e-5	

Bu örnekte ortak madde yazım ilkelerinden "*Madde doğru cevaba ipucu vermeyecek şekilde yapılandırılmalıdır."* ihlal edilmiştir. Çünkü öncül listesinde dördüncü sıradaki "d. Madde ayırt ediciliği"ne, cevap listesindeki dördüncü tanımda bulunan "ayırt etme gücü" ifadesi doğru eşleştirmeye ipucu vermektedir. Ayrıca maddenin yönergesinde seçeneklerin kaç defa kullanılabileceği ya da her tanımın kullanılmayabileceğine ilişkin yönlendirmeler bulunmamaktadır.

Eşleştirmeli madde yazım ilkeleri bağlamında incelendiğinde örnek maddede *"Listeler kendi içinde homojen olmalıdır."* ilkesi ihlal edilmiştir. Çünkü seçeneklerde ifade ve yapı bakımından farklılıklar mevcuttur. Üçüncü seçenek bir formül olup diğer seçeneklerden farklı yapıdadır. Diğer dört seçeneğin her biri birer cümle olmasına karşın birinci ve beşinci seçenek fiil cümlesi, ikinci ve dördüncü seçenek isim cümlesidir. Örnek maddede "*Cevap listesi ve öncül-önerme listesi arasında sistematik ipucu verecek örüntüler bulunmamalıdır.*" ilkesi ihlal edilmiştir. Çünkü öncül-önerme listesindeki sıra ile cevap listesindeki sıranın birbiriyle aynı olduğu görünmektedir. Bu doğrultuda maddenin cevap anahtarının belirli bir örüntü içerdiği söylenebilir. Örnek maddede *"Cevap listesi öncül-önerme listesinden uzun olmalıdır."* ilkesi ihlal edilerek öncül ve seçenek listesi aynı uzunlukta verilmiştir. Ayrıca öncüller karışık sırada verilip *"Öncül-önerme listesi veya cevap listesi alfabetik ya da sayı sırası gibi belirli bir sırada verilmelidir."* ilkesi ihlal edilmiştir. Öte yandan *"Öncül-önerme listesi rakamla, cevap listesi büyük harflerle belirtilmelidir*." ilkesi de ihlal edilerek öncül-önerme listesi küçük harfle, cevap listesi ise rakamlarla belirtilmiştir. Örnek maddenin bu kusurları giderilmiş hâli aşağıda sunulmuştur.



Yönerge: *Kavramlar* sütunundaki her bir kavramın solundaki boşluğa *Tanımlar* sütunundan ilgili kavramın açıklaması olduğunu düşündüğünüz ifadenin harfini yazınız. *Tanımlar* sütunundaki ifadelerin hepsi bir kez kullanılabilir ve her tanım kullanılmayabilir.	
Kavramlar	Tanımlar
_____1. Madde güçlüğü _____2. Madde ayırt ediciliği _____3. Madde varyansı _____4. Standart hata	A. Gözlenen puanların gerçek puandan sapma miktarıdır. B. Ölçme sonuçlarının tutarlılığı ve tesadüfi hatalardan arınıklık derecesidir. C. Maddeyi doğru yanıtlayan bireylerin maddeyi yanıtlayan tüm bireylere oranıdır. D. Ölçme sonuçlarının amaca hizmet etme derecesidir. E. Bir maddenin testi alan bireyler arasındaki farklılığı ortaya koyma derecesidir. F. Bir maddenin yüksek ve düşük başarı düzeyindeki bireyler arasındaki farkı ortaya koyma derecesidir.
Doğru Eşleştirme: 1-C, 2-F, 3E, 4-A	

3.2. Katkı Gerektiren Madde Türlerine İlişkin Madde Yazım İlkeleri

3.2.1. Kısa Cevaplı Maddelere İlişkin Madde Yazım İlkeleri

Kısa cevaplı madde yazımında ortak madde yazım ilkelerinin yanı sıra bu madde türüne özgü madde yazım ilkeleri de göz önünde bulundurulmalıdır. Tablo 5'te hem soru biçiminde hem de boşluk doldurma biçimindeki kısa cevaplı maddelere ilişkin madde yazım ilkeleri verilmiştir.

Tablo 5

Kısa Cevaplı Madde Türüne Özgü Madde Yazım İlkeleri

Kısa Cevaplı Maddelere Özgü Madde Yazım İlkeleri	• Madde, cevabı olabildiğince kısa olacak şekilde yapılandırılmalıdır. • Beklenen cevap sayısalsa maddede istenen kesinlik derecesi ve birimler belirtilmelidir. • Boşluk doldurma biçimindeki maddeler, tek doğru cevabı olacak şekilde hazırlanmalıdır. • Boşluk doldurma biçimindeki maddelerde en fazla iki boşluk bırakılmalıdır. • Boşluk doldurma biçimindeki maddelerde boşluk cümlenin sonuna doğru bırakılmalıdır. • Boşluk doldurma biçimindeki maddelerde hem madde içinde hem de testin tamamında boşlukların uzunluğu eşit olmalıdır. • Boşluk doldurma biçimindeki maddelerde boşluktan sonra doğru yanıta ipucu verecek ekler kullanılmamalıdır.

Boşluk doldurma ve soru biçimindeki kısa cevaplı maddeler bazı durumlarda birbirleri yerine kullanılabilir. Özellikle küçük yaş gruplarında boşluk doldurma yerine soru biçimindeki kısa cevaplı madde türü tercih edilmelidir (Popham, 2014). Boşluk doldurma biçimindeki maddelerde genelde bir boşluk bırakılması önerilirken maddenin uzunluğuna bağlı olarak en fazla iki boşluk bırakılabilir. Eğer ikiden fazla yanıt gerektiriyorsa boşluk doldurma yerine soru biçimindeki kısa cevaplı madde türü kullanılmalıdır (Chappuis ve ark., 2012).

Aşağıda soru biçimindeki kısa cevaplı madde türünde geliştirilmiş kusurlu bir madde örneği sunulmuştur.



Cronbach'ın alfa katsayısının kabul edilebilir bir güvenirlik düzeyinde olmadığını söylememek için ölçüt değer kaçtır?

Bu örnekte ortak madde yazım ilkelerinden *"Kanıya dayalı, uzmanlar arasında birlik bulunmayan, tartışmaya açık bir konu kullanılmamalıdır."* ihlal edilmiştir. Çünkü Cronbach'ın alfa katsayısının kabul edilebilir düzeyde olduğunu söylemek için ölçüt değer alanyazında farklılaşmakta olup bu konu kitabın 3. bölümünde de değinildiği gibi tartışmalıdır. Ortak madde yazım ilkelerinden ihlal edilen bir diğer ilke ise *"Maddede olumsuz ifade kullanılmasından kaçınılmalıdır."* ilkesidir. Dahası bu örnekte olduğu gibi çift olumsuz ifade kullanılması soruyu karmaşıklaştırmakta ve bireyin anlamasını zorlaştırmaktadır.

Kısa cevaplı madde yazım ilkeleri bağlamında incelendiğinde örnek maddede *"Beklenen cevap sayısalsa maddede istenen kesinlik derecesi ve birimler belirtilmelidir."* ilkesi ihlal edilmiştir. Çünkü sözel olarak 1'e yakın cevabının verilmesi de teorik olarak doğru olup birey böyle bir cevap verdiğinde puanlanması gerekir. Bu nedenle bireyden beklenen yanıtın sınırlarının belirlenmesi önerilmektedir. Örnek maddenin bu kusurları giderilmiş hâli aşağıda sunulmuştur.

✓ Nunnally'ye (1978) göre Cronbach'ın alfa katsayısı için ölçüt değer kaçtır? (Virgülden sonra iki basamaklı yazınız)

Aşağıda boşluk doldurma biçimindeki kısa cevaplı madde türünde geliştirilmiş kusurlu bir madde örneği sunulmuştur.

! Kaynağı, yönü ve miktarı tam olarak belirlenemeyen, ölçme sonuçlarına tesadüfen karışan hatalara ..... denir. Bu hatalar …… ve ……. etkiler.

Bu örnekte ortak madde yazım ilkelerinden *"Madde doğru cevaba ipucu vermeyecek şekilde yapılandırılmalıdır."* ihlal edilmiştir. Çünkü maddede bulunan "tesadüfen karışan hata" ifadesi ilk boşluk için doğru cevap olan "tesadüfi hata"ya doğrudan ipucu vermektedir. Diğer yandan *"Madde açık ve net olarak ifade edilmelidir."* ilkesi de ihlal edilmiştir. Çünkü ikinci ve üçüncü boşluk testi alan her birey tarafından aynı şekilde anlaşılmayabilecek şekilde yapılandırılmıştır. Beklenen yanıtlar "geçerlik" ve "güvenirlik" iken bu boşluklara "öğrenci", "gözlenen puan" ve "değerlendirme" gibi birçok farklı doğru yanıtı barındırabilir. Kısa cevaplı madde yazım ilkeleri incelendiğinde ise *"Boşluk doldurma biçimindeki maddelerde en fazla iki boşluk bırakılmalıdır."* ve *"Boşluk doldurma biçimindeki maddelerde hem madde içinde hem de testin tamamında boşlukların uzunluğu eşit olmalıdır."* ilkeleri ihlal edilmiştir. Örnek maddenin bu kusurları giderilmiş hâli aşağıda sunulmuştur.

✓ Kaynağı, yönü ve miktarı tam olarak belirlenemeyen hata türüne ..... denir. Bu tür hataların artması ölçme sonuçlarının psikometrik özelliklerinden ..... doğrudan düşürür.

3.2.2. Açık Uçlu Maddelere İlişkin Madde Yazım İlkeleri

Hem yanıtı sınırlandırılmış hem de yanıtı sınırlandırılmamış açık uçlu madde yazımında ortak madde yazım ilkelerinin yanı sıra bu madde türüne özgü madde yazım ilkeleri de dikkate alınmalıdır. Tablo 6'da yanıtı sınırlandırılmış ve yanıtı sınırlandırılmamış açık uçlu maddelere ilişkin madde yazım ilkeleri verilmiştir.

Tablo 6

Açık Uçlu Madde Türüne Özgü Madde Yazım İlkeleri

Açık Uçlu Maddelere Özgü Madde Yazım İlkeleri	▪ Ezbere dayalı bilgileri yoklayacak şekilde hazırlanmamalıdır (yanıtı sınırlandırılmamış maddelerde dikkat edilmelidir). ▪ Beklenen cevap uzunluğu net bir şekilde belirtilmelidir. ▪ Beklenen cevabın içeriğini net bir şekilde belirtmek için çerçeve, maddede verilmelidir. ▪ Beklenen cevap sayısalsa maddede istenen kesinlik derecesi ve birimler belirtilmelidir. ▪ Puanlama kriterleri yönergede verilmelidir.

Açık uçlu maddeler bireylerin ezbere dayalı bilgilerini ölçmekten ziyade bu bireylerin bilgilerini anlama, yorumlama, ilişkilendirmeyi ve kendi ifadeleriyle açıklamayı gerektiren şekilde tasarlanmalıdır. Diğer bir deyişle özellikle yanıtı sınırlandırılmamış açık uçlu maddeler hatırlama düzeyinde hazırlanmamalı ve açık uçlu maddelerle daha üst düzey becerilerin ölçülmesi amaçlanmalıdır. Ancak kullanışlılık açısından yanıtı sınırlandırılmış maddelerde hatırlama düzeyinde de soru hazırlanabilir.

Beklenen cevabın içeriğinin çerçevesinin çizilmesinde "belirtiniz ve gerekçelendiriniz", "…… ilkesine/kuralına göre uygulayınız ve "uygun bir sonuca bağlayınız" gibi terimler kullanarak bireylerin cevabı nasıl yapılandıracağına yönelik bilgi sağlamalıdır (Russell & Airasian, 2012). Böyle terimlerin kullanımı bireylerin odaklanmalarına katkı sağlayacağı gibi puanlamanın da geçerliğini ve güvenirliğini artıracaktır. Açık uçlu bir madde ile bir konunun tartışılması isteniyorsa ortaya konulan görüşlerden / argümanlardan ziyade, bunları destekleyen kanıtların odak noktasında olduğu maddede ya da yönergede belirtilmelidir (Nitko & Brookhart, 2014).

Aşağıda yanıtı sınırlandırılmış açık uçlu madde türünde geliştirilmiş kusurlu bir madde örneği sunulmuştur.



Maddeleri test formuna yerleştirirken dikkat edilecek ölçme ilkelerini yazınız.

Bu örnekte ortak madde yazım ilkelerinden *"Madde belirli bir konu / durum ile sınırlı olmalıdır."* ilkesi ihlal edilmiştir. Çünkü maddelerin ne tür maddeler olduğu; bilişsel, duyuşsal ve psikomotor ve bilişsel özelliklerden hangisini ölçtüğü ayrıntıları bilinmemektedir. Örneğin, bilişsel özellikleri kısa cevaplı maddelerle ölçen bir testin form düzeni ile duyuşsal özellikleri likert tipi ölçek maddeleri ile ölçen bir testin form düzeninde dikkat edilecek ölçme ilkeleri birbirinden farklılaşır. Bunun yerine örnekteki konu bir test türü ile sınırlandırılmalıdır. Örnek maddede sınırlandırılması gereken diğer bir unsur ölçme ilkelerinin sayısıdır. Çünkü örnekte açık uçlu madde yazım ilkelerinden *"Beklenen cevap uzunluğu net bir şekilde belirtilmelidir."* ve *"Beklenen cevabın içeriğini net bir şekilde belirtmek için çerçeve, maddede verilmelidir."* ihlal edilmiştir. Ancak böyle sorularda yanıtı sınırlandırırken kapsamı da temsil edecek şekilde sınırlandırılması gerekmektedir. Bu sınırlandırma testin amacına göre oran belirlenerek yapılmalıdır. Tanıma amacıyla yapılan sınıf içi bir ölçme durumunda en az yarısı yoklanmalıdır. Bunun aksine biçimlendirme amacıyla yapılan sınıf içi bir ölçme durumunda ise en az üçte ikisi yoklanmalıdır. ÖSYM'nin yaptığı birçok yüksek riskli sınavda olduğu gibi kontenjan sayısının sınırlı başvuran sayısının ise çok olduğu durumlarda tam doğru cevabın dörtte üçü yoklanmalıdır. Örneğin, kimi branşlarda kısa cevaplı ya da yanıtı sınırlandırılmış açık uçlu maddelerden oluşan bir test şeklinde uygulanan Yan Dal Uzmanlık Sınavı'nda (YDUS) bir hastalığın tedavisinde kullanılacak sekiz etken madde varsa bunlardan en az altısının yoklanması önerilmektedir. Örnek maddenin bu kusurları giderilmiş hâli aşağıda sunulmuştur.



Çoktan seçmeli bir başarı testi geliştirmede maddeleri test formuna yerleştirilirken dikkat edilecek ölçme ilkelerinden dördünü yazınız.

Aşağıda yanıtı sınırlandırılmamış açık uçlu madde türünde geliştirilmiş kusurlu bir madde örneği sunulmuştur.



Bilgisayar ortamında bireye uyarlanmış test ile çok aşamalı testleri karşılaştırınız.

Bu örnekte açık uçlu madde yazım ilkelerinden *"Ezbere dayalı bilgileri yoklayacak şekilde hazırlanmamalıdır."* ilkesi ihlal edilmiştir. Çünkü bilgisayar ortamında bireye uyarlanmış testlere ve çok aşamalı testlere yönelik doğrudan bilgi düzeyinde ve yalnızca hatırlamayı gerektiren bir maddedir. Diğer bir deyişle testi alan bireylerin bilgiyi kullanarak yorum katmasını gerektirmemektedir. Bunun aksine yanıtı sınırlandırılmamış açık uçlu

maddelerden üst düzey becerileri yoklaması beklenir. Örnek maddede sınırlandırılması gereken bir unsur cevabın uzunluğu ve çerçevesidir. Bu örnekte açık uçlu madde yazım ilkelerinden *"Beklenen cevap uzunluğu net bir şekilde belirtilmelidir."* ve *"Beklenen cevabın içeriğini net bir şekilde belirtmek için çerçeve, maddede verilmelidir."* ihlal edilmiştir. Çünkü örnekte yer alan iki yaklaşıma ilişkin özelliklerin hangi açıdan karşılaştırılacağı belirsizdir. Diğer yandan testi alan birey böyle bir maddeyi bir paragrafla açıklayabileceği gibi sayfalarca da yanıt verebilir. Örnek maddenin bu kusurları giderilmiş hâli aşağıda sunulmuştur.

✓ ÖSYM Başkanı olduğunuzu varsayınız. Kurum tarafından hâlihazırda uygulanan sınavlardan birini seçerek, bu sınavın hangi gerekçelerle ve hangi yöntem (bilgisayar ortamında bireye uyarlanmış test / çok aşamalı test) kullanılarak bireye uyarlanmış bir yapıya dönüştürülmesini uygun görürdünüz? Yanıtınızda seçtiğiniz sınavı psikometrik özellikler ve hesap verebilirlik ilkeleri açısından değerlendirerek önerdiğiniz dönüşümün üstün ve sınırlı yönlerini tartışınız. Cevabınızı iki sayfayı aşmayacak şekilde hazırlayınız.

4. SEÇME VE KATKI VE SEÇME GEREKTİREN MADDE TÜRLERİNİN ÜSTÜN VE SINIRLI YÖNLERİ

Seçme ve katkı gerektiren madde türleri; hazırlanması, uygulanması, puanlanması, objektiflik, şans başarısı, duyarlılık, kapsam geçerliği ve yanıtı şişirme durumları bağlamında birbiriyle karşılaştırılmış ve genel bir çerçevede Tablo 7'de verilmiştir. Bu tablodaki bilgiler her zaman geçerli olmayıp genel hatlarıyla seçme ve katkı gerektiren maddelerin birbirine göre üstün ve sınırlı yönleri ortaya koymaktadır.

Tablo 7

Seçme ve Katkı Gerektiren Madde Türlerinin Belirli Özellikler Yönünden Karşılaştırması

Özellik	Seçme Gerektiren Maddeler	Katkı Gerektiren Maddeler
Hazırlanması	Daha zor	Daha kolay
Uygulanması	Daha kolay	Daha zor
Puanlanması	Daha kolay	Daha zor
Objektiflik	Var	Yok
Şans başarısı	Daha yüksek	Daha düşük
Duyarlılık	Daha yüksek	Daha düşük
Kapsama dayalı kanıtlar	Daha yüksek	Daha düşük
Yanıtı şişirme	Yok	Var

Seçme gerektiren maddeler, katkı gerektiren maddelere göre hazırlanması açısından genel itibariyle kullanışlılığı düşük madde türleridir. Özellikle çoktan seçmeli maddelerin hazırlanması seçeneklerin geliştirilmesinin ayrı bir uzmanlık gerektirmesinden dolayı daha zordur. Benzer şekilde eşleştirmeli maddelerin hazırlanması kısa cevaplı maddelerin hazırlanmasından daha zordur. Ancak belirtildiği gibi bu durum her zaman geçerli değildir. Örneğin, açık uçlu madde türünün hazırlanması doğru-yanlış türünde bir maddenin hazırlanmasından daha zordur. Çünkü açık uçlu madde hazırlanması eş zamanlı olarak puanlama anahtarının (dereceleme ölçeği ve dereceli puanlama anahtarı gibi) da hazırlanmasını gerektirmektedir.

Seçme gerektiren maddelerin uygulanması katkı gerektiren maddelere göre daha kolaydır. Çünkü seçme gerektiren maddelerde testi alan bireyler cevaplarını yapılandırmak yerine sadece verilen seçenekler arasında seçim yapmaktadırlar. Bu özelliği nedeniyle seçme gerektiren bir maddenin uygulanması ile katkı gerektiren maddelerden yapılandırılmamış açık uçlu bir maddenin uygulanması birim zaman anlamında daha kısa süre alır. Uygulamadaki bu kolaylık aynı sürede sorulabilecek soru sayısını etkileyebileceğinden seçme gerektiren maddelerden elde edilen ölçümler ile daha fazla kapsama dayalı kanıtların elde edilmesi sağlanır ve geçerlik olumlu yönde etkilenir. Ancak unutulmamalıdır ki bu durum maddelerin farklı öğrenme çıktılarını / konuları ölçtüğünü varsaydığımız durum için geçerlidir. Çünkü aynı öğrenme çıktısına ilişkin çok sayıda sorulduğu bir durumda sadece soru sayısının artmasının kapsam geçerliğini artırdığı söylenemez. Diğer yandan seçme gerektiren madde türlerini kullanmak, daha fazla madde uygulayabilme olanağı sunduğundan bu maddelerden oluşan testlerin daha duyarlı ölçümler sağladığı söylenebilir.

Seçme gerektiren maddeler, katkı gerektiren madde türlerine göre puanlanması açısından daha objektif ve kolaydır. Bu nedenle geniş ölçekli testlerde seçme gerektiren maddelerden özellikle çoktan seçmeli maddelere sıklıkla başvurulur. Seçme gerektiren maddelerin optik okuyucularla çok hızlı puanlanabilmesi maliyeti düşürüp ulaşılabilirliği artırdığından açısından son derece kullanışlı olduğu söylenebilir. Bunun aksine katkı gerektiren maddeleri puanlaması oldukça güçtür ve subjektiftir. Bunun önüne geçmek için puanlama anahtarları geliştirilerek bunlardan yararlanılsa da puanlayıcı güvenirliği seçme gerektiren maddeler kadar yüksek değildir. Ayrıca açık uçlu maddeleri rubrikle dahi puanlamanın uzmanlık gerektirmesi kullanışlılığı düşürmektedir. Ancak dijital dönüşümlerle birlikte otomatik puanlama sistemleri geliştirilebilmekte olup bu sistemler bu tür maddelerin objektif puanlamasına katkı sağlayabilmektedir. Detaylı bilgiler bu kitabın 10. bölümünde sunulmuştur. Her ne kadar seçme gerektiren maddelerde, daha az tesadüfi hatanın karışmasını sağlayacak objektiflik söz konusu olsa da bu ölçümlere tesadüfi hata karıştıracak farklı bir hata kaynağının olmadığı anlamına gelmemektedir. Seçme gerektiren maddelerin katkı gerektiren maddelere göre önemli bir sınırlılığı şans başarısını farklı bir hata kaynağı

olarak barındırmasıdır. Katkı gerektiren maddelerde maddeyi tesadüfen doğru cevaplama olasılığı seçme gerektiren maddelere göre oldukça düşüktür ve seçme gerektiren maddelerle yapılan ölçümlere bu yönde tesadüfi hatalar daha fazla karışabilmektedir.

Katkı gerektiren maddelerden özellikle yanıtı sınırlandırılmamış açık uçlu madde türünde testi alan bireylerin yanıtı şişirerek ölçülmek istenen özellik dışında başka değişkenleri de yanıtlarına ekledikleri durumlar söz konusudur. Puanlama anahtarının olmadığı durumlarda daha uzun yanıt veren bireyin daha kısa yanıt veren bireye göre haksız yere daha fazla puan almasına sebep olabilir. Bu yüzden ölçümlere hata karışarak ölçümlerin güvenirlik ve geçerliği zedelenebilir. Bu doğrultuda puanlama anahtarının önemi büyüktür. Bu konuya ilişkin detaylı bilgilere kitabın 7. bölümünde yer verilmiştir.

4.1. Seçme Gerektiren Madde Türlerinin Üstün ve Sınırlı Yönleri

Seçme gerektiren madde türlerinin üstün ve sınırlı yönlerini belirleyen özellikleri karşılaştırmalı olarak Tablo 8'de verilmiştir.

Tablo 8

Seçme Gerektiren Madde Türlerinin Belirli Özellikler Yönünden Karşılaştırması

Özellik	Doğru-Yanlış	Eşleştirmeli	Çoktan Seçmeli
Madde yapısı	Tek önerme	Öncül ve seçenek listesi	Madde kökü ve seçenekler
Seçenek sayısı	2	Genellikle 5–8	Genellikle 3–5
Şans başarısı	%50	4 öncül, 6 seçenek için %6,67	5 seçenek için %20
Hazırlama	Kolay	Orta	Zor
Puanlama	Kolay	Kolay	Kolay

Seçme gerektiren maddeler, özellikle objektif puanlanabilmelerinin de etkisiyle hem klasik test kuramı hem de madde tepki kuramı bağlamında analiz edilebilir yapıları sayesinde geniş ölçekli sınavlarda vazgeçilmezdir. Seçme gerektiren maddeler arasında şans başarısı en yüksek olan doğru-yanlış madde türüdür ve yarı yarıya şans başarısı bulunmaktadır. Diğer bir deyişle "Doğru" ve "Yanlış" sayısının cevap anahtarında homojen dağıldığı bir doğru-yanlış maddelerden oluşan bir testte hepsini "Doğru" ya da hepsini "Yanlış" işaretleyen bir birey konuyu bilmediği hâlde 100 puan üzerinden 50 puan alabilir. Sıklıkla kullanılan çoktan seçmeli türündeki maddelerde ise dört seçenek için %25, beş seçenek için %20 şans başarısı mevcuttur. Yine konuyu bilmeyen bir birey doğru seçenek sayısının homojen dağıldığı (20 maddelik test için her bir seçeneğin

dörder kez doğru cevap olması durumu) bir testteki tüm maddeler için aynı seçeneği işaretlediğinde 100 puan üzerinden 20 puan alır. Bu iki madde türünde şans başarısını azaltmak için bazı önlemler alınabilir. Örneğin, doğru-yanlış türündeki maddeler için cevap yanlışsa doğrusunu yazmaları istenebilir, çoktan seçmeli madde ise işaretlediği seçeneği seçme nedeni sorulabilir. Ancak bu uygulamaların kullanışlılığı düşüktür çünkü puanlamayı zorlaştırmaktadır. Dolayısıyla bu yöntemler sınıf içi ölçme ve değerlendirme süreçlerinde uygulanabilir olsa da geniş ölçekli sınavlarda uygulanması oldukça güçtür. Geniş ölçekli sınavlarda ise düzeltme formülünün kullanılması önerilebilir. Düzeltme formülüne göre seçenek sayısının bir eksiği kadar yanlış cevap, bir doğru cevabı götürmektedir. Örneğin, doğru-yanlış türündeki maddelerde "Doğru" ve "Yanlış" olmak üzere iki seçenek bulunduğundan bir yanlış (2 seçenek-1) bir doğruyu götürmektedir. Beş seçenekli çoktan seçmeli bir maddede ise dört yanlış (5 seçenek-1) bir doğruyu götürmektedir. Bu doğrultuda ham puan hesaplanırken doğru cevap sayısından yanlış cevap sayısının seçenek sayısının bir eksiğine oranı çıkarılmaktadır. Yani yukarıda verilen örneklerde düzeltme formülü kullanıldığında doğru-yanlış testi için birey 50 puan yerine, çoktan seçmeli test için ise 20 puan yerine 0 puan almış olur. Buradan da anlaşılacağı üzere seçenek sayısı azaldıkça düzeltme formülünün etkisi büyük olmaktadır. Ancak sertifika sınavları gibi mutlak değerlendirme gerektiren ve seviye / yeterlilik belirlemeye yönelik geliştirilen testlerde [örneğin; Yabancı Dil Bilgisi Seviye Tespit Sınavı (YDS), İş Sağlığı ve Güvenliği Genel Müdürlüğü İş Yeri Hekimliği ve İş Güvenliği Uzmanlığı Sınavı (İSG), Diyanet İşleri Başkanlığı Mesleki Bilgiler Seviye Tespit Sınavı (DİB-MBSTS)] düzeltme formülü kullanılmamalıdır. Seçme ve yerleştirme amacıyla yapılan sınavlarda [örneğin; Yükseköğretim Kurumları Sınavı (YKS), Kamu Personel Seçme Sınavı (KPSS)] ise düzeltme formülünün kullanılması önerilir. Şans başarısı eşleştirmeli maddeler için değerlendirildiğinde, her bir seçeneğin birden fazla kez kullanıldığı durumda eşleştirmeli maddelerin şans başarısının en düşük olduğu seçme gerektiren madde türü olduğu söylenebilir. Bunun aksine seçeneklerin tekrarlı kullanılamadığı durumda ilk öncülün karşılığına gelen seçeneğe ilişkin şans başarısı düşük olsa da sonraki öncüllere geçildiğinde seçenekler azaldıkça şans başarısı artmaktadır.

Seçme gerektiren maddelerde hazırlama kolaylığı seçenek sayısıyla ilişkilidir. Doğru-yanlış türü maddelerde tek bir önerme yazıldığından ve "Doğru" ve "Yanlış" olmak üzere iki seçenek bulunduğundan bu tür maddeleri hazırlamak diğer seçme gerektiren madde türlerine göre daha kolaydır. Hazırlaması en zor seçme gerektiren madde türü ise çoktan seçmeli maddelerdir. Özellikle çeldirici hazırlamak uzmanlık gerektirmektedir ve hazırlaması zaman alıcıdır. Öncül ve seçenek listesinden oluşan eşleştirmeli maddelerde ise yazılan seçeneklerin tüm öncüller için ortak olarak kullanılıyor olması bu madde türünün hazırlanmasını çoktan seçmeli maddeye göre daha kolay kılmaktadır. Tüm seçme gerektiren madde türleri çoğunlukla ikili (1-0) olarak objektif puanlanabildiğinden puanlama açısından kullanışlıdırlar.

4.2. Katkı Gerektiren Madde Türlerinin Üstün ve Sınırlı Yönleri

Katkı gerektiren madde türlerinin üstün ve sınırlı yönlerini belirleyen özellikleri karşılaştırmalı olarak Tablo 9'da verilmiştir.

Tablo 9

Katkı Gerektiren Madde Türlerinin Belirli Özellikler Yönünden Karşılaştırması

Özellik	Kısa Cevaplı	Yanıtı Sınırlandırılmış	Yanıtı Sınırlandırılmamış
Cevap uzunluğu	Kısa	Orta	Uzun
Cevap sınırı	Yüksek	Orta	Düşük
Katkı düzeyi	Düşük	Orta	Yüksek
Hazırlama	Kolay	Orta	Zor
Cevap çeşitliliği	Düşük	Orta	Yüksek
Puanlama kolaylığı	Yüksek	Orta	Düşük
Puanlama güvenirliği	Yüksek	Orta	Düşük

Katkı gerektiren maddeler arasında olası doğru cevap çeşitliliği en yüksek olan madde türü yanıtı sınırlandırılmamış açık uçlu madde olup aynı zamanda bireyler tarafından en yüksek katkıyı gerektirmektedir. Bunun aksine kısa cevaplı madde türünde doğru cevap çeşitliliği oldukça kısır olmakla birlikte bu madde türü bireylerin en düşük düzeyde katkı verdiği madde türüdür. Puanlamada kullanılan puanlama anahtarlarının da hazırlama sürecine dâhil olduğu düşünüldüğünde hazırlaması en zor madde türü yanıtı sınırlandırılmamış açık uçlu maddeyken hazırlaması en kolay madde türü kısa cevaplı maddelerdir.

Katkı gerektiren maddeler için otomatik puanlama sistemlerinin gelişmesi ve bu sayede objektifliğin artmasıyla bu madde türleri, geniş ölçekli sınavlarda daha yaygın biçimde kullanılmaya başlanmıştır. Katkı gerektiren maddeler arasında kısa cevaplı madde türü puanlama kolaylığı en yüksek olan madde türüdür. Çünkü açık uçlu sorulara göre cevabının sınırları çok daha net bir biçimde bellidir. Bu durum, kısa cevaplı maddelerde objektif puanlamayı daha mümkün kılar ve dolayısıyla puanlayıcılar arası güvenirliği açık uçlu maddelere göre daha yüksektir. Benzer olarak açık uçlu maddeler arasında yanıtı sınırlandırılmış açık uçlu madde yanıtı sınırlandırılmamış açık uçlu maddeye göre cevabı daha sınırlı ve nettir. Dolayısıyla daha objektif puan alır ve puanlama güvenilirliği daha yüksektir.

5. SONUÇ

Bu bölümde bilişsel alana özgü madde türleri, bireyden beklenen yanıtın verilme biçimi esas alınarak seçme gerektiren ve katkı gerektiren maddeler şeklinde iki grupta ele alınmıştır. Bölüm boyunca seçme ve katkı gerektiren madde türlerinin özellikleri ile üstün ve sınırlı yönleri açıklanmış; uygun madde türünün seçimi ve seçilen madde türüne uygun maddelerin yazımında dikkat edilmesi gereken temel ilkeler üzerinde durulmuştur. Bunun yanı sıra, bu bölümde hem geniş ölçekli testlerde hem de sınıf içi ölçme ve değerlendirme uygulamalarında madde yazımına rehberlik edecek uygulamaya yönelik öneriler sunulmuştur.

Seçme gerektiren maddelerin katkı gerektiren maddelere göre en büyük sınırlılığı şans başarısına açık olmalarıdır. Bunun üstesinden geniş ölçekli testlerde düzeltme formülü kullanarak gelinebildiği gibi özellikle sınıf içi ölçmelerde kullanılabilecek başka yöntemler de alanyazında mevcuttur. Seçme gerektiren maddeler içerisinde en yaygın olarak kullanılan madde türü çoktan seçmeli maddelerdir. Bu madde türünün de farklı alt türleri bulunmaktadır. Bu madde türleri arasından daha yenilikçi olan çoklu seçim gerektiren çoktan seçmeli maddeler, şans başarısını düşüreceği için geleneksel çoktan seçmeli madde yerine önerilmektedir. Doğru-yanlış türü maddeler ise şans başarısına oldukça açık olup, kullanışlılığı düşürse de yanlış yanıtın doğrusu istenerek bu etki azaltılabilir. Genel olarak şans başarısı en düşük seçme gerektiren madde türü olan eşleştirmeli maddelerde, yönergede açıkça belirtilmesi koşuluyla bir seçeneğin birden fazla önerme için kullanılmasına izin verilmesi, şans başarısını daha da azaltacaktır.

Katkı gerektiren maddelerin seçme gerektiren maddelere göre en önemli üstünlüğü, üst düzey düşünme becerilerini değerlendirmeye olanak sağlamasıdır. Genel olarak, bireyden yanıtını daha fazla yapılandırmasının beklendiği madde türlerinde, ölçülen bilişsel süreçlerin de daha üst düzeyde olması beklenmektedir. Diğer bir deyişle katkı gerektiren maddelerde, bireyden beklenen yanıt uzunluğu, ölçülen beceri seviyesini belirlemede kritik bir rol oynar. Yanıtı sınırlandırılmamış açık uçlu madde türünde olduğu gibi daha uzun ve ayrıntılı yanıt gerektiren maddeler adayın analiz etme, değerlendirme ve yaratma becerilerini ölçmeye olanak tanır. Buna karşın, adayın katkısını daha sınırlı tutmayı gerektiren yanıtı sınırlandırılmış açık uçlu madde türü uygulama ve analiz etme düzeyinde; kısa cevaplı maddeler ise genellikle hatırlama ve anlama düzeyindeki becerilerin ölçülmesine hizmet eder. Katkı gerektiren madde türleri arasında, özellikle yanıtı sınırlandırılmamış açık uçlu maddelerin puanlanmasında subjektiflik önemli ölçüde artar ve puanlama hataları da buna bağlı olarak sıklaşır. Kısa cevaplı veya yanıtı sınırlandırılmış açık uçlu maddelere kıyasla, bu tür maddeler değerlendiricinin yorumu ve beklentisine daha fazla bağımlıdır. Bu subjektifliğin azaltılması için, puanlamanın ayrıntılı ve net ölçütlerle geliştirilen rubrikler ile yapılması gereklidir.

Yanıtı sınırlandırılmamış açık uçlu maddeler, adaydan uzun ve ayrıntılı yanıtlar vermesini gerektirdiğinden analiz etme, değerlendirme ve yaratma gibi üst düzey düşünme becerilerinin ölçülmesine olanak tanır. Buna karşılık, adayın yanıtını belirli sınırlar içinde oluşturmasını gerektiren yanıtı sınırlandırılmış açık uçlu maddeler daha çok uygulama ve analiz düzeyindeki becerilerin; kısa cevaplı maddeler ise genellikle hatırlama ve anlama düzeyindeki becerilerin ölçülmesinde kullanılır. Katkı gerektiren madde türlerinin en önemli sınırlılıklarından biri puanlamaya ilişkin güçlüklerdir. Özellikle yanıtı sınırlandırılmamış açık uçlu maddelerde, yanıtların farklı biçimlerde ifade edilebilmesi değerlendirici yorumunun puanlamaya daha fazla yansımasına neden olmakta, bu durum ise puanlamada subjektifliği ve hata olasılığını artırmaktadır. Bu nedenle, puanlama sürecinde ayrıntılı, açık ve önceden tanımlanmış ölçütlere dayalı dereceli puanlama anahtarlarının (rubriklerin) kullanılması, puanlayıcılar arasındaki tutarlılığın artırılması ve puanlama hatalarının azaltılması açısından büyük önem taşımaktadır.

Sonuç olarak, madde türlerinin üstün ve sınırlı yönlerinin göz önünde bulundurulması, sınırlı yönlerine yönelik alınabilecek önlemlerin ele alınması ve madde yazım ilkelerinin uygulanması, testlerin geçerlik ve güvenirliğinin artırılmasında kritik bir rol oynamaktadır.

KAYNAKÇA | REFERENCES

Al-lawama, M. & Kumwenda, B. (2023). Decreasing the options' number in multiple choice questions in the assessment of senior medical students and its effect on exam psychometrics and distractors' function. *BMC Medical Education*, *23*, Article 212. https://doi.org/10.1186/s12909-023-04206-3

Atalay-Kabasakal, K. & Gören, S. (2025). Using testlets in education: eTIMSS 2019 as an example. *Education and Science, 50*(Suplement Issue), 111-127. https://doi.org/10.15390/EB.2025.14104

Aydoğan, S. & Gündüz, T. (2024). Maddenin görsel içerip içermemesinin madde ve test istatistiklerine olan etkisinin incelenmesi. *Muğla Sıtkı Koçman Üniversitesi Eğitim Fakültesi Dergisi, 11*(2), 182-206. https://doi.org/10.21666/muefd.1219489

Barker, D. & Ebel, R. L. (1982). A comparison of difficulty and discrimination values of selected true-false item types. *Contemporary Educational Psychology*, *7*(1), 35-40. https://doi.org/10.1016/0361-476X(82)90005-4

Baykul, Y. (2000). *Eğitimde ve psikolojide ölçme: Klasik test teorisi uygulaması* (3. Baskı). ÖSYM.

Burton, S. J., Sudweeks, R. R., Merrill, P. F. & Wood, B. (1991). *How to prepare better multiple-choice test items: Guidelines for university faculty.* Bringham Young University Testing Services and the Department of Instructional Science. https://www.uwosh.edu/testing/wp-content/uploads/sites/164/2019/09/How-to-Prepare-Better-Multiple-Choice-Test-Items-Guidelines-for-University-Faculty.pdf

Butler, A. C. (2018). Multiple-choice testing in education: Are the best practices for assessment also good for learning? *Journal of Applied Research in Memory and Cognition, 7*(3), 323–331. https://doi.org/10.1016/j.jarmac.2018.07.002

Case, S. M. & Swanson, D. B. (2002). *Constructing written test questions for the basic and clinical sciences* (3. baskı). National Board of Medical Examiners. https://health.uconn.edu/faculty-development/wp-content/uploads/sites/69/2017/06/constructing_written_test_questions.pdf

Chappuis, J., Stiggins, R. J., Chappuis, S. & Arter, J. A. (2012). *Classroom assessment for student learning: Doing it right—Using it well* (2. baskı). Pearson.

Chiavaroli, N. (2017). Negatively-worded multiple choice questions: An avoidable threat to validity. *Practical Assessment, Research & Evaluation, 22*, Article 3. https://doi.org/10.7275/ca7y-mm27

Dudycha, A. L. & Carpenter, J. B. (1973). Effects of item format on item discrimination and difficulty. *Journal of Applied Psychology*, *58*(1), 116-121. https://doi.org/10.1037/h0035197

Ebel, R.L. & Frisbie, D.A. (1991). *Essentials of educational measurement* (5. baskı). Prentice-Hall, Englewood Cliffs.

Frisbie, D. A. (1992). *The measurement and evaluation of learning*. Waveland.

Haladyna, T. & Downing, S. (1989). Validity of a taxonomy of multiplechoice item-writing rules. *Applied Measurement in Education, 2*(l), 51-78.

Haladyna, T. M., Downing, S. M. & Rodriguez, M. C. (2002). A review of multiple-choice ıtem-writing guidelines for classroom assessment. *Applied Measurement in Education*, *15*(3), 309–333. https://doi.org/10.1207/S15324818AME1503_5

Harasym, P. H., Price, P. G., Brant, R., Violato, C. & Lorscheider, F. L. (1992). Evaluation of negation in stems of multiple-choice items. *Evaluation & The Health Professions, 15*(2), 198–220. https://doi.org/10.1177/016327879201500205

Millî Eğitim Bakanlığı. (2025, 4 Haziran). *Millî eğitim bakanlığından yeni bir izleme araştırması*. https://odsgm.meb.gov.tr/www/milli-egitim-bakanligindan-yeni-bir-izleme-arastirmasi/icerik/1441

Karatoprak-Erşen, R. & Gündüz, T. (2023). Seçme ve katkı gerektiren maddelerin yazımı ve düzenlenmesi için kontrol listeleri. *Buca Eğitim Fakültesi Dergisi, 58*, 2473-2493. https://doi.org/10.53444/deubefd.1279240

Kaya-Uyanık, G. & Gelbal, S. (2018). Madde tepki modellemesinde genellenebilirlik ile iki yüzeyli desenlerin incelenmesi. *Journal of Measurement and Evaluation in Education and Psychology, 9*(1), 17-32. https://doi.org/10.21031/epod.349718

Kocaoğlu, S. & Şahin, M. G. (2024). Investigating the effect of testlets consisting of open-ended and multiple-choice items on reliability via generalizability theory. *Journal of Measurement and Evaluation in Education and Psychology, 15*(1), 65-78. https://doi.org/10.21031/epod.1429423,

Nitko, A. J. & Brookhart, S. M. (2014). *Educational assessment of students* (7. baskı). Pearson.

Popham, W. J. (2014). *Classroom assessment: What teachers to know* (7. baskı). Pearson.

Rudolph, M. J., Daugherty, K. K., Ray, M. E., Shuford, V. P., Lebovitz, L. & DiVall, M. V. (2019). Best practices related to examination item construction and post-hoc review. *American journal of pharmaceutical education*, *83*(7), Article 7204. https://doi.org/10.5688/ajpe7204

Rush, B. R., Rankin, D. C. & White, B. J. (2016). The impact of item-writing flaws and item complexity on examination item difficulty and discrimination value. *BMC Medical Education*, *16,* Article 250. https://doi.org/10.1186/s12909-016-0773-3

Russell, M. K. & Airasian, P. W. (2012). *Classroom assessment: Concepts and applications* (7. baskı). McGraw-Hill.

Sireci, S. G., Wiley, A. & Keller, L. A. (1998). An empirical evaluation of selected multiple-choice item writing guidelines. *Applied Measurement in Education, 19*(3), 185–210. https://files.eric.ed.gov/fulltext/ED428122.pdf

Taşdelen-Teker, G. & Dogan, N. (2015). The effects of testlets on reliability and differential item functioning. *Educational Sciences: Theory and Practice*, *15*(4), 969-980. https://doi.org/10.12738/estp.2015.4.2577

Thanyapa, I. & Currie, M. (2014). The number of options in multiple choice items in language tests: Does it make any difference? Evidence from Thailand. *Language Testing in Asia, 4,* Article 8. https://doi.org/10.1186/s40468-014-0008-7

Thorndike, R. M. & Thorndike-Christ, T. (2010). *Measurement and evaluation in psychology and education* (8. baskı). Pearson.

Turgut, F. & Baykul, Y. (2013). *Eğitimde ölçme ve değerlendirme* (5. Baskı). Pegem

Waugh, C. K. & Gronlund, N. E. (2013). *Assessment of student achievement* (10. baskı). Pearson.

Yılmaz-Koğar, E. (2022). Sağlık alanından örneklerle çoktan seçmeli maddelerin üst düzey düşünme becerilerine yönelik tasarlanması için öneriler. *Journal of Nursology*, *25*(2), 117-125.

YAZARLAR HAKKINDA | ABOUT AUTHORS

ARŞ. GÖR. DR. TUBA GÜNDÜZ

Muğla Sıtkı Koçman Üniversitesi | tubagunduz[at]mu.edu.tr

ORCID: 0000-0002-0921-9290

Dr. Tuba Gündüz, Muğla Sıtkı Koçman Üniversitesi Eğitim Fakültesi Eğitim Bilimleri Bölümü Eğitimde Ölçme ve Değerlendirme Anabilim Dalında araştırma görevlisi olup 2021 yılından bu yana çalışmalarını Ölçme, Seçme ve Yerleştirme Merkezi Başkanlığı (ÖSYM) bünyesinde sürdürmektedir. Lisans eğitimini Hacettepe Üniversitesinde, yüksek lisans ve doktora eğitimini ise Gazi Üniversitesinde Eğitimde Ölçme ve Değerlendirme alanında tamamlamıştır. Dr. Gündüz'ün ulusal ve uluslararası hakemli dergilerde madde tepki kuramı, bilişsel tanı modelleri, test eşitleme, değişen madde fonksiyonu, geniş ölçekli değerlendirmeler ve madde yazım ilkeleri üzerine yayımlanmış çalışmaları bulunmaktadır. Bunun yanında ÖSYM ve TÜBİTAK tarafından desteklenen projelerde araştırmacı ve yürütücü olarak görev almıştır. Son yıllarda özellikle üretken yapay zekânın madde geliştirme süreçlerinde kullanımı, dijital ölçme ve değerlendirme sistemleri ve yenilikçi psikometrik yöntemler üzerine çalışmalar yürütmektedir. Ayrıca lisans düzeyinde Eğitimde Ölçme ve Değerlendirme ile Bilim ve Araştırma Etiği derslerini yürütmüştür.

RES. ASST. DR. TUBA GÜNDÜZ

Muğla Sıtkı Koçman University | tubagunduz[at]mu.edu.tr

ORCID: 0000-0002-0921-9290

Dr. Tuba Gündüz is a research assistant in the Department of Educational Measurement and Evaluation, Faculty of Education, Muğla Sıtkı Koçman University. She has been conducting her academic and professional work at the Center for Assessment, Selection and Placement (ÖSYM), since 2021. She received her bachelor's degree from Hacettepe University and completed both her master's and doctoral degrees in Educational Measurement and Evaluation at Gazi University. Dr. Gündüz has published studies in national and international peer-reviewed journals on item response theory, cognitive diagnostic models, test equating, differential item functioning, large-scale assessments, and item writing principles. She has also served as a researcher and principal investigator in projects funded by ÖSYM and TÜBİTAK. In recent years, her research has focused on the use of generative artificial intelligence in item writing, digital assessment systems, and innovative psychometric methods. She has also taught undergraduate courses in Educational Measurement and Evaluation and Scientific Research Ethics.

DOÇ. DR. YILDIZ YILDIRIM

Aydın Adnan Menderes Üniversitesi | yildizyldrm[at]gmail.com

ORCID: 0000-0001-8434-5062

Doç. Dr. Yıldız Yıldırım, ilk ve ortaöğrenimini Ankara'da tamamlamıştır. Lisans eğitimini ise 2012 yılında Gazi Üniversitesi, Gazi Eğitim Fakültesi, Ortaöğretim Fen ve Matematik Alanları Bölümünde tamamlamıştır. 2015 yılında Gazi Üniversitesi, Eğitim Bilimleri Enstitüsü, Eğitimde Ölçme ve Değerlendirme Bilim Dalında yüksek lisans derecesini; 2021 yılında ise aynı bilim dalında doktora derecesini almıştır. 2012-2013 yılları ve 2018-2023 yılları arasında Aydın Adnan Menderes Üniversitesinde, 2013-2018 yılları arasında ise Gazi Üniversitesinde Eğitimde Ölçme ve Değerlendirme Anabilim Dalında araştırma görevlisi olarak görev yapmıştır. 2023 yılından beri Aydın Adnan Menderes Üniversitesinde öğretim üyesi kadrosunda bulunan Doç. Dr. Yıldız Yıldırım aynı yıl itibariyle Ölçme, Seçme ve Yerleştirme Merkezi Başkanlığında (ÖSYM) Ölçme ve Değerlendirme Uzmanı olarak görev yapmaktadır. Yazarın meta-analiz, yapısal eşitlik modelleri, madde tepki kuramı, test eşitleme ve test geliştirme ile ilgili çalışmaları bulunmaktadır.

ASSOC. PROF. DR. YILDIZ YILDIRIM

Aydın Adnan Menderes University | yildizyldrm[at]gmail.com

ORCID: 0000-0001-8434-5062

Assoc. Prof. Dr. Yıldız Yıldırım completed her elementary and secondary education in Ankara. She earned her bachelor's degree in 2012 from Gazi University, Gazi Faculty of Education, Department of Secondary Education in Science and Mathematics. In 2015, she received her master's degree from Gazi University, Graduate School of Educational Sciences, Department of Assessment and Evaluation in Education; and in 2021, she earned her Ph.D. in the same field. She served as a research assistant in the Department of Measurement and Evaluation in Education at Aydın Adnan Menderes University from 2012 to 2013 and from 2018 to 2023, and in the Assessment and Evaluation in Education at Gazi University from 2013 to 2018. Assoc. Prof. Dr. Yıldız Yıldırım, who has been a faculty member at Aydın Adnan Menderes University since 2023, has also served as an Assessment and Evaluation Specialist at the Center for Assessment, Selection, and Placement Center (ÖSYM) since that same year. The author's research focuses on meta-analysis, structural equation modeling, item response theory, test equating, and test development.

7. BÖLÜM

PERFORMANS DEĞERLENDİRME ARAÇ VE YÖNTEMLERİ

PERFORMANCE ASSESSMENT TOOLS AND METHODS

Prof. Dr. Mehtap ÇAKAN

PERFORMANS, PERFORMANS DEĞERLENDİRME ARAÇ VE YÖNTEMLERİ

Prof. Dr. Mehtap ÇAKAN
Gazi Üniversitesi

Öz

Bu bölümün amacı performans kavramını tanımlayarak öğrenci performansının değerlendirilmesinde yaygın olarak kullanılan araç ve yöntemleri tanıtmaktır. Performans en genel anlamda bireyin bilgi, beceri ve yeteneklerini gerçek ya da gerçeğe yakın koşullarda ortaya koyduğu, gözlemlenebilir davranış ve eylemler bütünü olarak tanımlanabilir. Performansın ölçülmesinde performans görevi, proje ve portfolyo gibi araçlardan yararlanılmaktadır. Performans görevi, bireyin bilgi ve becerilerini gerçek yaşam koşulları ya da buna yakın koşullarda sergilemesine olanak sağlayan bir görevi ifade eder ve bu görevle ürün ve süreç birlikte değerlendirilebilir. Proje ise performans görevinden farklı olarak daha kapsamlı, karmaşık ve daha üst düzey becerileri kullanmayı gerektiren yapıdadır ve disiplinler arası çalışmaya daha uygundur. Ayrıca projede o alana ilişkin konuyu ve süreçlerini bireyin kendisinin belirlemesi, tasarlaması beklenir. Diğer yandan Portfolyo ise; bireyin öğrenme sürecindeki çabasını, gelişimini ve akademik başarısını belgelemek amacıyla, zaman içerisinde oluşturduğu çalışmaların sistematik ve amaçlı bir biçimde derlenmesi olarak tanımlanmaktadır. Bu araçların puanlanmasında ise rubrikler, derecelendirme ölçekleri ve kontrol listeleri kullanılmaktadır. Rubrik; bir performansın, ürünün ya da görevin önceden belirlenmiş ölçütler doğrultusunda ve her ölçüt için tanımlanmış başarı düzeylerine göre sistematik, nesnel ve tutarlı biçimde değerlendirilmesini sağlayan bir puanlama aracıdır. Derecelendirme ölçeği; bir davranışın, becerinin veya performansın belirli düzeylerde ne ölçüde gerçekleştirildiğini değerlendirirken, kontrol listeleri belirlenen ölçüt veya davranışların gerçekleştirilip gerçekleştirilmediğini var/yok ya da evet/hayır biçiminde belirlemeye yarayan puanlama araçlarıdır. Bu bölümde performans değerlendirme araçları ile bunların puanlanmasında kullanılan puanlama araçları örnekleriyle birlikte detaylı şekilde ele alınmıştır.

Anahtar Kelimeler

Performans Değerlendirme, Performansa Dayalı Durum Belirleme, Performans Görevi, Proje, Portfolyo, Rubrik, Derecelendirme Ölçeği, Kriter Performans Listesi, Kontrol Listesi

PERFORMANCE ASSESSMENT TOOLS AND METHODS

Prof. Dr. Mehtap ÇAKAN
Gazi University

Abstract

The purpose of this chapter is to define the concept of performance and introduce the tools and methods used in performance assessment. In the broadest sense, performance can be defined as the set of observable behaviors and actions through which an individual demonstrates their knowledge, skills, and abilities under real or near-real conditions. Performance is measured using tools such as performance tasks, projects, and portfolios. A performance task is an assessment task that allows an individual to demonstrate their knowledge and skills in a practical manner, and both the product and the process can be assessed through this task. A project, on the other hand, differs from a performance task in that it has a more comprehensive and complex structure and may require interdisciplinary work. Additionally, in a project, the individual is expected to determine the topic themselves. A portfolio is defined as the systematic and purposeful compilation of work produced over time to document an individual's efforts, development, and academic achievement during the learning process. Rubrics, rating scales / criterion performance list, and checklists are used to score these tools. A rubric is a scoring tool that enables the systematic, objective, and consistent assessment of a performance, product, or task based on predetermined criteria and defined levels of achievement for each criterion. A rating scales / criterion performance list, is a scoring tool used to assess the extent to which a behavior, skill, or performance is demonstrated at specific levels, while a checklist is a scoring tool used to determine whether specified criteria or behaviors have been met or not, using a present / absent or yes / no format. This section provides a detailed discussion of all these tools and methods, along with examples of the scoring tools used by these instruments.

Keywords

Performance Assessment, Performance-based Assessment, Performance Task, Project, Portfolio, Rubric, Rating Scale, Criterion Performance List, Checklist

1. GİRİŞ

Birey performansının ölçülmesi, yalnızca günümüzün ilgi, odak ya da ihtiyaç noktası olmayıp çağlar boyunca farklı gerekçelerle, farklı alanlarda ve ortamlarda gerçekleştirmek istenen bir çabayı işaret etmektedir. Bu bağlamda, bireyin doğumundan başlayarak, onun okul yılları boyunca, daha sonra meslek hayatı boyunca belli alanlardaki performansı çeşitli ihtiyaçlar çerçevesinde gözlemlenip değerlendirilmek istenmiştir. Okullar için ele aldığımızda, gerek öğrencinin o dersteki yeterliğine ilişkin kararlar vermek, gerek uygulanan programların etkililiğine ilişkin kararlar vermek, gerekse öğretmenin yeterliğine ilişkin çeşitli kararlara ulaşmak amacıyla öğrenci performansı değerlendirilmektedir. Beklendiği üzere, öğrenci performansının değerlendirilmek istenmesindeki temel amaç, öğrencinin o ders ya da programda istenen yetkinlik düzeyine ne derece ulaştığına ilişkin kanıtlara ulaşmak ve bu kanıtlara dayalı olarak geçerliği yüksek kararlar vermektir.

Kitabın bu bölümünde öncelikle performans tanımlanmıştır. Ardından performans değerlendirme bağlamında kullanılan puanlama araçlarına yer verilmiştir. Ek olarak, yine performans değerlendirme bağlamında kullanılan performans görevi, portfolyo ve proje gibi ölçme araçları sunulmuştur.

2. PERFORMANS NEDİR?

Alanyazında, performansa ilişkin pek çok tanıma rastlamak mümkündür. Ancak bu tanımlar incelendiğinde genel olarak ortak bazı noktalara vurgu yapıldığı görülebilmektedir. Bu ortak noktalardan yola çıkarak performansı şu şekilde tanımlamak mümkündür: Performans, bireye o alandaki bilgi, beceri ya da yeteneğini en az uygulama düzeyinde, genellikle gerçek yaşam koşulları ya da buna yakın koşullarda sergileme imkânı veren, yine çoğunlukla hem sürecin hem de ürünün gözlemlenebilmesine imkân tanıyan, çoğunlukla doğrudan gözlemlenebilen, bireyin gerek çevresini, gerekse bir materyali, objeyi ya da düşünceyi belli yönde manipüle etmesine izin veren bir takım eylemler, davranışlar serisidir. Tanımdan da anlaşıldığı üzere performans üst düzey zihinsel süreçlerin ve psikomotor davranışların sergilenmesine olanak tanımaktadır. Tablo 1'de öğrencilerin farklı derslere ilişkin performanslarını değerlendirmede dikkate alınabilecek performans örneklerine yer verilmiştir.

Tablo 1

Farklı Alanlara İlişkin Performans Değerlendirme Örnekleri

Konu Alanları	**Performans değerlendirme örnekleri**
Fen Bilimleri	Bir deneyi yapıp sonuçları rapor halinde düzenlemek
	Suyun kaldırma gücünü gösteren basit bir düzenek hazırlayıp, bu düzenek üzerinden konuyu anlatmak
	Ayın tüm evrelerini bir ay boyunca gözlemleyip, sonuçları kendi çektiği fotoğraflarla da destekleyerek derste sunmak
	Ampul, duy, kablo vb. malzemeleri kullanarak basit bir elektrik devresi kurarak ampulü yakmak
	Atık malzemeleri kullanarak günlük hayatta kullanılabilir bir alet üretmek
	Ankara'daki barajların son 10 yıldaki doluluk oranlarını araştırarak sonuçları yıllara ve mevsimlere göre grafik halinde sunmak
Matematik	Ailenin bir aylık harcamalarına dair fişleri inceleyerek, harcama kategorilerini tablo ve grafiklerle göstermek
	Gazetelerden birini seçerek son bir ayda çıkan şiddet olaylarının türlerini tablo ve grafiklerle göstermek
Türkçe	Hayvan haklarına ilişkin etkileyici bir öykü yazarak sınıfta okumak
	Kurgusal bir öykü yazmak
	Okuduğu bir romanın kahramanlarını betimleyen bir yazı hazırlamak
Sosyal Bilimler	Uygulanan sığınmacı politikasını komşularla ilişkiler çerçevesinde ele alarak gerekçeleriyle irdeleyip bir rapor hazırlamak ve sınıfta sunmak
	Bölgede süregelen savaşlardan birine ilişkin yazılı, sözlü basın vb. kaynakları taramak ve uluslararası savaş hukuku kuralları açısından inceleyerek bir rapor hazırlamak ve sınıfta sunmak
	Komşu ülkelerden birini seçerek bu ülkeyle ilişkileri geçmişten günümüze tarihi perspektiften inceleyen bir rapor hazırlamak
Müzik	Bir müzik enstrümanını kullanarak bir parçayı seslendirmek
	Sözlerini yazarak küçük bir beste yapmak
	Koro etkinliklerine katılmak
Resim	Kendi yaptığı resimleri sergilediği bir okul sergisi düzenlemek
	Sonbahar temalı bir resim yapmak
	Kendi çizimlerinin/resimlerinin de yer aldığı doğa ve çocuk temalı bir afiş hazırlamak
Beden Eğitimi	Günlük egzersiz yapma ile sağlıklı yaşam arasındaki ilişkiyi anlatan bir poster hazırlayıp sunmak
	Bir yaşlı bakım evinde uygulanabilecek günlük egzersiz programı tasarlayıp sınıfta sunmak
	İstenilen mesafeyi istenilen sürede koşmak/yüzmek
	Temel jimnastik hareketlerini sergilediği bir gösteri yapmak

Tablo 1'den anlaşıldığı üzere, çoğu zaman her ders ya da konu alanına ilişkin olarak öğrencinin performansını değerlendirmeyi gerektiren durumlar mevcuttur. Diğer yandan, tanım dikkate alındığında performansın ifade ettiği davranış yelpazesinin oldukça geniş olduğu görülmektedir. Bu yelpaze şarkı söylemekten; resim yapma, kompozisyon yazma, drama yapma, deney yapma, bir alana ilişkin proje hazırlama, bir konuya ilişkin eleştirel bir köşe yazısı hazırlama, bir konuda konferans verme, dans etme, yemek yapma ve bahçe peyzajı yapmaya kadar pek çok farklı forma uzanabilmektedir.

Performans; bireyin o alandaki bilgi, beceri ve yeteneğini sergilemesini gerektiren bir süreci işaret eder ve bu sürecin sonunda bir ürünü ortaya çıkarması beklenir. Dolayısıyla hem süreç hem de ürün boyutu vardır ve bu iki boyutun birlikte ele alınıp değerlendirilmesi değerlidir ve çoğu zaman gereklidir.

3. PERFORMANSIN DEĞERLENDİRİLMESİ

Performansın ölçülmesi ve değerlendirilmesi konusu bu alanda çalışan araştırmacıların odak noktalarından olmuştur. Performansın nasıl değerlendirileceğine ilişkin çeşitli görüş ve yaklaşımlar tartışılagelmektedir. Performans değerlendirme, alanyazında alternatif test teknikleri ya da tamamlayıcı test teknikleri (Çakan, 2008) adı verilen araç ve tekniklerle ölçülen ve otantik değerlendirme ya da performansa dayalı durum belirleme (Kutlu ve ark., 2010) olarak adlandırılan değerlendirme yaklaşımlarıyla birlikte anılan bir çabayı ifade etmektedir. Performans değerlendirme Uluslararası Eğitim Ansiklopedisi'nde testi alan bireyin karmaşık ve gerçek yaşam koşullarını içeren test durumlarına cevap olarak sunduğu ve bireyin bilgi, beceri ve yeteneklerine ilişkin gözlemlenebilir kanıtlar sunan bir takım ürün ya da performanslar üretmesini gerektiren etkinlik ya da etkinlikler bütünü olarak tanımlanmıştır. Bu bütüncül yaklaşım gerçek yaşam durumlarına çok benzeyen ortamlarda yetkinliği ölçer ve çok yönlü bir doğaya sahip olan 21. yüzyıl becerilerinin önemini vurgular. Diğer bir deyişle performans değerlendirme, bireyin sahip olduğu bilgi, beceri ve yeteneklere dair gözlemlenebilir bir takım somut ürün ya da performans kanıtları sergilemesine olanak tanıyan ve bunu bireye gerçek/olası gerçek yaşam koşulları içeren karmaşık bir görev ya da görevler takımı sunarak mümkün kılan etkinlikler bütünüdür. Performans değerlendirme, ölçtüğü davranış yelpazesinin geniş olması yönüyle bütüncül bir yaklaşımdır. Özellikle 21. yüzyıl becerileri olarak adlandırılan ve sınıf içi öğrenmelerin ve kuramsal bilgilerin gerçek yaşam problemlerini çözmeye yönelik olarak aktarılmasını teşvik eden ve bu görevlerin gereği olarak eleştirel düşünme, problem çözme, analiz etme, bilgileri sentezleme, karar verme gibi üst düzey zihinsel becerileri kullanmayı gerektirmesi açısından değerlidir.

Performans tanımlanırken vurgulandığı üzere, özellikle okullarda performans değerlendirmedeki temel amaç öğrenciye üst düzey zihinsel becerilerini (psikomotor becerilerin ön planda olduğu alanlarda üst düzey motor becerileri) sergileyebileceği ortamlar sunmak ve bu becerilerin gelişimini desteklemektir. Bunun bir gereği olarak öğrenciye verilen görevlerin söz konusu üst düzey becerileri kullanarak yerine getirilebilir görevler olması önemlidir.

Yine performans değerlendirmede verilen görevlerin bireyin gerçek yaşam koşullarına benzer (kurgusal durumlarda olabilir) koşulları içeren görevler olması beklenen bir durumdur. Çünkü performans değerlendirmenin önemli amaçlarından birisi de derste öğrenilen kuramsal bilgilerin pratik durumlara uyarlanarak anlamlı ve kalıcı hale getirilmesidir. Tanımdan da anlaşılacağı üzere performans değerlendirmede gözlemlenen bilgi, beceri ve yeterlikler çokludur. Tek bir performans görevi verilerek öğrencinin dönem boyunca edindiği pek çok bilgi, beceri ve yeteneğini birlikte kullanması ve sergilemesi sağlanabilmektedir. Tanımda vurgulanan "karmaşık görev" ifadesi üst düzey ve birden çok bilgi ve becerinin bir arada kullanılmasına vurgu yapmaktadır.

Performans değerlendirme, geleneksel ya da klasik olarak adlandırabileceğimiz pek çok ölçme aracıyla (çoktan seçmeli, doğru-yanlış, eşleştirme ve kısa cevaplı maddeler) yapılan ölçme ve değerlendirme anlayışından farklı olarak öğrenciden bir performansı gerçekleştirmesini ister. Performans görevi olarak öğrenciye verilen görev, öğrencinin gerçek hayatta karşısına çıkabilecek problem durumlarına mümkün olduğunca benzer olacak şekilde tasarlanır. Ayrıca bu görev genellikle bir bağlam ve çözülmesi gereken karmaşık bir problemi içerir.

Performans değerlendirme, zorlu bir görevi, bir yanıtı ve bir puanlama sistemini gerektirir (Johnson ve ark., 2008). Görev öğrenciyi düşünmeye, akıl yürütmeye ve bu doğrultuda bilgilerini, becerilerini ve deneyimlerini birleştirerek verilen problem durumuna çözüm yolları üretmeye teşvik edici niteliktedir. Zorlu görevle kastedilen şey, görevin başarıyla tamamlanabilmesi için öğrenciyi bu şekilde bir sürece mecbur bırakmasıdır. Bu doğrultuda verilen bir görev, öğrencinin problemi çözmek için somut materyallerle çalışmasını gerektirir ancak problemi çözmek için atılması gereken adımları belirtmez. Atılacak adımları, izlenecek süreçleri, öğrencinin düşünüp tasarlaması, kurgulaması beklenir. Örneğin, öğrenciye küçük bir pil motoru, kablo ve tekerlek verilir ve basit bir elektrikli araç yapması istenebilir. Öğrenciye bir takım geri dönüşüm materyalleri verilerek öğrenciden bu materyalleri kullanarak günlük hayatta kullanılabilecek bir alet ya da eşya yapması ya da tasarlamasının istenmesi ise bir başka örnektir.

Bazı durumlarda performans değerlendirmeleri, performansların ve ürünlerin bir kombinasyonunu içerir. Örneğin, deneyler; öğrencilerin hipotezler geliştirmeleri, deneyleri planlamaları ve yürütmeleri, bulguları özetlemeleri gibi bilimsel süreçlere katılmalarını gerektirir (Feuer & Fulton, 1993). Değerlendirmenin odak noktası deneyin uygulanması (yani bir gösterim), sonuçları açıklayan laboratuvar raporu (yani bir ürün) ya da her ikisi olabilir. Benzer şekilde bir yazma portfolyosu yalnızca nihai ürünleri yansıtıyorsa bir ürün olarak kavramsallaştırılabilir; ancak yazma süreci hakkında bilgi veren taslakların da çalışmaya eklenmesi, portfolyonun odağını süreci ve ürünü kapsayacak şekilde değiştirir (Johnson ve ark., 2008). Öğrencinin gerçek yaşama ilişkin matematiksel bir problemi çözmesi ve buna ilişkin adımları detaylarıyla açıklayarak raporlaması da öğretmene hem ürün hem de sürece ilişkin değerlendirme yapma olanağı sağlar. Değerlendirme odağının üründe mi süreçte mi olacağı istenen performans görevinin doğasına, değerlendirmenin amacına, öğrenciye ne ölçüde geribildirim sağlanmak istendiğine ve hatta süreçleri gözlemleme olanağının olup olmadığına göre değişebilmektedir.

Puanlama sistemi, öğrencilerden beklenen kritik bilgi ve akıl yürütmenin ne kapsamda olacağını belirler. Öğrenciden beklenen yeterliklerin neler olduğunu ifade eder. Süreçte gözlemlenmesi planlanan detaylar da dâhil olmak üzere, görevin gerektirdiği tüm performans yelpazesi puanlama sisteminde mutlaka kapsanmalıdır. Bu yapılan değerlendirmenin geçerliği açısından da önemlidir. Puanlama sistemi, doğru cevapların belirlenmesini, yapılan işlemlerin gerekçelendirilmesini, işlem süreçleri ya da adımlarını ve problem çözme yaklaşımları gibi detayları içerebilir.

Yukarıda bahsedildiği üzere, performansın ürün ve süreç boyutu olmak üzere iki boyutu mevcuttur. Her ikisinin birlikte değerlendirilmesi önemli ve gereklidir. Hangi alandaki performansın ölçüldüğüne göre bazen süreç boyutu daha fazla önem kazanır bazen de ürün boyutu daha fazla önem kazanır. Örneğin; fen bilgisi dersinde deney yapan bir öğrencinin deneyi başarılı bir şekilde tamamladı olarak kabul edilebilmesi için deneyi istenen sürede tamamlaması yeterli olmayacaktır. Deneyin başarılı tamamlandığını varsaymak için öğrencinin deney sırasında izlediği adımları dikkate almak da önemlidir. Örneğin, öğrenci deney sırasında güvenlik önlemi almamış ise (mesafe, yüzü ve gözleri korumak amaçlı gözlük kullanma vb.) deneyi istenen sürede tamamlaması yeterli olmayacaktır. Ürün kadar süreç de önemli olacaktır. Yine bir başka örneği ele alırsak; bomba imha biriminde görev almak isteyen bir polis adayının kurgusal bir bomba düzeneğini istenen sürede etkisiz hale getirmesi yeterli olmayacaktır. Bu işlem sırasında, süreçte izlediği adımlarda önemlidir. Örneğin, alanı güvenlik çemberine almak ve işleme başlamadan önce koruyucu kıyafetleri giymek de düzeneği etkisiz hale getirmek kadar önemlidir. Çünkü her iki örnekte de süreçte bu adımların izlenmemesi önemli zararlara ya da kayıplara neden olabilecektir. Ancak hangi adımların ya da süreçlerin izlendiği aşağıdaki örnekte öncelik olmayabilir.

Üniversitede ekoloji dersinden sorumlu öğretim üyesi dersi alan öğrencilerden Türkiye'deki artan kuraklık ve içme suyu sorununa çözüm olabilecek bir proje tasarlanmasını ister.

Burada öğrenciler, Türkiye'deki artan kuraklık ve içme suyu sorununa çözüm olabilecek, uygulanabilir bir proje üretebildikleri takdirde bu projeyi tasarlarken hangi süreçleri izlediklerinin belirleyici bir önemi yoktur ve bu süreçler bireylere göre farklılıklar gösterebilir. Dolayısıyla performansı değerlendirirken sürecin mi ürünün mü daha önemli olduğu ya da her iki boyutun eşdeğer önemde olup olmadığı istenilen performans alanına, performansın verilme amacına yani değerlendirmenin yapılış amacına göre değişebilecektir. Örneğin, verilen performans görevinin amacı öğrenciye geribildirim sağlamak yani biçimlendirmeye ve yetiştirmeye yönelik bir değerlendirme yapmak ise süreçlerin de gözlemlenmesi gereklidir çünkü bu sayede öğrencinin güçlü ve zayıf yönlerini etkili bir şekilde tespit edebilecektir. Bunun kararını öğretmenin vermesi gerekmektedir. Ancak odak noktası hangisi olursa olsun, her iki boyutun birlikte ele alınması ve değerlendirilmesi, aksi bir gerekçe yok ise o alandaki performansın gerçekleştirilme düzeyini ve niteliğini göstermesi açısından bilgilendirici olacaktır. Özellikle sınıf içi performans değerlendirmelerde bu durum önemlidir. Performansın ölçülmesinde süreç boyutunun da gözlemlenebilir olması performansa dayalı değerlendirme ya da performansa dayalı durum belirlemelerin geleneksel ölçme araçlarıyla yapılan çoğu ölçme ve değerlendirme işlemlerine bir üstünlüğüdür. Performansa dayalı değerlendirmeler, öğrencinin o alandaki ya da dersteki bilgi, beceri ve yeteneklerini doğrudan sergilemesine ve gözlemlenmesine olanak sağlaması nedeniyle eğitimde çok önemli bir yere sahiptir ve elbette geleneksel ölçme araçlarıyla birlikte birbirini tamamlar nitelikte kullanılmalıdırlar. Öğrencinin o alandaki yeteneklerinin doğrudan sergilenebilir ve doğrudan gözlemlenebilir olması öğrenciye ilişkin verilecek kararların daha geçerli olmasını sağlamaktadır. Birey gerçek performansını sergileme imkânına sahiptir. Çoğu durumda doğrudan gözlemlenebilir olması performans değerlendirmenin önemli bir avantajıdır. Bu sayede bireye ilişkin daha geçerli ve güvenilir kararlar verilebilmek mümkündür.

Performansa dayalı değerlendirmenin eğitime kazandırdığı belki de en büyük katkı, öğrencinin derste öğrendiği bilgi ve becerilerini gerçek yaşam problemleri üzerinde uygulama imkânı, transfer etme imkânı bulmasıdır. Böylelikle öğrenci, eğitim süresince edindiği kuramsal bilgilerini gerçek yaşamda uygulama imkânına sahip olur. Bu nedenle etkili bir performans değerlendirme düzeneğinin öğrenciye derste kazanılan bu bilgi ve becerilerini otantik, gerçek yaşam problemleri ya da olası gerçek yaşam problemleri üzerinde uygulama imkânı vermesi istenilen en önemli koşullardandır. Yapılan performans değerlendirmenin ne derece yeterli ve geçerli olduğu kullanılan ölçme

yönteminin iyi planlanmış, savunulabilir ve geçerli olmasına bağlıdır. Bu noktada, performansı ölçmek için kullanılan yöntemin ya da ölçme aracının, ölçülmesi amaçlanan bilgi, beceri ve yetenekleri ölçebilir nitelikte tasarlanan, dersin öğrenme çıktılarıyla uyumlu, onları istenilen düzeyde ve kapsamda ortaya çıkarmaya elverişli, uygun bir yöntem ya da araç olması gerekir. Bilindiği üzere, ölçmede kullanılan ölçme araçlarının ya da yöntemlerinin ölçülmek istenen öğrenme çıktılarına uygun olması belki de en önemli koşuldur (Çakan, 2008). Ölçülmek istenen kazanım / bilgi / beceri / yetenek / yeterlik ve kullanılan ölçme aracı ya da yönteminin doğru eşleşmesi geçerlik konusunun en önemli ön koşullarındandır.

Performansın ölçülmesi ve değerlendirilmesi gereklidir ancak kendi içinde güçlükleri de mevcuttur. Bunlardan birisi uygulanmasının ve değerlendirilmesinin uzun zaman gerektirmesidir. Özellikle kalabalık sınıf ortamlarında uygulanması uzun zaman gerektirir. Bu nedenle kullanışlılığı geleneksel değerlendirme yaklaşımlarına göre düşüktür. Yine uygulamanın maliyetli olması, puanlayıcı güvenirliğinin sağlanmasındaki zorluklar diğer güçlükler arasındadır (Kitabın 3. Bölümünde puanlayıcı güvenirliği detaylı bir şekilde incelenmiştir). Elbette bu güçlükler, performans değerlendirmenin yapılmamasının gerekçesi olmamalıdır. Öğretmen dönem boyunca hangi derste ve ne sıklıkta performansa dayalı değerlendirme yapacağına dikkatle karar vermelidir. Bazı dersler diğerlerine oranla daha yoğun performansa dayalı değerlendirmeleri gerektirmektedir. Örneğin, tarih dersine kıyasla fen bilgisi dersi performansa dayalı değerlendirmenin daha yoğun kullanıldığı derslerdir (Çetin & Çakan, 2010). Çünkü, pek çok öğrenme çıktısı bu şekilde değerlendirilmeyi gerektirmektedir. Dersin hangi öğrenme çıktılarının performans değerlendirmeyi gerektirdiğine, bu öğrenme çıktılarının hangi performans değerlendirme araç ya da yöntemiyle ölçülüp değerlendirileceğine, her bir performans değerlendirme araç ya da yönteminin geliştirilmesinde ölçülecek bilgi, beceri, ya da yeterliklerin neler olduğuna ve bunları doğru ve kapsamlı ölçebilecek şekilde geliştirilmesine, özellikle puanlama kriterlerinin yine değerlendirilecek bu bilgi, beceri ve yeteneklere uygun kapsam ve detayda olmasına dikkat edilmelidir.

4. PERFORMANS DEĞERLENDİRMEDE KULLANILAN BAŞLICA ÖLÇME ARAÇLARI

Eğitimde öğrenci performansının ölçülmesinde ve değerlendirilmesinde kullanılan pek çok ölçme aracı mevcuttur. Bunlardan en çok öne çıkanlar performans görevi, proje ve portfolyodur. Bunların tümü değişen derecelerde olmakla birlikte öğrencinin o alandaki performansını hem süreç hem de ürün boyutlarıyla gözlemleme ve değerlendirme imkânı sunmaktadırlar. Bu bölümde bu araçların her biri ayrıntılı olarak sunulmaktadır.

4.1. Performans Görevi

Performans görevi öğrencinin sınıf ortamında öğrenmiş olduğu özellikle üst düzey bilgi ve becerileri gerçek yaşam koşullarında, gerçek yaşam problemlerini çözerken kullanma imkânı bulduğu bir ölçme ve değerlendirme aracıdır. Bilgi ve becerilerin gerçek yaşama transfer edilmesini ve kullanılmasını sağladığı için öğrenci öğrenmelerinin kalıcı hale getirilmesine ve pekiştirilmesine imkân sağlar. Performans görevinde öğrenilenlerin transferi öncelikli hedeflerdendir. Sınıf içi öğrenmelerin gerçek yaşam durumlarına transferi amaçlanır. Öğrenci sınıfta öğrendiği bilgi ve becerilerini gerçek hayatla ilişkilendirdiği için öğrenmeleri anlamlı hâle gelir. Performans görevi öğrencinin sahip olduğu bilgi ve becerileri değerlendirme imkânı veren bir araç olmasının yanı sıra aynı zamanda öğrencinin verilen görevi yerine getirirken öğrendiği de bir etkinlikler bütünüdür. Performans görevi öğrencinin çok yönlü cevaplar vermesini gerektirir ve öğrenci cevaplarını farklı formlarda sergileme imkânı bulur. Öğrenci söz konusu performans görevini yerine getirirken o ders alanına ilişkin birden fazla öğrenme çıktısıyla ilişkili öğrenmelerini bir arada kullanma ihtiyacı duyar. Çünkü verilen görev bunu gerektirir. Çoktan seçmeli bir testte her maddenin dersin tek bir öğrenme çıktısına, davranışa yönelik olmasına dikkat edilirken burada sergilenmesi istenen performans sıklıkla birçok öğrenme çıktısının bir arada gerçekleştirilmesini gerektirir. Diğer bir deyişle öğrencinin o ders alanında öğrendiklerini kullanarak bir performans sergilemesi istenir ve bu performans birçok bilgi ve becerinin bir arada ve etkileşimli olarak kullanılmasını gerektirir.

Performans görevi sıklıkla ev ödevleri ile karıştırılır. Ev ödevleri derste öğretilen bilgi ve becerilerin evde tekrar edilmesini, pekiştirilmesini gerektiren aktivitelerdir. Örneğin matematik dersinde bölme işlemini öğreten bir öğretmen ev ödevi olarak bölme işleminin kullanılmasını gerektiren birkaç bölme problemi verebilir. Bu sayede öğrenci öğretilen bölme işlemine ilişkin kuralları uygulayarak sınıfta çözülen problemlere benzer problemleri evde çözer. Böylece öğrenci yeni öğrenmiş olduğu bilgileri tekrarlayarak pekiştirir. Kitabın 8. bölümünde ev ödevi ayrıntılı olarak ele alınmıştır. Performans görevinin öncelikli hedefi ise bilgi ve becerinin tekrar edilmesi değil öğrencinin bu bilgi ve becerileri gerçek hayata transfer ederek gerçek bir problemi çözerken kullanmasıdır. Bu anlamda performans görevleri genel olarak ödevlerden daha üst düzey davranışları ölçmeye ve değerlendirmeye yöneliktir. Dolayısıyla ödevler ile performans görevleri arasındaki ayrımın iyi bilinmesi gerekmektedir ki biri diğeriyle karıştırılmasın. Elbette ödev adı altında verilen ve üst düzey zihinsel becerilere yönelik etkinliklerde mevcuttur ancak genel olarak öğretmenlerin günlük etkinlik olarak verdikleri ödevlerin birincil kullanım amacı ve şekli sunulan bilgi ve becerilerin tekrar edilmesi ve pekiştirilmesidir. Benzer şekilde, performans görevleri yaygın olarak tek bir ders alanına ilişkin olarak verilmektedir. Bu yaygın bir uygulamadır. Ancak istenmesi halinde farklı disiplin ya da ders alanlarına ilişkin de verilebilmektedir.

İyi bir performans görevinin genel olarak şu özellikleri taşıması beklenmektedir:

1. Öncelikle, tüm diğer değerlendirme araçlarında olduğu gibi söz konusu dersin öğrenme çıktılarıyla uyumlu olmalıdır. Çünkü amaç bu öğrenme çıktılarının yoklanmasıdır.
2. Gerçek hayat problemleri ya da gerçek hayatta gerçekleşmesi olası senaryolar ya da simülasyon durumlarına yönelik olmalıdır.
3. Üst düzey zihinsel becerilere ya da 21. yüzyıl becerileri olarak nitelendirilen becerilere (eleştirel düşünme, problem çözme, iletişim becerileri, akıl yürütme vb.) yönelik olmalıdır. Daha alt düzey olan öğrenme çıktıları diğer ölçme ve değerlendirme araçlarıyla da yoklanabilirler.
4. Performansın değerlendirilmesinde dikkate alınacak olan kriterleri (hem süreç hem de ürüne yönelik) içeren detaylı bir puanlama kılavuzu olmalıdır (rubrik ya da derecelendirme ölçekleri) ve bu araçlar öğrencilerle başlangıçta paylaşılmalıdır. Bu kriterler son haline getirilmeden önce öğrenci ile paylaşılıp gerekli ekleme ya da çıkarmaların yapılması faydalıdır.
5. Öğrencinin performansı tamamlarken dikkat etmesi gereken (yapılmasına izin verilen ya da izin verilmeyen eylemler) kuralları içeren açık ve anlaşılır bir yönerge hazırlanmalıdır.
6. Görevin dersin hangi öğrenme çıktılarına yönelik olduğuna dair ayrıntılar netleştirilmelidir.
7. Görev kapsamına göre performans görevi sınıf içinde ya da sınıf dışında tamamlanabilecek nitelikte olabilmektedir.
8. Görev, verilen problem durumuna ilişkin yalnızca cevap değil kanıtlarını da içerecek nitelikte tasarlanmalıdır.

Performans görevi tanımlama, görev, yönerge ve puanlama olmak üzere dört ana bölümden oluşmaktadır. Bu bölümler aşağıda kısaca açıklanmıştır.

1. Tanımlama: Öğrencilere verilen performans görevinin hangi sınıf düzeyinde, hangi derse ait olduğu, görevin ilgili olduğu konu ya da üniteler, performans görevi ile yoklanması amaçlanan öğrenme çıktılarının neler olduğu (genellikle birden fazla öğrenme çıktısına ilişkin olmaktadır), performans görevinin tamamlanması ile hangi öğrenme çıktılarının gerçekleştirilmesinin beklendiği (örneğin, problem çözme ve iletişim) ve performans görevinin hangi puanlama aracı ile değerlendirileceği gibi genel bilgiler bu bölümde verilir.
2. Görev: Bu bölümde öğrencilere tanımlama bölümünde verilen detaylara uygun olarak çözüm bulmaları ya da araştırmaları gereken bir problem durumu verilir. Burada verilen görevlerin tanımlama bölümünde verilen kazanımlar ve öğrenme

çıktılarına uyumlu, onları yansıtan nitelikte olması şarttır. Aynı şekilde verilen görevlerin o yaş düzeyi öğrencilerin gerçekleştirebileceği görevler olmasına dikkat edilmelidir ki bu denge halihazırda dersin öğrenme çıktılarında dikkate alınmış olmalıdır. Burada sunulan göreve karar verilirken yukarıda belirtildiği üzere bir ev ödevi değil de performans görevi olduğundan emin olunmalıdır. Yani sınıf içi öğrenmelerin gerçek hayat durumlarına transferini gerektirecek, öğrencinin üst düzey zihinsel becerileri kullanmasını gerektirecek görevler olmalıdır. Aksi taktirde ev ödevi ile aynı işlevi görecektir ki bu da performans görevlerinin amacı değildir.

3. Yönerge: Bu bölümde öğrencilerin görevlerini yerine getirirken hangi noktalara ya da kurallara dikkat etmeleri gerektiği listelenir. Bu hususlar gerek görev öncesinde gerek görev sırasında ve gerekse görev sonrasında dikkate alınması ve uyulması gereken kuralları içerir. Örneğin, görevi yerine getirirken başkalarından yardım alabilecek mi? İnternet vb. kaynaklardan yararlanabilecek mi? Hangi araç, gereç ya da malzemeleri kullanabilecek? gibi detaylar netleştirilmelidir.
4. Puanlama: Bu bölümde, öğrencinin tamamlamış olduğu performans görevinin hangi kriterlere göre hangi düzeylerde puanlanacağına ilişkin kullanılacak puanlama aracı sunulur. Performans görevinin puanlanmasında rubrikler ya da derecelendirme ölçeklerinin kullanımı yaygındır. Hangi puanlama aracının kullanılacağına karar verirken değerlendirilmek istenen performansın türü, kapsamı, sınıf düzeyi (ilköğretim, ortaöğretim), görevin veriliş amacı gibi etkenler dikkate alınmalıdır. Burada kullanılacak puanlama araçlarının önceki bölümlerde değinildiği üzere söz konusu performansa ilişkin ürünü, süreci, çalışma sınıfta sunulacak ise sunu aşamasına ilişkin puanlanacak tüm davranışları kapsayacak nitelikte olması gereklidir. Gözlemlenecek davranışlar kapsam olarak yeterli olup, açık bir dille ifade edilmelidir.

Aşağıda yüksek lisans tezinde (Çetin, 2009, s. 125) kullanılmış olan ve Fen ve Teknoloji dersine yönelik hazırlanmış bir performans görevi örneği ve bu performans görevine ilişkin derecelendirme ölçeği sunulmuştur.

Tablo 2

5. Sınıf Fen ve Teknoloji Dersi Performans Görevi Örneği

Performans Görevi **Ders:** Fen ve Teknoloji **Sınıf:** 5 **Ünite:** Canlılar ve Hayat **Beklenen Performans:** Araştırma ve Gözlem Yapma **Kazanımlar:** 1a, 1b, 1c - 4a, 4b, 4c, 4d
Performans Görevi **Sevgili öğrenciler;** Sizden aşağıdaki konuyla ilgili araştırma yapmanız ve araştırmanızın sonunda konuyla ilgili bir rapor hazırlamanız beklenmektedir. 1. Yakın ve uzak çevrenizde yaşayan çeşitli canlılar nelerdir? Canlılar benzerlik ve farklılıklarına göre kaça ayrılmaktadır? Canlıların incelenmesinde canlıları sınıflandırmak ne gibi kolaylık sağlamaktadır? araştırıp raporlaştırın. 2. Çevrenizdeki hayvanların benzerlik ve farklılıklarını nelerdir? Hayvanlar bir omurgaya sahip olup/olmama açısından kaça ayrılmaktadır? Omurgalı hayvanları sınıflayan ve omurgalı hayvan sınıflarının genel özelliklerini araştırıp raporlaştırın.
Görevi Hazırlarken Dikkat Etmeniz Gerekenler **1)** Görevinizin içeriğini belirleyin, içeriğe uygun ve araştırmanızı zamanında sunabilecek bir plan yapın. **2)** Görevinizi yaparken aile büyüklerinizin ve çevrenizdeki diğer insanların bilgilerine başvurabilirsiniz. **3)** Görevinizi yaparken gözlemlerinizden, kitaplardan, ansiklopedi, internet vb. kaynaklardan yararlanabilirsiniz. **4)** Hazırladığınız raporu sınıfta daha etkili sergilemek amacıyla resimler, fotoğraflar, çizimler vb. ile zenginleştirmelisiniz. **5)** Görevinizin içeriğine uygun konu başlıkları seçmelisiniz. İçerikte bulunan bilgilerin doğru olmasına dikkate etmelisiniz. **6)** Görevinizi açık ve anlaşılır bir şekilde hazırlayın. Raporunuzun yazım ve düzenine özen gösterin. **7)** Hazırladığınız ödevinizi 2 hafta içinde tamamlamış ve sunmuş olmalısınız. **8)** Sunumunuz esnasında zamanı tasarruflu kullanmalısınız ve konuyla ilgili soruları yanıtlamalısınız. Akıcı düzgün bir anlatım sergilemelisiniz.

PERFORMANS GÖREVİ DERECELENDİRME ÖLÇEĞİ					
Öğrencinin Adı/Soyadı: **Okul:** **Puanlayıcı Öğretmen:** **Sınıf/şube:**					
Performans görevi: **Ölçütler**	**PERFORMANS DERECELERİ**				
ÖDEV İÇERİĞİ	**5**	**4**	**3**	**2**	**1**
1. Amaca uygun plan yapma.					
2. İçeriği amaca uygun hazırlama					
3. İçerikte doğru bilgilere yer verme					
4. Uygun konu başlıkları seçme					
5. Resim, fotoğraf, çizim gibi görsel materyallerle çalışmayı destekleme					
ÖDEVİN SUNUMU					
1. Konuyla ilgili çelişkili ifadeler kullanmama					
2. Araştırmanın içeriğine uygun açıklamalar yapma					
3. Konuya uygun örnekler verme					
4. Kullandığı materyaller konun içeriğine uygunluğu					
5. Sorulan soruların hepsine doğru yanıt verme					
6. Akıcı, düzgün bir anlatım sergileme					
7. Zamanı iyi kullanma					
ÖDEVİN DÜZEN VE TESLİMİ					
1. Çalışmasında açık ve anlaşılır bir anlatım kullanma					
2. Kullanılan kaynaklar kaynakça kısmına yazma					
3. Ödevin yazımına özen gösterme					
4. Düzenine özen gösterme					
5. Ödevi zamanında teslim etme ve sunma					
KAYNAK KULLANIMI					
1. Yeterince yazılı kaynaklardan yararlanma					
2. Yeterince görsel kaynaklardan yararlanma.					
3. Yeterince çevresindeki kaynak kişilerden yararlanma					
TOPLAM PUAN:					

4.2. Proje

Proje birey performansının belirlenmesi ve değerlendirilmesi amacıyla pek çok disiplin alanında kullanılmaktadır. Bu alanlardan birisi de eğitimdir. Okullarda öğrenci performansını değerlendirmek amaçlı olarak sıklıkla kullanılmaktadır. Proje hem bir öğrenme etkinliği hem de performansı değerlendirme aracı olarak kullanılabilmektedir. Bu bölümde performans değerlendirme aracı olma özelliği üzerinde durulacaktır. Proje, öğrencilerin sınıfta öğrendikleri bilgi ve becerileri kullanarak gerçek hayattaki yine gerçek bir probleme ilişkin yeni bir çözüm yolu bulmak ya da yeni bir ürün ortaya koymak amacıyla gerçekleştirdiği etkinlikleri ifade etmektedir.

Projeler üst düzey zihinsel becerileri uygulamayı gerektirir ve böylelikle bu becerilerin değerlendirilmesine olanak tanır. Proje ve performans görevleri birbirine benzerlikleri nedeniyle sıklıkla karıştırılır. Proje ve performans görevlerinin benzer olduğu özellikleri ve farklılaşan özellikleri mevcuttur. Bunlar kısaca:

1. Proje genellikle daha uzun sürelidir. Tamamlanması haftalar hatta aylar gerektirebilir. Fakat performans görevi, görevin içeriğine göre değişmekle birlikte birkaç ders saati ya da birkaç günde ya da haftada tamamlanabilir.
2. Proje konusunun kapsamı geniştir. O konuya ilişkin çok kapsamlı ve derin araştırma gerektirir. Oysa performans görevinin kapsamı çok daha dar ve somuttur.
3. Proje disiplinler arası çalışma/araştırma gerektirebilir. Oysa performans görevi çoğunlukla belli bir disiplin ya da konu alanına odaklıdır.
4. Proje daha karmaşık yapıdaki gerçek problemlere, durumlara ilişkin uygulama gerektirirken, performans görevi günlük hayatta karşılaşabilecek ancak çok daha dar kapsamlı bir problem durumuna yönelik olabileceği gibi kurgusal bir duruma ilişkin de olabilir. Örneğin, kuraklık konusu ile ilgili olarak proje yeraltı sularını korumaya dönük bir ürün/tasarım gerektirirken performans görevi evde kullanılan atık suların tekrar nasıl değerlendirileceğine ilişkin bir ürün/tasarım geliştirmeyi gerektirebilir.
5. Projenin önemli bir amacı öğrencinin istenen konuya ilişkin olarak farklı kaynaklardan bilgi toplayıp, araştırma yapıp özgün bir ürün ortaya koymasıdır. Bu ürün yazılı bir rapor, bir çözüm yöntemi ya da özgün bir araç olabilir. Oysa performans görevinin ana amacı öğrencinin derste öğrendiği bilgi ve beceriyi günlük hayata aktarması ve bunu yaparken problem çözme, eleştirel düşünme, iletişim gibi becerileri kullanmasıdır.
6. Projede öğrenci, ilgileri doğrultusunda o alana ilişkin bir konu seçebilir. Oysa performans görevinde görev, öğretmen tarafından tanımlanarak verilir.
7. Projede konu gerçek problemler veya durumlardır. Oysa performans görevinde günlük hayatta karşılaşılabilecek daha dar kapsamlı bir problem olabileceği gibi kurgusal bir duruma ilişkin problemlerde verilebilir.

8. Projede sürecin sonunda nasıl bir ürünün ortaya çıkacağı büyük ölçüde belirsizdir. Öğrenci sorularını kendi oluşturur ve bu sorulara çözüm arar. Konuya nasıl ve nerden yaklaştığına göre süreç sonunda farklı ürünler ortaya çıkarabilir. Oysa performans görevinde görev öğretmen tarafından tanımlandığı için ortaya çıkacak olan ürün büyük ölçüde yapılandırılmıştır ve bellidir.
9. Projede kontrol ve yönlendirme büyük ölçüde öğrencidedir. Öğrenci konu seçiminde, problemi ifade etmede, sorularını oluşturmada, izlenecek yollarda, hangi kaynakların kullanılacağında ve nerelerden bilgi edinileceği gibi tüm aşamalarda büyük ölçüde özgürdür. Performans görevinde görev büyük ölçüde tanımlanmış olduğu için büyük ölçüde öğretmen yönlendirmesi söz konusudur.

Bu farklılıklarla beraber hem proje hem de performans görevi öğrencinin problem çözme, eleştirel düşünme, yaratıcı düşünme ve araştırma yapabilme becerilerini ölçmeye uygundur. Projeler de diğer ölçme ve değerlendirme araçlarında olduğu gibi o ders ya da konu alanına ilişkin birtakım öğrenme çıktılarının yoklanması amacıyla kullanılmaktadır. Dolayısıyla bu amaçlara hizmet edecek şekilde tasarlanmalıdırlar.

Öğretmen performans görevi mi yoksa proje mi kullanması gerektiğine karar verirken dersin öğrenme çıktıları, öğrenme ihtiyaçları, öğretimin düzeyi, zaman gibi etkenler yanında yukarıda karşılaştırmada vurgulanan özgünlük, öğrencinin ne ölçüde sorumluluk almasını istediği, ürün ya da performansın ne derece özgün olması gerektiği gibi noktaları dikkate alınarak karar vermelidir. Tablo 3'te farklı derslerde kullanılabilecek proje örneklerine yer verilmiştir.

Tablo 3

Farklı Ders Alanlarına İlişkin Proje Örnekleri

Alan	**Proje Örnekleri**
Sosyal Bilgiler	Şehrindeki sokak hayvanlarının yaşam haklarını sürdürebilmesi için uygulanabilir bir plan tasarlamak.
	Yaşadığı şehrin turistik bölgelerini tanıtan bir video çekerek hazırlamak.
Türkçe	Yabancı öğrencilere Türkçe öğretimine ilişkin bir öğretim programı hazırlamak.
Matematik	Yaşadığı sokağı/caddeyi krokisine uygun olarak üç boyutlu geometrik cisimlerden faydalanarak tasarlamak
Fen Bilimleri	Bitkilerin farklı ısı, ışık ve su şartlarında büyümesini inceleyip sonuçları görselleriyle beraber raporlamak.
	Yaşadığı mahalledeki geri dönüşümleri daha üst seviyeye çıkarabilecek bir plan hazırlamak.
	Bir web sitesi tasarlamak.

4.3. Portfolyo

Performansa dayalı değerlendirmelerde ya da durum belirlemede sıklıkla kullanılan araçlardan bir diğeri de portfolyolardır. Butler ve McMunn (2006) portfolyoyu öğrencinin çabasını, ilerlemesini veya yeterlilik derecesini gösteren amaçlı ve bütünleşik bir öğrenci çalışmaları koleksiyonu olarak tanımlamıştır. Başka bir tanımla portfolyo, öğrencinin o derse ilişkin çabasını, gelişimini ya da başarısını tanımlamak amaçlı öğrenci çalışmalarının organize bir şekilde toplanması olarak ifade edilmektedir (Airasian, 2001; Danielson & Abrutyn, 1997; Johnson & Johnson, 2002).

Portfolyo eğitim dışında diğer iş kollarında, özellikle işe alımlarda (örneğin bir şirkete eleman alımında) öteden beri kullanıla gelmektedir. Portfolyoların sınıflarda kullanılması ise 1970'lere uzanmaktadır (Belgrad, 2013). Portfolyo değerlendirmenin "otantik" (gerçek ve anlamlı) bir değerlendirme olması nedeniyle sonraki yıllarda öğretmenler arasında kullanımı artarak devam etmiştir (Mertler, 2001). Portfolyolar dönem boyunca öğrencinin dersin farklı öğrenme çıktılarına ilişkin geniş bilgi, beceri ve performans yelpazesine ilişkin pek çok ürününü, çalışmaya ilişkin kanıtları ve dokümanları bir arada barındırmaktadır. Bu yönüyle performans görevi ve proje gibi diğer performansa dayalı durum belirleme ya da değerlendirme araçlarından farklılaşmaktadır.

Portfolyoya seçilen çalışmalar, öğrencinin o derse/konu alanına ilişkin sergilediği çabasını, gelişimini ve performansını toplu olarak sergilemeye ilişkindir. Portfolyoya dayalı durum belirlemenin ya da öğrencinin yeterlik düzeyini tespit etmenin başarısı, öğrencinin bir ya da birden fazla alandaki ürününü, süreci, çabayı ya da gelişimini gösteren çalışmalarının amaçlı bir şekilde toplanmasına bağlıdır (Montgomery, 2001). Dolayısıyla, portfolyolara eklenen çalışmalar gelişigüzel değil belli bir amaca yönelik olarak seçilmiş çalışmalardır. Diğer ölçme araçlarında olduğu gibi portfolyo da o dersteki öğrenme çıktılarına uygun şekilde kullanılmalıdır. Öğrencinin o ders ya da konu alanında gerçekleştirmesi beklenen öğrenme çıktısı ya da öğrenme hedeflerinin varlığını ve ne düzeyde gerçekleştirildiğini kapsamlı bir şekilde yansıtan, uygun nitelikte ve çeşitte çalışma örneği içerilmelidir.

Öğrencinin o dersteki çalışmalarına ait koleksiyonunun portfolyo olarak kabul edilebilmesi için 3 temel koşul öne sürülmüştür (Airasian, 2001; Gallagher, 1998). Bunlar:

1. Bu çalışmaların belli bir amaç doğrultusunda toplanması
2. Öğrencinin dosyaya hangi çalışmaların ekleneceği ya da süreç içerisinde hangi çalışmaların çıkarılacağı konularında aktif rol oynaması ve
3. Öğrencilerin kendi öğrenme süreçleri ve deneyimleri üzerinde düşünmeye aktif olarak katılmaları.

Portfolyoya eklenen çalışma ya da ürünler gelişigüzel değil belli bir amaç doğrultusunda ve sistematik olarak toplanmış olan materyaller, çalışmalar veya ürünlerdir. Portfolyo içerisinde sunular, projeler, deney raporları, günlükler, video ya da ses kaydı gibi dönem boyunca o derste gerçekleştirilen çalışmalara ve performansa ilişkin görsel, işitsel vb. örnekler saklanabilir. Tüm bu dokümanlar öğrencinin o derse ya da konu alanına ilişkin olarak neleri bildiğini ve neleri yapabileceğine dair kanıtlar sunmaktadır. Portfolyo içerisinde bu kanıtların yanı sıra puanlamada kullanılan rubrikler, derecelendirme ölçekleri, kontrol listeleri ve öğrenciye ait öz yansıtma, öz değerlendirme formları da bulunabilmektedir (Mertler, 2003).

Portfolyo diğer ölçme araçlarından farklı olarak öğrenciyi yalnızca sınırlı bir zaman sürecinde, anlık değerlendirmez ve aynı şekilde sınırlı bir içeriğe dayalı değerlendirmez. Aksine öğrencinin becerilerine ve yetkinliklerine dair kanıtları uzun bir zaman diliminde toplama ve inceleme imkânı sağlar. Portfolyo süreci öğrencilerin kendi öğrenme süreçleri hakkında düşüncelerini, değerlendirmelerini ve duygularını ifade etme imkânı sağlar. Öğrenci ne öğrendiğini, nasıl öğrendiğini, öğrenme sürecinde hangi güçlü/zayıf yanları olduğunu kendisi fark eder. Bu öğrencinin öz değerlendirme yapmasını ve kendi öğrenme süreçlerine daha bilinçli yaklaşmasını sağlar. Öğrenci kendi gelişimini değerlendirmek ve çalışmalarını istenilen ölçütlere uygun hale getirmek amacıyla bu çalışmaları gözden geçirme sorumluluğunu alır. Öğrencinin dosyayı sahiplenmesi, portfolyoya dayalı yapılan durum belirlemenin ve değerlendirme sürecinin başarısı için önemlidir (Oosterhof, 2001).

Öğrenci merkezli bir anlayış benimsendiği için portfolyo sürecinde öğretmenin asıl görevi öğrenciye kılavuzluk etmektir. Bu kılavuzluk, dosyaya eklenecek çalışmaların seçiminden tutunda öğrencinin öz yansıtma yaparak kendi çalışmalarına ilişkin duygu, düşünce ve deneyimlerini sağlıklı bir şekilde gözden geçirmesine ve değerlendirmesine yardımcı olunmasına kadar uzanmaktadır. Öğretmen gerek portfolyoya eklenecek çalışma örneklerinin türlerini belirlemede, gerekse bu çalışmaların nasıl değerlendirileceğine ilişkin kriterlerinin belirlenmesi sürecinde her bir öğrenci ile iş birliği içinde çalışmalıdır. Bu görüşmelerin uzunluğu, öğretmenin öğrenciyle ne sıklıkta görüştüğüne, öğrencinin portfolyo kullanımına ne derece aşina olduğuna ve sınıf düzeyine göre farklılaşmaktadır. Mertler (2003), öğretmenin diğer iş yüklerine bağlı olmakla beraber portfolyoyu gözden geçirme sıklığını yaklaşık 4 ile 6 hafta ve süre olarak da öğrenci başına ilköğretimde her bir seans için bir saat, ortaöğretimde ise muhtemelen daha uzun bir süre almasının beklendiğini ve bu süre uzunluğunun öğretmenlerin portfolyo kullanım tercihlerini azaltıcı ana etken olduğunu ifade etmiştir.

Portfolyolar, öğrencilere ve öğretmenlere dosya içerisindeki çalışmalar üzerine değerlendirme yorumları ekleyebilme imkânı da sağlamaktadır. Bu çerçevede öğrenciler, portfolyolarına dâhil ettikleri çalışmaları seçme nedenlerini açıklayabilir; bu çalışmaların güçlü yönlerini ve geliştirilmesi gerektiğini düşündükleri noktaları

değerlendirebilirler. Benzer şekilde, öğretmenler ise başta dersin kazanımları olmak üzere, öğretimin gerekli kıldığı niteliği de dikkate alarak öğrenci çalışmalarının kalite düzeyine ilişkin görüşlerini ifade edebilir, aynı şekilde çalışmanın daha iyi hale nasıl getirileceğine ilişkin önerilerde bulunabilirler (Chase, 1999).

Bu türden müdahaleler hem öğretmene hem de öğrenciye öğrenim süreci boyunca yapılan çalışmalar üzerine düşünme fırsatı tanımaktadır. Öğrencinin kendi çalışmalarına bu şekilde eleştirel yaklaşması, süreç içerisinde dosyaya hangi çalışmaların ekleneceği ya da çıkarılacağına dair zaman ayırması öğrencinin gelişimine katkıda bulunmaktadır (Mertler, 2003).

Öğrencilerin dosyalarına kendi öğretim hedefleri doğrultusunda diğer öğrencilerden farklı çalışma örneklerini ekleme imkânına sahip olmaları, portfolyoların öğrencilerin özel eğitim ihtiyaçlarına daha iyi cevap verebilmesini de mümkün kılmaktadır. Öğrenci ve öğretmen dosyaya sürekli erişime sahiptir. Genellikle dosya sınıfta kalır ve öğrenci müsait olduğunda inceleyebilir. Dosya diğer öğrenciler ve öğretmenlerde dâhil olmak üzere öğrencinin istediği kişi ile paylaşılabilir. Portfolyoda öğrencinin bitmiş ürünü yanında bu ürüne ilişkin taslak çalışmalar da bulunabilir. Örneğin, yazdığı bir makalenin önceki halleri ve son hali gibi. Bu durum öğrencinin çalışmasına ilişkin sürecin, öğrencinin bu süreç içerisindeki gelişiminin ve çabasının gözlemlenebilmesine imkân tanır. Bu çalışmalarda bitmiş çalışmalara ek olarak değerlendirilir (Mertler, 2001).

Portfolyolar, diğer ölçme araçlarından farklı olarak, öğrencilerin yapamadıklarını göstermeye değil, başarılı oldukları alanları sergileme noktasına vurgu yapar ve öğrencilerin başarılarına odaklanır. Öğrencilerin, en iyi çalışmalarına ve belirli bir zaman dilimindeki gelişimlerine önem verir. Özellikle vitrin portfolyolarda öğrenci en iyi çalışmalarını seçmeye teşvik edilir. Öğretmenle yapılan değerlendirme ve tartışmaların odak noktası çalışmada gözlemlenen eksikliklerden ziyade öğrencinin göstermiş olduğu güçlü yanlardır. Bu da öğrencilerde başarı duygusunu güçlendirmesi nedeniyle öğrenci açısından motive edici olarak kabul edilebilir.

Genel olarak portfolyoların üç temel bölümden oluştuğu söylenebilir. Bu bölümler; öz geçmiş, ürünler ve yansıtmadır. Özgeçmiş bölümünde dosyada yer alan ürünlerin oluşturulmasında izlenen süreçlere dair bilgiler; ürünler bölümünde öğrencinin performansını, becerilerini, çabasını gösteren çalışma örnekleri; yansıtma bölümünde ise öğrencinin portfolyoda sergilediği çalışma örneklerine ilişkin deneyimleri, görüşleri ve duyguları yer alır. Bunlar dışında portfolyoda genel olarak şunlara yer verilir; kapak sayfası, çalışma takvimi, içindekiler, özgeçmiş/öğrenci tanıtımı, ürün kontrol listesi, çalışma örnekleri/ürünler, öz değerlendirme-yansıtma, öğrenci ve öğretmen yorum kartları, portfolyo değerlendirme rubriği ya da derecelendirme ölçeği.

Farklı ihtiyaç ve amaçları (genellikle öğrencilerin önemli çalışmalarını sergileme, belli bir zaman süresince öğrenmelerin gelişimini izleme, öğrencilerin başarı durumlarını değerlendirme) gerçekleştirmek üzere farklı portfolyo türleri geliştirilmiştir. Ancak türü ne olursa olsun portfolyoların iki temel hedefi karşılaması gerektiği vurgulanmıştır (Gredler, 1999: 1) öğrencinin kendi çalışması üzerinde düşünmesini gerektirecek şekilde kişisel sorumluluk alacağı bir yapının tasarlanması 2) öğrencinin zaman içindeki gelişimini belgelemesi.

Kullanım amacına göre portfolyolar genellikle 4 farklı şekilde sınıflandırılmaktadır (Gredler, 1999; Mertler, 2003):

1. Derleme portfolyoları
2. Vitrin portfolyoları
3. Değerlendirme portfolyoları
4. Sınıf portfolyoları

4.3.1. Derleme Portfolyoları

Derleme portfolyosu; "durum belirleme portfolyosu", "gelişim portfolyosu"., "ilerleme portfolyosu" olarak da adlandırılmaktadır. Sınıfta yaygın kullanılan bir türdür. Öğrencinin belli bir zaman süresindeki gelişimini izlemeyi ve göstermeyi amaçlar. Öğrenci gelişiminin sürekli bir kaydı sağlanır (Gredler, 1999). Bu portfolyoların temel vurgusu çaba, ilerleme veya başarı alanlarından biri ya da hepsine ilişkin kanıt toplamaktır. Çaba ve ilerlemenin değerlendirilmesi başarıya göre daha güç olduğundan (ve not vermede başarı dışı etkenler olarak görüldüğünden) portfolyolar bu niteliklerin değerlendirilmesi ve belgelenmesi açısından kullanışlıdırlar. Bu portfolyolar diğer tür değerlendirme sonuçlarını da içerebilirler (Gredler, 1999). Örneğin; öğretmen kontrol listeleri, anekdot kayıtları, performans değerlendirme sonuçları ve hatta yazılı sınavların sonuçları gibi. Bu tür, yalnızca öğrenci ürünleriyle sınırlı değildir. Odak noktası, her öğrencinin gelişimi ve başarıları hakkında bütüncül nitelikte kanıtlar sunmaktır. Bir dönem ya da okul yılı boyunca öğrencinin gelişimini sürekli kayıt altına alır. Dolayısıyla süreç odaklı olduğu söylenebilir. Portfolyoya periyodik olarak yeni eklemeler yapılır ve gerekirse başka çalışmalar kaldırılır.

4.3.2. Vitrin Portfolyoları

Vitrin portfolyosu "en iyi çalışma portfolyosu" (Nitko, 2001), "ürün dosyası" (Montgomery, 2001) olarak da bilinir. Bu türün ana amacı öğrenci başarılarını vurgulamak, öne çıkarmak ve sergilemektir (Gredler, 1999). Bu dosya öğrencinin en iyi çalışmaları ya da en çok beğendiği çalışmalardan oluşan bir koleksiyondur ve öncelikle öğrenci tarafından seçilmelidir (Gredler, 1999). Vitrin dosyasında hem ürün hem de sürece dair kanıtlar eklenebilir. Asıl önemli olan, eklenen örneklerin öğrencinin en yüksek başarı seviyesini göstermesidir. Bu öğrencilerde büyük bir gurur ve başarı duygusu uyandırır ve neler yapabileceğini gösterme imkânı sunar (Mertler, 1999).

Öğrencinin kendi çalışmalarına dair duygu ve düşünceleri, bu tür portfolyoların genellikle önemli bir unsurudur. Öğrenciler en iyi çalışmalarından örnekler seçmekten sorumlu oldukları için, öğrenci yansıtmaları genellikle bunları neden "en iyi" çalışma olarak gördüğüne ve portfolyoya dâhil etmek üzere seçtiğine dair fikir verir (Rolheiser ve ark., 2000). Bu yönüyle ürün odaklı olduğu söylenebilir.

4.3.3. Değerlendirme Portfolyoları

Bu portfolyo türünün amacı ilk ikisinden farklı olarak, kamuoyuna veya devlet kurumlarına raporlama yapmaktır. Okullar ulusal ya da uluslararası değerlendirme programlarına katılırlar, zaman zaman yeni bir programın uygulanması için pilot seçilirler. Değerlendirme portfolyoları bu okulların, ya da uygulanan öğretim programlarının, yeni müfredatın, öğretim çalışmalarının etkinliği hakkında paydaşlara ya da sorumlulara rapor vermek için kullanılır. Dolayısıyla özel uygulamalara sahiptir.

4.3.4. Sınıf Portfolyoları

Öğrencilerin farklı başarılarına ilişkin özet dokümanları içerir (özet niteliğinde). Öğrencilerin çalışmalarına ilişkin duygu ve düşüncelerinin değerlendirmelerini içermesi şart değildir. Ana kullanım amacı velilere yönelik raporlama yapmaktır (Mertler, 1999).

4.3.5. Portfolyo Kullanımında Dikkat Edilecek Bazı Unsurlar

1- Öğrencilerin çalışmalarına dair öz değerlendirme, öz yansıtma yapmaları öğrenmenin bireyselleştirilmesi öğrenci gelişimi için önemlidir. Bu nedenle öğretmenler öz değerlendirme, öz yansıtma yapmanın amacını anlamaları açısından ihtiyaç duyan öğrencilere yardım, rehberlik etmelidirler. Bu amaç doğrultusunda öğretmen, öğrenciyi düşünmeye sevk edici bazı sorular yöneltebilir (Nitko, 2001). Öğrencinin bu bağlamda kendi öğrenmelerini, düşüncelerini, hislerini ve gelişimini sorgulamasına yardımcı olan olası sorular aşağıdakiler şeklinde olabilir (Mertler, 2001):

- Bu ödevi tamamlarken hangi adımları ve süreçleri izledin?
- Fikirlerini nereden edindin? Daha fazla bilgi topladıkça düşüncelerini nasıl gözden geçirdin?
- Bu projeyi/çalışmayı neden portfolyona koymaya karar verdin? Bu çalışmanı tamamladığın diğer çalışmalardan daha etkili kılan şey nedir?
- Çalışmanın güçlü yanları nelerdir?
- Çalışmanın zayıf yanları nelerdir? Neleri düzeltmen gerektiğini düşünüyorsun?

Bu ve benzeri sorular öğrenciyi düşünmeye sevk eder. Yine, bu tür sorular genellikle öğretmenin geliştirmiş olduğu ve çalışmalarını değerlendirmek amacıyla öğrenciye tamamlatılan öz değerlendirme ve öz yansıtma formlarında sunulabilirler. Öğrenci bu formları her bir çalışması için tamamlar ve bu şekilde çalışmalarını eleştirel bir bakış açısıyla ele alması sağlanır. Aşağıda bu amaçla öğrenciye tamamlatılabilecek yansıtma sorularından oluşan bir form verilmiştir.

Tablo 4

Öğrencilere Yöneltilebilecek Yansıtma Soruları Örnek Formu

Öğrenci Portfolyo Öz Yansıtma Formu **İsim:** **Tamamlanma tarihi:**
Bu çalışmamı en iyi çalışmam olarak seçtim, çünkü …… …… ……
Çalışmamda en çok beğendiğim noktalar şunlardır: …… …… ……
Bu çalışmamı daha iyi hale getirmek için geliştirmem gereken noktalar şunlardır: …… …… ……
Bu çalışmayı yaparken öğrendiğim en önemli şeyler şunlardır: …… …… ……
Bu çalışmayı yaparken en çok zorlandığım yerler şunlardır: …… …… ……
Bu çalışmaya kaç puan verirdin? Neden? …… …… ……

2- Portfolyoya eklenecek çalışma örneklerinin seçiminde mutlaka öğrenci de dâhil edilmelidir. Bu öğrencinin portfolyoyu sahiplenmesine, kişileştirmesine ve gelişimine imkân sağlar. Öğrenci kendi öğrenmelerine ilişkin sorumluluk alır ve farkındalığı artar.

3- Benzer şekilde, performans değerlendirme ölçütlerinin belirlenmesi sürecinde de öğrenci işe katılmalıdır. Öğrenci, çalışmalarının hangi ölçütlere göre değerlendirileceğini bilmelidir. Bu, öğrenci merkezli bir değerlendirme sistemi için önemlidir.

4- Portfolyo sadece nihai ürünleri değil, süreçlere ilişkin kanıt niteliğindeki belgeleri de barındırabilir. Bunun için yazılı çalışmaların, projelerin, problem çözme etkinliklerinin veya çizimlerin hem ilk hâlleri/taslak hem de nihai hâlleri portfolyoya eklenebilir (Airasian, 2001). Böylece öğretmen; öğrencinin hangi süreçlerden geçtiğine, nasıl bir gelişim izlediğine dair fikir sahibi olur.

5- Portfolyoda kapsanan çalışmaların sayısı ve niteliği, öğrencinin o derste ne derece yetkin ya da yeterli olduğuna dair geçerli bir karar vermeye yetecek standartlarda olmalıdır. Portfolyoya eklenebilecek çalışmalar neredeyse sınırsızdır. Airasian (2001), Montgomery (2001) ve Tanner (2001) portfolyoya eklenebilecek çalışmaları şu şekilde sıralamıştır: denemeler ve yazılı raporlar, şiirler, günlükler, mektuplar, haritalar, fotoğraflar, ses kasetleri ve video kayıtları, afişler, sanat çalışmaları, laboratuvar raporları, problem çözümüne yönelik yazılı çözümler, buluşlar. Mertler (2001) bunlara ek olarak şu ögelerinde eklenebileceğini ifade etmiştir: Afişler, laboratuvar raporları, sanat çalışmaları, bir problemin çözümüne yönelik yazılı çözümler, kısa sınavlar (quizler) ve testler. Görüldüğü üzere bir portfolyoda kapsanacak çalışma örneği türü son derece geniştir ve hem yazılı hem görsel hem de işitsel ürünleri içerebilir. Ancak akılda tutulması gereken belki de en önemli nokta şudur; önemli olan eklenen çalışmaların öğretim hedefleriyle uyumlu olması ve öğrencinin öğrenme sürecini yeterince yansıtmasıdır.

4.3.6. Portfolyoların Puanlanması

Portfolyoların puanlanması aşaması oldukça zahmetli ve zaman alıcı bir işlemdir. Ancak belli hassasiyetlere dikkat edilerek puanlama güvenirliğini ve geçerliğini arttırmak mümkündür. Portfolyonun puanlanması görevi birincil olarak öğretmene aittir. Ancak öğrencide dosya sahibi olması sebebiyle bu sürece dâhil edilmelidir. Bu dâhil ediş, gerek puanlama rubriklerinin ya da derecelendirme ölçeklerinin geliştirilmesi aşamasında gerekse performans kriterlerinin belirlenmesi aşamalarında olmalıdır (Airasian, 2001).

Portfolyonun değerlendirilmesinde temel olarak rubrikler ya da derecelendirme ölçekleri kullanılmaktadır. Detaylı, somut maddelerden oluşan ve değerlendirilecek performansa uygun maddelerin yer aldığı bir rubrik ya da derecelendirme ölçeği yapılan puanlamanın geçerliğini ve güvenirliğini artıracaktır. Portfolyoda bulunan çalışma örnekleri, genellikle performans değerlendirmelerinden ortaya çıkan karmaşık performanslar ya da ürünlerden oluşmaktadır. Bu nedenle puanlama araçlarının somut ve değerlendirilecek performansa uygun gözlemlenebilir maddelerden oluşması önemlidir. Portfolyo değerlendirmede, hem dosya içerisindeki çalışma örnekleri tek tek değerlendirilir hem de portfolyonun tamamına dönük bir değerlendirme yapılır. Genellikle tek tek çalışma örneklerini değerlendirmede analitik rubrikler ya da derecelendirme ölçekleri kullanılırken, portfolyonun bütününü değerlendirmede bütüncül bir rubriğin ya da derecelendirme ölçeğinin kullanımı yaygındır. Elbette bu karar değerlendirilecek olan çalışmanın yapısına ve değerlendirme kriterlerine göre değişebilmektedir. Dosya da ayrıca öğrenciye ait öz değerlendirmeler, öz yansıtma formları da değerlendirilecek diğer ögelerdir (Airasian, 2001; Mertler, 2001). Portfolyoların değerlendirilmesi aşamasının zaman alıcı ve zahmetli olması bu işlemlerden kaynaklanmaktadır.

4.3.7. Portfolyonun Başlıca Üstün Yönleri

Portfolyo kullanımının eğitim alanına getirdiği pek çok üstün yönü bulunmaktadır. Bunlardan başlıcaları aşağıda özetlenmiştir.

1- Öğrenci merkezlidir. Öğrenciler sürece aktif olarak katılır ve sorumluluk alırlar. Özellikle portfolyoya seçilecek etkinliklerin, projelerin, ürünlerin ya da çalışma örneklerinin seçiminde öğretmenle birlikte görev alırlar. Değerlendirme aşamasında da yine performans ölçütlerinin belirlenmesi, rubriklerin ya da derecelendirme ölçeklerinin geliştirilmesi aşamasında doğrudan ve önemli ölçüde katkı vermeleri beklenir. Yine öğrenciler kendi gelişimini değerlendirme ve çalışmalarını istenilen ölçütlere uygun hale getirmek amaçlı gözden geçirme konusunda da sorumluluk üstlenir, kendi çalışmalarına ilişkin öz değerlendirmeler, yansıtmalar yaparlar.

2- Öğrencinin o dersteki gelişiminin dönem boyunca nasıl bir seyir izlediği konusunda öğrenciye, öğretmene ve velilere bilgi sunar ve benzer şekilde sürecin gözlemlenmesine olanak sağlar. Portfolyolar öğrenci gelişiminin izlenmesine olanak tanır. Bu portfolyo içerisindeki çalışma örneklerinin tamamlanmış hâli dışında taslak/ilk hallerinin de kapsanmasıyla mümkün olacağı gibi öğrencinin o becerilere ilişkin gösterdiği performans konusunda dönemin başından sonuna kadar nasıl bir gelişim izlediği incelenerek de mümkün kılınır.

3- Öğrenciye kendi öğrenme süreci, çalışmaları hakkında deneyimlerini, düşüncelerini, duygularını ve değerlendirmelerini ifade etme imkânı sağlar. Öğrenci ne öğrendiğini, nasıl öğrendiğini, öğrenme sürecinde hangi güçlü/zayıf yönleri olduğu üzerinde düşünme imkânı bulur ve bu yönlerini fark ederek geliştirir. Bu öğrencinin, öz değerlendirme/yansıtma yapmasını ve öğrenmesine daha farkındalıklı bir şekilde yaklaşmasını sağlar.

4- Portfolyo diğer değerlendirme araçlarının aksine öğrencinin zayıf yanlarına değil güçlü yanlarına daha fazla odaklanır. En iyi çalışmalarını seçmesi istenir. Öğrenciye başarı duygusunu yaşatır.

5- Öğrenci, öğretmen ve veli iletişimine ve etkileşimine olanak sağlar. Veli çocuğunun neleri bildiği ve yapabileceği konusunda somut bilgiye ulaşır. Bu durum velinin gerek öğrenci gerekse öğretmenle olan diyaloğunu somut bilgiler çerçevesinde sürdürmesine olanak sağlar.

6- Öğrencinin o derse/konu alanına ilişkin öğrenme çıktılarını ne derece gerçekleştirdiğine ne derece yeterli olduğuna dair yeterli sayıda ve çeşitte dokümanları bir arada sunar (örneğin; yaptığı sunumlar, hazırladığı sergiler, projeler, problem çözümüne yönelik yazılı çözümler, günlükler**,** fotoğraflar, ses kasetleri ve video kayıtları, kısa sınavlar, gösteriler, köşe yazıları vb.). Bunların her biri öğrenci yeterliğine ilişkin kanıt niteliğinde dokümanlardır. Bu durum, öğretmenin öğrenci yeterliğine dair vereceği kararların daha geçerli olmasına katkı sağlar (Johnson & Johnson, 2002; Mertler, 2001).

7- Portfolyolar "otantik" (gerçek ve anlamlı) bir değerlendirme olanağı sağlar (Mertler, 2001). Diğer bir deyişle, sınıf içinde edinilen bilgi ve becerilerin gerçek yaşam koşullarına uyarlanmasını, uygulanmasını sağlar.

4.3.8. Portfolyonun Bazı Sınırlı Yönleri

Tüm ölçme ve değerlendirme araçlarında olduğu gibi portfolyo kullanımının da bazı sınırlılıkları mevcuttur. Bunlardan öne çıkanlar kısaca;

1- Portfolyoya dayalı değerlendirme sistemi zaman alıcı ve maliyetlidir. Özellikle öğretmenin portfolyoları gözden geçirmesi, değerlendirmesi aşaması oldukça zaman alıcıdır. Bunun ana nedeni, bu aşamada öğretmen hem tek tek dosya içindeki her bir çalışma örneğini değerlendirmek, hem de dosyanın tümüne ilişkin değerlendirme yapmak durumundadır. Bu da zaman alıcıdır ve oldukça emek harcamayı gerektirmektedir. Öğretmenin süreç içerisinde düzenli aralıklarla her bir öğrenciyle bireysel olarak bir araya gelmesi ve dosyaya seçilecek ve dosyadan çıkarılacak çalışmalara öğrenciyle birlikte karar vermesi aşaması da özellikle kalabalık sınıflarda yine zaman alıcıdır. Burada çalışma örneklerinin seçimi, bu çalışmalar halihazırda ders sürecinde zaten yapılan çalışmalar olduğu için değerlendirme aşamasına kıyasla daha az zaman gerektirebilmektedir.

2- Lojistik sorunlar (depolama, düzenleme). Portfolyoların sınıflarda muhafaza edilmeleri alan sorununu doğurmaktadır. Portfolyo dosyaları sınıflarda muhafaza edilirler ve istenildiği zaman o öğrencinin ve öğretmenin ulaşımına açıktır. Ancak, portfolyo dosyalarına giren çalışma örneklerinin çok çeşitli olduğu dikkate alındığında bir yer sorunu doğmaktadır. Montgomery (2001), materyallerin saklanması için çeşitli klasörler, albümler, ayakkabı kutuları, plastik kasalar, dosya dolapları, bilgisayar diskleri ve kompakt disklerde dâhil çeşitli seçeneklerin kullanılabileceğini ifade etmiştir. Her öğrenciye ait ayrı bir kutu, klasör veya bölme olması önemlidir. Bu öğrencilerin kutuları kişiselleştirebilmesine olanak sağlar (Gallagher, 1998). Elektronik portfolyoların kullanımı bu sorunu çözme anlamında önemli katkı sağlayacaktır.

3- Öğretmenler, sınıf içi öğretim etkinliklerini bu değerlendirme anlayışına uygun olacak şekilde yeniden yapılandırmak zorunda kalabilirler. Bu durum ders kitabı temelli etkinliklerden çok uygulamalı ve problem çözmeye dayalı çalışmalara yönelmeyi gerektirir. Bu nedenle ders içeriği ya da öğrenme çıktıları portfolyo kullanımına uygun olan derslerde kullanılması daha uygun olacaktır. Örneğin fen bilgisi, yazma dersleri gibi.

4- Standart bir puanlama sağlanması güçtür. Portfolyo kullanımına dair önemli bir sınırlılık, standart bir puanlama sağlanmasındaki güçlüktür (Belgrad, 2013; Marzano, 2017). Bu nedenle bazı yazarlar, portfolyo dosyalarının bütününe ilişkin bir puanlama yapılmamasını önermektedir (Marzano, 2017). Bu yazarlar, portfolyonun öğrenci başarısının kanıtlarını gösteren bir derleme, koleksiyon olarak ele alınmasının daha doğru olduğunu, çünkü, portfolyoyu oluşturan materyallerin, ürünlerin zaten hali

hazırda puanlanmış materyaller olduğunu yani ikinci puanlamaya gerek olmadığını vurgulamıştır (örneğin: sunular, kompozisyonlar, projeler, panolar vb.). Bunun yerine, öğretmen portfolyoda toplanmış olan çalışma örneklerini etkileyici bulabilir ve öğrencinin dersteki toplam puanını arttırma yoluna gidebilir.

5- Portfolyo puanlarının geçerlik ve güvenirliğini sağlamak testlere kıyasla daha karmaşık ve güçtür. Bu nedenle, portfolyolarda bir güvenirlik indeksi hesaplamaktan ziyade, güvenirliği artırıcı tedbirleri almak yaygın bir uygulamadır ve özellikle öğretmenler açısından daha önemlidir. Aynı durum yapılan ölçümlerin geçerliğini sağlama konusunda da aynı bakış açısını gerektirmektedir.

4.3.9. Portfolyo Geliştirme Aşamaları

Aşağıda portfolyo geliştirirken izlenecek adımlar kısaca özetlenmiştir:

1.Adım: İlk adım portfolyonun temel amacının ne olduğuna karar vermektir. Portfolyo öğrenci gelişimini belgelemek ya da öğrencinin en iyi çalışmalarını sergilemek amacıyla hazırlanabilir. Amacın ne olduğunu netleştirmek sonraki adımları doğrudan etkileyeceği için oldukça önemlidir. Örneğin, hangi çalışmaların seçileceği bu karara bağlıdır.

2. Adım: Portfolyo diğer değerlendirme araçları gibi o derse ait öğrenme çıktılarına, kazanımlara uygun olmalıdır. Dolayısıyla, bu adımda portfolyo ile hangi içerik, kavram ve becerilerin değerlendirileceği netleştirilmelidir. Diğer bir deyişle değerlendirilecek içerik, kavram ve beceriler doğrudan o dersin öğretim hedefleri ve amaçlarından türetilmelidir.

3. Adım: Öğrencilerin hangi tür ürün ve etkinlikleri hazırlayacağı kararlaştırılmalıdır. Seçilen çalışma ya da ürünler ikinci adımda belirlenen içerik, kavram ve becerilerin değerlendirilmesine imkân sağlayan çalışmalar olmalıdır. Bu noktada, kararlar çok katı olmamalı; öğrencilerin kendi ürünlerini geliştirmelerine ve seçmelerine esneklik tanınmalıdır. Ancak, öğrencilerin bu çalışmaları hangi çerçevede yapacağına dair temel noktalar belirlenmelidir. Ayrıca lojistik konular (örneğin portfolyoların nasıl ve nerede saklanacağı, düzenleneceği) bu aşamada netleştirilmelidir.

4. Adım: Etkili bir değerlendirme süreci planlanmalıdır. Daha önce değinildiği üzere, öğrencilerle düzenli olarak buluşma ve dosyaya eklenecek ya da dosyadan çıkarılacak çalışmalar üzerinde tartışma süreci özellikle kalabalık sınıflarda zaman almaktadır. Bu nedenle daha sonra zaman sorunu yaşamamak için öğretmenin dönemin başında uygulanabilir, etkili bir plan geliştirmesi büyük önem taşımaktadır.

5. Adım: Öğretmen, öğrencilerin sürece nasıl ve ne zaman dâhil edileceğini belirlemelidir. Daha önce değinildiği üzere, öğrenci özellikle dosyaya hangi çalışma örneklerinin gireceği konusunda ve performans kriterlerini belirleme, rubrikleri ya da derecelendirme ölçeklerini oluşturma aşamalarında sürece dâhil olmalıdır. Öğrencilerin katılım düzeyi, portfolyonun işlevini ve etkililiğini doğrudan etkilemektedir.

6. Adım: Değerlendirme rubriklerinin ya da derecelendirme ölçeklerinin geliştirilmesi. Portfolyoların amacına ve içeriğine uygun şekilde analitik ya da bütüncül rubrikler ya da derecelendirme ölçekleri hazırlanmalıdır. Portfolyo kullanım amacımız, bu yolla öğrencilerin dersin öğrenme çıktılarını gerçekleştirme durumlarını saptamak olduğu için rubrikler ya da derecelendirme ölçekleri hazırlanırken dersin öğrenme çıktılarını yansıtacak şekilde, somut, gözlemlenebilir performans maddeleri şeklinde hazırlanmalıdır. Maddeler o beceri ya da davranışları kapsayıcı olmalıdır.

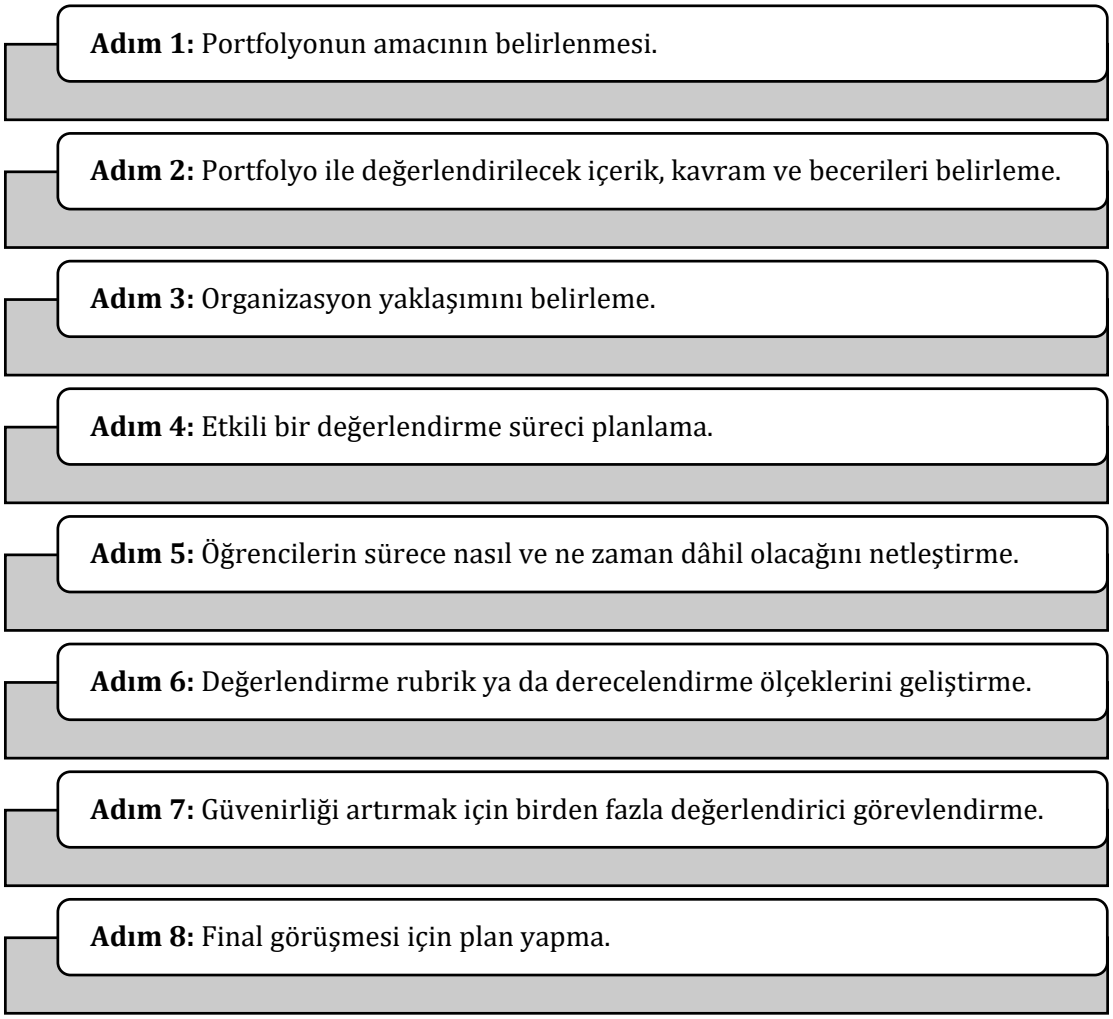
7. Adım: Puanlama güvenirliğini artırmak için mümkünse birden fazla değerlendirici sağlanmalıdır. Portfolyoların puanlanması genellikle klasik ölçme araçlarına kıyasla daha karmaşık ve güçtür. Portfolyo değerlendirmeleri için resmî bir güvenirlik standardı olmasa da birden fazla değerlendiricinin sürece dâhil olması, değerlendirmelerin tutarlılığını artırır ve puanlamaların nesnelliğine katkıda bulunur. Bir önceki adımda hazırlanan puanlama araçlarının detaylı, açık ve anlaşılır olması puanlayıcılar arası tutarlılığı arttırma anlamında belki de en önemli katkıyı sağlayacaktır.

8. Adım: Öğrencilerle dönem boyunca ya da yıl sonunda yapılacak olan son görüşme için plan yapılmalıdır. Dönem veya yıl sonunda yapılacak final görüşmeler, öğrencilere kendi gelişimlerini ve nihai başarılarını gözden geçirme fırsatı sunacaktır. Ancak bu görüşmeler daha önce değinildiği üzere zaman alıcıdır, bu yüzden öğretmenin önceden planlama yapması gerekir (Mertler, 2001).

Aşağıda portfolyo oluştururken izlenecek aşamalar kısaca şematik olarak sunulmuştur:

Şekil 1

Portfolyo Oluştururken İzlenecek Aşamalar



4.3.10. Portfolyo Değerlendirmelerinde Geçerlik ve Güvenirlik

4.3.10.1. Geçerlik

Portfolyo değerlendirmelerinin geçerliğine yönelik alınacak tedbirlerin neler olduğuna karar verirken aşağıdaki bazı sorulara cevap aramak faydalı olacaktır:

1. İçerik öğrenme hedefleriyle, dersin öğrenme çıktılarıyla uyumlu mu?
2. Portfolyo içindeki çalışma örnekleri öğrencinin gerçek becerilerini yansıtıyor mu?
3. Dersin öğrenme çıktılarına ilişkin farklı görev ve ürünler portfolyoda yer alıyor mu?
4. Sonuçlar öğrenci, öğretmen ve veli tarafından anlaşılabilir mi?

Gredler'e (1999) göre portfolyoların geçerliğini sağlamada üç önemli husus bulunmaktadır: Öğrenci yeterlilikleri, konu alanı temsili ve rubrik/derecelendirme ölçekleri -örnek uyumu.

1. **Öğrenci yeterlikleri (ölçülecek beceriler, davranışlar):** Öğrenciden beklenen yeterliliklerin neler olduğuna öncelikli olarak karar verilmelidir. Bunu yaparken dersin ya da konu alanının öğrenme çıktıları dikkate alınmalıdır. Bu öğrenme çıktılarıyla uyumlu olarak becerilere, yeterliklere karar verilmelidir. Portfolyoya eklenecek çalışma örneklerine ise bu yeterlikler doğrultusunda karar verilmelidir. Portfolyoya seçilen çalışma örnekleri öğrencinin becerilerini ve öğrenme çıktılarını yansıtır nitelikte olmalıdır.

2. **Konu alanı temsili (derse/konu alanına ilişkin öğrenme çıktıları, öğrenme hedefleri)** Portfolyoya seçilen çalışma örnekleri, materyaller, o dersin ya da konu alanının hedeflerini ve içeriğini doğru ve yeterli bir şekilde temsil etmelidir. Burada bahsedilen özellik, bir çoktan seçmeli testten beklenen kapsam geçerliğine benzetilebilir. Eğer çalışma örnekleri yalnızca sınırlı bir yeterliğe veya beceriye ilişkin olursa, öğrencinin genel gelişimi ve başarısı hakkında yanıltıcı kararlar kaçınılmaz olacaktır (kapsam geçerliği düşük olan bir çoktan seçmeli ya da yazılı yoklama sınavı sonuçlarının neden olacağı yanlış kararlar gibi).

   Bu bağlamda, portfolyoya eklenecek çalışma örneklerinin, materyallerin hangi öğrenci becerilerini temsil edeceği önceden netleştirilmelidir. Önce becerilere karar verilmeli daha sonra bu becerilere karşılık gelen çalışmaların neler olacağı netleştirilmelidir. Daha önce değinildiği üzere portfolyoda ses kasetleri, fotoğraflar, video kayıtları, günlüklerden projelere kadar farklı çeşitlilikte çalışma ya da materyal kapsanarak dersin farklı öğrenme çıktılarına göre öğrencinin farklı yönleri ortaya konulabilmektedir.

3. **Rubrik/derecelendirme ölçekleri-örnek çalışmaların uyumu (değerlendirme kriteri ile değerlendirilen ürün arasındaki uyum):** Kullanılan değerlendirme rubriği ya da derecelendirme ölçekleri ile portfolyodaki gerçek ürünler arasında uyum bulunmalıdır. Rubrikler ya da derecelendirme ölçekleri dersin öğrenme çıktıları dikkate alınarak hazırlanmalıdır ve puanlanması amaçlanan becerileri ya da davranışları gözlemlemeye uygun olmalıdır. Her bir çalışma örneğiyle temsil edilen beceriler ya da yeterlikler farklılık gösterebilir. Hazırlanan rubrik ya da derecelendirme ölçekleri o çalışmayı değerlendirmeye uygun nitelikte olmalıdır. Örneğin, görsel bir performansı değerlendirirken kullanılan rubrik ile bir öykü denemesini değerlendirirken kullanılan rubrik ve bu rubrikteki maddeler farklı olacaktır. Portfolyo değerlendirilirken kullanılan rubrik (puanlama ölçütleri), portfolyodaki içerikle doğrudan uyumlu olmalıdır. Genel olarak konuşursak, kullanılan puanlama aracındaki maddeler ile puanlanacak davranışların, becerilerin

örtüşmesi gerekmektedir ya da değerlendirme kriteri ile değerlendirilen ürün arasında uyum olmalıdır. Aksi bir durum geçerliği olmayan sonuçlara yol açacaktır. Dikkatli ve detaylı hazırlanmış bir rubrik ya da derecelendirme ölçeği yapılan değerlendirmenin nesnelliğini de artıracaktır.

4.3.10.2. Güvenirlik

Benzer şekilde, portfolyo değerlendirmelerinin güvenirliğine yönelik alınacak tedbirlerin neler olduğuna karar verirken aşağıdaki bazı sorulara cevap aramak faydalı olacaktır:

1- Aynı portfolyoyu farklı öğretmenler değerlendirirse benzer sonuçlar çıkar mı?

2- Aynı öğretmen portfolyoyu tekrar değerlendirildiğinde sonuçlar tutarlı çıkar mı? Zaman içinde tutarlılık korunuyor mu?

3- Ölçütler önceden öğrenciyle paylaşıldı mı?

4- Değerlendirme, öğretmenin kişisel yargılarından bağımsız mı? Nesnel ölçütler kullanıldı mı? (Bu soru aynı şekilde geçerlik için de sorulabilir.)

Portfolyo değerlendirmelerinde güvenirlik genellikle puanlayıcılar arası tutarlılık ile açıklanmaktadır (Gredler, 1999; Mertler, 2001). Güvenirlik, birden fazla puanlayıcının aynı öğrenci çalışmalarına benzer puanlar verip vermemesiyle ilgilidir. Farklı puanlayıcıların aynı portfolyoyu benzer şekilde puanlamaları yapılan puanlamanın güvenirliğine işaret etmektedir. Ancak öğretmenin puanlama için başka bir öğretmene ulaşması pratikte oldukça güçtür. Mertler (2001) bunun için bir yol önermiştir. Buna göre öğretmen, başka öğretmenlerden öğrencilerin portfolyolarından örnekleri incelemelerini isteyebilir. Bu öğretmenlerin verdiği puanlar, sınıf öğretmeninin puanlarıyla karşılaştırılarak tutarlılık ölçülebilir.

Ancak bunun mümkün olmadığı durumlarda (ki bu durum özellikle ülkemizde sınıf mevcutları dolayısıyla yaygındır), öğretmen puanlamada oldukça dikkatli olmalıdır. Öğretmenin puanlama güvenirliğini arttırmak için alacağı en önemli tedbirlerden biri rubriklerin ya da derecelendirme ölçeklerinin hazırlanma aşamasıdır. Hazırlanan rubriklerin değerlendirilecek çalışmaya ilişkin detayları içermesi ve açık ve anlaşılır şekilde hazırlanması önemlidir. Örneğin, rubrikte ya da derecelendirme ölçeğinde söz konusu çalışmanın hangi boyutlarının ve bu boyutlar altındaki hangi detayların puanlanacağına dikkatle karar verilmelidir. Bunun dışında puanlanacak çalışmanın yalnızca ürün boyutu mu puanlanacak yoksa sürece ilişkin detaylarda değerlendirmeye tabi tutulacak mı? buna karar verilmelidir. Yine çok önemli bir nokta ise rubrikte ya da derecelendirme ölçeğinde yer alan maddelerin gözlemlenebilir davranışlara/özelliklere yönelik olması gereğidir. Bu rubriğin başka bir öğretmen tarafından kullanılacağını varsayalım, bu durumda öğretmen maddeleri okuduğunda çalışmada neye bakacağı konusunda tereddüt yaşamamalıdır. Buda yapılan puanlamanın nesnelliğini arttırıcı bir tedbirdir.

5. PERFORMANS DEĞERLENDİRMEDE KULLANILAN PUANLAMA ARAÇLARI

Performans değerlendirmede kullanılan puanlama araçları yaygın olarak rubrikler ve kontrol listeleri olmak üzere iki kategoride ele alınmaktadır. Ancak rubriklere benzeyen fakat rubriklerde olduğu üzere performansın düzeylerini betimsel olarak ayrıntılı açıklamayan bir grup puanla araçları vardır ve bunların rubriklerden ayrı tutulması daha sağlıklı olacaktır. Bu nedenle, burada performans değerlendirmede kullanılan puanlama araçları üç kategoride ele alınacaktır:

1. Rubrikler (Dereceli Puanlama Anahtarı olarak adlandırılması yaygındır),
2. Derecelendirme Ölçeği ya da Kriter Performans Listesi (Rating Scale ya da Criterion Performance List),
3. Kontrol listeleri (Checklist)

Performansın değerlendirilmesinde kullanılan temel puanlama araçlarından biri rubriklerdir. Rubrikler ülkemizde "Dereceli Puanlama Anahtarı" adıyla da sıklıkla kullanılmaktadır. Ancak, yukarıda değinildiği üzere, rubrikler dışında olup rubriklere benzerliğiyle bilinen ve sıklıkla rubriklerle karıştırılan ve farklı isimlerle karşımıza çıkan puanlama araçları mevcuttur. Bunlar genellikle alanyazında derecelendirme ölçekleri ya da kriter performans listesi olarak adlandırılmaktadır (Brookhart, 2013; McTighe ve ark., 2020). Son olarak, tanımı konusunda büyük ölçüde fikir birliği sağlanmış olunan ve basit bir yapıya sahip olması açısından karışıklığa yer vermeyen kontrol listeleri vardır.

Hibbard ve arkadaşları (1996) kriter performans listesini, öğrenci performansını değerlendirmek için kullanılan temel ölçütlerin basit bir listesinden oluşan temel ve pratik bir araç olarak tanımlamıştır. McTighe ve arkadaşları (2020, s.73), kriter performans listelerinin, ölçütler veya temel özellikler/öğeler kümesinden, geribildirim ve yansıtma (reflection) için ayrılmış bir alandan ve bazı durumlarda öğretmenlerin çeşitli rubrik (dereceli puanlama anahtarı) öğelerini değerlendirmek için puanlar atamasına olanak tanıyan bir derecelendirme ölçeğinden oluştuğunu ifade etmiştir. İlköğretim düzeyinde yazma dersi için hazırlanmış bir kriter performans listesine aşağıda yer verilmiştir.

Tablo 5

İlköğretim Düzeyi Yazma Dersine Yönelik Örnek Kriter Performans Listesi

Kriter Performans Listesi	😀 Harika	☺ Neredeyse oldu	😐 Geliştirilmeli
1. Hikâyem için ilgi çekici bir mekân ve karakterlerim var.			
2. Hikâyemdeki problem okuyucularım için açık ve anlaşılır.			
3. Hikâyemdeki olaylar doğru sırayla verilmiştir.			
4. Çözüm, okuyucularım için açık ve anlaşılır olacaktır.			
5. Olanları anlatmak için birçok betimleyici kelime kullandım.			
6. Kelimelerim "zihinde bir resim oluşturur."			
7. Hikâyemle uyumlu bir başlığım var.			
Bir dahaki sefere hikâye yazarken neyi daha iyi yapmaya çalışacaksın? …… ……			

Kaynak: (McTighe ve ark., 2020, s. 75).

Bu bölümde benzer özelliklerinden dolayı sadeleştirmek adına "Derecelendirme Ölçekleri ya da Kriter Performans Listesi" olarak adlandırılan puanlama araçları kısaca "Derecelendirme Ölçekleri" olarak ifade edilmiştir. Ülkemizde de bu araçların isimlendirilmelerine ilişkin farklı kullanımlar mevcut olup, henüz tekbir isimlendirme üzerinde fikir birliği tam olarak mevcut değildir. Ancak Derecelendirme Ölçekleri/Kriter Performans Listesi olarak anılan araçlar rubrik değildir. Bu iki puanlama aracı temel özellikleri ve kullanıldıkları durumlar açısından ayrışmaktadırlar. Aşağıda, bu araçlar ayrıntılı olarak sunulmaktadır.

Alanyazında ve uygulamada sıklıkla rastlanan bir diğer kavram yanılgısı ise performans değerlendirmede kullanılan "puanlama araçlarının" birer "ölçme aracı" olduğunun düşünülmesidir. Oysa puanlama araçları bir ölçme aracı değildir. Aksine, herhangi bir ölçme aracının puanlama bölümünü oluştururlar ve bazı durumlarda birer puanlama anahtarıdırlar. Bu araçlar performans görevi, proje, portfolyo, gözlem benzeri performansı ölçme ve değerlendirmede kullanılan araçların puanlama bölümlerini oluştururlar (Bkz. Çetin & Çakan, 2010; Kutlu ve ark., 2010).

Aşağıda üç temel puanlama aracı olan rubrik, derecelendirme ölçekleri/kriter performans listesi ve kontrol listesi sırasıyla sunulmaktadır.

5.1. Rubrikler

Rubrik, öğrencilerin çalışmalarını değerlendirmek için kullanılan, performans düzeylerinin tanımlarıyla birlikte verilen tutarlı ölçütler bütünü olarak tanımlanmaktadır (Brookhart, 2013). Burada rubrik tanımına ilişkin olarak en dikkat çekici özellik "performans düzeylerinin tanımlarıyla birlikte verilen" vurgusudur, çünkü bu özellik rubrikleri derecelendirme ölçeklerinden ve kontrol listelerinden ayıran en ayırt edici özelliktir. Bilindiği üzere, rubrikler sıklıkla derecelendirme ölçekleri ile karıştırılmaktadır. Rubriklerin aksine derecelendirme ölçekleri ve kontrol listeleri, performansın kalitesine dair açıklamalar içermezler (Brookhart, 2013). Moskal ve Leydens (2000) rubriklerin derecelendirme ölçeklerinden türediğini ancak ölçütlerin anlaşılabilirliği bakımından farklılaştığını belirtmiştir. Brookhart (2013) rubrikler ile derecelendirme ölçekleri ve kontrol listeleri arasındaki farkı şu şekilde açıklamıştır:

1. Rubriklerin aksine, derecelendirme ölçeği sıklık ya da kaliteyi derecelendirir ancak ayrıntılı açıklamalar içermez. Kontrol listeleri de benzer şekilde performansın kalitesine ilişkin ayrıntı vermez yalnızca varlık ya da yokluğuna dair bilgi verir.
2. Rubrikler en belirgin anlamda öğrencilerin çalışmalarına ya da performanslarına ilişkin ölçütler ve performans düzeylerinin açıklamalarıyla tanımlanırlar. Performans düzeyinin tanımlanması önemlidir. Derecelendirme ölçekleri ve kontrol listeleri bu iki parçadan birini ya da her ikisini içermediği için rubrik olarak tanımlanamazlar.
3. Kontrol listeleri ve derecelendirme ölçekleri belirli ölçütlere dayanır. Bu ölçütler, kontrol listesi için öğrencinin bir işi yapıp yapmadığını, derecelendirme ölçeği için ise ne kadar iyi ve hangi sıklıkta yaptığını göstermeye yöneliktir. Performansın niteliğine dair ayrıntılı açıklamaya gerek duyulmadığı durumlarda rubrik yerine derecelendirme ölçekleri ve kontrol listelerini kullanmak çok daha pratik ve işlevseldir.

Rubrikler gözlemlenen davranışlarda sergilenen yanıtlar ya da davranışlar nitelik açısından ele alındığı durumlarda kullanıldığında oldukça faydalıdır ve anlamlıdır. Rubriklerde, becerinin yerine getirilme düzeyleri derecelendirme ölçeklerinde olduğu üzere sayısal olarak (5-4-3-2-1 gibi) nitelendirilmekten öte betimsel olarak tanımlanır ve detaylı bir açıklaması yapılır.

Rubrikler kendi içinde bütüncül ve analitik olarak ikiye ayrılmaktadır. Bütüncül rubriklerde söz konusu performans ya da performansa ilişkin çalışma örneği her bir boyut için tek tek ele alarak değerlendirilmez. Öğretmen çalışmayı bir bütün olarak ele alır ve öğrenci başarısına dair tek bir puan üretir. Çalışmanın tamamına ilişkin kalitenin daha önemli olduğu durumlarda ve öğrenciye geribildirim verme amacının olmadığı durumlarda bütüncül rubrik tercih edilebilir. Çalışmanın geneli değerlendirildiği için daha pratiktir ve daha az zaman alır. Özellikle kalabalık sınıflarda analitik rubriğe göre daha kullanışlıdır. Öğrencinin güçlü ve zayıf yanlarına ilişkin detaylı bilgi vermez. O nedenle öğrenciye detaylı bir geribildirim verilmesi amaçlanıyorsa analitik rubriğin kullanılması daha faydalı olacaktır. Aşağıda kompozisyon yazma becerisini değerlendirmede kullanılabilecek bir bütüncül rubrik örneği sunulmuştur:

Tablo 6

Kompozisyon Yazma Becerilerine Yönelik Bütüncül Rubrik Örneği

Bütüncül Rubrik **Öğrenci Adı ve Soyadı:** **Sınıf:** **Öğrenci Numarası:** **Tarih:**	
Yeterli (3)	Yazıda, konuya uygun ana fikir ve yardımcı fikirlere yer verilmiş, fikirler örnek ya da açıklamalarla desteklenmiş, anlatım tutarlı, yazıda özgün ya da yaratıcı bir bakış acısı mevcut, akıcı ve anlaşılır bir dil kullanılmış, yazım ve dil bilgisi kurallarına tümüyle uyulmuş
Geliştirilmesi Gerekli (2)	Yazıda konuya uygun ana fikir ve yardımcı fikirlere kısmen yer verilmiş, fikirler örnek ya da açıklamalarla kısmen desteklenmiş, anlatım çoğunlukla tutarlı, yazıda özgün ya da yaratıcı bir bakış acısı kısmen mevcut, dil genellikle akıcı ve anlaşılır, yazım ve dil bilgisi kurallarına genellikle uyulmuş
Yetersiz (1)	Yazıda konuya uygun ana fikir ve yardımcı fikirlere yer verilmemiş, fikirler örnekler ya da açıklamalarla desteklenmemiş, anlatımda çok sayıda tutarsızlık mevcut, yazıda özgün ya da yaratıcı bir bakış acısı mevcut değil, akıcı ve anlaşılır bir dil kullanılmamış, yazım ve dil bilgisi kurallarına çoğunlukla uyulmamış
TOPLAM PUAN:	

Yukarıda görüldüğü üzere bütüncül rubriklerde değerlendirilmek istenen performansa ilişkin alt boyutlara yer verilmez. Genel olarak performansın ya da ortaya konulmuş olan ürünün bütününe ilişkin bir değerlendirme yapılır.

Analitik rubriklerde ise bütüncül rubriklerin aksine, puanlayıcı öğrenci performansını veya ürününü birden fazla boyut için ayrı ayrı ele alarak istenen ölçütler doğrultusunda puanlar, değerlendirir. Analitik rubriklerde performans belirli boyutlara (örneğin içerik,

anlatım, özgünlük, dil bilgisi vb.) ayrıştırılır ve çalışma örneği her bir boyut için ayrı ayrı değerlendirilir. Örneğin, dört ayrı boyut varsa (içerik, anlatım, özgünlük, dil bilgisi) çalışma tek tek her bir boyut için olmak üzere toplamda dört kez değerlendirilir. Öğrenci başarısı bu boyutların her birinden alınan puanların toplamıyla belirlenir. Analitik rubrikler özellikle öğrenciye geribildirim verilmesi amaçlandığında tercih edilmelidir. Kullanılacak rubriğinin türüne karar vermede önemli olan diğer bir etken de değerlendirilen performansın ya da çalışmanın yapısıdır. Performans çok boyutlu bir yapıya sahip ise analitik rubrik tercih edilmelidir. Yine öğrenci gelişimi takip edilmek istendiğinde ayrıntılı geribildirim sağlaması nedeniyle analitik rubrikler tercih edilmelidir. Analitik rubrik kullanımı, yapılan puanlamanın tutarlılığını, duyarlılığını ve nesnelliğini de artırmada faydalı olacaktır. Yine, performansın bütününden ziyade alt boyutlardaki başarının önemli olduğu durumlarda analitik rubriğin kullanılması önemlidir. Aşağıda kompozisyon yazma becerisine ilişkin analitik rubrik örneği sunulmuştur:

Tablo 7

Kompozisyon Yazma Becerilerine Yönelik Analitik Rubrik Örneği

Analitik Rubrik **Öğrenci Adı ve Soyadı:** **Öğrenci Numarası:**			**Sınıf:** **Tarih:**	**Puan**
	Yeterli (3)	Geliştirilmesi Gerekli (2)	Yetersiz (1)	
İçerik	Konuya uygun, dikkat çekici bir başlık kullanılmış; ana fikir ve yardımcı fikirler açıkça belirtilmiş ve örneklerle desteklenmiş; giriş, gelişme ve sonuç bölümleri arasında tutarlılık mevcut.	Konuya uygun bir başlık kullanılmış; ana fikir ve yardımcı fikirler kısmen belirtilmiş ve kısmen örneklerle desteklenmiş; giriş, gelişme ve sonuç bölümleri arasında kısmen tutarlılık mevcut.	Başlık konuya uygun değil; ana fikir ve yardımcı fikirler çoğunlukla belirsiz; giriş, gelişme ve sonuç bölümleri arasında tutarlılık yok; gereksiz bilgiler mevcut.	
Dil ve Anlatım	Açık, anlaşılır ve akıcı bir dil kullanılmış; konuya uygun, doğru ve yeterli çeşitlilikte kelime kullanılmış; giriş, gelişme ve sonuç boyunca anlam bütünlüğü mevcut; noktalama işaretleri ve yazım kurallarına tümüyle uyulmuş.	Çoğunlukla açık, anlaşılır ve akıcı bir dil kullanılmış; kısmen konuya uygun, doğru ve yeterli çeşitlilikte kelime kullanılmış; giriş, gelişme ve sonuç boyunca kısmen anlam bütünlüğü sağlanmış; noktalama işaretleri ve yazım kurallarına genellikle uyulmuş.	Dil açık, anlaşılır ve akıcı değil; kullanılan kelimeler çoğunlukla konuya uygun değil ve yetersiz çeşitlilikte; giriş, gelişme ve sonuç arasında anlam bütünlüğü sağlanmamış; noktalama işaretleri ve yazım kurallarına çoğunlukla uyulmamış.	
Yaratıcılık ve Özgünlük	Yazıda özgün sayılabilecek fikir ya da farklı bakış açısı sunulmuş; yaratıcı fikirlere, özgün örnek ya da benzetmelere yer verilmiş.	Yazıda kısmen özgün sayılabilecek fikir ya da farklı bakış açısına yer verilmiş; kullanılan örnek ya da benzetmeler kısmen yaratıcı.	Yazıda ifade edilen fikir ya da bakış açısı özgün değil; örnek ya da benzetmelerde yaratıcılık yok.	
Sayfa Düzeni	Yazı düzenli ve okunaklı, sayfa temiz kullanılmış, kenar boşlukları doğru kullanılmış.	Yazı kısmen düzenli ve okunaklı, sayfa kısmen temiz kullanılmış, kenar boşlukları kısmen doğru kullanılmış.	Yazı düzenli ve okunaklı değil, sayfa temiz kullanılmamış, kenar boşlukları doğru kullanılmamış.	

Analitik rubrik örneğinde görüldüğü üzere, kompozisyon becerisi dört ayrı boyutta, üç düzeyli olarak puanlanmıştır. Burada öğrenciye yönelik geribildirim verilmek istendiğinde, öğrenciye her bir boyuttaki güçlü, zayıf ve geliştirilmesi gereken yönleri ayrıntılı olarak sunulabilmektedir.

5.2. Derecelendirme Ölçekleri

Derecelendirme ölçekleri, söz konusu çalışmaya ya da performansa ilişkin olarak belirlenmiş olan özelliklerin ve bu özelliklerin her birinin hangi düzeyde sergilendiğini işaretlemek için ayrılmış bölümlerden oluşan bir listedir (Brookhart, 2013). Derecelendirme ölçekleri de kontrol listeleri gibi bir listedir, ancak burada özelliğin hangi düzeyde gerçekleştirildiğini işaretlemeye yarayan bir derecelendirme vardır. Bu derecelendirme genellikle 4 ya da 5'li olabildiği gibi 3'lü olarak da kullanılabilmektedir. Bu yanıyla Brookhart (2013) derecelendirme ölçeklerinin iki boyutlu bir kontrol listesi olarak düşünülebileceğini ifade etmiştir. Derecelendirme ölçekleri, kontrol listelerinden farklı olarak özelliğin yapılma sıklığına ya da kalitesine vurgu yapar. Yani bir derecelendirme söz konusudur. Dolayısıyla yapılma sıklığı ya da kalitesi söz konusu olmayan davranışların derecelendirme ölçekleri ile puanlanması uygun değildir. Örneğin, bir sınıf öğretmeninin cümle kurma becerisini kontrol etmek amacıyla öğrencilerinden bir cümle yazmalarını istemesi ve ardından öğrencilerin kurdukları cümlelerde aşağıdaki davranışların varlığını derecelendirme ölçeğiyle gözlemlemesi uygun olmayacaktır:

- Cümlenin sonuna nokta koydu mu?
- Cümleye büyük harf ile başladı mı?

Bu davranışların kontrol listesiyle var/yok ya da yaptı/yapmadı şeklinde gözlemlenmesi uygun olacaktır. Çünkü söz konusu davranışlar bir sıklık ya da kaliteye ilişkin bir bilgi sunmamaktadır. Brookhart (2013) özellikle performans becerilerinin değerlendirilmesinde, sıklık ölçeklerinin çok faydalı olduğunu belirtmiştir. Örneğin, topluluk önünde konuşma becerisini değerlendirirken aşağıda sunulan üç davranış yapılma sıklığı açısından önemli bilgiler sunabilecektir:



"Göz teması kurar" ------ sık sık, ara sıra, nadiren, hiç

"Vurgulara dikkat eder" -----sık sık, ara sıra, nadiren, hiç

"Konuşmalarını örneklerle destekler" ----- sık sık, ara sıra, nadiren, hiç

Bu ve benzeri davranışların sergilenme sıklığı becerinin değerlendirilmesi açısından önemlidir. Eğitimde ve günlük hayatta pek çok davranışın doğası gereği yapılma sıklığı ya da yapılma kalitesini içinde barındırdığını söyleyebiliriz. Örneğin, derse katılım yapılma sıklığı açısından önemlidir ancak yazılan bir kompozisyon ise kalitesi açısından önemlidir.

Brookhart (2013) sıklık ölçeklerinin; ayrıca davranışların, çalışma alışkanlıklarının ve diğer öğrenme becerilerinin değerlendirilmesinde de mükemmel olduğunu ve birçok davranışın "her zaman, sık sık, bazen**,** hiç" şeklinde ifade edilerek iyi bir şekilde tanımlanabileceğini öne sürmüştür. İnsan davranışlarının doğası dikkate alındığında, okullardaki çoğu davranış da dâhil olmak üzere pek çok davranışın bu anlamda kontrol listesinden ziyade derecelendirme ölçekleriyle ifade edilmeye daha uygun olduğu söylenebilir. Örneğin; derse katılım, öz güven, akran zorbalığı, akranlarla ya da öğretmenle iletişim ya da tamamlanan bir ödevin, raporun niteliği gibi. Brookhart (2013) en az üç nedenden dolayı derecelendirme ölçeklerinde kullanılan kalite derecelerinin (örneğin; oldukça iyi, iyi, orta, zayıf gibi derecelendirmeler) öğrenme için çok az fayda sağladığı görüşünü savunmuştur ve bu görüşü şöyle gerekçelendirmiştir:

1. "Kalite dereceleri, kanıtı açıklamadan doğrudan hüküm verir. Yani bir adımı atlar. Örneğin, "Bu mükemmel çünkü ona mükemmel dedim" → açıklama yok. "Performans düzeyleri" gibi görünse de aslında tanımlayıcı değildir. Öğrenme sürecine rehberlik etmez, sadece yargı bildirir
2. Kalite derecelendirme ölçekleri bir yargıdır, açıklama değildir. Bu yüzden öğrenmeyi ilerletecek bilgi içermezler. Öğrenme sürecini durdururlar. Örneğin, bir öğrenci "iyi" seviyedeyse ama "mükemmel" olmak istiyorsa, kalite ölçeği ona ne yapması gerektiğini söylemez. "Mükemmel – İyi – Orta – Zayıf" gibi ölçekler en yaygın yanlış kullanımlardandır. Rubrik olması gereken yerlerde kullanıldığında, "standartlara dayalı" gibi gözüken ifadeler de aslında kalite ölçeklerine dönüşebilir. Eğer ölçek sadece kelimeler listesiyse — örneğin, "İleri, Yeterli, Yeterliliğe Yakın, Başlangıç" — ve yanında açıklama yoksa, bu rubrik değil, kalite ölçeğidir. Bazen "açıklama gibi görünen şeyler" aslında açıklama değildir. Örneğin, "Problemleri ileri düzeyde çözer" ifadesi, tanımlayıcı değil sadece bir kalite yargısıdır.
3. Kalite dereceleri öğretmenleri çoğu zaman görev tabanlı ölçütleri kullanmaya yönlendirir çünkü kalite dereceleri bu tür ölçütlere kolayca uygulanabilir. Örneğin, yazılı bir raporda görev tabanlı ölçütler şu şekilde olabilir: Giriş, Metin, Görseller, Kaynaklar. Siz sadece kaliteyi yargılarsınız aslında her bölüme bir not vermiş olursunuz. Rubriklerin kalite dereceleriyle karıştırılması ya da onlara dönüştürülmesi ise, rubriklerin asıl amacını ve anlamını yok eder" (s. 79).

Özetle, sıklık ölçekleri bir davranışın/özelliğin ne sıklıkla sergilendiğini gösterir, kalite dereceleri ise bir işin ne kadar iyi olduğunu söyler ancak açıklama yapmadığı için öğrenme sürecine sınırlı bir katkı sunar. Öğrenciye eksiği konusunda detaylı değil sınırlı bir geribildirim sağlar. Bunlar genellikle öğrenme yerine, görevin daha yüzeysel yönlerini gözlemler. Bununla beraber kontrol listesi ve sıklık ölçeği doğru yerde kullanıldığında faydalıdır. Kalite ölçeği ise rubrik gibi görünür ancak öğrenmeyi sınırlı oranda destekler, yargı verir ancak detaylı gerekçe sunma noktasında eksiktir. Öğrenci

ölçekteki bir madde üzerinden ve toplam puanlar üzerinden performans düzeyi hakkında bir fikir sahibi olur. Ancak, özellikle detaylı bir geribildirim sağlanması ve bu şekilde öğrenmenin desteklenmesi amaçlanıyor ise rubriklere kıyasla sınırlıdır. Aşağıda öğrencinin derse katılımına ilişkin örnek bir derecelendirme ölçeği sunulmuştur.

Tablo 8

Derse Katılıma İlişkin Örnek Derecelendirme Ölçeği Maddeleri

Derse Katılım Derecelendirme Formu **Öğrenci Adı ve Soyadı:** **Öğrenci Numarası:** **Sınıf:** **Tarih:**					
	Her zaman (5)	**Sık sık (4)**	**Genellikle (3)**	**Nadiren (2)**	**Hiçbir zaman (1)**
1. Derse hazırlıklı gelir	( )	( )	( )	( )	( )
2. Fikirlerini ifade eder	( )	( )	( )	( )	( )
3. Parmak kaldırır	( )	( )	( )	( )	( )
4. …..	( )	( )	( )	( )	( )

Örnekte de görüldüğü üzere burada öğrencinin derse katılımı bir gözlem formuyla değerlendirilmek isteniyor ve puanlama aracı olarak derecelendirme ölçeği kullanılıyor. Bu puanlama araçları vasıtasıyla ölçülen davranışlar doğası gereği yapılma kalitesi ya da sıklığı söz konusu olan davranışlardır. Örneğin, yukarıda verilen örnekte öğrencinin derse hazırlıklı gelme durumu için hazırlıklı gelir ya da gelmez şeklinde bir sınıflama yapılamaz çünkü dönem boyunca bir kez hazırlıklı gelen öğrenciyi ne hazırlıklı gelir sınıfına ne de hazırlıklı gelmez sınıfına dâhil etmek mümkün olmayacaktır. Bu davranışın bir yapılma sıklığı söz konusudur.

Rubriklere benzer şekilde, derecelendirme ölçekleri de gözlemlenmek istenen performansı genel hatlarıyla ya da tek tek her bir alt boyut altında puanlayacak şekilde hazırlanabilirler. Aşağıda Türkçe dersinde kullanılabilecek bütüncül olarak ifade edebileceğimiz genel ve alt boyutlardan oluşan analitik olarak ifade edebileceğimiz iki ayrı derecelendirme ölçeği örneği sunulmuştur:

Tablo 9

Türkçe Kompozisyon Yazma Becerisine İlişkin Örnek Bütüncül Derecelendirme Ölçeği Maddeleri

Kompozisyon Değerlendirme Formu				
Öğrenci Adı ve Soyadı: **Sınıf:** **Öğrenci Numarası:** **Tarih:** **(4) Çok iyi (3) İyi (2) Gelişmekte (1) Yetersiz**				
	(4)	(3)	(2)	(1)
1. Konuya uygun bir başlık kullanılmış				
2. Yazı bir ana fikir çerçevesinde ele alınmış				
3. Yardımcı fikirlere yer verilmiş				
4. Fikirleri destekleyici açıklamalara ya da örneklere yer verilmiş				
5. Görüşler kanıtlarıyla desteklenmiş				
6. Akıcı bir dil kullanılmış				
7. Anlaşılır bir dil kullanılmış				
8. Yazıda özgünlük ya da farklı bir bakış acısına yer verilmiş				
9. Dil bilgisi kurallarına uyulmuş				
10. Yazım kurallarına uyulmuş				
11. Sayfa düzenine uyulmuş				
TOPLAM PUAN:				

Verilen derecelendirme ölçeği örneğinde kompozisyon yazma becerisini alt boyutlara ayırmadan genel özelliklerini dikkate alarak puanlama yapılması amaçlanmaktadır. Örnekte görüldüğü üzere öğrenci "konuya uygun bir başlık kullanmış" olabilir ancak bu uygunluk derecesi "Çok iyi", "İyi", "Gelişmekte" ya da "Yetersiz" görülmüş olabilir. Aynı şekilde öğrenci dil bilgisi kurallarına belli oranda uymuş olabilir. Aşağıdaki örnekte bulunan ve kompozisyon yazma becerisini puanlamak üzere hazırlanan derecelendirme ölçeğinde ise farklı boyutlara ilişkin maddeler ya da davranışlar analitik olarak ayrı ayrı ele alınmış ve dört dereceli olarak değerlendirilmiştir. Ayrıca, her alt boyutun ayrı ayrı puanlanması amaçlanmıştır.

Tablo 10

Türkçe Kompozisyon Yazma Becerisine İlişkin Örnek Analitik Derecelendirme Ölçeği Maddeleri

Kompozisyon Değerlendirme Formu					
Öğrenci Adı ve Soyadı: **Sınıf:** **Öğrenci Numarası:** **Tarih:** **(4) Çok iyi (3) İyi (2) Gelişmekte (1) Yetersiz**					
		(4)	**(3)**	**(2)**	**(1)**
İçerik	1. Konu dikkat çekici				
	2. Konuyu uygun bir başlık yazılmış				
	3. Konuyu destekleyici örneklere yer verilmiş				
	4. Giriş, gelişme ve sonuç bölümlerine yer verilmiş				
	5. Ana fikir açık ve belirgin olarak vurgulanmış				
	6. Ana fikir örnekler ve açıklamalarla desteklenmiş				
	7. Yardımcı fikirler konuya uygun olarak verilmiş				
	8. Fikirler arasında tutarlılık var				
	9. Gereksiz bilgilere yer verilmemiş				
Dil ve Anlatım	10. Cümleler açık ve anlaşılır				
	11. Akıcı bir dil kullanılmış				
	12. Yazı boyunca tutarlı bir anlatım dili kullanılmış				
	13. Konuya uygun ve doğru kelimeler seçilmiş				
	14. Kelime çeşitliliği yeterli				
	15. Gereksiz kelime tekrarları yapılmamış				
	16. Fikirler bir mantık sırasıyla verilmiş				
	17. Giriş, gelişme ve sonuç arasında anlam bütünlüğü var				
	18. Yazım kurallarına uyulmuş				
	19. Noktalama işaretlerine uyulmuş				
Yaratıcılık ve Özgünlük	20. Yazıda özgünlük ve farklı bir bakış acısı var				
	21. Yaratıcı fikirlere yer verilmiş				
	22. Özgün örneklere ya da benzetmelere yer verilmiş				
Sayfa Düzeni	23. Yazı düzenli ve okunaklı				
	24. Sayfa temiz kullanılmış				
	25. Sayfa kenarları doğru kullanılmış				
TOPLAM PUAN:					

Yukarıda değinildiği üzere derecelendirme ölçekleri ihtiyaç görülmesi halinde alt boyutlar halinde de hazırlanabilmektedir. Kompozisyon yazma becerisi; içerik, dil ve anlatım, yaratıcılık ve özgünlük, son olarak sayfa düzeni olmak üzere dört alt boyutta ele alınmıştır. Her boyut altında ise o boyuta ilişkin dikkat edilmesi gereken maddelere yer verilmiştir. Bu haliyle öğrencinin her bir maddeden elde edeceği puan yanında her bir alt boyuttan elde edeceği puanlarda ayrı ayrı yorumlanabilmektedir.

5.3. Kontrol Listesi

Kontrol listelerinin en önemli ayrıcalığı gözlemlenen davranışların var/yok, evet/hayır, yaptı/yapmadı şeklinde yalnızca iki kategoride ölçülmesidir. Söz konusu davranış ya vardır ya da yoktur. Dolayısıyla bir davranışın gerçekleştirilmesinde yapılma kalitesi ya da sıklığı söz konusu ise kontrol listeleri kullanılamaz. Gözlemlenen davranışlar genellikle doğası gereği var ya da yok şeklinde sınıflanabilen davranışlardır. Brookhart (2013) kontrol listesini belirli özelliklerin yer aldığı ve bu özelliklerin var olup olmadığını işaretlemek için ayrılmış bölümlerden oluşan bir liste olarak tanımlamaktadır. Bu liste bir görevin parçalara ayrılmasından yani yapılacakların listelenmesinden oluşabilir. Değinildiği üzere söz konusu davranış ya da özellik ya vardır ya da yoktur. Örneğin; ödevini teslim etti/etmedi durumu, özellikle psikomotor davranışların gözlemlendiği beden eğitimi derslerinde bir yönergenin izlenip izlenmediği, ya da işlem sırası gibi.

Kontrol listeleri rubrikler ya da derecelendirme ölçekleri gibi yorum gerektirmez ve madde formatı bunlara kıyasla çok daha basit yapıdadır. Bu nedenle hazırlanmaları rubriklere ya da derecelendirme ölçeklerine kıyasla daha kolay ve daha hızlıdır. Brookhart (2013) kontrol listelerinin özellikle iki tür durumda kullanılmasının oldukça faydalı olduğunu vurgulamıştır:

1. Gözlemlenecek davranışlar ya da öğrenme çıktıları, nitelikleri açısından değil de yalnızca bir özelliğin varlığı ile tanımlandığı durumlarda hem öğretmenler hem öğrenciler için yararlıdır. Bazı öğrenme çıktıları bu şekildedir. Örneğin, bir ilkokul öğretmeni cümle kurma becerisini belirlemek amaçlı geliştirdiği kontrol listesinde aşağıdaki maddelere yer verebilir ve bu maddeleri "evet/hayır" şeklinde kontrol edebilir:

- Cümlenin sonuna nokta koydu mu?
- Cümleye büyük harf ile başladı mı?
- Cümlede bir özne var mı?
- Cümlede bir yüklem var mı?

2. Kontrol listeleri, öğrencilerin ödev için yönergeleri izleyip izlemediklerini, gerekli tüm parçaları ekleyip eklemediklerini ya da biçimsel gereklilikleri karşılayıp karşılamadıklarını kontrol etme noktasında hem kullanışlıdır hem de faydalıdır. William (2011) böylesi bir listenin "ön kontrol listesi" olarak adlandırılabileceğini ve öğrencilerin bu tür listeleri, çalışmayı teslim etmeden önce kullanmalarını önermiştir.

Dikkat edilirse ikinci kullanımda kontrol listesini kullanmadaki amaç yönergelerin takip edilip edilmediğini ve gerekli tüm unsurların mevcut olup olmadığını kontrol etmektir. Amaç projenin/çalışmanın kalitesini değerlendirmek değildir. Bu şekilde ön kontrol yapıldıktan sonra öğretmen bu çalışmaları ya da raporları niteliği açısından değerlendirme aşamasına geçer. Bu aşamada ise öğretmen çalışmanın istenen niteliklerde bir çalışma olup olmadığını söz konusu çalışmanın ya da performansın yapısına ve değerlendirmeyi yapış amacına göre bir kontrol listesi, derecelendirme ölçeği ya da rubrik aracılığıyla yapabilir. Örneğin, edebiyat dersinde öğretmen öğrencilerinden Yaşar Kemal'in Türk edebiyatına katkılarını inceleyen bir araştırma yapmalarını ve bunu bir rapor halinde sunmasını isteyebilir. Öğretmen bu çalışmaya ilişkin bir kontrol listesi oluşturabilir ve bu listedeki maddeler çalışma raporu yönergesinin gerektirdiği unsurların varlığını kontrol eden bir liste olur. Öğretmen bu durumda aşağıdaki maddelerden oluşan bir rapor kontrol listesi kullanabilir:

- Kapak sayfası var mı?
- İçindekiler sayfası var mı?
- Kaynakça sayfası eklenmiş mi?
- Rapor zamanında teslim edildi mi? vb.

Bu liste, araştırma raporu yönergelerinin gerektirdiği unsurların listesidir. Fakat raporun öğretmen tarafından değerlendirilmesi farklı şekilde gerçekleşir. Edebiyat öğretmeni, büyük ihtimalle bu raporu "Kapsam", "Özgünlük" "Kullanılan kaynaklar", "Dil bilgisi" ve "Anlatım" vb. gibi farklı boyutları içeren analitik bir rubrikle değerlendirebilecektir. Kontrol listesi yalnızca rapor yönergesine uyulup uyulmadığını, belli unsurların var olup olmadığını tespit etme aşamasında kullanılacaktır. Raporun niteliği ise bir rubrik aracılığıyla ortaya konulabilecektir.

Aşağıda kontrol listesine ilişkin bir örneğe yer verilmiştir. Örnekte aşçılık okulundaki öğrencilerin yemek yapma becerilerine ilişkin gözlemlenecek söz konusu davranışlar evet/ hayır şeklinde kontrol listesi aracılığıyla gözlemlenmektedir. Burada kullanılan ölçme aracı gözlem formudur ve puanlama aracı olarak kontrol listesi kullanılmaktadır.

Tablo 11

Yemek Yapımı Hazırlık Aşamasına İlişkin Örnek Gözlem Formu Maddeleri

Yemek Yapımı Hazırlık Gözlem Formu		
Öğrenci Adı ve Soyadı:		**Sınıf:**
Öğrenci Numarası:		**Tarih:**
	Evet (1)	**Hayır (0)**
1. Önlük giyilmiş	( )	( )
2. Bere takılmış	( )	( )
3. Maske takılmış	( )	( )
4. Malzemeler tezgâh üstüne çıkarılmış	( )	( )
5……		
6……		

Burada görüldüğü üzere öğrencilerin yemek yapmaya hazırlık aşamalarına ilişkin gözlemlenecek davranışlar listelenmiştir ve her bir davranışın puanlanmasında evet/hayır şeklinde iki kategori mevcuttur. Listelenen davranışlara bakıldığında bu davranışlar öğrenci tarafından ya gerçekleştirilmiştir ya da gerçekleştirilmemiştir. Bu davranışlarda yapılma derecesi ya da sıklığına ilişkin bir boyut bulunmamaktadır. Bu nedenle kontrol listesiyle puanlanması en uygun yaklaşımdır.

Kontrol listelerini öğrenci ve öğretmen kontrol listesi şeklinde kimin kullandığına göre ikiye ayırmak mümkündür. Öğrenci kontrol listeleri, öğrencinin karmaşık süreçleri ve pek çok alt beceriyi kullanmasını gerektiren karmaşık bir görevi yerine getirirken onlara kılavuzluk etmesi amacıyla hazırlanır. Örneğin, böyle bir kontrol listesiyle öğrencinin o görevi başarıyla yerine getirebilmesi için izlemesi gereken süreçler ve sıra konusunda öğrenciye kılavuzluk edilir. Kontrol listelerindeki maddeler istenilen göreve ilişkin süreç, ürün, çaba, organizasyon gibi ögelere ilişkin olabilir. Ancak kontrol listesinin dersin belirli öğrenme çıktılarının yoklanmasına ilişkin eylemleri, performansı değerlendirme amaçlı kullanıldığını dikkate alırsak böylesi bir durumda ana odak noktası yoklanması hedeflenen beceriler, içerik ve bağlam olacaktır. Diğer bir deyişle öğrenci istenen beceriyi, içeriği, kavramı gerçekleştirememiş ise o görevi yapma sırasında sergilediği çaba, düzen gibi davranışlar yeterli olmayacaktır. Hangi davranışların gözlemlendiğine göre değişmekle birlikte kontrol listesinde ağırlık ürün ve performansta ya da süreçte olabilecektir.

6. SONUÇ

Performans değerlendirme araçları bireylerin yalnızca bilgi düzeylerini değil, aynı zamanda uygulama becerilerini, süreç içerisindeki gelişimlerini ve farklı beceri alanlarındaki yeterliklerini ortaya koymayı sağlayan önemli ölçme araçlarıdır. Performans görevleri, projeler ve portfolyolar aracılığıyla bireylerin gerçek ya da gerçeğe yakın durumlarda sergiledikleri performans değerlendirilebilirken; rubrikler, dereceleme ölçekleri ve kontrol listeleri bu değerlendirme sürecinin daha sistematik, nesnel ve güvenilir biçimde yürütülmesine katkı sağlamaktadır. Bu doğrultuda performans değerlendirme yaklaşımlarının, bireylerin çok boyutlu öğrenme çıktılarını belirlemede gerekli ve önemli olduğu unutulmamalıdır.

KAYNAKÇA | REFERENCES

Airasian, P. W. (2001). *Classroom assessment: Concepts and applications* (5. baskı). McGraw-Hill Higher Education.

Belgrad, S. F. (2013). Portfolios and e-portfolios: Student reflection, self-assessment, and goal setting in the learning process. J. H. McMillan (Ed.), *Sage handbook of research on classroom assessment* (s. 331–346) içinde. Sage.

Brookhart, S. M. (2013). *How to create and use rubrics for formative assessment and grading.* Association for Supervision & Curriculum Development.

Butler, S. & McMunn, N. (2006). A *teacher's guide to classroom assessment: Understanding and using assessment to improve student learning*. Jossey-Bass.

Chase, C. I. (1999). *Contemporary assessment for educators*. Longman.

Çakan, M. (2008). Eğitim sistemimizde yaygın olarak kullanılan sınav türleri. Tekindal, S. (Ed.), *Eğitimde ölçme ve değerlendirme* (s. 92-125) içinde. Pegem Akademi.

Çetin, M. O. (2009). *Öğrencilerin çoktan seçmeli, yazılı yoklama ve performans görevleri ile ölçülen fen ve teknoloji dersi başarıları ve öğrencilerin performans görevlerine ilişkin görüşleri* (Tez No. 241807) [Yüksek lisans tezi, Abant İzzet Baysal Üniversitesi]. YÖK Ulusal Tez Merkezi. https://tez.yok.gov.tr/UlusalTezMerkezi/tezSorguSonucYeni.jsp

Çetin, M. O. & Çakan, M. (2010). Fen ve teknoloji dersi başarılarının farklı yaklaşımlarla ölçülmesi ve bu yaklaşımlara ilişkin öğrenci görüşleri. *Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi, 1*(2), 93-99.

Danielson, C. & Abrutyn, L. (1997). *An introduction to using portfolios in the classroom*. Association for Supervision and Curriculum Development.

Feuer, M. J. & K. Fulton. (1993). The many faces of performance assessment. *Phy Delta Kappan, 74*(6), 478.

Gallagher, J. D. (1998). *Classroom assessment for teachers*. Upper Saddle River.

Gredler, M. E. (1999). *Classroom assessment and learning*. Longman.

Hibbard, K. M., Van Wagenen, L., Lewbel, S., Waterbury-Wyatt, S., Shaw, S. & Pelletier, K. (1996). *A teacher's guide to performance-based learning and assessment*. Association for Supervision and Curriculum Development.

Johnson, R. L., Penny, J. A. & Gordon, B. (2008). *Assessing performance: Designing, scoring, and validating performance tasks.* Guilford Publications.

Johnson, D. W. & Johnson, R. T. (2002). *Meaningful assessment: A manageable and cooperative - process*. Allyn & Bacon.

Kutlu, Ö., Doğan, D. & Karakaya, İ. (2010). *Öğrenci başarısının belirlenmesi: Performansa ve portfolyoya dayalı durum belirleme*. Pegem Akademi.

Marzano, R. J. (2017). *Making classroom assessments reliable and valid: How to assess student learning.* Solution Tree.

McTighe, J., Doubet, K. J. & Carbaugh, E. M. (2020). *Designing authentic performance tasks and projects: Tools for meaningful learning and assessment*. Association for Supervision & Curriculum Development.

Mertler, C. A. (1999). Assessing student performance: A descriptive study of the classroom assessment practices of Ohio teachers. *Education*, *120*, 285-296.

Mertler, C. A. (2001). Designing scoring rubrics for your classroom. *Practical Assessment, Research & Evaluation*, 7, 1-10.

Mertler, C. A. (2003). Classroom assessment: A practical guide for educators. Taylor & Francis.

Montgomery, K. (2001). *Authentic assessment: A guide for elementary teachers*. Longman.

Moskal, B. M. & Leydens, J. A. (2000). Scoring rubric development: Validity and reliability. *Practical Assessment, Research & Evaluation*, *7*(1), 10. https://doi.org/10.7275/q7rm-gg74

Nitko, A. J. (2001). *Educational assessment of students* (3. baskı). Upper Saddle River.

Oosterhof, A. (2001). *Classroom applications of educational measurement* (3. baskı). Upper Saddle River.

Rolheiser, C., Bower. B. & Stevahn, L. (2000). The Portfolio Organizer Succeeding with Portfolios in Your Classroom. Association for Supervision and Curriculum Development.

Tanner, D. E. (2001). *Assessing academic achievement*. Allyn & Bacon.

William, D. (2011). *Embedded formative assessment*. Solution Tree.

YAZAR HAKKINDA | ABOUT AUTHOR

PROF. DR. MEHTAP ÇAKAN

Gazi Üniversitesi | mehtapcakan[at]gazi.edu.tr

ORCID: 0000-0001-6602-6180

Prof. Dr. Mehtap Çakan, lisans eğitimini Hacettepe Üniversitesi Eğitimde Ölçme ve Değerlendirme Bölümünde tamamlamıştır. Yüksek lisans ve doktora eğitimini Amerika Birleşik Devletleri'ndeki Louisiana State Üniversitesinde Eğitimsel Ölçme ve Değerlendirme alanında tamamlayan Çakan, doktora sürecinde psikoloji alanında da yan dal (minor) eğitimi almıştır. 2000 yılında doktora derecesini alarak lisansüstü eğitimini tamamlayan Çakan, 2000–2011 yılları arasında Abant İzzet Baysal Üniversitesinde görev yapmıştır. 2011 yılından bu yana ise Gazi Üniversitesi Gazi Eğitim Fakültesinde akademik çalışmalarını sürdürmektedir. Ulusal ve uluslararası hakemli dergilerde çok sayıda bilimsel çalışması yayımlanmış, ulusal ve uluslararası bildiriler sunmuş ve çeşitli dergilerde hakemlik yapmıştır. Ayrıca lisans ve lisansüstü düzeyde ölçme ve değerlendirme alanlarında dersler vermektedir.

PROF. DR. MEHTAP ÇAKAN

Gazi University | mehtapcakan[at]gazi.edu.tr

ORCID: 0000-0001-6602-6180

Prof. Dr. Mehtap Çakan received her bachelor's degree in Measurement and Evaluation in Education from Hacettepe University. She completed her master's and doctoral studies in Educational Measurement and Assessment at Louisiana State University in the United States. During her doctoral studies, she also completed a minor in Psychology. After earning her Ph.D. in 2000, she served as a faculty member at Abant İzzet Baysal University from 2000 to 2011. Since 2011, she has continued her academic career at the Gazi Faculty of Education, Gazi University. She has published numerous scholarly articles in national and international peer-reviewed journals, has presented papers at national and international conference and has served as a reviewer for various academic journals. In addition, she teaches undergraduate and graduate courses in the field of educational measurement and assessment.

8. BÖLÜM

ÖLÇME VE DEĞERLENDİRMEDE ÇEŞİTLİ ÖĞRENME YAKLAŞIMLARININ KULLANILMASI

USING VARIOUS LEARNING APPROACHES IN EDUCATIONAL MEASUREMENT AND EVALUATION

Doç. Dr. Erkan Hasan ATALMIŞ

Prof. Dr. Gökhan AKSU

ÖLÇME VE DEĞERLENDİRMEDE ÇEŞİTLİ ÖĞRENME YAKLAŞIMLARININ KULLANILMASI

Doç. Dr. Erkan Hasan ATALMIŞ
Manisa Celal Bayar Üniversitesi

Prof. Dr. Gökhan AKSU
Aydın Adnan Menderes Üniversitesi

Öz

Eğitimde ölçme ve değerlendirme, öğrencilerin yalnızca akademik başarılarını belirlemeyi değil, aynı zamanda öğrenme süreçlerini anlamayı ve geliştirmeyi de amaçlamaktadır. Bu nedenle son yıllarda, geleneksel sınavların yanında öğrenme sürecini merkeze alan alternatif ölçme araçları daha fazla kullanılmaya başlanmıştır. Bu bölümde, öğrenme yaklaşımlarına dayalı ölçme araçları kuramsal ve uygulamalı yönleriyle ele alınmıştır. İlk olarak, geleneksel ölçme araçlarının sınırlılıkları açıklanmış ve öğrenci merkezli değerlendirme anlayışının önemi vurgulanmıştır. Daha sonra yapılandırılmış grid, tanılayıcı dallanmış ağaç, kavram haritaları, zihin haritaları, kelime ilişkilendirme testi, ev ödevi, evde sınav, yansıtıcı günlük, öz değerlendirme, akran değerlendirme ve grup değerlendirme araçları ayrıntılı olarak tanıtılmıştır. Her bir aracın amacı, hazırlanma süreci, puanlama yöntemleri ile üstün ve sınırlı yönleri açıklanmış, ayrıca farklı ders alanlarından uygulama örneklerine yer verilmiştir. Bölümde, bu araçların öğrencilerin kavramsal öğrenmelerini, problem çözme becerilerini, eleştirel düşünmelerini, öz düzenleme becerilerini ve yansıtıcı düşünme düzeylerini değerlendirmede önemli katkılar sağladığı belirtilmiştir. Bunun yanında, alternatif ölçme araçlarının öğretmenlere etkili geri bildirim sağlayarak öğretim sürecinin planlanmasına destek olduğu vurgulanmıştır. Sonuç olarak, öğrenme yaklaşımlarını temel alan ölçme araçlarının yalnızca öğrencilerin ürünlerini değil, öğrenme süreçlerini de değerlendirmeye olanak sağladığı ifade edilmiştir.

Anahtar Kelimeler

Ölçme ve Değerlendirme, Alternatif Değerlendirme, Öğrenme Yaklaşımları, Biçimlendirici Değerlendirme, Öğrenci Merkezli Değerlendirme, Öğrenme Süreci

USING VARIOUS LEARNING APPROACHES IN EDUCATIONAL MEASUREMENT AND EVALUATION

Assoc. Prof. Dr. Erkan Hasan ATALMIŞ
Manisa Celal Bayar University

Prof. Dr. Gökhan AKSU
Aydın Adnan Menderes University

Abstract

Measurement and evaluation in education aims not only to determine students' academic achievement but also to understand and improve their learning processes. For this reason, alternative assessment tools that emphasize the learning process have become increasingly common alongside traditional examinations in recent years. This chapter examines assessment tools based on different learning approaches from both theoretical and practical perspectives. First, the limitations of traditional assessment methods are discussed, and the importance of student-centered assessment is emphasized. The chapter then introduces a variety of assessment tools, including structured grids, diagnostic branched trees, concept maps, mind maps, word association tests, homework, take-home examinations, reflective journals, self-assessment, peer assessment, and group assessment. For each tool, its purpose, development process, scoring procedures, strengths, and limitations are explained, together with practical examples from different subject areas. The chapter also discusses how these tools can be used to assess students' conceptual understanding, problem-solving skills, critical thinking, self-regulation, and reflective thinking. In addition, it highlights the role of alternative assessment tools in providing meaningful feedback to teachers and supporting the planning and improvement of the instructional process. Finally, it is emphasized that assessment tools based on different learning approaches enable educators to evaluate not only students' learning outcomes but also their learning processes.

Keywords

Educational Measurement and Evaluation, Alternative Assessment, Learning Approaches, Formative Assessment, Student-centered Assessment, Learning Process

1. GİRİŞ

Eğitimde ölçme ve değerlendirmenin amaçları arasında, öğrencilerin ne kadar bilgi edindiklerini belirlemenin ötesinde, öğrenme süreçlerinin derinliğini, kalıcılığını ve anlamını ortaya koymak da yer alır. Ancak uzun yıllar boyunca ölçme ve değerlendirme uygulamaları, büyük ölçüde sonuç odaklı, ürün merkezli ve sınav temelli yöntemlerle sınırlı kalmıştır. Bu tür geleneksel ölçme araçları; öğrenci performansını genellikle çoktan seçmeli, doğru–yanlış veya kısa yanıtlı testlerle ölçmekte, öğrencinin düşünme biçimini, problem çözme sürecini ve yaratıcılığını çoğunlukla göz ardı etmektedir (Black & Wiliam, 1998; Nitko & Brookhart, 2014). Bu durum, öğrenmenin yalnızca bilişsel yönüne odaklanan dar bir bakış açısını beraberinde getirmiştir. Oysa çağdaş eğitim anlayışı, öğrencinin aktif, sorgulayan ve üretken bir birey olarak öğrenme sürecine katılımını esas alır. Öğrencinin yalnızca "ne bildiği" değil, bu bilgiyi "nasıl kullandığı", "nasıl yapılandırdığı" ve "nasıl ifade ettiği" de öğrenmenin ayrılmaz parçalarıdır. Bu nedenle, öğrenme sürecini yalnızca sonuçlarla sınırlamayan farklı değerlendirme yollarına da ihtiyaç duyulmaktadır.

Günümüzde yapılandırılmış grid, tanılayıcı dallanmış ağaç, kavram ve zihin haritaları, kelime ilişkilendirme testleri, ev ödevleri ve yansıtıcı günlük gibi öğrenme yaklaşımlarını kapsayan ölçme araçları; öğrencilerin bilişsel yapılarının, kavramlar arası ilişkilerinin ve öğrenme stratejilerinin çok boyutlu biçimde incelenmesine olanak tanımakta ve bu yönüyle öğrenme süreçlerine ilişkin derinlemesine bilgi sunan etkili değerlendirme yaklaşımları olarak öne çıkmaktadır. Bu araçlar yalnızca öğrencinin bilgi/becerisini değerlendirmeye değil, aynı zamanda öğrenme sürecine dönüt sağlamaya, öğretim stratejilerini biçimlendirmeye ve öğrencinin kendi öğrenmesini izleme becerisini geliştirmeye de hizmet eder (Gulikers ve ark., 2004).

Bu ölçme araçlarının yaygınlaşmasındaki bir diğer önemli gerekçe, öğrenmenin bireysel farklılıkları dikkate alan, daha adil ve kapsayıcı değerlendirme yaklaşımlarına duyulan gereksinimdir. Geleneksel testler çoğu zaman tek bir doğru cevabı ararken, öğrenme yaklaşımlarını kapsayan bu araçlar öğrencinin farklı ifade biçimlerine, özgün düşünme yollarına ve bireysel gelişim hızına alan tanır. Ayrıca bu araçlar, öğrenciyi değerlendirme sürecinin pasif bir nesnesi olmaktan çıkararak onu sürecin etkin bir öznesi hâline getirir. Öz, akran ve grup değerlendirme gibi yaklaşımlar, öğrencinin kendi öğrenme sürecini izlemesini, güçlü ve zayıf yönlerini fark etmesini, öğrenme sorumluluğunu üstlenmesini ve nasıl daha etkili öğrenebileceğini anlamasını sağlar (Dochy ve ark., 1999). Bu tür araçların ölçme–değerlendirme sürecine kazandırdığı en önemli katkılardan biri de öğrenme ve öğretim arasındaki sınırları ortadan kaldırmalarıdır. Öğrenme yaklaşımlarına dayalı ölçme araçları, öğrencinin öğrenme sürecine ilişkin verileri yalnızca not verme amacıyla değil, öğretim süreçlerini geliştirmek amacıyla da kullanır.

Öğrencinin hatalarını fark etmesini, kavramsal eksikliklerini gidermesini ve düşünme biçimini yeniden yapılandırmasını sağlar. Böylelikle öğretmen, ölçmeyi bir kontrol aracı olmaktan çıkarıp bir öğrenme fırsatına dönüştürür. Bu anlayış, biçimlendirici değerlendirme (formative assessment) yaklaşımının temelini oluşturur (Stiggins, 2005). Sonuç olarak, ölçme ve değerlendirme kullanılan öğrenme yaklaşımlarını kapsayan bu ölçme araçları yalnızca ölçmenin biçimini değil, eğitimin felsefesini de dönüştürmektedir. Bu araçlar, öğretimi yönlendiren bir geri bildirim aracı, öğrenmeyi derinleştiren bir süreç ve öğrenciyi merkeze alan bir değerlendirme kültürünün temel bileşeni hâline gelmiştir.

Kitabın bu bölümde, çeşitli öğrenme yaklaşımlarını kapsayan ölçme araçları tanıtılmıştır. Bu araçların öncelikle amacı sunulmuş ardından hazırlanması ve puanlanması ile üstün ve sınırlı yönleri sunulmuştur. Bunun yanında ele alınan her bir yaklaşımın ölçme ve değerlendirme sürecine nasıl bütüncül bir katkı sundukları ayrıntılı biçimde tartışılmıştır.

2. YAPILANDIRILMIŞ GRİD (STRUCTURED GRID)

Yapılandırılmış grid, öğrencilerin düzeylerine uygun olarak 9, 12 ya da 16 hücreden oluşan ve bu kutucuklara tanım, terim, formül, önerme, sayı, şekil ya da görsel gibi çeşitli yanıtların yerleştirildiği, soruların ise bu yanıtların kutular içinde bulunacağı şekilde düzenlendiği bir teknik olarak ifade edilmektedir (Bahar ve ark., 2008). Grid, kavram hatalarını, yanlış öğrenmeleri ve anlamlı öğrenmeleri açığa çıkarması nedeniyle etkili bir değerlendirme yöntem olarak kabul edilir. Çoktan seçmeli maddeler temel alınarak geliştirilen ve bu süreçte şans başarısını oluşmasını engelleyen yapılandırılmış grid, puanlama aşamasında çoktan seçmeli testlerdeki gibi tek doğru yanıt odaklı bir değerlendirme içermez, kısmi bilgileri de puanlamaya dâhil edilir (Bahar, 2001, akt. Öztürk, 2011). Bu sebeple çoktan seçmeli testlerin önemli bir sınırlılığı olan şans başarısını büyük ölçüde azaltması, bilgili öğrenci ile bilgisi sınırlı olan öğrenci arasındaki farkın daha açık biçimde ortaya çıkmasını sağlar (Bahar, 2001).

Çalışmalar, bu tekniğin amacının anlamlı öğrenmenin belirlenmesinde, kavram yanılgılarının ortaya çıkarılmasında ve öğrenme eksikliklerinin tespit edilmesinde etkili olduğunu göstermektedir (Bahar ve ark., 2008). Ayrıca yöntemin kutulara farklı türde yanıtların yerleştirilebilmesi, hem görsel hem de sözel düşünmeyi destekleyici bir özellik sunmaktadır. Öğrencilerin bu tekniği bireysel olarak da uygulayabildikleri belirtilmektedir (Bahar ve ark., 2008; Eroğlu & Kelecioğlu, 2011). Alternatif kavramların belirlenebilmesi ve anlamlı öğrenmenin ne kadar etkili şekilde gerçekleştiğine ilişkin bulguları ortaya koyması, tekniğin eğitim uygulamalarında önemli bir değerlendirme aracı olarak öne çıkmasına katkı sağlamaktadır (Aydoğdu & Kesercioğlu, 2005).

2.1. Hazırlanışı ve Puanlanması

Yapılandırılmış gridde öğrencilerin bilgi düzeylerine uygun sorular ile bu soruların yanıtlarının yerleştirildiği kutucuklar hazırlanır. Bu kutucukların sayısı öğretmen tarafından hazırlanırken öğrenci düzeylerine uygun olarak değişmekte ve öğretmen konuyla ilgili soru hazırlayarak cevaplarını rastgele kutucuklara yerleştirir. Numaralı kutucuklardan oluşan yapının içinde bilginin sunulması, uygun kutucukların öğrenciler tarafından seçilmesi ve seçilen bu kutucukların mantıksal olarak sıralanmasını gerektiren birtakım sorulara cevap verilmesiyle oluşmuş bir araçtır (Johnstone ve ark., 2000). Öğrencilerden soruların doğru cevaplarını söylemeleri ve sıralamaları beklenir. Öğrenciler soruları cevaplarken gridin en önemli özelliğinden biri olan, bir kutunun birden fazla soruya ait olabilmesidir, bu nedenle yapılandırılmış gridlerde seçenek sayısı klasik çoktan seçmeli maddelere göre daha fazladır (Yazıcıoğlu, 2007). Bu durum, tüm maddelerin ortak bir seçenek havuzu üzerinden yapılandırılmasını gerektirir ve gridin hazırlanış mantığını oluşturan temel unsurlardan biridir. Grid tekniğinin çözümleme sürecinde iki ayrı puanlama yaklaşımından yararlanılmaktadır. Bu yaklaşımlar arasında literatürde daha sık tercih edilen yöntem aşağıda verilmiştir:

$$\frac{C1}{C2} - \frac{C3}{C4}$$

C1: Öğrencinin işaretlediği doğru hücre sayısı

C2: Gridde yer alan toplam doğru hücre sayısı

C3: Öğrencinin işaretlediği yanlış hücre sayısı

C4: Gridde bulunan toplam yanlış hücre sayısı

Bu formül uygulandığında elde edilen değerler –1 ile +1 aralığında bulunmaktadır. Hesaplanan puan önce 1 ile toplanarak negatif değerlerden arındırılarak, ardından 5 ile çarpılarak sonuçlar 0–10 aralığına dönüştürülmektedir (Johnstone ve ark., 2000). Yapılandırılmış grid tekniğiyle ilgili bir örnek Şekil 1'de verilmektedir.

Şekil 1

10. Sınıf İkinci Dereceden Bir Bilinmeyenli Denklemler Konusu İle İlgili Yapılandırılmış Grid Örneği

1 x^2- 10x+25=0	2 x^2+12x+35=0	3 $9x^2$-81x=0
4 x^2-9=0	5 x^2 +3x+4=0	6 x^2-2x+6=0
7 $4x^2$-16=0	8 x^2-4x+3=0	9 $5x^2$-125=0

1. Kutucuklardan hangisi/hangilerinin diskriminant değeri pozitiftir?
2. Kutucuklardan hangisi/hangileri (x+5).(x+7) şeklinde çarpanlarına ayrılır?
3. Kutucuklardan hangisi/hangilerinin çözüm kümesi elemanlarından biri 5'tir?
4. Kutucuklardan hangisi/hangileri tam karedir?

2.2. Üstün ve Sınırlı Yönleri

Reid'e (2003, s. 32) göre yapılandırılmış gridin avantajları:

- Hazırlık süreci, çoktan seçmeli maddelerin oluşturulmasına göre daha kolaydır.
- Öğrencilerin kavramı anlama düzeylerini test edebilmektedir.
- Öğrencinin doğru seçeneği tahmin ederek bulma olasılığını büyük ölçüde ortadan kaldırır.
- Öğrencilerin verdiği doğru cevaplar, temel kavramların ne ölçüde kavrandığını ortaya koyar.
- Öğrencilerin hatalı cevapları, kavram yanılgılarını ve yanlış öğrenmeleri ortaya çıkarır.
- Yanıtların açık biçimde sunulması, oldukça açıklayıcı ve yol göstericidir.
- Yapılandırılmış grid aracılığıyla kavramların farklı boyutlarına ilişkin yeni bir bakış kazandıracak çok sayıda soru sorulabilir.

Özatlı'ya (2006) göre yapılandırılmış gridin dezavantajları:

Yapılandırılmış grid, alışılmış sınav türlerinden farklı olduğu için öğrencilerin dikkatini çekmekte ve bulmaca niteliğinde düşünülmektedir. Ancak bu teknik uygulandıktan hemen sonra sınıf ortamında geri bildirim verilmediği durumda, öğrenciler bir süre sonra hangi maddeleri neden o şekilde cevaplandırdıklarını hatırlamakta güçlük çekebilmektedir. Bu nedenle uygulamanın ardından sonuçların derste değerlendirilmesi ve öğrencilerin zorlandığı ya da yanlış cevapladığı maddeler üzerinde tartışılmalıdır.

3. TANILAYICI DALLANMIŞ AĞAÇ (DIAGNOSTIC BRANCHED TREE)

Tanılayıcı Dallanmış Ağaç (TDA) geleneksel çoktan seçmeli madde türlerinden doğru-yanlış (D/Y) tipi maddelerin çoklu olarak sorulmasını sağlayan bir tekniktir. Burada dikkat edilmesi gereken husus geleneksel doğru-yanlış maddelerinde birbirinden bağımsız sorularken, TDA tekniğinde ise doğru-yanlış tipi maddeler ardışık ve birbiriyle ilişkili olarak oluşturulmaktadır. Aşağıda sırasıyla TDA tekniğinin amacı, hazırlanışı, puanlama süreci, avantajları ve dezavantajlarından bahsedilmektedir.

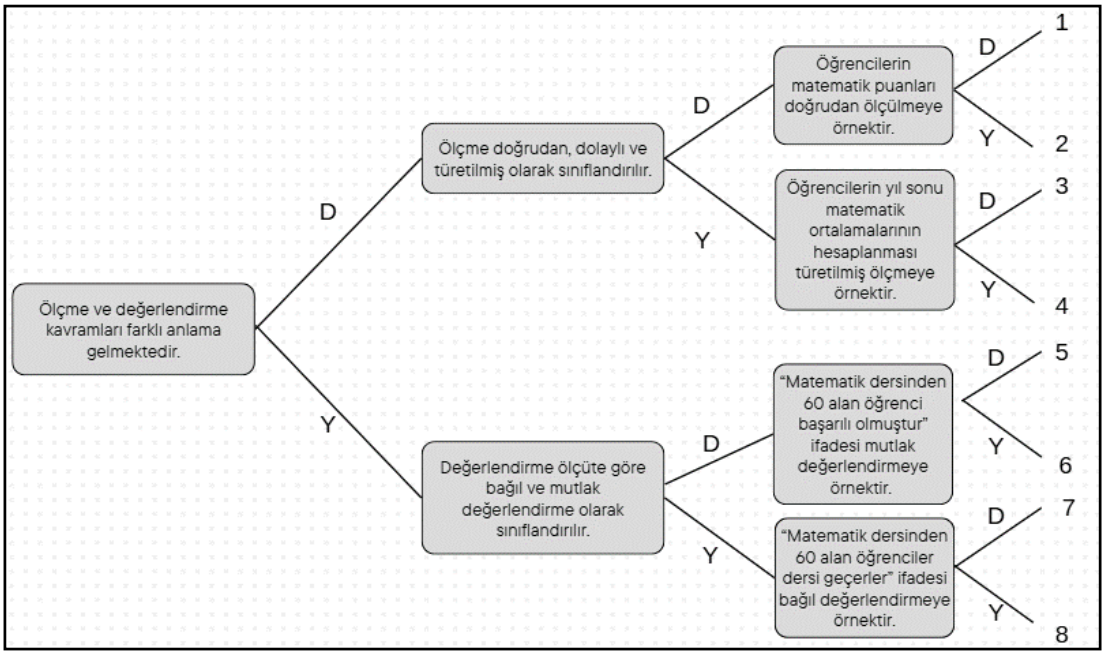
TDA tekniğinin temel amacı, öğrencilerin bir konudaki kavramsal anlayışını ve bilgiler arasındaki ilişkileri ortaya çıkarmaktır. Bu yöntem, konuların nedensellik ve süreçlerini anlama becerilerini geliştirirken, öğrencilerin bilgiyi mantıksal bir yapıda organize etme yeteneklerini değerlendirmeye olanak tanır (Şahin & Çepni, 2011). Aynı zamanda, öğrenme sürecinde ortaya çıkan boşlukları ve yanlış anlamaları tespit etmek için de etkili bir araç olan TDA, öğretmenlere öğretim etkinliklerini değerlendirme ve daha etkili öğrenme ortamları yaratma konusunda değerli geri bildirimler sunmaktadır (Bahar, 2001).

3.1. Hazırlanması ve Puanlaması

TDA tekniğini hazırlamak için öncelikle bir konuya ilişkin anahtar kavram ya da kavramlar belirlenir. Ardından bu kavrama yönelik bir önerme oluşturulur ve bu önerme modelde kök görevi görmektedir. Bu önermeye verilebilecek doğru veya yanlış cevap, gidilecek bir sonraki yolu belirlemektedir. Sonraki aşamada, muhtemel verilen tüm cevaplara uygun yeni önermeler oluşturulur ve modelde ağacın ilk dalları oluşturulmuş olur. Sonraki adımlarda benzer örüntü devam ederek TDA tamamlanmış olur. Kısaca TDA tekniğinde önermeler genelden özele doğru oluşturulmaktadır. TDA çizilirken yatay veya dikey şeklinde ve her dal doğru ve yanlış iki dala ayrılarak bir model çizilir. Şekil 2'de "ölçme ve değerlendirme" konusuna ilişkin TDA örneği gösterilmektedir.

Şekil 2

Ölçme ve Değerlendirme Konusuna İlişkin TDA Örneği



Şekil 2 incelendiğinde en soldaki metin kutudaki "ölçme ve değerlendirme kavramları farklı anlama gelmektedir" önermesinin (1. adım) doğru ya da yanlış olmasına göre iki önerme (2.adım) daha ortaya çıkmaktadır. Ardından bu iki önermenin her biri için tekrardan iki önerme yazılarak toplamda 4 önerme (3.adım) ile model sonlandırılmıştır. Buradaki örnekte 3 adım ilerlendiğinden $2^3 = 8$ çıkış ortaya çıkarken ancak bunlardan 1 tanesi doğru olmalıdır.

TDA tekniğinde her bir önermenin doğru ya da yanlış olmasına göre puanlama yapılır. Örneğin, Şekil 2'de doğru çıkış olan 2.çıkışa gelmek için 3 adımı tamamlamak gerekir. Daha açık bir ifadeyle ilk önerme olan "Ölçme ve değerlendirme kavramları farklı anlama gelmektedir." ifadesi "D"; ikinci önerme olan "Ölçme doğrudan, dolaylı ve türetilmiş olarak sınıflandırılır." ifadesi "D"; üçüncü önerme olan "Öğrencilerin matematik puanları dolaylı ölçmeye örnektir" ifadesi "Y" olarak seçilmelidir. Kısaca (D, D, Y) yolunu izleyen bir öğrenci 3 tam puan alır. Ancak bu yolu izlemeyen öğrenciler doğru bildikleri önerme kadar puan almaktadır. Mesela 8.çıkışa ulan bir öğrenci (Y, Y, Y) yolunu izleyerek, bu önermelerden birini (Matematik dersinden 60 alan öğrenciler dersi geçerler ifadesi bağıl değerlendirmeye örnektir) doğru bilmiş ve sadece 1 puan almış olur.

3.2. Üstün ve Sınırlı Yönleri

Öğrenciler TDA tekniğinde sadece bilgiyi ezberlemekle kalmayıp, aynı zamanda kavramlar arasındaki ilişkileri daha rahat şekilde anlayabilirler. Bu bağlamda, öğrencinin hangi konularda zorlandığını ve nerelerde yanlış anlamalar yaşadığını net bir

şekilde tespit ederek TDA tekniği öğrenme boşluklarını belirginleştirir. Görsel ve yapısal özelliği, karmaşık konuları anlaşılır bir çerçeveye oturtarak öğrenmeyi ve değerlendirmeyi kolaylaştırırken, öğrencilerin kendi öğrenme süreçlerine aktif katılımını teşvik ederek motivasyonlarını artırabilir. Yine öğrenciler TDA tekniğini kullanırken başlangıç önermelerde yanlış yapsa dahi, sonraki adımlarda yanlış yaptığının farkına vardığında geriye dönerek yanlışlarını düzeltebilir. Ayrıca geleneksel seçmeli sorularında gibi net doğru/yanlış puanlamanın aksine TDA tekniği öğrencilere kısmi puan alma imkânı tanır. Son olarak bilgisayar ortamında farklı yaş gruplarına uygulamaya ve puanlanmaya oldukça elverişlidir.

TDA tekniğinin de bir kısım sınırlılıkları bulunmaktadır. Özellikle karmaşık konular için doğru-yanlış şeklinde dallandırma süreci öğretmenler için zaman alıcı bir hazırlık süresi gerektirebilir. Öğrenciler TDA tekniğine ilk kez aşina olduklarında, ne istendiğini anlamaları ve uygulamaları zaman alabileceğinden zorlayıcı olabilir ve ek açıklama veya pratik gerektirebilir. TDA tekniğinde şans faktörü bulunmaktadır. Şekil 8.4 incelendiğinde şans faktörünün 8'de 1 diğer ifade ile %12,5 olduğu görülmektedir. TDA tekniğinde öğrenciler seçeneklere maruz bırakıldığından öğrencilerin analiz etme, problem çözme ve soyut düşünme gibi becerilerinin gelişmesine diğer bir ifade ile üst düzey becerini kullanmasına katkı sağlayamaz.

4. KAVRAM HARİTALARI (CONCEPT MAPS)

Öğretimin her kademesinde kullanılan kavram haritaları, belirli bir konuya ilişkin kavramlar arasındaki ilişkilerin görsel olarak sunulduğu araçlardır. Özellikle anlamlı bir öğrenmenin gerçekleştirilmesine katkı sunan kavram haritaları, öğrencilerin belirli bir konuya ilişkin kavramları nasıl anladığını, ilişkilendirdiğini ve organize ettiğini ortaya koymak için kullanılmaktadır. Aşağıda sırasıyla kavram haritalarının amacı, hazırlanışı, puanlama süreci, avantajları ve dezavantajlarından bahsedilmektedir.

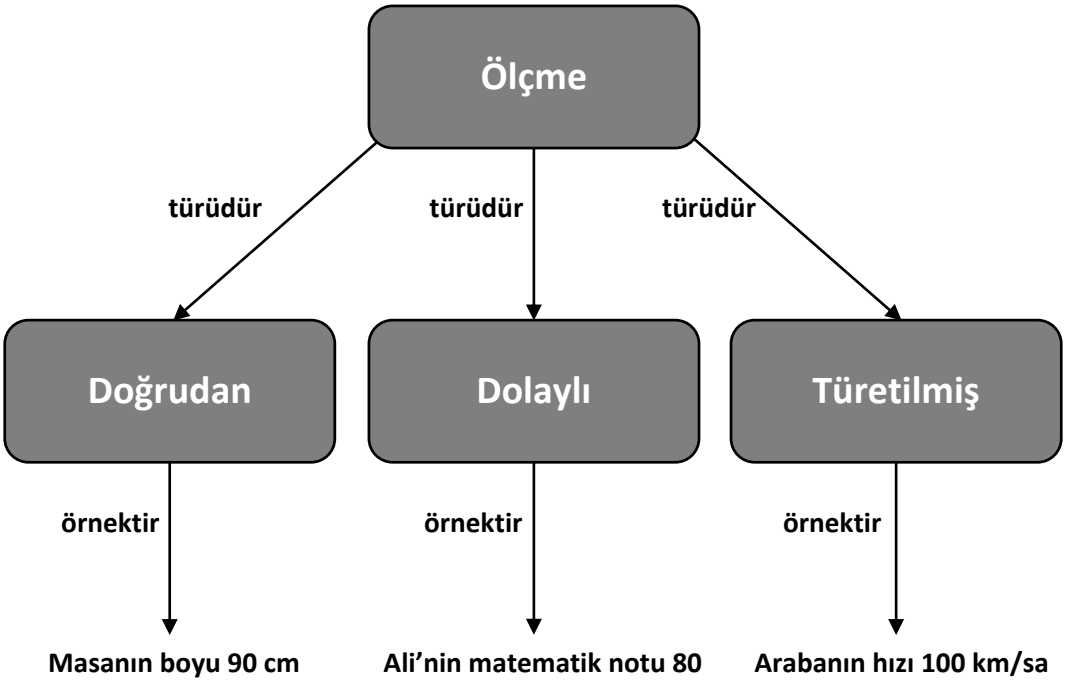
Bir öğretim yöntem ve tekniği olan kavram haritaları, öğrencilerin bir konuya ilişkin kavramları anlamlı bir şekilde öğrenmelerine yardımcı olmaktadır. Bu yöntem, öğrencilerin yeni bilgileri önceden edindikleri bilgilerle ilişkilendirmesini ve bilgiyi yeniden yapılandırmasını sağlayarak anlamlı öğrenmeyi desteklemektedir. Anlamlı öğrenmenin önemini vurgulayan Gestalt Kuramı'na göre, bir bütün parçalarının toplamından daha fazladır ve birey, bütünü parçalara ayırarak değil, bütünlük içinde algılamaktadır. Kavram haritaları da bu yaklaşımla uyumlu olarak, bir konuya ilişkin kavramları ilişkilendirerek konu bütünlüğünü sağlamaktadır. Bu sayede, öğrenciler bilgiyi hatırlama düzeyinin ötesine geçerek anlama düzeyinde öğrenme gerçekleştirir ve ezbere dayalı öğrenmeden uzaklaşarak anlamlı öğrenmeyi gerçekleştirirler. Kavram haritalarından elde bilgiler aynı zamanda öğrenci başarısının izlenmesi amacıyla ölçme ve değerlendirme yöntemi olarak da kullanılabilmektedir.

4.1. Hazırlanması ve Puanlanması

Kavram haritası oluşturmanın ilk adımı, verilen bir konunun geneli kapsayan ana kavramları belirleyerek, haritada bu kavramları genelden özele doğru diğer bir ifadeyle sıralamaktır. Ardından, bu ana kavramlara bağlı olarak daha özel ve alt kavramlar belirlenip mevcut hiyerarşiye yerleştirilir. Bu kavramlar arasındaki ilişkiler, çizgilerle birbirine bağlanırken, bu ilişkiler "neden", "sonuç", "çeşididir", "bir türüdür" ve benzer kelimelerle adlandırılarak kavramlar arası anlamlı bağlantılar kurulur. Son olarak, haritayı daha somut ve anlaşılır kılmak için ilgili örnekler eklenerek, kavramlar arasındaki ilişkiler ve bütünlük daha da zenginleştirilir. Kavram haritaları farklı şekillerde ve farklı amaçlar için kullanılır. Bunlara örnek olarak anlam çözümleme tabloları, olaylar zinciri dizinleri, balık kılçığı haritası, sınıflama haritası ve örümcek haritası verilebilir (Başol, 2013). Ölçme ve değerlendirme kavramlarına ve isim çeşitlerine ilişkin bir kavram haritaları Şekil 3 ve Şekil 4'te gösterilmektedir.

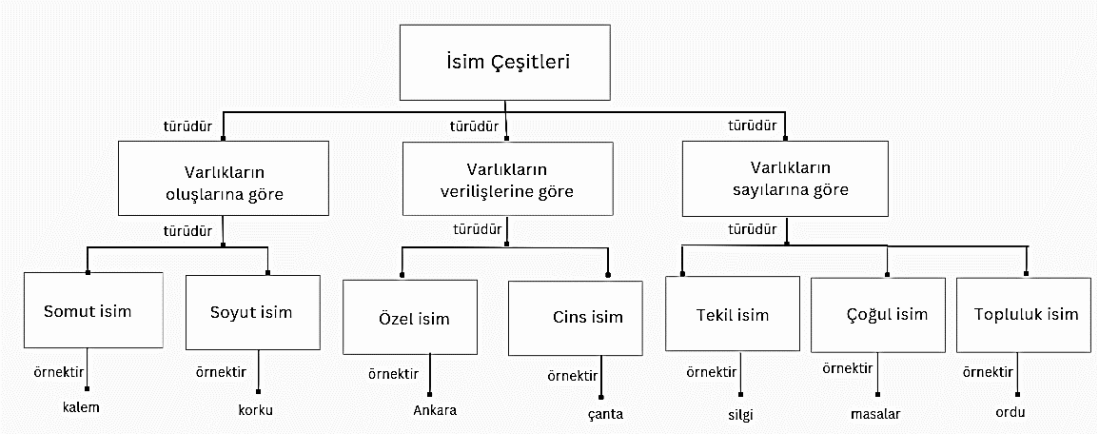
Şekil 3

Ölçme ve Değerlendirme Kavramlarına İlişkin Kavram Haritası



Şekil 4

İsim Çeşitlerine İlişkin Kavram Haritası



Kavram haritalarının değerlendirilmesinde yaygın olarak bütünsel ve ilişkisel olmak üzere iki temel sistem kullanılır. Novak ve Gowin (1984) tarafından geliştirilen bütünsel puanlama sisteminde puanlama bir bütün olarak ele alınır ve dört kısımdan oluşur: kavramlar, iki kavram arasındaki önerme (ok yönü) ve bağlantının adlandırılması, hiyerarşi basamakları ve örnekler. Bu yöntemde puanlama yapılırken iki kavram arasında önermenin doğruluğuna 1 puan; hiyerarşide aynı düzeyde bulunan kavramların doğru hiyerarşi basamağına bağlanması 5 puan; hiyerarşide farklı düzeyde bulunan kavramların arası yapılan çapraz bağlantıların doğruluğuna 10 puan; kavramlara uygun doğru örneklere ise 1 puan verilir. McClure ve arkadaşları (1999) tarafından geliştirilen ilişkisel puanlamada ise sadece kavramlar arası ilişkiler analiz edilir ve puanlama 3 kısımdan oluşur: iki kavram arasındaki ilişki, iki kavram arasındaki önerme (ok yönü) ve bağlantının adlandırılması dikkate alınır. Bu yöntemde puanlama yaparken her bir kısım 1 puan olmak üzere toplamda 0-3 puan üzerinden puanlama yapılır. Kavram haritaları kolay ve nesnel puanlama imkânı sunmaktadır.

4.2. Üstün ve Sınırlı Yönleri

Kavram haritaları, herhangi bir konuya ilişkin kavramları ve kavramlar arası ilişkileri görsel olarak sunduğundan öğrencilerin anlamlı öğrenme becerilerini artırır. Bu sayede onların yanlış kavram yanılgılarının ortaya çıkmasını, ardından bu yanılgıların giderilmesi hususunda öğrenci ve öğretmene yol gösterir.

Kavram haritalarının hazırlaması ve puanlaması zaman alabilmektedir. Öğrenciler kavram haritası yapmayı bilmiyorsa, bu durum ölçme işlemini zorlaştırabilir. Her ders veya konu için uygun bir yöntem olmayabileceği gibi, puanlamada öznel yargılar da devreye girebilir, özellikle öğretmenler arasında farklılıklar oluşması söz konusu olabilir. Bunun için birden fazla puanlayıcı tarafından puanlanması önerilebilir.

5. ZİHİN HARİTASI (MIND MAP)

Zihin haritası bir konuya ilişkin bilgilerin organize edilerek görsel olarak sunulduğu bir diyagram türüdür (Kortelainen & Vanhala, 2004). Bu sebeple kavram haritaları ile benzerlik göstermektedir. Ancak zihin haritası, kavram haritalarında olduğu gibi kavramlar arası ilişkiler doğrusal ve hiyerarşik olarak tanımlanmaz. Bunun yerine kavramlar (anahtar kelimeler) merkeze yerleştirilir ve ardından bağlantılarla ilgili alt fikirler bağlanır. Bu sebeple zihin haritaları merkezden dışa doğru okunur. Ayrıca bağlantılar daha çok renkler, resimler ve simgelerle desteklenir. Aşağıda sırasıyla zihin haritalarının amacı, hazırlanışı, puanlama süreci, avantajları ve dezavantajlarından bahsedilmektedir.

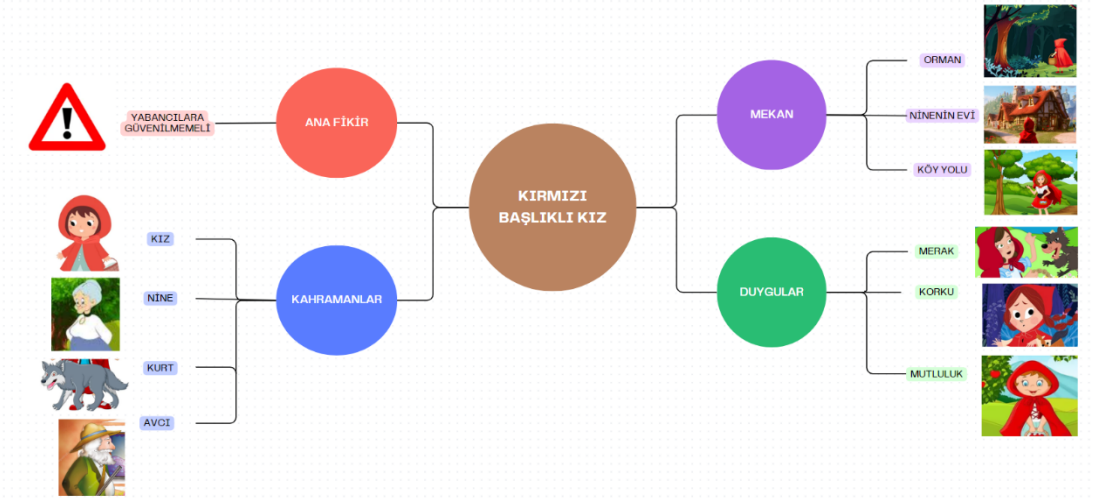
Zihin haritalarının öncelikle temel amacı, karmaşık bilgileri organize ederek daha anlaşılır bir yapı oluşturmaktır. Bu görsel ve ilişkisel yapı, öğrenmeyi kolaylaştırmakta ve bilginin anlamlı ve kalıcı olmasını sağlamaktadır. Tony Buzan tarafından 1960'larda not alma tekniği olarak geliştirilen zihin haritaları (Evrekli & Balım, 2010), ders kitapları ve toplantılarda konuya ilişkin bilginin özetlenmesine yardımcı olmakta, projeler, sunumlar ve yazılan raporlar için okuyucu ve dinleyiciler için bilginin ilgi çekici olmasını sağlamaktadır.

5.1. Hazırlanması ve Puanlaması

Zihin haritasını oluşturmak için bir kısım basamakları takip etmek gerekmektedir. Öncelikle zihin haritasının merkezine ana fikir ya da daha akılda kalması için ilgili sembol veya resim yerleştirilmelidir. Ardından bu ana fikriden yayılan dallar çizilir, her bir dal bir anahtar kelime ile desteklenir ve yine her dal üzerine bir kelime veya kısa bir ifade yazılır. Sonrasında ana dallar üzerinden daha ince alt dallar eklenir ve benzer işlemler bu aşamada da yapılır. Bu arada oluşturulan dallar beyin kıvrımlarına benzemesi açısında kıvrımlı şekilde çizilmeli ve daha akılda kalması için farklı renkli çizgilerle tasarlanmalı, sembol ve resimlerle desteklenmelidir (Türkyılmaz, 2017). Şekil 5'te "Kırmızı Başlıklı Kız" masalı konumlandırılarak ve Şekil 6'da "iskelet sistemine" yönelik zihin haritalarını gösterilmektedir.

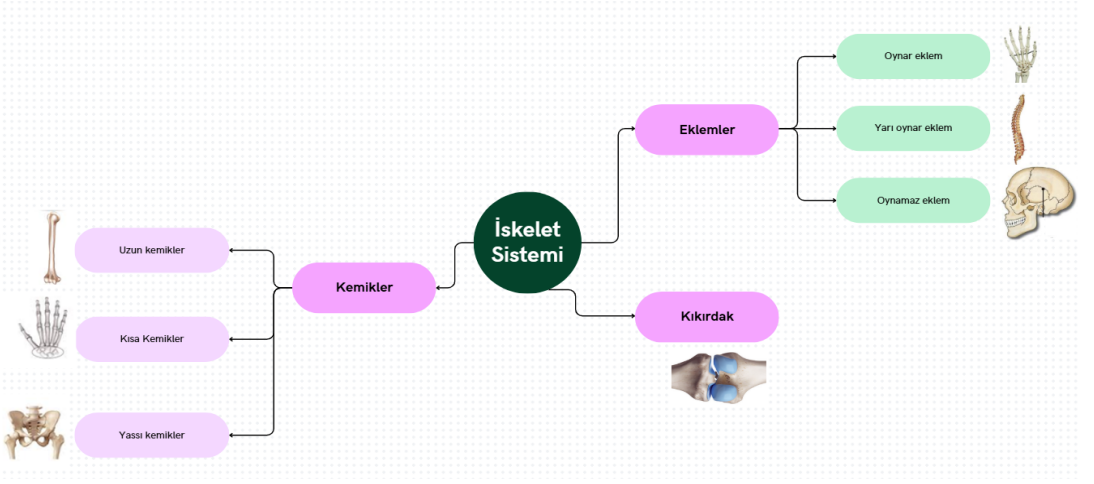
Şekil 5

Kırmızı Başlıklı Kız Zihin Haritası Örneği



Şekil 6

İskelet Sistemine Yönelik Zihin Haritası Örneği



Zihin haritalarının puanlanması, kavram haritalarından farklı olarak daha öznel ölçütlerle puanlanmaktır. Özellikle puanlamanın daha nesnel olması açısından daha önceden belirlenmiş ölçütlerle hazırlanmış dereceli puanlama anahtarının kullanılması yerinde olacaktır. Rubrik hazırlanırken şu ölçütler göz önüne alınabilir: üretilen fikirler (fikir sayısı, özgünlüğü ve çeşitliliği), organize etme yeteneği (fikirler arası ilişkilerin hiyerarşisi ve netliği), kullanılan anahtar kelimeler (anahtar kelimelerin ilişkilerle uyumu ve doğruluğu), zihin haritası hazırlama süreci (öğrencinin süreç içerisindeki çabası) ve görsel sunum (haritanın okunabilirliği ve estetik durumu).

5.2. Üstün ve Sınırlı Yönleri

Zihin haritaları, beynin doğal çalışma şekline uygunluğu sayesinde yaratıcılığı ve fikir üretimini büyük ölçüde destekler. Bilgiyi tek bir görselde oluşturarak bütünsel olarak sunduğundan hatırlamayı kolaylaştırır. Her bireyin kendi düşünce yapısına göre haritasını şekillendirmesine olanak tanıyan esnek ve kişiye özel yapısı, onu etkili bir araç haline getirir. Zihin haritaları, karmaşık bilgileri özetlemek ve önemli noktaları vurgulamak için ideal bir not alma ve özetleme aracıdır. Ayrıca, sürecin eğlenceli ve yaratıcı olması öğrencinin ilgisini çekerek güdülenmeyi artırır ve farklı seçenekleri görmeyi sağlayarak problem çözme ve karar verme süreçlerini destekler.

Zihin haritalarının kullanımında bazı dezavantajlar mevcuttur. Özellikle kavram haritaları gibi standartlaştırılmış bir puanlama sistemi uygulanmadığından puanlamada öznellik görülebilir. Zihin haritası tekniğinin uygulanması süreç almakta ve bu durum ilk başlarda öğrenme-öğretme sürecini zorlaştırabilmektedir. Yine çok soyut veya çok detaylı konular için etkili bir zihin haritası oluşturmak zaman ve çaba gerektirebileceğinden, bu tür karmaşık konular için zihin haritasını uygulamak zorlayıcı olabilir. Zihin haritası tekniği her konuya her zaman uygun olmayabilir; bazı konular daha doğrusal bir sunum gerektirirken, zihin haritası formatına daha az elverişli olabilir. Son olarak, öğrencilerin konudan uzaklaşıp çok fazla detaya girerek ana fikirden sapma riski olan "gereksiz dal" oluşturma durumu oluşturabilir.

6. KELİME İLİŞKİLENDİRME TESTİ (WORD ASSOCIATION TEST)

Kelime İlişkilendirme Testi (KİT), öğrencilerin belirli bir kelimeye veya kavrama yönelik çağrışım yaptığı diğer kelimeleri veya kavramları belirlemek ve bunlar arasında ilişkileri ortaya koymak amacı ile kullanılır (Bahar & Özatlı, 2003). Daha basit bir ifade ile KİT öğrencilerin bilişsel yapısını; bu yapı içindeki kavramları, yine bu kavramlar arası ilişkileri ve bu ilişkilerin anlamlı olup olmadığı tespit etmek amacı ile kullanılan bir tekniktir. Aşağıda sırasıyla KİT'in amacı, hazırlanışı, puanlama süreci, avantajları ve dezavantajlarından bahsedilmektedir.

KİT'in temel amacı, bireyin bir kelimeyle hangi kelimeleri ilişkilendirdiğini inceleyerek onun sahip olduğu anlam ağlarını ve kavramsal yapısını anlamaktır. Bu sayede bilginin zihninde nasıl organize edildiği bizlere sunmaktadır. Bunun yanında, bilginin beyinde nasıl depolandığı, işlendiği ve geri çağrıldığına dair ipuçları sunarak öğrenme ve bellek süreçlerini incelemek için de kullanılabilir. Yine bireylerin verilen bir kelimeye ne kadar çeşitli ve özgün çağrışımlar yapılabildiğini gözlemleyerek yaratıcılığı ve zihinsel esnekliği ölçmeyi hedefler. Ayrıca bireylerin belirli bir konuya ilişkin kavramları ve bu kavramlar arasındaki ilişkileri ölçerek ilgili konudaki bilgi düzeyini saptamak için de kullanılabilir. Bu doğrultuda hem öğretmen hem de öğrencilere öğrenme-öğretme sürecine rehber olabilir.

6.1. Hazırlanması ve Puanlaması

KİT'i hazırlarken öncelikle ilgili konunun temelini oluşturan 5-10 arası anahtar kavram belirler. Ardından her bir anahtar kavram 1 sayfaya gelecek şekilde bir test uygulaması hazırlanır. Burada dikkate edilmesi gereken husus her bir anahtar kavram her sayfada alt alta 10 defa yazılması ve öğrencilere her sayfayı doldurmak için belirli bir süre verilmesidir. Daha açık ifade ile öğrencilere 30 saniye verilerek her sayfadaki anahtar kavramla ilgili aklına gelen kelime yazması beklenir. Bu süre ilköğretim seviyesindeki öğrenciler için daha da uzatılabilir. Elde edilen sonuçları analiz etmek için anahtar kavramlara verilen cevap kelimelerinin çeşitleri ve sayılarının yanında bu anahtar kavramlara verilen ortak cevap kelimelerin sayısına yönelik frekans tablosu oluşturulur. Ve ardından bu frekans tablosundaki veriler kullanılarak kavram ağı haritası oluşturulur. Haritalama işleminde ise Bahar ve arkadaşları (1999) tarafından geliştirilen Kesme Noktası (KN) kullanılabilir.

KİT'in hazırlaması, frekans tablosunun oluşturulması ve haritalama sürecini daha iyi anlamak için Bahar ve arkadaşları (2014) tarafından hazırlanan çevreye ilişkin bilişsel yapılarının KİT örneğinin bir kısmını alıp buraya uygun şekilde uyarlayalım. Tablo 1 KİT tekniğinde öğrencilere 5 ayrı sayfada uygulanan 5 kavramın (hava kirliliği, su kirliliği, toprak kirliliği, erozyon ve kentleşme) 3'ü örnek olarak yer almaktadır.

Tablo 1

KİT Tekniğinde Uygulanan Sayfa Düzenleri Örneği

1. sayfa		2.sayfa		3.sayfa	
Hava kirliliği	……	Su kirliliği	……	Toprak kirliliği	……
Hava kirliliği	……	Su kirliliği	……	Toprak kirliliği	……
Hava kirliliği	……	Su kirliliği	……	Toprak kirliliği	……
Hava kirliliği	……	Su kirliliği	……	Toprak kirliliği	……
Hava kirliliği	……	Su kirliliği	……	Toprak kirliliği	……
Hava kirliliği	……	Su kirliliği	……	Toprak kirliliği	……
Hava kirliliği	……	Su kirliliği	……	Toprak kirliliği	……
Hava kirliliği	……	Su kirliliği	……	Toprak kirliliği	……
Hava kirliliği	……	Su kirliliği	……	Toprak kirliliği	……
Hava kirliliği	……	Su kirliliği	……	Toprak kirliliği	……

Öğrencilerden alınan cevaplardan sonra bu beş kavram ile öğrencilerin her bir kavrama verdikleri cevapların bir kısmının arasındaki frekans tablosu Tablo 2'de gösterilmektedir.

Tablo 2

KİT'te Anahtar Kavramlara İlişkin Cevap Kelimeleri Arasındaki Frekans Tablosu

	Anahtar Kavramlar				
Cevap Kelimeler	Hava Kirliliği	Su kirliliği	Toprak kirliliği	Erozyon	Kentleşme
Akciğer	17	-	-	-	9
Bina yığını	9	8	8	9	20
Duman	21	-	-	-	9
Fabrika	9	8	7	2	9
Gecekondu	6	7	2	1	21
Hastalık	9	18	3	-	8
İklim değişikliği	5	3	1	1	2
Kentleşme	25	11	-	7	-
Solunum	23	-	-	-	8
Trafik	13	1	2	-	12

Tablo 2'deki frekans tablosu kullanılarak oluşturulan kavram ağı haritası ise Şekil 7'de gösterilmektedir.

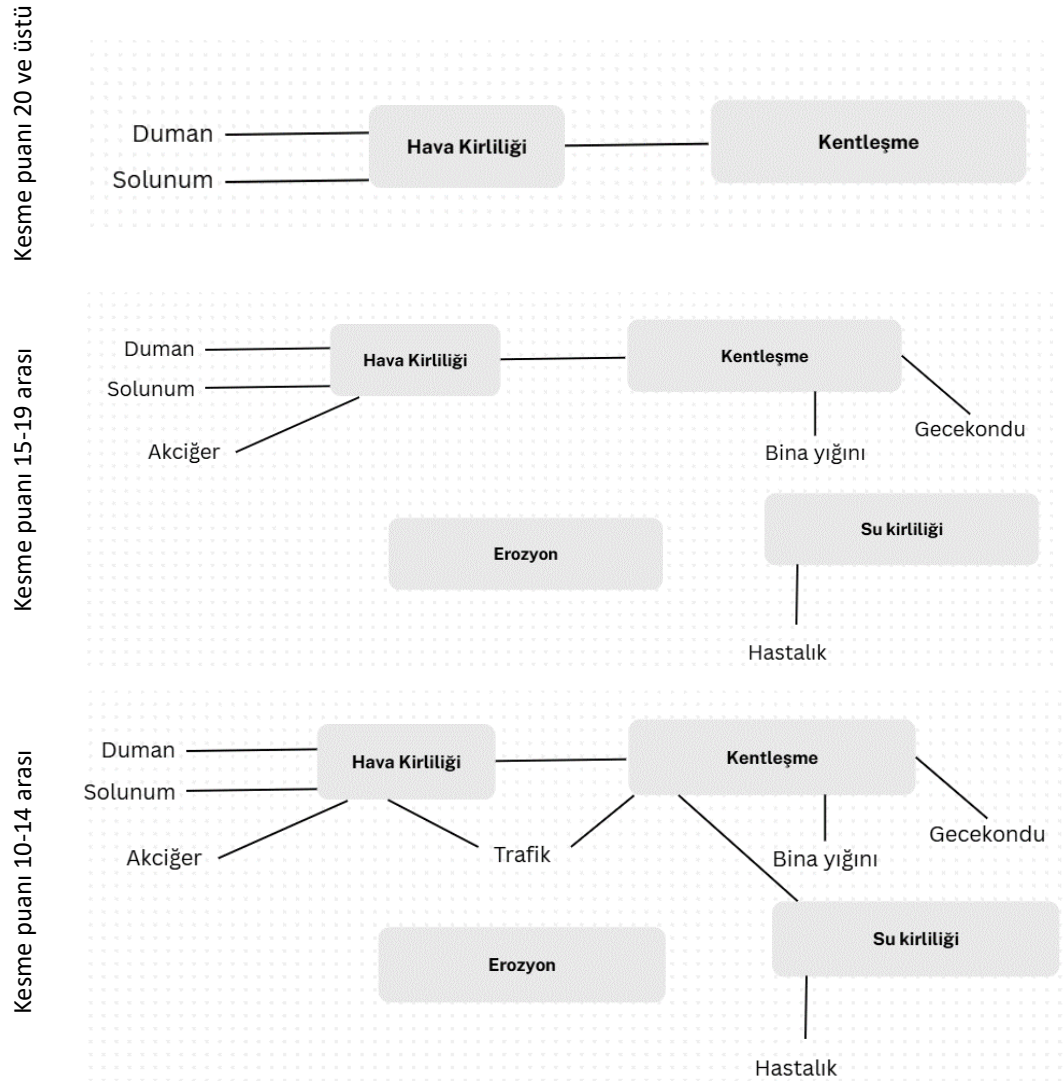
Kelime ilişkilendirme testlerinin puanlanması, testin amacına bağlı olarak hem nicel ve hem de nitel yöntemlerle puanlanabilir. Nicel puanlamada, ilk olarak öğrencilerin anahtar kavramlara uygun cevap kelimelerin sayısı olabilir. Burada öğrencinin "hız" veya "üretkenlik" becerileri ön plana çıkabilir. Ardından, üretilen kelimelerin anahtar kelimeyle ne kadar ilişkili olduğuna bakılabilir. Burada dikkat edilmesi gereken husus ise öğrencilerin cevaplarının doğruluk ve ilişki düzeyini önceden belirlenmiş bir kelime listesi üzerinden test edilebilmesidir. Ayrıca, herhangi bir öğrencinin ürettiği kelimeleri diğerlerinin ürettiğinden ne kadar farklı olduğu incelenerek özgünlük derecesi ölçülür. Nitel puanlama ise, öğrencilerin ürettiği kelimelerden oluşan anlam bütünlüğünü ve bu kelimelerin bir araya gelerek ne tür bir anlam çıkardığı analiz edilebilir. Bu bağlamda, üretilen kelimelerdeki tekrar eden temalar ve ilişkiler incelenerek kavramsal örüntüler belirlenir. Bunun için öğrencilerin belirli kelimelerin neden yazdığına yönelik açıklama istenebilir.

6.2. Üstün ve Sınırlı Yönleri

KİT kısa sürede tamamlanabilen, hızlı ve kolay uygulanabilir olmalarıyla dikkat çekmektedir. Bu sebeple farklı yaş grubuna da uygulanabilmektedir. Genellikle kağıt-kalem veya ücretsiz online araçlarla yapılabildiği için düşük maliyetlidir. KİT ile öğrenciler doğal çağrışım becerilerini kullanır ve yaratıcılık becerilerini geliştirir. Ayrıca, öğrencilerin bir konu hakkında ne kadar derinlemesine bilgi sahibi olduğunu ve bu bilgiyi nasıl ilişkilendirdiğini göstererek öğrenme yapısını anlamaya yardımcı olur.

Şekil 7

KİT Kavram Ağı Haritası



KİT'in kullanılmasının en önemli dezavantajı ise özellikle nitel değerlendirmeler için puanlamada öznelliğe yol açabilir ve farklı değerlendiriciler arasında tutarsızlıklara neden olabilir. Nicel olarak tam bir standartlaştırma yapmak her zaman kolay olmayıp, üretilen kelimelerin anahtar kavramlarla ilişki derecesini belirleme hususunda standartlaştırma zor olabilir. Kelimelerin anlamları ve çağrışımları kültürel ve dilsel farklılıklardan etkilenebileceğinden, testin farklı kültürlerde uygulanması durumunda farklılıklara sebep olabilir. Son olarak, KİT sadece kelime düzeyinde bilgiye odaklandığından, daha karmaşık dilbilgisel yapıları veya derinlemesine analizleri ölçmede sınırlılık oluşturabilir. Diğer bir ifade ile üst bilişsel becerilerin ölçülmesinde yeterli olmayabilir.

7. EV ÖDEVİ (HOMEWORK)

Ödevler eğitimde öğrenmeleri zenginleştirme, pekiştirme ve kalıcılığı artırmak için kullanılan bir yöntemdir ve ders dışı faaliyetler arasında gösterilmektedir. Öğrencilerin genellikle ev ya da kütüphane ortamında gerek bireysel gerekse grup olarak yaptıkları öğrenme etkinlikleridir. Ödevler her ne kadar daha çok öğrencinin okulda işlediği konuyu tekrar etme ve pekiştirme amaçlı kullanılsa da (Yar Yıldırım, 2018), işlenecek konunun öncesinde araştırma ve okuma faaliyetleriyle öğrencilerin hazır bulunuşluğunu artırmaya katkı sunmaktadır (Yakar, 2019). Bu bağlamda ödevin verilme amacına göre hazırlanışı, verilmesi, uygulanması ve puanlanma süreci değişim göstermektedir. Bunun sonucunda ödevin güvenirlik ve geçerliği etkilenmektedir.

Ev ödevlerinin başlıca amaçları arasında öğrencilerin geliştirilmesi gereken yaşam becerilerine katkıda bulunmak ve öğretmenlerin sınıf içi uygulamalarını yönlendirmelerine yardımcı olmak yer alır. Bu amaçlardan hareketle ev ödevlerine öğrencinin öğrendiklerini pekiştirmek ve derse karşı hazır bulunuşluğunu artırmak için başvurulabilir. Öte yandan ev ödevleri, öğrenciye sorumluluk bilincini kazandırmak, araştırma ve bilgiye ulaşma yollarını keşfetmek, ders dışı zamanının etkili kullanmak ve öğrencinin ders dışındaki öğrenme süreci hakkında geri bildirim vermeye destek olur. Ramdass ve Zimmerman (2011) ödevlerin öğrencilerin önemli yaşam becerilerinden öz düzenleme becerilerini geliştirdiğini belirtmektedir. Öğrenciler ödev tamamlama sürecinde kendilerini güdüleyerek, dikkat dağıtıcı ögeleri engelleyerek, ödevlerini tamamlamak için stratejiler kullanarak, zamanı yöneterek, hedef belirleyerek, performansları hakkında öz değerlendirme yaparak ve tatmini erteleyerek öz düzenleme yapmaktadırlar. Yine öğretmenler ödevlerle verdikleri geri bildirimlerle öğrencilerin güçlü ve zayıf yönlerini görmekte ve bu geri bildirime uygun olarak öğretmenler öğretim yöntemini gözden geçirerek öğrencin gelişmelerine önemli katkılar sunabilmektedir.

7.1. Hazırlanması ve Puanlanması

Ev ödevleri hazırlanırken geçerliği ve güvenirliği sağlamak amacıyla dikkat edilmesi gereken noktalar bulunmaktadır. Ev ödevleri, kazandırılması gereken öğrenme çıktısına ve becerilere uygun şekilde hazırlanmalıdır. Yine verilen ödevler öğrencinin sınıf ve yaş düzeyine, hazır bulunuşluk düzeyine ve ilgisine uygun olmalıdır. Aksi durumda öğrencinin başarısızlık duygusunu artırmaktadır. Bir diğer husus ise ödev içeriği açık ve net şekilde açıklanmalı, öğrencinin anlamakta zorlanacağı anlatımlar yer almamalıdır. Özellikle okul dışındaki zaman yönetimini göz önüne alınarak ödev tamamlama süresi makul şekilde planlanmalıdır. Öğrencilere farklı becerileri kazandırmak için ödevler sadece seçmeli ve yazmalı sorular olarak değil aynı zamanda araştırma, sunum, deney yapma gibi farklı özelliklerde verilmelidir. Öğrencilerin yaptıkları ödevler nesnel biçimde ve belirli bir zaman içerisinde değerlendirilmeli ve öğrenciye kısa sürede geribildirim verilmelidir.

Ödevlerin puanlanması özellikle güvenirlik bakımından çok önemlidir. Ödevler yazma ya da araştırma gerektiren etkinlikler içeriyorsa, puanlamada dereceli puanlama anahtarları (scoring rubric) kullanılabilir. Puanlamada; içeriğin doğruluğu ve yeterliliği, ödevin anlaşılır ve düzenli olması, zamanında teslim edilmesi, öğrencinin kendi ifadelerini kullanması, özgünlüğü gibi ölçütler kullanılmalıdır.

Tablo 3

Geometrik Cisimler Konusuyla İlgili Örnek Ev Ödevi

SORULAR	**CEVAPLAR**
1. SORU: Kenar uzunluğu 5 cm olan bir küpün hacmini bulunuz.	
2. SORU: Yarıçapı 2cm, yüksekliği 7 cm olan bir silindirin alanını bulunuz.	
3. SORU: Günlük hayatımızda geometrik cisimlere uygun örnekler listeleyeniz ve bu cisimlerin hangi geometrik cisme benzediğini ifade ediniz. (5 cisim örneği yeterlidir.)	

7.2. Üstün ve Sınırlı Yönleri

Öğrencilere verilen ev ödevleri öğrencilerin ve öğretmenlerin yanında velilere de yarar sağlamaktadır. Öğrenci açısında bakıldığında, öğrenilen bilgilerin kalıcılığını ve anlamlılığını artırır; öğrencinin sorumluluk, disiplin becerisi, öz değerlendirme, öz düzenleme, planlama ve zaman yönetimini gibi becerilerin kazanmasına katkı sağlayabilir. Öğretmen açısından bakıldığında öğrencinin güçlü ve geliştirilmesi yönlerini keşfederek etkili bir geribildirime katkı sunabilir. Ödevler velilerin öğrenme sürecine katılımı konusunda etkin bir rol oynar. Ödevler sayesinde veliler sınıf içi öğrenme sürecinde çocuğunun neleri başarabildiğini ve hangi konularda desteklenmesi gerektiğini fark ederler.

Ödevlerin avantajları olduğu gibi dezavantajları vardır. Özellikle kapsamlı ve uzun zaman alan ve ödevlerinin öğrencilerde bıkkınlık, isteksizlik yarattığı, öğrencilerde strese sebep olduğu belirtilmiştir (Galloway ve ark., 2013). Öğrencinin öğrenme düzeyinin üstünde verilen ev ödevlerinde öğrenciler aile desteği almakta, bu durum öğrencilerin hak etmedikleri puanları almasına sebep olmaktadır. Yine ödevlerin içerik kalitesi ve öğrencinin sadece bunun yanında dersin evde tekrarına dayanan ödevler ise öğrencinin analitik ve yaratıcı düşünme sürecinden uzaklaştırmasına yol açmaktadır.

8. EVDE SINAV (TAKE HOME EXAM)

Uygulama bakımından ev ödevlerine benzeyen evde sınav öğrencilerden belirli bir zaman diliminde (bazen birkaç saat, bazen birkaç gün ya da hafta olabilir) sınıf dışında (ev ve kütüphane gibi) verilen görevi tamamlanması beklenmektedir. Genellikle yükseköğretim düzeyinde kullanılan evde sınavlar öğrencileri daha fazla zaman

tanıyarak onların daha gerçek hayata yakın, karmaşık ve araştırma gerektiren problemlerin çözümüne yöneltmektedir. Bengtsson (2019) yaptığı sistematik çalışmada evde sınavların yüksek düzeyli taksonomi soruları içerdiğinden öğrenciler için gerçek bir öğrenme etkinliğine dönüştüreceğini savunmaktadır.

8.1. Hazırlanması ve Puanlanması

Evde sınavların hazırlanışında, öğrencilerin bilgiyi derinlemesine analiz etmelerini, araştırma yapmalarını ve eleştirel düşünme becerilerini kullanmalarını teşvik edecek, açık uçlu ve genellikle daha karmaşık görevler içeren sorular tasarlanmalıdır. Bu aşamada, sınavın süresi, erişilebilecek kaynaklar (kitaplar, makaleler, internet vb.) ve teslim formatı gibi tüm detaylar net bir şekilde belirtilerek öğrencilere rehberlik edilecek küçük bir yönerge hazırlanmalı ve intihal gibi durumların önüne geçilmesi için gerekli uyarılar yapılmalıdır.

Puanlama sürecinde ise ev ödevindeki bir puanlama anahtarı (rubrik) kullanılmalıdır. Bu puanlama anahtarında çalışmanın içeriğinin doğruluğu ve derinliği, argümanların gücü, kullanılan kaynakların uygunluğu ve doğru atıfları, alıntılanma stili (APA, MLA gibi), bilginin sentezlenerek sunulma biçimi, dilin akademik uygunluğu ve genel yapının (çerçevenin) uygunluğu gibi ölçütler yer almalıdır. Bu sayede hem değerlendirme sürecinde objektiflik ve tutarlılık sağlanır hem de öğrencilerin kendi öğrenme süreçleri hakkında yapıcı geri bildirimler alarak gelişimleri desteklenmektedir.

Evde sınav uygulamaları, öğrencilerin sınıf ortamı dışında verilen bir görevi belirli bir zaman diliminde tamamlamasına dayanan alternatif bir değerlendirme türü olarak kullanılmaktadır. Bu yöntemde öğrencilerden, araştırmaya dayalı, kapsamlı ve analitik nitelikte cevaplar oluşturmaları beklenir. Evde sınav tekniğinin kullanımına ilişkin örnek Tablo 4'de verilmiştir.

Tablo 4

Evde Sınav Uygulamasına İlişkin Hukuk Senaryo Örneği

Bir kat mülkiyetine sahip bir binada, bağımsız bölüm malikleri A ve B kişileri arasında ortak alanların kullanımı konusunda bir uyuşmazlık ortaya çıkmıştır. B kişisi, binanın giriş katında bulunan ortak depolama alanının, yönetim planında belirtilen kullanımı dışında A kişisi tarafından özel kullanım alanı haline getirildiğini iddia etmektedir. A kişisi ise bahsedilen depo alanının yıllardır tüm maliklerin bilgisi dâhilinde kendisi tarafından kullanıldığını, bu nedenle fiili durumun kazanılmış hak oluşturduğunu savunmaktadır. Bu uyuşmazlık, yönetim planındaki hükümler, Kat Mülkiyeti Kanunu'nda ortak alanların hukuki çerçevesine uygun düzenlemeler ve meşru kullanım sınırlılıkları kapsamında değerlendirilmelidir.

Aşağıdaki sorular, öğrencinin kanıt inceleme**,** mevzuat uygulama**,** eleştirel bir şekilde hukuki çözümleme becerisini ölçmeyi amaçlayan, açık uçlu maddeler şeklinde tasarlanmıştır.

1. Soru: Verilen senaryoya göre ortak alanların hukuki niteliğini Kat Mülkiyeti Kanunu hükümlerine dayanarak açıklayınız. Depolama alanının hukuki statüsünü değerlendirin.

2. Soru: A'nın ileri sürdüğü "fiili kullanımın kazanılmış hak doğurduğu" yönündeki iddiasını hukuki açıdan inceleyiniz. Bu iddia ortak alanlar bakımından geçerliği var mıdır? Gerekçeleriyle birlikte açıklayınız.

3. Soru: B'nin iddia ettiği "yönetim planına aykırılık" iddiasını değerlendirin. Yönetim planının kat malikleri açısından bağlayıcılığı nedir? İlgili madde veya hükümler üzerinden açıklayınız.

4.Soru: Benzer bir durumun yeniden yaşanmaması için bina yönetimine önerilebilecek hukuki ve idari önlemleri yazınız.

8.2. Üstün ve Sınırlı Yönleri

Evde sınavların üstün yönleri arasında verilen bir konunun öğrenciler tarafından derinlemesine incelenmesidir. Bu süreç aynı zamanda onların eleştirel düşünme ve araştırma becerilerini geliştirmelerine katkı sağlamaktadır. Ayrıca evde sınavlar, öğrencilerin yazılı ifade becerilerinin gelişmesine yardımcı olurken, aynı zamanda öğrencilere zaman yönetimi ve öğrenme süreçlerini planlama sorumluluğu kazandırır. Farklı türde akademik kaynakları bulma ve kullanma konusunda deneyim kazandırması da önemli bir diğer olumlu yönüdür. Sınıf içi sınav baskısının ve stresinin bulunmaması, öğrencilerin daha sakin ve rahat bir ortamda performans göstermelerine olanak tanır.

Bununla birlikte evde sınavların bazı sınırlı yönleri de vardır. En büyük dezavantajlardan biri, öğrencilerin başkalarının çalışmalarını kendi fikirleriymiş gibi sunmalarına neden olabilecek intihal riskidir; bu nedenle güçlü kontrol mekanizmaları gerektirir. Ayrıca öğrenciler arasında aşırı yardımlaşma ya da kopya çekme gibi işbirliği sorunları yaşanabilir ve tüm öğrencilerin eşit kaynaklara erişememesi fırsat eşitsizliklerine yol açabilir. Özellikle öğretmenler için çok sayıda evde sınav değerlendirilmesi zaman alıcı olabilir. Yine soruların net hazırlanmadığı, sürenin yeterli verilmediği veya puanlama kriterlerinin belirsiz olduğu durumlarda yöntemin verimliliği düşebilir.

9. YANSITICI GÜNLÜK (REFLECTIVE JOURNAL)

Eğitim–öğretim sürecinde yenilikçi ve esnek bir yöntem olarak kullanılan yansıtıcı günlükler; öğrencilerin öğrenme süreçlerine ilişkin düşüncelerini görünür kılabildikleri, belirli bir konuya dair duygu ve görüşlerini açık bir şekilde ifade ederek yazılı hâle getirdikleri kişisel kayıt araçları şeklinde de nitelendirilebilmektedir (Akkoyunlu ve ark., 2016; Dummer ve ark., 2008). Bu açıklamalar, yansıtıcı günlüklerin gündelik yaşamda tutulan klasik günlüklerden farklı bir işlev ve yapıya sahip olduğunu göstermektedir (Arslan, 2017). Moon (2007), günlüklerin çeşitli amaca hizmet edebileceğini vurgulayarak bunları; yaşantıların kaydedilmesi, deneyimden öğrenmenin artırılması, anlamlandırmayı ve anlamın ortaya konmasını destekleme, eleştirel düşünmeyi geliştirme, üstbilişsel farkındalığı artırma, aktif katılımı teşviği artırma, yansıtma becerisini ilerletme, problem çözmeyi geliştirme, kişisel gelişime katkı sağlama, davranış değişimini destekleme ya da tedavi edici amaçlarla kullanma, yaratıcılığı besleme, yazma becerisini ve kendini ifade etme kapasitesini iyileştirme, grup içi iletişimi artırmak, araştırma veya projelerin planlanması ve yürütülmesine yardımcı olma gibi çok çeşitli işlevlerle açıklamıştır. Yansıtıcı metinler Moon'un (2009) sınıflandırmasına göre dört düzeyde ele alınmaktadır: tanımlayıcı yazma, kısmen yansıtıcı özellik taşıyan tanımlayıcı yazma, yansıtıcı yazma ve ileri düzey yansıtıcı yazma. Hubbs ve Brand'a (2010) göre yansıtıcı günlük, bireyin inançları, deneyimleri ve değerleriyle ilgili içsel düşünme süreçlerini belirli bir düzen içinde kayda geçirmesine sağlayan bir araç olarak tanımlanmaktadır.

Yansıtıcı günlüklerin temel amacı Malthouse ve Roffey'e (2013) göre, öğrencilerin kendi öğrenme süreçlerine ilişkin deneyimlerini süreç odaklı biçimde değerlendirmelerine olanak tanımaktır. Bu sayede bireyler edindikleri deneyimleri bir sonraki öğrenme aşamalarında etkili biçimde kullanabilir, böylece akademik gelişimlerini daha bilinçli şekilde planlayabilirler (Akt. Arslan, 2017).

9.1. Hazırlanması ve Puanlanması

Yansıtıcı günlüklerin oluşturulma süreci, Raw ve arkadaşlarının (2005) belirttiğine göre üç aşamada gerçekleşmektedir. İlk aşamada, bir görev sırasında hoş olmayan ya da rahatsızlık veren bir durumun fark edilmesi söz konusudur. İkinci basamak, söz konusu

görevin çözümlenmesi, üçüncü adım ise görevi başarıyla yerine getirebilmek için yeni düşünceler geliştirilmesidir. Bu üçlü süreç her yazım döngüsünde yinelenmekte olup, bu yolla öğrenenin kendi öğrenme sürecine ilişkin düşünme ve değerlendirme yapması desteklenmektedir. Bu tür günlüklerin etkili biçimde kullanılabilmesi için öğrencilere önceden belirli bir yönlendirme ve hazırlayıcı eğitim verilmesi gerekmektedir. Böyle bir hazırlık, yansıtıcı günlüklerin daha güvenilir ve anlamlı bilgiler sunmasını sağlamaktadır. Bu doğrultuda, yansıtıcı günlük oluştururken dikkat edilmesi gereken temel noktalar aşağıdaki gibi ifade edilmektedir:

1. Öğrenme sürecinde karşılaşılan bir aksaklık ya da engeli fark etme,
2. Gerçekleştirilen göreve ilişkin performansı inceleyip çözümleme,
3. Belirlenen öğrenme sürecindeki performansı geliştirebilmek için uygun stratejiler planlama (Raw ve ark., 2005).

Bu aşamalar yansıtıcı günlük uygulamasının temelini oluşturmaktadır. Moon'a (2009) göre, günlüklerde yer alan değerlendirmeler aracılığıyla, öğrenicilerin kendi öğrenme süreçlerine yönelik farkındalık düzeylerinin belirlenebileceğini ifade eder ve bu durumu yansıtıcı yazma düzeyleri olarak adlandırır. Bu yazma düzeyleri aşağıdaki biçimde sınıflandırılmaktadır (Akt. Melanlıoğlu & Demir, 2016):

1. Düzey: Tanımlayıcı yazma: Yazma üretimleri sınırlı miktarda yansıtma içerir. Öğrenme sürecine ilişkin değerlendirme yapmak yerine, süreçte izlenen adımlar sıralanmaktadır.

2. Düzey: Az yansıtma içeren tanımlayıcı yazma: Yansıtma oranı düşük olan üretimlerdir. Genellikle soru–cevap biçiminde kısa açıklamalar bulunur. Öğrenme sürecine ilişkin analiz yeterli değildir.

3. Düzey: Yansıtıcı yazma: Öğrenici sürece dair bir görüşe sahiptir ancak bu görüş tüm süreci kapsamayıp ilgi veya dikkat gibi nedenlerle belirli aşamalara odaklanmaktadır. Bu aşamalara ilişkin çözümlemeler yapılmakta, yansıtıcı değerlendirmeler bulunmakta ve farkındalık gelişmeye başlamaktadır. Ancak bu farkındalık tüm süreci kapsayacak düzeyde değildir.

4. Düzey: İleri düzey yansıtıcı yazma: Öğrenme sürecini izlemiş, süreç içerisinde karşılaşılan sorunlara ilişkin alternatif çözüm yolları üretmiş ve buna yönelik kapsamlı değerlendirmeler yapmış metinlerdir. Derin bir yansıtma mevcuttur ve üstbilişsel stratejiler etkin biçimde kullanılmaktadır. Mevcut bilgi ve deneyimler, yeni öğrenmelere yön vermektedir.

Yansıtıcı metinler Moon'un (2009) sınıflandırmasından yola çıkılarak yazılar belirlenen yansıtma seviyelerine göre söz konusu dört kategoriden birine yerleştirilmiş ve 1'den 4'e kadar puanlanmıştır. Bu puanlama 1. Düzeye göre sınırlı üretim içerdiğinden 1.00–1.75**,** 2. Düzeye göre yüzeysel açıklamalar içerdiği için 1.76–2.50**,** 3. Düzeye göre kısmi

farkındalık ve analiz barındırdığından 2.51–3.25, son düzeyde ise derinlemesine çözümleme ve üstbiliş stratejilerinin kullanımını içerdiğinden 3.26–4.00 puan aralığında değerlendirilmektedir. Bu yapı, öğrencilerin yansıtıcı düşünme düzeylerinin sistematik biçimde ölçülmesini mümkün kılmaktadır.

Tablo 5'de 10. sınıf kimya dersinde işlenen *"Asit–Baz Tepkimelerinin Gözlemlenmesi"* etkinliğine yönelik hazırlanan yansıtıcı günlük maddeleri örnek olarak sunulmuştur.

Tablo 5

Asit–Baz Tepkimelerinin Gözlemlenmesi Konusu İle İlgili Yansıtıcı Günlük Örneği

Öğrenci Adı ve Soyadı:

Tarih:

Deneyin Adı: Asit-Baz Tepkimelerinde Gözlenebilir Değişimler

Kazanım: Kimyasal tepkime türlerinin oluşum sürecine ilişkin bilimsel gözlem yapabilme (Kim.10.1.3.)

1. Bu etkinlikten neler öğrendim?
2. Deney plan sürecine tamamen uyabildim mi, değişiklik yapmak durumda kaldığım noktalar oldu mu?
3. Deney esnasında hoşuma giden yerler (örneğin; Renk değişimi gözlemi, deney düzenekleri) gibi neler olabilir?
4. Deney esnasında zorlandığım yerler (örneğin; Ph değerlerini yorumlama, ölçüm yapma, güvenli çalışma adımları gibi) neler olabilir?
5. Benzer deneylerde daha etkili performans sergileyebilmek için kişisel öğrenme sürecimi nasıl düzenlemeyi, geliştirmeyi ve planlamayı düşünüyorum?

9.2. Üstün ve Sınırlı Yönleri

Yansıtıcı günlüklerin eğitim sürecine sağladığı üstünlükler Moon (1999a) belirttiği amaca hizmet ettiği ifadeler; öğrencilerde eleştirel düşünmeyi teşvik ederek öğrenmenin niteliğini derinleştirmesidir Öğrencilerin kendi öğrenme süreçlerini anlamalarına imkân tanıması, öğrenmenin sahiplenilmesini ve sürece aktif katılımı artırması bakımından da önemli bir avantajdır. Ayrıca yansıtıcı günlükler, öğrencinin mesleki uygulamalarını geliştirmesine, kişisel gelişim sürecinde öz değerlendirme becerileri kazanmasına ve sezgisel düşünmeyi daha etkili kullanarak yaratıcılığını geliştirmesine katkıda bulunmaktadır. Bunun yanında, yazılı ifade becerilerini güçlendirmesi ve öğrencilerin kendilerini konuşma yoluyla ifade etmekte zorlandıkları durumlarda alternatif bir anlatım kanalı sunması da bu yöntemin diğer olumlu yönleri arasındadır. Ek olarak, grup içi etkileşimi desteklemesi ve yansıtıcı bir öğrenme kültürü oluşturması, yöntemin pedagojik değerini artıran unsurlardır (Akt. Can & Altuntaş, 2016).

Yansıtıcı günlük uygulamasında çeşitli sınırlılıklar da gözlenmektedir. Öğrenciler, yazma sürecini çoğu kez hem zaman gerektiren hem de zorlayıcı bir süreç olarak tanımlamakta ve bu durum yöntemin sürdürülebilirliğini olumsuz etkileyebilmektedir (Kok & Chabeli, 2002; Can & Altuntaş, 2016). Özellikle derslerin sonunda kısa bir süre içerisinde günlük yazma beklentisi, uygulamanın yorucu**,** tekrarlayıcı ve motivasyonu düşüren bir niteliğe bürünmesine yol açmaktadır. Kok ve Chabeli'nin (2002) belirttiği üzere, günlük yazma sürecine ilişkin beklentilerin açık biçimde ifade edilmemesi, öğrenciler açısından belirsizlik oluşturmakta ve yazma sürecine yönelik istekliliği azaltabilmektedir. Bu sebeple, yansıtıcı günlüklerin etkili ve verimli şekilde kullanılabilmesi için öğrencilere sürecin nasıl yürütüleceğine ilişkin açık ve anlaşılır şekilde yönlendirme sağlanması, yazma sürecinin uygun bir zaman aralığına yayılması ve değerlendirmenin hangi ölçütlere göre yapılacağının net biçimde ifade edilmesi önem taşımaktadır.

10. ÖZ DEĞERLENDİRME (SELF-ASSESSMENT)

Öz değerlendirme**,** öğrencilerin kendi öğrenmelerini ve performanslarını değerlendirme süreci olarak tanımlanmaktadır. Bu yöntemde birey, kendi öğrenme sürecini veya ürünlerini gözden geçirerek güçlü, zayıf ve eksik yönlerini belirlemeye çalışır. Öz değerlendirme genellikle performans çalışmaları, projeler ve portfolyolar gibi öznel ölçme araçları üzerinden yürütülmektedir.

Bireyin kendi öğrenmesini değerlendirme sürecine aktif olarak katılması, onun öğrenme sorumluluğunu üstlenmesini sağlar. Böylece öğrenci, kendi gelişim sürecinde daha bilinçli ve yönlendirici bir rol üstlenir (Cheng & Warren, 1999; Palchikov & Bound, 1989; Roberts, 2006). Öz değerlendirmenin temel amacı; öğrenmeyi ve akademik başarıyı artırmak, öz-düzenleme becerilerini teşvik etmek**,** öğrencilerin kendi öğrenmelerini izleme, değerlendirme ve yönetme eğilimlerini geliştirmektir.

Son dönemlerde öğrencilerin öz değerlendirme sürecinde çeşitli sorunlarla karşılaştıkları görülmekte ve bu süreci desteklemek amacıyla yapay zekâ temelli sohbet sistemlerinden faydalanılması önerilmektedir. University of Florida CITT tarafından hazırlanan rehberde belirtildiği üzere, öğrenciler öz değerlendirme yaparken zorlandıklarında yapay zekâ destekli bir sohbet uygulamasını kullanmaları, süreci daha etkili bir biçimde yönetmelerine fayda sağlayabilmektedir. Bu tür dijital araçlar, öğrencilere kendi çalışmalarını değerlendirmelerine yardım ederek yönlendirici sorular sorarak üst düzey düşünme becerilerini geliştirmek için harekete geçirmekte ve öğrencilerin yapay zekânın sunduğu dönütleri gözden geçirip bunlara katılım seviyelerini sorgulamaları, öz değerlendirme alışkanlığı geliştirmelerine ve çalışmalarında hangi noktalarda dikkatli olmaları gerektiğini daha iyi fark etmelerine imkân sunmaktadır (University of Florida CITT, 2025).

10.1. Hazırlanması ve Puanlanması

Öz değerlendirme sürecinde aşağıdaki hususlar dikkate alınmalıdır:

1. Öz değerlendirmenin amacının ne olduğunun, nasıl uygulandığının ve sağlayacağı katkıların açıklanması,
2. Değerlendirme kriterlerinin geliştirilmesi,
3. Kriterlere göre değerlendirmenin nasıl yapılacağının açıklanması ve örnek uygulamaların yapılması.

Tablo 6

Öz Değerlendirme Formu Örneği

ÖZ DEĞERLENDİRME FORMU

Adı ve Soyadı:……

Bu form kendinizi değerlendirmek amacıyla hazırlanmıştır. Çalışmanızı dikkatli ve en doğruyu yansıtacak şekilde dolduracak şekilde seçeneğe X ile işaretleyiniz. Bu öz değerlendirme formundan alınabilecek en yüksek puan 18, en düşük puan 6'dır.

BECERİLER	DERECELER		
	Her zaman (3)	Nadiren (2)	Hiçbir zaman (1)
1. Yönergeyi izledim.			
2. Çalışmalarımda farklı materyaller kullandım.			
3. Anlamadığım yerlerde öğretmenime danıştım.			
4. Çalışmayı verilen zamana uygun olarak gerçekleştirdim.			
5. Modelime uygun şemayı oluşturdum.			
6. Model sunumumu gerçekleştirdim			

Aşağıdaki maddeyi yaptığınız proje çalışmasına göre cevaplayınız.
Bu projeden neler öğrendim?
……
……
……
……
……

10.2. Üstün ve Sınırlı Yönleri

Literatürde öz değerlendirmenin çeşitli üstünlükleri ve sınırlılıklarının olduğu Rolheiser ve Ross'a (2021) göre ifade edilmektedir.

Üstün Yönleri	Sınırlı Yönleri
• Öğrencilerin dikkatini öğrenme hedeflerine yönlendirerek başarılarını artırır, öğrenme eksiklerini fark etmelerine ve motivasyonlarının yükselmesine yardımcı olur. • Geleneksel testlerin ifade edemediği öğrencilerin içsel süreçleri, hedefleri ve başarıya ilişkin inançları hakkında öğretmene daha detaylı bilgi sağlar.	• Öz değerlendirme sürecinde öğretmenlerin değerlendirme kontrolünü öğrencilerle paylaşması zor olabilir ve öğretmen–öğrenci ilişkilerinde değişim gözlemlenir. • Hem öğretmenlerin hem öğrencilerin öz değerlendirmenin ne olduğunu anlaması ve süreci doğru uygulamayı öğrenmesi zaman alır.

11. AKRAN DEĞERLENDİRME (PEER ASSESSMENT)

Akran değerlendirme, öğrencilerin gerçekleştirdikleri etkinliklerin veya ürünlerin, önceden belirlenmiş ölçütler doğrultusunda arkadaşları tarafından değerlendirilmesi sürecidir. Bu değerlendirme türü özellikle grup çalışmaları, performans görevleri, projeler ve portfolyo uygulamalarında kullanılmaktadır. Ayrıca sözlü sunumlar, tiyatro, koro gibi performans gerektiren etkinliklerde de akran değerlendirmesi sıklıkla başvurulan bir yöntemdir. Akran değerlendirme, öğrencilerin yalnızca kendi öğrenme süreçlerine değil, aynı zamanda başkalarının öğrenmelerine de aktif biçimde katkı sağlamalarına olanak tanır. Bu süreçte öğrenciler, başkalarının çalışmalarını değerlendirirken eleştirel düşünme, empatik yaklaşım ve nesnel değerlendirme becerilerini geliştirirler.

Tablo 7

OGM Materyal Sınıf İçi Akran Değerlendirme Formu Örneği

ÖZ VE AKRAN DEĞERLENDİRME Bu form grup olarak çalışmalarınızı değerlendirmek üzere hazırlanmıştır. Aşağıdaki her ifadeyi okuyunuz. Bu ifadelere göre, önce grupdaki arkadaşlarınızı son sütunda ise kendinizi değerlendiriniz. Buna göre; **5 = Çok İyi, 4 = İyi, 3 = Orta, 2 = Kabul edilebilir, 1 = Geliştirilmesi gerekli** olarak dereceleyiniz. Değerlendirme yapan öğrencinin adı soyadı:							
Performans görevi öncesi / süreci / sonrası davranışlar	**1. Öğrencinin Adı .....**	**2. Öğrencinin Adı .....**	**3. Öğrencinin Adı .....**	**4. Öğrencinin Adı .....**	**5. Öğrencinin Adı .....**	**6. Öğrencinin Adı .....**	**Ben**
Konu ile ilgili temel kavramları bilir/bilirim.							
Deney araç-gereçlerini tanır/tanırım.							
Deneye ön hazırlık çalışmaları yapmıştır/yaptım.							
Deneyi, ilgili kavramları kapsayacak şekilde planlamıştır/planladım.							
Konu ile ilgili basit araç-gereç ve materyalleri seçmiştir/seçtim.							
Deney düzeneğini doğru bir şekilde hazırlamıştır/hazırladım.							
Deney basamaklarını doğru bir şekilde gerçekleştirmiştir/gerçekleştirdim.							
Deneyde elde edilen verileri doğru kaydetmiştir/kaydettim.							

11.1. Üstün ve Sınırlı Yönleri

Topping (2009) çalışmasında akran değerlendirmenin çeşitli avantajlarını vurgulamıştır:

- Öğrencilerin akranlarından geri bildirim almalarını destekler.
- Kavramların daha kalıcı ve kolay öğrenmelerine katkı sağlar.
- Öğrencilerin yazma becerilerinin gelişimine imkân tanır.
- Grup çalışmalarına katılımı artırarak iş birlikli öğrenmeyi geliştirir.
- Öğretmenlerin değerlendirme sürecinde zamandan tasarruf yapmalarına yardımcı olur.

Literatürde akran değerlendirmesinin avantajları kadar bazı dezavantajları da olduğu belirtilmektedir. Topping'e (1998) göre dezavantajları şu şekilde ifade etmiştir:

- Öğrenciler, akranlarıyla aralarındaki ilişki olumsuz etkilenmesi endişesiyle sorumluluk almaktan kaçınabilir ve bu durum değerlendirmeyi zayıflatabilir.
- Arkadaşlık ilişkilerinin kötü amaçlı kullanılması, sürecin objektif bir değerlendirme olmaktan çıkmasına neden olabilir.
- Akranların birbirini değerlendirmesi, ölçmenin geçerlilik ve güvenilirliğini azaltabilir.

12. GRUP DEĞERLENDİRME (GROUP ASSESSMENT)

Bu tür değerlendirme uygulamalarında, öğretmenin önceden belirlediği ve belirli bir yeterlik düzeyine sahip öğrencilerden oluşan bir grup, tüm katılımcıların çalışmalarını inceleyip puanlamaktadır. Değerlendirme süreci; çalışmaların gözden geçirilmesi, uygun ölçütlere göre değerlendirilmesi ve sonuçların düzenlenmesi aşamalarını içerir. Öğretmen ise bu süreçte yalnızca gözlemci olarak bulunur ve doğrudan müdahalede bulunmaz (Bora, 2024).

Tablo 8

OGM Materyal Sınıf İçi Grup Değerlendirme Formu Örneği

GRUP DEĞERLENDİRME FORMU-2												
Grup içi arkadaşlarınızın sosyal sorumluluk ve toplum hizmeti çalışmalarındaki tutum ve davranışlarını düşünerek aşağıdaki maddelere katılım düzeyinizi puanlama anahtarına göre işaretleyiniz:												
Puanlama Anahtarı												
1= Zayıf	**2= Kabul Edilebilir**			**3= Orta**				**4= İyi**			**5= Çok İyi**	
Öğrencinin Adı ve Soyadı:	Çalışmaya hazır oluş	Başkalarının fikirlerine saygı	Sorunları paylaşma	Grup arkadaşlarını destekleme	Tartışmalara katkı	Görüşleri gerekçelendirme	Farklı görüşlere saygı duyma	Görev ve işbirliği	Zamanı verimli kullanma	Ödevleri tamamlamama	Ödevleri saklama	Toplam Puan
1.												
2.												
3.												
4.												
5.												

12.1. Üstün ve Sınırlı Yönleri

Grup değerlendirmesinin eğitim ve öğretim sürecine sağladığı faydalar arasında, özellikle öğrencilerin işbirliği içinde çalışma alışkanlığı kazanmaları ve karşılıklı iletişim becerilerinin gelişmesi önemli bir yer tutmaktadır (Race, 2001, akt. Kutlu ve ark., 2010, s. 105). Öğrencilere ortak çalışma ortamlarının sunulması, hem süreçte hem de ürün aşamasında öğrenme fırsatı sunar. Ayrıca, öğrencinin değerlendirme sürecine dâhil olacağını bilmesi, kendisini daha sorumlu hissetmesine ve sürece olumlu yaklaşmasına yardımcı olur. Bu durum, bireyin kendi görüşlerinin dikkate alındığını hissetmesini sağlayarak, öz güven ve aidiyet duygusunu desteklemektedir.

Göçer ve Çaylı'nın (2017) çalışmalarında ifade edildiği gibi, grup değerlendirme süreci bazı dezavantajlar içermektedir. Öğrencilerin değerlendirme yaparken arkadaşlık ilişkilerinden olumsuz etkilenebileceğini düşünerek değerlendirme sonuçlarında yanlılığa neden olabileceğini vurgulamaktadır. Grup içi sosyal dinamiklerin değerlendirme sürecine doğrudan etkilemesi, ölçmenin objektifliğini zayıflatabilmektedir. Ayrıca, grup çalışmalarında sorumluluk paylaşımının eşit olmaması, bazı öğrencilerin yeterli fayda sağlamadan yüksek puan elde etmesine imkân tanımaktadır. Öğrencilerin değerlendirme becerilerinin tam olarak gelişmemiş olması da sürecin güvenilirliğini azaltan bir diğer durumdur. Belirtilen dezavantajlar grup değerlendirmenin dikkatle planlanması ve öğrencilere değerlendirme kriterlerinin açık ve anlaşılır şekilde öğretilmesi gerekliliğini ortaya koymaktadır.

13. SONUÇ

Bu bölümde, öğrenme sürecini merkeze alan farklı ölçme ve değerlendirme araçları ele alınmıştır. Geleneksel ölçme araçlarının yanında kullanılan yapılandırılmış grid, tanılayıcı dallanmış ağaç, kavram haritaları, zihin haritaları, kelime ilişkilendirme testi, ev ödevi, evde sınav, yansıtıcı günlük, öz değerlendirme, akran değerlendirme ve grup değerlendirme gibi araçların farklı amaçlara hizmet ettiği görülmektedir. Bu araçların her biri, öğrencilerin öğrenmelerini farklı yönleriyle ortaya koymakta ve öğretmenlere öğrencilerin gelişimleri hakkında daha ayrıntılı bilgi sağlamaktadır.

Öğrenme yaklaşımlarına dayalı bu araçların en önemli özelliklerinden biri, öğrencinin yalnızca ulaştığı sonucu değil, öğrenme sürecini de değerlendirmeye olanak sağlamasıdır. Öğrenciler bu süreçte kendi öğrenmelerini gözden geçirebilmekte, eksiklerini fark edebilmekte ve öğrenme sorumluluğunu daha fazla üstlenebilmektedir. Özellikle öz, akran ve grup değerlendirme uygulamaları öğrencilerin değerlendirme sürecine aktif olarak katılmalarını desteklemektedir.

Bununla birlikte, bu araçların etkili bir şekilde kullanılabilmesi için amacın, değerlendirme ölçütlerinin ve uygulama sürecinin öğrencilere açık biçimde anlatılması gerekmektedir. Özellikle öz değerlendirme ve akran değerlendirme gibi uygulamalarda

öğrencilerin süreci doğru anlamaları ve değerlendirme ölçütlerini bilmeleri önem taşımaktadır. Ayrıca her aracın kendine özgü üstün ve sınırlı yönleri bulunduğundan, kullanılacak ölçme aracının dersin amacı ve öğrencilerin özellikleri dikkate alınarak belirlenmesi gerekmektedir.

Sonuç olarak, ölçme ve değerlendirme yalnızca öğrencilerin başarılarını belirlemek için kullanıldığı düşünülemez. Ölçme sonuçları, öğrencilerin öğrenmelerini geliştirmek ve öğretim sürecini yeniden düzenlemek için de kullanılmalıdır. Bu açıdan bakıldığında, farklı ölçme araçlarının uygun biçimde kullanılması hem öğrencilerin öğrenme süreçlerini daha iyi anlamaya hem de öğretmenlerin öğretimlerini geliştirmelerine katkı sağlayacaktır. Böylece ölçme ve değerlendirme, öğrenmenin sonunda yapılan bir işlem olmaktan çıkarak öğrenmeyi destekleyen doğal bir sürece dönüşebilecektir.

KAYNAKÇA | REFERENCES

Akkoyunlu, B., Telli, E., Çetin, M. N. & Dağhan, G. (2016). Öğretmen eğitiminde yansıtıcı günlüklere ilişkin öğretmen adaylarının görüşleri. *Turkish Online Journal of Qualitative Inquiry (TOJQI), 7*(4), 312-330.

Arslan, M. M. (2017). Adayların yansıtıcı günlüklerinde öğretmenlik uygulamasına yönelik farkındalıkları. *Bartın Üniversitesi Eğitim Fakültesi Dergisi, 6*(3), 1017-1026.

Aydoğdu, M. & Kesercioğlu, T. (2005). *İlköğretimde fen ve teknoloji öğretimi*. Anı Yayıncılık.

Bahar, M. (2001). Çoktan seçmeli testlere eleştirel bir yaklaşım ve alternatif metotlar. *Kuram ve Uygulamada Eğitim Bilimleri Dergisi, 9*, 4-6.

Bahar, M., Johnstone, A. H. & Sutcliffe, R. G. (1999). Investigation of students' cognitive structure in elementary genetics through word association tests. *Journal of Biological Education*, *33*(3), 134-141.

Bahar, M., Nartgün, Z., Durmuş, S. & Bıçak, B. (2008). *Geleneksel-alternatif ölçme ve değerlendirme öğretmen el kitabı*. Pegem Akademi.

Bahar, M., Nartgün, Z., Durmuş, S. & Bıçak, B. (2014). *Geleneksel-tamamlayıcı ölçme ve değerlendirme teknikleri: Öğretmen el kitabı*. Pegem Akademi.

Bahar, M. & Özatlı, N. S. (2003). Kelime iletişim test yöntemi ile lise 1. sınıf öğrencilerinin canlıların temel bileşenleri konusundaki bilişsel yapılarının araştırılması. *Balıkesir Üniversitesi Fen Bilimleri Enstitüsü Dergisi*, *5*(2), 75-85.

Başol, G. (2013). *Eğitimde ölçme ve değerlendirme*. Pegem Akademi.

Bengtsson, L. (2019). Take-home exams in higher education: A systematic review. *Education Sciences*, *9*(4), 267.

Black, P. & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy & Practice, 5*(1), 7–74.

Bora, M. (2024). *Alternatif değerlendirme yöntemleri ışığında resim dersi kapsamında yapılmış örnek bir değerlendirme uygulaması*. Kesit Akademi

Can, R. & Altuntaş, B. (2016). Yansıtıcı günlük yazma uygulamasına ilişkin öğrenci görüşleri. *Ana Dili Eğitimi Dergisi, 4(*1), 53-63.

Cheng, W. & Warren, M. (1999). Peer and teacher assessment of oral and written tasks of a group project. *Assessment and Evaluation in Higher Education, 24*(3), 310– 314.

Dochy, F., Segers, M. & Sluijsmans, D. (1999). The use of self-, peer and co-assessment in higher education: A review. *Studies in Higher Education*, *24*(3), 331–350.

Dummer, T. J. B., Cook, I. G., Parker, S. L., Barrett, G. A. & Hull, A. P. (2008) Promoting and assessing 'Deep learning' in geography fieldwork: An evaluation of reflective field diaries. *Journal of Geography in Higher Education, 32*(3), 459-479.

Eroğlu, M. G. & Kelecioğlu, H. (2011). Kavram haritaları ve yapılandırılmış grid ile elde edilen puanların geçerlik ve güvenirliklerinin incelenmesi. *Hacettepe Üniversitesi Eğitim Fakultesi Dergisi, 40*, 210- 220.

Evrekli, E. & Balım, A. G. (2010). Fen ve teknoloji öğretiminde zihin haritası ve kavram karikatürü kullanımının öğrencilerin akademik başarılarına ve sorgulayıcı öğrenme becerileri algılarına etkisi. *Batı Anadolu Eğitim Bilimleri Dergisi*, *1*(2), 76-98.

Galloway, M., Conner, J. & Pope, D. (2013). Nonacademic effects of homework in privileged, high-performing high schools. *Journal of Experimental Education*, *81*(4), 490–510.

Göçer, A. & Çaylı, C. Ç. (2017). Türkçe eğitiminde öğrenen özerkliğine dayalı değerlendirme uygulamalarına genel bir bakış. *Karadeniz Sosyal Bilimler Dergisi, 9*(2), 121-138.

Gulikers, J., Bastiaens, T. & Kirschner, P. (2004). A five-dimensional framework for authentic assessment. *Educational Technology Research and Development*, *52*(3), 67–86.

Hubbs, D. & Brand, C. F. (2010). Learning from the ınside out: A method for analyzing. *Reflective Journals in the College Classroom. Journal of Experiential Education*, *33*(1), 56-71.

Johnstone, A. H., Bahar, M. & Hansell, M. H. (2000). Structural communication grids: A valuable assessment and diagnostic tool for science teachers. Journal of Biological Education, *34*(2), 87-89.

Kok, J. & Chabeli, M. M. (2002). Reflective journal writing: How it promotes reflective thinking in clinical nursing education: A Students' Perspective. Curationis, *25*(3), 35- 42.

Kortelainen, T. & Vanhala, M. (2004). Portfolio, peer evaluation, and mind map in an introductory course of information studies. *Journal of Education for Library and Information Science, 45*(4), 273285.

Kutlu, Ö., Doğan, C. D. & Karakaya, İ. (2010). *Öğrenci başarısının belirlenmesi performans ve portfolyoya dayalı durum belirleme* (3. Baskı). Pegem Akademi Yayıncılık.

McClure, J. R., Sonak, B. & Suen, H. K. (1999). Concept map assesment of classroom learning: Reliability, validity and logistical practicality. *Journal of Research in Science Teaching*, *36*(4), 475-192.

Melanlıoğlu, D. & Demir, T. (2016). Türkçe öğrenen yabancılar örneğinde yansıtıcı günlük kullanma ve yazma öz-yeterliği arasındaki ilişki. *Mustafa Kemal Üniversitesi Sosyal Bilimler Enstitüsü Dergisi, 13*(35).

Moon, J. (2007). Getting the measure of reflection: considering matters of definition and depth. *Journal of* Radiotherapy in Practice, *6*(4), 191-200.

Moon, J. (2009). The use of graduated scenarios to facilitate the learning of complex and difficult-to-describe concepts. *Art, Design & Communication in Higher Education, 8*(1), 57-70.

Nitko, A. J. & Brookhart, S. M. (2014). *Educational assessment of students* (7. baskı). Pearson.

Novak, J. D. & Gowin, R. (1984). *Learning how to learn.* Cambridge University Pres.

Özatlı, N. S. (2006) *Öğrencilerin biyoloji derslerinde zor olarak algıladıkları konuların tespiti ve boşaltım sistemi konusundaki bilişsel yapılarının yeni teknikler ile ortaya konması*. [Yüksek Lisans Tezi], Balıkesir Üniversitesi.

Öztürk, P. (2011). The effect of usage of concept maps, structured grid and diagnostic tree technics to teach the "Living things and energy relations unit" on 8th grade of primary school students? Attitudes towards science and technology lesson. [Yayınlanmamış Yüksek Lisans Tezi]. https://tez.yok.gov.tr/UlusalTezMerkezi/tezSorguSonucYeni.jsp

Palchikov, N. N. & Bound, D. (1989). Student self-assessment in higher education: a meta-analysis. *Review of Educational Research, 59*, 395-430.

Ramdass, D. & Zimmerman, B. J. (2011). Developing self-regulation skills: The important role of homework. *Journal of Advanced Academics, 22*, 194–218.

Raw, J., Brigden, D. & Gupta, R. (2005). Reflective diaries in medical practice. *Reflective Practice, 6*(1), 165-169.

Reid, N. (2003). Getting started in pedagogical research in the physical science. Learning and Teaching Support Network.

Roberts, T. (2006) Self, peer and group assessment in e-learning-introduction. PA Publishers.

Rolheiser, C. & Ross, J. A. (2001). Student self-evaluation: What research says and what practice shows. R. D. Small & A. Thomas (Ed.), *Plain talk about kids* (s. 43–57) içinde. Center for Development and Learning.

Stiggins, R. J. (2005). From formative assessment to assessment FOR learning: A path to success in standards-based schools. *Phi Delta Kappan*, *87*(4), 324–328.

Şahin, Ç. & Çepni, S. (2011). Developing of the concept cartoon, animation and diagnostic branched tree supported conceptual change text: "Gas pressure". *Eurasian Journal of Physics and Chemistry Education, 1*(1), 25-33.

Topping, K. (1998). Peer assessment between students in colleges and universities. *Review of Educational Research, 68* (3), 249–276.

Topping, K. J. (2009). Peer assessment. *Theory Into Practice, 48*, 20–27.

Türkyılmaz, H. (2017). Bir öğretim materyali olarak zihin haritalarının Din Kültürü ve Ahlak Bilgisi dersinde kullanımı. *Dokuz Eylül Üniversitesi İlahiyat Fakültesi Dergisi*, *45*, 41-80.

University of Florida Center for Instructional Technology & Training. (t.y.). *Designing effective peer and self-assessment*. https://citt.it.ufl.edu/resources/course-development-resources/assessing-student-learning/designing-effective-peer-and-self-assessment/

Yakar, L. (2019). Tamamlayıcı ölçme ve değerlendirme teknikleri III. N. Doğan (Ed.), *Eğitimde Ölçme ve Değerlendirme* (s. 246 – 269) içinde. Pegem Akademi.

Yar Yıldırım, V. (2018). Öğretmen, öğrenci ve velilerin ortaokul düzeyinde verilen günlük ödevler hakkındaki görüşleri. *Millî Eğitim Dergisi, 47*(220), 201-224.

Yazıcıoğlu, E. B. (2007). Çoktan seçmeli testler ile yapılandırılmış gridlerin psikometrik özellikler açısından karşılaştırılması. [Yüksek lisans tezi, Ankara Üniversitesi]. http://tez2.yok.gov.tr/

YAZARLAR HAKKINDA | ABOUT AUTHORS

DOÇ. DR. ERKAN HASAN ATALMIŞ

Manisa Celal Bayar Üniversitesi | erkan.atalmis[at]cbu.edu.tr

ORCID: 0000-0001-9610-491X

Doç. Dr. Erkan Hasan Atalmış, 2004 yılında Ege Üniversitesi Eğitim Fakültesinden mezun olmuş, 2006 yılında Balıkesir Üniversitesi Eğitim Yönetimi alanında tezsiz yüksek lisans eğitimini tamamlamıştır. 2008 yılında Millî Eğitim Bakanlığı bursu ile gittiği Amerika Birleşik Devletlerinde University of Kansas Eğitim Psikolojisi Anabilim Dalında doktora eğitimine başlamış ve 2014 yılında doktora derecesini almıştır. 2019 yılında doçent unvanını alan Atalmış, halen Manisa Celal Bayar Üniversitesi Spor Bilimleri Fakültesinde öğretim üyesi olarak görev yapmaktadır. Araştırma alanları arasında eğitimde ölçme ve değerlendirme, psikometri, meta-analiz, madde tepki kuramı, test geliştirme, geniş ölçekli değerlendirmeler ve eğitim istatistikleri yer almaktadır. Ulusal ve uluslararası hakemli dergilerde çok sayıda bilimsel çalışması yayımlanmış; ulusal ve uluslararası araştırma projelerinde araştırmacı ve yürütücü olarak görev almıştır. Lisans ve lisansüstü düzeyde ölçme ve değerlendirme, araştırma yöntemleri, istatistik ve veri analizi alanlarında dersler vermektedir.

ASSOC. PROF. DR. ERKAN HASAN ATALMIŞ

Manisa Celal Bayar University | erkan.atalmis[at]cbu.edu.tr

ORCID: 0000-0001-9610-491X

Assoc. Prof. Dr. Erkan Hasan Atalmış graduated from the Faculty of Education at Ege University in 2004 and completed a non-thesis master's degree in Educational Administration at Balıkesir University in 2006. In 2008, he was awarded a scholarship by the Turkish Ministry of National Education to pursue doctoral studies in the United States. He began his Ph.D. in Educational Psychology at The University of Kansas and earned his doctoral degree in 2014. He was awarded the title of Associate Professor in 2019 and is currently a faculty member at the Faculty of Sport Sciences, Manisa Celal Bayar University. His research interests include educational measurement and evaluation, psychometrics, meta-analysis, item response theory, test development, large-scale educational assessments, and educational statistics. He has published numerous articles in national and international peer-reviewed journals and has served as both a researcher and coordinator in various national and international research projects. He teaches undergraduate and graduate courses in educational measurement and evaluation, research methods, statistics, and data analysis.

PROF. DR. GÖKHAN AKSU

Adnan Menderes Üniversitesi | gokhanaksu[at]adu.edu.tr

ORCID: 0000-0003-2563-6112

Prof. Dr. Gökhan Aksu, 2007 yılında Dokuz Eylül Üniversitesi Ortaöğretim Matematik Öğretmenliği Bölümünden mezun olmuş, 2012 yılında Adnan Menderes Üniversitesi Eğitim Programları ve Öğretim alanında yüksek lisans eğitimini tamamlamıştır. 2014 yılında Hacettepe Üniversitesi Eğitimde Ölçme ve Değerlendirme Anabilim Dalında doktora eğitimine başlamış ve 2018 yılında doktor unvanını almıştır. Akademik yaşamında 2023 yılında doçent, 2025 yılında profesör unvanını alan Aksu, halen Adnan Menderes Üniversitesi Eğitim Fakültesinde öğretim üyesi olarak görev yapmaktadır. Araştırma alanları arasında eğitimde ölçme ve değerlendirme, Programme for International Student Assessment (PISA), büyük veri analitiği, veri madenciliği, çok değişkenli istatistik ve programlama yer almaktadır. Ulusal ve uluslararası hakemli dergilerde çok sayıda bilimsel çalışması yayımlanmış; çeşitli araştırma projelerinde araştırmacı olarak görev almıştır. Lisans ve lisansüstü düzeyde istatistik ve ölçme-değerlendirme alanlarında dersler vermektedir.

PROF. DR. GÖKHAN AKSU

Adnan Menderes University | gokhanaksu[at]adu.edu.tr

ORCID: 0000-0003-2563-6112

Prof. Dr. Gökhan Aksu graduated from the Department of Secondary School Mathematics Education at Dokuz Eylül University in 2007. He received his master's degree in Curriculum and Instruction from Adnan Menderes University in 2012. In 2014, he began his doctoral studies in the Department of Educational Measurement and Evaluation at Hacettepe University and earned his Ph.D. in 2018. He was promoted to Associate Professor in 2023 and to Full Professor in 2025. He is currently a faculty member at the Faculty of Education, Adnan Menderes University. His research interests include educational measurement and evaluation, the Programme for International Student Assessment (PISA), big data analytics, data mining, multivariate statistics, and programming. He has published numerous articles in national and international peer-reviewed journals and has served as a researcher in various national and international research projects. He teaches undergraduate and graduate courses in statistics and educational measurement and evaluation.

9. BÖLÜM

SINAV ÇEŞİTLERİ VE SINAV PLANI

TYPES OF EXAMINATIONS AND TEST PLANNING

Prof. Dr. Seher YALÇIN

SINAV ÇEŞİTLERİ VE SINAV PLANI

Prof. Dr. Seher YALÇIN
Ankara Üniversitesi

Öz

Bu bölümde test ve sınav kavramları, sınav çeşitleri, sınav planlama süreçleri ve sınav güvenliği ele alınmaktadır. Testler, bireylerin belirli bilgi, beceri veya yeterlik düzeylerini ölçmek amacıyla kullanılan araçlar olup çoktan seçmeli, açık uçlu veya eşleştirmeli maddelerden oluşabilmektedir. Sınavlar ise testleri de kapsayabilen, öğrencilerin belirli bir ders, dönem ya da öğrenme sürecindeki performanslarını değerlendirmeye yönelik daha kapsamlı uygulamalardır. Testler; standart ve standart olmayan testler ile geniş ölçekli ve küçük ölçekli uygulamalar şeklinde sınıflandırılabilmektedir. Bu bölümde bu sınıflamalar örneklerle açıklanmış ayrıca uluslararası geniş ölçekli uygulamalar da tanıtılmıştır. Sınavlar ise sonuçlarının bireyler ve kurumlar üzerindeki etkilerine göre yüksek ve düşük riskli sınavlar olarak değerlendirilmektedir. Ayrıca sınavlar, amaçlarına ve uygulama yöntemlerine bağlı olarak farklı biçimlerde yapılandırılabilmekte; geleneksel kâğıt-kalem uygulamalarının yanı sıra elektronik ve bilgisayar tabanlı ortamlarda da gerçekleştirilebilmektedir. Sınavların verilen kararlara göre seçme, sınıflama, yerleştirme ve yeterlik sınavları olarak sınıflandırmak da mümkündür. Bu bölümde, farklı sınav çeşitlerine değinilmiş Türkiye ve dünyadaki uygulama örneklerine yer verilmiştir. Bölümde ayrıca etkili bir ölçme ve değerlendirme sürecinin temel unsurlarından biri olan sınav planlamasının önemi vurgulanmaktadır. Sınav planı, değerlendirmenin amacı, kapsamı, ölçülecek öğrenme çıktıları, bilişsel düzeyler ve kullanılacak soru türleri arasında sistematik bir ilişki kurulmasını sağlar. Bunun yanında sınav güvenliği, sınav içeriğinin ve sonuçlarının yetkisiz erişim, paylaşım, hile ve diğer tehditlere karşı korunmasına yönelik uygulamalar çerçevesinde ele alınmaktadır. Sınav uygulama sürecinde dikkat edilmesi gereken süreç hakkında bilgi verilmiş, sınav evraklarının dağıtımından adayların kimlik kontrolüne, sınavın yürütülmesinden gözetim süreçlerine kadar izlenmesi gereken temel adımlar açıklanmıştır. Son olarak, sınav sonuçlarına ilişkin itiraz süreçleri değerlendirilmekte ve bu mekanizmaların ölçme ve değerlendirme uygulamalarında adalet, şeffaflık ve güvenilirliğin sağlanmasındaki rolü tartışılmaktadır.

Anahtar Kelimeler

Sınav, Test, Sınav Türleri, Sınav Planı, Sınav Güvenliği, Sınav İtirazı

TYPES OF EXAMINATIONS AND TEST PLANNING

Prof. Dr. Seher YALÇIN
Ankara University

Abstract

This chapter addresses the concepts of tests and examinations, types of examinations, examination planning processes, and examination security. Tests are instruments used to measure individuals' levels of knowledge, skills, or competencies and may consist of multiple-choice, constructed-response, or matching items. Examinations, on the other hand, are broader assessment practices that may incorporate tests and are designed to evaluate students' performance within a specific course, academic term, or learning process. Tests can be classified as standardized and non-standardized tests, as well as large-scale and small-scale assessments. These classifications are explained in this chapter through illustrative examples, and major international large-scale assessment programs are also introduced. Examinations are further categorized as high-stakes or low-stakes according to the consequences of their results for individuals and institutions. In addition, examinations may be structured in various ways depending on their purposes and modes of administration, ranging from traditional paper-and-pencil formats to electronic and computer-based environments. Based on the decisions they support, examinations can also be classified as selection, classification, placement, and proficiency examinations. Accordingly, this chapter discusses a variety of examination types and presents examples of their implementation in Türkiye and around the world. The chapter also emphasizes the importance of examination planning as a fundamental component of effective measurement and evaluation processes. An examination blueprint provides a systematic framework that aligns the purpose and scope of the assessment with the intended learning outcomes, cognitive levels to be measured, and item formats to be employed. Furthermore, examination security is discussed in terms of practices aimed at protecting examination content and results from unauthorized access, disclosure, cheating, and other potential threats. Information is also provided regarding the procedures that should be carefully followed during the administration of examinations. Key steps are described, including the distribution of examination materials, candidate identification procedures, the conduct of the examination, and proctoring processes. Finally, procedures for appealing examination results are examined, and the role of these mechanisms in ensuring fairness, transparency, and reliability in measurement and evaluation practices is discussed.

Keywords

Exams, Tests, Types of Exams, Exam Schedule, Exam Security, Exam Appeals

1. GİRİŞ

Sınav kavramı, çoğu zaman bir test uygulamasıyla özdeşleştirilse de aslında sınavlar bireylerin bilgi, beceri ve yetkinliklerini ortaya koymalarını sağlayan çok yönlü değerlendirme araçlarıdır. Alanyazında 'test' ve 'sınav' terimleri birbirlerinin yerine kullanılmakla birlikte aralarında belirgin farklılıklar vardır. Testler genellikle kısadır ve belirli bir konuyu ölçmeye odaklanır. Öte yandan sınavlar daha kapsamlıdır, resmidir ve genellikle bir dönemin ya da programın tümünü değerlendirmede kullanılır. Bu nedenle sınavlar daha yüksek etkiye sahiptir ve sonuç odaklıdır. Testler bir araç olarak görülmelerinin ötesinde, bir dizi soru aracılığıyla bir davranış örneklemini sistematik biçimde ölçmek için kullanılan standart prosedürler olarak da değerlendirilebilir (Linn, 2008). Aynı zamanda testing (test etme) eyleminin öznesi ve bu sürecin somutlaşmış biçimi olarak da görülebilir. Nitekim Crocker ve Algina (1986) testi, spesifik bir alandaki davranışların örneklemini belirlemede kullanılan standart bir prosedür olarak tanımlamıştır. Eğitim uygulamalarında testler, öğrencilerin belli görevleri tamamlama yeteneklerini, bir beceri veya konu alanındaki yeterliklerini belirlemek için kullanılan araçlardır. Testler çoktan seçmeli sorulardan oluşabileceği gibi açık uçlu veya eşleştirmeli türünde sorulardan da meydana gelebilir.

Sınavlar ise testleri de içerebilen, ancak daha çok öğrencinin belli bir ders, dönem ya da öğrenme sürecindeki genel performansını değerlendirmeye yönelik uygulamalardır. Sınavlar, eğitim sürecinin ayrılmaz ve işlevsel bir parçası olarak öğrenmenin hem değerlendirilmesini hem de yönlendirilmesini sağlar. Sınavlar, çoğunlukla öğrencilerin beklenen bilgi, beceri veya yeterlik düzeylerine ulaşıp ulaşmadığını belirlemek amacıyla kullanılır. Dolayısıyla öğretmenlerin yaptığı yazılı sınavlar ya da üniversitelerdeki vize ve final sınavları bu kapsamda ele alınabilir. Dahası sınavlar, öğretim süreçlerinin etkililiğini değerlendirmeye, eksik kalan alanları belirlemeye ve öğretimi iyileştirmeye yönelik önemli geri bildirimler sağlar. Sınavlar ve testler; amaçlarına, uygulama kapsamına ve karar süreçlerine katkılarına göre çeşitli şekillerde sınıflandırılır.

Kitabın bu bölümünde öncelikle amacına ve uygulama yöntemine göre alanyazında sıkça kullanılan sınav çeşitleri üzerinde durulmuştur. Ardından sınav uygulama planı başlığına yer verilmiş; bu başlık altında önce sınav güvenliği ile sınav süreçlerinden bahsedilmiştir. Akabinde ise sınav sonuçlarına yönelik itirazlar konusuna değinilmiştir.

2. SINAV ÇEŞİTLERİ

2.1. Standart ve Standart Olmayan Testler

Testlere ilişkin ayırımlardan biri standart ve standart olmayan testlerdir. Bu ayrıma esas teşkil eden güvenirlik, geçerlik ve karşılaştırılabilirlik gibi ölçmeye ait temel özelliklerdir. Testler hazırlayanın uzmanlığı, yapılandırılma ve uygulanma biçimlerine göre standart ve standart olmayan testler şeklinde ikiye ayrılır. Standart testler**,** genellikle uzman bir ekip tarafından (alan uzmanı, ölçme-değerlendirme uzmanı ve dil uzmanı) hazırlanır ve bu testleri geliştirmek uzun zaman alır. Bu yönüyle standart testler, standart olmayan testlere kıyasla daha maliyetli olup genellikle yayınevleri ve geniş ölçekli eğitim kurumları için geliştirilir. Standart test geliştirme sürecinde ideal olan pilot uygulamalara başvurmaktadır. Ancak kimi zaman güvenlik riskleri sebebiyle pilot uygulama mümkün olmaz. Bu gibi durumlarda hazırlanan test maddeleri çok sayıda bağımsız denetmenin görüşüne sunulur ve pilot uygulamanın yapılamamasından kaynaklanabilecek olası negatif etkiler en aza indirilmeye çalışılır. Standart testleri uygulayan kişiler, sınavın her yerde ve her zaman aynı koşullar altında yürütülmesinden sorumlu iken bu testlerde görev alan puanlayıcılar aynı puanlama ölçütlerini izlemekle yükümlüdürler. Dolayısıyla standart test süreçlerinde görev yapacak uygulayıcılar ve puanlayıcılar özel olarak eğitilirler. Temsil gücü olan norm tabloları sayesinde standart testler öğrenciler, sınıflar, okullar, eyaletler ve bölgeler arasında karşılaştırma yapılmasını sağlar (Kubiszyn & Borich, 2016).

Standart testler, hem yüksek riskli kararlar almak (örneğin sınıf geçme, mezuniyet vb.), hem öğrencilerin güçlü ve zayıf yönlerini belirlemek hem de belli program ve müfredatları değerlendirmek amacıyla kullanılabilmektedir. Bu testler; başarıyı ölçmek amacıyla hazırlanabileceği gibi yetenek, ilgi, kişilik vb. özellikleri ölçmek amacıyla da kullanılabilmektedir. Örneğin öğrenciler, okullar veya bölgeler arasında ya da zaman içinde boylamsal gelişime yönelik karşılaştırmalar yapmak için standart başarı testlerinden sıklıkla yararlanılmaktadır (Kubiszyn & Borich, 2016). Standart başarı testleri genellikle çoktan seçmeli, eşleştirme ve doğru-yanlış türündeki objektif puanlanabilen maddeleri içermekle beraber, günümüzde bu testlerde açık uçlu soruların kullanımı da giderek yaygınlaşmaktadır. Ölçme, Seçme ve Yerleştirme Merkezi Başkanlığı (ÖSYM) tarafından uygulanmaya başlanan Elektronik İngilizce Yeterlik Sınavı [Electronic-Test of English Proficiency (e-TEP)] açık uçlu soruların kullanıldığı standart testlere örnek olarak verilebilir (ÖSYM, 2025).

Standart başarı testlerinin uluslararası alanda en bilinen örneği Iowa Temel Beceriler Testi'dir (ITBS). Anaokulundan 8. sınıfa kadar olan öğrencilere uygulanan bu test; Kelimeler, Kelime Analizi, Dinleme, Okuma, Anlama, Dil (yazım, büyük harf kullanımı, noktalama ve dil bilgisi), Matematik (kavramlar, problem çözme ve işlem becerileri), Sosyal Bilgiler, Fen Bilimleri, Bilgi Kaynakları ve Temel ile Genel Test Bataryası (birden

çok alt testten oluşan birleşik ölçümler) alanlarında puanlar sağlamaktadır. Standart başarı testlerine başka bir örnek olarak Stanford Başarı Testi Serisi (SAT-10) verilebilir. Bu test, anaokulundan 12. sınıfa kadar uygulanmaktadır (Kubiszyn & Borich, 2016).

Türkiye Cumhuriyeti Millî Eğitim Bakanlığı (MEB) tarafından ilk kez 2016 yılında 8. sınıf öğrencilerine uygulanan Akademik Becerilerin İzlenmesi ve Değerlendirilmesi (ABİDE) çalışması da standart testler arasında yer almaktadır. ABİDE, iki yılda bir 4, 8 ve 10. sınıf seviyelerinde öğrenim gören öğrencilerin Türkçe, matematik ve fen bilimlerindeki bilgi ve becerilerini ulusal düzeyde inceleyen bir izleme araştırmasıdır. Bu çalışma, öğrencilerin kazandıkları bilgi ve becerileri günlük yaşamda kullanabilme durumlarını; gerçek yaşam senaryoları üzerinden eleştirel düşünme, problem çözme, muhakeme ve yorum yapma gibi üst düzey düşünme becerileri aracılığıyla değerlendirmeyi hedeflemektedir (MEB, 2019). Alan uzmanlarının hazırladığı sorular, ölçme-değerlendirme ve dil uzmanları tarafından gözden geçirilmekte, gerekli düzenlemeler yapılarak son hâline getirilmektedir. Asıl uygulama öncesinde pilot çalışma gerçekleştirilmekte, pilot uygulamadan elde edilen test ve madde istatistikleri incelenerek başarı testlerindeki sorular ile anket formlarındaki maddelerde gerekli revizyonlar yapılmaktadır (İlgün-Dibek, 2021). ABİDE uygulamasına katılan öğrenciler Türkçe, matematik, fen bilimleri ve sosyal bilgiler alanlarındaki performanslarına göre beş yeterlik düzeyine (temel altı, temel, orta, orta üstü ve ileri) ayrılmaktadır. ABİDE uygulamasından elde edilen bulgular öğretim programlarındaki eksiklerin giderilmesine katkı sağlamakta, eğitim alanında yapılan araştırmalar için önemli bir kaynak oluşturmaktadır (MEB, 2019).

Standart olmayan testler ise genellikle belirli bir öğrenci grubunun öğrenme düzeyini ölçmek üzere öğretmenler tarafından geliştirilir, uygulanır ve puanlanır. Bu bakımdan standart olmayan testler "öğretmen yapımı testler" olarak da adlandırılır. Bu testler, boşluk doldurmadan doğru-yanlışa, eşleştirmeden çoktan seçmeli ve açık uçluya kadar farklı türde sorular içerebilmektedir. Öğretmen yapımı testler; uygulama ve puanlama yöntemleri ile hazırlanışına verilen önem bakımından çoğunlukla esnek ya da değişkendir. Farklı öğretmenler testlerini hazırlarken daha özenli ya da daha rahat olabilir, sınav için daha uzun ya da daha kısa süre tanıyabilir ve puanlamada daha katı ya da daha cömert davranabilirler. Öğretmenler aynı testi tekrar tekrar kullanabilir ve diledikleri takdirde her kullanımda testte değişikliğe gidebilirler. Bu durumda, zaman içinde öğrenci performanslarının karşılaştırılması pek mümkün olmaz (Kubiszyn & Borich, 2016). Öğretmenler tarafından hazırlanan testlerde test geliştirme sürecindeki tüm adımlar tam anlamıyla takip edilmediği için bu testlerden elde edilen puanların geçerliği ve güvenirliği nispeten düşük olabilmektedir.

2.2. Geniş Ölçekli ve Küçük Ölçekli Testler

Testler, uygulandıkları kapsam ve hedef kitleye göre geniş ölçekli ve küçük ölçekli testler olarak şeklinde iki başlıkta incelenebilir. Geniş ölçekli uygulamalar, eğitim sistemindeki genel performans düzeyleri ve eğilimler hakkında bilgi sağlayarak politika yapımında yol gösterici olur. Bu değerlendirmeler, sistem düzeyinde başarı eğilimlerine ilişkin veriler ve iç görüler sunarak tüm sistemi kapsayan reformlara yol gösterir (Clarke & Luna Bazaldua, 2021).

Geniş ölçekli ve küçük ölçekli değerlendirmeleri birbirinden ayıran en belirgin fark büyüklük olmakta birlikte bu iki sınav çeşidi arasında amaç, maliyet, sonuçların genellenebilirliği, veri analizi türü ve karmaşıklığı bakımından da önemli farklar bulunmaktadır (Anderson & Postlethwaite, 2007). Bu beş temel özellik açısından iki sınav çeşidi arasında gözlenen farklılıkları şöyle detaylandırmak mümkündür: Örneklem büyüklüğü açısından bir sınavın küçük ölçekli olması, uygulama yapılan grubun yaklaşık 30 okulu veya 60 sınıfı aşmaması anlamına gelir. Bu sayı bir ülkede bir sınıf düzeyindeki çeşitliliği tam anlamıyla yansıtmaya yetmeyeceğinden küçük ölçekli çalışmalar genellikle ülkenin tamamını değil yalnızca bir bölgeyi veya ilçeyi kapsar. Geniş ölçekli çalışmalarda ise örneklem ülke genelindeki okullar arasından seçilerek oluşturulur ve genellikle 100'den fazla okul ve her okulda 20–30 öğrenci olması beklenir. Çalışmanın amacı açısından küçük ölçekli testler genellikle keşif amaçlı olup ölçme araçlarının ve yeni öğretim yöntemlerinin geliştirilmesine odaklanır (Anderson & Postlethwaite, 2007). Geniş ölçekli uygulamalar ise ülke içi veya ülkeler arası karşılaştırma yapılmasını, eğitim politikalarına ilişkin kararlar alınmasını sağlar. Maliyet ve zaman açısından karşılaştırıldığında küçük ölçekli uygulamalar, geniş ölçekli uygulamalara kıyasla daha düşük maliyetlidir ve hızlıdır. Ülke genelinde 200 okulu kapsayan bir uygulamayı yürütmek oldukça pahalı olduğu gibi verilerin yönetimi (veri girişi, veri temizliği ve veri analizi) bakımından da maliyetli ve zaman alıcıdır. Veri analizi türü ve karmaşıklığı dikkate alındığında geniş ölçekli uygulamalarda genelde küçük ölçekli uygulamalardan çok daha karmaşık istatistiksel analizlere ihtiyaç duyulur (Anderson & Postlethwaite, 2007). Öğrenci, öğretmen, okul ve veli gibi farklı kaynaklardan ve düzeylerden gelen veriler söz konusu olduğundan tüm veriyi bir arada değerlendiren ve bu yolla hatayı azaltan çok düzeyli istatistiksel modelleri kullanmak gerekir. Son olarak küçük ölçekli çalışmalarda örnekleme hataları büyüktür, dolayısıyla genellemeler yapmak zordur (Anderson & Postlethwaite, 2007). Diğer taraftan geniş ölçekli çalışmalarda ülkeyi temsil eden bir örneklem seçildiğinden elde edilen bulgular evrene genellenebilir olmaktadır. Bu da sonuçların genellenebilirliği yönüyle iki sınav çeşidini farklı kılmaktadır.

Yukarıda bahsedilen farklar göz önünde bulundurulduğunda geniş ölçekli testler, ulusal ya da uluslararası düzeyde çok sayıda katılımcıya aynı anda uygulanan, zaman ve para açısından maliyetli, karmaşık analizler gerektiren ve sistem düzeyinde genellenebilir kararlar alınmasını amaçlayan uygulamalar bütünü olarak tanımlanabilir. Bu tür testler,

politikaların yönlendirilmesi, sistem performansının değerlendirilmesi ya da öğrenciler arası karşılaştırmalar yapılması gibi makro düzey hedefler taşır. Yükseköğretim Kurumları Sınavı (YKS), Liselere Geçiş Sınavı (LGS) ve ABİDE gibi ulusal düzeyde uygulanan sınavlar ile Uluslararası Öğrenci Değerlendirme Programı (PISA-The Programme for International Student Assessment), Uluslararası Matematik ve Fen Eğilimleri Araştırması (TIMSS-Trends in International Mathematics and Science Study) gibi uluslararası değerlendirme programları bu kapsama girer.

Uluslararası geniş ölçekli ilk uygulama, 1964 yılında Birinci Uluslararası Matematik Araştırması (First International Mathematics Study- FIMS) adıyla 12 ülkenin katılımıyla gerçekleştirilmiştir. Türkiye'nin de katıldığı uluslararası düzeyde öğrenci başarılarını belirlemek amacıyla eğitim araştırmaları yürütülmektedir. Bu araştırmalardan bazıları: PISA, TIMSS, Uluslararası Okuma Becerilerinde Gelişim Projesidir (PIRLS-Progress in International Reading Literacy Study).

PISA, üç yılda bir öğrencilerin matematik, fen ve okuma alanlarında bilgi ve becerilerini değerlendirmeye yönelik yapılan bir çalışmadır. Her uygulama döneminde, bu üç beceri alanlarından biri uygulamanın odağında yer almaktadır. Son yapılan PISA 2025 uygulaması, fen bilimleri alanı odaklıdır. TIMSS uygulamaları dört yılda bir 4. ve 8. sınıf öğrencilerinin matematik ve fen başarısını ölçmektedir. İlk uygulama 1995 yılında yapılmıştır. Türkiye, araştırmaya sekizinci sınıf düzeyinde 1999, 2007, 2011, 2015, 2019 ve 2023 yıllarında; dördüncü sınıf düzeyinde ise 2011 ve 2015 yıllarında, 5. sınıf düzeyinde 2019 ve 2023'te katılmıştır. PIRLS'de ilkokul 4. sınıf öğrencilerinin farklı metin türlerine göre okuduğunu anlama becerileri ve okuma alışkanlıkları ölçülmektedir. Uluslararası Eğitim Başarılarını Değerlendirme Kuruluşu (IEA) tarafından üye ve diğer katılımcı ülkelerde beş yıllık dönemler halinde uygulanan bu tarama çalışmasının (MEB, 2021) ilki, 2001 yılında yapılmıştır. PIRLS döngüleri 2006, 2011, 2016, 2021 yıllarında devam etmiş, Türkiye 2001 ve 2021 uygulama döngülerine katılmıştır.

Dünya genelindeki geniş ölçekli değerlendirmeler çeşitli açılardan farklılık göstermektedir. Bu farklılıklar Şekil 1'deki gibi özetlenebilir (Clarke & Luna Bazaldua, 2021)**:**

Şekil 1

Geniş Ölçekli Değerlendirme Uygulamaları Arasında Gözlenebilen Farklılıklar

Uygulanan sınıf düzeyi veya yaş	Çoğu, ilkokul ya da ortaokul öğrencilerini hedef alır.
Hedef kitlenin kapsanma biçimi	Çoğu, öğrencilerden ve okullardan temsil gücü olan örneklemlerden bilgi toplar.
Kapsanan dersler veya beceri alanları	Çoğu değerlendirme dil (ya da okuryazarlık) ve matematik (ya da sayısal beceri) alanlarını içerir, ancak diğer dersler veya bilgi alanları da
Uygulama biçimi	Çoğu değerlendirme kâğıt-kalem formatında yapılmakla birlikte bilgisayar tabanlı ve çevrimiçi uygulamalara geçiş son 10 yılda artmıştır.
Toplanan arka plan verileri	Genelde öğrencilerin ev koşulları ve öğretmen özelliklerine ilişkin bilgiler toplanır.
Uygulama sıklığı	Değerlendirmeler genellikle iki ila beş yılda bir uygulanır.

Küçük ölçekli testler ise belirli bir okul, sınıf ya da öğretmen düzeyinde uygulanan, genellikle öğretim süreçlerini izleme, öğrenci gelişimini değerlendirme ve öğrenmeye dönüt sağlama amacı güden uygulamalardır. Öğretmenlerin kendi geliştirdiği sınavlar, dönem sonu başarı testleri ya da sınıf içi performans değerlendirmeleri bu grupta yer alır. Geniş ölçekli testlerin geçerlik ve güvenirlik açısından yüksek teknik standartlara sahip olması beklenirken küçük ölçekli testlerde bağlamsal uygunluk ve pratiklik ön plandadır (Anderson & Postlethwaite, 2007).

2.3. Yüksek ve Düşük Riskli Sınavlar

Sınavlar, ortaya koydukları sonuçların özellikleri ve bu sonuçlara dayanarak verilen kararların önem derecesi bakımından birbirinden farklılık gösterir. Sonuçlarının bireyler ya da kurumlar üzerindeki etkisine göre sınavlar yüksek ve düşük riskli olarak sınıflandırılır. Yüksek riskli sınavlar, bireylerin eğitim hayatları ya da mesleki gelecekleri üzerinde belirleyici etkilere sahiptir. YKS ve Kamu Personeli Seçme Sınavı (KPSS) yüksek riskli sınavlara örnek olarak gösterilebilir. Yüksek riskli sınav sonuçları; öğrencilerin sınıf geçmesi veya mezun olması, işe girmesi, öğretmen ve müdürlerin değerlendirilmesi, sözleşmelerinin uzatılması ve nakit teşvikler, okullar ve bölgeler için kamusal övgü veya yergi, emlak değerleri ve daha birçok eğitim dışı konu üzerinde etkili olabilmektedir (Kubiszyn & Borich, 2016).

Bazı öğretmenler, öğrencilerinin yüksek riskli sınavlardaki puanlarını yükseltmek için üzerlerinde bir baskı hissederler. Bu baskı öğretmenleri, yüksek riskli sınavlardaki soruları pratik etkinlikler olarak sınıf içi eğitimlerinde kullanmaya yöneltebilmektedir.

Öğretmenlerin sınav hazırlığı süreçlerinde sıklıkla sınavlardaki test maddelerine odaklandıkları gözlenmekte, bu da sınav sonuçlarının geçerliğini düşürücü etki yaratabilmektedir. Öğretmenlerin test odaklı yaklaşımlarla sınavlardaki sorulara yönelik öğretime yönelmesinde politika yapıcıların rolü büyüktür. Yüksek-riskli sınavlarda ölçülen bilgi ve becerilerin açık/ayrıntılı biçimde tanımlanmaması, öğretmenleri test maddeleri üzerinden yürütülen öğretime (madde öğretimine) yöneltmekte, bu durum sınavların geçerliğini zedelemektedir. Bu nedenle politika yapıcıların, yalnızca içeriği net tanımlanmış sınavları desteklemeleri gerektiği, aksi halde hem öğretmenlerin hem de öğrencilerin kaybeden taraf olacağı unutulmamalıdır (Popham, 2001).

Sınav sonucunun testi alan öğrenci için önemli olmadığı uygulamalar düşük riskli sınav olarak adlandırılır (Akyol ve ark., 2021). Düşük riskli sınavlar, genellikle öğrenci gelişimini izleme, öğrenme süreçlerine dönüt sağlama ve öğretim stratejilerini belirleme gibi daha sınıf içi odaklı, yönlendirici amaçlar taşır. Sonuçları, bireyler ya da kurumlar üzerinde ciddi etkilere yol açmaz. Öğretmenler tarafından yapılan kısa sınavlar ve deneme testleri ile gelişimsel değerlendirme araçları bu gruba girer. Düşük riskli sınavların dezavantajı (Akyol ve ark., 2021), bazı öğrencilerin bu sınavları ciddiye almaması ve buna bağlı olarak performanslarının gerçek yeteneklerini yansıtmamasıdır. Yani düşük riskli sınavlar ölçülen grubun motivasyon eksikliğinden kaynaklı hatalara açıktır. Düşük riskli sınavların en bilinen örneklerinden biri PISA'dır. PISA sonuçlarının uygulamaya katılan öğrenciler için doğrudan hiçbir etkisi yoktur. Dolayısıyla PISA politika yapıcılar için önem arz eden bir sınav olabilse de öğrenciler için düşük riskli bir sınavdır. ABİDE, TIMSS ve PIRLS gibi diğer geniş ölçekli uygulamalar da bu grupta yer almaktadır.

2.4. Amacına Göre Sınavlar

Sınavlar, ölçme-değerlendirme sürecinde hangi tür kararların alınacağına bağlı olarak farklı amaçlarla yapılandırılabilir. Mehrens ve Lehman (1991) testle elde edilen veriler ile alınabilecek kararları Şekil 2'deki dört başlıkta ele almaktadır.

Şekil 2

Test Verilerine Dayalı Olarak Alınabilecek Kararlar

Öğretimle İlgili Kararlar	Rehberlik Kararları	İdari Kararlar	Araştırma ve Değerlendirme Kararları
Öğrencilerin öğrenme düzeylerinin tanılanması, öğrenme çıktılarının izlenmesi ve müfredatın öğrenciler üzerindeki etkisinin değerlendirilmesi	Öğrencilerin mesleki, eğitsel ve kişisel planlamaları	Öğrencilerin seçilmesi, sınıflandırılması, yerleştirilmesi ve müfredatın planlanması	Eğitim uygulamaları, politikalar ya da programların etkililiğini sınamaya yönelik yapılan araştırmalar

Şekil 2'de özetlenen kararlar sınavların çeşitlenmesine yol açmıştır. Özellikle idari kararlar, eğitimde ölçme ve değerlendirme alanında sınavların en belirgin kullanım alanlarından biridir. Bu kapsamda sınavları seçme, sınıflama, yerleştirme ve yeterlik sınavları olarak sınıflandırmak mümkündür.

- Seçme sınavları, bireylerin belirli bir program ya da kuruma kabul edilip edilmeyeceğine karar vermek için uygulanır. Örneğin YKS ve KPSS bu amaca hizmet eder.
- Sınıflama sınavları, bireylerin hangi tür programa veya alana yönlendirileceğini belirlemek amacıyla kullanılır. Örneğin liselerde öğrencileri sayısal, sözel, eşit ağırlık gibi alanlara yönlendirmeye yardımcı olması amacıyla yapılan sınavlar bu kategoride yer alır.
- Yerleştirme sınavları, bireylerin belli bir programın hangi düzeyinde (hazırlık sınıfı, temel düzey ve ileri düzey gibi) eğitim alacağına karar verilmesine yardımcı olur. ÖSYM tarafından yapılan Tıpta Uzmanlık Eğitimi Giriş Sınavı (TUS) bunun tipik bir örneğidir. Bu sınavın sonuçlarına göre, tıpta uzmanlık eğitimine kabul edilecek adaylar puanları ve tercihleri doğrultusunda mevcut kontenjanlara yerleştirilmektir.
- Yeterlik sınavları, bireylerin belirli bir alanda asgari bilgi ve beceriye sahip olup olmadığını saptamak için kullanılır. Yabancı dil yeterlilik sınavları, ÖSYM'nin yaptığı Hukuk Mesleklerine Giriş Sınavı (HMGS) ve İdari Yargı Ön Sınavı (İYÖS) yeterlik sınavlarına örnek olarak verilebilir. HMGS hukuk fakültesi mezunlarının avukatlık ya da noterlik stajına başlayabilmesi ve hâkimlik-savcılık yardımcılığı sınavına başvurabilmesi için bir ön koşul sınavken İYÖS, siyasal bilgiler, idari bilimler, iktisat ve maliye gibi alanlardan dört yıllık lisans düzeyinde mezun olan kişilerin idari yargı hâkimliği/savcılığı gibi mesleklere geçebilmek için girdiği ön sınavdır.

Bu tür sınavlardan elde edilen sonuçların güvenilir ve geçerli olabilmesi için sınavın hangi bilgi ve beceri alanını yokladığının açıkça tanımlanması gerekir. Söz gelimi öğrencileri belirli düzeydeki derslerden birine (başlangıç, orta ve ileri seviye İngilizce gibi) yerleştirmeye dönük sınav geliştirenler, öğrencilerin bu düzeydeki derslere hazır olarak katılabilmeleri için gerekli bilgi ve beceri düzeyini dikkatle tanımlamalıdır. Sınavı hazırlayanlar, ilgili derste gerekli olan ön bilgi ve becerileri ayrıntılı olarak belirtmezlerse öğrencileri yanlış kurslara yönlendirebilirler, haliyle uygulanan sınavın geçerliği düşük olur (Schmeiser & Welch, 2006).

Sınav puanlarına dayanarak çıkarım yapılacak alanların doğru seçilmesi, bir sınavın birden fazla amaca uygunluğunu değerlendirmeye de yardımcı olur. Çünkü sınavlar genellikle belli bir amaç (örneğin yerleştirme, izleme) ve belli bir hedef grup için tasarlanır. Çıkarım yapılacak alanların önceden net bir şekilde belirlenmesi belli bir

amaca yönelik olarak geliştirilen bir sınavı, başka amaçlar için kullanmanın neden uygun olmayacağına da ortaya koyar. Örneğin bir eyalet, öğrencilerin lise diploması alabilmesi için bitirme sınavını geçmelerini şart koşabilir. Bu sınav eyaletin belirlediği içerik standartlarına göre, lise mezuniyeti için yeterli kabul edilen bilgi düzeyini tanımlamalıdır. Bu performans standartları genellikle bir standart belirleme süreciyle oluşturulur. Ancak, aynı testin üniversiteye giriş hakkında kararlar almak için de kullanılması uygun değildir. Çünkü eyalet standartları genellikle lise diploması için gerekli bilgiyi yansıtır ve bu düzey genel itibariyle öğrencilerin üniversitede başarılı olabilmeleri için gereken ön bilgiden farklıdır. Hedef bilgi alanları önemli ölçüde farklı olduğundan aynı testin bu iki amaca hizmet etmesi çok zor, hatta imkânsızdır (Schmeiser & Welch, 2006). Benzer biçimde Türkiye'de farklı sınavlar başka başka amaçlara hizmet eder. Örneğin, MEB tarafından yapılan kazanım değerlendirme sınavları öğrencilerin öğretim programındaki kazanımlara ne ölçüde ulaştıklarını belirlemeyi amaçlarken ÖSYM'nin yaptığı YKS öğrencilerin yükseköğretime geçişteki göreli başarılarını ölçer. Aynı testin hem öğrenme düzeyini değerlendirme hem de seçme amacıyla kullanılması, ölçme kapsamı ve geçerlik açısından çeşitli problemler doğurur.

Mehrens ve Lehman (1991) ölçmenin en önemli rolünün, bireylerin gelişimini desteklemeye yönelik öğretimsel kararlara katkı sağlamak olduğunu vurgulamakta, bununla birlikte seçim ve sınıflama kararlarının da eğitim sisteminde kaçınılmaz olduğunu belirtmektedir. Eğitim kurumlarında başvurunun fazla fakat kontenjanın sınırlı olduğu etkinlikler (örneğin özel yetenek sınavları veya bursluluk sınavları) seçim işlevini zorunlu kılar. Bu nedenle ölçme-değerlendirme, yalnızca gelişimi destekleyen değil aynı zamanda seçme ve sıralama işlevini de yerine getiren kritik bir araçtır.

Tablo 1'de farklı sınav amaçları ve kullanımları için sınav puanlarından yapılabilecek çıkarım türlerine örnekler sunulmaktadır. Burada listelenen amaçların hem bireylerle ilgili eğitimsel kararları hem de müfredat ve öğretimle ilgili kararları kapsadığına dikkat edilmelidir (Schmeiser & Welch, 2006).

Tablo 1

Farklı Sınav Amaçları ve Kullanımları için Sınav Puanlarından Yapılabilecek Çıkarımlar (Schmeiser & Welch, 2006)

Sınavın Amacı		**Kullanım Bağlamı**	**Çıkarımlar**
Bireysel	Yerleştirme	Derse yerleştirme Rehberlik	Belirli bir derse "öğrenmeye hazır" şekilde katılmak için gerekli bilgi
	Tanılama	Telafi Zenginleştirme Öğretimsel müdahaleler	Çeşitli içerik alanlarında görece güçlü ve zayıf yönler Belirlenmiş içerik standartlarına göre çeşitli alanlarda bilgiye ne ölçüde hâkim olunduğu
	Seçme	Üniversiteye giriş	Üniversite öğrenimine hazır olmayı gösteren bilgi ve beceriler
	Sınıflama	Lisanslama / Sertifikalandırma	Güvenli ve etkili mesleki uygulama için gerekli bilgi
	İlerleme	Mezuniyet / Çıkış Ders sonu Sınıf geçme Zaman içindeki gelişim	Zaman içinde müfredat alanı veya belirlenmiş standartlara göre bilgi değişimi Bir dersin, sınıfın veya eğitim düzeyinin tamamlanmasıyla ulaşılan bilgi düzeyi
Öğretimsel / Program Temelli	Öğretimsel Uyarlama	Ders öncesi test	Öğrencilerin ders hazırbulunuşluğuna ilişkin başarı dağılımı (öğretim planlamasında kullanılmak üzere)
	Öğretimsel Etkililik	Ders sonrası test Ders / program değerlendirme ve geliştirme	Kabul edilebilir ders başarısı standardını karşılamak için gerekli bilgi Dersin içerik alanının, diğer okullardaki "benzer" derslerin içerikleriyle karşılaştırılması
	Program Etkililiği	Konu alanındaki dersler boyunca eğitimsel ilerleme	İçerik alanlarında, belirlenmiş ilerleme beklentilerine göre zaman içinde başarı artışı

2.5. Uygulama Yöntemine Göre Sınavlar

Sınavlar, uygulama yöntemine göre geleneksel (kâğıt-kalem) ve elektronik (dijital, bilgisayar tabanlı) sınavlar olarak ikiye ayrılabilir. Yani, test ya kâğıt üzerinde ya da bir ekranda öğrenciye sunulur. *Kâğıt-kalem testi*, kâğıda basılı olarak uygulanan geleneksel testleri ifade eder. Bu testler çoğunlukla kitapçık biçiminde hazırlanır. Testlerin kendisi kâğıt üzerinde hazırlanıp gözden geçirilse de günümüzde yaygın olarak bilgisayar veya

internet aracılığıyla oluşturulmakta, ardından basılarak uygulanmaktadır (Foster, 2016). Görece kolay uygulanabilmesi ve düşük teknoloji gerektirmesi nedeniyle halen birçok kurumda kâğıt-kalem testleri tercih edilmektedir. Geniş örneklemlerin katıldığı bazı sınavlarda teknolojik imkânların yeterli olmaması da kâğıt-kalem uygulamalarının yoğun biçimde devam etmesinde etkili bir faktördür. YKS, KPSS ve LGS kâğıt-kalem formatında uygulanan sınavlara örnek olarak verilebilir.

Elektronik sınavlarda (e-sınavlar), geleneksel kâğıt-kalem tabanlı yöntemler yerine dijital teknolojiler kullanılır. Bu sınavlar genellikle bilgisayarlar, tabletler veya diğer dijital cihazlar aracılığıyla uygulanır. Elektronik sınavlar internet üzerinden çevrimiçi olarak (çevrimiçi sınavlar) yapılabileceği gibi yerel bir ağ veya test sistemi aracılığıyla çevrimdışı olarak da (bilgisayar tabanlı testler) yürütülebilir. Bilgisayar tabanlı testlerde, test içeriği uygulama başlamadan önce tamamen indirilir. Genellikle bir sınav merkezinde veya okul bilgisayar laboratuvarında, masaüstü, dizüstü veya mobil cihazlar üzerinden yerel bir sunucu ağı aracılığıyla uygulanır. Test başlamadan önce test içerikleri ve yönergeler programın sunucularından yerel sunucuya indirilir. Sınav sırasında veya sonrasında öğrenci yanıtları yine bu yerel ağ üzerinden toplanır ve daha sonra programın ana sunucusuna aktarılır. İnternet tabanlı elektronik testlerde, sürekli bir internet bağlantısı aracılığıyla sorular öğrenci tarafından kullanılan bilgisayar cihazına ihtiyaç duyulduğu anda aktarılır. Her öğrenci yanıtı da anında internet üzerinden uzak bir sunucuya iletilir (Foster, 2016). Teknik altyapı gereklilikleri, siber güvenlik önlemleri ve eşit erişim gibi konular, elektronik sınav uygulamalarını dikkatle planlamayı gerektirir. Günümüzde birçok kurum, iki yöntemi bir arada kullanmakta ya da elektronik sınavlara kademeli geçiş yapmaktadır.

Türkiye'de MEB'in yaptığı Özel Motorlu Taşıt Sürücüleri Kursiyerleri Sınavı (MTSKS) 2011 yıllından beri elektronik olarak gerçekleştirilmektedir. Yine ÖSYM tarafından yapılan Yabancı Dil Sınavı (YDS) 2014 yılı itibariyle elektronik ortamda bilgisayar tabanlı olarak da uygulanmaya başlanmıştır. PISA ve TIMSS gibi geniş ölçekli uluslararası değerlendirmeler de artık bilgisayar ortamında uygulanmaktadır. PISA'da 2018 yılında ilk defa bireye uyarlanmış test (adaptive test) uygulamasına geçilmiştir. Buna göre, okuma değerlendirmesinin önceki bölümlerinde iyi performans gösteren öğrencilerin değerlendirmenin ilerleyen bölümlerinde daha zor sorularla karşılaşma olasılıkları daha yüksektir. Benzer şekilde, değerlendirmenin önceki bölümlerinde kötü performans gösteren öğrencilerin sonraki bölümlerinde daha kolay sorular alma olasılıkları daha yüksektir. PISA'da kullanılan uyarlamalı test yaklaşımı, aslında bilgisayar ortamında bireye uyarlanmış test uygulamalarından biridir.

Bilgisayar ortamında bireye uyarlanmış test (Computerized Adaptive Testing-BOBUT), uygulamalarının temelinde bireyin önceki yanıtlarından hareketle kestirilen yetenek düzeyine en uygun maddelerin seçilerek uygulanması anlayışı vardır. Böylece test bireylerin yetenek düzeyine göre bireyselleştirilerek sunulmaktadır (Embretson &

Reise, 2000). Bu yaklaşım sayesinde bireyler, kendi düzeylerine uygun olmayan sorularla uğraşmak zorunda kalmaz; bu da daha az sayıda maddeyle yetenek düzeyini belirlemeye olanak tanır (Kalender, 2009). Yani yüksek yetenek düzeyindeki bireyler kolay sorularla karşılaşmazken düşük yetenek düzeyindeki kişiler de zor sorularla karşılaşmamaktadır. BOBUT hakkında detaylı bilgiler kitabın 15. bölümünde yer almaktadır.

PISA, 2018 yılından itibaren her dönemde bir alanı çok aşamalı test olarak uygulamıştır. Önce okuma alanı ile başlamış ve 2022'de matematik okuryazarlığında öğrencinin önceki yanıtlarına göre sonraki soruların belirlendiği dinamik bir yapı geliştirmiştir. Fen okuryazarlığında bireyselleştirilmiş test ise PISA 2025'te kullanılmıştır. TIMSS 2023 ise ilk kez tüm ülkelerde bilgisayar tabanlı uygulanmış; bu sayede dijital etkileşim, çoklu materyal kullanımı ve farklı madde türleri mümkün olmuştur. TIMSS 2023'te bireye tam uyarlamalı bir BOBUT yerine, grup düzeyinde uyarlama sağlayan, grup adaptif desen benimsenmiş, ülkelerin öğrenci düzeyine göre uyarlanmış kitapçık rotasyonları uygulanmıştır. Geleneksel rotasyon yapısı geliştirilerek farklı zorluk derecelerine sahip kitapçıkların ülkeler arasında değişen oranlarda dağıtılması sağlanmıştır. Böylece geçerlik ve öğrenci ilgisi arttırılmıştır. Bilgisayar tabanlı uygulama ile test güvenliği güçlendirilmiş, puanlama objektifleştirilmiş ve otomatik veri kaydı ve puanlama sağlanmıştır (MEB, 2024).

3. SINAV UYGULAMA PLANI

Etkin ve geçerli bir ölçme süreci için sınavın önceden planlanması büyük önem taşır. Sınav planı, değerlendirmenin hangi amaçla yapılacağını, hangi içeriklerin ne düzeyde kapsanacağını, hangi bilişsel düzeydeki hedeflerin ölçüleceğini ve hangi tür soruların kullanılacağını sistematik biçimde ortaya koyar. Planlama sürecinde öncelikle ölçme amacına karar verilir. Bu amaç doğrultusunda uygun kazanımlar ve bu kazanımları temsil edecek soru sayıları belirlenir. Sınav kapsamı ve kapsam ağırlıkları netleştirilerek öğretim programının temsiliyeti sağlanır. Ardından soruların bilişsel düzeylerine (hatırlama, anlama, uygulama, analiz vb.) göre dağılımı yapılır ve soru türleri (çoktan seçmeli, açık uçlu, eşleştirme vb.) kararlaştırılır. Tüm bu unsurlar, sınavın güvenirliğini ve geçerliğini doğrudan etkiler. Puanlama ölçütlerinin belirlenmesi, uygulama sürecinin organizasyonu ve sınav sonrası yapılacak analizler de sınav planlamasının önemli bir parçasıdır. İyi hazırlanmış bir sınav planı, öğretmenlerin ve ölçme uzmanlarının hem daha bilinçli hem de daha adil değerlendirmeler yapmasına olanak tanır.

4. SINAV GÜVENLİĞİ

Sınav güvenliği, test içeriği ve sınav sonuçlarını zarara uğratabilecek her türlü tehdit ve ihlale karşı alınan önlemleri ifade eder (International Test Commission [ITC], 2014). Sınav güvenliği, test içeriğinin ve puanlarının yetkisiz erişim, paylaşım veya hile girişimlerinden kaynaklanabilecek zararlara karşı direnci veya korunma derecesidir

(Foster, 2016). Kopya çekme ve test hırsızlığı yüksek riskli testlerde güvenliği sağlamak için iki önemli hususdur. *Kopya çekme* bireysel ya da organize girişimler yoluyla puanları yapay olarak yükseltmeyi hedeflerken *test hırsızlığı* veya *korsanlığı* test içeriğini yasa dışı yollarla ele geçirerek paylaşmayı ya da satmayı amaçlar. Bu bağlamda, tehdit, savunmasızlık ve ihlal kavramları sınav güvenliğini anlamak açısından kritik öneme sahiptir. *Tehdit* potansiyel zarar kaynağını, *savunmasızlık* bu tehdide ne derece açık olunduğunu, *ihlal* ise söz konusu tehdidin başarıya ulaşarak zarara yol açtığı durumu ifade eder (ITC, 2014).

Yüksek riskli testlerde iki farklı türde test sahteciliği vardır. Bunlar kopya çekme ve test hırsızlığıdır. *Kopya çekme,* genellikle tüm test sahteciliğinin eş anlamlısı olarak kullanılır. Ancak bu yanlış bir kullanım olup kopya çekme kişinin kendi çabasıyla veya başkaları adına hak ettiğinden daha yüksek bir test puanı elde etmesini sağlayan her türlü çabayı ifade eder. *Test hırsızlığı* ise daha yüksek bir puan almaya yönelik değil test içeriğinin (sorular, cevaplar ve sorularla ilişkili diğer veriler [ör. grafikler] dahil) tamamını veya bir kısmını ele geçirmeye dönük bir eylemdir. Test içeriğinin çalınması veya korsanlığı, sınav sorularının önceden bilinmesini sağlamak, kopya çekmek amacıyla paylaşmak veya satılmak için yapılır (Foster, 2016).

Kâğıt-kalem tabanlı sınavlarda en büyük riskler arasında kopya çekme, sınav kitapçıklarının çalınması veya içeriklerinin izinsiz paylaşılması yer alır. Bu nedenle kurumlar sınav sorularının hazırlandığı bilgisayarların güvenliğinin sağlanması, sınav materyallerinin güvenli dağıtımı, saklanması ve erişimi konusunda titizlikle hareket etmelidir. Test kitapçıklarının kilitli dolaplarda saklanması, yalnızca yetkili personelin erişimine izin verilmesi ve dağıtım sürecinde güvenilir prosedürlerin takip edilmesi sınav güvenliğini sağlamak için alınabilecek kritik önlemlerden bazılarıdır. Ayrıca gözetmenlerin görevlerini titizlikle yapması, öğrencilerin kimlik doğrulamalarının dikkatle gerçekleştirilmesi ve sınav sırasında kurallara ne derece uyulduğunun özenle takip edilmesi (örneğin sınav salonun kamere ile kayıt altına alınması, gizli gözlemciler ile sınavların izlenmesi) sınav güvenliğini artıran temel adımlardır (ITC, 2014; ÖSYM, 2016).

Kurumlar ayrıca sınav sürecinin tüm aşamalarında — sınav öncesi, sınav anı ve sınav sonrası — güvenlik planları oluşturmalıdır (ITC, 2014). Güvenlik planı, bir sınavın güvenli biçimde yürütülmesini sağlamak amacıyla uyulması gereken kuralları içeren resmi bir dokümandır. Bu planın sınav sürecinde güvenliğin korunmasını garanti altına alabilmesi hem adayların hem de görevli personelin uyması gereken tüm kuralları ayrıntılı şekilde açıklamasına bağlıdır. Hazırlanan planda sınava katılanların sorumlulukları, uygulanacak yasal düzenlemeler ve yapılacak sözleşmeler ile olası güvenlik ihlallerine ilişkin açıklamalar yer almalı; test güvenliğinden kimlerin sorumlu olduğu ve bunun nasıl sağlanacağı netleştirilmelidir. Yasal düzenlemeler ve sözleşmeler bölümünde ise adayların sınav öncesinde imzalaması gereken gizlilik taahhütleri olup olmadığı, sınavı hazırlayan ve uygulayan tarafların karşılıklı yükümlülükleri ve sınav

sorularının üçüncü kişilerle paylaşılması durumunda doğacak yasal sonuçların ne olacağı belirtilmelidir. Güvenlik ihlalleriyle ilgili kısımda ise sınav sırasında veya sonrasında yaşanabilecek kural ihlallerine karşı uygulanacak yaptırımlar açıklanmalıdır. Örneğin kural ihlali yapan bir adayın puanının iptal edilip edilmeyeceği, yeniden sınava girme hakkının olup olmayacağı, sınav sorularını internette paylaşması durumunda hangi yasal işlemlerin uygulanacağı, bu paylaşımların nasıl engelleneceği veya çevrimiçi sınavlarda, sınav sırasında farklı bir ekran açıldığında sınavın geçersiz sayılıp sayılmayacağı gibi hususlar bu bölümde açıkça belirtilmelidir (Doğan & Aksu, 2021). ÖSYM 2011 yılından beri yaptığı sınavlarda, güvenlik önlemlerinin bir parçası olarak sınav salonları kamera ile izlenmekte (ÖSYM, 2011), kamera kayıtlarından kural ihlali yaptığı tespit edilen adayların sınavı ÖSYM Yönetim Kurulu tarafından geçersiz sayılabilmektedir (ÖSYM, 2013a). Ayrıca, 6114 sayılı ÖSYM Kanunu'nda 'kopya çektiği tespit edilen adayın sınavı iptal edilir ve iki yıl süreyle ÖSYM'nin sınavlarına giremez' hükmü yer almaktadır (Resmî Gazete, 2011).

Sınavın güvenliğini en üst düzeye çıkarmak için uygulama öncesinde, süresince ve sonrasında dikkat edilmesi gereken çeşitli hususlar vardır. Sınav öncesinde materyallerin çoğaltılması ve taşınması güvenli şekilde yürütülmelidir. Sınav sırasında gözetmenler, adayların kopya girişimlerini görmezden gelmemeli, olası kural ihlallerine müdahalede bulunmalı ve gerekli prosedürleri işletmelidir. Sınav sonrasında ise test kitapçıkları ve cevap kâğıtları güvenli şekilde toplanmalı, taşınmalı ve saklanmalıdır. Bu çabalar, tüm kopya ya da içerik hırsızlığını önlemeyi garanti etmese de bu risklerin ortaya çıkma olasılığını önemli ölçüde azaltmaktadır (ITC, 2014). ÖSYM sınavlarında da pek çok güvenlik önlemi alınmaktadır. Örneğin sınav evrakları, şehirlerarası nakil araçlarıyla emniyet ve jandarma koordinasyonunda, zaman ayarlı elektromekanik kilit sistemi ve GPS takip yazılımlarıyla yüksek güvenlik altında taşınmaktadır. Nakil araçlarının güzergâhları, hızları ve tüm duraklamaları anlık olarak ÖSYM Sınav Koordinasyon Merkezi'nde izlenmekte; araçlar güvenlik güçlerinin gözetiminde korunmaktadır. Sınav sabahı elektronik kilitler kameralar önünde açılarak evrak şehir içi nakil araçlarına devredilmekte ve binalara ulaştırılmaktadır. Sınav bitiminde evrak yine kilitli kutularla toplanarak aynı güvenlik önlemleriyle ÖSYM'ye geri götürülmektedir (ÖSYM, 2025b).

Teknoloji tabanlı sınavlar, bir yandan ölçme sürecinde hız, otomatik puanlama ve erişim kolaylığı gibi avantajlar sunarken diğer yandan yeni güvenlik risklerini beraberinde getirmektedir. Dijital kopyalama, küçük cihazlarla içerik sızdırma ve internet üzerinden yayma gibi tehditler bu riskler arasındadır. Bu nedenle güvenlik planları sınav öncesi, sırası ve sonrasındaki tüm süreçleri kapsayacak şekilde tasarlanmalı; kimlik doğrulama, madde geliştirme, sınav yönetimi, puanlama ve verilerin saklanması gibi aşamalarda duyarlı materyallerin korunmasına yönelik düzenlemeleri içermelidir. Böylelikle, sınav güvenliğine yönelik çabalar hem bireysel hile girişimlerini hem de daha sistematik test sahtekârlıklarını en aza indirmeyi hedefler (ITC, 2014).

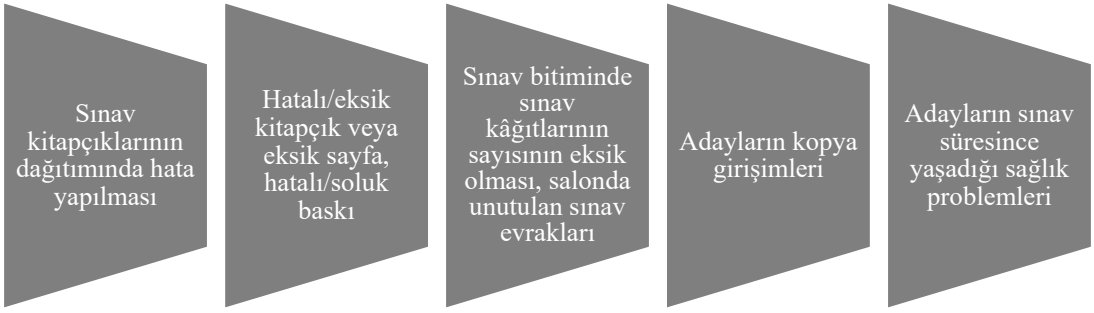
Son yıllarda çevrimiçi sınavlarda, cep telefonuyla soruların paylaşılması, uzak masaüstü yazılımları aracılığıyla ekran görüntülerinin aktarılması veya kamerayla soru paylaşılması gibi çeşitli yöntemlere başvurulduğu görülmektedir. Bu tür hile girişimlerini önlemek ve adayları caydırmak amacıyla yerinde denetim, ekran izleme, göz izleme ve biyometrik doğrulama gibi uygulamaların hataya geçirilmesi önerilmektedir. Adayların sınav süresince verdikleri yanıt örüntülerinin ve performanslarının incelenmesi de kopya girişimlerine dair ipuçları sunabilmektedir. Adayın soruları olağan dışı kısa sürede yanıtlaması, şansa bağlı doğru yanıtlama oranlarının beklenen düzeylerin dışında gerçekleşmesi ya da sınav sırasında olağan dışı test davranışlarının sergilenmesi üzerinde durulması bulgular arasında yer almaktadır (Doğan & Aksu, 2021).

5. SINAV SÜRECİ

Sınav evraklarının uygulama yapılacak binaya getirilip teslim edilmesiyle sınav süreci başlamış olur. Adayların sınav salonuna girişi, kimlik kontrolü yapılarak gerçekleştirilir. Adaylar, önceden belirlenen oturma düzenine göre sıralarına yerleştirilir. Sınav kuralları adaylara açıklanır. Sınav sırasında gözetmenler, sınav uygulama rehberinde belirtilen sınav akışını adım adım takip eder. Gözetmenlerin sınav süresince yaşayabileceği bazı durumlar şunlardır (Sindre & Vegendla, 2015):

Şekil 3

Gözetmenlerin Sınav Süresince Yaşayabileceği Bazı Durumlar



Bunlara ek olarak, Türkiye'deki merkezî sınav uygulamaları düşünüldüğünde görevli değişiklikleri, sınav yönergelerine hâkimiyet eksikliği, kimlik kontrolü ve güvenlik süreçlerinden kaynaklanan gecikmeler, adayların sınav salonuna yasak materyalle girmeye çalışmaları, sınav salonlarının fiziksel koşullarına ilişkin sorunlar veya sınav evraklarının teslim sürecinde yaşanan aksamalar gözetmenlerin karşılaşabileceği durumlar arasında yer almaktadır.

Sınav uygulama sürecinde dikkat edilmesi gereken hususlar ise şöyle sıralanabilir: tüm öğrenciler sınav uygulama ilkelerine eksiksiz riayet etmelidir. Sınav sırasında adayın dersine giren öğretmen ve okul personeli salonda bulunmamalıdır. Sınav materyalleri yalnızca değerlendirmeye katılacak öğrencilere verilip sınav sonrasında geri

toplanmalıdır. Sınavın tamamlanmasına ilişkin tüm yönergeler açıkça iletilmeli, öğrencilerin cevaplarını nasıl kaydedeceklerini anlamaları ve sınav sürelerine sıkı şekilde uyulmaları sağlanmalıdır (Clarke & Luna-Bazaldua, 2021). Uygulayıcılar her sınav için uygulama öncesi eğitimler almalı, sınav uygulama rehberinde yer alan her aşamayı sıkı sıkıya takip etmelidirler.

Sınav uygulayıcılarının, güçlü organizasyon becerilerine, okullarda çalışma deneyimine sahip olmaları ve sınav protokollerini tam olarak uygulamaya bağlılık göstermeleri beklenir. Sınavların uygulandığı okullarda, okul irtibat görevlisi, ulusal geniş ölçekli değerlendirme ekibinin okul bazındaki iletişim kişisidir. Bu kişi, okul personelinin değerlendirme hakkında bilgilendirilmesine yardımcı olur ve sınavın okulda düzenli şekilde uygulanabilmesi için hazırlıkları ve lojistiği koordine eder (Clarke & Luna-Bazaldua, 2021). Sınav uygulama sürecinin sağlıklı bir şekilde yürütülmesi oldukça önemlidir. Uygulama sürecinde yapılan hatalar adayların sınav sonuçlarına hata karışmasına ve sınavlarının iptal edilmesine neden olabilir.

6. İTİRAZLAR

Sınav sonuçlarına yönelik itirazlar, değerlendirme sürecinin adil, şeffaf ve güvenilir olmasının temel unsurlarından biridir. Adaylar veya ilgili taraflar, sınav sonuçları hakkında şüphe duyduklarında ya da puanlama hatası, teknik aksaklık ya da usulsüzlük iddiasında bulunduklarında resmi itiraz mekanizmalarını kullanabilirler. Sınav yapan kurumların; itirazların nasıl yapılacağına, hangi süre içinde itiraz başvurularının alınacağına ve itirazların nasıl değerlendirileceğine dair açık/net yönetmeliklerinin bulunması bu noktada önem taşımaktadır. Örnek teşkil etmesi amacıyla aşağıda Türkiye'de yaygın olarak sınav yapan iki kurumun itiraz sürecine ilişkin yönetmelikleri ve açıklamaları sunulmuştur.

ÖSYM'nin "Soruların Hazırlanması, Soru Havuzunun Oluşturulması ve Soruların Güvenliğinin Sağlanmasına İlişkin Yönetmelik"te "Soruların analizi" başlığı altında yer alan Madde 10'da (ÖSYM, 2013b):

> "*…Adaylar sorular ile ilgili itirazları soru yazarına ve bilimsel denetmenlere ve gerektiğinde haricen görevlendirilecek konunun uzmanı öğretim üyelerine ileterek soru ile ilgili görüş ve düşünceleri talep edebilir. Elde edilen bulgular sorunun yanlış veya eksik olduğunu gösterdiği takdirde, ilgili soru ÖSYM Yönetim Kurulu tarafından iptal edilebilir. Hatalı veya eksik soru hazırlayan veya denetleyenler uyarılabileceği gibi havuzdan da çıkarılabilir.*"

şeklinde açıklamalar yer almaktadır. Kurum, sınav sorularının bilimsel açıdan doğru olup olmadığının incelenmesine ilişkin taleplerin, sınav tarihinden itibaren üç iş günü içerisinde Genel Amaçlı Dilekçe Örneği ile yapılması gerektiğini de yayımlamıştır (ÖSYM, 2025c).

MEB (2025a)nın yayımladığı güncel "Millî Eğitim Bakanlığı Merkezî Sistem Sınav Yönergesi"nde Sınav itirazı başlığındaki Madde 34'te:

> *"(1) Merkezî sistem sınavlarına yapılan itirazlarda:*
>
> *a) Sınav soruları ve sınav sonuçlarına yönelik itirazlar, sınav protokolü veya kılavuzunda belirtilen usul ve esaslar doğrultusunda yapılır. Kılavuz veya protokolünde itiraz süresi belirlenmeyen sınavlarda itiraz süresi; sorulara itiraz için soruların ilanından sonra 5 (beş) takvim günü, sonuçlara itiraz için sonuçların ilanından sonra 5 (beş) takvim günüdür,*
>
> *…*
>
> *d) Süresi geçtikten sonra yapılan itirazlar ile sınav adı, aday ad ve soyadı, T.C. kimlik numarası, iletişim bilgileri, imza ve ekinde itiraz ücretinin yatırıldığını gösteren belge olmayan dilekçeler işleme alınmaz.*
>
> *e) İtirazlar, sınav protokol veya kılavuzunda/duyurusunda belirtilmişse usulüne uygun olarak e-İtiraz modülünden alınır ve cevaplandırılır.*
>
> *f) Sınav protokol veya kılavuzunda/duyurusunda aksi belirtilmedikçe itiraz yolu ile cevap kâğıdı inceleme talepleri, Genel Müdürlük onayı ile Genel Müdürlükte veya il millî eğitim müdürlüğünün onayı ile ölçme değerlendirme merkezi müdürlüklerinde oluşturulan komisyonlar vasıtasıyla karşılanır."*

açıklamaları yer almaktadır.

MEB (2025b), sorulara, cevap anahtarlarına ve sonuçlara yapılacak itirazlar, soruların, cevap anahtarlarının ve sonuçların https://www.meb.gov.tr internet adresinde yayımlanmasından itibaren en geç 5 (beş) takvim günü içinde Ölçme, Değerlendirme ve Sınav Hizmetleri Genel Müdürlüğü resmî internet sayfasında yayımlanan sınav itiraz bölümünden elektronik olarak yapılabileceğini duyurmuştur.

İtiraz süreci, sınavın geçerliğini ve güvenirliğini korumanın yanı sıra, katılımcıların haklarının korunmasına da hizmet eder. Sınav yapan kurumların, itiraz sürecini nesnel bir şekilde değerlendirmesi önemlidir. Böylece, hem sınav yapan kurumların itibarı güçlenir hem de ölçme sonuçlarının kabul edilebilirliği artar.

7. SONUÇ

Testler ve sınavlar, eğitim sistemlerinde öğrenme çıktılarının değerlendirilmesi, öğrencilerin bilgi, beceri ve yeterlik düzeylerinin belirlenmesi ve eğitim süreçlerine ilişkin kararların desteklenmesinde temel araçlar olarak kullanılmaktadır. Bununla birlikte, her sınavın belirli bir amaç ve hedef kitle doğrultusunda geliştirilmesi gerektiği unutulmamalıdır. Sınav puanlarından yapılacak çıkarımların kapsamının önceden net biçimde tanımlanması, ölçme sonuçlarının doğru yorumlanmasını sağlamakta ve belirli

bir amaç için geliştirilen sınavların farklı amaçlarla kullanılmasından kaynaklanabilecek geçerlik sorunlarının önüne geçmektedir.

Sınavlar, ortaya koydukları sonuçların özellikleri ve bu sonuçlara dayanarak verilen kararların önem derecesi bakımından birbirinden farklılık gösterir. Sonuçlarının bireyler ya da kurumlar üzerindeki etkisine göre sınavlar yüksek ve düşük riskli olarak sınıflandırılır. Sınavlar, kapsamları ve kullanım amaçları bakımından da farklı özellikler göstermektedir. Özellikle geniş ölçekli değerlendirmeler, çok sayıda katılımcıyı kapsamaları, yüksek maliyet ve karmaşık analiz süreçleri gerektirmeleri nedeniyle küçük ölçekli uygulamalardan ayrılmaktadır. Sınavlar, uygulama yöntemine göre geleneksel (kâğıt-kalem) ve elektronik (dijital, bilgisayar tabanlı) sınavlar olarak ikiye ayrılabilir. Özellikle teknolojik gelişmelerle uluslararası geniş ölçekli uygulamaların bilgisayar tabanlı olarak yapılması bu alandaki önemli gelişmelerdendir.

Etkili bir ölçme ve değerlendirme süreci yalnızca nitelikli test geliştirmeyi değil, aynı zamanda güvenli ve standartlaştırılmış sınav uygulamalarını da gerektirmektedir. Özellikle dijitalleşmeyle birlikte çevrimiçi sınavlarda ortaya çıkan güvenlik riskleri, sınav güvenliğine yönelik yeni önlemlerin geliştirilmesini zorunlu kılmıştır. Bunun yanında sınav uygulayıcılarının yeterli eğitim ve deneyime sahip olmaları, sınav protokollerine eksiksiz uymaları ve uygulama sürecini titizlikle yürütmeleri, ölçme sonuçlarının doğruluğu açısından kritik bir rol oynamaktadır.

Son olarak, sınav sonuçlarına yönelik itiraz mekanizmaları ölçme ve değerlendirme sistemlerinin ayrılmaz bir parçasıdır. Şeffaf, adil ve nesnel biçimde yürütülen itiraz süreçleri, katılımcı haklarının korunmasına katkı sağlamakta; aynı zamanda sınav sonuçlarının güvenilirliğini, kabul edilebilirliğini ve sınavı uygulayan kurumlara duyulan güveni artırmaktadır. Bu bağlamda, sınavların planlanmasından uygulanmasına, güvenliğinin sağlanmasından sonuçların raporlanması ve itirazların değerlendirilmesine kadar tüm süreçlerin bilimsel ilkelere dayalı olarak yürütülmesi, nitelikli ölçme ve değerlendirme uygulamalarının temel koşulu olarak değerlendirilebilir.

KAYNAKÇA | REFERENCES

Akyol, P., Krishna, K. & Wang, J. (2021). Taking PISA seriously: How accurate are low-stakes exams? *Journal of Labor Research*, *42*, 184–243. https://doi.org/10.1007/s12122-021-09317-8

Anderson, L. W. & Postlethwaite, T. N. (2007). *Program evaluation: Large-scale and small-scale studies*. International Institute for Educational Planning (UNESCO) *Education Policy Series*, *8*. https://unesdoc.unesco.org/ark:/48223/pf0000181752

Clarke, M. & Luna Bazaldua, D. (2021). *Primer on large-scale assessments of educational achievement*. World Bank.

Crocker, L. & Algina, J. (1986). *Introduction to classical & modern test theory*. Holt, Rinehart and Winston.

Doğan, N. & Aksu, G. (2021). Çevrimiçi (online) değerlendirme. N. Doğan (Ed.), *Eğitimde ölçme ve değerlendirme* (s. 362-390) içinde. Pegem Akademi.

Embretson, S. E. & Reise, S. P. (2000). *Item response theory for psychologists.* Lawrence-Earlbaum Associates.

Foster, D. (2016). Testing technology and its effects on test security. F. Drasgow (Ed.), *Technology and testing: Improving educational and psychological measurement*, (s. 235-254) içinde. Routledge. https://doi.org/10.4324/9781315871493-11

International Test Commission (2014). *International guidelines on the security of tests, examinations, and other assessments*. [www.intestcom.org]

İlgün Dibek, M. (2021). Ölçme araçlarının sınıflandırılması. Ö. Yıldırım & S. Kula Kartal (Ed.), *Eğitimde ölçme ve değerlendirme* (s. 91-123) içinde. Lisans Yayıncılık.

Kalender, İ. (2009). Başarı ve yetenek kestirimlerinde yeni bir yaklaşım: Bilgisayar ortamında bireyselleştirilmiş testler (Computerized adaptive tests - CAT). *CITO Eğitim: Kuram ve Uygulama Dergisi, 5*, 39-48.

Kubiszyn, T. & Borich, G. (2016). *Educational testing and measurement* (11. baskı). Wiley, Hoboken.

Linn, R. L. (2008). *Measurement and assessment in teaching*. Pearson Education India.

Mehrens, W. A. & Lehmann, I. J. (1991). *Measurement and evaluation in education and psychology*. Houghton Mifflin Company.

MEB. (2019). *Akademik Becerilerin İzlenmesi ve Değerlendirilmesi: 8. Sınıflar Raporu*. https://abide.meb.gov.tr/meb_iys_dosyalar/2025_01/20161850_abide20188.sinifrapor.pdf adresinden 17 Temmuz 2025 tarihinde alınmıştır.

MEB. (2021). *PIRLS 2021 broşürü.* https://pirls.meb.gov.tr/meb_iys_dosyalar/2021_01/12204200_PIRLS_2021_BroYuru.pdf adresinden 17 Temmuz 2025 tarihinde alınmıştır.

MEB. (2024). *TIMSS 2023 Türkiye Raporu*. https://odsgm.meb.gov.tr/meb_iys_dosyalar/2024_12/04111224_timss_2023_rapor_0412.pdf adresinden 17 Temmuz 2025 tarihinde alınmıştır.

MEB. (2025a). Millî Eğitim Bakanlığı merkezî sistem sınav yönergesi. *https://mevzuat.meb.gov.tr/dosyalar/2244.pdf* adresinden 20 Eylül 2025 tarihinde alınmıştır.

MEB. (2025b). İnternet duyurusu. İtiraza İlişkin Sınavlar. https://eitiraz.meb.gov.tr/ adresinden 20 Eylül 2025 tarihinde alınmıştır.

ÖSYM. (2016). ÖSYM'de 'sınav gözlemcisi' dönemi. https://www.osym.gov.tr/TR,12785/osymde-sinav-gozlemcisi-donemi-12052016.html adresinden 18 Temmuz 2025 tarihinde alınmıştır.

ÖSYM. (2011). *Sınav salonlarında kamera ile kayıt uygulaması* [Basın duyurusu]. http://osym.gov.tr/TR,1034/sinav-salonlarinda-kamera-ile-kayit-uygulamasi-16032011.html adresinden 18 Temmuz 2025 tarihinde alınmıştır.

ÖSYM. (2013a). *Sınav uygulamalarına ilişkin güvenlik tedbirleri* [Duyuru]. https://www.osym.gov.tr/TR,630/sinav-uygulamalarina-iliskin-guvenlik-tedbirleri-30012013.html adresinden 18 Temmuz 2025 tarihinde alınmıştır.

ÖSYM. (2013b). *Soruların hazırlanması, soru havuzunun oluşturulması ve soruların güvenliğinin sağlanmasına ilişkin yönetmelik*. Resmî Gazete. https://www.mevzuat.gov.tr/mevzuat?MevzuatNo=17289&MevzuatTur=7&MevzuatTertip=5 adresinden 19 Temmuz 2025 tarihinde alınmıştır.

ÖSYM. (2025a). e-TEP technical report: e-TEP 2025/1 Evaluation Analyses. ÖSYM. https://www.osym.gov.tr/TR,33659/e-tep-2025-1-technical-report.html adresinden 20 Eylül 2025 tarihinde alınmıştır.

ÖSYM. (2025b). YKS soru kitapçıklarının sınav merkezlerine sevkiyatı başladı [Duyuru]. https://www.osym.gov.tr/TR,33287/yks-soru-kitapciklarinin-sinav-merkezlerine-sevkiyati-basladi-20062025.html adresinden 20 Eylül 2025 tarihinde alınmıştır.

ÖSYM. (2025c). İnternet duyurusu. Adaylar tarafından dilekçe gönderilmesi ve işlem ücretleri. https://www.osym.gov.tr/TR,32938/adaylar-tarafindan-dilekce-gonderilmesi-ve-islem-ucretleri-09012025.html adresinden 20 Eylül 2025 tarihinde alınmıştır.

Popham, W. (2001). Teaching to the test. *Educational Leadership*, *58*, 16-20.

Resmî Gazete. (2011). *6114 sayılı Ölçme, Seçme ve Yerleştirme Merkezi Hizmetleri Hakkında Kanun*. https://www.mevzuat.gov.tr/mevzuat?MevzuatNo=6114&MevzuatTur=1&MevzuatTertip=5 adresinden 2 Kasım 2025 tarihinde alınmıştır.

Schmeiser, C. B. & Welch, C. J. (2006). Test development. Brennan, R. L. (Ed.), *Educational measurement* (s. 307-354) içinde. Westport, CT: Praeger.

Sindre, G. & Vegendla, A. (2015, Kasım). *E-exams and exam process improvement.* The UDIT / NIK 2015 conference, Ålesund, Norveç.

YAZAR HAKKINDA | ABOUT AUTHOR

PROF. DR. SEHER YALÇIN

Ankara Üniversitesi | yalcins[at]ankara.edu.tr

ORCID: 0000-0003-0177-6727

Prof. Dr. Seher Yalçın, Ankara Üniversitesi Eğitim Bilimleri Fakültesi Ölçme ve Değerlendirme Anabilim Dalında öğretim üyesidir ve aynı zamanda üniversitenin Ölçme ve Değerlendirme Uygulama ve Araştırma Merkezi Müdürü olarak görev yapmaktadır. Lisansüstü eğitimini Ankara Üniversitesi Eğitimde Ölçme ve Değerlendirme alanında tamamlayan Yalçın, 2011'den bu yana akademik çalışmalarını sürdürmektedir. 2016 yılında Hollanda Tilburg Üniversitesi Yöntem ve İstatistik Bölümünde ziyaretçi araştırmacı olarak bulunmuş, burada ileri ölçme modelleri ve istatistiksel yöntemler üzerine çalışmalar yapmıştır. 2019 yılında doçent, 2025 yılında profesör unvanını almıştır. Uzmanlık alanları arasında bilgisayar ortamında bireye uyarlanmış testler, madde yanlılığı ve geniş ölçekli değerlendirmeler yer almaktadır. Uluslararası saygın dergilerde çok sayıda makalesi yayımlanmış; ulusal ve uluslararası birçok projede araştırmacı olarak görev almış ve bu çalışmalarıyla çeşitli akademik ödüllere layık görülmüştür. Ayrıca ölçme ve değerlendirme, araştırma yöntemleri, istatistik ve geniş ölçekli uygulamalarla ilgili lisans ve lisansüstü düzeyde dersler de vermektedir.

PROF. DR. SEHER YALÇIN

Ankara University | yalcins[at]ankara.edu.tr

ORCID: 0000-0003-0177-6727

Prof. Dr. Seher Yalçın is a faculty member in the Department of Measurement and Evaluation at the Faculty of Educational Sciences, Ankara University. She also serves as the Director of the Measurement and Evaluation Application and Research Center at Ankara University. Prof. Yalçın completed her graduate studies in Educational Measurement and Evaluation at Ankara University and has been actively engaged in academic research since 2011. In 2016, she was a visiting researcher at the Department of Methodology and Statistics at Tilburg University in the Netherlands, where she conducted research on advanced measurement models and statistical methods. She was awarded the title of Associate Professor in 2019 and promoted to Full Professor in 2025. Her research interests include computerized adaptive testing, item bias, and large-scale assessments. She has published numerous articles in prestigious international journals and has participated as a researcher in many national and international research projects. Her scholarly contributions have been recognized with various academic awards. In addition to her research activities, she teaches undergraduate and graduate courses in measurement and evaluation, research methods, statistics, and large-scale assessment applications.

10. BÖLÜM

PUANLAMA YÖNTEMLERİ

SCORING METHODS

Doç. Dr. Murat Doğan ŞAHİN

PUANLAMA YÖNTEMLERİ

Doç. Dr. Murat Doğan ŞAHİN
Anadolu Üniversitesi

Öz

Bu bölümde eğitimde kullanılan puanlama yöntemleri, puanın nasıl üretildiğini merkeze alan epistemolojik bir bakış açısıyla ele alınmıştır. Bu doğrultuda yöntemler; puanlama sürecine dayalı yaklaşımlar, otomatik puanlama ve karşılaştırma yargılarıyla puanlama olmak üzere üç temel başlık altında incelenmiştir. İlk bölümde manuel ve dijital puanlama, ikili ve kısmî puanlama ile ağırlıklı puanlama yaklaşımları; kullanım alanları, avantajları, sınırlılıkları ve psikometrik özellikleri bakımından değerlendirilmiştir. Ayrıca puanlama sürecinin ölçme amacına, ölçülen özelliğe ve uygulama koşullarına göre nasıl farklılaşabileceği tartışılmıştır. İkinci bölümde otomatik puanlama sistemlerinin tarihsel gelişimi, geleneksel kural ve özellik temelli modellerden derin öğrenme ve üretken yapay zekâ temelli yaklaşımlara uzanan dönüşüm çerçevesinde açıklanmıştır. Kural temelli, makine öğrenmesi, sinir ağı ve ontoloji temelli yöntemlerin çalışma ilkeleri özetlenmiş; CNN, LSTM ve Transformer mimarilerinin yazılı performansların değerlendirilmesindeki rolleri ele alınmıştır. Büyük dil modellerinin otomatik puanlamaya sağladığı olanaklar, zero-shot, few-shot ve fine-tuning yaklaşımları kapsamında incelenmiş; otomatik puanlama performansının değerlendirilmesinde *QWK* başta olmak üzere kullanılan ölçütler ve bu ölçütlerin yorumlanmasına ilişkin sınırlılıklar tartışılmıştır. Son bölümde ise karşılaştırma yargılarıyla puanlama yaklaşımı tanıtılmış; mutlak puanlama anlayışından farklı olarak performansların ikili karşılaştırmalar yoluyla değerlendirilmesine dayanan bu yöntemin kuramsal temelleri açıklanmıştır. Böylece bölüm, geleneksel insan puanlamasından yapay zekâ destekli sistemlere ve karşılaştırmalı yargı temelli yaklaşımlara kadar güncel puanlama yöntemlerini bütüncül bir bakış açısıyla değerlendirerek, farklı ölçme amaçları ve uygulama koşulları için uygun puanlama yaklaşımlarının seçimine yönelik kuramsal ve uygulamalı bir çerçeve sunmaktadır.

Anahtar Kelimeler

Manuel ve Dijital Puanlama, Kısmî Puanlama, Ağırlıklı Puanlama, Otomatik Puanlama, Üretken Yapay Zekâ, Karşılaştırma Yargılarıyla Puanlama

SCORING METHODS

Assoc. Prof. Dr. Murat Doğan ŞAHİN
Anadolu University

Abstract

This chapter examines scoring methods used in educational assessment from an epistemological perspective that focuses on how scores are generated. Accordingly, the methods are organized into three main categories: process-based scoring approaches, automated scoring, and comparative judgement. The first section discusses manual and digital scoring, dichotomous and partial-credit scoring, and weighted scoring approaches in terms of their applications, advantages, limitations, and psychometric properties. It also examines how scoring procedures vary according to the purpose of assessment, the construct being measured, and the conditions under which the assessment is administered. The second section explains the historical development of automated scoring systems, tracing their evolution from traditional rule-based and feature-based models to deep learning and generative artificial intelligence-based approaches. It summarizes the operating principles of rule-based, machine learning, neural network, and ontology-based methods, and discusses the roles of CNN, LSTM, and Transformer architectures in the evaluation of written performances. The opportunities provided by large language models for automated scoring are examined within the context of zero-shot, few-shot, and fine-tuning approaches. In addition, the chapter discusses the metrics used to evaluate automated scoring performance —particularly the *QWK* — and the limitations associated with interpreting these metrics. The final section introduces the comparative judgement approach to scoring and explains its theoretical foundations as an alternative to absolute scoring, in which performances are evaluated through pairwise comparisons. Overall, the chapter provides a comprehensive perspective on contemporary scoring methods, ranging from traditional human scoring to AI-assisted systems and comparative judgement approaches. It offers both a theoretical and practical framework for selecting appropriate scoring methods for different assessment purposes and implementation contexts.

Keywords

Manual and Digital Scoring, Partial-credit Scoring, Weighted Scoring, Automated Scoring, Generative Comparative Judgement

1. GİRİŞ

Bir performansın puanlanmasında kullanılabilecek çeşitli yöntemler bulunmaktadır. Hangi yöntemin kullanılacağı ölçmeye konu olan özellik başta olmak üzere ölçme işleminin amacı, grubun büyüklüğü ve değerlendirme işlemini yapan kurumun-bireyin operasyonel kabiliyeti gibi bir dizi değişkene göre farklılaşabilmektedir. Bu bölümde, eğitimde ve ölçme değerlendirme alanında sık kullanılan puanlama yöntemleri incelenmiştir. Söz konusu yöntemleri farklı başlıklarda ele almak mümkündür. Bu bölümde puanlama süreçlerini tarihsel olarak sınıflamak (örneğin, geleneksel ve yenilikçi gibi) tercih edilmemiştir. Çünkü teknolojideki hızlı gelişime paralel olarak bugün yenilikçi görünen bir uygulamanın kısa sürede yerini başka uygulamalara bırakması ve birkaç yıl içinde geleneksel olarak ifade edilmesi muhtemeldir. Buna bağlı olarak mevcut bölümde puanlama yöntemleri "puanlama sürecine dayalı yaklaşımlar", "otomatik puanlama" ve "karşılaştırma yargılarıyla puanlama" olmak üzere üç kısımda sunulmuştur. Bu ayrım, puanlama sürecinin epistemolojik olarak ele alınmasıyla ortaya çıkmıştır. Puanlama yaklaşımlarının epistemolojisi merkeze alındığında yöntemler teknolojiye veya uygulama biçimine göre kategorize edilmez; puanın bir bilgi olarak nasıl üretildiği, hangi varsayımlara dayandığı ve performansa ne tür anlamlar yüklediği temele alınarak bir sınıflandırma yapılır.



Puanlama süreçleri, cevaplayıcı kararlarına dayalı olarak da yürütülebilir. Bu puanlama biçimi sınava katılanların kesinlikle yanlış olduğuna yönelik düşüncelerini ya da işaretledikleri seçeneğin doğruluğuna olan güvenlerini esas alır (Özdemir, 2003). Bu bölüm kapsamında yanıtlayıcıları merkeze alan yöntemlere yer verilmemiştir.

2. PUANLAMA SÜRECİNE DAYALI YAKLAŞIMLAR

Puanlama sürecine dayalı yaklaşımlar; puanın önceden belirlenmiş kurallara göre üretildiği, karar akışının açık, öngörülebilir ve şeffaf olduğu puanlama sistemlerini ifade eder. Bu yaklaşımda puan, algoritmik tahminlerden değil, tanımlı bir değerlendirme sürecinin sonucundan türetilir. Bu nedenle belirlenen puanlama yöntemi kâğıt üzerinde ya da dijital ortamda olması fark etmeksizin aynı biçimde uygulanır. Bir başka ifadeyle, puanlamanın yapıldığı ortamın yanı sıra kullanılan puanlama sürecinin niteliği ve puanın nasıl üretildiğine ilişkin ilkeler bu yaklaşımın sınıflandırılmasında belirleyici rol oynar. Bu doğrultuda puanlama sürecine dayalı yaklaşımlar puanın verildiği ortam, puanın veriliş biçimi olmak, otomatik puanlama ve karşılaştırma yargılarıyla puanlama olmak üzere dört kısımda incelenmiştir.

2.1. Puanlama Yapılan Ortama Göre Puanlama Türleri

Bu başlık altında, puanlamanın geleneksel metotlarla (kâğıt-kalem vb.) ya da dijital ortamlarda yapılmasını esas alan yöntemlere değinilmiştir.

2.1.1. Manuel Puanlama

Manuel puanlama, öğrenci yanıtlarının veya performanslarının doğrudan bir insan tarafından puanlanmasıdır. Bu yöntemde puanlayıcı, yanıtları inceleyerek önceden belirlenen puanlama ölçütlerine göre karar verir. Açık uçlu yanıtlar, üretimsel dil becerileri (yazma ve konuşma), sunumlar, yaratıcılık gerektiren performanslar ve uygulamalı performans görevleri gibi ürünlerin değerlendirilmesi insan puanlayıcılar tarafından yapılabilir. Özellikle yapılandırılmış eğitim ortamlarındaki puanlama işlemleri genellikle bu yolla yapılır.

Puanlama sürecine esneklik kazandırması manuel puanlamanın kayda değer bir üstünlüğüdür. Ancak bu yöntemin bazı sınırlılıkları da vardır. Öncelikle, süreç zaman alıcıdır ve büyük öğrenci gruplarında uygulanması oldukça zahmetlidir. Ayrıca, puanlamada subjektiflik ortaya çıkabilir; aynı performansın farklı puanlayıcılar tarafından farklı şekilde puanlanması, güvenirlik sorunlarına yol açabilir. Bu tür sorunları azaltmak için puanlama anahtarları, kontrol listeleri, dereceleme ölçekleri veya rubriklerden (dereceli puanlama anahtarlarından) yararlanılması önerilir. Fakat özellikle K-12 düzeyinde yapılan performans değerlendirmelerinde bir puanlama anahtarına bağlı kalmaktan ziyade genel izlenimle puanlama yaygındır. Bu da sonuçların adilliği ve objektifliği konusunda endişeleri beraberinde getirmektedir.

Manuel puanlamada yoğun bir insan emeği söz konusudur; puanlama sürecindeki tüm işlemler insan eliyle yapılır. Puanlanacak performans sayısının az olduğu, nispeten düşük önem arz eden uygulamalarda ve dijital araçlardan yararlanmak mümkün olmadığında manuel puanlama yapılması makul görülebilir. Manuel puanlama, teknolojik altyapının yetersiz olduğu durumlarda zorunlu bir seçenek olmasının yanında, özellikle üst düzey düşünme becerilerinin ölçülmesinde ve nitelikli geri bildirim sunulmasında hâlâ önemli bir yere sahiptir. Nitekim bölümün devamında üzerinde durulacak yapay zekâ uygulamalarının üst düzey becerilerinin değerlendirilmesinde kullanılabilmesi çok maliyetli olabilmekte ve kimi zaman bu maliyetlere rağmen beklenen sonuçlar alınamamaktadır.

2.1.2. Dijital Puanlama

Dijital değerlendirme veya elektronik değerlendirme; sınavların, testlerin, anketlerin ya da ölçme işlemlerinin bilgisayar, tablet ve cep telefonu gibi dijital aygıtlar aracılığıyla sunulmasıdır (Assessment Systems, 2025). Sınavların dijitalleştirilmesi, yenilikçi ölçme-değerlendirme tasarımlarını kullanmayı mümkün kılmaktadır (OECD Digital Education Outlook, 2023). Bu durum puanlama süreçlerine de yansımaktadır. Dijital puanlama, öğrenci performanslarının dijital araçlar, yazılımlar veya sistemler aracılığıyla puanlanmasını ifade eder. Bu yöntem, değerlendirme sürecinde insan puanlayıcıların rolünü azaltabilir, destekleyebilir ya da tamamen ortadan kaldırabilir. Dijital puanlama sistemleri, teknolojik altyapının gelişmesiyle birlikte özellikle büyük ölçekli sınavlarda, çevrim içi ölçme ortamlarında ve merkezî değerlendirme uygulamalarında yaygın olarak kullanılmaktadır.

Dijital puanlama, zaman ve maliyet açısından verimlilik sağlayan, çok sayıda verinin kısa sürede işlenmesine olanak tanıyan bir yaklaşım sunar. Bu bağlamda kullanışlılığı artıran bir yöntem olduğu söylenebilir. Bununla birlikte, sistemin doğruluğu ve güvenilirliği kullanılan yazılımın niteliğine, algoritmaların hassasiyetine ve sistemin tasarımına bağlıdır.

2.1.2.1. Optik Okuyucu ile Puanlama

Genellikle çoktan seçmeli testlerde kullanılan, yanıtların özel formlara işaretlendiği ve bu formların optik okuyucu cihazlar tarafından otomatik olarak değerlendirildiği bir dijital puanlama yöntemidir. Bu sistem, özellikle kalabalık gruplara yapılan uygulamalarda kısa sürede ve hatasız biçimde puanlamanın esas olduğu durumlar için ideal bir çözümdür. Bu da çoktan seçmeli maddeleri geniş ölçekli sınavlar için oldukça kullanışlı hale getirir.

Optik okuyucu sistemleri, işaretlenen cevapları belirli bir anahtarla karşılaştırarak doğru ve yanlışları belirler; böylece hızlı ve objektif bir puanlama süreci sağlar. İnsan hatasını en aza indirir, değerlendirme sürecini büyük ölçüde otomatikleştirir ve raporlama işlemlerini kolaylaştırır. Bununla beraber optik okuyucu sistemlerinin sınırlı bir değerlendirme kapasitesi vardır. Sadece seçme gerektiren maddeleri (Bkz. 6. Bölüm) işleyebildiği için açık uçlu, yorum gerektiren, yaratıcı ya da analitik düşünme becerilerini ölçmeye yönelik yanıtları puanlayamaz. Sonuç olarak, optik okuyucu ile puanlama; hız, doğruluk ve maliyet açısından avantajlı bir sistem sunmakla birlikte madde türü çeşitliliği bakımından sınırlayıcıdır. Bundan dolayı, bu yöntemin uygunluğu ölçülmek istenen beceri düzeyi ve sınavın amacı doğrultusunda dikkatle ele alınmalıdır.

2.1.2.2. Performansın Dijital Ortamda Puanlanması

Bazen ölçme işlemine konu olan özellikleri çoktan seçmeli maddelerle ölçmek mümkün olmaz. Örneğin, yazma ya da konuşma gibi performans gerektiren özellikler çoktan seçmeli maddeler ile ölçülemez. Söz konusu özellikler performans görevleri aracılığıyla ölçülürken sınavın tüm süreçleri ya da sadece değerlendirme basamağı dijital ortamda gerçekleştirilebilir. Özellikle bilgisayar tabanlı sınavlarda öğrencinin yanıtı dijital ortamda sözlü ya da yazılı olarak (ölçülen özelliğe göre) alınır; böylece değerlendirme süreci de çoğunlukla rubriklere dayalı olarak bu dijital ortamda ilerler.

Kimi zaman ise sınav kâğıt-kalem biçiminde yapılır ancak yanıtlar dijital ortama aktarılarak sadece puanlama süreci bu ortamda gerçekleşir. İster tüm süreç isterse yalnızca puanlama aşaması dijital ortamda yürütülsün performansın belirli düzeyleri olduğunda kısmî puanlamalar yapılabilir. Burada, değerlendirme otomasyonunun özellikleri ön plana çıkar. Buna göre sınavın önem düzeyi dikkate alınarak her bir yanıt sadece bir ya da birden çok puanlayıcı tarafından puanlanabilir. Yüksek risk taşıyan sınavlarda puanlamalar genelde birden fazla puanlayıcı tarafından gerçekleştirilir. Puanlayıcı sayısının birden fazla olduğu bu tür durumlarda, puanlayıcıların birbirinden bağımsız ve bir rubriğe dayalı olarak yaptıkları puanlamalar arasındaki uyuma göre süreç devam eder ve performansa nihai puan verilir. Sınıf içi ölçme-değerlendirme süreçlerinde ise çoğu zaman sadece bir puanlayıcı performansa değer biçer.

Özellikle bilgisayar tabanlı sınavlarda, madde türü çoktan seçmeli ya da doğru-yanlış türünde olduğunda optik okuyucuya gerek kalmaz; öğrenci yanıtladığı anda yanıtın doğru/yanlış olduğu değerlendirilebilir. Bu özellikleriyle dijital ortamlar çoktan seçmeli testlerin bireye uyarlanmış biçimde yapılabilmesini mümkün kılar.

2.2. Puanın Veriliş Biçimine (Puanlama Kuralına) Göre Puanlama Türleri

Puanlama sürecinde işletilen kural ya da başka bir ifadeyle puanın veriliş biçimi, kullanılan ölçme aracının özelliğiyle ilgilidir. Ölçme aracında çoktan seçmeli, doğru-yanlış ya da açık uçlu madde türlerinden hangisine yer verildiği puanlama biçimini etkiler.

2.2.1. İki Kategorili Puanlama (1-0 Puanlama)

Kısmi puanlamanın kullanılmadığı durumlarda maddeye verilen yanıt ya tam doğru ya da tamamen yanlış kabul edilir. Bu tür durumlarda elde edilen veri türü "1-0" olarak anılır. Yanıt doğruysa 1, yanlışsa veya yanıt yoksa 0 puan verilir (Kurz, 1999). Doğru-yanlış türündeki maddeler 1-0 şeklinde puanlanır. Bunun yanında çoktan seçmeli ve kısa cevaplı maddelerde de çoğunlukla 1-0 şeklinde puanlama yapılır. Maddelerin bu şekilde puanlandığı ölçme araçlarında toplam puanı hesaplamak için doğru yanıt sayısı her doğru yanıt için ön görülen katsayıyla çarpılır. Şans başarısını azaltmak amacıyla düzeltme işlemi uygulandığında ise seçenek sayısına bağlı olarak her yanlış yanıt, doğru

sayının belirli bir yüzdede düşmesine neden olur. Böylece "net", yani düzeltilmiş puan hesaplanır ve sonrasında belirlenen katsayıyla çarpılır. 1-0 şeklinde puanlanan maddeler genelde aynı ağırlığa sahiptir; daha açık bir anlatımla her maddenin nihai test puanına katkısı eşittir.

2.2.2. Kısmi Puanlama (Çok Kategorili Puanlama)

Özellikle geniş ölçekli uygulamalarda kullanışlılık ilkesi gereği çoğunlukla 1-0 şeklinde puanlanan maddelere başvurulur. Bu tür puanlama, ölçülen özelliğin bireyde ya tam olarak var olduğunu ya da hiç olmadığını, bir başka ifadeyle bireyin ya başarılı ya da başarısız olduğunu kabul eder. Ancak çoğu zaman soruyla yoklanmak istenen özellik, bireylerde farklı düzeylerde görülebilir. Örneğin, yazılı ya da sözlü anlatım becerisinin ölçülmesi hedefleniyorsa 1-0 şeklinde puanlanan bir madde türü kullanmak geçerlik problemi yaratır. Dolayısıyla bu tür beceriler ölçülürken performans görevlerinin kullanılması ve kısmi puanlama yapılması tercih edilir.

Öğrenci performansının farklı düzeylerde puanlanmasına kısmî puanlama adı verilir. Önceden hazırlanan bir cevap anahtarı ya da dereceli puanlama anahtarına dayanarak bireylerin yanıtları farklı düzeylerde puanlanabilir. Kısmi puanlamanın düzeyleri, ölçme işleminin amacına ve ölçülen özelliğe göre farklılık gösterebilir. Kimi zaman bir performans "yanlış-geçersiz yanıt, kısmen doğru, tam doğru" olmak üzere üç seviyeye ayrılır. Üst düzey becerilerin ölçüldüğü durumlarda ise düzey sayısı artabilir. Kısmi puanlamada düzey sayısı ölçülen özellik kadar, sınavın amacına ve eldeki imkânlara da bağlıdır. Sınıf içinde yapılan ölçme-değerlendirme işlemlerinde daha esnek bir puanlama yaklaşımı mümkünken geniş ölçekli uygulamalarda, özellikle yüksek riskli sınavlarda, puanlama güvenirliğini yüksek tutmak amacıyla düzey sayısı daha katı biçimde belirlenir. Bu noktada duyarlılık ile birbirinden net biçimde ayrılabilecek düzey sayısı arasında bir denge gözetmek gerekir.

> Anahtarla puanlamada bütünsel ve analitik yaklaşım tercih edilebilir. Ölçme işlemine konu olan özellik bütünsel değerlendirmeye uygunsa holistik (bütünsel) bir puanlama aracı kullanılır. Öte yandan özellik aşamalık arz ediyor ya da alt bileşenlerden oluşuyorsa bunlardan her birini ayrı ayrı puanlamak gerekir ki bu durumda analitik bir puanlama aracı tercih edilir (Kula-Kartal, 2021).

Kısmi puanlama, çoktan seçmeli maddelerde de kullanılabilir; buna göre tam doğru seçeneğin yanında kısmi puanlama yapılacak seçenekler de belirlenebilir. Doğru yanıtı verene kadar yanıtlama, tamamen yanlış bilgiden başlayarak çeşitli düzeylerde kısmi ve tam bilgiye ulaşmayı esas alan eleme puanlaması ve adaylara seçeneklerin herhangi bir sayıda doğru olabileceğinin bildirildiği çoklu seçim formatı (Bkz. 6. Bölüm), çoktan seçmeli maddelerde kullanılması mümkün olan kısmi puanlama yöntemleridir (Kurz, 1999). Uygulanmasındaki karmaşıklığa karşın güvenirlik ve geçerlik açısından getirilerinin sınırlı olması bu metotların yaygınlaşmasına engel teşkil etmektedir.

2.2.3. Ağırlıklı Puanlama

Klasik Test Kuramı'na (KTK) dayalı uygulamalarda çoktan seçmeli maddeler, yer aldığı alt testte genel olarak eşit ağırlığa sahiptir. Fakat bazen testteki maddelerden bazılarının diğerlerine göre daha yüksek ya da daha düşük ağırlığa sahip olması gerekli görülür. Bu gereklilik, çoktan seçmeli testlerin 1-0 şeklinde puanlanmasının getirdiği sınırlılıklardan kaynaklanır. Abu-Sayf (1979), dört temel sınırlılıktan bahsetmektedir. Bunlardan ilki, testin yönergelerine dayanarak tahmini teşvik etmenin görece yüksek oranda hata varyansı oluşturduğu ve bu formülün kısmi bilgiyi hesaba katmadığı yönündeki psikometrik argümandır. Pragmatik argüman, temkinli davranarak bilmediği soruları boş bırakan adayların risk alanlara kıyasla cezalandırıldığı yönündedir. Ahlaki argüman tahmin yürütmenin yanlış ve bu davranışı ödüllendirmenin sakıncalı olduğu görüşüne dayanırken siyasi argüman adayları tahminde bulunmaya teşvik etmenin çoktan seçmeli testlere olan güvenlerini kaybetmelerine yol açacağını iddia eder. Bunun yanında kimi maddelerin diğerlerinden daha kritik becerileri ölçmesi, daha zor veya ayırt edici olması ya da ikili puanlamanın yanlış yapan tüm yanıtlayıcıları ölçülen beceriden aynı düzeyde yoksun sayması (Rowley & Traub, 1977) gibi sınırlılıklar çoktan seçmeli maddelerde ağırlıklandırmayla puanlamaya yönelik çalışmalara zemin hazırlamıştır. 1900'lü yılların başlarından itibaren bu konudaki araştırmalar hız kazanmıştır. Bununla birlikte Stanley ve Wang (1970) o güne değin yapılan çalışmaları incelediğinde ağırlıklandırmanın test puanlarının güvenirlik ve geçerliğinde kayda değer bir iyileşme yaratmadığını ortaya koymuştur. Ağırlıklı puanlama genellikle önsel ve sonsal olmak üzere iki başlıkta ele alınır.

2.2.3.1. Önsel Ağırlıklı Puanlama

Kimi zaman uzmanlar, testte yer alan maddelerin bir kısmının daha kritik becerileri ölçtüğü ya da ilgili maddeleri yanıtlayabilmek için daha uzun süreli/üst düzey bir performans ortaya konulması gerektiği görüşünde olurlar. Ya da test sadece çoktan seçmeli maddelerden oluşmuyor, açık uçlu maddeler de içeriyor olabilir. Böyle durumlarda, uygulama öncesinde uzman görüşlerine dayanarak maddelerin nihai puanlamadaki ağırlıkları belirlenebilir ve kararlaştırılan bu ağırlıklar doğrultusunda her birey için nihai puan hesaplanır. Ancak bu yöntemde belirlenen ağırlıkların keyfî olmadığından emin olunmalıdır. Bunun için tek bir alan uzmanının görüşünden hareketle karar vermekten kaçınılmalı, birden fazla uzmanın görüşüne başvurulmalıdır.

Önsel olarak özellikle çok kategorili puanlama ile ağırlıklandırılan maddelerin güvenirliğe etkisi tartışma konusudur. Benton (2018) yaptığı geniş çaplı araştırmada, çok kategorili puanlanan maddelerin aynı toplam puan değerinde ancak bireysel olarak puanlanmış maddelere kıyasla (söz gelimi, 0-4 arası puanlanabilen tek bir madde ile her biri 1-0 şeklinde puanlanan dört madde) güvenirliğe katkısının daha düşük olduğunu belirlemiştir. Düşük puanlı maddelerin ağırlığı artırılıp çok kategorili olanların

azaltılmasıyla elde edilen puanların ilk puanlarla korelasyonun çok yüksek olduğu, düzeltme sonucunda ulaşılan çıktıların böyle bir işlemi yapmaya gerekçe oluşturmayacağı ilgili çalışmada rapor edilen dikkat çekici bir diğer bulgudur.

2.2.3.2. Sonsal Ağırlıklı Puanlama

Testteki maddelerin ağırlıklandırılmasına ilişkin önsel güçlü bir dayanak olmadığında ya da bu tür bir dayanağın keyfiyet arz edeceği düşünülüp ampirik kanıtlara gereksinim duyulduğunda ağırlıklandırma için testin uygulanmasıyla elde edilen istatistiklerden yararlanılabilir. Sonsal ağırlıklandırma denilen bu yöntemde, maddeler testin uygulanması sonucunda ulaşılan istatistiklere dayalı olarak ağırlıklandırılır. Burada akla ilk gelen yöntem, maddenin ayırt edicilik değeri ile puanın çarpılmasıdır. Böyle bir durumda madde 1-0 şeklinde puanlanmaz. Yanlış yanıta 0 puan verilirken doğru yanıta 0-1 arasında kalan ve değeri maddenin ayrıt edicilik indeksine (*r*) eşit olan bir puan atanır. Birnbaum (1968) çalışmasında bireylerin test puanlarını, doğru yanıtlarını madde ayıt ediciliğiyle çarparak elde etmiştir. Bir başka yöntem ise maddenin güçlüğünü esas almaktır. Bu durumda madde güçlük indeksi (*p*) arttıkça, yani madde kolaylaştıkça bu maddeden alınan puan düşer; tersine madde zorlaştıkça madde puanı artar. Ancak bir maddenin güçlük değeri onun işlerliğine ilişkin tek başına bilgi vermediğinden bu yöntemin kullanımı konusunda çekinceler vardır. Günümüzde hala yaygın olarak başvurulan bir başka yöntem şans başarısının düzeltme amacıyla kullanımına dayanır. Bu yöntemde bireyin yanlış sayısı testteki maddelerin seçenek sayısının bir eksiğine (örneğin, 5 seçenekli bir madde için 4) bölünür ve elde edilen değer doğru sayısından çıkarılarak testte puanlamaya esas teşkil eden "net" belirlenir. Kelley (1947) ise sonsal ağırlıklandırma için testten alınan puanlara ait güvenirlik kestiriminin kullanıldığı bir regresyon modeli önermiştir (akt: Yurdugül, 2010).

Benton (2018), maddeleri ağırlıklandırırken güvenirliğin maksimize edilmesini esas alan bir yöntem kullanmıştır. Elde ettiği sonuçlar, güvenirlikteki yükselmenin testin yordama geçerliği açısından çoğunlukla bir artışla sonuçlanmadığını göstermiştir. Kılıç ve Doğan (2019) ise yeni bir ağırlıklandırma yöntemi önermiştir. Buna göre bireyin testteki doğru sayısı testte yer alan madde sayısına bölünür ve bu bölme işleminden elde edilen değere maddenin güçlüğü eklenir. Hesaplanan nihai değer 1'in altında çıkarsa bireyin aldığı puan 1 ya da 0 olarak aynen devam eder, herhangi bir işlem yapılmaz. Nihai değer 1'e eşit ya da 1'den büyükse bireyin ilgili maddeyi doğru yanıtlayıp yanıtlamadığına göre hareket edilir; birey maddeye doğru yanıt vermişse 1'in üzerine maddenin güvenirlik değeri eklenirken yanlış yanıt vermişse herhangi bir işlem yapılmaz. Söz gelimi bir maddenin güçlük değeri 0,36, ayırt edicilik indeksi 0,56 ve bireyin testteki ortalaması (doğru yanıtlarının sayısı/testteki madde sayısı) 0,70 olsun. Bu durumda Güçlük + Bireyin test ortalaması (0,36 + 0,7 = 1,06) 1'in üzerinde çıkar. Dolayısıyla bireyin doğru yaptığı bu madde için 1'in üzerine madde güvenirliği eklenir ve bireyin maddeden aldığı puan 1+0,27=1,27 olur. Araştırmacılar, bu yöntemin testin güvenirliğine ve açıklanan varyans oranına katkıda bulunduğu sonucuna ulaşmışlardır.

Yaşar ve arkadaşları (2021) kullandıkları yöntemde maddelerin güçlük indeksine dayalı bir ağırlıklandırma yapmayı denemişler ve uyguladıkları ağırlıklandırmanın testin aritmetik ortalamasında düşüş, standart sapmasında ise artış ile sonuçlandığını rapor etmişlerdir. Diğer bir deyişle bu yöntem uygulandığında testin bireyler arasındaki farklılıkları ortaya koymada daha başarılı olduğunu ortaya koymuşlardır.

KTK uygulamalarında ağırlıklandırma, sınıf içi uygulamalar ya da düşük önem arz eden sınavlar dışında pek tercih edilen bir yöntem değildir. Bir maddenin güçlük ya da ayırt edicilik indeksine göre puanlamaya etkisinin farklı olması gerektiği düşünülüyorsa bir diğer ölçme kuramı olan Madde Tepki Kuramı'ndan (MTK) yararlanmak gerekir. Sonsal ağırlıklandırma yapılan birçok çalışmada MTK temelli uygulamalara da başvurulmaktadır. Bununla birlikte bu konuya ilişkin detaylı bilgilere kitabın 2. bölümünde yer verilmiştir.

2.3. Otomatik Puanlama

Otomatik puanlama, bir performansın insan puanlayıcıların değerlendirmelerine benzer biçimde analiz edilmesini ve puanlanmasını sağlayan, yapay zekâ veya istatistiksel modeller kullanan bilgisayar temelli bir sistemdir. Bu sistemler, özellikle doğal dil işleme ve makine öğrenmesi tekniklerinden yararlanarak yanıtların dilsel, anlamsal ve yapısal özelliklerine göre değerlendirilmesini esas alır. Bu yolla açık uçlu maddelere verilen yanıtların daha hızlı ve ekonomik biçimde puanlanması mümkün olabilmektedir. Otomatik puanlama sonucu elde edilen puanların geçerliğinin sağlanması bu süreçteki en önemli husustur. Zira birçok algoritmayla otomatik değerlendirme yapılarak bir puan elde edilebilmektedir. Bu puanın öğrencinin gerçek performansını ortaya koyması otomatik puanlama algoritmalarının işlerliği için en önemli ölçüttür. Kullanılan algoritmanın puanlama sırasında metindeki hangi özellikleri referans alacağı ya da farklı özellikleri aynı anda kullanırken nasıl bir ağırlıklandırma yapacağı algoritmanın performansını etkilemektedir.

Temeli 1960'lara dayanmakla beraber otomatik puanlamadaki önemli gelişmeler ve araştırmalar yaklaşık 20 yıllık bir geçmişe sahiptir. Otomatik puanlamanın erken örneklerinde genellikle dilbilgisi, sözcük dağarcığının karmaşıklığı ve cümle yapısı gibi iyi tanımlanmış sınırlı sayıdaki dilsel özellik puanlamada kullanılmıştır. Bu özellikler, ölçülmek istenen yapıyı yakından temsil edecek ve kapsayacak biçimde alan uzmanları tarafından titizlikle seçilir. Bu nedenle elde edilen puanlar yüksek düzeyde açıklanabilir ve yorumlanabilir niteliktedir. Bu erken dönem yaklaşımlarda modelde kullanılan özellikler ölçülmek istenen yapıyla doğrudan ilişkili olup her bir özelliğe atanan ağırlıklar açık biçimde görülebilmekte ve modelin bu özelliklerden hareketle nasıl bir puanlama süreci izlediği adım adım takip edilebilmektedir. Dolayısıyla bu tür modellerin ürettiği puanlar yüksek derecede şeffaf, açıklanabilir ve yorumlanabilir nitelik taşımaktadır (Zhang ve ark., 2024). Fakat bir metindeki anlam örüntüsünden ziyade

metinsel özelliklere yöneldiğinden (uzunluk, belirli kelimelerin kullanımı vb.) bu modeller kimi zaman güçlü anlamsal içeriklere hak ettiğinin altında bir değer biçebilmektedir (Aydın ve ark., 2025). Ayrıca geleneksel modeller, yanıtların bir bölümünün önceden insan puanlayıcılar tarafından puanlanmasını zorunlu kılar ve bu verilere dayanarak insan puanlayıcıların kararlarını taklit edebilecek örüntüleri öğrenir.

Otomatik puanlama üzerine yapılan çalışmaların Page (1966) tarafından geliştirilen ve bir kompozisyonu dilbilgisi, sözcük seçimi ve cümle yapısı gibi özelliklere göre değerlendiren Project Essay Grader (PEG) ile başladığı söylenebilir. Foltz ve arkadaşları (1999) ise Intelligent Essay Assessor (IEA) adlı sistemi kullanarak örtük anlamsal çözümleme (latent semantic analysis) yöntemiyle içerik metnin değerlendirmesini yapmıştır. Shermis ve arkadaşları (2001) tarafından yayımlanan PEG'in geliştirilmiş bir versiyonu, dilbilgisine odaklanmış ve sistemin puanlama performansını insan puanlayıcılar ile karşılaştırmıştır. Powers ve arkadaşları (2002), E-rater sistemini; Rudner ve arkadaşları (2006), Intellimetric'i; Rudner ve Liang (2002) ise Bayesian Essay Test Scoring System'i (BETSY) geliştirmiştir. Bu sistemler, bir yazma performansını (bu örneklerde kompozisyonu) puanlamak için biçem (style) ve içerik özelliklerine odaklanan doğal dil işleme (NLP) tekniklerini kullanmaktadır (akt: Ramesh & Sanapudi, 2022).

1990'lı yıllarda geliştirilen otomatik puanlama sistemlerinin büyük çoğunluğu, örüntü eşleme (pattern matching) ve istatistik temelli yaklaşımları izlemiştir. Sonraki yıllarda kullanılan puanlama sistemleri regresyon temelli ve doğal dil işleme tekniklerini işe koşmaya başlamıştır. 2014 yılından itibaren geliştirilen otomatik puanlama sistemleri, sözdizimsel (syntactic) ve anlamsal (semantic) özellikleri içeren derin öğrenme tekniklerini kullanarak önceki modellere kıyasla daha başarılı sonuçlar elde etmiştir (Dong ve ark., 2017).

Geniş dil modellerindeki-GDM (Large Language Models-LLM) gelişmelerle birlikte daha önce kural ya da özellik temelli olarak iki kısımda ele alınan geleneksel yöntemlerin yerini hızla üretken yapay zekâ (ÜYZ) temelli otomatik puanlama almaya başlamıştır (Aydın ve ark., 2025; Bui & Barrot, 2025; Garcia-Varela ve ark., 2025; Lee ve ark., 2024). ÜYZ temelli otomatik puanlama uygulamalarıyla birlikte, geleneksel uygulamalardaki metodolojik ve dil bilimsel yetkinliklere sahip olma zorunluluğu önemli ölçüde azalmış, böylece otomatik puanlama çalışmalarında ve uygulamalarında kayda değer bir artış gözlenmiştir. Otomatik puanlama bölümünün bundan sonraki kısmında, buraya kadar ana hatlarıyla açıklanan yöntemler farklı başlıklarda derinlemesine ele alınmıştır.

2.3.1. Geleneksel Otomatik Puanlama Yöntemleri

Bu başlık altında ÜYZ temelli uygulamaların öncesinde kullanılan otomatik puanlama yaklaşımlarına yer verilmiştir. Bu yöntemleri kural temelli ve özellik temelli yöntemler şeklinde iki kısımda ele almak mümkündür.

Kural Temelli Yöntemler, metindeki anlam örüntüsünden ziyade bu anlam örüntüsüyle ilişkili olduğu düşünülen sayılabilir metinsel özelliklere odaklanır. Bu yaklaşımın altında, "yazının niteliği ölçülebilir özelliklerle temsil edilebilir" varsayımı yatar. Değerlendiricilerin yazılı performanslarını incelerken dikkate aldıkları "trins" (insan puanlayıcıların sezgisel olarak aradığı temel yazma özellikleri) olarak adlandırılan içsel özelliklerin belirlenmesini içerir. Ancak trins doğrudan gözlemlenemediğinden bunları temsil eden ölçülebilir göstergeler "prox"ların bilgisayar programları aracılığıyla hesaplanmasına çalışılır (Dikli, 2006). Örneğin, bir yazıdaki toplam kelime sayısı, "akıcılık" (fluency) trini için bir prox (yani dolaylı gösterge) kabul edilir. Erken dönem otomatik puanlama sistemlerinde yüzeysel özelliklere yapılan bu vurgu, içerik açısından metnin niteliğini artırmadan yalnızca biçimsel değişikliklerle (örneğin noktalama işaretleri ekleyerek) puan yükseltmeye yönelik hile stratejilerine sistemi açık hale getirmiştir. Bu sınırlılığı ortadan kaldırmak için sonraki yıllarda, yazma performansını puanlarken daha yordayıcı dilsel özellikler üzerinde duran başka doğal dil işleme araçları geliştirilmiştir. Söz gelimi Coh-Metrix, sözcüksel incelik (lexical sophistication), sözdizimsel karmaşıklık ve metin bağdaşıklığını (cohesion) inceleyen kapsamlı göstergelerden biridir (Yeung, 2025). Kısacası kural temelli yöntemlerde uzmanlar tarafından yazma performansının nesnel göstergeleri olarak kabul edilen metinsel özellikler önceden belirlenir ve puanlama işlemi buna göre yapılır; ileri düzey istatistiksel modellerden yararlanılmaz. Kural temelli yaklaşımlarda alan uzmanlarınca belirlenen içerik özelliklerinin puanlamada kullanılması, puanlamanın şeffaflığını artırdığı biçiminde yorumlanır.

Özellik temelli yöntemler ve makine öğrenmesi temelli yöntemleri aynı çatı altınca incelemek makul olacaktır. Çünkü özellik temelli yaklaşımlar makine öğrenmesine dayanan yaklaşımların erken biçimi olarak nitelendirilebilir. Bu yaklaşımda modellerin puanlanmış yanıtlarla eğitilmesi gerekir. Model, insanlar tarafından puanlanmış eğitim verisine dayalı olarak algoritmalar oluşturur ve bu algoritmaları kullanarak kalan yanıtları puanlar. Makine öğrenmesi temelli yöntemler kural temelli yöntemlerdeki özelliklerden bazılarını kullanabilir. Ancak bu durum önceden belirlenmemiştir; modelin eğitim sürecinde kullandığı ve insan puanlayıcıların verdiği puanları yordama düzeyi yüksek metinsel özellikler puanlamada rol oynayabilir. Makine öğrenmesi temelli yöntemler (1) *regresyon temelli*, (2) *sınıflama temelli* ve (3) *sinir ağı temelli modeller* ile (4) *ontolojik temelli yaklaşımlar* olmak üzere dört ana başlıkta sıralanabilir (Ramesh & Sanampudi, 2022):

Regresyon Temelli Modeller, en temel tanımıyla bağımlı/yordanan değişkenin en az aralık düzeyinde olduğu, yani yazılı performansa verilen puanların aralık düzeyinde kabul edildiği durumlarda kullanılır. Temelde bu istatistiksel metodoloji esas olmakla birlikte modelin performansını artırmak amacıyla farklı uygulamalardan yararlanılır. Metinler arası anlamsal benzerliği ortaya koymak amacıyla metin ve derlem (corpus) temelli

uygulamalar (Mohler & Mihalcea, 2009), metindeki farklı özelliklerin (features) her yazılı performans için vektörize edilmesine dayanan uygulamalar (Adamson ve ark., 2014), Ridge regresyon temelli uygulamalar (Sultan ve ark., 2014), bulanık ontoloji temelli (Darwish & Mohamed, 2020) ve metin madenciliği temelli uygulamalar (Süzen ve ark., 2020) regresyon temelli yaklaşımlara örnek olarak gösterilebilir.

Sınıflama Temelli Modeller, bağımlı/yordanan değişkenin aralık düzeyinin altında olduğu durumlarda kullanılır. Destek vektör makinelerinden yararlanılan yanıt temelli ve word2vec kullanılarak kosinüs benzerliğine odaklanan referans temelli uygulamalar (Sakaguchi ve ark., 2015), rassal orman temelli uygulamalar (Mathias & Bhattacharyya, 2018) ve *XGboost* temelli uygulamalar (Ke ve ark., 2019), bu yaklaşımlara örnek olarak gösterilebilir.

Sinir Ağı Temelli Modeller kısımda ise özellikle Evrişimsel Sinir Ağları – (Convolutional Neural Networks - CNN), Kısa-Uzun Süreli Bellek – (Long-short Term Memory- LSTM) ve hatta transformer temelli BERT gibi uygulamalar ön plana çıkar. Basit ileri beslemeli ya da yüzeysel sinir ağları geleneksel olarak makine öğrenmesi altında sınıflandırılmıştır. Ancak CNN ve LSTM ağları gibi derin öğrenme mimarilerinin ortaya çıkmasıyla bu modeller geleneksel makine öğrenmesi sınırlarının ötesine geçmiş ve giderek artan biçimde yapay zekâ temelli otomatik puanlama sistemleri bağlamında ele alınmaya başlanmıştır.

Ontoloji Temelli Yaklaşımlar**,** yazılı performansları anlamsal veya kavramsal ağlar biçiminde temsil eden, yüzeysel metin özelliklerinin ötesine geçerek kavramlar arası anlam ilişkilerini değerlendiren modellerdir. Sözcükler ve ifadeler alan-temelli ontolojilerle eşlenerek, metnin kavramsal tutarlılığı ve anlam ilişkileri değerlendirilir. Destek vektör regresyonu, rassal orman ve cümle benzerliği ağları gibi graf[1] temelli tekniklerle anlamsal benzerlik sayısallaştırılmıştır (Mohler ve ark., 2011; Ramachandran ve ark., 2015; Zupanc ve ark., 2017; Ajetunmobi & Daramola, 2017). Geleneksel yöntemlerde kullanılan metinsel özellikler Tablo 1'de sunulmuştur.

[1] Graf temelli yaklaşımlar (graph-based approaches), metni kavramlar (düğümler-nodes) ve bu kavramlar arasındaki anlamsal ilişkiler (kenarlar-edges) biçiminde temsil eden yöntemlerdir. Bu modeller, öğrencinin yazısındaki kavramsal yapı ve bağlantı örüntülerini analiz ederek yanıtın konuya ve beklenen bilgi yapısına (ontolojiye) ne ölçüde uygun olduğunu değerlendirir.

Tablo 1

Özellik Türleri [Ramesh ve Sanampudi'den (2022) uyarlanmıştır]

İstatistiksel Özellikler	Biçem (Stil) Tabanlı Özellikler	İçerik Tabanlı Özellikler
Kelime sayısına göre metnin uzunluğu	Cümle yapısı	Belge içindeki cümleler arası bağdaşıklık
Cümle sayısına göre metnin uzunluğu	Sözcük türleri	Cümlelerin örtüşmesi (prompt ile)
Ortalama cümle uzunluğu	Noktalama	Bilginin uygunluğu
Ortalama kelime uzunluğu	Dilbilgisel yapı	Kelimelerin anlamsal rolleri
N-gram	Mantıksal bağlaçlar Sözcük dağarcığı	Doğruluk Tutarlılık Temel kavramları ifade eden cümle

Tablo 1'de özetlendiği üzere otomatik puanlama sistemlerinde kullanılan özellikler genellikle istatistiksel, biçemsel ve içerik temelli olarak sınıflandırılabilir. Hangi özelliklerin kullanıldığının yanı sıra bu özelliklerin bilgisayar ortamında nasıl temsil edildiği bir başka deyişle nasıl kullanıldığı da önemlidir. *Bag-of-Words* ve TF-IDF gibi sıklık temelli modellerden, anlamsal temsiller (embeddings) sağlayan derin öğrenme yaklaşımlarına kadar farklı metin temsil yöntemleri, modelin dili nasıl "gördüğünü" belirler. Bu yöntemler Tablo 2'de özetlenmiştir.

Tablo 2

Metin Temsil Yöntemleri

Yöntem	Tanım	N-gram ile İlişkisi	Otomatik Puanlamada Kullanımı
Bag of Words (BoW)	Metni, kelimelerin sıklıklarına göre temsil eden, kelime sırasını dikkate almayan yöntemdir (Jurafsky & Martin, 2023).	N=1 olan özel bir durum olarak görülebilir. N-gram modelinin temelini oluşturur.	Yazılardaki kelime çeşitliliğini ve sıklık örüntülerini analiz etmek için erken otomatik puanlama sistemlerinde kullanılmıştır.
TF-IDF (Term Frequency – Inverse Document Frequency)	Kelimelerin bir belgede ne kadar önemli olduğunu, diğer belgelerdeki sıklığına göre ağırlıklandıran yöntemdir (Manning ve ark., 2008).	BoW temsiline dayanır, ancak kelimelere anlamlı ağırlıklar atar.	Önemli kelimeleri vurgulamak ve yüzeysel metin özelliklerinin ötesine geçmek için kullanılır.
N-gram*	Metindeki ardışık n kelimeden oluşan dizileri temel alan temsil yöntemidir (Shermis & Burstein, 2013).	BoW'un genişletilmiş hâlidir; kelime sırasını ve yerel bağlamı kısmen yakalar.	Dilsel akış, sözdizimi ve bağdaşıklık özelliklerini yakalamak için kullanılır.
Word Embeddings (Word2Vec, GloVe, BERT)	Kelimeleri çok boyutlu anlamsal uzaylarda vektör olarak temsil eden yöntemlerdir (Mikolov ve ark., 2013).	N-gram'ın ötesine geçerek anlamsal (semantik) ilişkileri doğrudan öğrenir.	Derin öğrenme temelli modellerde metnin anlamını yakalamak için kullanılır.

* N-gram'lar hem dilsel bir özellik hem de bir metin temsil yöntemi olarak ele alınabilir. Özellik türleri bağlamında ölçülebilir bir dilsel örüntü, temsil yöntemleri bağlamında ise bu örüntülerin sayısal olarak kodlanma biçimi olarak değerlendirilmiştir.

Bu yöntemler, otomatik puanlamanın ilk dönemlerinde özellik mühendisliğine dayalı sistemlerin temelini oluşturmuştur. Ancak derin öğrenme mimarilerinin gelişmesiyle birlikte, bu özellikleri elle tanımlama gerekliliği ortadan kalkmış, modeller anlamı doğrudan öğrenmeye başlamıştır. Sıradaki başlıkta, bu dönüşümün dayandığı yapay zekâ temelli yaklaşımlardan bahsedilmiştir.

2.3.2. Yapay Zekâ Destekli Otomatik Puanlama

Geleneksel otomatik puanlama yöntemleri açıkça belirtilmiş kurallara ya da özellik mühendisliğine odaklanan yordayıcı yaklaşımlara dayanırken yapay zekâ destekli yaklaşımlar metnin içindeki anlam örüntüsünü ortaya koymaya yönelik ileri düzey matematiksel modeller üzerine kuruludur. Makine öğrenmesi temelli modellerde de bahsi geçen derin öğrenme yaklaşımları, günümüzde kullanılan yapay zekâ destekli otomatik puanlama yaklaşımlarının öncül adımları olarak nitelendirilebilir.

2.3.2.1. Derin Öğrenme Temelli Otomatik Puanlama

Bu yöntemlerden ilk akla gelen CNN ve LSTM'dir. CNN, yerel veya ardışık girdilerdeki "öğrenilebilir filtreler" (convolutions) aracılığıyla örüntüleri yakalayan derin öğrenme mimarisidir. LSTM ise ardışık verilerde uzun süreli bağımlılıkları/örüntüleri yakalamaya yönelik bir RNN (tekrarlayan sinir ağı-recurrent neural nets) yapısı olup geleneksel RNN'lerin "uzun bağlamı öğrenememe" sorununu gidermek üzere geliştirilmiştir (Zhou ve ark., 2015). CNN, yazılı metin içindeki yerel dilsel örüntülere (örneğin; kelime dizimleri, n-gram benzeri yapılar) odaklanırken LSTM bağlamı, cümleler ve kelimeler arasındaki ardışıklığı yakalamayı hedefler; yani bu örüntülerin bağlam içindeki ilişkilerini öğrenir (Taghipour & Ng, 2016).

Bu modellerin otomatik puanlamadaki işleyişi en genel biçimiyle şöyle açıklanabilir: Öncelikle yazılı performans, derin öğrenme modelinin işleyebileceği sayısal bir temsile dönüştürülür. Bu dönüşüm genellikle her kelimenin belirli bir vektörle temsil edildiği vektör temsili[2] katmanı aracılığıyla gerçekleştirilir (Pennington ve ark., 2014). Böylece metindeki her kelime, anlamsal ve sözdizimsel ilişkilerini koruyan çok boyutlu bir uzayda yer alır. Model bu vektörleri kullanarak kelimeler arasındaki benzerlikleri, örüntüleri ve bağlamsal ilişkileri öğrenebilir.

Bu temsiller daha sonra CNN katmanlarına aktarılır. CNN katmanları, metindeki kısa kelime dizileri veya *n-gram* düzeyindeki örüntüler üzerinde çalışarak yüzeysel fakat güçlü biçimde ayırt edici dilsel özellikleri yakalar. Bu yapı, metni küçük dilsel parçalara ayırarak her bir parçadaki kalıpları tarar ve böylece yazının biçimsel düzeni ile söylemsel organizasyonuna ilişkin ipuçlarını ortaya çıkarır. Örneğin; "in conclusion" (sonuç olarak) veya "as a result of" (-nin sonucu olarak) gibi söylemsel belirteçler, metnin yapısal

[2] Her kelime, modelin dilin anlamını öğrenebilmesi için sayısal bir biçimde vektör temsili (embedding) olarak kodlanır.

olgunluğu ve retorik bütünlüğü hakkında dolaylı bilgi sağlayan örüntülerdir. CNN tabanlı otomatik puanlama modelleri bu tür yüzeysel örüntüleri yakalayabilmektedir (Dong ve ark., 2017). Bu tarama sürecinin sonucunda, metnin yüzey yapısında yer alan *n-gram* düzeyindeki örüntüler otomatik olarak belirlenir ve daha ileri katmanlarda kullanılmak üzere giderek daha soyut ve yüksek düzeyli dilsel özelliklere dönüştürülür (Zhou ve ark., 2016).

Elde edilen yerel özellikler daha sonra LSTM katmanına aktarılır; burada amaç metin düzeyinde daha uzun bağımlılıkların öğrenilmesidir (Tashu ve ark., 2022; Zhou ve ark., 2016). LSTM, RNN'lerin aksine mesafeli bağlamı koruyabilir; yani sözcükler arasındaki uzaklık artsa da aralarındaki bağlam LSTM sayesinde anlaşılabilir (Greff ve ark., 2017). Bu sayede model, cümleler arasında kurulan mantıksal geçişleri, neden-sonuç ilişkilerini ve anlatıdaki bütünlüğü çözümleyebilir (Ghojogh & Ghodsi, 2023). LSTM katmanları, özellikle bireyin metinde bir fikri nasıl geliştirdiğini veya argümanını nasıl yapılandırdığını tanımada etkilidir. CNN katmanlarından gelen yerel örüntü bilgisi ile LSTM'nin bağlamsal hafızası birleştiğinde model hem mikro (kelime-dizi) hem makro (paragraf-bağlam) düzeyde anlamlı dil yapılarını tanıyabilir (Yin ve ark., 2017).

Son aşamada LSTM katmanının çıktısı özetlenir. Bu özetleme işlemi, modelin farklı zaman adımlarından elde ettiği bilgiyi tek bir vektörde topladığı "havuzlama" (pooling) adımıyla gerçekleştirilir. Böylece model, metni genel bir bütün olarak temsil eden tek bir özet vektör oluşturur. Bu vektör daha sonra "yoğun katman"a (dense layer) aktarılır ve burada öğrenilmiş ağırlıklar aracılığıyla metne nihai bir puan verilir. Bu mimari sayesinde model, insan değerlendiricilerin yaptığına benzer şekilde hem yerel dilsel örüntüleri hem de bütünsel anlatı yapısını dikkate alan tutarlı bir puanlama gerçekleştirebilir (Lopez, 2017; Roy & Kumar, 2021). Dolayısıyla derin öğrenme tabanlı yaklaşımlar, önceki kural ve özellik temelli modellerin sınırlı temsil gücünü aşarak metnin biçimsel ve anlamsal boyutlarını bütüncül biçimde değerlendirebilmektedir.

2.3.2.2. Derin Öğrenmeden Üretken Yapay Zekâ'ya Geçiş

CNN ve LSTM gibi derin öğrenme modelleri, metinleri sayısal temsiller üzerinden işlemeye odaklanırken üretken yapay zekâ yaklaşımları, özellikle *Transformer* mimarisine dayalı modeller, dilin kendine özgü bağlamsal doğasını doğrudan modelleme kapasitesi sunar (Vaswani ve ark., 2017). LSTM ve benzeri sıralı modeller, zaman adımlarındaki bilgi aktarımında sınırlılıklar gösterirken (Liu ve ark., 2021), *Transformer* mimarisi bu sınırlılıkları aşarak paralel işlem gücü ve öbek hâlinde bağlam analizine olanak tanır (Vaswani ve ark., 2017). Dolayısıyla LSTM gibi sıra temelli modellerin öğrenme kapasitesini aşan durumlarda, *Transformer* tabanlı büyük dil modelleri otomatik puanlamada yeni bir paradigma oluşturmuştur.

Transformer Mimarisi ve Temel İlkeleri

Transformer mimarisi, günümüz ÜYZ modellerinin temel yapıtaşlarından biridir. İlk olarak Encoder–Decoder biçiminde tanımlanmış olan bu mimaride, metin önce bir kodlayıcı (encoder) katmandan geçirilir. Ardından çözümlenmiş temsiller bir çözücüye (decoder) aktarılır ve hedef çıktıya dönüştürülür; yani son aşamada, modelin hedeflediği sonuç (örneğin çeviri, özet veya bu bölüm özelinde puan) üretilir (Vaswani ve ark., 2017). Bu yapı içinde en kritik unsur self-attention mekanizmasıdır: Her bir *token* (kelime ya da sembol), diğer tüm token'lerle dinamik biçimde ilişki kurarak hem yerel (kelimenin metinde kendisine fiziksel olarak yakın olan kelimelerle ilişkisi) hem de bütünsel (kelimenin metnin genelindeki anlam ilişkileri) bağlamları eşzamanlı olarak öğrenebilir (Vaswani ve ark., 2017; IBM, 2024). *Transformer* mimarisinde, sıranın anlamını korumak için positional encoding adı verilen konumsal yerleştirme işlemi uygulanır; bu sayede kelimeler arasındaki bağıntılar sıralı bir şekilde değil, bağlama göre modellenir (OpenReview, 2020). Böylece, geleneksel sıralı modellerin (örneğin LSTM) zamana bağlı bağımlılıkları yakalamadaki sınırlılıklarının ötesine geçilir, bağımlılıklar artık sıraya değil bağlama göre ele alınmış olur (Vaswani ve ark., 2017). Bu durum otomatik puanlama açısından metinlerin yalnızca diziler olarak değil anlam ve bağlam ilişkileriyle bir bütün olarak değerlendirilmesine olanak tanımış ve modellerin dilsel bütünlüğü insan benzeri biçimde çözümleyebilmesini mümkün kılmıştır (Devlin ve ark., 2019).

2.3.2.3. Üretken Yapay Zekâ (ÜYZ) Temelli Otomatik Puanlama Modelleri

Bahsi geçen özellikleriyle GDM, önceki yaklaşımlara kıyasla çok daha düşük düzeyde özellik mühendisliği ve ön-tanımlı kural gereksinimiyle otomatik puanlama yapmayı mümkün kılmaktadır. Kural temelli yaklaşımlarda, puanlamanın geçerliğini sağlamak için alan uzmanlarınca puanlama kriterlerinin ve kurallarının önceden ayrıntılı biçimde tanımlanması zorunludur. Makine öğrenmesi temelli yaklaşımlarda sistemin öğrenebilmesi için insan puanlayıcıların verdiği puanlardan oluşan etiketli veri setleri kullanılarak bir tahmin algoritması geliştirilir. ÜYZ temelli yaklaşımlarda ise zaman ve emek açısından maliyetli bu süreçlere gerek kalmadan yüksek doğruluk düzeyinde puanlama yapılabilir. GDM'ler, dilin bağlamsal yapısını doğrudan modelleyebildiği için doğru kurgulanmış bir istem (prompt) aracılığıyla, hangi ölçütlere göre değerlendirme yapacağını kavrayıp uygun puanı üretebilmektedir. Böylelikle modelden artık önceden tanımlanmış kuralları veya yüzeysel istatistiksel eşleşmeleri değil geniş bağlam ilişkilerini esas alan bütüncül bir yargıda bulunması beklenmektedir.

Son yıllarda GDM'lerin gelişmesi, otomatik puanlamada geleneksel makine öğrenmesi yaklaşımlarına alternatif olarak istem-tabanlı puanlamayı, yani sıfır-atım (*zero-shot)* veya birkaç-atım (*few-shot)* değerlendirmeyi mümkün kılmıştır. Zero-few shot kavramları aslında otomatik puanlamaya özgü değildir. *Zero-shot* senaryolar modelin yalnızca verilen yönergeye dayanarak istenen bir işlemi yapmasını, *few-shot* senaryolar ise istenen işlemi yapmak için modele birkaç örnek gösterilmesini esas alır (Brown ve

ark., 2020). Xiao ve arkadaşları (2024) ÜYZ temelli otomatik puanlama yöntemlerini rubriksiz *zero-shot,* rubrikli *zero-shot* ve rubrikli *few-shot* olmak üzere üç kısımda ele almıştır (akt; Aydın ve ark., 2025): **(1)** Rubriksiz *zero-shot* yaklaşımında modele yalnızca istem ve puanlanacak yazılı performans sunulur; **(2)** rubrikli *zero-shot* yaklaşımında isteme ek olarak değerlendirmede esas alınacak rubrik de verilir; **(3)** rubrikli *few-shot* yaklaşımında ise rubrikle birlikte birkaç örnek performans ve bu performansların insan puanlayıcılar tarafından verilmiş puanları modele gösterilerek değerlendirme yapılması sağlanır.

Fine-tuning (ince ayar) ise istem-tabanlı bu yaklaşımların ötesine geçerek, modelin görev spesifik puanlama ölçütlerini doğrudan öğrenmesini sağlayan ek bir uyarlama süreci sunar (OpenAI, 2023). *Zero-shot* veya *few-shot* senaryolar büyük dil modellerinin genel kapasitesine dayanırken *fine-tuning* belirli bir yazma türüne, rubriğe veya hedef performans özelliklerine ilişkin örüntüleri modelin parametrelerine işler. Bu sayede model puan tahmini yapmakla kalmaz, aynı zamanda puanın gerekçesine ilişkin açıklamalar üretebilecek düzeyde göreve uyarlanmış bir değerlendirme yetkinliği kazanır (Xiao ve ark., 2024). Ancak bu işlem oldukça maliyetlidir.

2.3.2.4. Otomatik Puanlama Performansının Değerlendirilmesi

Otomatik puanlamaya ilişkin en önemli hususlardan biri, puanlama işleminin performansını-hatasızlığını (accuracy) ortaya koymak; bir diğer ifadeyle geçerliğine ilişkin kanıt elde etmektir. Otomatik puanlamanın performans kıstası ise insan puanlayıcılarla uyum düzeyidir. Bu uyumu ortaya koymada en çok kullanılan istatistik Quadratic Weighted Kappa-*QWK*'dır. *QWK*, iki puanlayıcının (örneğin, insan–insan veya insan–yapay zekâ) aynı dereceli ölçek üzerinde ne kadar tutarlı puan verdiğini ölçen bir uyum katsayısıdır (Williamson ve ark., 2012)**.** Buradaki 'quadratic' kavramı, ağırlıklandırmanın doğrusal değil farkların karesine dayalı olmasını ifade eder. Haliyle puanlayıcılar arasındaki puan farkı büyüdükçe, uyum katsayısındaki düşüş daha belirgin hâle gelir. *QWK'nin* Haberman (2019) tarafından sunulan varyans–kovaryans temelli eşdeğer gösterimi Eşitlik 1'de verilmiştir:

$$QWK = \frac{2\,\mathrm{Cov}(X,Y)}{\mathrm{Var}(X) + \mathrm{Var}(Y) + (E[X] - E[Y])^2} \qquad \textit{Eşitlik 1}$$

Eşitlik 1'de *Cov(X,Y)*, iki puanlayıcının verdiği puanlar arasındaki kovaryansı; *Var(X)* ve *Var(Y)* her bir puanlayıcının puan dağılımlarının varyansını; *E[X]* ve *E[Y]* ise beklenen (ortalama) puanları ifade eder. Payda, iki puanlayıcı arasındaki toplam puanlama değişkenliğini temsil ederken pay bu değişkenlik içinde ortak hareket eden kısmı gösterir. Böylece *QWK* puanlayıcılar arasındaki uyumu, puan farklarının büyüklüğünü de dikkate alarak ±1 arasında değerler alacak biçimde ortaya koyar. Otomatik puanlama performansının yeterli düzeyde olduğunun kabul edilebilmesi için bu değerin 0.70 ve üzerinde olması beklenir (Williamson ve ark., 2012).

Otomatik puanlama uygulamalarında geleneksel Kappa yerine *QWK'nın* tercih edilmesinin nedeni, Kappa'nın nominal veri varsayımı gereği tüm puan farklarını aynı düzeyde uyumsuzluk olarak değerlendirmesidir. Oysa *QWK*, ordinal ölçek yapısını dikkate alarak puan farklarını büyüklüklerine göre ağırlıklandırır. Örneğin bir puanlayıcının 6, diğerinin 7 vermesi görece küçük bir uyumsuzluk iken birinin 2 diğerinin 7 vermesi belirgin bir uyumsuzluk olarak ele alınır. Bu ayrım *QWK* ile ortaya konulabilmektedir.

QWK, otomatik puanlama modellerinde insan puanlarıyla uyumu değerlendirmek için en yaygın kullanılan ölçütlerden biridir ve çoğu çalışmada Pearson korelasyonuna benzer biçimde yorumlanır (Cohen, 1968; Williamson ve ark., 2012). Ancak *QWK*'nin değeri yalnızca model başarısından değil puan dağılımlarının ortalama ve varyans eşleşmesinden, ölçeğin değer aralığından ve puanlayıcıların belirli kategorilerde yoğunlaşmasından da etkilenir. Yüksek prevalans (puan kategorilerinden bazılarında aşırı yığılma) veya yanlılık durumlarında *QWK* yapay biçimde yükselebilir ya da düşebilir (Lewis & Casabianca, 2023; Sim & Wright, 2005; Xu & Lorber, 2014). Ayrıca puanlama ölçeği daraldıkça küçük farklılıkların ağırlığı arttığı için *QWK* genellikle daha düşük değerler alır (Brenner & Kliebsch, 1996; Parker ve ark., 2013). İnsan puanlayıcılar arasındaki tutarlılık düşük olduğunda da *QWK*, modelin gerçek performansını olduğundan daha zayıf gösterebilir (Loukina ve ark., 2020). Bu nedenle *QWK* tek başına mutlak bir doğruluk göstergesi olarak alınmamalı; bağlam, dağılımlar ve puanlayıcı güvenilirliği göz önünde bulundurularak yorumlanmalıdır (McCaffrey ve ark., 2025).

10.2.4. Karşılaştırma Yargılarıyla Puanlama

Buraya kadar açıklanan puanlama türleri, performansın mutlak bir ölçüte dayanarak puanlanmasını esas alır; her performansın-yanıtın diğerlerinden bağımsız olarak bir değeri vardır ve puanlayıcı (ister insan ister algoritma) bu değeri bir ölçüte dayalı olarak biçebilir. Bununla birlikte mutlak puanlama sürecinde incelenen performans dışında başka performanslar da yapılan puanlamaları etkileyebilir. Bu durumda puanlayıcı, yaptığı puanlamada sadece önceden belirlenmiş ölçütleri değil diğer performansları da bir kıyas olarak kullanmış olur. Laming (2004) bunu "mutlak değerlendirme yoktur; tüm değerlendirmeler birinin diğeriyle karşılaştırılmasına dayanır" sözüyle ifade etmiştir. Analitik puanlamadan yararlanılan kimi durumlarda ise parçaların toplamı, bütünden farklı bir şey ifade edebilmektedir. Yaratıcılığı ölçmeye çalıştığımız bir uygulamada, puanların güvenirliğini artırabilmek için analitik puanlama yaptığımızı düşünelim. Burada analitik puanlama aracının her bir bileşeni özenle hazırlanmış olsa dahi bu parçalardan elde edilen toplam puan her zaman yaratıcılığı yansıtmayabilir (Jones & Davies, 2023).

Thurstone'nin karşılaştırmalı yargı yaklaşımını esas alan karşılaştırmalı yargıyla puanlama (comparative judgement) - (KYP), esas itibariyle model temelli bir yaklaşımdır. Bununla beraber modelin kullandığı verinin üretiliş biçimi ontolojik olarak farklılaştığından KYP'yi bölüm kapsamında ayrı bir başlık şeklinde ele almak daha uygun bulunmuştur.

KYP yönteminde puanlayıcıya iki farklı bireyin performansı sunulur ve puanlayıcıdan yalnızca bütünsel açıdan hangi performansın diğerinden daha üstün olduğuna karar vermesi beklenir. Her performans bu şekilde çok sayıda ikili karşılaştırma sürecinden geçer ve bir matrise kaydedilir. Modelleme ise *Rasch* modeline oldukça yakın bir yapıya sahip olan Bradley-Terry-Luce (BTL) (Bradley & Terry, 1952; Luce, 1959) ile yapılır. KYP'nin çıkış noktası, mutlak bir puana karar vermektense göreli bir yargıda bulunmanın daha kolay olduğu fikridir (Bramley, 2005). Ayrıca iki puanlayıcı belirli bir performansı değerlendirirken aynı puanı vermeyebilir; fakat iki performanstan hangisinin diğerinden "daha iyi" olduğu konusunda uzlaşmaları daha olasıdır (Bramley, 2007; Steedle & Ferrara, 2016). Nitekim ikili karşılaştırmalarla yapılan değerlendirmelerde, puanlayıcıların yalnızca kısa ve yüzeysel bir eğitimle süreci etkili biçimde yürütebildikleri belirtilmektedir (Gürel ve ark., 2025). KYP uygulaması performansların eşleştirilmesi, yetenek kestirimi, güvenirlik ve sonlandırma kuralı şeklinde dört temel bileşen üzerinden açıklanabilir. Bu bileşenlerin her biri aşağıda ayrı bir başlık olarak sunulmuştur.

Yetenek Kestirimi Thurstone (1927) karşılaştırmalı yargı yasasıyla gizil bir ölçek üzerinde nesnelerin göreli konumlarını tahmin etmeye yönelik ilk yaklaşımı ortaya koymuştur. Ardından Bradley ve Terry (1952) ile Luce (1959), karşılaştırmalı yargı verilerinin analizinde lojistik fonksiyonların nasıl kullanılabileceğini göstermiş, Andrich (1978) ise Thurstone'un yaklaşımının Rasch lojistik modeliyle örtüştüğünü ifade etmiştir. Karşılaştırma modeli Eşitlik 2'deki tanımlanır:

$$P_k(a_j) = p(X_j k = 1 | a_j, a_k) = exp(a_j - a_k) / \left(1 + exp(a_j - a_k)\right) \qquad \textit{Eşitlik 2}$$

Burada $P_{\mathrm{k}}(a_{\mathrm{j}})$ ya da $p(X_{\mathrm{jk}} = 1 | a_{\mathrm{j}}, a_{\mathrm{k}})$, j performansının k performansına tercih edilme olasılığını ifade eder. X_{jk}=1 durumu, j performansının k'ye göre seçildiğini gösterir. Yetenek değerleri α_j ve α_k ilgili performanslara ait logit birimindeki başarı düzeylerini temsil eder. Yetenek kestirimi ise maksimum olabilirlik metoduyla yapılır.

Performansların Eşleştirilmesi konusunda öne çıkan yöntem yarı-rassal eşlemedir. Bu yöntemde her bir performans bir diğeriyle daha önceki eşleştirmeler dışarıda bırakılarak rassal olarak eşleştirilir. İsviçre satrancı olarak anılan yöntemde ise amaç eşleşmeyi uyarlamalı olarak yapabilmektir. Toplam karşılaştırma sayısını azaltırken güvenirliği yüksek tutmayı mümkün kılan başka uyarlamalı yöntemler de bulunmaktadır (ayrıntılar için bkz: Gürel ve ark., 2025).

KYP'de güvenirlik, ölçeğin ayırt etme gücünü ifade eden SSR (scale separation reliability) ile hesaplanır. SSR, değerlendiricilerin performans sıralamalarına ilişkin ne düzeyde hemfikir olduğunda ilişkin bilgi sunar. Formülü şu şekildedir: (Gustafsson, 1977):

$$\mathrm{SSR} = (\sigma^{a2} - \mathrm{MSE}) / \sigma^{a2} \qquad \textit{Eşitlik 3}$$

Sonlandırma Kuralı KYP'de çoğunlukla belirli bir karşılaştırma sayısına ya da ortalama karşılaştırmaya dayalı bir sonlandırma kuralı işletilir (Cromptvoets ve ark., 2020). Farklı yetenek düzeylerine sahip performanslara eşit sayıda karşılaştırma yapılması, uç noktalardaki performansların kestirim standart hatasının artmasına yol açmaktadır (Cromptvoets ve ark., 2020; Uysal ve ark., 2024). Bunun önüne geçebilmek içni Verhavert ve arkadaşları (2022), sonlandırma kuralında SSR değerinin kullanılmasını önermiştir. Bu kurala göre belirli bir güvenirlik düzeyine ulaşıldığında süreç durdurulur. Ancak yüksek SSR bile tüm kestirimlerin aynı standart hata düzeyine sahip olmasını garanti etmez. Bu nedenle sonlandırma sürecinin BBT'de olduğu gibi, doğrudan kestirim standart hatasına bağlı olarak yürütülmesi önerilir (Gürel ve ark., 2025).

3. SONUÇ

Puanlama yöntemleri, eğitimde ölçme ve değerlendirme sürecinin en kritik bileşenlerinden birini oluşturmaktadır. Ölçme aracının niteliği, ölçülmek istenen özellik, değerlendirmenin amacı ve uygulama koşulları, hangi puanlama yaklaşımının tercih edileceğini belirleyen temel unsurlar olarak görülebilir. Bu nedenle tek bir puanlama yönteminin tüm değerlendirme durumları için en uygun çözüm olduğunu söylemek mümkün değildir. Teknolojik gelişmeler, özellikle otomatik puanlama alanında önemli dönüşümleri beraberinde getirmiştir. Kural ve özellik temelli sistemlerden derin öğrenme ve büyük dil modellerine uzanan bu gelişim, üretimsel becerilerin daha hızlı, tutarlı ve ölçeklenebilir biçimde değerlendirilmesini mümkün kılmaktadır. Bununla birlikte, yapay zekâ temelli sistemlerin sağladığı yüksek performans, puanlama sürecine ve sonuçlarına ilişkin tartışmaları ortadan kaldırmamaktadır. Bu nedenle otomatik puanlama sistemlerinin, özellikle geçerlik konusunda vermesi gereken önemli sınavlar vardır. Diğer yandan, karşılaştırma yargılarıyla puanlama gibi alternatif yaklaşımlar, özellikle bütüncül değerlendirmenin ön planda olduğu performans görevlerinde önemli avantajlar sunmaktadır. Yapay zekâyla karşılaştırmalı yargılamanın birlikte kullanılmasına yönelik çalışmalar, gelecekte hibrit puanlama sistemlerinin yaygınlaşabileceğine işaret etmektedir. Böyle sistemlerin, psikometrik modellerle yapay zekânın hız ve ölçeklenebilirliğini birleştirerek değerlendirmelerin güvenirliğine ve geçerliğine katkı sunabileceği düşünülmektedir. Sonuç olarak, puanlama yöntemlerindeki gelişim yalnızca teknolojik ilerlemelerle değil, aynı zamanda ölçme kuramındaki gelişmelerle de şekillenmektedir. Gelecekteki çalışmaların, farklı puanlama yaklaşımlarını karşılaştırmalı olarak incelemesi ve bu yöntemlerin geçerlik, güvenirlik, adillik ve uygulanabilirlik boyutlarını birlikte ele alması, eğitimde daha nitelikli değerlendirme sistemlerinin geliştirilmesine önemli katkılar sağlayacaktır.

KAYNAKÇA | REFERENCES

Andrich, D. (1978). Relationships between the Thurstone and Rasch approaches to item scaling. *Applied Psychological Measurement*, *2*(3), 451–462. https://doi.org/10.1177/014662167800200319

Benton, T. (2018, Ekim). *Exploring the relationship between optimal methods of item scoring and selection and predictive validity*. Paper presented at the 19th Annual Conference of the Association for Educational Assessment in Europe, Arnhem-Nijmegen, The Netherlands.

Benton, T. (2021). Comparative judgement for linking two existing scales. *Frontiers in Education*, *6*, 775203. https://doi.org/10.3389/feduc.2021.775203

Bradley, R. A. & Terry, M. E. (1952). Rank analysis of incomplete block designs: I. The method of paired comparisons. *Biometrika*, *39*(3–4), 324–345. https://doi.org/10.1093/biomet/39.3-4.324

Bramley, T. (2005). A rank-ordering method for equating tests by expert judgment. *Journal of Applied Measurement*, *6*(2), 202–223.

Bramley, T. (2007). Paired comparison methods. P. Newton, J. Baird, H. Goldstein, H. Patrick & P. Tymms (Ed.), *Techniques for monitoring the comparability of examination standards* (s. 246–294) içinde. Qualifications and Curriculum Authority.

Brenner, H. & Kliebsch, U. (1996). Dependence of weighted kappa coefficients on the number of categories. *Epidemiology*, *7*(2), 199–202. https://doi.org/10.1097/00001648-199603000-00016

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Amodei, D. & Ryder, N. (2020). *Language models are few-shot learners*. In Advances in Neural Information Processing Systems.

Cohen, J. (1968). Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Educational and Psychological Measurement*, *33*(1), 37–46. https://doi.org/10.1177/001316446802800104

Crompvoets, E. A. V., Béguin, A. A. & Sijtsma, K. (2020). Adaptive pairwise comparison for educational measurement. *Journal of Educational and Behavioral Statistics*, *45*(3), 316–338. https://doi.org/10.3102/1076998619890589

Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (s. 4171–4186) içinde. Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1423

Dikli, S. (2006). An overview of automated scoring of essays. *Journal of Technology, Learning, and Assessment, 4*(3). https://files.eric.ed.gov/fulltext/EJ843855.pdf

Dong, F., Zhang, Y. & Yang, J. (2017). Attention-based recurrent convolutional neural network for automatic essay scoring. Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017) (s. 153–162) içinde. https://frcchang.github.io/pub/conll17.fei.pdf

Ghojogh, B. & Ghodsi, A. (2023). Recurrent neural networks and long short-term memory networks: Tutorial and survey. arXiv. https://arxiv.org/abs/2304.11461

Greff, K., Srivastava, R. K., Koutník, J., Steunebrink, B. R. & Schmidhuber, J. (2017). LSTM: A search space odyssey. *IEEE Transactions on Neural Networks and Learning Systems, 28*(10), 2222–2232. https://doi.org/10.1109/TNNLS.2016.2582924

Gustafsson, J.-E. (1977). The Rasch model for dichotomous items: Theory, applications and a computer program. Göteborg University.

Gürel, S., Şahin, M. D., Uysal, İ., İbileme, A. İ. & Gündüz, T. (2025). Karşılaştırma yargılarıyla puanlamada uyarlamalı eşleme ve standart hata sonlandırma kuralı: Yazma becerisinin ölçülmesine yönelik bir uygulama. *Eğitim ve Bilim, Ek Sayı 1,* 93–110.

Haberman, S. J. (2019). *Measures of agreement versus measures of prediction accuracy* (Research Report No. RR-19-20). ETS. https://doi.org/10.1002/ets2.12258

IBM. (2024, Ocak 10). What is self-attention? IBM Think. https://www.ibm.com/think/topics/self-attention

Jones, I. & Davies, B. (2023). Comparative judgement in education research. *International Journal of Research & Method in Education, 47*(2), 170–181. https://doi.org/10.1080/1743727X.2023.2242273

Jurafsky, D. & Martin, J. H. (2023). Speech and language processing (3. baskı). Draft online edition.

Kula Kartal, S. (2021). Performans görevlerine ve portfolyoya dayalı durum belirleme. M. D. Şahin (Ed.), *Eğitimde ölçme ve değerlendirme* (s. 175-206) içinde. Nobel Akademi.

Kurz, T. B. (1999). *A review of scoring algorithms for multiple-choice tests* (Report No. ED 428 076). EDRS Publications.

Laming, D. (2003). *Human judgment: The eye of the beholder*. Thomson Learning.

Lewis, C. & Casabianca, J. M. (2023). On the misuse of quadratic weighted kappa in automated scoring research. *Educational Measurement: Issues and Practice*, *42*(1), 60–72. https://doi.org/10.1111/emip.12552

Liu, H., Shen, Y. & Chen, <J. (2021). Understanding the limitations of LSTMs for long sequences. *IEEE Access, 9*, 63785–63797. https://doi.org/10.1109/ACCESS.2021.3074639

Lopez, M. M. (2017). *Deep learning applied to NLP*. arXiv preprint. https://doi.org/10.48550/arXiv.1703.03091

Loukina, A., Madnani, N., Zechner, K. & Cahill, A. (2020). *Using simulated data to examine metrics for automated scoring evaluation*. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (s. 6072–6082) içinde. https://doi.org/10.18653/v1/2020.acl-main.540

Luce, R. D. (1959). *Individual choice behaviour: A theoretical analysis*. John Wiley & Sons.

Manning, C. D., Raghavan, P. & Schütze, H. (2008). *Introduction to information retrieval*. Cambridge University.

McCaffrey, D. F., Casabianca, J. M. & Johnson, M. S. (2025). Measuring the accuracy of true score predictions for AI scoring evaluation. *Journal of Educational Measurement*. Advance online publication. https://doi.org/10.1111/jedm.7001

Mikolov, T., Chen, K., Corrado, G. & Dean, J. (2013). Efficient estimation of word representations in vector space. In NIPS Workshop.

OpenAI. (2023). Fine-tuning GPT models. https://platform.openai.com/docs/guides/fine-tuning

OpenReview. (2020). On position embeddings in BERT. https://openreview.net/pdf/be0283e323f1b118c975dbc46f7f75c59b467fe0.pdf

Özdemir, D. (2003). Çoktan seçmeli testleri puanlama yöntemlerine bir bakış [An overview of methods for scoring multiple choice tests]. *Eğitim Araştırmaları Dergisi, 4*(12), 121–122.

Page, E. B. (1966). The imminence of... grading essays by computer. *Phi Delta Kappan, 47*(5), 238–243. http://www.jstor.org/stable/20371545.

Parker, P., May, L. & Stahl, J. (2013). The impact of score scale length on reliability and agreement indices. *Journal of Applied Measurement, 14*(2), 123–134.

Pennington, J., Socher, R. & Manning, C. D. (2014). *GloVe: Global vectors for word representation*. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP) (s. 1532–1543) içinde. Association for Computational Linguistics.

Rowley, G. L. & Traub, R. E. (1977). Formula scoring, number-right scoring, and test-taking strategy. *Journal of Educational Measurement*, *14*(1), 15–22.

Roy, P. K. & Kumar, A. (2021). Convolutional neural network for text: A stepwise working guidance. Proceedings of Yukthi 2021 – The International Conference on Emerging Trends in Engineering içinde. https://www.researchgate.net/publication/356795494

Shermis, M. D. & Burstein, J. (2013). Handbook of automated essay evaluation. Routledge.

Sim, J. & Wright, C. C. (2005). The kappa statistic in reliability studies: Use, interpretation, and sample size requirements. *Physical Therapy, 85*(3), 257–268. https://doi.org/10.1093/ptj/85.3.257

Stanley, J. C. & Wang, M. D. (1970). Weighting test items and test-item options: An overview of the analytical and empirical literature. *Educational and Psychological Measurement, 30*(1), 21–35.

Taghipour, K. & Ng, H. T. (2016). A neural approach to automated essay scoring. Proceedings of the Conference on Empirical Methods in Natural Language Processing (s. 1882–1891) içinde.

Tashu, T. M., Maurya, C. K. & Horváth, T. (2022). *Deep learning architecture for automatic essay scoring*. arXiv preprint. https://arxiv.org/abs/2206.08232

Thurstone, L. L. (1927). A law of comparative judgment. *Psychological Review, 34*(4), 273–286. https://doi.org/10.1037/h0070288

Uysal, İ., Gürel, S., Şahin, M. D., İbileme, A. İ. & Yıldırım, Y. (2024, Eylül). *Açık uçlu maddelerin karşılaştırmalı yargıyla puanlanmasında sabit sayıda uyarlamalı ve rassal eşlemeye dayalı bir simülasyon çalışması*. 9. Eğitimde ve Psikolojide Ölçme ve Değerlendirme Kongresi, Anadolu Üniversitesi, Eskişehir.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł. & Polosukhin, I. (2017). Attention is all you need. I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan & R. Garnett (Ed.), Advances in Neural Information Processing Systems (Vol. 30). Curran Associates içinde. https://doi.org/10.5555/3295222.3295349

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł. & Polosukhin, I. (2017). *Attention is all you need.* arXiv preprint arXiv:1706.03762. https://doi.org/10.48550/arXiv.1706.03762

Verhavert, S., Furlong, A. & Bouwer, R. (2022). The accuracy and efficiency of a reference-based adaptive selection algorithm for comparative judgment. *Frontiers in Education, 6*, 785919. https://doi.org/10.3389/feduc.2021.785919

Williamson, D. M., Xi, X. & Breyer, F. J. (2012). A framework for evaluation and use of automated scoring. *Educational Measurement: Issues and Practice, 31*(1), 2–13. https://doi.org/10.1111/j.1745-3992.2011.00223.x

Xu, R. & Lorber, M. F. (2014). Interrater agreement statistics with skewed data: Evaluation of weighted kappa. *Psychological Methods*, *19*(4), 524–539. https://doi.org/10.1037/met0000014

Yaşar, M., Kula Kartal, S. & Aybek, E. C. (2021). Scoring methods for multiple choice tests: How does the item difficulty weighted scoring change students' test results? *Bartın University Journal of Faculty of Education*, *10*(2), 309–324. https://doi.org/10.1016/buefad.878504

Yin, W., Kann, K., Yu, M. & Schütze, H. (2017). *Comparative study of CNN and RNN for natural language processing*. arXiv preprint arXiv:1702.01923. https://arxiv.org/abs/1702.01923

Yurdugül, H. (2010). Farklı madde puanlama yöntemlerinin ve farklı test puanlama yöntemlerinin karşılaştırılması. *Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi*, *1*(1), 1–8.

Zhou, C., Sun, C., Liu, Z. & Lau, F. C. M. (2015). *A C-LSTM neural network for text classification*. arXiv. https://arxiv.org/abs/1511.08630

Zhou, C., Sun, C., Liu, Z. & Lau, F. C. M. (2016). *Text classification improved by integrating bidirectional LSTM with two-dimensional max pooling*. https://aclanthology.org/C16-1329.pdf

YAZAR HAKKINDA | ABOUT AUTHOR

DOÇ. DR. MURAT DOĞAN ŞAHİN

Anadolu Üniversitesi | mdsahin[at]anadolu.edu.tr

ORCID: 0000-0002-2174-8443

Doç. Dr. Murat Doğan ŞAHİN, 1984 yılında Kütahya'da doğmuş, ilk ve ortaöğrenimini burada tamamlamıştır. 2006 yılında Hacettepe Üniversitesi Fen ve Matematik Alanları Eğitimi Bölümünden mezun olmuş; 2011 yılında aynı üniversitede Ölçme ve Değerlendirme alanında yüksek lisansını, 2017 yılında ise doktorasını tamamlamıştır. 2012–2017 yılları arasında Anadolu Üniversitesi Açıköğretim Fakültesi Test Araştırma Biriminde uzman olarak görev yapmıştır. 2018 yılından bu yana Anadolu Üniversitesi Eğitim Fakültesi Eğitimde Ölçme ve Değerlendirme Anabilim Dalında öğretim üyesi olarak çalışmalarını sürdürmektedir. Araştırmaları ağırlıklı olarak açık uçlu yanıtların puanlanmasına yönelik yenilikçi yaklaşımlar, yapay zekâ destekli değerlendirme sistemleri, bireye uyarlanmış testler ve dijital ölçme uygulamaları üzerine odaklanmaktadır.

ASSOC. PROF. DR. MURAT DOĞAN ŞAHİN

Anadolu University | mdsahin[at]anadolu.edu.tr

ORCID: 0000-0002-2174-8443

Associate Professor Dr. Murat Doğan ŞAHİN was born in Kütahya in 1984 and completed his elementary and secondary education there. He graduated from the Department of Science and Mathematics Education at Hacettepe University in 2006; he completed his master's degree in Measurement and Evaluation at the same university in 2011 and his Ph.D. in 2017. Between 2012 and 2017, he served as a specialist at the Test Research Unit of the Faculty of Open Education at Anadolu University. Since 2018, he has been continuing her work as a faculty member in the Department of Measurement and Evaluation in Education at the Faculty of Education, Anadolu University. His research focuses primarily on innovative approaches to scoring open-ended responses, artificial intelligence-supported assessment systems, individualized tests, and digital assessment applications.

11. BÖLÜM

ÖLÇME SONUÇLARINA UYGULANAN İSTATİSTİKSEL İŞLEMLER

STATISTICAL ANALYSIS OF MEASUREMENT RESULTS

Doç. Dr. Eren Can AYBEK

ÖLÇME SONUÇLARINA UYGULANAN İSTATİSTİKSEL İŞLEMLER

Doç. Dr. Eren Can AYBEK
Pamukkale Üniversitesi

Öz

Bu bölümün amacı ölçme sonuçlarını anlamlandırabilmek için kullanılan test ve madde istatistiklerini açıklamaktır. Bu bağlamda, bir dağılımın hangi noktaya eğilim gösterdiğinin ölçüsü olan merkezi eğilim ölçülerinin yanında dağılımdaki verilerin o merkez noktadan ne kadar yayıldığının göstergesi olarak yayılım ölçüleri aktarılmıştır. Ayrıca farklı ölçme sonuçlarının karşılaştırılmasına olanak sağlayan standart z ve T puanları açıklanmıştır. Test puanları ile hesaplanan bu istatistiklerin yanında testte yer alan maddelerin niteliklerini belirlemek için kullanılan madde istatistik ve parametreleri de yine bu bölümün konuları arasındadır. Bir maddenin, uygulandığı gruba ne kadar zor geldiğinin bir ölçüsü olarak madde güçlüğü; yine bir maddenin ölçmeye çalıştığı özelliğe az ya da çok sahip olan bireyleri ayırmada ne derecede başarılı olduğunun bir ölçüsü olarak madde ayırt edicilik indeksine yer verilmiştir. Bunların yanında yanıtların seçeneklere dağılımı da bölüm kapsamında ele alınan bir diğer konu olmuştur. Madde analizleri öncelikle Klasik Test Kuramı bağlamında ele alınmış olsa da Madde Tepki Kuramı temelinde de madde parametreleri hakkında bilgi verilmiştir. Bölüm, test istatistikleri ve madde analizine dair kuramsal bilgiler verdikten sonra Shiny Item Analysis yazılımı ve örnek bir veri seti kullanarak kuramsal olarak ele alınan bilgilerin uygulamalarını içermektedir. Özellikle Madde Tepki Kuramı için varsayımların test edilmesi, madde – veri uyumunun incelenmesi, madde parametrelerinin kestirimi ve yorumlanması hem istatistiksel değerler hem de grafikler üzerinden yorumlanmış ve açıklanmıştır.

Anahtar Kelimeler

Test İstatistikleri, Madde Analizi, Klasik Test Kuramı, Madde Tepki Kuramı, Shiny Item Analysis

STATISTICAL ANALYSIS OF MEASUREMENT RESULTS

Assoc. Prof. Dr. Eren Can AYBEK
Pamukkale University

Abstract

The aim of this chapter is to explain the test and item statistics used to interpret measurement results. In this context, measures of central tendency, which indicate the point toward which a distribution tends to cluster, are presented alongside measures of variability, which reflect the extent to which the observations in a distribution are dispersed around that central point. Standardized z and T scores, which enable the comparison of different measurement results, are also explained. In addition to these statistics calculated from test scores, the chapter addresses the item statistics and parameters used to determine the characteristics of the items included in a test. Item difficulty is presented as an indicator of how difficult an item is for the group to which it is administered, whereas the item discrimination index is discussed as an indicator of the extent to which an item successfully distinguishes between individuals who possess different levels of the attribute being measured. The distribution of responses across the available options is another topic covered in the chapter. Although item analyses are primarily discussed within the framework of Classical Test Theory, information about item parameters based on Item Response Theory is also provided. After presenting the theoretical foundations of test statistics and item analysis, the chapter demonstrates their practical applications using the Shiny Item Analysis software and a sample dataset. Particular attention is given to testing the assumptions of Item Response Theory, examining item–data fit, and estimating and interpreting item parameters. These processes are explained and interpreted using both statistical results and graphical representations.

Keywords

Test Statistics, Item Analysis, Classical Test Theory, Item Response Theory, Shiny Item Analysis

1. GİRİŞ

Kitabın ilk bölümünde tanımlandığı gibi ölçme, bireylerin eğitime konu olan özelliklerinin gözlenip sonuçların sembollerle, özellikle de sayı sembolleriyle ifade edilmesidir (Turgut, 1997). Bu gözlem sonuçları üzerinde yapılacak istatistiksel işlemler, hem ölçülen bireylerin ölçmeye konu özelliğe ne düzeyde sahip olduğuna hem de kullanılan ölçme aracının niteliğine dair bilgiler sağlar. Örneğin, "Testi alan öğrencilerin başarısı nasıldır? Başarı düzeyleri birbirine benziyor mu? Grubun genelinden aşırı sapma gösteren öğrenciler var mı? Testin 3. maddesi gruba zor mu geldi? Testteki 6. madde ölçülen özelliğe sahip olan ve olmayan öğrencileri iyi bir şekilde ayırabilmiş mi? sorularına yanıt bulabilmek için istatistiksel işlemlere başvurulur.

Kitabın bu bölümünde öncelikle, bir testten alınan puanlar ile testin puanlanabilen en küçük birimi olan maddeden elde edilen puanlar üzerinde yapılan istatistiksel işlemler üzerinde durulmuştur. Daha sonra uygulama başlığı altında söz konusu istatistiksel işlemlerin nasıl gerçekleştirileceği açıklanmıştır. Akabinde uygulanan analizlere ait sonuçların nasıl yorumlanacağına ilişkin örnekler sunulmuştur.

2. TEST PUANLARI ÜZERİNDE YAPILAN İSTATİSTİKSEL İŞLEMLER

Bir test ister uluslararası düzeyde isterse küçük bir sınıfta uygulanmış olsun ölçme sonuçlarını anlamlandırmak için belli istatistiksel değerlere ihtiyaç duyulur. On kişilik bir sınıfın başarı düzeyinin nasıl olduğu ortaya konmak isteniyorsa her öğrencinin başarı puanını ayrı ayrı söylemek yerine tüm sınıfa ait tek bir sayısal değer verilmesi tercih edilir. Örnekteki bir öğrencinin test puanı "veri" ya da "gözlem" iken (Arıcı, 2006), tüm sınıfın puanları bir araya gelerek bir "dağılım" oluşturur (Evans & Ildstad, 2001). Test puanlarından anlam çıkarmak için bireylere ya da maddelere ait verilere ayrı ayrı odaklanmak yerine dağılıma ait özelliklerle ilgilenilir.

Bir testten alınan toplam puanları sıralama düzeyinde mi yoksa eşit aralıklı düzeyde mi kabul etmek gerektiğine yönelik tartışmalar 1940'lı yıllara kadar uzanmaktadır. Crocker ve Algina (2008) bu tartışmaların tarihsel gelişimini ele almış, bu kapsamda farklı araştırmacıların görüşlerinden bahsetmiştir. Stevens (1946) psikolojik testler ile sıralama düzeyinde ölçme yapıldığını, bu nedenle bu testlerden alınan toplam puanlar üzerinde aritmetik ortalama ve standart sapma hesaplamanın yanlış olacağını vurgulamıştır. Burke (1963) sıralama – eşit aralıklı ayrımını ölçme yönelimli (measurement-directed) ve ölçmeden bağımsız (measurement-independent) olarak sınıflamıştır. Ölçme yönelimli grup psikolojik testlerden elde edilen puanlarla yapılacak çıkarımsal istatistiklerde parametrik olmayan tekniklerin kullanılması gerektiğini savunurken, ölçmeden bağımsız grup normal dağılım gibi varsayımların karşılanması halinde toplam puanların eşit aralıklı kabul edilebileceğini savunmuştur. Lord ve Novick (1968) ise bu ikileme daha faydacı yaklaşmış, yerleştirme veya performansın yordanmasında verileri eşit aralıklı kabul etmek daha çok yarar sağlıyorsa toplam puanların eşit aralıklı kabul edilebileceğini belirtmiştir. Crocker ve Algina (2008) bu tartışmalara değindikten sonra Lord ve Novick'in (1968) görüşüne katıldıklarını ifade etmiştir. Günümüzde de testten alınan toplam puanlar normallik varsayımını karşılıyorsa genelde eşit aralıklı düzeyde kabul edilmektedir. Bu tartışmalar ışığında bölümün bundan sonraki kısımlarında testten alınan puanların eşit aralıklı ölçme düzeyinde olduğu kabul edilmiştir.

Bir dağılımı tanımlamak ve anlamlandırmak için iki temel ölçüye ihtiyaç duyulur. Bunlar dağılımdaki verilerin merkezinin ne olduğu ve bu verilerin merkezden ne kadar uzağa dağıldığıdır (Köklü ve ark., 2007). Sırasıyla merkezî eğilim ölçüleri ve yayılım ölçüleri olarak adlandırılan bu ölçüler bölümün devamında detaylı şekilde ele alınmıştır. Bahse konu ölçüleri birbiriyle karşılaştırabilir kılmak adına, Tablo 1'de örnek bir veri matrisine yer verilmiş ve tüm ölçüler bu matristen elde edilmiştir. Tabloda doğru yanıtlar 1; yanlış ve boş yanıtlar ise 0 olarak gösterilmiştir. Bu tür puanlamalar iki kategorili puanlama ya da 1 – 0 puanlama olarak adlandırılmaktadır.

Tablo 1

Beş Maddelik Bir Teste İlişkin Yanıt Matrisi

Öğrenci	Madde 1	Madde 2	Madde 3	Madde 4	Madde 5	Toplam
Ö1	1	1	1	1	1	5
Ö2	1	1	0	1	1	4
Ö3	1	1	1	1	0	4
Ö4	1	1	0	0	1	3
Ö5	1	1	1	0	0	3
Ö6	1	1	0	1	0	3
Ö7	1	1	0	0	1	3
Ö8	1	0	1	0	0	2
Ö9	1	0	0	0	1	2
Ö10	1	0	0	0	0	1

2.1. Merkezî Eğilim Ölçüleri

Dağılımın hangi noktaya eğilim gösterdiğini temsil eden ölçülerdir. Bu başlık altındaki ölçülerin her biri, merkezî eğilimi bulmayı amaçlasa da taşıdıkları anlamlar farklıdır. En sık kullanılan merkezî eğilim ölçüleri ortalama (eğitimdeki ölçmelerde genellikle aritmetik ortalama), ortanca ve tepe değerdir.

Dağılımda yer alan tüm değerlerin aritmetik ortalamayı etkilediği unutulmamalıdır. Başka bir deyişle aritmetik ortalama uç değerlerden etkilenir. Dolayısıyla dağılımda uç değerler varsa aritmetik ortalama yanlış yorumlamalara neden olabilecek bir merkezi eğilim ölçüsü haline gelir.

Aritmetik Ortalama ($\bar{X}$): Bir dağılımdaki tüm gözlemlerin toplanıp gözlem sayısına bölünmesiyle hesaplanır. *X* bir gözlemi, *n* ise gözlem sayısını temsil etmek üzere, aritmetik ortalamanın matematiksel gösterimi Eşitlik 1'de verilmiştir (Kaptan, 1998).

$$\bar{X} = \frac{\sum X}{n} \qquad \textit{Eşitlik 1}$$

Tablo 1'de yer alan örnek dağılımda, öğrencilerin puanlarına ait aritmetik ortalamayı elde etmek için tüm puanlar toplanıp gözlem sayısına bölünmelidir. Buna göre puanların aritmetik ortalaması:

$\bar{X} = \frac{5+4+4+3+3+3+3+2+2+1}{10} = \frac{30}{10} = 3$ olarak bulunur.

Ortanca: Medyan olarak da isimlendirilen bu istatistik bir dağılımdaki değerlerin tam orta noktasını gösterir. Yani ortanca, gözlemlerin sayısal değerlerinden değil sıralamasından etkilenir. Dolayısıyla bir gözlemin 5,50 ya da 100 olmasından ziyade, gözlemin kaçıncı sırada olduğu önemlidir. Bir dağılımın ortancası bulunurken dağılımdaki değerler küçükten büyüğe sıralanır ve ortadaki değer tespit edilir. Eğer dağılımda tek sayıda gözlem varsa ortada tek bir değer olacağından ortanca kolayca

belirlenir. Ancak dağılımda çift sayıda gözlem varsa ortada tek bir değer bulunmaz, bu durumda ortadaki iki değerin aritmetik ortalaması alınarak ortanca değere ulaşılır (Arıcı, 2006).

Tablo 1'deki toplam puanlar büyükten küçüğe doğru 5, 4, 4, 3, **3, 3**, 3, 2, 2, 1 şeklinde sıralanır. Bu dağılımın tam ortasında koyu ile yazılmış olan iki adet 3 gözlemi bulunur. Ortada iki adet gözlem olduğu için bu iki gözlemin aritmetik ortalaması dağılımın ortancasıdır. Dolayısıyla bu dağılımın ortancası (3+3) / 2 = 3'tür. Gözlem sayısının çok fazla olduğu durumlarda ortancayı bu örnekteki gibi kolayca bulmak mümkün değildir. Bunun yerine *n* gözlem sayısını temsil etmek üzere Eşitlik 2 kullanılarak ortancanın *kaçıncı gözlem* olduğu, diğer bir deyişle konumu (L) bulunabilir.

$$L = \frac{n+1}{2} \qquad \textit{Eşitlik 2}$$

Buna göre, örnekte yer alan 10 gözlem için ortancanın konumu (10 + 1) / 2 = 5,5 bulunur. Yani 5. ile 6. gözlemin ortası (aritmetik ortalaması) bu dağılımın ortancasıdır. Dağılımın 5. ve 6. gözlemleri sırasıyla 3 ve 3'tür. Bu iki değerin aritmetik ortalaması da 3 olduğundan dağılımın ortancası 3'e karşılık gelir.

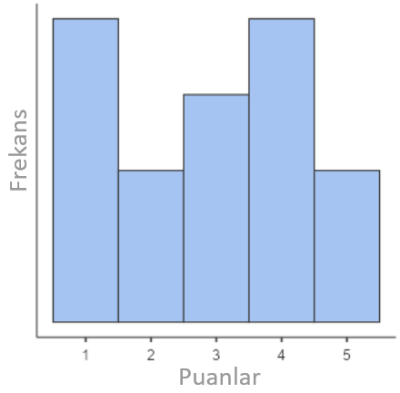


Ülkelerin gelirleri üzerine raporlama yapılırken genellikle aritmetik ortalama yerine ortanca tercih edilir. Bunun nedeni, ülke nüfusundaki az sayıda olan çok zenginlerin, gelir ortalamasını yukarı çekmesinin getireceği yanlış yorumlama riskidir. Gelirin çarpık dağılımı nedeniyle aritmetik ortalama yerine ortancanın raporlanması daha çok tercih edilir.

Tepe Değer: Tepe değer (mod), bir dağılımda en çok tekrar eden gözlemi temsil eder. Tablo 11.1'deki toplam puanlar incelendiğinde 1 adet 5; 2 adet 4; 4 adet 3; 2 adet 2 ve 1 adet 1 puanın olduğu görülür. Dolayısıyla bu seride en çok tekrar eden toplam puan değeri 3'tür ve dağılımın tepe değeri 3 olarak yorumlanır. Tepe değer de tıpkı ortanca gibi uç değerlerden etkilenmez. Bunun yerine değerlerin tekrar etme sıklığından etkilenir. Bir dağılımda tek bir tepe değer olabileceği gibi iki ya da daha fazla tepe değer de bulunabilir (Turgut & Baykul, 2010). Tepe değer, özellikle en sık tekrar eden yanıtın ya da puanın belirlenmesinde; dağılımın hangi noktalarda yoğunlaştığının incelenmesinde kullanışlılık sağlar. Örneğin; öğrenci puanlarının 1, 1, 1, 1, 2, 2, 3, 3, 3, 4, 4, 4, 4, 5, 5 olduğu bir durumda, dağılımda en çok tekrar eden puanlar 4'er defa tekrarlanan 1 ve 4 olur. Bu puanlar için aritmetik ortalama hesaplandığında 43 / 15 = 2,87 değeri elde edilir. Öte yandan bu dağılımın ortancası 3'tür. Oysaki bu dağılımdaki veriler Şekil 1'deki gibi görselleştirildiğinde dağılımdan başka bir anlam çıkmaktadır.

Şekil 1

Örnek Dağılıma Ait Histogram



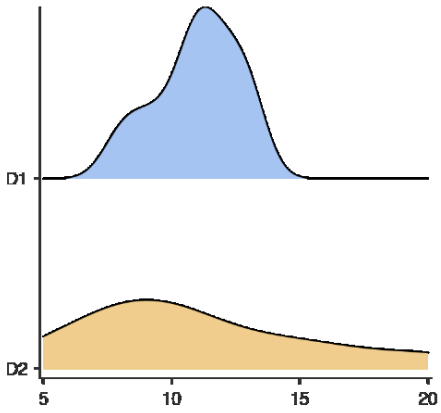
Şekil 1'de görüldüğü üzere dağılımın aritmetik ortalaması 2,87 ve ortancası 3 olmasına rağmen öğrenciler 1 ve 4 puanda yoğunlaşmıştır. Böyle dağılımlar bimodal (çift modlu / tepe değerli) olarak adlandırılmaktadır. Bu anlamda bir dağılımın tepe değeri, aritmetik ortalama ve ortancanın sağlayamadığı bilgileri de sunabilmektedir.

2.2. Yayılım Ölçüleri

Bir dağılımı yalnızca merkez noktası ile tanımlamak yeterli değildir. Dağılımdaki gözlemlerin merkez noktadan ne kadar uzağa dağıldığını da tespit etmek, dağılım hakkında daha fazla bilgi verir. Bu durum, dünya üzerindeki bir konumun yalnızca enlem ile değil aynı zamanda boylamla tanımlanmasına benzetilebilir. Örnek üzerinden açıklamak amacıyla Şekil 2'de aritmetik ortalaması aynı olan iki dağılıma ait yoğunluk grafiği verilmiştir.

Şekil 2

İki Dağılıma Ait Yoğunluk Grafiği



Şekil 2'deki iki dağılımın da aritmetik ortalaması 11'dir. Buna karşın üstte yer alan D1 dağılımındaki veriler 11 noktasına daha yakın dağılırken altta bulunan D2 dağılımındaki veriler 11 noktasından çok daha uzağa dağılmıştır. Yayılım ölçüleri, bir dağılımdaki gözlemlerin birbirinden ne kadar uzağa yayıldığı hakkında bilgi sağlamaktadır. Bu ölçüler aynı zamanda dağılım ölçüleri olarak da anılmaktadır. Dağılım ölçüleri birer homojenlik/heterojenlik ölçüsü olarak da nitelendirilmektedir.

Dağılım Açıklığı (Ranj): Dağılım açıklığı ya da ranj bir dağılımdaki en büyük ve en küçük değer arasındaki farka eşittir (Arıcı, 2006). Dağılıma ait kabaca ve hızlı bir fikir edinmek isteniyorsa kullanışlı bir ölçüdür. Dağılım açıklığı hesaplanırken en küçük ve en büyük değer kullanıldığı için verilerin hangi aralıkta değiştiğine dair test geliştiricisine fikir verir. Örneğin; öğrenci puanlarının 50 – 90 aralığında değiştiğini ve dağılım açıklığının 40 olduğunu bulan bir test geliştiricisi dağılımdaki en başarısız öğrencinin bile 50 puan aldığını, bunun yanında tüm maddelere doğru yanıt veren (dolayısıyla 100 puan alan) öğrenci bulunmadığını, en başarılı ile en başarısız öğrenci arasında ise 40 puanlık bir fark olduğunu söyleyebilir.

Varyans: Dağılımdaki gözlemlerin aritmetik ortalamadan ne kadar uzaklaştığına ilişkin bir ölçü olup gözlemlerin değişkenliği hakkında bilgi verir (Woolley, 1998). *X*, gözlenen değeri; $\bar{X}$ aritmetik ortalamayı ve *n* toplam gözlem sayısını temsil etmek üzere bir dağılımın varyansı Eşitlik 3 kullanılarak hesaplanır

$$s^2 = \frac{\sum(X - \bar{X})^2}{n - 1} \qquad \textit{Eşitlik 3}$$

Tablo 1'deki toplam puan dağılımına ait varyans hesaplanmak istendiğinde ve bu dağılımın aritmetik ortalamasının 3 olduğu hatırlanırsa;

$$s^2 = \frac{(5-3)^2 + (4-3)^2 + (4-3)^2 + (3-3)^2 + (3-3)^2 + (3-3)^2 + (3-3)^2 + (2-3)^2 + (2-3)^2 + (1-3)^2}{10-1}$$

$$s^2 = \frac{(2)^2 + (1)^2 + (1)^2 + (0)^2 + (0)^2 + (0)^2 + (0)^2 + (-1)^2 + (-1)^2 + (-2)^2}{9}$$

$$s^2 = \frac{4 + 1 + 1 + 0 + 0 + 0 + 0 + 1 + 1 + 4}{9} = \frac{12}{9} = 1{,}33 \text{ olarak bulunur.}$$

Ancak bu hesaplamada toplam puanların birimi "puan" olarak kabul edilirse hesaplanan varyansın birimi "puan^2" olur. Tam da bu sebeple, varyans istatistiksel testlerde sıklıkla kullanılmasına karşın bir dağılımı tanımlayıp yorumlamak için pek işlevsel ölçü değildir.

Standart Sapma: Varyans biriminin, hesaplandığı gözlem biriminin karesi olması nedeniyle dağılıma dair yorumda bulunmak güçleşir. Bu nedenle birimi, toplanan verilerdeki gözlemlerin birimi ile aynı olacak bir ölçüye ihtiyaç duyulur. Standart sapma bu ihtiyaca cevap vermekte ve varyansın karekökü alınarak elde edilmektedir. Tablo 1'deki toplam puanlara ait varyansın 1,33 olduğu hatırlanırsa bu dağılıma ait standart sapma değeri $s = \sqrt{1{,}33} = 1{,}15$ olarak elde edilir.

Gerek varyans, gerekse standart sapma verilerin aritmetik ortalamadan sapma miktarını gösteren ölçüler olduğundan bu istatistiklere ait değerler büyüdükçe gözlemler ortalamadan uzaklaşır; küçüldükçe gözlemler ortalamaya yaklaşır. Bu bakımdan Şekil 2'de yer alan iki dağılımdan D1; D2'ye göre daha küçük bir standart sapmaya, dolayısıyla daha küçük bir varyansa sahiptir. Standart sapmanın küçülmesi, gruptaki öğrencilerin birbirine benzer ya da homojen olduğu şeklinde yorumlanırken standart sapmanın büyümesi gruptaki öğrencilerin birbirinden farklı ya da grubun heterojen olduğu şeklinde yorumlanır.

2.3. Standart Puanlar

Farklı ölçme sonuçlarını karşılaştırırken ham puanlar yerine ölçme sonuçlarının ortak bir birime dönüştürüldüğü standart puanlar kullanılır. Örneğin; matematik testinden 20, Türkçe testinden 15 doğrusu olan bir öğrencinin matematik testinde daha başarılı olduğunu söylemek yanıltıcı olabilir. Öyle ki bu öğrenci matematik testinde toplam 40 maddeden 20 doğru yaparken Türkçe testinde 20 maddeden 15 doğru yapmış olabilir. İşte, ham puanların standart puanlara dönüştürülmesi bu gibi farklı ölçme sonuçlarını karşılaştırılabilir kılar. Bunun için yaygın olarak kullanılan standart puan, z-puanıdır. Bu puan aynı zamanda öğrencilerin kendi içinde karşılaştırılmasına da olanak sağlar ve bir öğrencinin ortalamanın kaç standart sapma uzağında olduğu hakkında bilgi verir.

Standart z-Puanı: Ortalaması 0 ve standart sapması 1 olan bir dağılıma sahiptir ve bireyin ortalamanın kaç standart sapma uzağında olduğunu gösterir. Bir kişiye ait *z*-puanı; *X*, gözlenen puan; $\bar{X}$, aritmetik ortalama ve *S*, standart sapma olmak üzere Eşitlik 4'teki gibi hesaplanır (Gulliksen, 1950):

$$z = \frac{X - \bar{X}}{S} \qquad \textit{Eşitlik 4}$$

Tablo 1'de yer alan örnekteki toplam puanların aritmetik ortalamasının 3 ve standart sapmasının 1,15 olduğu hatırlanırsa her bir öğrencinin *z*-puanı aşağıdaki gibi hesaplanabilir:

Tablo 2

Öğrencilere Ait z-Puanları

Öğrenci	X	$X - \bar{X}$	$z = \frac{X - \bar{X}}{S}$
Ö1	5	2	1,74
Ö2	4	1	0,87
Ö3	4	1	0,87
Ö4	3	0	0,00
Ö5	3	0	0,00
Ö6	3	0	0,00
Ö7	3	0	0,00
Ö8	2	-1	-0,87
Ö9	2	-1	-0,87
Ö10	1	-2	-1,74

Yukarıda bahsedildiği gibi *z*-puanı, ortalaması 0 ve standart sapması 1 olan bir dağılımdır. Örnekte hesaplanan *z*-puanlarının da aritmetik ortalamasının 0 ve standart sapmasının 1 olduğu görülebilir. Standart *z*-puanına bakılarak bireyin ortalamanın kaç standart sapma uzağında olduğu şeklinde bir yorum yapılabilir. Örnekteki 1. öğrenci (Ö1), aritmetik ortalamanın 1,74 standart sapma üzerindeyken 9. öğrenci (Ö9) aritmetik ortalamanın 0,87 standart sapma altında bir performans sergilemiştir. Ö4-Ö7 arasındaki öğrencilerin puanı ise tam olarak aritmetik ortalamada olduğu için *z*-puanları sıfır olarak bulunmuştur.

Yukarıda açıklanan yöntem ham puanları doğrusal bir şekilde *z*-puanına dönüştürmektedir. Bu nedenle, hesaplanan *z*-puanları ham puanların sahip olduğu dağılımı korur. Öte yandan ham puanlar standart normal dağılım ile elde edilen yüzdelikler üzerinden doğrusal olmayan bir şekilde normalleştirilerek de *z*-puanına dönüştürülebilir. Bu durumda dönüşüm yapılan *z*-puanı dağılımı, normale yakın bir dağılım sergiler. Sözü edilen dönüşümü yapmak için öncelikle ham puanlardan yüzdelik dilimler oluşturulur, ardından her bir ham puanın yüzdelik dilimi standart normal dağılım altında kalan alan olarak ele alınır ve bu alana karşılık gelen *z*-puanı belirlenir (Crocker & Algina, 2008).

Standart T-Puanı: Standart *z*-puanı teorik olarak -∞ ile +∞ arasında değişmekle birlikte, genelde -3 ile +3 arasında kalan değerler alır. Bu değerler test geliştiricisi tarafından kolayca yorumlanabilse de testi alanlar açısından yorumlaması güçtür. Bundan dolayı testi alanlar tarafından da kolay bir şekilde yorumlanabilen, ortalaması 50 ve standart sapması 10 olan standart T-puanı ileri sürülmüştür. T-puanı, Eşitlik 5'teki gibi *z*-puanı 10 ile çarpılıp sonuca 50 eklenerek elde edilmektedir (Gulliksen, 1950).

$$T = 50 + 10.z \qquad \textit{Eşitlik 5}$$

Buna göre, *z*-puanının hesaplandığı örnekteki her bir öğrenci için T-puanı hesaplanırsa Tablo 3'teki bulgular elde edilir:

Tablo 3

Öğrencilere Ait z ve T-Puanları

Öğrenci	X	X - $\bar{X}$	$Z = \frac{X - \bar{X}}{S}$	T = 50 + 10.Z
Ö1	5	2	1,74	67,4
Ö2	4	1	0,87	58,7
Ö3	4	1	0,87	58,7
Ö4	3	0	0,00	50,0
Ö5	3	0	0,00	50,0
Ö6	3	0	0,00	50,0
Ö7	3	0	0,00	50,0
Ö8	2	-1	-0,87	41,3
Ö9	2	-1	-0,87	41,3
Ö10	1	-2	-1,74	32,6

Görüldüğü üzere, hesaplanan T-puanları testi alanlar için daha alışıldık, dolayısıyla daha kolay yorumlanabilir hale gelmiştir.

3. MADDE PUANLARI ÜZERİNDE YAPILAN İSTATİSTİKSEL İŞLEMLER

Test puanları, testi oluşturan madde puanları ile yakından ilişkilidir. Test puanları üzerinde yapılan istatistiksel işlemler tüm gruba (örneğin, bir sınıfın başarı düzeyi) ait bilgi sunarken madde puanları üzerinde yürütülen istatistiksel işlemler, maddelerin uygulamanın yapıldığı grupta nasıl çalıştığı hakkında bilgi verir. Madde güçlüğü ile madde ayırt ediciliği sıklıkla kullanılan iki madde istatistiği olup madde standart sapması, madde varyansı, madde güvenirliği gibi istatistikler de bulunmaktadır. Sıklıkla kullanılan madde istatistikleri, kitabın 2. bölümünde bahsedilen iki ana kurama göre aşağıda ayrı ayrı detaylı olarak ele alınmıştır.

ÖSYM'nin yaptığı birçok testin raporlanmasında standart puanlar kullanılmaktadır. Bu testlerden biri olan Akademik Personel ve Lisansüstü Eğitimi Giriş Sınavı'nın (ALES) Başvuru Kılavuzu incelendiğinde adayların ALES puanının aşağıdaki eşitlikle hesaplandığı görülmektedir (ÖSYM, 2025):

$$ALES\ puanı = 70 + \frac{30[2(AP - X) - S]}{2(B - X) - S}$$

AP: Adayın ağırlıklı puanı
X: Ağırlıklı puanların ortalaması
S: Ağırlıklı puanların standart sapması
B: Ağırlıklı puanların en büyüğü

Aynı kılavuzda adayın ağırlıklı puanının hesaplanması hakkında, "Adayların her testteki sorulara verdikleri doğru ve yanlış cevaplar ayrı ayrı toplanacak, doğru cevap sayısından yanlış cevap sayısının dörtte biri çıkarılarak ham puanlar elde edilecektir. Her test için adayların ham puanlarının ortalaması ve standart sapması ayrı ayrı hesaplanarak adayların ham puanları, ortalaması 50 ve standart sapması 10 olan standart puanlara dönüştürülecektir." ifadesi yer almaktadır. Dolayısıyla, adayların ham puanları öncelikle T-puanına (ortalaması 50, standart sapması 10) dönüştürülmektedir. Bunu takiben dönüştürülen bu T-puanları her bir puan türüne göre ağırlıklandırılarak adayın ağırlıklı puanı elde edilmektedir. Ardından tüm adayların ağırlıklı puan ortalaması, standart sapması ve en yüksek ağırlıklı puan kullanılarak ALES puanı raporlanmaktadır. Okuyucular dikkat edecek olursa bu sürecin hiçbir noktasında maddenin ne kadar zor olduğu hesaplamaya dâhil edilmemektedir. Ancak halk arasında "zor sorunun daha çok puanlandığına" yönelik bir yanılgı vardır. Kitabın 2. bölümünde bu yanılgıya ait bir tartışmaya değinilmiştir.

3.1. Madde Güçlüğü

Bir testteki maddelerin, testi alan gruba ne kadar zor geldiğini belirlemek önemlidir. Maddeyi yanıtlayan grubun maddenin ölçtüğü özelliğe ne derece sahip olduğunu ortaya koymanın yanı sıra, testin geliştirilme amacına uygun olan maddeleri tespit edebilmek için de madde güçlüğünü hesaplamak gerekir. Aşağıda hem KTK temelinde madde güçlük indeksi, hem de MTK temelinde madde güçlük parametresine yer verilmiştir.

Klasik Test Kuramına Göre

Bir maddenin, uygulandığı gruba ne kadar kolay ya da zor geldiğini belirlemek için maddenin güçlüğü incelenir. Madde güçlüğü, maddeyi doğru yanıtlayan birey sayısının, maddeyi yanıtlayan tüm bireylere oranını ifade eder. Bu oran 1-0 şeklinde puanlanan maddeler için aynı zamanda madde puanlarının ortalamasına eşittir (Erkuş, 2012). Maddeyi grubun tamamı doğru yanıtlarsa madde güçlüğü 1değerini alır, diğer taraftan gruptan hiç kimse maddeye doğru yanıtlamazsa maddenin güçlüğü 0 olur. Dolayısıyla bu indeksin adı maddenin *güçlüğü* olsa da indeks değeri arttıkça madde kolaylaşır. İndeks değeri 0,50 olduğunda ise madde tam orta güçlüktedir. Madde güçlük indeksi p_j sembolü ile gösterilir ve j madde numarasını temsil etmek üzere Eşitlik 6 aracılığıyla hesaplanabilir:

$$p_j = \frac{\overline{X}_j - min_j}{max_j - min_j} \qquad \textit{Eşitlik 6}$$

Bu eşitlikte $\overline{X}_j$, güçlüğü hesaplanmak istenen maddenin ortalama puanını; min_j o maddeden alınabilecek en düşük puanı ve max_j ise o maddeden alınabilecek en yüksek puanı temsil eder. Buna göre, 1-0 şeklinde puanlanan bir madde için $max_j = 1$ ve $min_j = 0$ olacağından Eşitlik 6 otomatik olarak $\overline{X}_j$'ye dönüşür. $\overline{X}_j$ hesaplanırken her bir puan (ya 0 ya da 1) toplanıp toplam kişi sayısına bölündüğünden 1-0 puanlanan maddeler için güçlük, "*doğru yanıtlayan kişi sayısı / toplam kişi sayısı*" formülüyle hesaplanır. 1-0 dışında puanlanan maddeler (örneğin, açık uçlu maddeler) içinse o maddeden alınan puan ortalamalarından, maddeden alınabilecek en düşük puan (genellikle 0) çıkarılıp elde edilen fark maddenin beklenen ranjına bölünür. Örneğin, alınabilecek en yüksek puanın 10, en düşük puanın 0 olduğu bir açık uçlu maddeden alınan puanların aritmetik ortalaması 6,70 ise bu maddenin güçlüğü: (6,70 – 0) / (10 – 0) = 0,67 olarak bulunur.

Yukarıda aktarılan genel eşitlik (eşitlik 6) haricinde, en başarılı ve en başarısız %27'lik katılımcı grubunu dikkate alarak da madde güçlüğünü hesaplamak mümkündür. Alt ve üst %27'lik gruplar yöntemine göre madde güçlüğü Eşitlik 7 ile hesaplanır.

$$p_j = \frac{n_{d\ddot{u}} + n_{da}}{n_{\ddot{u}} + n_a} \qquad \textit{Eşitlik 7}$$

$n_{d\ddot{u}}$*:* Üst %27'lik grupta maddeye doğru yanıt veren kişi sayısı
n_{da}*:* Alt %27'lik grupta maddeye doğru yanıt veren kişi sayısı
$n_{\ddot{u}}$ *ve* n_a*:* Üst ve alt %27'lik gruplardaki kişi sayısı.



Alt ve üst gruptaki kişileri belirlerken mutlaka %27'lik dilimi referans almak gibi bir zorunluluk yoktur. Madde analizi için en sık başvurulan dilim %27 olmakla birlikte; Murphy ve Davidshofer (2014) alt ve üst grubun testin uygulandığı örneklemin %25 ile %35'ini kapsayabileceğini belirtmiş, bu aralıktaki farklı yüzdelik dilimlerin benzer sonuçlar üreteceğini vurgulamıştır.

Test geliştirme sürecinde madde güçlüğünün nasıl olması gerektiği testin amacına bağlıdır. Bu noktada madde varyansı, yani maddenin testi alan bireyler arasındaki farklılığı ne derecede ortaya koyabildiği ön plana çıkmaktadır. Madde varyansı bir maddenin güçlüğü ile ilişkili olup "pj × qj" ile hesaplanır. Bu işlemdeki *qj*, 1 - pj'dir. Yani bir maddenin varyansı pj(1 – pj) eşitliği ile bulunur. Bu bağlamda, madde varyansı madde güçlüğünün 0,50 olduğu durumda en yüksek değerini alır ve 0,50(1 - 0,50) = 0,25 olur. Bunun dışındaki madde güçlüklerinde madde varyansı 0,25'ten daha küçük çıkar. Örneğin, madde güçlüğünün 0,10 olduğu durumda madde varyansı 0,10(1 – 0,10) = 0,09 olarak elde edilir. Madde güçlüğünün 0,10 olması; bu maddenin testi alan grubun %10'u tarafından doğru, %90'ı tarafından ise yanlış yanıtladığını gösterir. Bahse konu madde testi alanlar arasındaki farklılığı ortaya koymakta, güçlüğü 0,50 olan bir maddeye kıyasla yetersiz kalmaktadır.

Öte yandan özellikle çoktan seçmeli maddelerde şans başarısı söz konusudur. Örneğin, beş seçenekli bir maddenin şansla doğru yanıtlanma olasılığı 1 / 5 = 0,20'dir. Bu durum göz önüne alındığında bireyler arasındaki farklılığı en çok ortaya çıkarması beklenen madde güçlük düzeyinde bir düzeltme gerekir. Bunun için, *m* seçenek sayısını göstermek üzere madde güçlüğüne 0,50 / m eklenmelidir (Crocker & Algina, 2008). Buna göre beş seçenekli bir madde için bireyler arasındaki farklılığı en fazla ortaya koyması beklenen madde güçlüğü 0,50 değil, 0,50 + 0,50 / 5 = 0,60'tır. Lord (1952) ise yaptığı simülasyon çalışmasında hesaplanan bu değerden biraz daha kolay olan maddelerin seçilmesi halinde en yüksek test güvenirliğine ulaşıldığını ifade etmiştir.

Yukarıdaki hesaplama, madde varyansını en yüksek yapması beklenen madde güçlüğüne yönelikti. Benzer şekilde şans başarısı dikkate alınarak hesaplanan madde güçlüğünde bir düzeltme yapılmak istenirse Eşitlik 8 kullanılabilir (Erkuş, 2003):

$$p_j' = \frac{p_j - c}{1 - c}$$

p_j': *Düzeltilmiş madde güçlüğü*
c: *Maddenin şansla doğru yanıtlanma olasılığı*

Eşitlik 8

Buna göre, beş seçenekli bir madde için madde güçlüğü 0,40 olarak elde edildiyse düzeltilmiş madde güçlüğü: p_j' = (0,40 – 0,20) / (1-20) = 0,25 bulunur. Dikkat edilirse düzeltilmiş madde güçlük indeksi (0,25), ilk hesaplanan madde güçlüğü indeksine (0,40) göre azalmış, daha açık bir anlatımla maddenin gerçekte daha zor olduğu belirlenmiştir. Bu eşitlik az önce beş seçenekli bir madde için varyansı en yüksek değere ulaştırması beklenen 0,60 değerine uygulanırsa (0,60 – 0,20) / (1 – 0,20) = 0,50 değeri elde edilir. Öte yandan Lord (1952) maddenin şansla doğru yanıtlanma olasılığına dayalı madde güçlüğünün hesaplanmasında, özellikle düzeltme formülü kullanıldığı durumda (adayın doğru sayısından yanlış sayısının seçenek sayısının bir eksiğine bölünerek çıkarılması halinde), adayların tahminde bulunma davranışının azaldığına; bu nedenle hesaplamada tam olarak 1/m kullanmanın uygulamadaki sınırlılığına dikkat çekmiştir.

Testteki maddelerin güçlük indekslerinin toplamı ise testin ortalamasını vermektedir. Dolayısıyla bireyler arasındaki farklılığı en üst düzeyde ortaya koymak için ya tüm maddeler orta güçlükte olmalı ya da madde güçlük indeksleri farklı değerler alsa da (zor ya da kolay maddeler) ortalama test güçlüğünün 0,50 civarında olması sağlanmalıdır (Erkuş, 2003). Maddeler, güçlük indekslerine göre kolay-orta-zor şeklinde bir sınıflandırmaya tabi tutulmak istendiğinde, herkes tarafından kabul gören sınırlar bulunmamaktadır. Popham (2014), 0,20'den küçük değerlerin zor; 0,80 ve üzeri değerlerin kolay madde olarak sınıflandırıldığını ifade etmiş ama bu sınıflandırmaya karşı nötr bir tavır takınmıştır. Çünkü bir maddeye ait güçlüğün 0,90 olarak bulunması, maddenin uygulandığı grubun %90'ı tarafından doğru yanıtlandığını, dolayısıyla gruba kolay geldiğini bildirmesinin yanında grubun %90'ının maddenin ölçtüğü kazanıma ulaştığını da gösterir. Bu bakımdan yalnızca indeks değerini yorumlamak yerine, maddenin hangi bağlamda uygulandığına göre bir inceleme yapmak gerektiğini ifade etmiştir.

Alanyazında madde güçlüğünü sınıflandırmada kullanılabilecek keskin eşik değerlerden kaçınılsa da uygulayıcıların farklı kesme noktaları tanımladığı görülmektedir. Örneğin, Assessment System Corporation; 0,00 – 0,30 aralığını çok zor, 0,31 – 0,50 arasını zor, 0,51 – 0,70 arasını orta, 0,71 – 0,90 arasını kolay ve 0,91 – 1,00 arasını çok kolay madde olarak yorumlamaktadır (Thompson, 2020). Bu anlamda madde güçlük indeksi için 0,00 – 0,20 çok zor, 0,21 – 0,40 zor, 0,41 – 0,60 orta, 0,61 – 0,80 kolay ve 0,81 – 1,00 çok kolay kategorizasyonu pratikte kullanılabilecek bir sınıflandırmadır.

Madde Tepki Kuramına Göre

Örtük özellikler kuramı olarak da adlandırılan Madde Tepki Kuramı (MTK), bireyin bir maddeye vereceği yanıtın olasılığı ile yetenek düzeyi arasındaki ilişkiyi genellikle lojistik bir fonksiyonla inceler (Hambleton & Swaminathan, 1985). MTK'nin KTK ile benzerlikleri, KTK'ye olan üstünlük ve sınırlılıkları kitabın 2. bölümünde detaylı biçimde verilmiştir. Bu bölümde yalnızca MTK'ye dayalı madde güçlüğü ile ayırt ediciliğine ilişkin bazı hatırlatmalar yapılıp hesaplama ve yorumlama üzerinde durulmuştur.

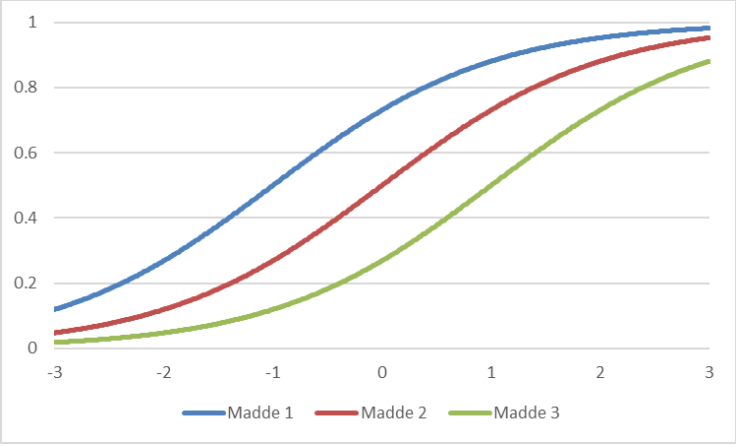
MTK'ye göre madde güçlüğü, maddeyi doğru yanıtlama olasılığının %50 olduğu yetenek düzeyini temsil eder ve *b parametresi* olarak adlandırılır (Baker & Kim, 2017). Örneğin, bir maddenin *b parametresi* -1 ise yetenek düzeyi -1 olan bir bireyin bu maddeyi doğru yanıtlama olasılığı %50'dir. Anlaşılacağı üzere MTK'de madde parametresi, yetenek düzeyi ile temsil edilir. Başka bir deyişle madde ve birey özellikleri aynı ölçekte yer alır.

MTK'ye göre madde güçlük parametresinin kestirimi, KTK'de yapılan madde güçlük indeksi hesaplanmalarından çok daha karmaşıktır. Bölümün sadeliğini koruyabilmek adına aşağıda *b parametresi*nin kestirimi yerine, bu parametrenin yorumu üzerine durulmuştur. Ancak ilgili okuyucular Marjinal En Yüksek Olabilirlik (Marginal Maximum Likelihood), Birleşik En Yüksek Olabilirlik (Joint Maximum Likelihood) gibi yöntemleri araştırabilir.

MTK'de *b parametresi*nin yetenek düzeyi ile aynı ölçekte ifade edilmesi nedeniyle *b parametresi*nin değeri azaldıkça madde kolaylaşır (maddeyi %50 olasılıkla doğru yanıtlamak için daha az yetenek gerekir), *b parametresi*nin değeri arttıkça madde zorlaşır. Buna bağlı olarak KTK ve MTK'deki madde güçlüğü değerleri ters şekilde yorumlanır. *b parametresi* yetenek düzeyi ile aynı ölçekte ele alındığından teorik olarak -∞ ile +∞ arasında uzanabilir. Ancak pratikte genellikle -3 ile +3 ya da -4 ile +4 arasında değerler alır. Şekil 3'te *b parametresi* sırasıyla -1, 0 ve 1 olan üç maddeye ait madde karakteristik eğrileri verilmiştir.

Şekil 3

Üç Maddeye Ait Madde Karakteristik Eğrisi



Madde karakteristik eğrileri incelenirse Madde 1'e ait eğri daha solda; madde 3'e ait eğri ise en sağdadır. Bunun nedeni, eğrilerin orta noktalarının madde 1 için -1.00 ve madde 3 için 1.00'da yer almasıdır. Buna göre en kolay madde 1. madde iken (çünkü bu maddeyi doğru yanıtlamak için gerekli yetenek düzeyi daha düşüktür), 3. madde en zor maddedir (bu maddeyi doğru yanıtlamak için gerekli yetenek düzeyi daha yüksektir).



Bu noktalar aynı zamanda madde karakteristik eğrilerinin dönüş yaptığı noktalara denk gelmektedir. *b parametresi*nin olduğu noktada madde karakteristik eğrisinin yönünü değiştirdiğine dikkat edilmelidir.

3.2. Madde Ayırt Ediciliği

Bir testte yer alan maddelerin ölçülen özelliğe az ya da çok sahip olanları ayırabilmesi istenir. Örneğin, testin genelinde başarılı olan öğrenciler maddeyi doğru yanıtlarken testin genelinde başarısız olan öğrencilerin maddeyi yanlış yanıtlaması beklenir. Bir maddenin bunu ne derece başardığını belirlemek için madde ayırt edicilik indeksi kullanılır. Aşağıda hem KTK temelinde madde ayırt edicilik indeksine hem de MTK temelinde madde ayırt edicilik parametresine yer verilmiştir.

Klasik Test Kuramına Göre

KTK'ye göre madde ayırt edicilik indeksi r_{jx} sembolü ile gösterilir. Bu sembolde *r* korelasyonu; *j* maddeyi ve *x* toplam puanı temsil eder. Dolayısıyla madde ayırt edicilik indeksi bir maddeden alınan puanlarla, testten alınan toplam puanlar arasındaki korelasyon katsayısı olarak ifade edilir. Madde ayırt edicilik indeksinin hesaplanmasında kullanılacak korelasyon katsayısı, madde türüne ve test geliştiricisinin altta yatan varsayımlarına bağlıdır. İkili puanlanan (1 – 0) maddeler için madde puanları iki kategorili, toplam puan ise süreklidir. Haliyle iki kategorili süreksiz bir değişken ile sürekli bir değişken arasındaki korelasyon katsayısının hesaplanması gerekir. Bu amaçla çift serili (r_{bis}) ya da nokta çift serili (r_{pbis}) korelasyon katsayısından yararlanılır. Madde puanlarının (1 – 0) gerçek süreksiz olduğu varsayılıyorsa nokta çift serili; yapay süreksiz (maddenin ölçtüğü özellik miktarı aslında sürekli, ancak yapay olarak 1 ve 0 olduğu) olduğu varsayılıyorsa çift serili korelasyon katsayısı kullanılır (Crocker & Algina, 2008). Maddeler çok kategorili puanlanıyorsa (örneğin, 0 – 10 arası puan), doğrudan Pearson momentler çarpımı korelasyon katsayısı ya da Spearman sıra farkları korelasyon katsayısı kullanılabilir. Ayrıca korelasyonda "madde – toplam" yerine düzeltilmiş madde toplam korelasyonunun kullanılması önerilir. Düzeltilmiş madde toplam korelasyonu, "madde – kalan" korelasyonu olarak da isimlendirilir ve herhangi bir maddeden alınan puanlar ile o madde haricinde testin kalan maddelerinden alınan toplam puan arasındaki korelasyonu gösterir. Madde – toplam korelasyonu yerine madde – kalan korelasyonun tercih edilmesinin nedeni, toplam puana ayırt ediciliği incelenen maddeye ait puanların da dâhil edilmesinin hesaplanacak madde korelasyonunu yapay olarak yükseltmesidir (Metsamuuronen, 2020).

Bunun yanında gruplar arası farkı (ya da oranları) temel alan alt – üst %27'lik gruplar yöntemi, Brennan indeksi ve Brennan'ın özel durum indeksi gibi madde ayırt edicilik indeksleri de vardır. Bu yöntemler korelasyon katsayısına dayalı olmadığı için r_{jx} yerine D veya B gibi harflerle gösterilir. Alt – üst %27'lik gruplar yöntemi, hesaplama kolaylığı nedeniyle özellikle elle hesaplamada tercih edilen bir yöntemdir ve Eşitlik 9 aracılığıyla hesaplanır (Popham, 2014).

$$D = p_{ü} - p_{a} \qquad \textit{Eşitlik 9}$$

Bu eşitlikte $p_{ü}$, üst grubun maddeyi doğru yanıtlama oranını; p_a ise alt grubun maddeyi doğru yanıtlama oranını temsil eder. Bu eşitliği kullanmak için öncelikle testi alan bireyler toplam puanlarına göre büyükten küçüğe doğru sıralanır. Ardından testi alan toplam kişi sayısının %27'si belirlenir ve puanı en yüksek olan %27'lik grup ile en düşük olan %27'lik grup seçilir. Daha sonra her bir madde için üst grupta maddenin doğru yanıtlanma oranından, alt grupta maddenin doğru yanıtlanma oranı çıkarılır. Bu yöntemde kullanılan %27 oranı ise Kelley (1939) tarafından gruptaki bireyler arası farkı en fazla ortaya çıkaran oran olarak bulunmuş ve ardından diğer araştırmacılar tarafından kullanılmaya başlanmıştır.

Brennan'ın indeksinde ise amaç bir maddenin mutlak bir ölçütün altında ve üstünde yer alan bireyleri ne kadar iyi ayırt ettiğini belirlemektir. Bu yönüyle daha önce aktarılan katsayılardan ayrılmaktadır. Şimdiye kadar yer alan madde ayırt edicilik indeksleri bir maddenin grubun genelini nasıl ayırdığına odaklanırken Brennan indeksi grubun belli bir kesiminin diğerlerinden nasıl ayrıldığına odaklanır ve Eşitlik 10 ile hesaplanır (Brennan, 1972).

$$B = \frac{U}{n_1} - \frac{L}{n_2}$$

n_1: Ölçüt puanın üzerindeki birey sayısı
n_2: Ölçüt puanın altındaki birey sayısı
U: Ölçüt puanın üzerinde puan alan bireylerden maddeye doğru yanıt veren sayısı
L: Ölçüt puanın altında puan alan bireylerden maddeye doğru yanıt veren sayısı

Eşitlik 10

Brennan indeksi sayesinde bir maddenin belli bir yeterlik düzeyinin üzerindeki ve altındaki bireyleri ne derece ayırt edebildiği saptanabilir. Arıkan ve Aybek (2022) maddenin belli bir yeterlik düzeyi yerine testi alan grubun belli bir bölümünü birbirinden ne derece ayırabildiğini belirlemek için Brennan'ın indeksini özel koşullar için uyarlamış ve önerdikleri uyarlamanın özellikle az sayıda kişinin seçilmesinin amaçlandığı durumlarda diğer indekslere göre daha etkili olduğunu ortaya koymuşlardır. Buna göre özel durumlar için Brennan'ın indeksi Eşitlik 11'deki gibi hesaplanmaktadır.

$$B_{x-y} = p_x - p_y$$

p_x: Üst %x grup için madde güçlüğü ya da maddeye doğru yanıt verme oranı
p_y: Kalan %y grup için madde güçlüğü ya da maddeye doğru yanıt verme oranı

Eşitlik 11

Bu eşitliğe göre, örneğin grubun en başarılı %10'luk grubunu kalan %90'lık grubundan ne derece ayırdığını belirlemek için:

$$B_{10-90} = p_{10} - p_{90}$$

eşitliği kullanılır. En başarılı %10'luk grubun maddeyi doğru yanıtlama oranından, kalan grubun maddeyi doğru yanıtlama oranı çıkarılır. Özellikle az sayıda öğrencinin seçildiği testler için B_{x-y}, belli bir grubun diğerlerinden ayırt edilmesi bakımından diğer madde ayırt edicilik indekslerine göre daha doğru sonuçlar vermektedir (Arıkan & Aybek, 2022).

Madde ayırt edicilik indeksi bir korelasyon katsayısı olduğundan -1 ile +1 arasında değerler alabilir. Madde ayırt edicilik indeksinin negatif olması, maddenin ters çalıştığını gösterir. Başka bir deyişle başarısı düşük olanlar maddeyi doğru yanıtlarken başarısı yüksek olanlar maddeyi yanlış yanıtlamıştır. Bu durumda akla ilk gelmesi gereken, cevap anahtarının yanlış hazırlanmış olması ihtimalidir. Cevap anahtarında herhangi bir problem yoksa maddenin testi alan öğrenciler tarafından anlaşılmamış ya da yanlış anlaşılmış olma olasılığı üzerinde durulmalıdır.

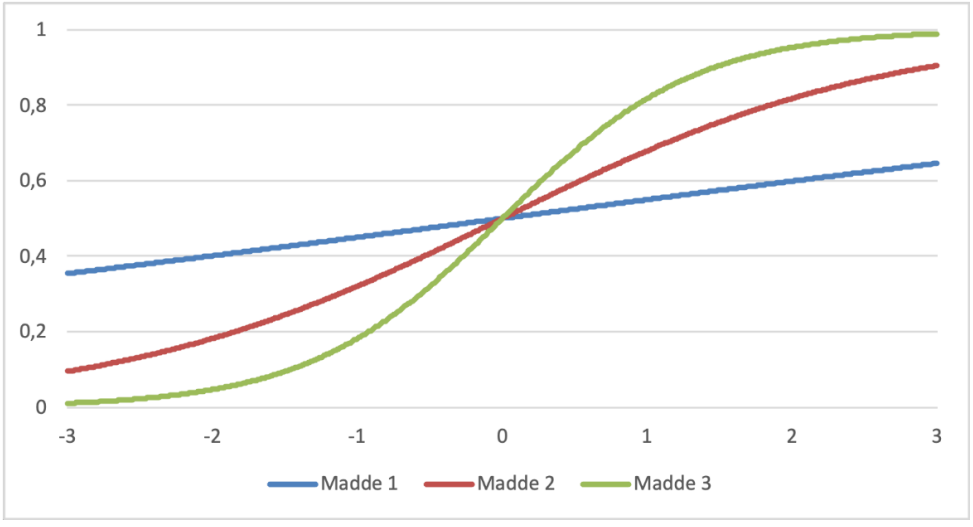
Test geliştirme sürecinde madde ayırt edicilik indeksinin 0,40'tan büyük olması beklenir. Madde ayırt edicilik indeksi 0,30 ile 0,39 arasında ise maddede düzeltme gerekmez ya da çok az düzelme gerekir. Madde ayırt edicilik indeksinin 0,20 ile 0,29 arasında olması halinde maddede düzeltme gerekirken 0,19 ve altındaki indekse sahip maddelerin testten çıkarılması gerekir (Popham, 2014). Bunun yanı sıra bir madde tüm katılımcılar tarafından doğru yanıtlandığında madde ayırt edicilik indeksi 0 olur. Bu durumda sıralanan ölçütler maddenin testten çıkarılması gerektiğine işaret eder. Fakat özellikle sınıf içi uygulamalarda bir maddenin bütün öğrenciler tarafından doğru yanıtlanması, tüm öğrencilerin maddenin ölçtüğü kazanıma ulaştığı anlamına gelir. Dolayısıyla yukarıda bahsedilen ölçütler testin kullanım amacı bağlamında değerlendirilmelidir.

Madde Tepki Kuramına Göre

MTK'de madde ayırt ediciliği için *a parametresi* kullanılır. Tek boyutlu MTK modellerinde iki kategorili ya da çok kategorili puanlanan maddeler için tek bir *a parametresi* bulunur. Bu parametrenin değeri büyüdükçe maddenin ayırt ediciliği artar, ancak bu değer çok yüksek olduğunda madde oldukça dar bir yetenek aralığı için ayırt edici olur ve bu aralık dışında kalan bireylerin büyük bölümü bu maddeyi ya doğru ya da yanlış yanıtlar (Baker, 2001). Şekil 4'te *b parametresi* 0 olan; *a parametresi* ise sırasıyla 0,20, 0,75 ve 1,50 olan üç maddeye ait madde karakteristik eğrileri verilmiştir.

Şekil 4

Ayırt Edicilikleri Farklı Üç Maddeye Ait Madde Karakteristik Eğrileri



Üç maddeye ait karakteristik eğrileri incelendiğinde Madde 1'e ait eğrinin oldukça yatay seyrettiği görülmektedir. Dolayısıyla bu madde, farklı yetenek düzeyindeki bireyleri ayırt etmede oldukça yetersizdir. Öyle ki yetenek düzeyi -2 olan (evrenin %2'sinden daha başarılı) bir bireyin bu maddeyi doğru yanıtlama olasılığı yaklaşık %40 iken yetenek düzeyi +2 olan (evrenin %97'sinden daha başarılı) bir bireyin bu maddeyi doğru

yanıtlama olasılığı yaklaşık %60'tır. Başka bir deyişle yetenek düzeyinin farklılaşması bu maddeyi doğru yanıtlama olasılığında önemli bir değişikliğe yol açmamaktadır. Öte yandan madde 3 incelendiğinde yetenek düzeyi -2 olan bir bireyin bu maddeyi doğru yanıtlama olasılığının yaklaşık %10; yetenek düzeyi +2 olan bir bireyin ise bu maddeyi doğru yanıtlama olasılığının yaklaşık %90 olduğu anlaşılmaktadır. Buna göre, bireylerin yetenek düzeyleri bu örnekteki 3. maddenin doğru yanıtlanma olasılığını önemli düzeyde etkilemektedir.

Bir maddenin *a parametresi* için Baker (2001) tarafından eşik değerler önerilse de pratikte *a parametresi*nin 1'den büyük olması istenir. Özellikle bir maddenin, birey hakkında ne düzeyde bilgi verdiğini temsil eden madde bilgi fonksiyonu hesaplanırken *a parametresi*nin karesi çarpan olarak hesaplamaya dâhil edilir. Bu nedenle *a parametresi*nin 1'den küçük olması halinde, hesaplanan bilgi fonksiyonu oldukça küçük değerler almaya başlar. Aşağıda üç farklı *a parametresi*nin madde bilgi fonksiyonuna katkıları hesaplanmıştır:

Tablo 4

a Parametresi ve Madde Bilgi Fonksiyonuna Katkı Miktarları

	a = 0,85	*a = 1,00*	*a = 1,15*
Madde bilgi fonksiyonuna katkı (a^2)	0,723	1,000	1,323

Görüldüğü üzere, *a parametresi* 1'den küçük değerler aldığında madde bilgi fonksiyonuna katkısı kendi değerinden daha az olurken 1'den büyük değerler aldığında madde bilgi fonksiyonuna katkısı kendi değerinden daha büyük olmaktadır. Öte yandan çok yüksek değere sahip bir *a parametresi* maddenin oldukça dar bir aralığı ayırt ettiği şeklinde yorumlanır.

Madde Tepki Kuramına Göre Diğer Madde Parametreleri

Detayları kitabın 2. bölümünde verilen ikili puanlanan maddeler için kullanılan MTK modellerinde *a* ve *b* parametrelerinin yanında *c* ve *d* parametresi de kullanılmaktadır. Bilindiği üzere *c* parametresi, maddenin ölçtüğü özelliğe sahip olmayan bireyin o maddeyi şansla doğru yanıtlamasına ilişkin bilgi verirken *d* parametresi maddenin ölçtüğü özelliğe sahip olan bireyin dikkatsizlik gibi nedenlerle maddeyi yanlış yanıtlama olasılığına ilişkin bilgi sağlar.

Kullanılan MTK modeli, *a, b, c* ve *d* parametrelerinden kaç tanesinin değişken olarak ele alındığına göre adlandırılır. Yalnızca *b* parametresi değişken olarak alınıyorsa 1 Parametreli Lojistik Model (PLM), *b* ve *a* parametreleri değişken olarak alınıyorsa 2PLM; *b, a* ve *c* parametreleri değişken olarak alınıyorsa 3PLM; *b, a, c* ve *d* parametreleri değişken olarak alınıyorsa 4PLM adını almaktadır. Pratikte 1, 2 ve 3PLM kullanılsa da 4PLM daha çok kuramsal çalışmalarda kullanılmaktadır.

3.3. Yanıtların Seçeneklere Dağılımı (Çeldiricilerin Analizi)

Maddelere ait güçlük ve ayırt edicilik indeksleri ya da parametreleri maddelerin nasıl çalıştığı hakkında fikir verir. Hesaplanan indeksler ya da parametreler doğrultusunda bazı maddelerde düzeltmelere gidilebilir. Bu noktada yalnızca madde kökü değil seçeneklerin, özellikle de çeldiricilerin incelenmesi, madde hakkında daha derin fikir sahibi olunmasını sağlar. Daha açık bir anlatımla, çoktan seçmeli bir maddede çeldiriciler yeterince iyi değilse başarısız öğrenci için seçilebilir olmaz, dolayısıyla madde ayırt ediciliği düşer. Bu nedenle çeldiricilerin nasıl çalıştığını incelemek önemlidir. Özellikle testi alan bireylerin çok düşük bir kısmı tarafından seçilen çeldiriciler kontrol edilmeli, gerekli görülürse değiştirilmelidir. Ayrıca, özellikle başarısız gruptaki öğrencilerin doğru yanıtlayıp başarılı gruptaki öğrencilerin yanlış yanıtladığı maddelerde çeldiriciler analiz edilip olası sorunlar giderilmelidir.

İyi bir çeldirici başarısız öğrenciler tarafından, başarılı öğrencilere göre daha çok tercih edilmelidir. Bu bakımdan madde ayırt ediciliğinde kullanılan teknikler çeldirici analizi için de kullanılabilir. Şöyle ki alt – üst %27'lik gruplar yönteminde en başarılı %27'lik grubun doğru yanıtlama yüzdesinden, en başarısız %27'lik grubun doğru yanıtlama yüzdesi çıkarılıyordu. Aynı işlem, doğru yanıtlayan yüzdesi yerine bir çeldiriciyi seçme yüzdesi için yapıldığında beklenen durum üst grubun çeldiriciyi seçme yüzdesinin, alt grubunkinden küçük olmasıdır. Bir başka deyişle ideal olan alt gruptaki kişilerin daha yüksek bir oranda çeldiriciye gitmiş olmasıdır. Tahmin edileceği üzere bu durumda çeldiricilere ait istatistiklerin negatif değer alması beklenir (Gronlund, 1977). Eğer pozitif bir değer ortaya çıkıyorsa, yani bir çeldiriciye üst gruptaki öğrenciler, alt gruptaki öğrencilere göre daha fazla yöneliyorsa bu çeldirici ters çalışmış anlamına gelir. Böyle bir sonuçla karşılaşıldığında önce maddede bir hata olup olmadığı kontrol edilmeli, daha sonra çeldirici gözden geçirilmelidir.

Bazen bir çeldirici kimse tarafından seçilmeyebilir. Bu durumda çeldirici çalışmamış demektir. Sınıf içi öğrenci başarısının belirlenmesi amacıyla yapılan bir testte bu tür bir sonuçla karşılaşılması, öğrencilerin maddenin ölçmeye çalıştığı kazanıma ulaştığı şeklinde yorumlanabilir. Ancak çok sayıda öğrencinin / adayın aldığı standart testlerde, her bir çeldiricinin seçilmiş olması istenir. Haladyna ve Downing (1993), testi alan grubun %5'i tarafından seçilmemiş olan bir çeldiricinin, kimse tarafından seçilebilir olmadığını, dolayısıyla bu çeldiricinin çalışmadığını ifade etmiştir. Aynı zamanda bu bölümün Uygulama kısmında gösterilen Şekil 17 gibi bir madde için seçenek yanıt eğrisi elde edildiğinde, beklenen durum başarı arttıkça çeldiricinin seçilme oranının düşmesidir.

3.4. Madde Güvenirlik İndeksi

Madde güvenirlik indeksi, madde standart sapmasının ve madde – toplam korelasyonun çarpımı ile elde edilen bir katsayıdır. Hatırlanacak olursa 1 – 0 puanlanan maddelerde, madde standart sapması $\sqrt{p_j q_j}$ formülüyle hesaplanmaktadır. Bu formülden anlaşılacağı gibi aynı ayırt ediciliğe sahip iki maddeden varyansı (dolayısıyla standart sapması) daha büyük olan, testten alınan puanların güvenirliğine daha çok katkı sağlar. Öte yandan testteki maddeler orta güçlükte – ya da birbirine yakın güçlükte – ise maddelerin standart sapmaları da birbirine yakın olur. Böyle bir durumda test oluşturulurken sadece madde ayırt ediciliğini dikkate alma ile madde güvenirliğini dikkate alma arasında büyük bir fark oluşmaz (Crocker & Algina, 2008).

4. UYGULAMA

Bu bölümde hazır bir veri seti kullanarak test ve madde istatistikleri elde edilecektir. Bunun için başvurulabilecek çeşitli yazılımlar mevcuttur. Bununla beraber açık kaynaklı olması ve kullanım kolaylığı sunması bakımından burada Shiny Item Analysis (Martinkova & Hladka, 2023; Martinkova & Drabinova, 2018) yazılımı tercih edilmiştir. Shiny Item Analysis (SIA), KTK ve MTK temelli analizler, madde kalibrasyonları, değişen madde fonksiyonu, faktör analizi gibi birçok analizi yapabilen bir R paketidir. Bunun yanında Shiny arayüzü sayesinde herhangi bir kodlama bilgisi gerektirmeden kolayca kullanılabilmektedir. Dileyen okuyucular SIA'yı *shinyitemanalysis.org* üzerinden Quick Start (Hızlı Başlat) bağlantısına tıklayarak çalıştırabilir; dileyenler ise R'da:

```
install.packages("ShinyItemAnalysis")
library(ShinyItemAnalysis)
startShinyItemAnalysis()
```

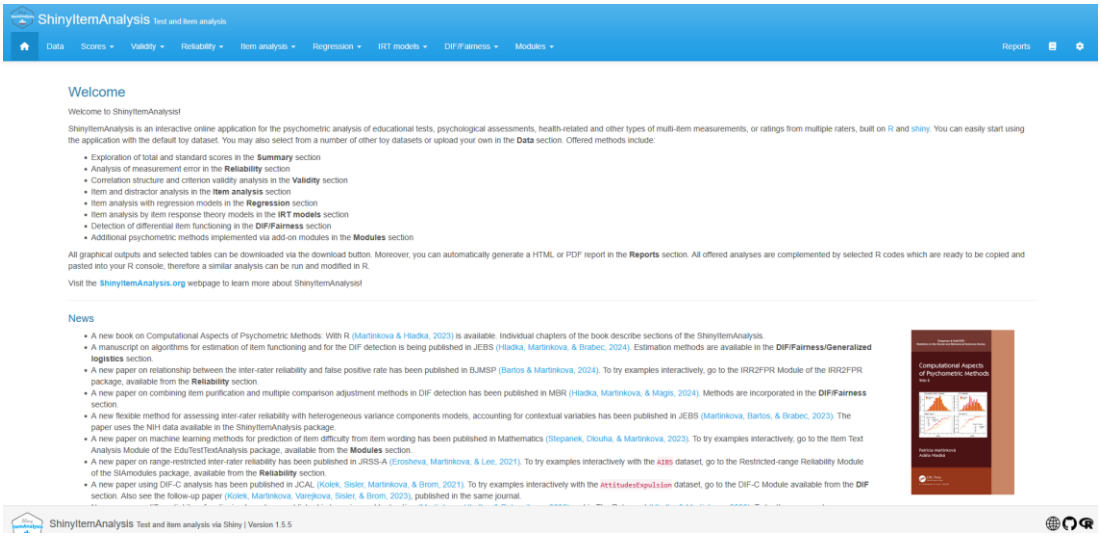
komutlarını kullanarak SIA'yı yükleyebilir ve ardından SIA'yı çağırıp uygulamayı başlatabilir.

4.1. Giriş Ekranı

SIA başlatıldığında Şekil 5'teki gibi bir ekranla karşılaşılır.

Şekil 5

SIA Karşılama Ekranı



Bu ekran bir karşılama ekranı olup SIA'nın kısa bir tanıtımını ve yazılımla ilgili güncel gelişmeleri içerir. Veri yükleme ve analiz işlemlerinin yapılacağı sayfaya ise Şekil 6'da yer alan üst menüden ulaşılabilir.

Şekil 6

SIA Üst Menüsü



Bu menüdeki ögeler işlevleri ile birlikte aşağıda listelenmiştir:

- *Data:* Veri yükleme
- *Scores:* Test istatistikleri (Merkezî eğilim ve dağılım ölçüleri, standart puanlar)
- *Validity:* Ölçüte ve içyapıya dayalı geçerlik kanıtları (korelasyon matrisi, faktör analizi)
- *Reliability:* İç tutarlılık anlamında güvenirlik katsayıları
- *Item analysis:* Madde analizleri (Madde güçlüğü, ayırt ediciliği, madde - yanıt grafikleri)
- *Regression:* Lojistik regresyon (özellikle MTK ile ilişkisini incelemek amacıyla)
- *DIF/Fairness:* Değişen Madde Fonksiyonu, Yanlılık
- *Modules:* Kullanıcılar tarafından eklenen modüller

4.2. Data (Veri) Ekranı

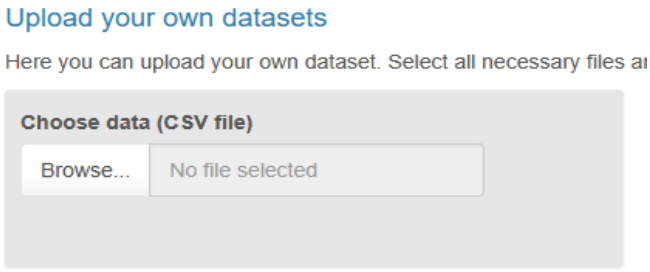
Data sekmesinde; Data (Veri), Basic Summary (Temel özet) ve Data Exploration (Veriyi keşfet) şeklinde üç alt sekme bulunur. Data sekmesinden kullanılacak veriler seçilir/yüklenir. Basic summary sekmesinde verilere ait temel özet bilgiler (en küçük ve en büyük değerler, kişi sayısı gibi) sunulur. Data exploration sekmesinde ise yüklenen veriler incelenebilir. Böylece doğru veri ile çalışılıp çalışılmadığı kontrol edilir.

Data alt sekmesinde kullanıcılar kendi verilerini yükleyebilecekleri gibi yazılım içerisindeki hazır veri setlerinden de yararlanabilirler. Tüm okuyucuların aynı okuma deneyimini elde edebilmesi adına bu bölümde yazılım içerisinde yer alan hazır verilerden GMAT verisi ile çalışılmıştır. Ancak aşağıda kendi verisini yüklemek isteyen okurların bunu nasıl yapabileceği de açıklanmıştır.

Çalışılacak veriyi yüklemek için "Upload your own datasets" (kendi veri setinizi yükleyin) başlığı altındaki "Choose data" (Veri seç) bölümü kullanılır (Şekil 7). SIA, yalnızca CSV formatındaki verilerle çalışmaya izin verdiğinden verinin öncelikle bu formatta kaydedilmesi gerekir. Microsoft Excel gibi birçok yazılım verinin CSV formatında kaydedilmesine izin vermektedir.

Şekil 7

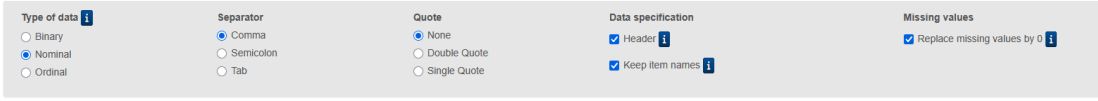
Veri Yükleme Alanı



Çalışılacak veri seçildikten sonra veri yapısına ilişkin bazı bilgilerin tanımlanması gerekir. Bunun için Şekil 8'deki bölümden tercihler yapılmalıdır.

Şekil 8

SIA Veri Dosyası Özelliklerinin Tanımlanması



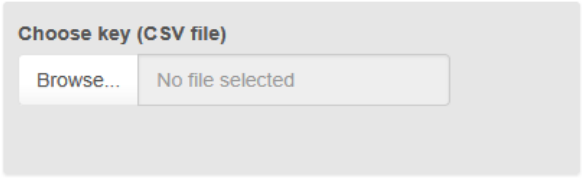
Bu bölümde kullanıcının tercih edebileceği seçenekler ve anlamları şöyledir:

- *Type of data (veri tipi):* Nasıl bir veri yüklendiği seçilir. 1 - 0 (binary), kategorik yani A-B-C-D gibi (nominal) ya da 1-2-3-4-5 gibi sıralama (ordinal) seçeneklerinden biri işaretlenmelidir.
- *Seperator (ayraç):* CSV dosyasında sütunların nasıl ayrıldığının seçilmesi gerekir. Virgül (comma), noktalı virgül (semicolon) ya da sekme (tab) seçeneklerinden biri işaretlenmelidir. CSV dosyasının nasıl ayrıldığından emin olunamaması halinde, CSV dosyası not defteri ile açılarak kontrol sağlanabilir. Microsoft Excel varsayılan olarak noktalı virgül ayracı ile CSV dosyası oluşturmaktadır.
    - *Quote (tırnak işareti):* Veride metin varsa bu metinlerin veride nasıl belirtildiğinin girildiği bölümdür. Verideki metin "Ölçme, Seçme ve Yerleştirme Merkezi Başkanlığı" şeklindeyse Single Quote yani tek tırnak; ""Ölçme, Seçme ve Yerleştirme Merkezi Başkanlığı"" şeklindeyse Double Quote yani çift tırnak seçilmelidir. Metinlerin veride nasıl yer aldığı bilinmiyorsa CSV dosyası not defteri benzeri bir uygulama ile açılarak kontrol edilebilir.
- *Data specification (veri özellikleri):* Verinin ilk satırı değişken isimlerinden oluşuyorsa "Header" seçili olmalıdır. Değişken isimlerinin aynen korunması isteniyorsa "Keep item names" seçeneği seçili kalmalıdır, aksi halde SIA tarafından otomatik olarak değişken ismi atanacaktır.
- *Missing values (kayıp değerler):* Verideki kayıp değerlerin sıfır olarak puanlanması / kodlanması isteniyorsa "Replace missing values by 0" seçili olmalıdır.

Bir sonraki girdi seçeneği Şekil 9'da gösterilen ve cevap anahtarının yüklenebileceği bölümdür.

Şekil 9

Cevap Anahtarı Yükleme Alanı

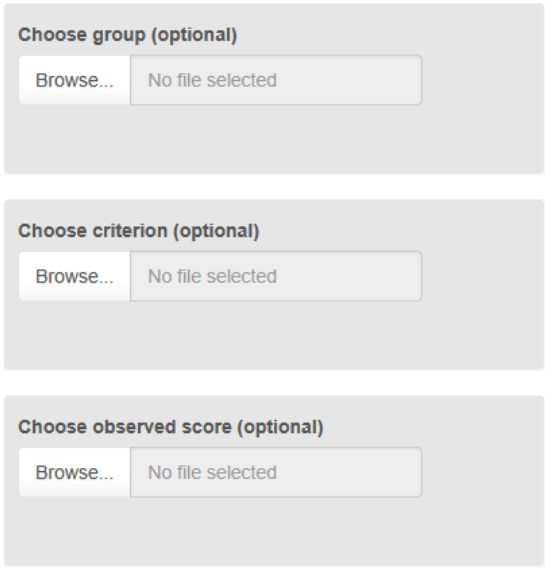


Eğer yüklenen veri 1 – 0 olarak puanlanmış veri değil de doğrudan A - B - C - D gibi yanıtlardan oluşan bir veriyse bu bölüme cevap anahtarının yüklenmesi gerekir. SIA cevap anahtarına göre otomatik olarak puanlamayı yapar. Ancak A ve B kitapçığı gibi bir desteği bulunmamaktadır. Bu nedenle yüklenecek verilerin tamamının aynı kitapçığa göre düzenlenmiş olması şarttır. Ayrıca A - B - C - D gibi kategorik bir veri ve cevap anahtarı yüklenecekse bir üstteki Type of data kısmından Nominal seçili olmalıdır.

Daha sonra Değişen Madde Fonksiyonu / Yanlılık gibi çalışmalarda kullanmak üzere grup verisinin; ölçüte dayalı geçerlik çalışmasında kullanmak üzere ölçüt puanların ve gözlenen puanların yüklenebileceği alan gelmektedir (bkz. Şekil 10).

Şekil 10

Grup, Ölçüt ve Gözlenen Puan Yükleme Alanı



Bu verilerin yüklenmesi zorunlu olmayıp ihtiyaç duyulması halinde yüklenir. Her bir verinin CSV türünde olması gerektiği unutulmamalıdır. Yüklenmek istenen veriler seçildikten sonra sağ alttaki Upload data (Veriyi yükle) düğmesine tıklanarak veriler yüklenebilir.

Bu bölümde örnek olarak GMAT verisi ile çalışılacağı belirtilmişti. Dolayısıyla "Training datasets" altından GMAT verisi seçilerek uygulamanın sonraki aşamalarına devam edilmiştir.

GMAT verisinin nasıl bir yapıda olduğunu görmek için Data sekmesi altından sırasıyla Basic summary ve Data exploration alt sekmeleri incelenir. Basic summary alt sekmesinde; sırasıyla ana veri seti, puanlanmış veri seti, grup, ölçüt değişken ve gözlenen puanlara ait bir özet sunulur. Data exploration alt sekmesinde ise doğrudan verilerin ham hali incelenebilir. Böylece özellikle kendi verisini yükleyen kullanıcılar verinin doğru bir şekilde yüklenip yüklenmediğini görebilirler.

4.3. Test İstatistikleri

Test puanlarına ait aritmetik ortalama, ortanca, standart sapma gibi değerlerin yanında standartlaştırılmış z ve T-puanlarına ait bilgileri elde etmek için Scores (puanlar) sekmesinden yararlanılır. Scores altında Total scores (toplam puanlar) ve Standard scores (standart puanlar) olmak üzere iki seçenek yer alır. Total scores seçildiğinde açılan sayfada Şekil 11'de görülen temel test istatistikler elde edilebilir.

Şekil 11

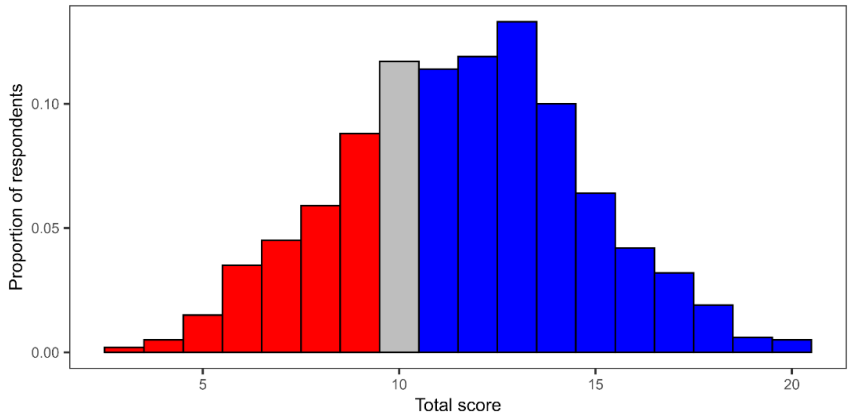
Temel Test İstatistikleri

n	n_c	Min	Max	Mean	Median	SD	Skewness	Kurtosis
2000	2000	3.00	20.00	11.60	12.00	3.11	0.00	2.75

Burada öncelikle çalışılan verinin kaç kişiden oluştuğu (n), kayıp değerler çıkarıldıktan sonra kaç kişinin verisinin yer aldığı (n_c), en küçük değer (min), en büyük değer (max), aritmetik ortalama (mean), ortanca (median), standart sapma (SD), çarpıklık (skewness) ve basıklık (kurtosis) değerleri yer almaktadır. Örnekte 2000 kişilik bir veriyle çalışıldığı, en küçük puanın 3 ve en büyük puanın 20 olduğu görülmektedir. Ortalama 11,6 ve ortanca 12 olarak elde edilmiştir. Standart sapma ise 3,11'dir. Ortalama ve ortancanın birbirine yakınlığı, çarpıklığın sıfıra çok yakın olması; basıklığın ise 3 civarında olması nedeniyle puanların normale yakın olduğu söylenebilir. Bu durum tablonun hemen altında yer alan histogramla da desteklenmektedir (bkz. Şekil 12).

Şekil 12

Test Puanlarına İlişkin Histogram



Histogramdaki gri alan kesme noktasını, kırmızı alan kesme noktasının altını, mavi alansa kesme noktasının üstünde puan alanları temsil etmektedir. Kesme noktası kullanıcı tarafından histogramın hemen üzerinde yer alan kaydırma çubuğundan ayarlanabilmektedir.

Standard scores menüsüne gelindiğinde ise her bir ham puanın (total score) sırasıyla yüzdelik dilimi (percentile), yüzlük puanı (success rate), z ve T-puanları görülebilmektedir (bkz. Şekil 13). Şekil 13'te görüldüğü gibi SIA standart puanları her bir öğrenci için ayrı ayrı vermek yerine, her bir ham puana karşılık gelen standart puanları raporlamaktadır.

Şekil 13

Test Puanlarına İlişkin Standart Puanlar

Table by score

Total score	Percentile	Success rate	Z-score	T-score
3.00	0.00	15.00	-2.77	22.30
4.00	0.01	20.00	-2.45	25.52
5.00	0.02	25.00	-2.13	28.74
6.00	0.06	30.00	-1.80	31.96
7.00	0.10	35.00	-1.48	35.18
8.00	0.16	40.00	-1.16	38.40
9.00	0.25	45.00	-0.84	41.62
10.00	0.37	50.00	-0.52	44.84
11.00	0.48	55.00	-0.19	48.06
12.00	0.60	60.00	0.13	51.28
13.00	0.73	65.00	0.45	54.50
14.00	0.83	70.00	0.77	57.72
15.00	0.90	75.00	1.09	60.94
16.00	0.94	80.00	1.42	64.16
17.00	0.97	85.00	1.74	67.38
18.00	0.99	90.00	2.06	70.60
19.00	1.00	95.00	2.38	73.82
20.00	1.00	100.00	2.70	77.04

Validity (geçerlik) menüsü üzerinden ölçüt geçerliği ve korelasyon matrisi ile içyapıya dayalı geçerlik kanıtına (faktör analizi aracılığıyla) ulaşılabilmektedir. Ölçüte dayalı geçerlik kanıtı elde edebilmek için Data sekmesinde "Choose criterion" kısmına veri yüklenmiş olmalıdır. Örnekte her bir öğrenci için ölçüt puanlar da girilmiş olduğundan *Criterion Validity* menüsüne girildiğinde bir saçılım grafiği ve ardından ölçüt ile test puanları arasındaki korelasyon katsayısı raporlanmaktadır. Correlation structure kısmında ise maddeler arası korelasyon katsayıları raporlanmakta, aynı zamanda kolay yorumlama için sıcaklık grafiği (heatmap) sunulmaktadır. Korelasyon katsayısı olarak Pearson, Spearman ya da Polychroic alternatiflerinden biri seçilebilmektedir. Maddeler 1-0 şeklinde puanlanıyorsa tetrakorik korelasyon katsayı kullanılmalı, bunun için Polychroic seçeneği işaretlenmelidir. *Factor analysis* menüsünde ise testin nasıl bir faktör yapısını ölçtüğüne dair inceleme yapılabilmektedir. Bu menüde temel bileşen analizi (PCA) veya faktör analizi (FA); Pearson ya da polikorik korelasyon matrisine dayalı olarak yapılabilmektedir. Faktör analizi ve korelasyon katsayılarının detaylı sunumu bu bölümün kapsamına girmediğinden ilgili kısmın anlatımı sınırlandırılmıştır.

Reliability (güvenirlik) menüsünde ise iç tutarlılık anlamındaki güvenirlik katsayıları elde edilebilmektedir. Bu amaçla Spearman-Brown tahmini, iki yarıya bölme yöntemi ve Cronbach alfa katsayıları hesaplanmaktadır. Spearman-Brown tahmininde "Test X sayıda benzer maddeden oluşsaydı güvenirlik katsayısı ne olurdu?" ya da "Güvenirlik katsayısının Y olması isteniyorsa test buna benzer kaç maddeden oluşmalıydı?"

sorularına yanıt aranmaktadır. Örnek veride 20 madde bulunmakta ve güvenirlik katsayısı 0,55 olarak hesaplanmaktadır. *"Number of items in new data"* (Yeni verideki madde sayısı) kısmına 30 yazılırsa hemen altında 0,647 değeri görülür. Bunun anlamı, teste benzer nitelikte 10 madde daha eklenirse, yani testin toplamda 30 maddeden oluşursa güvenirlik katsayısının 0,647 olması beklenmektedir. Hemen alttaki "Reliability of new data" (Yeni verinin güvenirliği) kısmına ise 0,80 yazılırsa, hemen altında 66 madde olarak hesaplama yapılır. Yani 0,80 güvenirlik katsayısına ulaşılması bekleniyorsa, testtekilere benzer 46 madde daha teste eklenmeli ve toplam 66 maddelik bir test oluşturulmalıdır. *Split-half method* (iki yarıya bölme yöntemi) menüsünde ise iki yarıya bölme yöntemine göre güvenirlik katsayısı hesaplanmaktadır. Bu menüde ilk yarı – son yarı, tek – çift, rastgele, ortalama ve en kötü seçenekleri bulunmaktadır. Ayrıca 10000 farklı şekilde test iki yarıya bölünerek her birinin sonucundan bir histogram oluşturularak raporlanmaktadır. *Cronbach alfa* menüsünde ise teste ait Cronbach alfa katsayısı hesaplanmaktadır. 1 – 0 puanlanan maddelerde Cronbach alfa, aynı veri seti için hesaplanacak KR-20 katsayısına eşit olacağından sayfada KR-20 için ayrı bir menüye yer verilmemiştir. Güvenirlik ve geçerlik kanıtları, faktör analizinin yapılması gibi konular bu bölümün kapsamı dışında kaldığından ilgili menüler ile ilgili açıklamalar bu kadarla sınırlandırılmıştır.

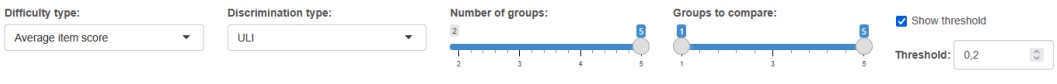
4.4. Madde Analizi

SIA ile madde analizleri temelde *Item Analysis* (madde analizi) menüsünden yapılmaktadır. Bunun yanında *DIF/Fairness* (değişen madde fonksiyonu / yanlılık) menüsünden maddelerin farklı gruplarda benzer çalışıp çalışmadığı incelenebilmektedir.

Item Analysis menüsü altındaki ilk seçenek *Traditional Item Analysis'dir* (geleneksel madde analizi). Bu menü altında KTK temelli madde güçlüğü ve ayırt ediciliğine ilişkin istatistikler elde edilebilmektedir. Şekil 14'te yer alan seçenekler aracılığıyla Şekil 15'te verilen grafiğe ilişkin ayarlamalar yapılabilmektedir.

Şekil 14

Madde Analizi Seçenek Alanı



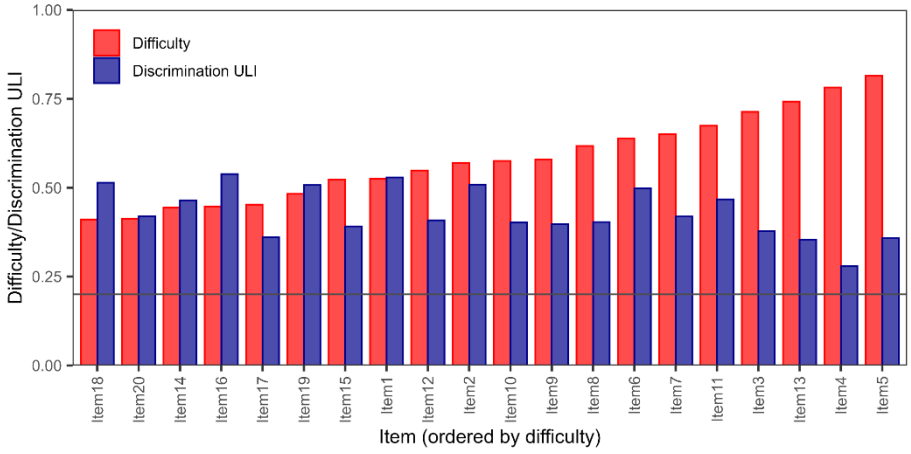
Difficulty type (güçlük tipi) kısmında iki seçenek bulunmaktadır: *Average scaled score*, madde puan ortalamasının madde ranjına bölünmesiyle elde edilen ölçeklenmiş puanı verirken *Average item score* maddenin ortalama puanını vermektedir. Çok kategorili puanlanan maddeler (örneğin, açık uçlu maddeler) için average scaled score 0 - 1 aralığında madde güçlüğünü sunmaktadır. İki kategorili puanlanan maddeler içinse her iki seçenek aynı sonucu vermektedir.

Discrimination type (ayırt edicilik tipi) kısmında ise dört seçenek yer almaktadır. *ULI: Upper-Lower Index*, alt-üst gruplar yöntemi ile elde edilen indekstir. *RIT*, madde – toplam korelasyonunu; *RIR* ise madde – kalan korelasyonunu verir. *None* seçeneği seçildiğinde ise madde ayırt ediciliği raporlanmaz. *Number of groups* ve *groups to compare* seçenekleri alt – üst gruplar yöntemi seçildiğinde aktif hale gelir. *Number of groups*, verinin kaç gruba bölüneceğini; *groups to compare* ise ayırt edicilik indeksinin hangi iki grubun karşılaştırılmasına göre hesaplanacağını belirler. Dolayısıyla her ne kadar bu indeks alt – üst gruplar yöntemi olarak anılsa da örneklem beş gruba bölünüp maddenin dördüncü ve beşinci grubu birbirinden ayırıp ayırmadığı incelenebilmektedir.

Son olarak Şekil 15'teki grafiğe yatayda bir kesme çizgisi eklemek için *Show threshold* seçeneği işaretlenerek alttaki kutucuktan kesme çizgisinin hangi değerde olması gerektiği belirtilir. Tüm bu ayarlamalardan sonra Şekil 15'teki gibi bir grafik elde edilebilir.

Şekil 15

Madde Analizi Sonuçları – I



Şekildeki grafik maddeleri güçlüklerine göre (kırmızı renkli çubuklar) sıralamakta, aynı zamanda ayırt edicilik indeksini de (mor çubuklar) göstermektedir. Bu grafiğin altında ise Şekil 16'daki gibi bir tablo yer almaktadır.

Şekil 16

Madde Analizi Sonuçları – II

	Diff.	Avg. score	SD	Min.	Obs. min.	Max.	Obs. max.	Prop. max.	RIT	RIR	I-C cor.	ULI	Rel.	Val.	α-drop	Missed [%]	Not-reached [%]
Item1	0.52	0.52	0.50	0	0	1	1	0.52	0.38	0.23	0.09	0.43	0.19	0.05	0.53	0.00	0.00
Item2	0.57	0.57	0.50	0	0	1	1	0.57	0.37	0.22	0.09	0.42	0.18	0.04	0.53	0.00	0.00
Item3	0.71	0.71	0.45	0	0	1	1	0.71	0.27	0.13	0.09	0.29	0.12	0.04	0.55	0.00	0.00
Item4	0.78	0.78	0.41	0	0	1	1	0.78	0.25	0.12	0.12	0.25	0.10	0.05	0.55	0.00	0.00
Item5	0.81	0.81	0.39	0	0	1	1	0.81	0.33	0.21	0.10	0.29	0.13	0.04	0.54	0.00	0.00

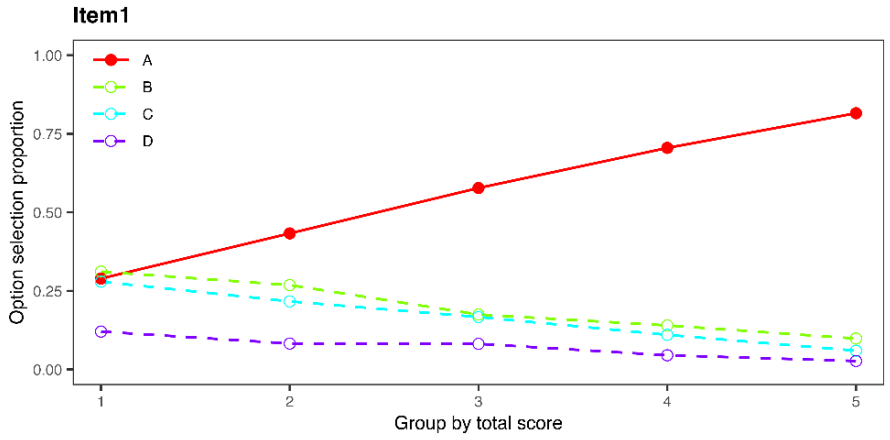
Bu tablodaki *Diff.* madde güçlüğünü; *Avg. score* ortalama puanı temsil etmektedir. 1-0 puanlanan maddeler için *Diff.* ve *Avg. score* sütunları aynı değere sahip olur. *SD*, maddenin standart sapmasını; *min.* o maddeden alınabilecek en düşük puanı; *Obs. min*, o maddeden alınan en düşük puanı; *max.* alınabilecek en yüksek puanı; *Obs. max* alınan en yüksek puanı göstermektedir. *Prop. max.* ise maddeden en yüksek puanı alan kişilerin oranını göstermektedir. 1-0 puanlanan maddeler için *Prop. max.* madde güçlüğünü bildirmektedir.

RIT, madde - toplam korelasyonunu; *RIR* ise madde – kalan korelasyonunu göstermektedir. *I-C cor.* ise madde ile ölçüt puan arasındaki korelasyonu, *ULI*: Alt - üst gruplar yöntemine göre ayırt edicilik indeksini göstermektedir. *Rel.* madde güvenirliğini; *Val.* ise madde geçerlik indeksini vermektedir. *α-drop*, madde çıkarılırsa Cronbach alfa katsayısının nasıl değişeceğini, *Missed [%]* kayıp değer yüzdesini ve son olarak *Non-reached [%]* ise bu maddeye ulaşamayan öğrencilerin yüzdesini temsil etmektedir.

Yine *Item Analysis* menüsü altında yer alan *Item Response Curves* (madde tepki eğrileri) yolu izlenirse testi alan bireylerin toplam puanlarına göre üç gruba ayrıldığı ve her seçeneğin bu gruplarda hangi oranda seçildiği bir grafik görüntülenir. Grup sayısı ayarından, grup sayısı üç yerine beş yapıldığında Şekil 17'deki gibi bir grafik elde edilir.

Şekil 17

Seçeneklerin Analizi



Bu grafikte çeldiriciler kesikli çizgi, anahtar yanıt (doğru cevap) ise düz çizgi ile temsil edilmektedir. Grafik incelenerek çeldiricilerin ve doğru yanıtın farklı gruplarda seçilme oranları görülebilmektedir. Buna göre grup daha başarılı hale geldikçe çeldiricilerin seçilme oranı azalmakta ve anahtar yanıtın seçilme oranı artmaktadır. Her seçenek öğrenciler tarafından belli oranlarda işaretlenmiş olup çalışmayan çeldirici yoktur. Yalnızca D çeldiricisi en başarısız grup tarafından bile diğer çeldiricilere göre daha az tercih edilmiştir.

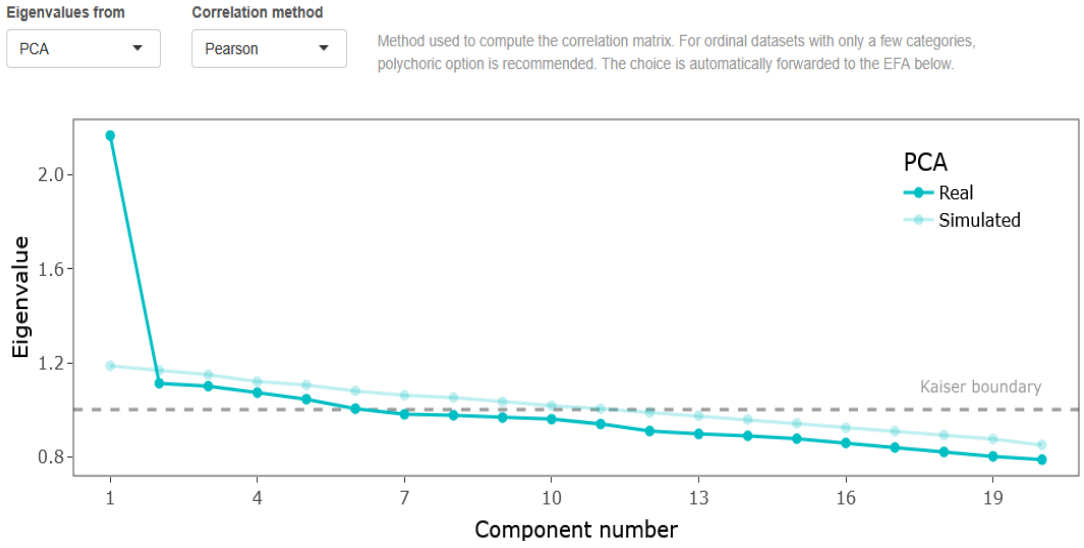
4.5. MTK Analizleri

Varsayımların Kontrolü

Tek boyutlu MTK modelleri ile çalışıldığında ilk varsayım maddelerin tek bir başat faktörü ölçüyor olmasıdır. Bunu kontrol etmek için faktör analizinden yararlanılabilir. Faktör analizi bölümün kapsamı dışında kaldığından burada yalnızca yamaç birikinti grafiği kullanılarak tek boyutluluk incelenecektir. Bunun için *Validity* menüsü altındaki *Factor analysis* yolu kullanılacaktır. Şekil 18'deki gibi açılan sayfada varsayılan olarak *Temel Bileşenler Analizi (Principal Component Analysis-PCA)* ve maddeler arası korelasyon katsayısı için Pearson momentler çarpım korelasyon katsayısı seçilidir.

Şekil 18

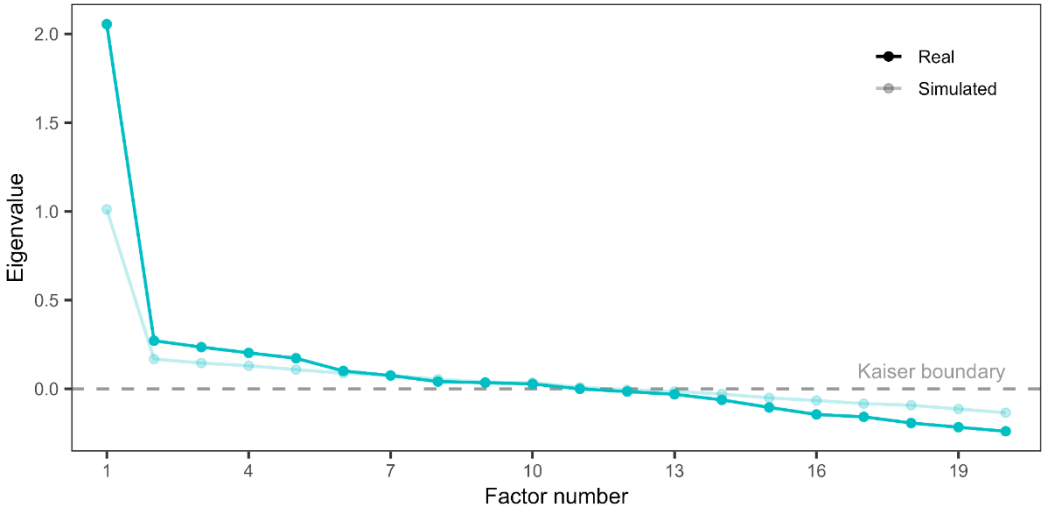
Temel Bileşen Analizine Göre Yamaç-Birikinti Grafiği



Ancak psikolojik ölçmelerde, maddeler ile altta yatan gizil bir yapı ortaya konulmaya çalışıldığından temel bileşen analizi yerine faktör analizinin yapılması tercih edilir. Ayrıca örnekteki veri 1-0 şeklinde puanlandığı için korelasyon yöntemi olarak tetrakorik korelasyon katsayısı kullanılmalıdır. Bu amaçla Eigenvalues from (özdeğerler) ve Correlation method (korelasyon yöntemi) kutucuklarında sırasıyla FA ve Polychoric ifadeleri seçilmelidir. Bu seçimlerin ardından yamaç birikinti grafiği otomatik olarak Şekil 19'daki gibi güncellenir.

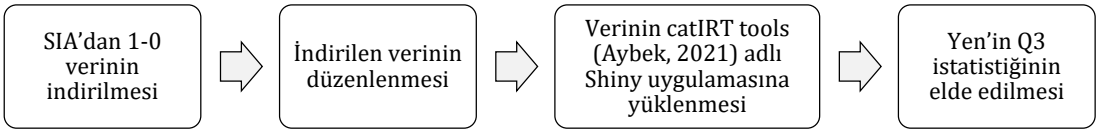
Şekil 19

Faktör Analizine Göre Yamaç-Birikinti Grafiği



Bu yamaç birikinti grafiğinde Real (gerçek) ile gösterilen, gerçek veriden elde edilen öz değerler iken Simulated (simülatif) ile gösterilen paralel analiz sonucudur. Her ikisine göre de yirmi maddenin tek bir başat faktör altında toplanabildiği anlaşıldığından tek boyutluluk varsayımının sağlandığı söylenebilir. İlgisi olan okurlar bu grafiğin altındaki tabloları inceleyerek bütün faktör analizini tamamlayabilirler.

Her ne kadar tek boyutluluk varsayımı karşılandığında yerel bağımsızlığın sağlandığı yönünde karar verilebileceğini belirten kaynaklar olsa da yerel bağımsızlığın testinde Yen'in Q3 istatistiği sıklıkla kullanılmaktadır. Ne yazık ki SIA, henüz Yen'in Q3 istatistiğini vermemektedir. Bu nedenle Yen'in Q3 istatistiği farklı bir Shiny uygulaması ile incelenecektir. Yen'in Q3 istatistiğini elde etmek için sırasıyla aşağıdaki adımlar izlenecektir:



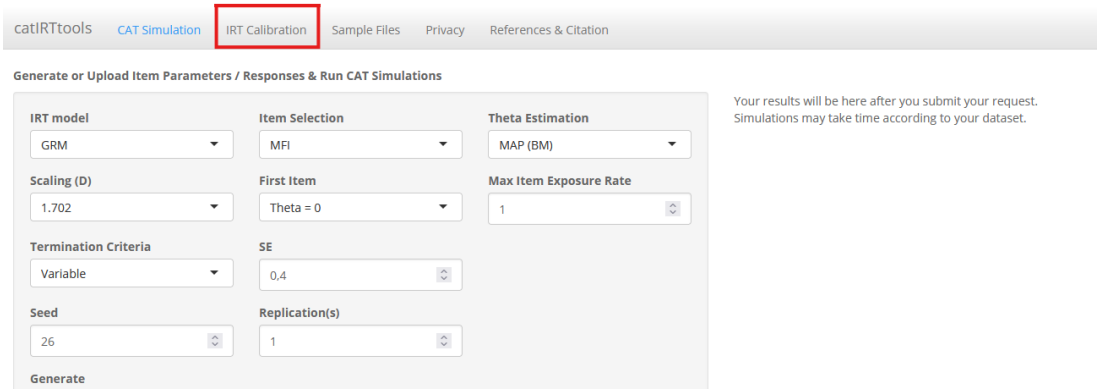
Yen'in Q3 istatistiği için başvurulacak olan catIRT tools uygulamasına https://eptlab.com/catirt adresinden çevrimiçi erişim sağlanabilir ya da https://github.com/ecaybek/catIRTtools adresinde yer alan kaynak kodlar indirilerek uygulamanın kullanıcı bilgisayarında da çalıştırılması mümkündür.

Öncelikle SIA'da Data (veri) sekmesinden Data exploration (veriyi keşfet) alt sekmesine tıklanmalıdır. Ardından Scored / binarized data (puanlanmış veri) başlığı altındaki Download scored dataset (puanlanmış veri setini indir) butonuna tıklanır. Bu butona tıklandığında 1-0'a dönüştürülmüş veri .csv formatında bilgisayara indirilir.

catIRT tools uygulaması açıldığında Şekil 20'deki gibi bir ekran görüntüsü ile karşılaşılacaktır.

Şekil 20

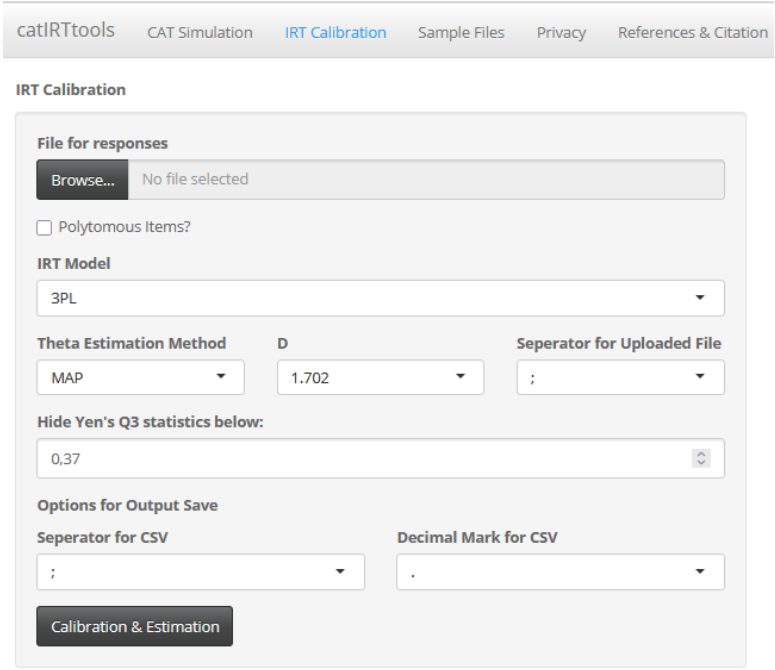
catIRT Tools Ana Ekranı



Uygulamanın ana sayfası, kullanıcıların Bilgisayar Ortamında Bireye Uyarlanmış Test simülasyonu yapmalarına olanak sağlamaktadır. Yen'in Q3 istatistiği için Şekil 20'de kırmızı ile gösterilen IRT Calibration (MTK Kalibrasyonu) sekmesine gidilmelidir. Bu sekmeye ait ekran görüntüsü Şekil 21'de verilmiştir.

Şekil 21

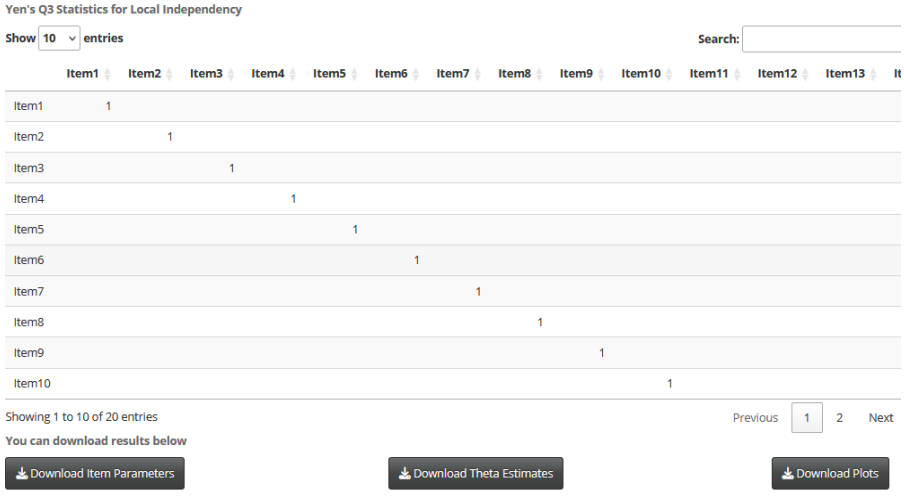
MTK Kalibrasyon Sekmesi



SIA'dan indirilen csv dosyası File for responses (Yanıtlar için dosya) alanına atanmalıdır. Örnekteki dosya 1-0 puanlanan bir veri olduğu için Polytomous items (çok kategorili puanlanan madde) kısmına herhangi bir tik atılmamalıdır. IRT Model kısmından kullanılması planlanan MTK modeline ilişkin seçim yapılmalıdır. Theta estimation method (Teta kestirim yöntemi) kısmından kullanılması planlanan yetenek kestirim yöntemi; D kısmından ölçekleme sabiti (1 ya da 1,702) ve yüklenen csv dosyasında sütunların hangi işaretle ayrıldığı (Seperator for Uploaded File) seçilmelidir. SIA'dan indirilen csv dosyası, sütun ayracı olarak virgül (,) kullandığı için bu bölümde virgül seçilmelidir. Son olarak Hide Yen's Q3 statistics below (Aşağıdakinden küçük Q3 istatistiklerini gizle) kısmına Yen'in Q3 istatistiği için arzu edilen kesme değeri girilmelidir. Bu değer girilmezse tüm madde çiftleri için Yen'in Q3 istatistiği raporlanacağından tabloyu yorumlamak güçleşir. Ölçüt olarak genellikle .20 kullanıldığından bu kısma 0,20 değeri girilebilir. Daha sonra Calibration & Estimation (Kalibre et & kestirim yap) butonuna tıklanır. Sonuçların elde edilmesi verinin büyüklüğüne göre zaman alıcı olabilir. Sonuçlar elde edildiğinde Şekil 22'deki gibi bir tablo ve butonlar görünür.

Şekil 22

Yen'in Q3 İstatistikleri



İlk gösterilen tabloda madde çiftleri arasındaki Yen'in Q3 istatistikleri raporlanmaktadır. Analiz öncesinde 0,20 değerinin altındaki Q3 istatistiklerinin gösterilmemesi istendiği için madde çiftleri arasında herhangi bir değer görülmemektedir. Her maddenin kendisiyle olan Q3 istatistiği ise 1 olarak raporlanmaktadır, ancak bu dikkate alınmamalıdır. Ayrıca en alttaki butonlardan sırasıyla madde parametreleri, bireylere ait yetenek kestirimleri ve madde karakteristik eğrileri indirilebilir. Ancak SIA, MTK bağlamında daha çok özellik sunduğu için bu uygulamadan yalnızca yerel bağımsızlık kontrolü için yararlanılmıştır. Sonuç olarak hiçbir madde çiftinin yerel bağımsızlığı bozmadığına karar verilmiş ve MTK analizlerine SIA üzerinden devam edilmiştir.

Model Karşılaştırma

Verinin hangi MTK modeli ile en yüksek uyumu gösterdiğini belirlemek için SIA'da IRT models (MTK modelleri) sekmesi altındaki Model comparison (model karşılaştırma) yolu izlenir. Açılan sayfada Şekil 23'te yer alan karşılaştırma tablosu incelenir.

Şekil 23

MTK Modeli Karşılaştırma Sonuçları

Table of comparison statistics

Row **BEST** indicates which model has the lowest value of given information criterion.

	AIC	BIC	logLik
1PL	51254.424	51372.043	-25606.212
2PL	51192.115	51416.151	-25556.057
3PL	51200.297	51536.351	-25540.148
4PL	51232.136	51680.209	-25536.068
BEST	**2PL**	**1PL**	

Bu tabloda Akaike bilgi ölçütü (Akaike Information Criterion - AIC), Bayes bilgi ölçütü (Bayesian Information Criterion - BIC) ve logLikelihood değerleri raporlanmaktadır. Ayrıca AIC ve BIC ölçütlerine göre veri ile en yüksek uyumu gösteren model BEST satırında gösterilmektedir. Örnekteki bulgulara göre AIC'e göre 2 PLM; BIC'e göre 1 PLM veri ile en yüksek uyumu göstermektedir. Bu noktada hangi modelle analize devam edileceği test geliştiricisine bağlıdır. Test geliştiricisi aşağıdaki yollardan biri ile ilerlemeyi seçebilir:

1. Madde ayırt edicilik parametrelerini önemli görerek 2PLM ile devam edebilir
2. Daha az sayıda parametreyi değişken tutarak daha genellenebilir bulgular elde etmek için 1PLM ile devam edebilir
3. Her ikisine göre de analizleri yaparak en uygun modele daha sonra karar verebilir.

Uygulamada hem *a* hem de *b* parametrelerinin nasıl değiştiğinin gözlenebilmesi için bu bölümde 2PLM ile devam edilmiştir. Ancak yukarıdaki gerekçelerle 1PLM ile de devam edilmesi yanlış değildir. İki parametreli lojistik modele göre madde kalibrasyonlarının yapılabilmesi için IRT models sekmesi altındaki Dichotomous models (İki kategorili puanlanan modeller) yolu izlenir. Varsayılan olarak Summary (özet) alt sekmesi açılacaktır. Bu sayfada Şekil 24'teki gibi iki adet seçenek sunulmaktadır.

Şekil 24

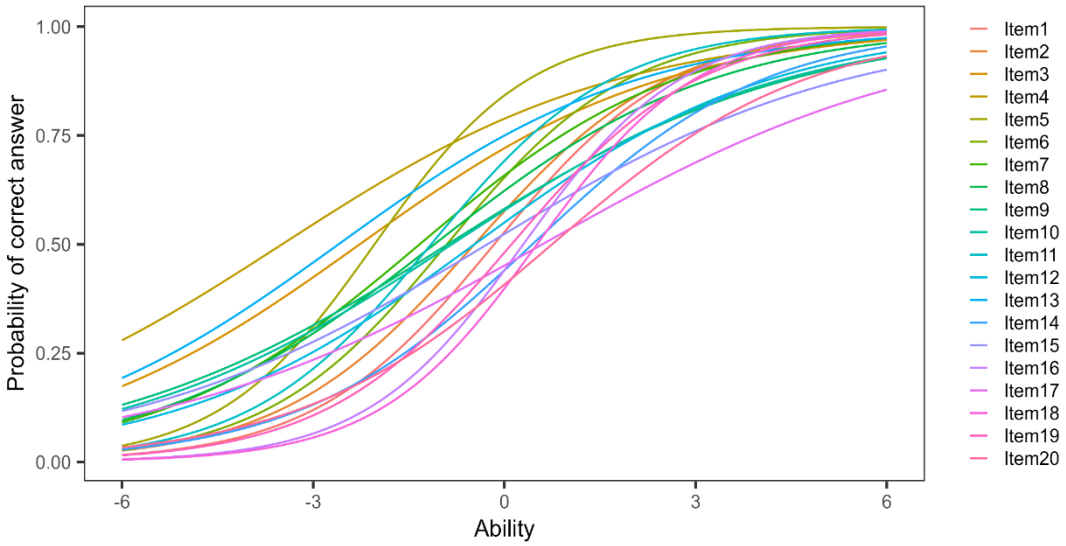
MTK Modeli Seçim Alanı



Model seçeneğinden, maddelerin ölçeklenmesinin istendiği MTK modeli seçilmelidir. Örnek uygulama için bu kısım 2PL olarak seçilmiştir. Parametrization kısmı ise IRT olarak kalmalıdır. Bu sayfa bir özet sayfası olduğu için tüm maddelere ait madde karakteristik eğrileri tek bir grafik üzerinde yer almaktadır. Benzer şekilde tüm maddeler için madde bilgi fonksiyonları da tek bir grafikte sunulmaktadır. Ayrıca test bilgi fonksiyonu ve standart hata değeri de bir grafikte verilmektedir. Tüm bu grafikler sırasıyla Şekil 25 ile Şekil 27 arasında sunulmuştur.

Şekil 25

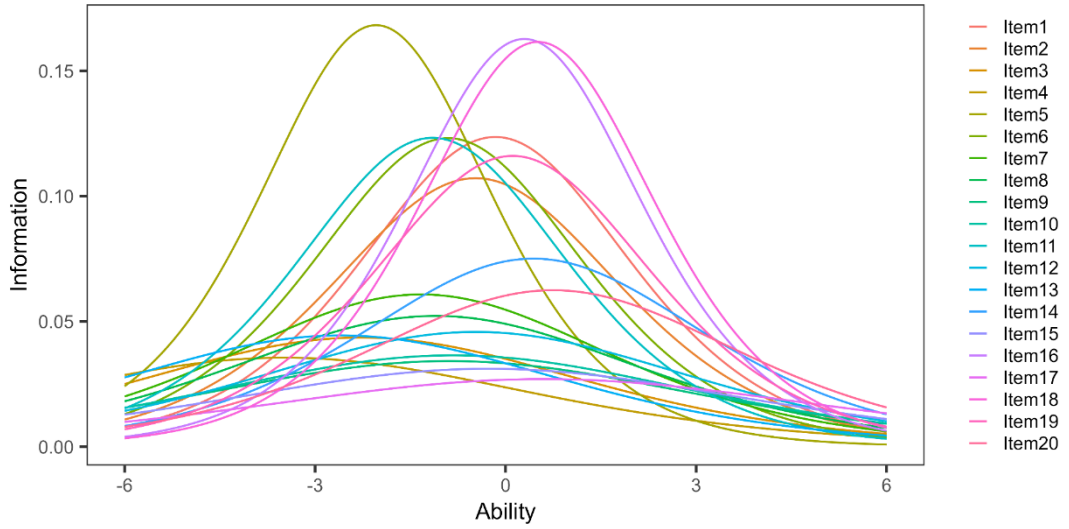
Tüm Maddelere Ait Madde Karakteristik Eğrileri



Bu grafikte her bir maddeye ait madde karakteristik eğrileri yer almaktadır. 2PLM kullanıldığı için tüm maddelerin *a* parametreleri birbirinden farklıdır, bu nedenle her bir eğri farklı eğime sahiptir. Grafiğin x ekseni bireylere ait yetenek düzeyini; y ekseni ise buna karşılık gelen doğru yanıtlama olasılığını temsil etmektedir.

Şekil 26

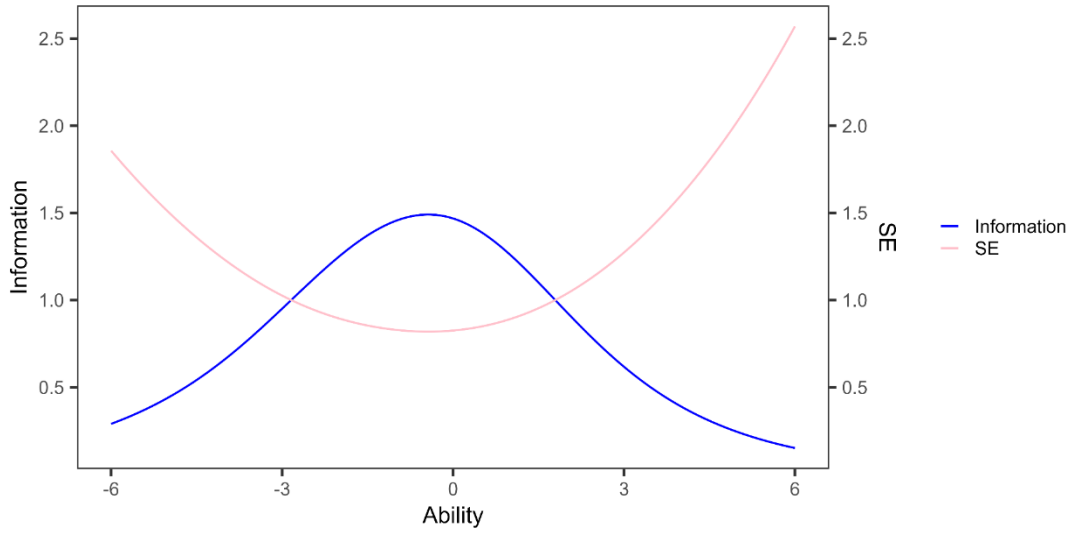
Tüm Maddelere Ait Madde Bilgi Fonksiyonları



Bu grafikte ise maddelere ilişkin bilgi fonksiyonları yer almaktadır. Grafiğin x ekseni yetenek düzeyini gösterirken y ekseni buna karşılık gelen bilgi miktarını temsil etmektedir. Maddenin belli bir yetenek düzeyi için sağladığı bilgi miktarı, maddeye ait *a* parametresinin karesi ile doğrudan ilişkilidir. Haliyle, *a* parametresi küçük olan maddelerin sağladığı bilgi miktarı da düşük olmaktadır. Bu durum Şekil 26'da açık bir şekilde görülmektedir (basık olan çizgiler).

Şekil 27

Test Bilgi ve Standart Hata Fonksiyonları



Bu grafikte ise mavi eğim test bilgi fonksiyonunu yansıtmaktadır. Test bilgi fonksiyonuna bakılarak x ekseni üzerindeki farklı yetenek düzeyleri için bu testteki tüm maddeler kullanıldığında ulaşılacak bilgi miktarının ne olduğu belirlenebilmektedir. Görüldüğü üzere test en çok orta yetenek düzeyindeki bireyler için bilgi sağlarken uçlardaki bireyler (çok başarılı ya da çok başarısız) için daha az bilgi sunmaktadır. Pembe eğri ise standart hatayı temsil etmektedir. Grafikte görüldüğü gibi standart hata, bilgi ile ters orantılıdır. Bu grafiklerin ardından kestirilen madde parametreleri ve madde – model uyum istatistiklerinin yer aldığı bir tablo sunulmaktadır. Bu tabloya ait ekran görüntüsü Şekil 28'de verilmiştir.

Şekil 28

Madde Parametreleri

	a	SE(a)	b	SE(b)	c	SE(c)	d	SE(d)	SX2-value	df	p-value
Item1	0.70	0.08	-0.16	0.07	0.00	NA	1.00	NA	6.54	13.00	0.92
Item2	0.65	0.07	-0.47	0.09	0.00	NA	1.00	NA	16.96	13.00	0.20
Item3	0.42	0.07	-2.27	0.38	0.00	NA	1.00	NA	29.22	14.00	0.01
Item4	0.38	0.08	-3.49	0.68	0.00	NA	1.00	NA	11.37	14.00	0.66
Item5	0.82	0.10	-2.04	0.21	0.00	NA	1.00	NA	14.79	12.00	0.25

Görüldüğü gibi Şekil 28; *a*, *b*, *c*, *d* parametreleri ve bu parametrelere ait standart hata değerleri ile S_χ^2 istatistiklerini içermektedir. 2PLM ile ölçekleme yapıldığından tüm maddeler için *c* parametresi 0 ve *d* parametresi 1 olarak atanmıştır. Öte yandan madde - model uyumunu gösteren S_χ^2 istatistiklerine ait *p* değerleri incelenip model ile uyum göstermeyen maddeler tespit edilmelidir. Bunun için ölçüt kabul edilecek alfa düzeyine test geliştiricisi karar verir. Bu uygulamada I. tip hatayı azaltmak amacıyla alfa düzeyi 0,01 olarak seçilmiştir. Buna göre tüm maddelerin model - veri uyumunu sağladığı anlaşılmıştır.



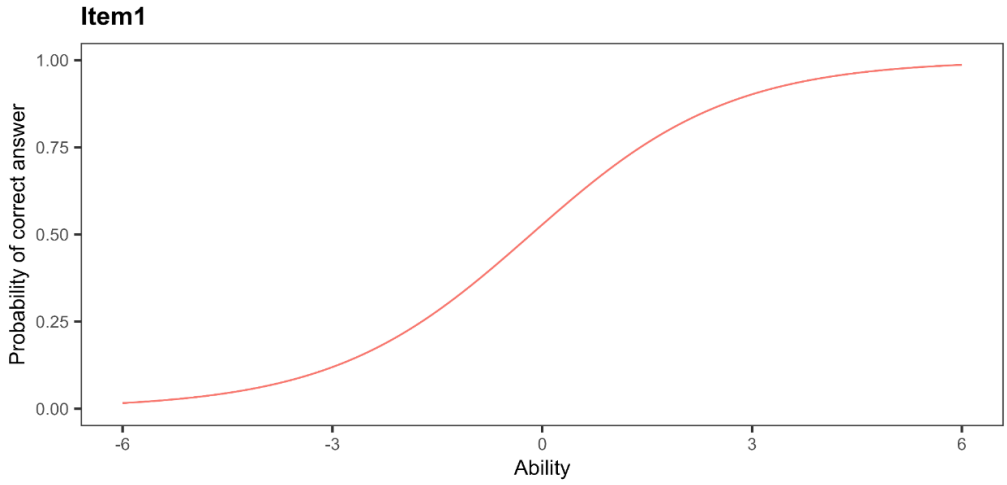
Eğer alfa düzeyi 0,05 seçilirse, 3 numaralı maddenin testten çıkarılması gerekir.

Madde parametrelerinin ardından her bir kişiye ait toplam ham puan, *z* ve T-puanları ile MTK'ye göre yetenek kestirimlerinin yer aldığı bir tablo özeti sunulmaktadır. Alttaki Download abilities (yetenekleri indir) butonu ile testi alan tüm bireylere ait puanlar indirilebilmektedir. Son olarak standartlaştırılmış *z*-puanları ile MTK yetenek kestirimleri arasındaki korelasyon katsayısı sunulmaktadır. Örnek uygulama için *z*-puanları ile teta yetenek kestirimleri arasındaki Pearson momentler çarpım korelasyon katsayısı 0,982 olarak raporlanmıştır.

Item (madde) alt sekmesinde ise her bir madde için ayrı ayrı madde karakteristik eğrileri, madde bilgi fonksiyonları ve madde parametreleri verilmektedir. Örnek olarak Madde 1 için madde karakteristik eğrisi ve madde bilgi fonksiyonu sırasıyla Şekil 29 ve Şekil 30'da sunulmuştur.

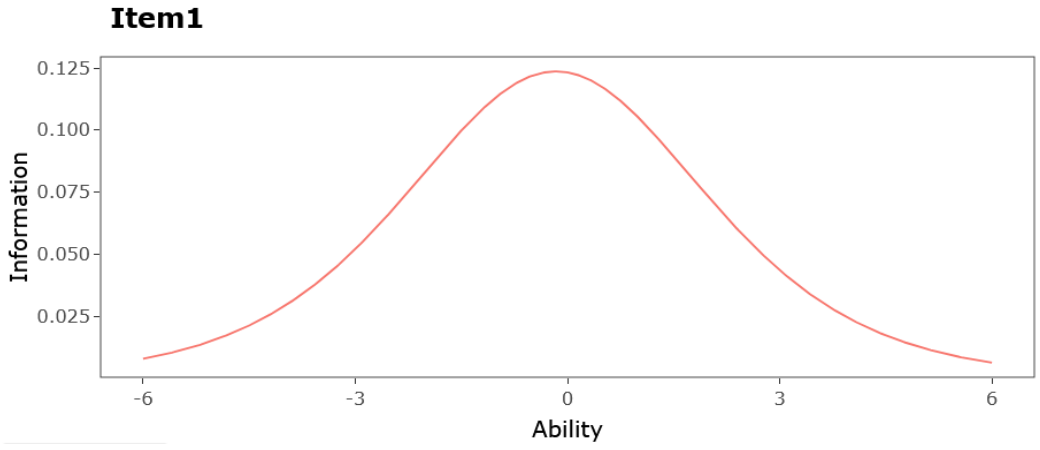
Şekil 29

Madde 1 için Madde Karakteristik Eğrisi



Şekil 30

Madde 1 için Madde Bilgi Fonksiyonu



Görüldüğü üzere bu defa tek bir madde için madde karakteristik eğrisi ve madde bilgi fonksiyonu sunulmaktadır. Özellikle madde havuzunda çok sayıda madde olduğunda tüm maddeleri aynı grafik üzerinde incelemek güçleşir. Bu durumda her maddeyi ayrı ayrı incelemek test geliştiricisine kolaylık sağlar.

5. SONUÇ

Bu bölüm ölçme sonuçları üzerinde yapılan istatistiksel işlemlerden test ve madde istatistiklerine odaklanmış hem KTK hem de MTK bağlamında yapılabilecek analizler Shiny Item Analysis ile örnek bir uygulama üzerinden gösterilmiştir. Test istatistikleri başlığı altında ele alınan merkezî eğilim ve yayılım ölçüleri birer betimsel istatistik olup ölçme işleminden elde edilen puanlar ile yapılabilecek hipotez testleri (T-testleri, varyans analizleri, regresyon, vb.) bölümün kapsamı dışında tutulmuştur. Dileyen okuyucular temel istatistik üzerine yazılmış kaynaklar aracılığıyla bu analizlere ilişkin okuma yapabilirler.

KAYNAKÇA | REFERENCES

Arıcı, H. (2006). *İstatistik: Yöntemler ve uygulamalar*. Meteksan.

Arıkan, S. & Aybek, E. C. (2022). A special case of Brennan's index for tests that aim to select a limited number of students: A Monte Carlo simulation study. *Educational Measurement: Issues and Practice, 41*, 35-49. https://doi.org/10.1111/emip.12528

Aybek, E. C. (2021). catIRT tools: A "Shiny" application for item response theory calibration and computerized adaptive testing simulation. *Journal of Applied Testing Technology, 22*(1). 12-24.

Baker, F. (2001). *The basics of item response theory*. ERIC Clearinghouse.

Baker, F. B. & Kim, S. H. (2017). *The basics of item response theory using R*. Springer.

Brennan, R. L. (1972). A generalized upper-lower item discrimination index. *Educational and Psychological Measurement, 32*(2), 289–303. https://doi.org/10.1177/001316447203200206

Crocker, L. & Algina, J. (2008). *Introduction to classical and modern test theory*. Cengage Learning.

Erkuş, A. (2003). *Psikometri üzerine yazılar*. Türk Psikologlar Derneği.

Erkuş, A. (2012). *Psikolojide ölçme ve ölçek geliştirme - I: Temel kavramlar ve işlemler*. Pegem.

Evans, C. H. & Ildstad, S. T. (2001). *Small clinical trials: Issues and challenges*. National Academy.

Gronlund, N. E. (1977). *Constructing achievement tests*. Prentice-Hall.

Gulliksen, H. (1950). *Theory of mental tests*. John Wiley and Sons.

Haladyna, T. M. & Downing, S. M. (1993). How many options is enough for a multiple-choice test item? *Educational and Psychological Measurement, 53,* 999-1010. https://doi.org/10.1177/0013164493053004013

Hambleton, R. & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Springer.

Kaptan, S. (1998). *Bilimsel araştırma ve istatistik teknikleri*. Tekışık Ofset.

Kelley, T. L. (1939). The selection of upper and lower groups for the validation of test items. *Journal of Educational Psychology, 30*(1), 17-24. https://doi.org/10.1037/h0057123

Köklü, N., Büyüköztürk, Ş. & Çokluk-Bökeoğlu, Ö. (2007). *Sosyal bilimler için istatistik*. Pegem.

Lord, F. M. (1952). The relation of the reliability of multiple-choice tests to the distribution of item difficulties. *Psychometrika, 17*(2), 181-194. https://doi.org/10.1007/BF02288781

Martinkova, P. & Hladka, A. (2023). *Computational aspects of psychometric methods*. Chapman and Hall/CRC. https://doi.org/10.1201/9781003054313

Martinkova, P. & Drabinova, A. (2018). ShinyItemAnalysis for teaching psychometrics and to enforce routine analysis of educational tests. *The R Journal, 10*(2), 503-515. https://doi.org/10.32614/RJ-2018-074

Metsamuuronen, J. (2020). Somers' D as an alternative for the item–test and item-rest correlation coefficients in the educational measurement settings. *International Journal of Educational Methodology, 6*(1), 207–221. https://doi.org/10.12973/ijem.6.1.207

Murphy, K. R. & Davidshofer, C. O. (2014). *Psychological testing: Principles and applications* (6. baskı). Pearson.

Popham, W. J. (2014). *Classroom assessment: What teachers need to know*. Pearson.

Thompson, N. (2020). *What is classical item difficulty (p value)?. https://assess.com/classical-item-difficulty-p-value/*

Turgut, M. F. (1997). *Eğitimde ölçme ve değerlendirme metotları.* Yargıcı.

Turgut, M. F. & Baykul, Y. (2010). *Eğitimde ölçme ve değerlendirme*. Pegem.

Woolley, T. W. (1998). A glossary of common statistical terms. *Journal of Pharmacy Practice, 11*(3). 219-225. https://doi.org/10.1177/089719009801100311

YAZAR HAKKINDA | ABOUT AUTHOR

DOÇ. DR. EREN CAN AYBEK

Pamukkale Üniversitesi | eaybek[at]pau.edu.tr

ORCID: 0000-0003-3040-2337

Eren Can Aybek lisans derecesini 2009 yılında Balıkesir Üniversitesi Fizik Öğretmenliği bölümünden almıştır. Yüksek lisans ve doktora eğitimini ise Ankara Üniversitesi Ölçme ve Değerlendirme bölümünde tamamlamıştır. 2017 yılından beri Pamukkale Üniversitesi Eğitim Fakültesi Eğitimde Ölçme ve Değerlendirme Anabilim Dalında öğretim üyesi olarak görevini sürdürmektedir. Aynı zamanda 2021 yılında Boğaziçi Üniversitesi Bireye Uyarlanmış Test Laboratuvarında araştırmacıdır. Çalışma alanları çoğunlukla Bilgisayar Ortamında Bireye Uyarlanmış Testler (BOBUT) ve Madde Tepki Kuramını (MTK) içermektedir. Ayrıca otomatik puanlama, BOBUT simülasyonu ve faktör analizine yönelik çeşitli psikometrik yazılımlar geliştirmiştir.

ASSOC. PROF. DR. EREN CAN AYBEK

Pamukkale University | eaybek[at]pau.edu.tr

ORCID: 0000-0003-3040-2337

Eren Can Aybek received his bachelor's degree in Physics Education from Balıkesir University in 2009. He completed both his master's and doctoral studies in Measurement and Evaluation at Ankara University. Since 2017, he has been serving as a faculty member in the Department of Educational Measurement and Evaluation at the Faculty of Education, Pamukkale University. Since 2021, he has also been a researcher at the Adaptive Testing Laboratory at Boğaziçi University. His research interests primarily include Computerized Adaptive Testing (CAT) and Item Response Theory (IRT). He has also developed several psychometric software applications for automated scoring, CAT simulation, and factor analysis.

12. BÖLÜM

DEĞERLENDİRME VE YERLEŞTİRME

ASSESSMENT AND PLACEMENT

Dr. Öğr. Üyesi Sungur GÜREL

Dr. Öğr. Üyesi Muhammet YORULMAZ

DEĞERLENDİRME VE YERLEŞTİRME

Dr. Öğr. Üyesi Sungur GÜREL
Siirt Üniversitesi

Dr. Öğr. Üyesi Muhammet YORULMAZ
Başkent Üniversitesi

Öz

Bu bölümde, eğitimde ölçme sonuçlarının yargıya dönüştürüldüğü "değerlendirme" sürecini ve adayların belirli kontenjanlara sahip programlara atanmasını sağlayan "yerleştirme" mekanizmaları iki ayrı başlık altında incelenmektedir. İlk başlıkta, değerlendirme türleri kapsamlı bir şekilde sınıflandırılmaktadır. Amacına göre değerlendirmeler; öğrencilerin hazır bulunuşluğunu ölçen tanıma-yerleştirmeye (ör. lise ve üniversiteye geçiş sınavları), süreci iyileştiren ve öğrenciye geri bildirim sağlayan biçimlendirmeye (formative) ve öğretim süreci sonunda not vermeyi amaçlayan değer biçmeye (summative) yönelik olarak üçe ayrılmıştır. Ölçütüne göre ise öğrencinin önceden belirlenmiş sabit bir başarı sınırını aşmasına dayanan mutlak değerlendirme ile öğrencinin başarısını sınıf veya grup ortalamasıyla kıyaslayan bağıl (normatif) değerlendirme ele alınmıştır. Ayrıca, eğitim politikalarına ve hesap verebilirliğe yön veren geniş ölçekli değerlendirmeler (PISA, vb.) ile bireysel gelişimi destekleyen sınıf içi uygulamaların birbirini tamamlayan unsurlar olduğu vurgulanır. İkinci başlıkta, ekonomi bilimindeki "eşleştirme teorisine" dayanan öğrenci yerleştirme problemi ele alınmaktadır. Öğrencilerin sınav puanı, puan türü, tercih sıralamalarına ve kontenjana göre adil ve sistematik bir şekilde yerleştirilebilmesi için gerekli temel politikalar açıklanmıştır. Örnek yerleştirme algoritması olan "Ertelenmiş Kabul" algoritması senaryolar üzerinden adım adım detaylandırılarak anlatılmaktadır.

Anahtar Kelimeler

Amacına Göre Değerlendirme, Ölçütüne Göre Değerlendirme, Geniş Ölçekli Değerlendirme, Öğrenci Yerleştirme, Ertelenmiş Kabul Algoritması

ASSESSMENT AND PLACEMENT

Asst. Prof. Dr. Sungur GÜREL
Siirt Üniversitesi

Asst. Prof. Dr. Muhammet YORULMAZ
Başkent Üniversitesi

Abstract

In this chapter, the "assessment" process—wherein educational measurement results are transformed into judgments—and the "placement" mechanisms, which facilitate the assignment of candidates to programs with specific quotas, are examined under two distinct headings. Under the first heading, types of assessments are comprehensively classified. Based on their purpose, assessments are divided into three categories: placement assessments that measure students' readiness (e.g., high school and university entrance exams), formative assessments that enhance the learning process and provide feedback to students, and summative assessments aimed at assigning grades at the end of the instructional process. Based on the evaluation criteria, the text discusses criterion referenced assessment, which relies on a student exceeding a predetermined and fixed proficiency threshold, and normative assessment, which compares a student's performance against a class or group average. Furthermore, it is emphasized that large-scale assessments (such as PISA), which guide educational policies and ensure accountability, and classroom-based practices, which support individual development, are complementary elements. Under the second heading, the student placement problem, grounded in "matching theory" within the field of economics, is addressed. The foundational policies required for the fair and systematic allocation of students based on their exam scores, score types, preference rankings, and institutional quotas are elucidated. Finally, the "Deferred Acceptance" algorithm, serving as an exemplary placement algorithm, is explained in detail and step-by-step through practical scenarios

Keywords

Assessment Purpose, Evaluation Criteria, Large-Scale Assessment, Student Placement, Deferred-acceptance Algorithm

1. GİRİŞ

Ölçme ve değerlendirme sürecinin nihai adımında, elde edilen ölçme sonuçları bir ölçüt ya da ölçüt takımı ile karşılaştırılarak yargıya varılır. Bu aşamada alınan kararlar, özellikle yerleştirme amacı güden yüksek riskli sınavlarda, adaylar hakkında son derece kritik ve bağlayıcı olabilmektedir. Bireyler hakkında geçerli, güvenilir ve adil kararlar alabilmek için değerlendirmede kullanılacak ölçüt ve yöntemler ile değerlendirme sürecini izleyen yerleştirme kararlarında esas alınacak yaklaşımlar doğru ve tutarlı biçimde belirlenmelidir.

Kitabın bu bölümünde öncelikle değerlendirme terimi derinlemesine ele alınmış; amacına ve kullanılan ölçüte göre değerlendirme türlerinden bahsedilmiştir. Sonrasında ulusal ve uluslararası düzeyde yürütülen geniş ölçekli değerlendirmeler üzerinde durulmuştur. Bölüm yerleştirme başlığı ile sonlandırılmıştır.

2. DEĞERLENDİRME

Değerlendirme, en yalın haliyle ölçme sonuçlarını bir ölçütle karşılaştırarak bir yargıya ulaşma (Yıldırım, 1966) anlamına gelir. Ancak Erkuş (2016) bunu değerleme olarak isimlendirmiş, değerlendirmeyi ise bir kurumun/sistemin hedeflerine ulaşıp ulaşmadığı; bir bireyin ya da grubun süreç içinde nasıl bir gelişim gösterdiği; aynı süreçte çeşitli gruplar arasında farklılıklar olup olmadığı ve varsa hangi açılardan farklılıklar bulunduğu gibi irdelemeleri içeren geniş bir kavram olarak tanımlamıştır. Bu bölümde değerlendirme (assessment) terimi, öğrenci davranışlarına odaklanarak ölçme işlemi ile başlayan ve ölçümleri bir ölçüt ile karşılaştırarak yargıya ulaşma biçiminde kavramsallaştırılmıştır.

Eğitimde yapılan değerlendirmeler; öğrenci başarısının belirlenmesi, öğrencilerin yetenek ve ilgilerine uygun alanlara yöneltilmesi, öğrenme eksiklerinin saptanması, öğretim programının, öğretimin, ölçme araçlarının ve ölçütlerin değerlendirilmesi gibi farklı amaçlar taşır (Turgut & Baykul, 2019). Değerlendirme öğrenciye durum bildirme, öğrenciyi güdüleme, öğretmene öğrenciler hakkında bilgi verme, öğretmenin kendisini değerlendirmesine yardımcı olma, idarecilere ve ebeveynlere öğrenci hakkında bilgi verme gibi rollere sahiptir (Turgut, 1969).

Binbaşıoğlu (1983) ve Tekin (1979) değerlendirmeyi amacına göre *(i)* tanıma-yerleştirmeye, *(ii)* biçimlendirme-yetiştirmeye ve *(iii)* değer biçmeye yönelik değerlendirme olmak üzere üç gruba ayırır. Değerlendirme, kullanılan ölçüte göre ise *(i)* mutlak ve *(ii)* bağıl değerlendirme şeklinde sınıflandırılır. Burada geniş ölçekli değerlendirme ve yerleştirme başlıklarına geçmeden önce kitabın birinci bölümünde yer verilen bu başlıklar hatırlatılacaktır.

Amacına Göre Değerlendirme

Öğretim sürecinde değerlendirmeler farklı amaçlarla gerçekleştirilebilir. Binbaşıoğlu (1983) ve Tekin'in (1979) anlayışıyla hareket edilirse;

Tanıma-Yerleştirmeye Yönelik Değerlendirme

Bu aşamada öğretim sürecinin başında öğrencinin ilgili öğretim sürecinden faydalanmak için gereken önkoşul öğrenmelere sahip olup olmadığının belirlenmesi amaçlanır (Tekin, 1979). Yani yapılan değerlendirme, öğrencinin öğretim süreci ile kazandırılması hedeflenen davranışları edinebilecek hazır bulunuşluk düzeyinde olup olmadığını tespit etme gayesi taşır. Örneğin bir yabancı dil kursuna yazılan bir birey öncelikle dil yeterlik düzeyinin saptanması için bir sınava alınır, bu sınavın sonuçlarına göre birey en yüksek faydayı sağlayacak gruba yerleştirilir. Kademeler arası geçişe temel olan Liselere Geçiş Sınavı (LGS) ile Yükseköğretim Kurumları Sınavı (YKS) ülkemizde geniş ölçekli tanıma-yerleştirmeye yönelik değerlendirme örnekleri arasında yer alır.

Türkiye Yüzyılı Maarif Modeli (TYMM) (MEB, 2025) ile öğretim programının içeriğinde önemli bir yer alan ön değerlendirme süreci ise tanıma-yerleştirmeye yönelik bir sınıf içi değerlendirme örneğidir. TYMM'de ön değerlendirme süreçleriyle hedeflenen öğrencinin bir öğrenme birimi için gereken ön bilgilere, programda yer alan ifadesiyle temel kabullere, ne düzeyde sahip olduğunun belirlenmesi amaçlanmaktadır. Bu adımda öğrencinin hangi bilgi ve becerilere sahip olduğunun yanı sıra ilgi ve ihtiyaçları da tespit edilir. Bu bağlamda tanıma-yerleştirmeye yönelik değerlendirmede öğrencinin sadece bilişsel özellikleri test edilmez, duyuşsal ve psikomotor özellikleri de sınanır. Bu değerlendirmelerden elde edilen çıktılar, öğrenci ihtiyacına en uygun öğrenme ortamını ve sürecini oluşturmada kılavuz niteliği taşır.

Biçimlendirme-Yetiştirmeye Yönelik Değerlendirme (Formative Assessment)

Öğretim sürecinin bir parçası olarak görülür ve temelde öğrenme eksikliklerinin tespit edilmesi ya da öğretimin etkililiğinin belirlenmesi amaçlanır. Bu değerlendirme ile öğrencinin öğretim süreci içerisinde kazandırılması öngörülen hedef davranışların ne kadarına sahip olduğu sürekli olarak kontrol edilir. Böylece öğrencilerin daha iyi öğrenebilmeleri ve öğrenme eksiklerini giderebilmeleri için onlara geri bildirim verilebilir (Tekin, 1979). Biçimlendirici değerlendirme; sınavın haricinde bünyesinde pek çok bileşen barındıran, öğrencinin de sürece dâhil olduğu, öğretim süreci içerisinde herhangi bir zamanda gerçekleştirilebilen, öğrencilere geri bildirim sağlayan ve öğretime yönelik düzenlemeler içeren bir süreçtir (McMillan, 2018). Cizek (2010, s. 8) biçimlendirici değerlendirmenin 10 niteliğini şu şekilde sıralar:

Şekil 1

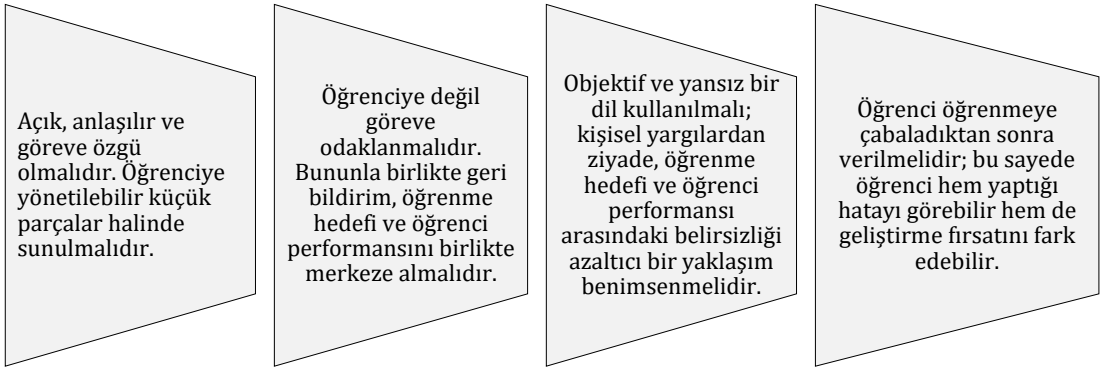
Biçimlendirici Değerlendirmenin Temel Nitelikleri

Öğrencilerin kendi öğrenimlerinden sorumlu olmalarını gerektirir.
Açık ve belirgin öğrenme hedeflerini iletir.
Öğrenim bağlamının ötesinde, uygulanabilirliği olan değerli eğitim çıktılarına odaklanır.
Öğrencinin mevcut bilgi/becerilerini ve istenen hedeflere ulaşmak için gerekli adımları belirler.
İstenen hedeflere ulaşmak için planlar geliştirilmesini gerektirir.
Öğrencileri, öğrenme hedeflerine yönelik olarak kendi ilerlemelerini izlemeye teşvik eder.
Öğrencinin çalışmalarını değerlendirmede kullanılacak ölçütleri veya puanlama anahtarları dâhil olmak üzere, öğrenme hedeflerine ilişkin örnekler sunar.
Akran ve öğrenci öz değerlendirme ile öğrenme etkinliklerine entegre edilmiş değerlendirmeler dahil olmak üzere sık sık değerlendirme yapmayı gerektirir.
Değerlendirme; öğrenme hedefleriyle ilgili, yargılayıcı olmayan, özelleştirilmiş ve zamanında sunulan geri bildirimler içerir.
Öğrenciye çalışmalarını gözden geçirme, iyileştirme ve anlayışını derinleştirme fırsatları sunar.
Öğrencilerin çalışmaları üzerinde öğrenmeyi öğrenme ve yansıtma yapmalarını teşvik eder.

Biçimlendirici değerlendirme döngüsü içinde öğrenci sürekli olarak "(a) Nereye gidiyorum, (b) Şu an neredeyim ve (v) Hangi stratejiler hedeflerime ulaşmama yardımcı olur?" sorularına cevap arar (Moss & Brookhart, 2019): Bu sorulardan ilki öğretimsel hedefi vurgular. İkincisi, öğrencinin mevcut durumu hakkında bilgi sağlar. Üçüncüsü ise öğretimsel hedef ile mevcut durumu arasındaki mesafeyi nasıl kısaltacağına dair öğrenciyi aydınlatır. Bu soruların cevabı öğrenciye geri bildirim olarak ulaştırılır. Shute (2008) etkili bir geri bildirimde aranan özelliklere dayanarak geri bildirim verirken dikkat edilmesi gereken unsurları Şekil 2'deki gibi özetler:

Şekil 2

Geri Bildirimde Bulunurken Dikkat Edilmesi Gereken Unsurlar



TYMM (MEB, 2025) ile öğretim programının öğrenme-öğretme yaşantıları arasında sunulan öğrenme kanıtları (ölçme ve değerlendirme) bölümü, biçimlendirici değerlendirme uygulamalarını da içermektedir. TYMM, öğrencilerin süreç içindeki öğrenme düzeylerinin takip edilmesi için biçimlendirici değerlendirme uygulamalarını merkeze almış ve onlara etkili geri bildirim verilmesini önermiştir. Öğretmenleri, öğretim programlarında sunulan biçimlendirici değerlendirme uygulamalarının yanı sıra kendi belirleyecekleri uygulamaları da kullanmaları yönünde teşvik etmiştir. Sınıf içi biçimlendirici değerlendirme uygulamalarında öğrencinin mevcut durumunu belirlemek amacıyla soru-cevap uygulamalarından, kısa sınavlardan, izleme testlerinden, yapılandırılmış sınıf içi gözlemlerden, ev ödevlerinden, çıkış biletlerinden, öğrenci ürün dosyalarından, performans görevlerinden ve öz değerlendirmelerden faydalanılabilir (İnce Aracı, 2023).

Sınıfta biçimlendirici değerlendirme uygulamalarını gerçekleştirmek pek çok zorluğu da beraberinde getirmektedir. Biçimlendirici değerlendirme, kısa sınavlar ve ödevlerin sayısını artırmaktan ibaret kalırsa etkisi azalır. Biçimlendirici değerlendirme gücünü yargılayıcı olmayan, zamanında verilen, düzeltmeye olanak sağlayan etkin geri bildirimlerden alır. Geçerliği ve güvenirliği yüksek puanlar üreten biçimlendirici değerlendirme uygulamalarının geliştirilmesi, uygulanması ve uygulama sonuçlarına dayalı geri bildirim sunulması önemli ölçüde zaman, emek ve maddi kaynak gerektirir (Cizek, 2010).

Değer Biçmeye Yönelik Değerlendirme

Öğrencinin öngörülen hedef davranışların ne kadarına sahip olduğunun belirlenmesi amacıyla genellikle öğretim sürecinin sonucunda yapılır. Değer biçmeye yönelik değerlendirme ile öğrenci, öğretmen ve program hakkında yargıda bulunulur (Tekin, 1979). Öğrenci hakkındaki kararlar not verme ile gerçekleştirilir ve öğrenci başarısı

özetlenir. Öğretim sürecinin sonunda gerçekleştirildiği için bu yaklaşıma sonuç odaklı değerlendirme (summative assessment) ismi de verilir.

Ölçütüne Göre Değerlendirme

Verilen kararın referans noktasına ölçüt ismi verilir. Referans noktası olarak kullanılan ölçüt önceden belirlenmiş ve sabit bir değer ise mutlak değerlendirme, grubun başarı düzeyi ile karşılaştırılarak saptanan bir değer ise bağıl değerlendirme söz konusu olmaktadır.

Mutlak Değerlendirme

Değerlendirme işleminde değeri önceden kesin olarak belirlenmiş, bu değerin grubun başarısından etkilenmediği bir ölçüt kullanılıyorsa yapılan değerlendirme mutlak değerlendirme olarak isimlendirilir (Tekin, 1979). Kriter dayanıklı olarak da adlandırılan bu değerlendirme türünde, öğrenciler sınava girmeden önce kaç puan alırlarsa başarılı sayılacaklarını net bir şekilde bilirler. Öğrenci hakkında verilecek karar, diğer öğrencilerin başarısından bağımsızdır. Karar, öğrencinin programdaki asgari yeterlikleri sağlayıp sağlamadığına ya da önceden belirlenen sınır değeri aşıp aşmadığına dayalı olarak verilir. Millî Eğitim Bakanlığı (MEB) Okul Öncesi Eğitim ve İlköğretim Kurumları Yönetmeliği'ne (MEB, 2014) göre ortaokul ve imam-hatip ortaokulunda yıl sonu başarı puanı 100 tam puan üzerinden belirlenir. Türkçe dışındaki dersler için 0-49.99 puanlar başarısız, 50.00 ve üzeri puanlar başarılı olarak değerlendirilir. Türkçe dersinde ise 0-69.99 puanlar başarısız, 70.00 ve üzeri puanlar başarılı sayılır. Yani ortaokullarda öğrenci başarısının belirlenmesi amacıyla Türkçe dışındaki dersler için 50 puan, Türkçe dersi için 70 puan mutlak ölçüt olarak kullanılmaktadır. MEB Ortaöğretim Kurumları Yönetmeliği'ne (MEB, 2023) göre ortaöğretimde de öğrenci başarısını değerlendirmede benzer bir ölçüt kullanılmaktadır. Ortaöğretim kurumlarında başarı için kullanılan mutlak ölçüt 50 puandır. 100 üzerinden 0-49.99 arasında puan alan öğrenciler "geçmez" olarak değerlendirilir ve başarısız kabul edilir. Ayrıca ortaöğretim kurumlarında öğrenci, puanına göre derecelendirilir. 50.00-59.99 geçer, 60.00-69.99 orta, 70.00-84.99 iyi, 85.00-100 pekiyi derecesini alır.

Mutlak değerlendirme sadece başarılı/başarısız kararını vermek için gerçekleştirilmez. Bir yetkinlik belirtmek amacıyla gerçekleştirilen sertifikalandırma sınavlarında da mutlak değerlendirme tercih edilmelidir. ÖSYM tarafından yürütülen Elektronik İngilizce Yeterlik Sınavı (Electronic-Test of English Proficiency [e-TEP]) buna örnek olarak verilebilir. E-TEP adayların okuma, dinleme, konuşma ve yazma alanlarında İngilizce yeterliklerini "Diller için Ortak Avrupa Başvuru Metni" ile uyumlu bir şekilde ölçen ve B1, B2 ve C1 seviyelerinde sertifikalandıran bir sınavdır. Her bir bölümde 30 puan olmak üzere toplam 120 puan üzerinden değerlendirilen bu sınavda 98 puan ve üzeri alan adaylar C1, 74-98 arası puan alan adaylar B2 ve 50-74 arası puan alan adaylar B1 seviyesinde sertifikalandırılmaktadır (ÖSYM, 2025). Aday, sınava girmeden önce kaç puan alırsa hangi seviyede sertifikalandırılacağını bilir ve bu durum sınava giren diğer adayların başarısından bağımsızdır.

Bağıl Değerlendirme

Norm dayanıklı değerlendirme olarak da ifade edilen bu değerlendirme türünde öğrencinin başarısı, üyesi olduğu sınıfın başarısı ile karşılaştırılır (Tekin, 1979). Bu şekilde öğrencinin başarısının diğer öğrencilere kıyasla hangi seviyede olduğu belirlenir. Kullanılacak bağıl ölçüt, normal dağılım eğrisi altında hesaplanan yüzdelikler ya da standart puan olabilir. Söz gelimi 5 kategorili bir değerlendirme yapıldığında sınıftaki en başarılı %10'luk öğrenci en yüksek kategoride, devamındaki %20'lik öğrenci bir alt kategoride, bir sonraki %40'lık öğrenci orta kategoride, arkasından gelen %20'lik öğrenci ortanın altı kategorisinde ve en başarısız %10'luk öğrenci en alt kategoride değerlendirilebilir (Oyar & Gündüz, 2023). Ancak sınıftaki öğrenci başarısı normal dağılım göstermiyorsa yüzdeliklere göre yapılan değerlendirme yanıltıcı olabilir (Tekin, 1979). Ayrıca bu anlayış sınıf ortalamasının hangi seviyede olduğunu göz ardı eder. Standart puana dayalı olarak bağıl değerlendirme yapıldığında ise öğrenci notları öncelikle T ya da Z puanı gibi standart bir puana dönüştürülür. Sonrasında ilgili standart puana karşılık gelen değerlendirme yapılır. Hatta sınıfların başarı ortalaması farkını da göz önüne alarak aynı standart puan daha başarılı sınıflarda daha başarılı olarak değerlendirilebilir (Akın Arıkan, 2021). Günümüzde pek çok üniversite ders geçme sisteminde hem sınıf başarı ortalamasını hem de standart puanı temel alarak bağıl değerlendirme yapmaktadır (Karagöz & Alpaydın, 2024). Bununla birlikte, bağıl değerlendirme yapmanın öğrenciler arası rekabeti tetikleme potansiyeli olduğu unutulmamalıdır.



Karagöz ve Alpaydın (2024) Türkiye'deki 205 yükseköğretim kurumunun değerlendirme sistemini incelemiş ve 202'sinin değerlendirme sistemi kullandığını raporlamıştır. Bu raporda üniversitelerin 66'sının sadece mutlak, 39'u sadece bağıl değerlendirme yaptığını, 97'sinin ise hem mutlak hem bağıl değerlendirme sistemlerini kullandığını saptamıştır.

Geniş Ölçekli Değerlendirme

Bölümde bu noktaya kadar sınıf içi ölçmelere dayalı olarak yapılan değerlendirmeler merkeze alınmıştır. Ancak günümüzde ulusal ya da uluslararası düzeyde yürütülen ölçmelere dayalı olarak da çeşitli değerlendirmeler gerçekleştirilmektedir. International Association for Evaluation Educational Achievement (IEA) ve Organisation for Economic Co-operation and Development (OECD) organizasyonları IEA Trends in International Mathematics and Science Study (TIMSS), IEA Progress in International Reading Literacy Study (PIRLS), IEA International Civic and Citizenship Education Study (ICCS), IEA International Computer and Information Literacy Study (ICILS) ve OECD Programme for International Student Assessment (PISA) gibi geniş ölçekli değerlendirmeleri

uluslararası düzeyde yürütmektedir (Rocher & Hastedt, 2020). Türkiye'deki öğrenciler de MEB'in desteğiyle TIMSS, PIRLS ve PISA çalışmalarına katılmakta, ayrıca MEB Akademik Becerilerin İzlenmesi ve Değerlendirilmesi (ABİDE) programı ile 2016'dan günümüze değin iki yıllık dönemler halinde ulusal düzeyde bir geniş ölçekli değerlendirme yürütmektedir. Bahsi geçen bu uygulamaların detayına kitabın 9. bölümünde yer verilmiştir.

Geniş ölçekli değerlendirmeler, politika yapıcılar ve diğer eğitim paydaşlarının eğitim reformu için kanıta dayalı karar verme süreçlerinde kullandıkları en önemli araçlardan biridir (Rocher & Hastedt, 2020). Geniş ölçekli değerlendirmeler genellikle sistem düzeyindeki eğitim amaçlarına hizmet eder. Bu amaçlar arasında eğitimin kalitesini izlemek, hesap verebilirliği sağlamak, politika kararları almak, ulusal veya uluslararası düzeyde öğrenci performansını karşılaştırmak sayılabilir (Linn, 2000; OECD, 2019). Geniş ölçekli değerlendirmelere temel olacak ölçme araçları büyük bir titizlikle hazırlanır. Bu ölçme araçlarıyla gruplar arası ve zamana dayalı karşılaştırmaların yapılabilmesi önceliklendirilir ve yüksek güvenirlik ile geçerlikte ölçümler elde edilmesi hedeflenir. Yüksek güvenirlik ve tutarlık sağlamak amacıyla geniş ölçekli sınavlarda yaygın olarak objektif olarak puanlanabilen çoktan seçmeli maddeler ile standartlaştırılmış ölçekler aracılığıyla puanlanabilen yanıtı sınırlandırılmış maddeler tercih edilir (Wu, 2010). Yine geçerlik ve güvenirliği artırmak; test güvenliğini, adaleti ve uygulama birliğini sağlamak amacıyla geniş ölçekli değerlendirmelere ilişkin ölçme uygulamaları kontrolün yüksek olduğu standart koşullar altında gerçekleştirilir. Geniş ölçekli değerlendirme sonuçları öğrenciler hakkında bireysel sonuçlar üretebilmekle birlikte, daha çok sistem düzeyinde sonuçlar ortaya koyar. Bu nedenle, geniş ölçekli değerlendirmelerde ulaşılan sonuçlar okul sistemlerini değerlendirmek, öğretim programı revizyonlarına ışık tutmak, eğitim politikalarını şekillendirmek ve uluslararası düzeyde karşılaştırmalar yapmak gibi amaçlarla kullanılabilir (Linn, 2000).

Sınıf içi ve geniş ölçekli değerlendirmeler birbirinin tamamlayıcısı olarak ele alınmalıdır. Sınıf içi değerlendirmeler bireysel öğrencinin gelişiminin desteklenmesini merkeze alırken geniş ölçekli değerlendirmelerin odağında eğitim politikaları vardır. Yapısı gereği geniş ölçekli değerlendirmeler öğrenciler hakkında sınıf içi değerlendirmeler kadar kapsamlı bilgi sunmaz. Öte yandan geniş ölçekli değerlendirmeler farklı örneklem grupları arasındaki yeterlik farklılıklarını, öğrenci başarısındaki güçlü ve zayıf yönlerin genel anlamda neler olduğunu yansıtabilir (Pellegrino ve ark., 2001). Geniş ölçekli değerlendirme sonuçlarının eğitim politikasını şekillendirmesiyle sınıf içi öncelikler de şekilleneceği için bu iki değerlendirme karşılıklı olarak birbirini besler denilebilir. Ancak geniş ölçekli değerlendirmeler büyük önem taşıdığında sınıf içi öğretimin geniş ölçekli değerlendirme format ve içeriğini yansıtacak şekilde daralması ve sınava yönelik öğretim yapılması ciddi bir endişe kaynağıdır (NRC, 2011).

3. YERLEŞTİRME

Yerleştirme araştırmaları, ekonomi bilimi alanında uygulamalı mekanizmalar tasarlama çabaları sonucunda ortaya çıkan eşleştirme teorisine (matching theory) dayanmaktadır. Eşleştirme teorisinde, eşleştirme kavramı nakledilecek organlar, lojmanlar, yazlık evler gibi bölünemez nesnelerin tahsisi ya da değişimi anlamına gelebileceği gibi iş arayan kişiler-işçi arayan firmalar, öğrenciler-okullar, evlenmek için eş arayan kadınlar-erkekler gibi iki taraflı bir araya getirme işlemlerini de ifade eder (Sönmez & Ünver, 2011). Yerleştirme alanındaki öncü çalışmalardan biri günümüze kadar birçok araştırmaya ilham vermiş olan Gale ve Shapley'e aittir. Araştırmacılar geliştirdikleri tekrarlı bir algoritma sayesinde kararlı bir eşleştirme yapılabileceğini göstermişlerdir (Gale & Shapley, 1962). Geliştirilen algoritma uzun yıllar Amerika Birleşik Devletleri'nde intörn doktorların hastanelere eşleştirmesinde kullanılmıştır (Roth, 1984). Bu çalışmalar sonradan gelen araştırmacıları etkilemiş ve günümüze kadar birçok çalışma yapılmasına öncülük etmiştir. Görüldüğü gibi basit görünen yerleştirme işleminin arkasında uzun yıllara dayanan köklü bir bilimsel literatür bulunmaktadır. Alanyazındaki bilimsel çalışmalar teorik düzeyde kalmamış pratik olarak da çeşitli ülkelerde üniversite ve lise yerleştirmeleri ile organ nakli gibi gerçek hayattaki eşleştirme problemlerinde de uygulamaya konulmuştur. Gerçek hayat problemlerine uygulanan çözümlere öncülük ettikleri için Lloyd Shapley ile Alvin Roth, 2012 yılında İsveç Kraliyet Akademisi tarafından ekonomi alanında Nobel ile ödüllendirilmişlerdir (Nobel Prize, 2012).

Öğrenci yerleştirme (student placement) problemi, öğrencilerin merkezi bir otorite tarafından belirli kontenjanlara sahip okullara (veya üniversite bölümlerine / programlarına) atanmasını sağlayan bir eşleştirme (matching) modelidir (Sönmez & Ünver, 2011). ÖSYM ile ilgili hâlihazırda yürürlükte olan mevzuattaki 6114 sayılı kanunda "yerleştirme" tanımı şu şekilde yapılmıştır (ÖSYM Hizmetleri Hakkında Kanun, 2011). "Değerlendirme sonucunda oluşan puanlar, tercihler, önceden belirlenen kontenjan ve şartlar dikkate alınarak adayların yükseköğretim kurumlarına veya ilgili kurum ve kuruluşlara yerleştirilme işlemlerini ifade eder". Yani sınavlar yapılmış, puanlar hesaplanmış, program kontenjanları ile giriş özel koşulları önceden belirlenmiş ve adaylardan girmek istedikleri programların tercih sıraları alınmıştır. Bu işlemlerden sonra yerleştirme işlemine başlanır. Aslında ilk bakışta yerleştirme işlemi basit bir problem gibi görünebilir. Gerçekten de tek bir kritere göre yerleştirme yapılmak isteniyorsa yerleştirme problemi kolayca çözülebilir. MEB'in ilköğretime yeni başlayacak öğrencileri otomatik olarak evlerine en yakın okullara kaydetmesi bu duruma örnek olarak verilebilir. Okula yeni başlayacak öğrencilerin kaydı ulusal veri tabanındaki güncel adres bilgileri doğrultusunda okulların fiziki kapasitesi ve ulaşım imkânları da dikkate alınarak otomatik şekilde yapılır (MEB, 2014). MEB'in yaptığı

ilköğretime kayıt işlemi de öğrencileri uygun okullara "yerleştirme" işlemidir. Tek kriter, öğrenci ikametinin okul kayıt alanı içinde bulunup bulunmadığı olunca yerleştirme işlemi kolayca yapılabilmektedir. Fakat okulların yerleştirmeye esas puan türleri, öğrenci puanları, tercihleri ve çeşitli öncelikler işin içine girdiğinde yerleştirme işlemi karmaşık bir hâl almakta ve çözülmesi gereken bir probleme dönüşmektedir. Bu yönüyle yerleştirme günümüzde, bilim insanlarının ve araştırmacıların ilgisini çeken bir olgudur.

Nobel ile ödüllendirilecek kadar önemli olan ve geçmişi uzun yıllara dayanan yerleştirme alanyazının tamamını bir kitap bölümüne sığdırmak mümkün değildir. Bu yüzden kitabın bu bölümünde temel amaç, yerleştirme işleminin tanıtımını yapıp temel çalışma prensiplerini açıklamaktır. İlgili alandan olmayan okuyucuların da yerleştirme işlemlerini kolayca anlayabilmeleri için yerleştirmede takip edilen süreçler ile adımlar mümkün olduğu kadar basit bir dille anlatılmaya çalışılmıştır. Bölümü bitiren okuyucuların yerleştirme işlemlerinin temel çalışma prensiplerini açıklayabilmesi hedeflenmiştir.

3.1. Temel Tanımlar

Yerleştirme işlemlerinde sıklıkla kullanılan tanımlar aşağıda açıklanmaya çalışılmıştır. Tanımlar yerleştirme işlemlerindeki temel politikaları da açıklamaktadır. Temel politikalar, gerçek yaşamdaki yerleştirme işlemleri için kritik öneme sahiptir.

Kararlılık (Stability)

İki taraflı eşleşme modelinde bulunan uygun çözüm için kullanılır. Bir eşleştirme işlemi sonunda, mevcut eşleşmelere, alternatif olarak daha iyi bir eşleşme varsa, çözüm "kararsız" olarak nitelendirilir. Böyle bir durum oluşmamışsa çözüm kararlı kabul edilir (Gale & Shapley, 1962). Örneğin, üniversite programlarına yerleştirme işlemleri sonucunda bir adayın tercihinde daha üst sıralardaki bir programa yerleştirilmek yerine daha alt tercihlerinden bir programa yerleştirildiği varsayılsın. Adayın yerleştirilmediği, daha çok istediği programda da kontenjanın tam dolmadığını ya da daha az puanlı bir adayın yerleştiği durum söz konusu olsun. Bu durumda hem aday hem de program için mevcut eşleştirmeden daha iyi bir eşleşme vardır ve çözüm kararsız olarak nitelendirilir.

Adillik (Fairness)

Yerleştirme işlemlerinin en önemli özelliğidir. Daha yüksek puana sahip aday tercihine yerleşirken daha düşük puanlı bir adayın onun yerine geçememesi, yerleştirmenin adil olduğu anlamına gelir (Balinski & Sönmez, 1999). Örneğin ÖSYM yerleştirmeleri adil yerleşim açısından şeffaf ve güvenilirdir; her yerleştirme sonunda programların taban puanları ilan edilmektedir.

Pareto verimliliği (Pareto Efficiency)

Yerleştirme işlemlerinin doğal politika hedeflerinden biridir. Eşleştirme sürecinde yer alan taraflar yani katılımcılar, literatürde ajan olarak tanımlanır. Ekonomik ajanlar ise refahları dikkate alınan taraflardır. Pareto verimliliği, bulunan çözümde bir ekonomik ajanın refahını başkasının refahına zarar vermeden iyileştirmenin mümkün olmadığını ifade eder (Abdulkadiroglu & Andersson, 2022). Örneğin üniversite yerleştirmesinde bir adayı yerleştiği programdan daha çok istediği, tercihlerinde daha üst sıraya yazdığı başka bir programa yerleştirme imkânının olup olamadığı araştırılmak isteniyor olsun. İşte burada sözü edilen değişikliği diğer herhangi bir adayı yerleştiğinden daha alt sıralarda tercih ettiği bir programa yerleştirmeden yapmanın imkânsız olduğu çözümler Pareto verimli kabul edilir.

Stratejik Dayanıklılık (Strategy-Proofness)

Yerleştirme işlemlerinde, tercih beyan eden katılımcılar stratejik ajan olarak tanımlanır. Stratejik dayanıklılık, bu tercihlerin manipülasyondan uzak gerçekten dürüstçe beyan edilmesi anlamına gelir. Bir yerleştirme işlemi, adayların dürüstçe beyan ettikleri gerçek tercihlerinin, kendileri için en iyi sonucu (maksimum faydayı) doğuracağını düşündürüyorsa stratejik olarak dayanıklıdır (Abdulkadiroglu & Andersson, 2022).

3.2. Yerleştirme Algoritması

Literatürde öncü olan ve en çok bilinen yerleştirme algoritması yukarıda bahsedilen Gale-Shapley tarafından geliştirilmiş olan ertelenmiş kabul (deferred-acceptance) algoritmasıdır. Gale-Shapley çalışmalarında algoritmayı kolay anlatabilmek için evlilik eşleşmesi örneğinden yararlanmışlardır. Geliştirdikleri algoritma sayesinde bir grup erkek ve bir grup kadın kümesi arasında kararlı bir evlilik eşleşmesinin her zaman bulunabileceğini iddia etmişlerdir. Algoritmanın örnek evlilik eşleşmesi problemindeki adımları aşağıdaki gibidir (Gale & Shapley, 1962).

1. Her erkek en beğendiği kadına evlenme teklif eder.
2. Birden fazla teklif alan her kadın, teklif edenler arasından en beğendiğini "geçici" olarak kabul edip "beklemeye" alır ve diğerlerini reddeder. Kadın henüz son kararını vermez, çünkü daha çok beğendiği birinden teklif gelebilir.
3. Reddedilen erkekler ikinci tercihlerine evlilik teklif ederler.
4. Teklif alan her kadın, yeni teklif edenler ve beklemedeki erkek arasından en beğendiğini yine "beklemeye" alır diğerlerini reddeder.
5. Bu süreç, her kadın en az bir teklif alana kadar yani erkeklerin tüm teklifleri red ya da kabul edilene dek devam eder. Bir kadın teklif almadığı sürece reddedilmeler ve yeni teklifler olacaktır. Her erkek aynı kadına birden fazla teklifte bulunamadığı için her kadın mutlaka bir teklif alır.

6. Son kadın teklifini alır almaz "kur yapma" yani evlilik eşleşme süreci biter ve her kadın, geçici kabul ettiği beklemedeki erkeği kabul eder.

Teklifler hemen kabul edilmeyip en beğenilen geçici kabul ile beklemeye alındığından algoritma, ertelenmiş kabul (deferred-acceptance) algoritması olarak isimlendirilir. Daha iyi anlaşılabilmesi için algoritma bir erkek ve bir kadın kümesi üzerinden adım adım aşağıdaki gibi örneklendirilebilir.

Erkekler kümesi = {Alp, Cem, Ali, Veli, Efe}, Kadınlar kümesi = {Oya, Ahu, Ece, Eda, Naz}

Erkeklerin ve kadınların tercihlerinin aşağıdaki gibi olduğu varsayılmıştır.

Erkeklerin tercihleri (En çok beğenilenden en az beğenilene doğru)	
Alp	: Oya > Ahu > Ece > Eda > Naz
Cem	: Oya > Ece > Ahu > Naz > Eda
Ali	: Oya > Ahu > Naz > Eda > Ece
Veli	: Naz > Ahu > Ece > Eda > Oya
Efe	: Ahu > Eda > Oya > Naz > Ece

Kadınların tercihleri (En çok beğenilenden en az beğenilene doğru)	
Oya	: Cem > Alp > Efe > Veli > Ali
Ahu	: Alp > Ali > Efe > Veli > Cem
Ece	: Veli > Cem > Alp > Ali > Efe
Eda	: Efe > Veli > Ali > Cem > Alp
Naz	: Ali > Efe > Veli > Cem > Alp

1. Tekrar (Erkekler en çok beğendiklerine evlilik teklif ederler)

Alp → Oya, Cem → Oya, Ali → Oya, Veli → Naz, Efe → Ahu

Oya, teklif edenler arasından kendi beğeni sırasına göre Cem'i geçici kabul edip beklemeye alır; Alp ve Ali'yi reddeder.

Naz, Veli'yi geçici kabul eder.

Ahu, Efe'yi geçici kabul eder.

Birinci tekrar sonunda durum aşağıdaki gibi olur:

Geçici Eşleşmeler: (Cem–Oya), (Veli–Naz), (Efe–Ahu)
Boşta kalanlar: Alp, Ali

2. Tekrar (Reddedilen erkekler sonraki tercihlerine teklif ederler)

Alp → Ahu, Ali → Ahu

Ahu'nun mevcut geçici eşleşmesi Efe'dir, fakat Alp'i, Efe'den daha çok beğendiği için Alp'i geçici olarak kabul edip Efe'yi reddeder. Bu durumda Ali ve Efe reddedilir. İkinci tekrar sonunda durum aşağıdaki gibi olur:

Geçici Eşleşmeler: (Cem–Oya), (Alp–Ahu), (Veli–Naz)
Boşta kalanlar: Ali, Efe

3. Tekrar (Reddedilen erkekler sonraki tercihlerine teklif ederler)

Ali → Naz, Efe → Eda

Naz'ın tercihi: Ali > Efe > Veli > … şeklindedir. Mevcut geçici eşleşmesi olan Veli'dir, fakat Ali'yi daha çok beğendiği için Ali'yi geçici olarak kabul eder. Bu durumda Veli reddedilir. Eda, Efe'yi geçici kabul eder. Üçüncü tekrar sonunda durum aşağıdaki gibi olur:

Geçici eşleşmeler: (Cem–Oya), (Alp–Ahu), (Ali–Naz), (Efe–Eda)

Boşta kalanlar: Veli

4. Tekrar (Reddedilen erkekler sonraki tercihlerine teklif ederler)

Veli → Ahu

Ahu'nun tercihi: Alp > Ali > Efe > Veli > … şeklindedir. Mevcut geçici eşleşmesi Alp'i, Veli'ye tercih ettiğinden Veli reddedilir. Bu durumda geçici eşleşmeler değişmez. Dördüncü tekrar sonunda durum aşağıdaki gibi olur:

Geçici eşleşmeler: (Cem–Oya), (Alp–Ahu), (Ali–Naz), (Efe–Eda)

Boşta kalanlar: Veli

5. Tekrar (Reddedilen erkekler sonraki tercihlerine teklif ederler)

Veli → Ece

Ece evlilik teklif eden Veli'yi (zaten en çok beğendiği kişi) geçici kabul eder. Artık reddedilmiş boşta herhangi bir erkek kalmadığı, tüm teklifler için red ya da kabul kararı verildiği için algoritma durur. Beşinci tekrarın sonunda nihai kararlı eşleşme aşağıdaki gibi olur.

Alp – Ahu
Cem – Oya
Ali – Naz
Efe – Eda
Veli – Ece

Bu eşleşmede karşılıklı olarak birbirini tercih eden daha iyi hiçbir eşleşme yoktur. Dolayısıyla nihai evlilik eşleşme kümesinin kararlı olduğu söylenebilir. Örneğin, nihai çözümde Alp ve Oya eşleştirilmemişlerdir. Alp'in tercih listesinde, Oya'yı sonuçtaki eşleşmesine göre daha üst sırada tercih etmiştir. Alp birinci aşamada Oya'ya teklif etmiş ve Oya tarafından daha çok beğendiği biri lehine reddedilmiştir. Bu durumda Oya, mevcut eşini Alp'e tercih etmiştir ve çözümde kararsızlık durumu oluşmaz.

3.3. Üniversite Programlarına Öğrenci Yerleştirme İşlemi

Üniversite programlarına öğrenci yerleşim problemine geri dönüldüğünde problemin unsurları matematiksel olarak aşağıdaki gibi gösterilebilir.

N = Tercih yapan aday sayısı

M = Tercih edilebilecek program sayısı

S = En fazla tercih sayısı

a = { a_1, …, a_N} tercih yapan adaylar kümesi

p = {p_1, …, p_M} tercih edilebilecek programların kümesi

k = {k_1, …, k_M} tercih edilebilecek programların kontenjanları

t = {t_1, …, t_M} tercih edilebilecek programların puan türleri

r_{ij} = i. adayın j. tercihi i = {1,..N} , j = {1,..24}

Matematiksel olarak öğrenci yerleşimi, her adayı bir programa kontenjan kısıtını göz önünde tutarak yerleştiren ya da onu açıkta bırakan bir fonksiyondur.

Yine daha kolay anlaşılabilmesi için somut örnek aşağıdaki gibi kurgulanabilir.

Adaylar = { Cem, Oya, Alp, Ahu, Ali, Eda, Naz }
Programlar = {p_1, p_2, p_3 }

Programların Kontenjanları = {2, 2, 1}
Programların Puan türleri = {SAY, SÖZ, EA}

Pratikteki gerçek duruma benzemesi için kontenjan toplamı tercih yapan aday sayısından daha düşük olarak örneklendirilmiştir. Bu durumda öğrenciler kümesinde 7 aday vardır. Programlar kümesinde ise 3 programın toplam kontenjanı 5'tir. Dolayısıyla yerleştirme sonunda en az 2 aday açıkta kalacaktır. Adayların puanlarının Tablo 1'deki gibi olduğu varsayılmıştır. Gerçek hayattaki yerleştirme puanlarında ondalık değerler ve daha fazla puan türü olduğu aşikârdır. Örnek puan tablosunda problemi karmaşıklaştırmayıp basit kılmak adına ondalıklı haneler ile TYT ve DİL gibi diğer puan türlerine yer verilmemiştir.

Tablo 1

Örnek Puan Tablosu

Adaylar	SAY	SÖZ	EA
Cem	480	520	510
Oya	300	540	500
Alp	540	270	460
Ahu	470	480	535
Ali	455	410	430
Eda	430	500	495
Naz	440	495	470

Adayların tercihlerinin de aşağıdaki gibi olduğu varsayılmıştır. Yine gerçek duruma benzemesi düşüncesiyle adayların tercih sayıları birbirinden farklı tutulmuştur.

Cem : p_2 (520)> p_3 (510) > p_1 (480)
Oya : p_2 (540)> p_3(500)
Alp : p_1 (540)> p_3 (460)
Ahu : p_3 (535) > p_2 (480) > p_1 (470)
Ali : p_1 (455) > p_3 (430) > p_2 (410)
Eda : p_2 (500) > p_3 (495)
Naz : p_2 (495) > p_3 (470)

Bu durumda programlar kendilerini tercih eden adayları en yüksek puandan en düşük puana doğru sıraya koyduklarında programların tercihleri aşağıdaki gibi olur.

p_1 (SAY) : Alp(540) > Cem(480) > Ahu(470) > Ali(455)
p_2 (SÖZ) : Oya(540) > Cem(520) > Eda(500) > Naz(495) > Ahu(480) > Ali(410)
p_3 (EA) : Ahu(535) > Cem(510) > Oya(500) > Eda(495) > Naz(470) > Alp(460) > Ali(430)

1. **Tekrar** (Programlar kendilerini tercih eden adaylar içinde en yüksek puanlı olanlardan kontenjanları kadar adayı geçici kabul ederler)

p_1 → Alp, Cem
p_2 → Oya, Cem
p_3 → Ahu

Cem, hem p_1 hem de p_2 geçici kabul almıştır ancak tercih listesinde p_2 önce olduğu için p_1 reddedilir. Birinci tekrar sonunda durum aşağıdaki gibi olur:

Geçici Eşleşmeler:

p_1 → Alp
p_2 → Oya, Cem
p_3 → Ahu

Açıkta kalan adaylar: Ali, Eda, Naz

2. **Tekrar** (Reddedilen programlar boş kalan kontenjanları için sonraki en yüksek puanlı adayları almaya çalışır.)

p_1 → Ahu (sıradaki en yüksek)

Ahu, hâlihazırda p_3'te geçici kabul almıştır. Ahu'nun tercih listesinde $p_3 > p_1$ olduğu için p_1'i reddedilir. Bu durumda p_1'in hâlâ boş kontenjanı vardır. İkinci tekrar sonunda durum aşağıdaki gibi olur:

Geçici Eşleşmeler:

p_1 → Alp
p_2 → Oya, Cem
p_3 → Ahu

Açıkta kalan adaylar: Ali, Eda, Naz

3. **Tekrar** (Reddedilen programlar boş kalan kontenjanları için sonraki en yüksek puanlı adayları almaya çalışır.)

p_1 → Ali (sıradaki en yüksek)

p_1, Ali'nin zaten ilk tercihi olduğu için kabul edilir. Programların reddedilen kontenjanı kalmadığı ve tüm kontenjanlar dolduğu için algoritma durur. Nihai eşleşme aşağıdaki gibi olur.

p_1 (SAY) : Alp(540) > Ali(455)
p_2 (SÖZ) : Oya(540) > Cem(520)
p_3 (EA) : Ahu(535)

Nihai eşleşme sonunda Eda ve Naz herhangi bir programa yerleştirilememiş, açıkta kalmışlardır. Açıkta kalan adaylar, programlara yerleşen adayların taban puanlarından daha düşük puanlara sahiptir. Yedi aday ve üç programla basit olarak işletilen bu algoritma her yıl milyonlarca adayın, on binlerce programın yüz binlerce kontenjanına yerleştirilmesinde kullanılan üniversite yerleşim sisteminin küçük bir özetidir. Örnekte açıklanan algoritma, veri tabanı desteği ve günümüz yüksek seviyeli programlama dillerinin ileri veri yapıları ile çok sayıda iç içe döngü komutları kodlanarak geliştirilen yerleştirme yazılımlarının çekirdek kısmını oluşturmaktadır. Geliştirilen yazılımlar, çok sayıda döngünün tamamlanmasıyla kararlı bir yerleştirme sonucunun üretilmesine imkân sağlamaktadırlar. Üniversite yerleştirmelerinde, tüm adayların yararlanabilmesi için ilan edilen kontenjanlara genel kontenjan adı verilmektedir. İlan edilen genel kontenjanların yanı sıra özel durumu olan adaylar için de sınırlı sayıda kontenjan ilanı da yapılmaktadır. Örneğin ortaöğretim okullarını birincilikle bitiren adaylar için okul birinciliği kontenjanı ayrıca ilan edilir. Okul birinciliği, Yükseköğretim Kanunu ile tanımlanan özel bir kontenjan türüdür (Yükseköğretim Kanunu, 1981). Bu durumda tanımlanan hem genel hem de özel tüm kontenjanlar için bir öncelik sırası uygulamak gerekir. Söz gelimi Türkiye'deki üniversite merkezi yerleştirmesinde genel kontenjan ile toplam beş farklı kontenjan türü mevcuttur. Adayların puanları ve tercihleri göz önünde tutularak yerleştirmedeki öncelik sırası önce genel sonra diğer özel kontenjanlar olarak belirlenmiştir (ÖSYM, 2025). Yerleştirme algoritması genel kontenjanlara benzer şekilde özel kontenjanlar içinde işletilerek yerleştirme yapılır.

4. SONUÇ

Bu bölümün birinci başlığı altında, ölçme sonuçların yargıya dönüştürüldüğü değerlendirme süreci özetlenmiştir. Öncelikli olarak değerlendirmeler amacına göre sınıflandırılmış ve tanıma-yerleştirmeye yönelik değerlendirme, biçimlendirici değerlendirme ve değer biçmeye yönelik değerlendirme tanıtılmıştır. Sonrasında değerlendirmeler ölçütüne göre sınıflandırılmış ve mutlak değerlendirme ile bağıl değerlendirme ele alınmıştır. Son olarak eğitim politikalarına yön veren geniş ölçekli değerlendirmeler sunulmuştur. Bu bölümün ikinci başlığı altında, sınav sonucuna bağlı olarak adayları programlara yerleştirme işleminin temel prensipleri üzerinde durulmuştur. Yerleştirme işlemlerinin temelinde bulunan ertelemeli kabul algoritması kurgulanan örnekler üzerinden işletilerek açıklanmaya çalışılmıştır. Kitap bölümünü tamamlayan okuyucunun değerlendirmenin eğitimdeki işlevini kavrayarak yerleştirme algoritmasının temel çalışma prensiplerini anlaması ve açıklayabilmesi beklenmektedir. Programcı kimliğine ya da altyapısına sahip okuyucuların ise bu bölümü okuduktan sonra açıklanan algoritmayı kodlayabileceklerdir.

KAYNAKÇA | REFERENCES

Abdulkadiroglu, A. & Andersson, T. (2022). *School Choice* (Working Paper Series, Issue 29822). https://doi.org/10.3386/w29822

Akın Arıkan, Ç. (2021). Değerlendirme ve not verme. M. D. Şahin (Ed.), *Eğitimde ölçme ve değerlendirme* (s. 261-294) içinde. Nobel Akademi.

Balinski, M. & Sönmez, T. (1999). A tale of two mechanisms: *Student placement. Journal of Economic Theory*, *84*(1), 73–94. https://doi.org/10.1006/JETH.1998.2469

Binbaşıoğlu, C. (1983*). Eğitimde ölçme ve değerlendirme* (4. Baskı). Binbaşıoğlu Yayınevi.

Cizek, G. J. (2010). An introduction to formative assessment: History, characteristics, and challenges. H. L. Andrade & G. J. Cizek (Ed.*), Handbook of formative assessment* (s. 3-18) içinde. Routledge.

Erkuş, A. (2006). *Sınıf öğretmenleri için ölçme ve değerlendirme kavramlar ve uygulamalar*. Ekinoks Eğitim ve Danışmanlık Hiz.ve Bas. Yay. Dağ. San. Ve Tic. Ltd. Şti.

Gale, D. & Shapley, L. (1962). College admissions and the stability of marriage. *The American Mathematical Monthly*, *69*(1), 9–15.

İnce Aracı, F.G. (2023). Biçimlendirici ve özetleyici değerlendirme kaynakları. Tabak, H. & Özdemir, V. (Ed.). *Sınıf içi öğrenmeleri ölçme ve değerlendirme* (s. 245-277) içinde. Pegem Akademi.

Karagöz, N. & Alpaydın, Y. (2024). Türkiye'deki yükseköğretim kurumlarının not ve başarı sistemlerinin analizi. *Yükseköğretim ve Bilim Dergisi*, *14*(3), 403-412. https://doi.org/10.5961/higheredusci.1356938

Linn, R. L. (2000). Assessments and accountability. *Educational Researcher*, *29*(2), 4–16. https://doi.org/10.3102/0013189X029002004

McMillan, J. H. (2018). *Classroom assessment: Principles and practice that enhance student learning and motivation* (7. baskı). Pearson Education.

MEB, *Millî Eğitim Bakanlığı okul öncesi eğitim ve ilköğretim kurumları yönetmeliği (2014)..* https://www.mevzuat.gov.tr/mevzuat?MevzuatNo=19942&MevzuatTur=7&MevzuatTertip=5

MEB. (2014). Millî Eğitim Bakanlığı okul öncesi eğitim ve ilköğretim kurumları yönetmeliği. Resmî Gazete, Sayı: 29072.

MEB. (2023). Millî Eğitim Bakanlığı ortaöğretim kurumları yönetmeliği. Resmî Gazete, Sayı: 31262.

MEB. (2025). Türkiye Yüzyılı Maarif Modeli Öğretim Programları Ortak Metni.

Moss, C. M. & Brookhart, S. M. (2019). *Advancing formative assessment in every classroom: A guide for instructional leaders* (2. baskı). ASCD.

National Research Council. (2011). *Incentives and test-based accountability in education*. National Academies Press.

NobelPrize. (2012). *The Prize in Economic Sciences 2012 - Popular information - NobelPrize.org*. https://www.nobelprize.org/prizes/economic-sciences/2012/popular-information/

OECD. (2019). *PISA 2018 assessment and analytical framework*. OECD Publishing. https://doi.org/10.1787/b25efab8-en

Oyar, E. & Gündüz, T. (2023). Öğrenme çıktılarının değerlendirilmesi: not verme. Tabak, H. & Özdemir, V. (Ed.), *Sınıf içi öğrenmeleri ölçme ve değerlendirme* (s. 279-308) içinde.. Pegem Akademi.

ÖSYM Hizmetleri Hakkında Kanun, Pub. L. No. 6114, 27863 (2011). https://www.mevzuat.gov.tr/mevzuat?MevzuatNo=6114&MevzuatTur=1&MevzuatTertip=5

ÖSYM. (2011). Ölçme, Seçme ve Yerleştirme Merkezi Hizmetleri Hakkında Kanun, Kanun Numarası: 6114, https://www.mevzuat.gov.tr/mevzuat?MevzuatNo=6114&MevzuatTur=1&MevzuatTertip=5

ÖSYM. (2025). Yükseköğretim Kurumları Sınavı (YKS) Yükseköğretim Programları ve Kontenjanları Kılavuzu. https://dokuman.osym.gov.tr/pdfdokuman/2025/YKS/TERC%C4%B0H/tkilavuz_d09082025i.pdf

ÖSYM. (2025). Yükseköğretim Kurumları Sınavı (YKS) Yükseköğretim Programları ve Kontenjanları Kılavuzu. https://dokuman.osym.gov.tr/pdfdokuman/2025/YKS/TERC%C4%B0H/tkilavuz_d09082025i.pdf

ÖSYM. (2025, Ağustos 21). e-TEP (Electronic-Test of English Proficiency) – YDS (Yabancı Dil Bilgisi Seviye Tespit Sınavı) eşdeğerlik çerçevesi. https://dokuman.osym.gov.tr/pdfdokuman/2025/e-TEP/esdegerlik_21082025.pdf

Pellegrino, J. W., Chudowsky, N. & Glaser, R. (Ed.). (2001). Knowing what students know: The science and design of educational assessment. National Academies Press.

Rocher, T. & Hastedt, D. (2020). *International large-scale assessments in education: a brief guide IEA Compass: Briefs in Education No. 10*. Amsterdam, The Netherlands: IEA

Roth, A. E. (1984). The evolution of the labor market for medical ınterns and residents: A case study in game theory. *Journal of Political Economy*, *92*(6), 991–1016. https://doi.org/10.1086/261272

Shute, V. J. (2008). Focus on formative feedback. *Review of Educational Research*, *78*(1), 153–189.

Sönmez, T. & Ünver, M. U. (2011). Matching, allocation, and exchange of discrete resources. *Handbook of Social Economics*, *1*(1B), 781–852. https://doi.org/10.1016/B978-0-444-53187-2.00017-6

Tekin, H. (1979). *Eğitimde ölçme ve değerlendirme* (2. Baskı). Mars Matbaası.

The Nobel Prize. (2012). *The Prize in Economic Sciences 2012 - Popular information - NobelPrize.org.* https://www.nobelprize.org/prizes/economic-sciences/2012/popular-information/

Turgut, M. F. & Baykul, Y. (2019). *Eğitimde ölçme ve değerlendirme* (8. baskı). Pegem Akademi Yayıncılık.

Turgut, M.F. (1969). *Öğrenci başarısının değerlendirilmesi*. Millî Eğitim Bakanlığı Talim ve Terbiye Dairesi Araştırma ve Değerlendirme Bürosu.

Wu, M. (2010). *Comparing the similarities and differences of PISA 2003 and TIMSS. OECD Education Working Papers*, 32. OECD Publishing. https://doi.org/10.1787/5km4psnm13nx-en

Yıldırım, C. (1966). *Eğitimde ölçme ve değerlendirme*. Öğretmeni İşbaşında Yetiştirme Bürosu Yayınları.

YÖK. (1981). Yükseköğretim Kanunu. Kanun Numarası: 2547. https://www.mevzuat.gov.tr/mevzuat?MevzuatNo=2547&MevzuatTur=1&MevzuatTertip=5

Yükseköğretim Kanunu, Resmi Gazete 21 (1981). https://www.resmigazete.gov.tr/arsiv/17506.pdf

YAZARLAR HAKKINDA | ABOUT AUTHORS

DR. ÖĞR. ÜYESİ SUNGUR GÜREL

Siirt Üniversitesi |s.gurel[at]siirt.edu.tr

ORCID: 0000-0003-3425-858X

Dr. Sungur GÜREL, İstanbul Üniversitesi'nden lisans ve Florida Üniversitesi'nden yüksek lisans ve doktora dereceleri ile mezun olmuştur. Dr. Gürel, Siirt Üniversitesi Eğitim Fakültesi Ölçme ve Değerlendirme Anabilim Dalı'nda öğretim üyesidir. Dr. Gürel pek çok ulusal ve uluslararası projede nicel veri analizi uzmanı ve araştırmacı olarak görev yapmıştır. Çok sayıda uluslararası hakemli makale ve bildiriye imza atan Dr. Gürel'in uzmanlık alanları arasında eğilim puanlarının eğitimde kullanımı, yapısal eşitlik modellemesi, hiyerarşik doğrusal modelleme, gözlemsel ölçme, yabancı dil olarak Türkçe ve İngilizce dil yeterliliklerinin ölçülmesi ve puanlayıcı güvenirliği listelenebilir.

ASST. PROF. DR. SUNGUR GÜREL

Siirt University | s.gurel[at]siirt.edu.tr

ORCID: 0000-0003-3425-858X

Dr. Sungur GÜREL earned his bachelor's degree from Istanbul University and his master's and doctoral degrees from the University of Florida. Dr. Gürel is a faculty member in the Department of Measurement, Assessment and Evaluation at the Faculty of Education, Siirt University. Dr. Gürel has served as a quantitative data analysis expert and researcher on numerous national and international projects. Dr. Gürel has authored numerous international peer-reviewed articles and conference papers. His areas of expertise include the use of propensity scores in education, structural equation modeling, hierarchical linear modeling, observational measurement, the assessment of Turkish and English language proficiency as foreign languages, and rater reliability.

DR. ÖĞR. ÜYESİ MUHAMMET YORULMAZ
Başkent Üniversitesi | muhammed[at]baskent.edu.tr
ORCID: 0000-0002-4904-9396

Muhammet Yorulmaz, Başkent Üniversitesi Mühendislik Fakültesi, Endüstri Mühendisliği Bölümü öğretim üyesidir. Gazi Üniversitesi'nde tamamladığı lisans ve lisansüstü eğitiminden sonra akademik kariyerine "Bilişim Sistemleri", "İnsan Bilgisayar Etkileşimi", "Karar Destek sistemleri" ve "Veri Bilimi ve Analitiği" konuları üzerine odaklanarak devam etmiştir. Muhammet Yorulmaz, birçok ulusal ve uluslararası bilimsel yayına, kongre sunumuna ve araştırma projesine katkı sunmuştur.

ASST. PROF. DR. MUHAMMET YORULMAZ
Başkent University | muhammed[at]baskent.edu.tr
ORCID: 0000-0002-4904-9396

Muhammet Yorulmaz is a faculty member in the Department of Industrial Engineering, Faculty of Engineering at Başkent University. Following his undergraduate and graduate education completed at Gazi University, he has pursued his academic career by focusing on "Information Systems", "Human-Computer Interaction", "Decision Support Systems", and "Data Science and Analytics". Dr. Yorulmaz has contributed to numerous national and international scientific publications, conference presentations, and research projects.

13. BÖLÜM

TÜRKİYE YÜZYILI MAARİF MODELİ'NDE TANIMLANAN BECERİLERİN ÖLÇÜLMESİ

MEASUREMENT OF THE SKILLS DEFINED IN THE CENTURY OF TÜRKİYE EDUCATION MODEL

Doç. Dr. Sami PEKTAŞ

Dr. Öğr. Üyesi Semih AŞİRET

TÜRKİYE YÜZYILI MAARİF MODELİ'NDE TANIMLANAN BECERİLERİN ÖLÇÜLMESİ

Doç. Dr. Sami PEKTAŞ
Ankara Müzik ve Güzel Sanatlar Üniversitesi

Dr. Öğr. Üyesi Semih AŞİRET
Mersin Üniversitesi

Öz

Bu bölümde Türkiye Yüzyılı Maarif Modeli'nde tanımlanan becerilerin ölçülmesi, öğretim programlarının beceri temelli yapısı ve öğrenme kanıtları bağlamında ele alınmaktadır. Bölümün amacı, TYMM'de yer alan beceri çerçevesinin ölçme ve değerlendirme süreçlerine nasıl yansıdığını açıklamak; öğrenme çıktıları, süreç bileşenleri ve ölçme-değerlendirme uygulamaları arasındaki ilişkiyi ortaya koymaktır. Bu doğrultuda öncelikle beceri ve yeterlik kavramları açıklanmakta, becerilere ilişkin uluslararası çerçeveler özetlenmekte ve K12 Beceriler Çerçevesi: Türkiye Bütüncül Modeli ile TYMM arasındaki ilişki açıklanmaktadır. Ardından TYMM'de yer alan kavramsal beceriler, alan becerileri, öğrenme çıktıları ve süreç bileşenleri ölçme ve değerlendirme açısından incelenmektedir. Bölümde, TYMM öğretim programlarında ölçme ve değerlendirmenin "öğrenme kanıtları" başlığı altında yapılandırıldığı; tanılayıcı, biçimlendirici ve düzey belirleyici değerlendirme yaklaşımlarının öğrenme-öğretme süreciyle bütünleştirildiği vurgulanmaktadır. Ayrıca, becerilerin öncelikle sınıf içi uygulamalarla geliştirilip değerlendirilmesi gerektiği belirtilmekte; bununla birlikte LGS ve YKS gibi yüksek riskli sınavlarda beceri temelli öğrenme çıktılarının çoktan seçmeli maddelerle ölçülmesine ilişkin olanaklar ve sınırlılıklar tartışılmaktadır. Sonuç olarak, TYMM kapsamında ölçme ve değerlendirmenin yalnızca öğrenme sonunda yapılan bir kontrol işlemi olmadığını; öğrenme çıktıları, süreç bileşenleri ve öğrenme kanıtları temelinde öğrencinin beceri gelişimini izleyen, öğretimi yönlendiren ve öğrenme sürecini destekleyen bütüncül bir yapı olarak ele alınması gerektiğini ortaya koymaktadır.

Anahtar Kelimeler

Türkiye Yüzyılı Maarif Modeli, Öğretim Programı, Becerilerin Ölçülmesi, Sınıf İçi Ölçme Değerlendirme Uygulamaları, Yüksek Riskli Sınavlar

MEASUREMENT OF THE SKILLS DEFINED IN THE CENTURY OF TÜRKİYE EDUCATION MODEL

Assoc. Prof. Dr. Sami PEKTAŞ
Ankara Music and Fine Arts University

Asst. Prof. Dr. Semih AŞİRET
Mersin University

Abstract

This chapter examines the assessment of the skills defined in the Century of Türkiye Education Model (CTEM) in relation to the skills-based structure of the curricula and evidence of learning. The aim of the chapter is to explain how the skills framework in CTEM is reflected in measurement and evaluation processes and to clarify the relationship among learning outcomes, process components, and assessment practices. To this end, the concepts of skill and competence are first explained, international frameworks for skills are summarized, and the relationship between the K-12 Skills Framework: Türkiye Holistic Model and CTEM is evaluated. The chapter then examines the conceptual skills, domain-specific skills, learning outcomes, and process components in CTEM from an assessment and evaluation perspective. It emphasizes that assessment and evaluation in CTEM curricula are structured under the heading of "evidence of learning" and that diagnostic, formative and summative assessment approaches are integrated into the teaching-learning process. In addition, the chapter argues that skills should primarily be developed and assessed through classroom practices, while also discussing the possibilities and limitations of measuring skill-based learning outcomes through multiple-choice items in high-stakes examinations such as LGS and YKS. In conclusion, the chapter shows that measurement and evaluation within CTEM should not be regarded merely as an end-of-learning control procedure, but rather as a holistic structure that monitors students' skill development, guides instruction, and supports the learning process on the basis of learning outcomes, process components, and evidence of learning.

Keywords

Century of Türkiye Education Model, Curriculum, Skill Assessment, Classroom Assessment Practices, High-Stakes Tests

1. GİRİŞ

Günümüzde eğitim sistemlerinden beklenen rol, yalnızca bilginin öğrenciye aktarılmasıyla sınırlı değildir. Öğrencilerin bilgiyi anlamlandırmaları, farklı bağlamlarda kullanmaları, eleştirel biçimde değerlendirmeleri ve yeni ürün ya da çözümler geliştirmeleri çağdaş eğitimin temel hedefleri arasında yer almaktadır. Dijitalleşme, yapay zekâ, otomasyon ve toplumsal yaşamda yaşanan hızlı değişim, bu beklentiyi son yıllarda daha görünür hâle getirmiştir. Bu nedenle eğitimde yaşanan dönüşümü, bilginin önemini yitirmesi olarak değil, bilginin beceriler aracılığıyla kullanılmasını, dönüştürülmesini ve yeniden yapılandırılmasını öne çıkaran bir yönelim olarak değerlendirmek daha doğru olacaktır. Bilginin öğretmen ya da ders kitabı gibi otorite kabul edilen kaynaklardan öğrenciye doğrudan aktarıldığı öğretmen merkezli yaklaşım, 20. yüzyıl sınıflarının belirgin görünümlerinden biri olmuştur ve günümüzde de bazı sınıf içi uygulamalarında varlığını sürdürmektedir (Teo, 2019). Ancak öğrenciyi çoğunlukla pasif bir bilgi alıcısı olarak konumlandıran bu anlayış, eğitim bilimleri alan yazınında uzun süredir sınırlı görülmekte, bunun yerine öğrencinin bilgiyi sorgulayan, ilişkilendiren, gerekçelendiren ve üreten etkin bir özne olarak konumlandırıldığı yaklaşımlar öne çıkmaktadır (National Research Council [NRC], 2000; Sawyer, 2022).

Bu dönüşüm, eğitim alan yazınında yaygın biçimde "21. yüzyıl becerileri" kavramı etrafında tartışılmaktadır. Bununla birlikte, bu kavramın işaret ettiği becerilerin tümüyle yeni olduğu söylenemez. Eleştirel düşünme, problem çözme, yaratıcılık, iş birliği ve bilgi okuryazarlığı gibi becerileri edindirmek önceki dönemlerde de eğitimin önemli amaçları arasında yer almıştır. Yeni olan, dijitalleşme, toplumsal değişim ve iş yaşamındaki dönüşümle birlikte bireysel ve toplumsal başarının bu becerilere giderek daha fazla bağlı hâle gelmesidir (Greiff & Kyllonen, 2016; Rotherham & Willingham, 2010; Silva, 2009). Artan bu talep, becerilerin eğitim politikalarına ve öğretim programlarına daha sistematik biçimde dâhil edilmesi yönünde küresel bir eğilimi de beraberinde getirmiştir (OECD, 2019; Voogt & Roblin, 2012). Bu nedenle 21. yüzyıl becerilerinin dağınık bir beceri listesi olarak değil, kuramsal ve işlevsel ilişkileri bulunan bir yeterlikler bütünü olarak ele alınması gerekmektedir. Bu bağlamda 21. Yüzyıl Becerileri Ortaklığı'nın (Partnership for 21st Century Skills, P21) çerçevesi, 21. Yüzyıl Becerilerinin Değerlendirilmesi ve Öğretimi (Assessment and Teaching of 21st Century Skills, ATC21S) Projesi ve OECD'nin Öğrenme Pusulası 2030'u (OECD Learning Compass 2030) gibi çalışmalar kapsamında tanımlanan çerçeveler, öğrencilerin ihtiyaç duyacağı beceri ve yeterlikleri sınıflandırılmış kategoriler altında yapılandırmaktadır. Ayrıca bu çerçeveler, öğrencilerin 21. yüzyılın toplumsal ve ekonomik yapısında yön bulmalarını, değişime uyum sağlamalarını ve topluma katkıda bulunmalarını amaçlayan öğretim ve değerlendirme yaklaşımlarına yön vermektedir.

Kitabın bu bölümünde öncelikle beceri kavramı, beceri-yeterlik ilişkisi ve alan yazındaki başlıca beceri ve yeterlik çerçeveleri ele alınmıştır. Ardından Türkiye Yüzyılı Maarif Modeli'nin (TYMM) temel yaklaşımı, modelde yer alan beceri çerçevesi ve öğretim programlarına yansıma biçimi açıklanmıştır. İzleyen bölümde TYMM öğretim programlarında benimsenen ölçme ve değerlendirme yaklaşımı incelenmiş; becerilerin sınıf içi ölçme ve değerlendirme uygulamalarında nasıl ele alınabileceği çeşitli örnekler üzerinden gösterilmiştir. Son olarak yüksek riskli sınavlarda TYMM'de tanımlanan becerilerin ölçülmesine ilişkin olanaklar, sınırlılıklar ve dikkat edilmesi gereken ilkeler tartışılmıştır.

2. BECERİ VE YETERLİKLERE İLİŞKİN KAVRAMSAL ÇERÇEVE

2.1. Beceri Tanımı

İlgili alan yazın incelendiğinde "beceri (skill)" kavramı için evrensel düzeyde üzerinde uzlaşılmış tek bir tanımın bulunmadığı ve farklı araştırmacıların beceriyi çeşitli yönleriyle ele aldığı görülmektedir. Beceri kavramı bazı kaynaklarda "yeterlik (competence)" kavramıyla eş anlamlı kullanılırken (Ananiadou & Claro, 2009; Voogt & Roblin, 2012; Westera, 2001), bazı çalışmalarda daha somut, dar kapsamlı ve uygulamaya yönelik olarak ele alınmaktadır (Le Deist & Winterton, 2005; Weinert, 2001). Yeterlik, bireyin belirli bir bağlamda karşılaştığı karmaşık taleplere etkili biçimde yanıt verebilmek için bilgi, beceri, değer, tutum gibi yapıları uygun biçimde bir araya getirerek kullanabilme kapasitesi (OECD, 2019) olarak tanımlanmıştır. Rychen (2004) yeterliği bilgi, bilişsel beceriler, tutumlar ve diğer bilişsel olmayan bileşenleri kapsayan karmaşık bir eylem sistemi olarak tanımlarken beceriyi motor ve/veya bilişsel eylemleri gerçekleştirme yeteneği olarak tanımlamıştır. CC2020 (Computing Curricula Guidelines for the 2020s) projesinde de yeterliğe ilişkin kavramsal bir yapı ortaya konmuştur. Bu yapıda yeterlik; bilgi, beceri ve eğilimden oluşmaktadır ve bir görev bağlamında bu üç boyutun birbiriyle ilişkilendirilmesi gerektiği belirtilmektedir. Bu yapıda beceri "bir görevi aktif olarak yerine getirmek için bilgiyi uygulama yeteneği" olarak tanımlanmıştır. AB Konseyi'nin 2018 tarihli Yaşam Boyu Öğrenme İçin Temel Yeterlikler Tavsiye Kararı'nda (Council of the European Union, 2018) yeterlik; bilgi, beceri ve tutumların bir birleşimi olarak tanımlanırken beceri ise süreçleri yürütme ve mevcut bilgileri kullanarak sonuçlara ulaşma kapasitesi olarak tanımlanmıştır.

İlgili alan yazında becerinin farklı tanımları yer almaktadır. Thornhill-Miller ve arkadaşları (2023) beceriyi uygulama, deneyim ve eğitim yoluyla edinilen, görevleri etkili biçimde yerine getirmeyi sağlayan öğrenilmiş kapasite veya yetenek olarak tanımlamaktadır. Lucas ve Claxton (2009) ise beceriyi en yalın hâliyle "öğrenilmiş yararlı bir şey yapma kapasitesi" olarak ifade etmektedir. Lamri ve Lubart (2023) becerilerin çok boyutlu yapısına dikkat çekerek onları, bağlam içinde karşılaşılan problemleri çözmek ve görevleri doğru zamanda, uygun kaynakları bir araya getirerek yerine

getirme kapasitesi olarak tanımlamıştır. Bu bakış açısına göre beceriler; uygulama, deneyim ve eğitim yoluyla gelişen dinamik yapılar olup bireyin performansını belirleyen önemli bir bileşendir. DeKeyser (2020) de beceriyi bir görevi tamamlamak için bilgiyi işleme, anlama, yorumlama ve kullanma yeteneği olarak ele almıştır. Ayrıca tekrar ve uygulama yoluyla becerilerin giderek otomatikleştiğini ve performansta yeterliliğin görünür hâle geldiğini ifade etmiştir.

Bu tanımlardaki çeşitliliğe karşın alan yazında beceri, çoğunlukla yeterliğin bir bileşeni olarak daha sınırlı ve işlemsel bir yapıyı ifade edecek biçimde konumlandırılmaktadır. Örneğin etkili iletişim kurma yeteneği; bireyin dil bilgisinden, pratik bilgi teknolojisi becerilerinden ve iletişim kurduğu kişilere yönelik tutumlarından beslenen bir yeterliktir (OECD, 2005). Görüldüğü üzere bu örnekte beceri (bilgi teknolojilerini kullanma), bilgi (dil bilgisi) ve tutumla birlikte yeterliğin bileşenlerinden yalnızca birini oluşturmaktadır. Bununla birlikte işlemsel bilgi ya da görevi etkili biçimde gerçekleştirmeye yönelik motivasyon gibi bazı özellikler olmadan bir becerinin uygulanabilirliğini düşünmek de zordur. Bu doğrultuda Dreyfus ve Dreyfus (1986), becerinin özünde "nasıl yapılacağını bilme"yi (knowing-how), başka bir ifadeyle işlemsel bilgiyi (procedural knowledge) içerdiğini; olgulara ve kurallara ilişkin "ne olduğunu bilme" (knowing-that) biçimindeki bildirimsel bilginin (declarative knowledge) ise tek başına performansı açıklama için yeterli olmadığını ifade etmektedir.

Bu yorumlardan hareketle beceri, yalnızca bilginin hatırlanmasıyla sınırlı olmayıp, bilginin bir görev bağlamında etkili biçimde kullanılabilme kapasitesi olarak değerlendirilmektedir. Bununla birlikte beceri, bireyin karmaşık taleplere yanıt verme kapasitesini tek başına açıklamakta yetersiz kalmaktadır. Beceri; bilgi, tutum ve eğilimlerle bütünleştiğinde yeterliğe dönüşmektedir. Bu çerçevede beceri, yeterliğin gözlenebilir ve uygulamaya dönük boyutunu oluşturmakta; yeterlik ise becerinin hangi bağlamda, hangi bilgi ve tutumlarla birlikte kullanılacağını belirleyen daha geniş kapsamlı bir yapıyı ifade etmektedir.

2.2. Beceri ve Yeterlik Çerçeveleri

Beceri ve yeterlik alanında yapılan çalışmalar ülkelerin kendi sosyoekonomik, kültürel ve politik bağlamları doğrultusunda farklı çerçeveler geliştirdiğini göstermektedir. 21. yüzyıl becerilerine ilişkin geliştirilen çerçeveler incelendiğinde bu becerilerin tanımlanış biçiminden ziyade, nasıl çerçevelendiği konusunda çeşitlilik olduğu görülmektedir (Care, 2018). Alan yazında beceri tanımlarının farklılaşması, kavramın çeşitli kurum ve kuruluşlar tarafından farklı anlamlarda kullanılması (Saavedra & Opfer, 2012) ve kavramın geniş ve zaman zaman tutarsız biçimde ele alınması (Dede, 2010) çerçeveler arasındaki bu farklılığın temel nedenleri arasında değerlendirilebilir. Dolayısıyla tek ve evrensel bir sınıflandırmadan ziyade çoklu ve birbiriyle örtüşen çerçevelerin olduğu anlaşılmaktadır. Kotsiou ve arkadaşları (2022) tarafından yapılan derleme çalışmasında 99 farklı beceri çerçevesi incelemiş ve bu çerçevelerde beceri olarak adlandırılan 341

farklı terimin yer aldığı belirlenmiştir. Üstelik bu terimlerin bir bölümünün tutum ve bilgi türlerine karşılık gelmesi, kavram kullanımındaki tutarsızlığın boyutunu ortaya koymaktadır.

Bu alandaki öncü çalışmalardan biri, 2002 yılında ABD'de kurulan 21. Yüzyıl Becerileri Ortaklığıdır. Bu girişim, iş dünyası temsilcileri, eğitim liderleri ve politika yapıcılardan oluşan bir koalisyon tarafından oluşturulmuştur. Başlangıçta Partnership for 21st Century Skills adıyla kurulan P21'in adı, 2015 yılında Partnership for 21st Century Learning (21. Yüzyıl Öğrenme Ortaklığı) olarak değiştirilmiştir (Partnership for 21st Century Learning, 2015). Ortaklık tarafından geliştirilen 21. Yüzyıl Öğrenme Çerçevesi, temel ders bilgisini öğrenmenin temeli olarak kabul etmektedir. Çerçevede öğrenci çıktıları üç temel beceri alanı altında ele alınmaktadır. Bunlar, öğrenme ve yenilik becerileri, bilgi, medya ve teknoloji becerileri ile yaşam ve kariyer becerileridir. Çerçevenin en bilinen boyutu ise öğrenme ve yenilik becerileri kapsamında yer alan "4C" becerileridir. Bu beceriler eleştirel düşünme (critical thinking), yaratıcılık (creativity), iletişim (communication) ve iş birliği (collaboration) olarak tanımlanmaktadır (Soulé & Warrick, 2015).

Uluslararası ölçekte etkili olan bir diğer çerçeve, Cisco, Intel ve Microsoft'un desteğiyle Melbourne Üniversitesi öncülüğünde yürütülen 21. Yüzyıl Becerilerinin Değerlendirilmesi ve Öğretimi (ATC21S) Projesi'dir (Care, 2018; Griffin & Care, 2015). Örgün eğitimin öğrencileri hızla değişen iş gücü taleplerine yeterince hazırlayamadığı görüşünden hareket eden proje, aralarında P21'in de bulunduğu on iki çerçevenin analizine dayalı bir üst-model geliştirmiştir. Bu modelde on beceri; düşünme biçimleri, çalışma biçimleri, çalışma araçları ve dünyada yaşama olmak üzere dört kategori altında toplanmış ve her beceri, bilgi, beceri, tutum, değer ve etik (KSAVE) bileşenleri açısından işevuruklaştırılmıştır (Binkley ve ark., 2012). Proje, becerilerin öğretimi ile değerlendirilmesi arasındaki ilişkiyi kurması bakımından alan yazınında önemli bir konuma sahiptir.

OECD'nin beceri sınıflandırmasına yönelik çalışmaları görece erken bir tarihte başlamıştır; Ananiadou ve Claro (2009) OECD ülkeleri için hazırladıkları raporda 21. yüzyıl becerilerini bilgi, iletişim ve etik/sosyal etki boyutlarında ele almıştır. Kurumun bu alandaki en kapsamlı çalışması ise Eğitim ve Becerilerin Geleceği 2030 Girişimi'dir. Bu girişimin ana çıktısı olan Öğrenme Pusulası 2030 (OECD Learning Compass 2030); temel yapı taşları, bilgi, beceri, tutum ve değerler, dönüştürücü yeterlikler, öğrenci faaliyeti ve öngörü-eylem-yansıtma döngüsünden oluşan bir öğrenme çerçevesi sunmaktadır (OECD, 2019). Çerçevede dönüştürücü yeterlikler; yeni değer yaratma, sorumluluk alma ile çatışma, gerilim ve ikilemleri uzlaştırma olarak tanımlanmakta ve öğrencilerin yalnızca değişime uyum sağlamaları değil, bireysel ve toplumsal iyi oluş yönünde geleceği biçimlendirmeleri amaçlanmaktadır. Pusulada beceri, bir hedefe ulaşmak için süreçleri yürütme ve bilgiyi sorumlu biçimde kullanabilme yeteneği olarak tanımlanmaktadır. Ayrıca, bilişsel ve üstbilişsel, sosyal ve duygusal ile fiziksel ve pratik

beceriler birbirinden ayırt edilmektedir (OECD, 2019). Farklı bir sınıflandırma sunan Global Cities Education Network (GCEN) becerileri bilişsel, kişilerarası ve kişi içi olmak üzere üç üst düzey kategori altında toplamaktadır (Soland ve ark., 2013). İş gücü bakış açısıyla konuyu ele alan çalışmalar arasında ise Lippman ve arkadaşları (2015) tarafından iş gücü başarısıyla ilişkilendirilen beş temel yumuşak beceri ile Wagner'in (2010) 21. yüzyıl yaşamı, iş dünyası ve vatandaşlık için gerekli gördüğü yedi hayatta kalma becerisi öne çıkmaktadır.

3. TÜRKİYE YÜZYILI MAARİF MODELİ

Beceri ve yeterliklerin tanımlanması, kazandırılması ve ölçülmesine yönelik girişimler, çağın değişen toplumsal, ekonomik ve teknolojik gereksinimlerinden kaynaklanmaktadır. Bu yönelim, ülkelerin millî eğitim sistemlerini söz konusu gereksinimlere uyumlu hâle getirmek amacıyla kapsamlı beceri modelleri geliştirmesini doğurmuştur. Finlandiya, Singapur, İskoçya, Avustralya gibi ülkeler becerileri; değerler, disiplinler arası öğrenme ve üst düzey düşünme süreçleriyle bütünleştiren ulusal çerçeveler geliştirmiş ve bunları öğretim programlarına yansıtmıştır. Küresel düzeydeki bu paradigma değişimine paralel olarak, ülkemizde de öğretim programları geliştirilmeden önce ulusal bir beceri çerçevesi oluşturulmuştur. Söz konusu çerçeve, "K12 Beceriler Çerçevesi: Türkiye Bütüncül Modeli" adıyla hayata geçirilmiştir.

3.1. K12 Beceriler Çerçevesi: Türkiye Bütüncül Modeli

Türkiye'de Millî Eğitim Bakanlığı (MEB) ve UNICEF iş birliğinde yürütülen K12 Beceriler Çerçevesi Projesi kapsamında "K12 Beceriler Çerçevesi: Türkiye Bütüncül Modeli" (Aşkar & Altun, 2023) geliştirilmiştir. Geliştirilen çerçeve, millî eğitim sisteminin değişen ihtiyaçlarına yanıt vermeyi amaçlamaktadır. Bu çerçevenin oluşturulmasının temel gerekçeleri arasında öğrencilere 21. yüzyıl becerilerinin kazandırılmasına duyulan ihtiyaç, bilgi temelli öğrenme yaklaşımından beceri temelli yaklaşımlara geçiş gerekliliği ve becerilerin öğrenenlere nasıl aktarılacağı konusunda ortak bir anlayış oluşturma isteği yer almaktadır. Ayrıca becerilere geçiş sürecine rehberlik edecek sistematik bir yapıya duyulan ihtiyaç, mevcut sistemde bütüncül bakış açısının eksikliği ve öğretim programlarında becerilerin tutarlı ve ortak bir çerçevede tanımlanma gerekliliği de bu modelin geliştirilmesinin temel gerekçelerini oluşturmaktadır.

K12 Beceriler Çerçevesi'nde beceri, "bir çalışma veya öğrenme alanında edinilen mantıksal veya sezgisel nitelikli düşünme yaklaşımı ile el becerisi, yöntem ve araç-gereç kullanmayı gerektiren her tür edim veya eylem" olarak tanımlanmıştır. Yeterlik ise "belirli bir konuda bireylerin sahip olmaları gereken bilgi, beceri ve eğilimleri gösteren bütüncül yapılar" olarak tanımlanmıştır. Tanımdan da anlaşılacağı gibi yeterlik çerçevesi; bilgi (alan bilgisi), beceri ve eğilimden oluşmaktadır. Beceri çerçevesi ise kavramsal beceriler, sosyal-duygusal öğrenme becerileri, alan becerileri ve eğilimlerden oluşmaktadır. Bununla birlikte K12 Beceriler Çerçevesi'nde okuryazarlık becerileri de ayrıca bir beceri alanı olarak tanımlanmıştır. Bu çerçeve, TYMM öğretim programında yer alan becerilerin temelini oluşturmuştur.

3.2. TYMM Beceri Çerçevesi

K12 Beceriler Çerçevesi'nde belirlenen yapı güncellenerek TYMM öğretim programlarına aktarılmıştır. TYMM kapsamında kavramsal beceriler (temel, bütünleşik ve üst düzey), alan becerileri (Türkçe, matematik, fen bilimleri, sosyal bilimler), okuryazarlık becerileri ve sosyal-duygusal öğrenme becerileri, K12 Beceriler Çerçevesi referans alınarak yeniden düzenlenmiş ve bu becerilere ilişkin süreç bileşenleri tanımlanmıştır. 2025 yılında güncellenen TYMM'de çerçeve genişletilerek fiziksel beceriler de tanımlanmıştır. Ayrıca alan becerilerine sanat, beden eğitimi, oyun ve spor, bilişim teknolojileri, tasarım, din eğitimi ve öğretimi ile yabancı dil alan becerileri ve yabancı dil destek becerileri eklenmiştir (MEB, 2025). Böylece K12 Beceriler Çerçevesi'nde oluşturulan kuramsal yapı, TYMM aracılığıyla öğretim programlarına yansıtılmıştır. Aşağıda TYMM'de tanımlanan kavramsal beceriler ve alan becerileri sırasıyla ele alınmıştır.

3.2.1. Kavramsal Beceriler

Kavramsal beceriler, gözlenebilir "temel" beceriler ile soyut fikirlerin ve karmaşık süreçlerin uygulanmasında kullanılan "bütünleşik" ve "üst düzey düşünme" becerilerini kapsayan bütüncül bir yapıyı ifade etmektedir. Tüm boyutları ortak bir yapıda tanımlamak amacıyla "kavramsal beceri" terimi kullanılmıştır (MEB, 2025). Temel ve bütünleşik beceriler, aşamalılık içerisinde olmasa da üst düzey düşünme becerilerinin gelişimi için gerekli alt yapıyı oluşturmaktadır.

Kavramsal becerilerden temel beceriler; sayma, okuma, yazma, kesme gibi karmaşık bilişsel süreçler gerektirmeyen, doğrudan gözlenebilir ve uygulanabilir eylemleri ifade etmektedir. TYMM'de temel beceriler, doğası gereği birincil düzey eylemler olduklarından, TYMM bünyesinde ayrıca süreç bileşenlerine ayrıştırılmamıştır. Bütünleşik beceriler ise temel düzeyde gerçekleştirilen eylemleri aşarak süreç modellemesine dayanan daha karmaşık ve çok adımlı davranışları ifade etmektedir. TYMM'de bu kapsamda 20 bütünleşik beceri tanımlanmıştır. Bu beceriler arasında herhangi bir hiyerarşi bulunmamaktadır. Her bir bütünleşik beceri için ayrıca "süreç bileşeni" adı verilen bileşenler tanımlanmıştır. Beceri, doğası gereği bir süreç içermektedir. Süreç bileşeni, beceri edinim sürecindeki işlem adımlarını ifade etmektedir. TYMM'nin ayırt edici yönlerinden biri, becerileri yalnızca tanımlamakla kalmayıp bunları gözlenebilir ve ölçülebilir hâle getiren süreç bileşenlerini de ortaya koymasıdır. Şekil 1'de bütünleşik becerilerden "karşılaştırma becerisi" ve buna ait süreç bileşenleri yer almaktadır (MEB, 2025).

Şekil 1

Karşılaştırma Becerisi ve Süreç Bileşenleri (MEB, 2025)

KB2.7. Karşılaştırma Becerisi: Birden fazla kavram veya durum arasındaki benzerlik ve farklılıkların belirlenmesini ifade eder.

Süreç Bileşenleri

- KB2.7.SB1. Birden fazla kavram veya duruma ilişkin özellikleri belirlemek
- KB2.7.SB2. Belirlenen özelliklere ilişkin benzerlikleri listelemek
- KB2.7.SB3. Belirlenen özelliklere ilişkin farklılıkları listelemek

Becerilere ilişkin tanımlanan süreç bileşenleri, yapısal olarak ardışık bir düzen gösterebildiği gibi ilişkisel bir örüntü de sergileyebilmektedir. Şekil 1'de yer alan karşılaştırma becerisinin süreç bileşenleri incelendiğinde birinci (KB2.7.SB1) ve ikinci (KB2.7.SB2) süreç bileşenleri arasında belirgin bir ardıllık bulunduğu görülmektedir. Buna göre öğrencinin karşılaştırma yapabilmesi için öncelikle ilgili kavram veya duruma ait özellikleri belirlemesi gerekmektedir. Buna karşılık, karşılaştırma becerisinin ikinci (KB2.7.SB2) ve üçüncü (KB2.7.SB3) süreç bileşenleri arasında kesin bir ardıllık yerine ilişkisel bir yapı söz konusudur. Öğrenci, belirlediği özelliklere dayalı olarak önce benzerlikleri ardından farklılıkları listeleyebileceği gibi aynı süreci ters sırayla da gerçekleştirebilir. Bu nedenle TYMM'de bazı becerilerin süreç bileşenlerinin katı bir hiyerarşiden ziyade esnek bir ilişkisellik içinde işleyebildiği söylenebilir.

Üst düzey düşünme becerileri ise birden fazla bütünleşik beceriyi içeren, çok boyutlu zihinsel süreçler gerektiren eylemler olarak tanımlanmıştır (MEB, 2025). TYMM'de karar verme, problem çözme ve eleştirel düşünme olmak üzere üç üst düzey düşünme becerisi tanımlanmıştır. Bütünleşik becerilere benzer şekilde üst düzey düşünme becerilerinin de süreç bileşenleri tanımlanmıştır. TYMM'ye göre bütünleşik beceriler, üst düzey düşünme becerilerinin alt yapısını oluşturmaktadır. Bu nedenle bütünleşik becerilerdeki süreç bileşenlerinden farklı olarak üst düzey düşünme becerilerinin süreç bileşenleri bütünleşik becerilerden oluşmaktadır. Şekil 2'de problem çözme becerisi ve buna ilişkin süreç bileşenleri yer almaktadır (MEB, 2025).

Şekil 2

Problem Çözme Becerisi ve Süreç Bileşenleri (MEB, 2025)

KB3.2. Problem Çözme Becerisi: Karşılaşılan sorunla baş edebilmek için duruma ilişkin yapılandırma, özetleme, gözleme dayalı / mevcut bilgiye / veriye dayalı tahmin etme, akıl yürütme, yansıtma / değerlendirmede bulunmayı kapsayan zihinsel işlemleri yürütebilmeyi ifade eder.

Süreç Bileşenleri

- KB3.2.SB1. Problemi yapılandırmak
- KB3.2.SB2. Problemi özetlemek
- KB3.2.SB3. Problemin çözümüne yönelik gözleme dayalı / mevcut bilgiye / veriye dayalı tahminde bulunmak
- KB3.2.SB4. Problemin çözümüne yönelik önermeler üzerinden akıl yürütmek
- KB3.2.SB5. Problemin çözümüne ilişkin yansıtma/değerlendirmede bulunmak

Şekil 2 incelendiğinde sırasıyla yapılandırma (KB2.13.), özetleme (KB2.3.), mevcut bilgiye/veriye dayalı tahmin etme (KB2.12.), muhakeme (akıl yürütme) (KB2.16.) ve yansıtma (KB2.15.) veya değerlendirme (KB2.17.) bütünleşik becerilerinin, problem çözme becerisinin süreç bileşenleri olarak kullanıldığı görülmektedir. Ayrıca bu süreç bileşenleri arasında ardıllığın olduğu söylenebilir.

3.2.2. Alan becerileri

Bilgi ve beceri, çoğu durumda birbirinden ayrılması güç olan, hatta çoğu zaman bütünleşik şekilde işleyen iki tamamlayıcı unsurdur. TYMM'de alan becerilerinin oluşturulması da bu bütüncül anlayışa dayanmaktadır. Alan becerileri, kavramsal beceriler ile alana özgü kavramların etkileşiminden doğmakta ancak eylem boyutunda çekirdek becerilerin ötesine geçen ve farklı beceri türlerini kapsayan daha geniş bir yapıyı temsil etmektedir. 2024 yılında yayımlanan TYMM Ortak Metni'nde Türkçe, matematik, fen bilimleri ve sosyal bilimler olmak üzere dört alan becerisi tanımlanmıştır. 2025 yılında güncellenen modelde ise bu yapı genişletilmiştir. Modele; sanat, beden eğitimi, oyun ve spor, bilişim teknolojileri ve yazılım, tasarım, din eğitimi ve öğretimi, yabancı dil alan becerileri ve yabancı dil destek becerileri eklenmiş ve böylece tanımlanan toplam alan becerisi sayısı 10'a yükselmiştir (MEB, 2025).

Alan becerilerinin oluşturulması sürecinde her alan için bütünleşik beceriler belirlenmiş ve bu becerilerin süreç bileşenleri sistematik biçimde tanımlanmıştır. Alana özgü bütünleşik becerilerin geliştirilmesinde öncelikle kavramsal becerilerde yer alan bütünleşik beceriler dikkate alınmış ancak bu becerilerin yeterli olmadığı durumlarda alanın özgün yapısını karşılayan yeni bütünleşik beceriler tanımlanmıştır. Örneğin matematik alan becerileri; matematiksel muhakeme, matematiksel problem çözme,

matematiksel temsil, veri ile çalışma ve veriye dayalı karar verme ile matematiksel araç ve teknoloji ile çalışma olmak üzere beş alan becerisinden oluşmaktadır. Bu bağlamda matematiksel muhakeme alan becerisi, "*bilgi veya varsayımlar kullanarak mantıksal çıkarımlarda bulunma süreci*" olarak tanımlanmış ve çözümleme (KB2.4), yorumlama (KB2.14), çıkarım yapma (KB2.10) ile matematiksel doğrulama/ispat yapma (MAB1.1) bütünleşik becerilerinden oluşmuştur. Bu beceriler incelendiğinde ilk üç bütünleşik becerinin kavramsal beceri setinden alındığı, ispat yapmaya ilişkin becerinin ise alana özgü olarak tanımlandığı görülmektedir. Ayrıca söz konusu bütünleşik becerilerinin süreç bileşenleri de alanın doğasına uygun şekilde yapılandırılmıştır. Kavramsal becerilerden aktarılan bütünleşik becerilerde süreç bileşenleri ya doğrudan alınmış ya da alan bağlamına uyarlanarak kullanılmıştır. Örneğin çözümleme becerisi kavramsal becerilerde iki süreç bileşeninden oluşmakta ve matematiksel muhakeme becerisi altında yer alan çözümleme becerisinde bu iki süreç bileşeni birebir korunmaktadır. Ancak yorumlama becerisinde, kavramsal becerilerde yer alan ilk süreç bileşeni (*KB2.14.SB1. Mevcut olay/konu/durumu incelemek*) matematik alanındaki yorumlama becerisinin süreç bileşenleri arasında yer almamaktadır. Bu durum, alan becerilerinin oluşturulmasında kavramsal bütünleşik becerilerin esas alındığını ancak her alanın özgün gereksinimleri doğrultusunda uyarlama ve seçici aktarım yapıldığını göstermektedir.

Alan becerilerinin belirlenmesinde alanların kendi yapısal özelliklerinden kaynaklanan farklılıkların ortaya çıkması doğal bir durumdur. Örneğin Türkçe alan becerileri; anlama (dinleme/izleme ve okuma) ve anlatma (konuşma ve yazma) olmak üzere dört temel beceriden oluşmaktadır. Türkçe alan becerileri oluşturulurken diğer alanlarda olduğu gibi kavramsal beceriler ile alana özgü bütünleşik beceriler birlikte kullanılmıştır. Buna ek olarak sosyal duygusal öğrenme becerileri içinde yer alan benlik becerisinin bir alt boyutu olan kendine uyarlama (öz-yansıtma) becerisinin de Türkçe alan becerilerinde bütünleşik beceri olarak yer aldığı görülmektedir. Bu bölüm kapsamında odak, alan ve kavramsal becerilere yöneltildiğinden diğer beceri alanlarına ilişkin ayrıntılı açıklamalara ayrıca yer verilmemiştir.

3.3. TYMM Öğretim Programları ve Yapısı

Öğretim programları; hedef, içerik, öğrenme-öğretme süreci ve değerlendirme boyutlarından oluşan bütüncül bir sistem olarak tasarlanır (Ornstein & Hunkins, 2018). Bu sistem statik bir yapı değildir; toplumsal değişimler, teknolojik dönüşümler ve pedagojik yeniliklere uyum sağlayabilmek için sürekli olarak güncellenmesi gereken dinamik bir yapıya sahiptir (Demirel, 2020). Güncel eğitim yaklaşımları, öğretim programı yenilemenin yalnızca içeriği güncellemeyi değil, aynı zamanda öğretim stratejilerinin, öğrenme ortamlarının ve ölçme ve değerlendirme süreçlerinin de çağın gerektirdiği biçimde yeniden yapılandırılmasını zorunlu gördüğünü ortaya koymaktadır

(Darling-Hammond, 2017). Bu nedenle becerilerin öğretim programlarına entegrasyonu, öğretim programlarının tüm bileşenlerinde sistematik ve çok boyutlu bir dönüşüm gerektiren kapsamlı bir süreçtir.

Türkiye'de bu dönüşümün somut bir yansıması olarak Talim ve Terbiye Kurulu Başkanlığı tarafından 23 Mayıs 2024 tarihli, 20 sayılı karar ile TYMM'ye göre hazırlanan öğretim programları yürürlüğe girmiştir (MEB, 2024). Önceki öğretim programlarından farklı olarak TYMM'ye göre hazırlanan öğretim programlarında taksonomik bir sınıflama yerine, TYMM Öğretim Programları Ortak Metni'nde (MEB, 2025) tanımlanan beceri çerçevesi esas alınmıştır. TYMM öğretim programlarının yapısal kurgusu incelendiğinde, önceki öğretim programlarından farklı bir düzenleme ile karşılaşılmaktadır. TYMM öğretim programları;

(i) öğrenme çıktıları ve süreç bileşenleri,

(ii) içerik çerçevesi,

(iii) öğrenme kanıtları (ölçme ve değerlendirme),

(iv) öğrenme-öğretme yaşantıları (temel kabuller, ön değerlendirme süreci, köprü kurma, öğrenme-öğretme uygulamaları),

(v) farklılaştırma (zenginleştirme ve destekleme) ve

(vi) öğretmen yansıtmaları

şeklinde altı temel bölümden oluşmaktadır.

Öne çıkan değişikliklerden biri, öğrencilere kazandırılması hedeflenen bilgi ve becerilerin terminolojik açıdan yeniden tanımlanmasıdır. 2005 yılı öncesinde "hedef" veya "hedef davranış" olarak ifade edilen bu unsurlar 2005 sonrasında "kazanım" terimiyle anılmış, TYMM öğretim programlarında ise "öğrenme çıktısı" kavramıyla yapılandırılmıştır. Öğrenme çıktıları ve süreç bileşenleri içerik çerçevesi ile TYMM Öğretim Programları Ortak Metni'nde tanımlanan kavramsal ve alan beceriler ve bunlara ilişkin süreç bileşenlerinden oluşturulmuştur. Böylelikle öğrenme çıktıları, daha sistematik ve izlenebilir bir şekilde ortaya konmuştur.

3.4. TYMM Öğretim Programlarında Ölçme ve Değerlendirme

Önceki öğretim programlarında ölçme ve değerlendirme, "Öğretim Programlarında Ölçme ve Değerlendirme Yaklaşımı" başlığı altında genel ilkeler düzeyinde sunulurken TYMM öğretim programlarında bu alan hem TYMM Öğretim Programları Ortak Metni'nde "Öğrenme Kanıtları" başlığı altında hem de her dersin öğretim programlarının "Uygulama Esasları" bölümünde "Ölçme ve Değerlendirme Uygulamaları" başlığıyla ayrıntılı biçimde ele alınmıştır. Ayrıca her ders ve sınıf düzeyinde ünite, tema ve öğrenme alanları için "öğrenme kanıtları (ölçme ve değerlendirme)" başlığı altında, ilgili öğrenme çıktılarının değerlendirilmesine yönelik ölçme ve değerlendirme uygulamalarına dair

öneriler sunulmuştur. Öğrenme kanıtlarında her öğrenme çıktısı için en az bir öğrenme kanıtına yer verilmiş; bazı durumlarda ise bir öğrenme çıktısı için birden fazla öğrenme kanıtı önerilmiştir.

TYMM öğretim programlarında tanılayıcı, biçimlendirici ve düzey belirleyici (sonuç) değerlendirme türleri arasında dengeli bir yaklaşım benimsendiği söylenebilir. Öğrenme yaşantıları bölümünde her ünite, tema veya öğrenme alanı için öğrencilerin ön bilgilerini belirlemeye yönelik tanılayıcı değerlendirme araçlarına yer verilmiştir. Böylece öğretim sürecinin başlangıcında öğrencilerin hazır bulunuşluk düzeylerinin belirlenmesi amaçlanmıştır. TYMM Öğretim Programları Ortak Metni'nde (MEB, 2025) beceri odaklı sonuç değerlendirme uygulamalarına da vurgu yapılmakta; öğretim programlarının çıktılarının beceri gelişimi temelinde izlenmesi gerektiği belirtilmektedir. Bununla birlikte aynı metinde öğretim programlarının geliştirilme sürecinde biçimlendirici değerlendirmenin öne çıkarıldığı açık biçimde ifade edilmektedir (MEB, 2025).

Bu çerçevede TYMM'nin biçimlendirici değerlendirmeyi öğretim programlarının temel esaslarından biri olarak benimsemesi, öğrenmenin sürekli izlenmesine ve geliştirilmesine dayalı çağdaş ölçme ve değerlendirme anlayışıyla örtüşmektedir. Biçimlendirici değerlendirme, öğrenme-öğretme sürecinin ayrılmaz bir parçasıdır. Öğrencilerin öğrenme durumlarını planlı ve sürekli biçimde izlemeyi, öğretimi de elde edilen verilere göre düzenlemeyi amaçlar. Leahy ve arkadaşları (2005), biçimlendirici değerlendirmeyi, öğretim süreci boyunca saniyesi saniyesine, dakikası dakikasına, saati saatine, günü gününe ve haftası haftasına öğrenmeye dair kanıt toplama süreci olarak tanımlamaktadır. Bu yaklaşım, ölçme ve değerlendirmeyi yalnızca öğrenmenin sonunda gerçekleştirilen bir kontrol işlevi olmaktan çıkararak sürecin her aşamasına yayılan, öğrencilerin gelişimini yönlendiren ve öğrenmeye rehberlik eden bir araç hâline getirmektedir (Penuel & Shepard, 2016). Öğretim sırasında elde edilen öğrenme kanıtları, öğretim sürecinin öğrencilerin ihtiyaçlarına uygun şekilde uyarlanmasına olanak tanımaktadır. Öğrencilerin yanıtları, ürünleri veya performansları bu süreçte incelenerek öğretmenin öğretimi yeniden yapılandırmasına imkân vermektedir. Böylece öğretim süreci daha esnek, öğrenci merkezli ve öğrenme odaklı bir yapıya kavuşmaktadır. Bu yönüyle biçimlendirici değerlendirme, yalnızca mevcut öğrenme düzeyini belirlemekle kalmamakta aynı zamanda gelecekteki öğrenme yaşantılarının şekillendirilmesinde de önemli bir rol üstlenmektedir (Shepard, 2019).

3.4.1. Öğrenme Kanıtları

Geleneksel ölçme ve değerlendirme anlayışında, çoğunlukla sürecin sonunda, öğrenme-öğretme yaşantılarının ardından gerçekleştirilen bir uygulama olarak konumlandırılmıştır. TYMM'de ise Wiggins ve McTighe'nin (2005) ortaya koyduğu tersine tasarım anlayışına benzer bir anlayışla, ölçme ve değerlendirme "Öğrenme

Kanıtları" başlığı altında sürecin başında, öğrenme-öğretme yaşantılarından önce yer almaktadır. Bu anlayış, öğretim sürecinin baştan itibaren öğrenme çıktılarıyla uyumlu şekilde tasarlanmasına imkân tanımaktadır.

Öğrenme kanıtlarının önceden belirlenmesi "Öğrencinin hedeflenen öğrenme çıktısına ulaştığını nasıl anlayabiliriz?" sorusuna sistematik bir yanıt aramayı kolaylaştırmaktadır. Böylece hem ölçülecek bilgi ve beceriler netleşmekte hem de öğrenme süreci bu hedeflere ulaşmayı destekleyecek biçimde yapılandırılmaktadır. TYMM öğretim programlarında öğrenme kanıtlarında önerilen ölçme ve değerlendirme uygulamalarının öğrenme-öğretme süreçlerine doğal akış içinde bütünleştirilerek öğretmen-öğrenci etkileşiminin ayrılmaz bir bileşeni hâline getirilmesi öngörülmüştür. Değerlendirmenin öğretimden ayrı bir süreç olarak ele alınmasından öğretimsel bütünleşmeye geçiş olarak tanımlanan bu yapı, ölçme ve değerlendirmenin yalnızca öğretim programının tamamlayıcı bir unsuru değil aynı zamanda öğrenme sürecinin merkezî bir parçası olduğunu göstermektedir. Bu bütüncül yaklaşım "pedagojik üçlü" olarak adlandırılan öğretim-öğrenme-değerlendirme sistemi içinde ölçme ve değerlendirmenin rolünü yeniden tanımlamakta ve onu öğretim programlarının merkezî bir bileşeni hâline getirmektedir (Gordon Commission on the Future of Assessment in Education [Gordon Komisyonu], 2013).

Birden fazla kanıtı içeren bir sistemde, her kanıt türü kendi içinde belirli bir işlev üstlenmekte ve farklı türdeki kanıtlar birbirini tamamlayarak öğrencilerin öğrenme çıktılarına ne ölçüde ulaştıklarına ilişkin bütüncül bir tablo sunmaktadır. TYMM öğretim programlarında öğrenme çıktılarının oluşturulmasında bilgi ile birlikte becerileri de merkeze alan bir anlayış benimsenmiştir. Öğrenme kanıtları bölümünde her öğrenme çıktısı için en az bir öğrenme kanıtı önerilmiş ayrıca ölçme ve değerlendirme yöntemlerinde çeşitliliğe yer verilmiş, performans görevleri de ön plana çıkarılmıştır. Bu çerçevede performansa dayalı değerlendirmeler önemli bir konuma sahiptir. Öğrencilerden bir yanıt ya da ürün oluşturmalarını veya açık uçlu görevleri yerine getirmelerini gerektiren bu değerlendirmeler, becerilerin ölçülmesinde temel araçlardan biri olarak kabul edilmektedir. TYMM öğretim programlarında benzer biçimde performans görevleri ana kanıt türü olarak ele alınmıştır. Her ders ve sınıf düzeyinde, her ünite, tema veya öğrenme alanında en az bir performans görevinin uygulanması esas alınmıştır.

TYMM'de benimsenen bu yaklaşım, geleneksel "madde paradigması"ndan "etkinlik paradigması"na geçişin açık bir göstergesi olarak değerlendirilebilir. Etkinlik temelli ölçme anlayışı, öğrencilerin gerçek yaşam durumlarına benzer karmaşık görevlerdeki performanslarını gözlemlemeyi mümkün kılmaktadır (DiCerbo & Behrens, 2012). Böylece ölçme, yalnızca doğru yanıta ulaşmayı değil aynı zamanda öğrencilerin becerileri ne ölçüde kullanabildiklerini de değerlendirme kapsamına almaktadır. TYMM'de performans görevlerinin her ünite, tema ve öğrenme alanında zorunlu

tutulması, bu anlayışın somutlaşmış bir örneği olarak değerlendirilebilir. Bu noktada TYMM öğretim programlarındaki ölçme ve değerlendirme uygulamalarının genel çerçevesini ortaya koymak gerekmektedir.

3.4.2. Ölçme ve Değerlendirme Uygulamaları

Eğitimciler, öğrencilerin ne bildiklerini ve ne yapabildiklerini anlamak amacıyla ölçme ve değerlendirme yapar ancak hiçbir ölçme yöntemi öğrencinin zihnine doğrudan erişim sağlamaz. Eğitimde yapılan ölçmeler, boy ya da kilo ölçümü gibi doğrudan değildir ve bu nedenle ölçme hatasına daha açık bir yapı sergiler. Zira ölçülmek istenen özellikler çoğu zaman doğrudan gözlemlenemeyen bilişsel süreçler ve zihinsel temsilleri içerir. Bu nedenle ölçme araçları, öğrencilerin davranış ve performanslarından hareketle bilgi ve beceri düzeyleri hakkında çıkarım yapmaya elverişli veriler üretecek biçimde tasarlanır. Ölçme kadar ölçme sonuçlarına dayalı değerlendirme yapmak başka bir anlatımla "neyin, nasıl değerlendirileceğine" karar vermek karmaşık bir süreçtir. Bu nedenle de eğitim ortamında yapılacak ölçme ve değerlendirme uygulamaları öncesinde ölçülecek yapının doğası, bağlamı ve hedeflerinin iyi bir şekilde tanımlanması ve buna uyumlu biçimde yapılandırılması gerekir.

Becerilerin öğretim programlarında yer alması, bu becerilerin ne ölçüde edindirildiğinin güvenilir, geçerli ve adil biçimde değerlendirilmesini zorunlu kılar. Ancak becerilerin ölçülmesi bilgi ve kavram gibi daha temel yapıları ya da alt zihinsel bileşenleri değerlendirmeye kıyasla daha güçtür. Bu güçlüklerin temelinde, becerilerin çok boyutlu ve gelişimsel yapıları ile gözlenen davranışın bilgi mi yoksa beceri kaynaklı mı olduğunun ayırt edilmesindeki zorluklar yer alır (Care & Kim, 2018). Bu bağlamda "TYMM'de tanımlanan beceri çerçevesini ve öğrenme çıktılarını yansıtan özgün görevler nasıl tasarlanmalı, beceriler anlamlı ve etkili biçimde nasıl ölçülmeli, öğrenci performansı hangi ölçütlere göre puanlanmalı ve bu sonuçlar öğrenme-öğretim sürecine nasıl dahil edilmeli?" gibi sorular beceri temelli değerlendirmenin güncel tartışmalarının odağındadır.

TYMM bağlamında ölçme ve değerlendirmenin hem hesap verebilirlik odaklı sonuç değerlendirme süreçlerini hem de öğrenmeye rehberlik eden biçimlendirici değerlendirme uygulamalarını dengeli biçimde içermesi esastır. Beceriler söz konusu olduğunda ölçme süreci, kullanılan aracın niteliği kadar ölçmenin amacına göre de farklılaşmaktadır. Bu nedenle bir becerinin değerlendirilmesinde hangi araçların kullanılacağı; ölçmenin izleme, yönlendirme ya da düzey belirleme amacı taşıyıp taşımadığına bağlı olarak değişebilir. Bununla birlikte amaç ve bağlam ne olursa olsun tüm ölçme ve değerlendirme uygulamalarının geçerlik, güvenirlik ve adillik ilkelerini karşılaması zorunludur; aksi durumda elde edilen sonuçlar hem öğrenme süreçleri hem de karar mekanizmaları açısından yanıltıcı olabilir.

Becerilerin ölçülmesine ilişkin çalışmalar (ör., Griffin ve ark., 2012; Gordon Komisyonu, 2013; Ercikan & Oliveri, 2016) değerlendirme yaklaşımının sadece sonuç odaklı modelden, öğrenme süreçlerini destekleyen ve eğitimin niteliğini geliştirmeyi amaçlayan bir modele kaydığını göstermektedir. ATC21S çerçevesi, becerilerin büyük ölçüde performansa dayalı araçlarla ölçülmesini; ölçme ve değerlendirmenin kullanışlı geri bildirim üretmesini ve öğretmen-öğrenci kapasitesinin gelişimine katkı vermesini vurgular (Griffin ve ark., 2012). Benzer biçimde Gulikers ve arkadaşları (2004), odağın becerilerde olmasının nedenini, öğrencilerin bu becerileri gerçek yaşam bağlamlarında uygulayabilme kapasitesini geliştirmek olarak tanımlar. Pellegrino ve Hilton (2012), 21. yüzyıl yeterliklerini transfer edilebilir bilgi ve beceriler olarak tanımlamakta, bir alanda öğrenilenlerin yeni durum ve problemlere aktarılabilmesini derin öğrenmenin göstergelerinden biri olarak değerlendirerek değerlendirmelerin yalnızca ürüne değil, öğrenenlerin neyi anladıkları ve ne yapabildiklerine ilişkin daha derin bilgi üretmeye odaklanması gerektiğini belirtmektedir. Blömeke ve arkadaşları (2015) ise yeterliği, bireyin sahip olduğu yatkınlıklar (bilgi, beceri, tutum) ile gözlenen performans arasında uzanan ve durumu algılama, yorumlama ve karar verme süreçlerinin aracılık ettiği bir süreklilik olarak modellemektedir. Bu nedenle ölçmenin tek tür araca dayanmak yerine bu sürekliliğin farklı noktalarını yakalayacak biçimde çeşitlendirilmesi gerektiğini savunmaktadır. Dolayısıyla beceri ölçümü yalnızca sonuca odaklanmamalı, öğrenme eksikliklerini görünür kılan, transferi teşvik eden ve düzey belirleyici olduğu kadar biçimlendirici nitelik taşımalıdır (Blömeke ve ark., 2015; Pellegrino & Hilton, 2012).

Ölçme sürecinin ilk ve en önemli adımı, ölçülmek istenen yapının açık bir şekilde tanımlanmasıdır. Bu tanım, görev veya madde geliştirme süreçlerinde ortak bir anlayış zemini oluşturur. TYMM'de ölçülmesi hedeflenen yapılar "öğrenme çıktıları" aracılığıyla işlevsel hâle getirilmiştir. Bu çıktılar, bilgi ve becerinin bütünleşik biçimde sergilenmesini gerektiren performans ifadeleri olarak tanımlanmıştır. Ancak yalnızca çıktılar değil bu çıktılara ulaşmayı sağlayan süreç bileşenleri de değerlendirmeye dahil edilmelidir. Beceriler süreç odaklı yapılar olduğu için öğrencinin nihai performansı kadar o performansa götüren işlem adımları da değerlendirme kapsamına alınmalıdır. Bu nedenle TYMM'de geliştirilecek araçların hem öğrenme çıktılarının içeriğini hem de süreç bileşenlerini kapsayacak biçimde tasarlanması beceri odaklı ölçmenin temelini oluşturur.

Becerilerin ölçümüne ilişkin önemli bir diğer tartışma, değerlendirmede alana özgü (domain-specific) mü yoksa alandan bağımsız (domain-general) görev ve maddelerin mi kullanılacağıdır. Bilişsel becerilerin bağlama bağlı olup olmadığı sorusu alan yazında uzun süredir tartışılmaktadır (Perkins & Salomon, 1989). Alandan bağımsız yaklaşımı savunanlar, yaratıcı düşünme, problem çözme ve eleştirel düşünme gibi becerilerin belirli alanlara bağlı olmadığını ileri sürmektedir (Ennis, 1989; Sternberg, 2003). Bu yaklaşımda gerekli içerik bilgisi görev içinde kontrollü biçimde sunulur ve amaç,

içerikten olabildiğince bağımsız olarak becerinin kendisini ölçmektir. Buna karşılık alana özgü yaklaşımı savunanlar, genel bilişsel yetenekler rol oynasa da üst düzey performans için alana özgü bilgi ve uzmanlığın belirleyici olduğunu, dolayısıyla beceri ile bilginin iç içe geçtiğini ve bu becerilerin içerikten bağımsız ölçülemeyeceğini vurgulamaktadır (Baer, 2012; Tricot & Sweller, 2014). Üçüncü bir konum ise becerilerin hem alandan bağımsız hem de alana özgü bileşenleri olduğunu kabul eden karma yaklaşımdır (Baer & Kaufman, 2005; Plucker & Beghetto, 2004). Bu tartışmanın ölçme açısından doğrudan sonuçları vardır: Alandan bağımsızlık iddiası, puanların bağlamlar arası genellenebileceği iddiasıdır ve transfer kanıtı gerektirir. Oysa performans değerlendirme araştırmalarının yerleşik bulgularından biri, görev örnekleme varyansının büyük olduğu, yani bir görevdeki performansın diğer görevlere sınırlı biçimde genellendiğidir (Shavelson ve ark., 1993). Bu nedenle beceri ölçümünde tek bir göreve dayalı puanlardan alan-genel çıkarımlar yapmak geçerlik açısından risklidir.

Hem alan-bağımsız hem de alan-özgü yaklaşımlar, becerilerin ölçümünde bazı sınırlılıklar taşımaktadır. Alan-bağımsız görevler, gerçek yaşamın gerektirdiği içerik ve bağlamdan uzaklaştığında becerilerin günlük yaşamdaki kullanımıyla ilişki kurmak güçleşmektedir. Alan-özgü değerlendirmelerde ise performanstaki yetersizliğin alan bilgisi eksikliğinden mi yoksa hedeflenen becerinin yeterince gelişmemiş olmasından mı kaynaklandığını ayırt etmek her zaman kolay olmamaktadır (Ercikan & Oliveri, 2016). Bu nedenle beceri puanlarının farklı alanlara genellenebilmesi, ayrıca kanıtlanması gereken bir geçerlik sorunu olarak ele alınmalıdır. Alan-genel problem çözme becerilerinin önemini vurgulayan çalışmalar da bu becerilerin alan-özgü bilgi ve stratejilerle birlikte değerlendirilmesi gerektiğine işaret etmektedir (Greiff ve ark., 2014). ATC21S kapsamında geliştirilen görevler ise karmaşık becerilerin ölçümünde görev tasarımının ve süreç verilerinin önemini göstermektedir. Ancak bu tür değerlendirmelerde ön bilgi, görev bağlamı ve alan bilgisi etkisinin dikkatle kontrol edilmesi gerekmektedir (Care ve ark., 2015). Benzer biçimde Avustralya Eğitim Araştırmaları Konseyi (ACER) tarafından geliştirilen beceri geliştirme çerçevelerinde, becerilerin genel düzeyde tanımlanabileceği, buna karşın becerinin uygulanmasının öğrenme alanı ve disiplin bağlamından bağımsız düşünülemeyeceği vurgulanmaktadır (Scoular, 2025). TYMM'de öğrenme çıktıları bilgi-beceri bütünlüğü ile tanımlandığından becerilerin değerlendirilmesinde alana-özgü yaklaşım genel olarak daha uyumlu olduğu ifade edilebilir. Değerlendirmede yalnızca becerinin sergilenme düzeyi değil bu becerinin hangi bilgi temeli üzerinde nasıl yapılandırıldığı da dikkate alınmalıdır.

Öğrenme çıktılarının bir parçası olan bilgi ve becerilerin gerçek yaşam durumlarında kullanılmasını esas alan değerlendirme anlayışı, yapı geçerliğine ilişkin kanıtları güçlendirebilir. İçyapıya dayalı kanıtlar, bir değerlendirmenin ölçmeyi amaçladığı yapıyı ne ölçüde temsil ettiğiyle ilgilidir. TYMM bağlamında bu durum, görev ve maddelerde kullanılan bağlamların ölçülmek istenen beceriyi gerçekçi biçimde temsil etmesini,

görevlerin gerçek yaşam koşullarıyla ilişkilendirilmesini ve ilgili beceri ile süreç bileşenlerinin değerlendirmeye dâhil edilmesini gerektirmektedir. Otantiklik, ölçülen becerinin gerçek hayattaki kullanım durumlarına yakınlık düzeyini ifade etmektedir. Ancak otantiklik, yapı geçerliğini tek başına garanti etmemekte; görev tasarımı, puanlama ölçütleri ve yapılacak puan yorumlarıyla birlikte ele alındığında geçerlik kanıtlarını desteklemektedir (Birenbaum, 2003). Akademik dürüstlük açısından da otantik görevlerin tek başına yeterli bir güvence sağlamadığı, özellikle yapay zekâ araçlarının yaygınlaşmasıyla değerlendirme tasarımının süreç kanıtları, sözlü savunma, ürün geliştirme aşamaları ve bağlama özgü öğrenci üretimleriyle desteklenmesi gerektiği vurgulanmaktadır (Ellis ve ark., 2020; Kofinas ve ark., 2025). Bu nedenle TYMM'de performans görevleri, açık uçlu maddeler ya da çoktan seçmeli maddeler kullanılsa da görev ve bağlamların gündelik yaşamla ilişkilendirilmesi; beceri yapısına uygunluk, yapı temsili ve öğrenme yaşantılarını destekleme açısından önemli bir ilke olarak değerlendirilmektedir.

Değerlendirme tasarımı; amaç (öğrenmeyi destekleme, bireysel başarıyı belirleme, program etkililiğini izleme), uygulama bağlamı (sınıf içi değerlendirmeler, geniş ölçekli sınavlar) ve uygulama koşulları (zaman, kaynak, araçlar) tarafından belirlenir. Bu değişkenler, ölçme ve değerlendirmenin içeriğini ve biçimini şekillendirir. Bu bölümde TYMM bağlamında becerilerin ölçülmesi, sınıf içi ve geniş ölçekli değerlendirmeler olmak üzere iki başlıkta ele alınmıştır.

3.4.3. Sınıf İçi Ölçme Değerlendirme Uygulamaları

TYMM öğretim programları, ölçme ve değerlendirme süreçlerini ağırlıklı olarak sınıf içi uygulamalarla bütünleştiren bir anlayışla tasarlanmıştır. TYMM öğretim programlarında ölçme ve değerlendirmenin temel amacı öğrencilere yönelik öğrenme süreçlerini izlemek, yönlendirmek ve öğretimi geliştirmektir. TYMM öğretim programlarında yer alan değerlendirme önerileri; öğretmenlere sınıf içi değerlendirmeleri planlama, yürütme ve öğrenmeyi destekleyecek şekilde yorumlama konusunda rehberlik etmektedir. Bu bölümde TYMM öğretim programlarında yer alan becerilerin sınıf içi ortamda değerlendirilmesine yönelik temel ilkeler ve dikkat edilmesi gereken noktalar açıklanmıştır.

3.4.4. Temel Öğrenme Kanıtı Olarak Performans Görevleri ve Değerlendirilmesi

TYMM'de performans görevleri, biçimlendirici değerlendirme yaklaşımıyla ele alınmakta ve işlevi, yalnızca öğrencinin mevcut düzeyini ölçmekten öteye taşımaktadır. Öğrencilerin ürün, yanıt ve performanslarının süreç içinde incelenmesi; öğretmene öğretimi yeniden düzenleme, öğrenciye ise hedeflenen öğrenme çıktısına ulaşması için zamanında ve eylem odaklı geri bildirim verme olanağı sağlamaktadır. Böylece performans görevleri, öğrenmenin sonunda uygulanan sonuç odaklı bir kontrol mekanizması olmaktan çıkar ve öğrenme sürecine dahil edilerek öğrencinin gelişimini

sürekli destekleyen bir araç hâline gelir. TYMM'de performans görevlerinin her ünite, tema ve öğrenme alanında zorunlu tutulması, biçimlendirici değerlendirmenin programın merkezine yerleştirildiğini göstermektedir. Bu nedenle öğretim programlarında önerilen performans görevleri; okul dışında yapılacak bir görev olmaktan ziyade, öğrenme-öğretme süreciyle bütünleşik bir sınıf içi etkinlik olarak tasarlanmalıdır. Dolayısıyla görevlerin mümkün olduğunca sınıf içi ortamda yürütülmesi beklenmektedir. Öğretim programlarında önerilen performans görevleri örnek niteliğinde sunulmuştur; okulların fiziki imkânları ve öğrencilerin hazır bulunuşlukları doğrultusunda eş değer ve uygun görevler geliştirilebilir. Ancak temel ilke, performans görevlerinin "ilginç bir etkinlik" olmanın ötesinde belirli öğrenme çıktılarının ve ilgili süreç bileşenlerinin ölçülmesine hizmet eden araçlar olarak kurgulanmasıdır. Performans görevleri ile ilgili detaylı bilgiler için kitabın 7. bölümünü inceleyiniz.

Performans görevlerinin sistematik ve güvenilir biçimde değerlendirilebilmesi uygun ölçme araçlarının (özellikle dereceli puanlama anahtarlarının) kullanımını gerektirmektedir. Performans değerlendirmelerinde, karmaşık becerilere ilişkin gözlemlerin geçerli ve güvenilir şekilde yapılıp yapılamayacağı başlıca güçlük alanıdır. Bu noktada alan yazın, performans görevlerinin çoğunlukla dereceli puanlama anahtarı ile değerlendirilmesini önermektedir (Jonsson & Svingby, 2007). Bu yaklaşım, söz konusu araçların tasarımı, açıklığı ve kullanım yetkinliğini önemli hâle getirmektedir. TYMM'de önerilen öğrenme kanıtları incelendiğinde en sık kullanılan aracın performans görevi, bu görevlerin değerlendirilmesinde de çoğunlukla dereceli puanlama anahtarları olduğu görülmektedir (Aşiret & Su Özenir, 2025).

Dereceli puanlama anahtarları bir çalışmadan beklenen performansı tanımlayan ölçütler ile bu ölçütlere ilişkin performans düzeyi tanımlarından oluşur (Reddy & Andrade, 2010). Bu araçlar, öğretmen ve öğrenciye açık, yapılandırılmış ve tutarlı bir değerlendirme çerçevesi sunmakta; performans değerlendirmelerinin psikometrik özelliklerini güçlendirirken öğrencinin gelişim sürecini izlemeye ve öğrenmeyi desteklemeye de katkı vermektedir (Wiliam, 2011). Bu nedenle dereceli puanlama anahtarlarının, öğrenme çıktıları ve süreç bileşenlerini kapsayan ölçütler üzerinden hazırlanması yalnızca puanlama amacıyla değil öğrencilerin öz-düzenleme becerilerini destekleyen geri bildirim aracı olarak kullanılması önemlidir (Andrade, 2010). Brookhart (2018), özellikle genel amaçlı dereceli puanlama anahtarlarında, düzey tanımlarında "iyi/zayıf/mükemmel" gibi yargılayıcı etiketler yerine betimsel dil (ör. başlangıç, geliştirilmeli, yeterli, örnek düzeyde) kullanılmasını önermektedir. Betimsel dil, öğrencilerin mevcut durumlarını ve ulaşmaları gereken hedefleri daha net görmelerini sağlar. TYMM'de de bu doğrultuda, öğrencilerin becerileri edinim sürecinde gösterdikleri ilerlemeye yönelik geri bildirim verilmesi esastır. Uygulamada, analitik dereceli puanlama anahtarları biçimlendirici değerlendirme; bütünsel dereceli puanlama anahtarları ise daha çok düzey belirlemeye yönelik değerlendirme için

uygundur (Brookhart, 2018). TYMM'de öğrenme çıktılarının gözlenebilir süreç bileşenleri hâlinde tanımlanması, her bir bileşenin ayrı bir ölçüt olarak puanlanmasına ve öğrenciye bileşen düzeyinde geri bildirim sunulmasına olanak tanıdığından, model yapısı gereği analitik dereceli puanlama anahtarlarının kullanımına daha elverişlidir. Performans görevleri, biçimlendirici değerlendirmenin temel araçlarından biridir. Bununla birlikte çok yönlü kanıt toplamak için farklı araçlardan da yararlanılmalıdır. Bu nedenle aşağıda performans görevlerinin yanı sıra kullanılabilecek diğer kanıt türleri açıklanmıştır.

3.4.5. Diğer Kanıtlar

Öğrenci öğrenmesine ilişkin kanıt toplama sürecinde, elde edilen performansın, öğrencinin tipik yeterlik düzeyini mi yansıttığı yoksa tesadüfî bir duruma mı bağlı olduğunun ortaya konulması önemlidir. Bu nedenle öğretimle ilgili kararlar alınmadan önce birden fazla kaynaktan kanıt elde edilmelidir. Farklı kaynaklardan toplanan veriler ne kadar çeşitlilik gösterirse, öğrenmeye ilişkin çıkarımların güvenilirliği de o ölçüde artar (Griffin ve ark., 2010). Tüm ölçme ve değerlendirme uygulamaları esasen ilgili alanın bir örneklemesidir; bu alanın yeterince temsil edilmesi beklenebilir ancak hiçbir ölçme ve değerlendirme uygulaması ölçülecek alanı tamamen kapsayamaz (Messick, 1989). Bu nedenle yalnızca tek bir araca bağlı kalmak öğrencilerin öğrenme çıktılarına ilişkin sınırlı bir bakış sunar.

Dolayısıyla TYMM öğretim programlarında temel kanıt olarak ele alınan performans görevleriyle birlikte çalışma yaprakları, öz değerlendirme, çıkış biletleri, açık uçlu maddeler gibi farklı ölçme ve değerlendirme araçlarından elde edilen verilerin birlikte değerlendirilmesi, öğrenmenin çok boyutlu yapısını daha doğru yansıtır. Bu yaklaşım, TYMM ortak metninde yer alan yöntem çeşitliliği ilkesine de uygundur. Öğretim programlarında performans görevleri ana kanıt olmakla birlikte ek kanıt türleri de kullanılmıştır. Örneğin Aşiret ve Su-Özenir (2025), ortaöğretim düzeyinde yedi dersin öğretim programlarını ölçme ve değerlendirme ilkeleri açısından incelemiş ve toplam 84 farklı öğrenme kanıtının kullanıldığını raporlamıştır. Ancak çeşitlilik tek başına yeterli değildir. Kullanılan kanıtların öğrenme çıktılarıyla doğrudan ilişkili olması, öğrencilerin gelişim düzeyine uygunluğu ve biçimlendirici değerlendirme yaklaşımıyla uyumlu biçimde kullanılması büyük önem taşır. Ölçme ve değerlendirme araçları doğrudan tasarımlarına göre biçimlendirici ya da değer biçmeye (sonuç) yönelik olmaz; asıl belirleyici olan kullanım amacıdır. Örneğin, öğrencilere yöneltilen açık uçlu sorular, not verme amacıyla kullanıldığında değer biçmeye yönelik değerlendirmeye hizmet ederken; aynı sorular grup çalışması kapsamında akran değerlendirmesi ve geri bildirim üretmek için kullanıldığında biçimlendirici bir işlev kazanır. Bu bağlamda biçimlendirici değerlendirmenin etkili olabilmesi için öğrencilerin sürece etkin biçimde katılması gerekmektedir. Öz ve akran değerlendirme gibi çeşitlendirilmiş kanıt toplama süreçleri, öğrencileri sadece verinin nesnesi değil, değerlendirme sürecinin aktif bir öznesi kılarak biçimlendirici değerlendirmenin etkisini artırmaktadır.

3.4.6. Öğrencilerin Değerlendirme Sürecine Aktif Katılımı

Biçimlendirici değerlendirme yalnızca öğretmen tarafından yürütülen bir süreç değil öğrencilerin etkin katılımını içeren çok boyutlu bir yaklaşım olarak ele alınır. Özellikle öz değerlendirme ve akran değerlendirme aracılığıyla öğrencilerin sürece dahil edilmesi onları pasif alıcı olmaktan çıkarıp aktif özne hâline getirir (Black & Wiliam, 2009). Öz değerlendirme; öğrencilerin kendi bilgi, beceri ya da performanslarını açık ölçütlere göre izleyip değerlendirdikleri ve bu yargıları öğrenmeyi geliştirmeye yönelik eylemlerle bütünleştirdikleri bir süreçtir (Andrade, 2019). Bu uygulamalar, basitçe kendi çalışmasına not vermeden öğretmen ya da akran geri bildiriminin de dahil edildiği daha karmaşık tasarımlara kadar farklı biçimlerde yürütülebilir (Panadero ve ark., 2016). Akran değerlendirme benzer statüdeki öğrenenlerin önceden belirlenmiş ölçütlere göre birbirlerinin çalışmalarını veya performanslarını değerlendirdiği bir süreçtir (Topping, 1998). Bu süreç; öğrencilerin değerlendirme ölçütlerini anlamalarına (Panadero & Jonsson, 2013), kanıta dayalı değerlendirme kararları almalarına (Nicol & Macfarlane-Dick, 2006), değerlendirme yargısı geliştirmelerine (Carless & Boud, 2018) ve akademik başarılarını artırmalarına (Yan ve ark., 2022) yardımcı olur. Ayrıca öz ve akran değerlendirme; öğrencilerin sosyal-duygusal gelişimini, özerkliğini ve öz-düzenleme becerilerini destekler (Li ve ark., 2020).

Öğrencilerin değerlendirme sürecine aktif katılımı, TYMM öğretim programlarının da temel vurgularından biridir. Bu doğrultuda programlarda; dereceli puanlama anahtarları, dereceleme ölçekleri, kontrol listeleri, öz/akran/grup değerlendirme formları, KWL formu (bildiklerim, bilmek istekilerim, öğrendiklerim), çıkış biletleri, öğrenme günlükleri gibi öz-yansıtma odaklı ölçme ve değerlendirme uygulamalarına yer verilmiştir. Bu araçların etkililiği için, öğretmen-öğrenci etkileşimindeki rollerin karşılıklı olarak anlaşılması gerekir. Öğretmen, öğrenme süreçlerine ilişkin kanıtları toplar ve bu kanıtlara göre öğretimi uyarlamakla yükümlüdür. Öğrenci ise yanıt, ürün ve yansımalar üreterek sürece aktif katılım sağlar. Ortak anlayış gelişmediğinde öğrencilerin öğretmene bağımlı kalma riski artar; bu da özerk, yaşam boyu öğrenen birey olma potansiyelini sınırlar (Heritage, 2010). Bu nedenle öğrencilerin öğrenme çıktıları ile bu çıktılara ulaşma ölçütlerini içselleştirmelerini ve öğrenmelerine ilişkin yargı geliştirmelerini destekleyen bir öğrenme ortamı inşa edilmelidir. Ayrıca, öğrencilerin bu süreçte "acemi" oldukları gözetilerek sistematik yansıma yapabilmeleri için yeterli zaman ve fırsat tanınmalıdır.

4. TYMM'DE YER ALAN BECERİLERİN GENİŞ ÖLÇEKLİ TEST UYGULAMALARINDA ÖLÇÜLMESİ

TYMM, becerilerin öncelikle sınıf içi öğrenme-öğretme süreçlerinde geliştirilmesini ve ölçülmesini önceleyen bir yaklaşım sunar. Bununla birlikte Türkiye'de Liselere Geçiş Sistemi (LGS), Yükseköğretim Kurumları Sınavı (YKS) gibi geniş ölçekli testlere çok sayıda öğrencinin katılması ve puanlamada nesnellik/kullanışlılık gereksinimleri

nedeniyle çoktan seçmeli maddeler ağırlıkla kullanılmaktadır. Bu nedenle, beceri temelli öğrenme çıktılarının bu tür uygulamalarda da geçerli biçimde ölçülmesi gerekir. Becerilerin geniş ölçekli testlerle ölçülmesi modelin doğrudan hedefi olmasa da, ölçme sisteminin mevcut yapısı gereği kaçınılmaz bir gereklilik olarak görülmektedir.

Bu bağlamda, TYMM'de yer alan öğrenme çıktılarının çoktan seçmeli madde biçiminde nasıl ölçüleceği üzerinde durulması gereken önemli konulardan biridir. Her ne kadar TYMM'de tanımlanan beceriler arasında hiyerarşi bulunmasa da sınıflama, yorumlama, karşılaştırma, çözümleme gibi becerileri içeren öğrenme çıktıları çoktan seçmeli maddelerle daha uygun biçimde ölçülebilir. Buna karşılık problem çözme, eleştirel düşünme gibi üst düzey beceriler ile çok adımlı akıl yürütme gerektiren öğrenme çıktıları söz konusu olduğunda ölçme tasarımı önemli ölçüde karmaşıklaşır. Bununla birlikte beceri düzeyi ne olursa olsun TYMM'deki öğrenme çıktılarının etkili biçimde ölçülebilmesi için bağlam temelli çoktan seçmeli maddelerin geliştirilmesi gerekmektedir. Çünkü beceriler gerçek yaşam durumlarından bağımsız değildir; anlamlı bir bağlam içinde ortaya çıkar. Yalnızca "soru yazma tekniği" olmayan bu süreç içerik çerçevesinin ve becerinin analizi, süreç bileşenlerinin incelenmesi, bağlam tasarımı ile süreç bileşenlerinin ilişkisel ve hiyerarşik yapılandırılmasını kapsayan bütüncül bir ölçme ve değerlendirme tasarımıdır.

TYMM'de öğrenme çıktılarının geçerli biçimde değerlendirilmesinde gerçek yaşamla ilişkilendirilmiş bağlamlar önemli rol oynar. Nitelikli bir bağlam, öğrenciyi yalnızca ön bilgisini hatırlamaya değil verilen bilgiyi işlemeye, dönüştürmeye ve uygulamaya yönlendirir. Bu bağlamlar grafik, tablo, metin, harita, vaka ya da gerçek yaşam senaryoları içerebilir. Gündelik yaşamla ilişkili bir bağlam, öğrencinin önceki öğrenmelerini yeni durumlara transfer edebilmesine ve öğrenme çıktılarını farklı içerik/alanlara genelleyebilmesine imkân sağlar (Demirtaşlı, 2010). Ölçme aracında yer alan bu bağlamlar ölçülmek istenen beceriyi harekete geçiren işlevsel bir bileşen olmalıdır. Öğretimde olduğu gibi ölçmede de anlamlı bağlamlar kalıcılık için önemlidir. Bu nedenle bağlamların ilgi çekici, anlamlı, yaş düzeyine uygun ve bilişsel olarak uyarıcı olması; ayrıca içerik düzeyinin öğrenciye uygunluğu gözetilmelidir. Gerekli görülen durumlarda, bağlamlarda öğretim programındaki anahtar kelimeler, genellemeler ya da ilkelere de yer verilebilir. Bağlam, ölçülmek istenen beceriyi harekete geçirecek nitelikte değilse, madde yazımı ölçülmek istenen yapıyı temsil etme niteliğini kaybeder. Örneğin problem çözme becerisini ölçmeyi amaçlayan bir bağlam, öğrencinin bu beceriyi işe koşmasını gerektiren bir problem durumu içermiyorsa madde, ölçülmek istenen yapıyı yeterince temsil edemez. Bu durum, madde ile hedeflenen beceri arasındaki uyumu zayıflatır ve yapıya ilişkin geçerlik kanıtlarını sınırlar. Bağlamın niteliği kadar gereksiz bilişsel yük oluşturmaması da önemlidir. Gereksiz ayrıntılar okuma yükünü artırır, öğrencinin dikkatini ölçülmek istenen yapıdan uzaklaştırır ve ölçme sonuçlarının geçerliğini olumsuz etkileyebilir. Bu nedenle bağlam, hedeflenen beceriyi desteklemeli;

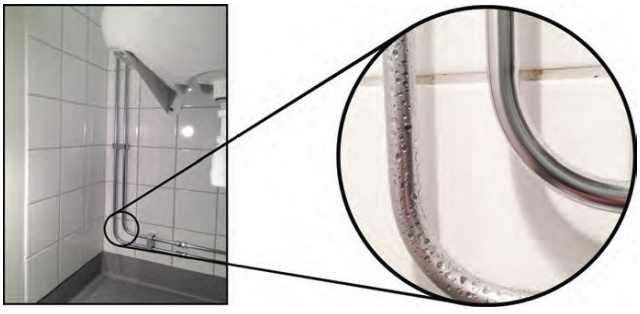
öğrenciyi gereksiz detaylarla meşgul ederek hedeflenen beceriyi sergilemesini zorlaştırmamalıdır. Aksi hâlde artan bilişsel yük, aracın geçerliğini zayıflatır ve sonuçlara dayalı kararların güvenirliğini düşürür.

Şekil 3'te bir bağlam ve bu bağlama bağlı bir çoktan seçmeli madde örneği verilmiştir. Maddede, öğrenciye bir banyo ortamı sunulmuş, görsel bir durum üzerinden fiziksel bir olgunun açıklanması istenmiştir. Böylece madde; hâl değişimleri, yoğuşma gibi soyut bilimsel kavramları, öğrencinin karşılaşabileceği somut bir durumla ilişkilendirmektedir. Öğrenciden doğrudan bilgiyi hatırlaması yerine gözlenen olguya ilişkin çıkarım yapması beklenmektedir. Bağlamda gereksiz bilgi veya görsel kullanımından kaçınılmıştır. Görsel, sorunun çözümüne doğrudan katkı sunmasa da görsel bir uyarıcı olarak bilişsel işlemleme sürecini desteklemekte ve sorunun daha açık ve anlaşılır olmasına katkı sunmaktadır.

Şekil 3

Gerçek Yaşam Bağlamına Dayalı Çoktan Seçmeli Madde Örneği

Bir banyoda biri soğuk, diğeri sıcak su borusu olmak üzere iki borudan su akmaktadır. Ali banyo yaptığında borulardan birinin üzerinde şekildeki gibi küçük su damlacıkları oluştuğunu fark eder. Boruyu kontrol ettiğinde ise su sızıntısı olmadığını görür.



Hal değişimleri bilgisine göre borunun üzerinde su damlacıklarının oluşmasının <u>en olası</u> nedeni aşağıdakilerden hangisidir?

A) Borunun içindeki suyun boru malzemesinden sızması

B) Sıcak su borusunun ısı vermesiyle havadaki su moleküllerini buharlaştırması

C) Soğuk boru çevresindeki nemli havanın boru yüzeyinde yoğuşması

D) Soğuk metal boru yüzeyi ile havadaki su molekülleri arasındaki kimyasal etkileşim

E) Borunun havadan aldığı ısı nedeniyle yüzeyinde nem tutması

Not: Bağlam, 2016-2017 öğretim yılı İsveç Ulusal Sınavı 9. sınıf fizik dersi A1 testindeki 8. soruda yer alan bağlam uyarlanarak hazırlanmıştır.

Ölçülecek öğrenme çıktısına uygun bir bağlam oluşturulduktan sonraki adım, bu bağlama dayalı ve ölçülecek öğrenme çıktısının süreç bileşenlerini de kapsayacak maddelerin yazılmasıdır. Özellikle okuma yükü, sınav süresi, sınavda sınırlı sayıda madde yer alması gibi uygulama koşulları dikkate alındığında bir bağlama dayalı birden fazla madde yazmak pratik ve işlevsel bir yaklaşımdır. Ancak bu tür bağlama dayalı "ortak köklü" maddelerde iki temel hatayla sıkça karşılaşılmaktadır. Bunların ilki, maddelerin bağlamdan bağımsız olarak çözülebilmesidir. Oysa öğrencinin maddeyi doğru yanıtlayabilmesi için bağlamda sunulan bilgilerden yararlanması zorunludur. İkinci hata ise maddelerin cevaplanmasının birbirine bağımlı olmasıdır (testlet etkisi). Bazı durumlarda bir madde, diğerinin ön koşulu hâline gelebilmekte ya da bir maddede verilen yanıt başka bir madde için ipucu oluşturabilmektedir.

TYMM'de öğrenme çıktılarının süreç bileşenlerinin de değerlendirilmesi amaçlandığı düşünüldüğünde bu tür etkileşimlerin ortaya çıkma olasılığı daha da artmaktadır. Örneğin, KB2.8.Sorgulama becerisinin ilk süreç bileşeni "*KB2.8.SB1. Merak ettiği konuyu tanımlamak*", ikinci süreç bileşeni ise "*KB2.8.SB2. İlgili konu hakkında sorular sormak*"tır. Bu iki süreç bileşeni ayrı maddelerle ölçülmeye çalışıldığında birinci süreç bileşenini ölçen maddenin, ikinci madde için ipucu oluşturması kaçınılmaz olabilir. Bu nedenle madde yazım sürecinde tek bir kalıp ya da şablona bağlı kalmak, süreci mekanikleştirerek maddelerin niteliğini düşürebilir ve maddeler arasında ipucu etkisi oluşturabilir. Bu tür sorunları önlemek ve geçerliği artırmak için madde geliştirme sürecinde farklı stratejiler kullanılabilir. Bu stratejiler; bağlam çeşitliliği, ek bağlam kullanma, bazı süreç bileşenlerinin bağlam içine yerleştirilmesi, birden fazla süreç bileşenini aynı anda ölçülmesi olabilir.

Bununla birlikte, bazı öğrenme çıktılarının her süreç bileşeninin ayrı maddelerle ölçülmesine gerek olmayabilir. Belirli süreç bileşenleri, ayrı bir madde yerine bağlamın doğal bir parçası olarak yapılandırılabilir. Örneğin tarih dersinde "*TAR.9.1.1. Tarih öğrenmenin bireye ve topluma faydalarını yorumlayabilme*" öğrenme çıktısının ilk süreç bileşeni olan "*a) Tarih öğrenmenin bireye ve topluma faydalarını kaynaklardan inceler.*" için ayrı bir madde yazmak yerine bağlamda farklı türde tarihî veya güncel kaynaklar sunulabilir. Bu durumda öğrenci, diğer süreç bileşenlerine yönelik maddeleri yanıtlayabilmek için zaten bu kaynaklarla etkileşime girmek durumunda kalacaktır. Böylece ilk süreç bileşeni, bağlamın işlevsel bir unsuru hâline gelmiş olur. Benzer şekilde, sorgulama becerisindeki "merak ettiği konuyu tanımlama" süreç bileşeni de bağlam aracılığıyla sunulabilir. Bu yaklaşım, özellikle sınırlı madde sayısıyla farklı süreç bileşenlerinin ölçülmesi gereken durumlarda hem bağlamı işlevselleştirir hem de madde geliştirme sürecine esneklik kazandırır.

4.1. Alan Bağımlı ve Alan Bağımsız Çoktan Seçmeli Madde Örnekleri

Çoktan seçmeli madde takımı, ortak bir temel bağlam etrafında yapılandırılmış birden fazla çoktan seçmeli maddeden oluşmaktadır. Bazı madde takımlarında, temel bağlamı koruyarak farklı durumsal ek bağlamlar sunmakta ve böylece aynı problem durumu içinde farklı süreç bileşenlerinin ölçülmesine olanak tanımaktadır.

Şekil 4'te Türk Dili ve Edebiyatı alanına ilişkin, problem çözme becerisinin kullanıldığı örnek madde takımı sunulmuştur. Bu örnekte, 12. sınıf düzeyinde "edebî dönem ve akımların ayırt edilmesi, metin-bağlam-özellik ilişkilerinin kurulması ve öğretimin bu doğrultuda biçimlendirilmesi" odağında yapılandırılmış bir bağlam kullanılmıştır. Bağlam, öğrencilerin edebî dönemlere ilişkin kuramsal bilgileri metin örnekleriyle ilişkilendirmekte zorlandıklarını ortaya koyan sınıf içi bir uygulamaya dayanmaktadır.

Şekil 4

Problem Çözme Becerisine Yönelik Çoktan Seçmeli Madde Takımı Örneği

1-4. soruları aşağıda verilen metin ve tablodan yararlanarak çözünüz.

Türk dili ve edebiyat öğretmeni Zeynep Hanım, 12. sınıf öğrencilerinin dönem başında hazırbulunuşluk düzeylerini belirlemek amacıyla onlara Tanzimat Dönemi sonrasına yönelik bazı sorular yöneltmiş ve öğrencilerin cevaplarını tahtaya yazmıştır.

Sorular	**Öğrenci cevapları**
Tanzimat döneminin temel amacı neydi?	Toplumu eğitmek, fayda amacı, hak-hürriyet temaları
Servet-i Fünun döneminde edebî dil nasıldı?	Ağır ve sanatlı dil, bireysel konular, batı etkisi
Fecr-i Âti sanat anlayışı neyi savunurdu?	Sanat toplulukların üzerinde görülür, sanatçılar birliği
Millî Edebiyat sanatçıları hangi ilkeleri benimsiyordu?	Sade Türkçe, halk için yazma, millî konular

Ardından öğretmen etkileşimli tahtaya dört farklı şiir parçası yansıtmış ve öğrencilerden bu şiirlerin hangi edebî döneme ait olabileceğini tahmin etmelerini istemiştir. Öğrencilerin çoğunluğunun hemfikir olduğu cevapları tabloda yer almaktadır.

Şiir parçaları	**Öğrenci cevapları**
1. Hicran-ı müebbed gibi dermanı bulunmaz, Bir derd ile kıvrandırıyor işte bugün yaz. Ruhumda karanlık geceler hükmünü sürmekte, Bir lâhza teselli bulacak hâl kalmamış az.	Fecr-i Âti
2. Gülşen-i dûnumuzun bir gülü yoktur açılmaz, Ey vatan, aşkın ile düştüğüm bu hâlden geçilmez. Feryâdımın âvâzesi göklere aks etse de, Sen bağrını açmazsan ben sana nasıl erişilmez?	Servet-i Fünun
3. Ben bir Türk'üm; dinim, cinsim uludur. Sinem özüm ateş ile doludur. İnsan olan vatanının kuludur, Türk evlâdı evde durmaz, giderim!	Tanzimat
4. Artık ne bir ses, ne bir nefes var; İçimde çökmüş ebedî bir hazan. Sanki bu loş perdeler arkasında Kaybolmuş bir âlemdir zaman.	Millî Edebiyat

Zeynep Hanım, yaptığı uygulama sonrasında öğrencilerle ilgili genel bir durum değerlendirmesi yaparak öğretimini biçimlendirmeye karar vermiştir.

1. Bu metne göre öğrencilerin edebî dönemleri öğrenme sürecinde karşılaştıkları <u>temel problem</u> aşağıdakilerden hangisidir?

A) Edebî dönemlerin tarihsel sırasını karıştırmaları ve kronolojik olarak doğru sıralayamamaları
B) Edebî dönemlere ilişkin kuramsal bilgileri metnin bağlamı, dili ve temasıyla bütünleştirmekte yetersiz kalmaları
C) Edebî dönemlerde öne çıkan bilgileri ve temel metinlerin eşleştirmesini hatırlayamamaları
D) Öğrencilerin derslerde işledikleri edebî metinleri tanımakta ve öğrenmekte zorlanmaları
E) Öğrencilerin metinlerin âhenk ve vezin özelliklerindeki farklılığı belirleme konusunda eksiklerinin bulunması

Ölçülen süreç bileşeni: KB3.2.SB1. Problemi yapılandırmak

2. Metinde öğrencilerin edebî dönemleri öğrenme sürecindeki sorun aşağıdakilerden hangisinde <u>en iyi</u> şekilde özetlenmiştir?

A) Öğrenciler edebî dönemlerin isimlerini karıştırdıkları ve tarihsel sıralamayı unuttuklarından dolayı sınavlarda başarısız olmakta ve edebiyat dersinden uzaklaşmaktadır.
B) Öğrenciler akımları birer özellik listesi şeklinde ezberlemekte ancak bu özelliklerin metinde nasıl göründüğünü, akımın ortaya çıkış nedenlerini ve dönem ruhunu kavrayamadıklarından metinleri derinlemesine analiz edememektedir.
C) Edebi akımlar çok karmaşık kavramlar içerdiği için öğrenciler bunları anlamakta zorlanmakta ve bu durum edebiyata olan ilgilerini tamamen yok etmektedir.
D) Öğretmenler geleneksel yöntemlerle ders anlattığı için öğrenciler edebî dönemleri sıkıcı bulmakta ve derse katılım göstermemektedir.
E) Öğrenciler edebi metinleri yeterince okumadıkları için akımlar hakkında hiçbir bilgi edinememekte ve sınav sorularını yanıtlayamamaktadır.

Ölçülen süreç bileşeni: KB3.2.SB2. Problemi özetlemek

Zeynep Öğretmen, dersini biçimlendirme öncesinde öğrencilerden verdikleri cevabın gerekçelerini belirtmelerini istemiştir.

Şiir parçaları	**Öğrenci cevapları**
1. Hicran-ı müebbed gibi dermanı bulunmaz, Bir derd ile kıvrandırıyor işte bugün yaz. Ruhumda karanlık geceler hükmünü sürmekte, Bir lâhza teselli bulacak hâl kalmamış az.	Fecr-i Âti çünkü sanat toplum ilişkisi ön plandadır.
2. Gülşen-i dûnumuzun bir gülü yoktur açılmaz, Ey vatan, aşkın ile düştüğüm bu hâlden geçilmez. Feryâdımın âvâzesi göklere aks etse de, Sen bağrını açmazsan ben sana nasıl erişilmez?	Servet-i Fünun çünkü ahenk aruz vezni ile sağlanmış ve dili ağırdır.
3. Ben bir Türk'üm; dinim, cinsim uludur. Sinem özüm ateş ile doludur. İnsan olan vatanının kuludur, Türk evlâdı evde durmaz, giderim!	Tanzimat Dönemi'ne aittir çünkü vatan konusuna odaklıdır.
4. Artık ne bir ses, ne bir nefes var; İçimde çökmüş ebedî bir hazan. Sanki bu loş perdeler arkasında Kaybolmuş bir âlemdir zaman.	Millî Edebiyat Dönemi'ne aittir çünkü sade bir dil kullanılmıştır.

3. Öğrenci cevaplarının ardından öğretmenin problemin çözümü için aşağıdakilerin hangisini yapması beklenir?

A) Tanzimat ve Servet-I Fünun Dönemi şiirlerinde karıştırılan belirleyici özelliklere yönelik farklı şiirler incelemesi
B) Şiir-dönem eşleştirmelerinde dil özelliklerinin belirleyici yönünü ön plana çıkaran çalışmalar gerçekleştirmesi
C) Şiir incelemelerinde tema, dil, biçim özelliklerini birlikte ele alınmasını sağlayan çözümlemeler yapması
D) Edebiyatın dönemlere ayrılmasında konuların önemine yönelik edebî gelenek anlayışını açıklaması
E) Millî Edebiyat Dönemi'ne ilişkin farklı içerik ve dil özelliğinde kaleme alınmış şiirleri karşılaştırmalı olarak çözümlemesi

Ölçülen süreç bileşeni: KB3.2.SB3. Problemin çözümüne yönelik gözleme dayalı/mevcut bilgiye/veriye dayalı tahminde bulunmak.
KB3.2.SB4. Problemin çözümüne yönelik önermeler üzerinden akıl yürütmek

Zeynep Öğretmen, yaptığı etkinlikler sonrasında yeni konuya geçmiş ve süreci şu şekilde ifade etmiştir:

"Bu dersimizde edebî akımların ortaya çıktığı dönemi ve nedenlerini inceleyeceğiz. Sonra her akımı yansıtan farklı metinler tahlil edecğiz. Örneğin bugün işleyeceğimiz akımındaki günlük dil kullanımını, sade anlatımı, beklenmedik bağlantıları Orhan Veli ve Melih Cevdet Anday şiirlerinde de tahlil edeceğiz. Ayrıca bu şiirleri daha sonra Cemal Süreya, Turgut Uyar gibi şairlerin şiirleriyle karşılaştırarak bir sonuca ulaşacağız."

4. Buna göre Zeynep Öğretmen'in amacı aşağıdakilerin hangisidir?

A) 11. sınıf ve 12. sınıf öğretim programında yer alan edebî dönemleri karşılaştırarak öğrencilerin bir genellemeye ulaşmasına olanak tanımak

B) Cumhuriyet ile Millî Edebiyat Dönemi şiirlerinin sahip olduğu özelliklere yönelik bilgilerin beceriye dönüşmesine yardımcı olmak

C) Şairlerin bireysel üsluplarını tespit etmeleri ve şiirlerdeki temaları sınıflandırarak değerlendirme yapmaları için öğrencilere rehberlik etmek

D) Cumhuriyet dönemi şiirindeki sade dil kullanımının edebî gelenek içindeki yerini inceleyerek, bu dönem şiirlerinin edebiyatta neden ön planda olduğunu göstermek

E) Garip akımının özelliklerini anlamaları, özellikleri metinlerle ilişkilendirmeleri ve İkinci Yeni akımı ile karşılaştırarak değerlendirme yapmalarını sağlamak

Ölçülen süreç bileşeni: KB3.2.SB5. Problemin çözmüne ilişkin yansıtma/değerlendirmede bulunmak

Birinci madde, "*KB3.2.SB1 Problemi yapılandırmak*" süreç bileşenini ölçmektedir. Problemi yapılandırma, düzensiz bilgileri anlamlı bir çerçeve haline getirmeyi gerektirir. Bu maddede öğrenciden, verilen metin ve tablolar üzerinden öğrencilerin edebî dönemleri öğrenme sürecinde yaşadıkları temel sorunu belirlemesi beklenmektedir. Öğrencinin bu soruyu yanıtlayabilmesi için öncelikle bağlamdaki verileri incelemesi, öğrencilerin verdikleri cevapları bütüncül olarak değerlendirmesi ve problemi doğru biçimde yapılandırması gerekmektedir. Böylece öğrencinin, karşılaştığı durumu bir "problem" olarak tanımlaması ve problemi oluşturan unsurları ayırt etmesi istenmektedir.

İkinci madde, "*KB3.2.SB2 Problemi özetlemek*" süreç bileşenini ölçmektedir. Problemi özetlerken öğrenciden yapılandırılmış bir problemin, temel bileşenleri korunarak daha basit ve anlaşılır bir ifadeye dönüştürülmesi beklenmektedir. Bu maddede öğrenciden, birinci maddede yapılandırılan problemi daha sade ve anlaşılır bir ifadeyle özetlemesi beklenmektedir. Seçenekler, öğrencilerin yaşadığı sorunu farklı düzeylerde ve farklı bakış açılarıyla ifade eden açıklamalardan oluşturulmuştur. Öğrenciden, bağlamdaki ayrıntılardan hareketle problemi en kapsayıcı ve sade biçimde ifade eden seçeneği belirlemesi beklenmektedir.

Üçüncü madde, "*KB3.2.SB3 Problemin çözümüne yönelik gözleme dayalı/mevcut bilgiye dayalı tahminde bulunmak*" ve "*KB3.2.SB4 Problemin çözümüne yönelik önermeler üzerinden akıl yürütmek*" süreç bileşenlerini birlikte ölçmektedir. Maddede, temel bağlama ek olarak öğrenci cevaplarını içeren ikincil bir bağlam yer almaktadır. Bağlamda, öğrencilerin şiir-dönem eşleştirmelerine ilişkin verdikleri gerekçeler sunulmuştur. Öğretmenin bu verilerden hareketle öğretimini nasıl biçimlendirmesi gerektiği sorgulanmıştır. Bu maddede öğrenciden beklenen, sunulan cevapları inceleyerek öğrencilerin kavram yanılgılarını veya eksik anlamalarını tespit etmesi ve bu tespitlerden hareketle en uygun öğretimsel müdahaleyi belirlemek üzere akıl yürütmesidir. Böylece öğrenci hem gözleme dayalı çıkarım yapmakta hem de bu çıkarım üzerinden akıl yürütmektedir.

Dördüncü madde, "*KB3.2.SB5 Problemin çözümüne ilişkin yansıtma/değerlendirmede bulunmak*" süreç bileşenini ölçmektedir. Bu maddede Zeynep öğretmenin dersin devamında izlediği öğretim süreci verilmiştir. Öğrenciden bu sürecin amacını değerlendirmesi istenmiştir. Öğrenciden istenen, öğretmenin seçtiği metinleri, karşılaştırmaları ve izlediği yolu dikkate alarak yapılan öğretimsel düzenlemenin hangi öğrenme amacına hizmet ettiğini belirlemektedir.

İkinci örnekte 9. Sınıf matematik dersinde "*MAT.9.5.5. Eşlik ve benzerlikle ilgili çıkarım ve teoremleri içeren problemleri çözebilme*" öğrenme çıktısına yönelik madde takımı yer almaktadır. Bu öğrenme çıktısı matematik alan becerilerinden "*MAB2. Matematiksel Problem Çözme*" becerisi kullanılarak hazırlanmıştır. Bu öğrenme çıktısının süreç bileşenleri Şekil 5'te verilmiştir.

Şekil 5

9. Sınıf Matematik Dersi Öğrenme Çıktısı

MAT.9.5.5. Eşlik ve benzerlikle ilgili çıkarım ve teoremleri içeren problemleri çözebilme
a) Problemin verilen ve istenenlerine ilişkin parçaları belirler.
b) Problemde verilenler, istenenler ve gerekli işlemler arasındaki ilişkileri belirler.
c) Problemin parçaları arasındaki ilişkileri problem bağlamına uygun olarak dönüştürür.
ç) Matematiksel temsillere dönüştürdüğü problemi kendi ifadeleri ile açıklar.
d) Problemin çözümünü gerçekleştirmek için stratejiler oluşturur.
e) Belirlediği stratejiyi çözüm için uygulayarak problemi çözer.
f) Problemin çözümünü kontrol eder.
g) Problemin çözümü için geliştirdiği, kullandığı stratejilerdeki kısa yolları ve çözüme ulaştırmayan stratejileri belirleyerek çözüme ilişkin deneyimini gözden geçirir.
ğ) Çözüme ulaştıran stratejilerden hangilerinin hangi tür problemlere uygulanabileceğine ilişkin çıkarım yapar.
h) Ulaştığı çıkarımların geçerliliğini matematiksel örneklerle değerlendirir.

Şekil 5'te yer verilen 9. Sınıf matematik dersi öğrenme çıktısına yönelik hazırlanan Şekil 6'daki madde takımında, birinci soru a, b, c ve e süreç bileşenlerini ölçecek biçimde yapılandırılmıştır. Öğrencinin öncelikle maddenin çözümüne ulaşmak için verilen ve istenen bilgileri belirlemesi gerekmektedir (a süreç bileşeni). Ardından problemde verilenler, istenenler ve gerekli işlemler arasındaki ilişkiden yola çıkarak binanın yüksekliği ile gölge uzunluğu arasındaki orantısal ilişkiyi kurması; çubuk için verilen yükseklik-gölge bilgisinden hareketle benzerlik/oran mantığını kullanarak gerekli uzunluğu elde etmesi beklenir (b süreç bileşeni). Bu süreçte öğrencinin, problemdeki büyüklükler arasındaki ilişkileri problem bağlamına uygun biçimde matematiksel ilişkilere dönüştürmesi gerekmektedir (c süreç bileşeni). Son aşamada ise öğrencinin belirlediği çözüm yolunu uygulayarak uçuş mesafesi olan [AC] uzunluğunu hesaplaması beklenir (e süreç bileşeni). Ayrıca öğrenci, hesapladığı sonucu seçeneklerle karşılaştırarak doğrulama sürecini örtük biçimde gerçekleştirmektedir. Bu durum f süreç bileşeninin dolaylı olarak ölçülmesine olanak tanımaktadır.

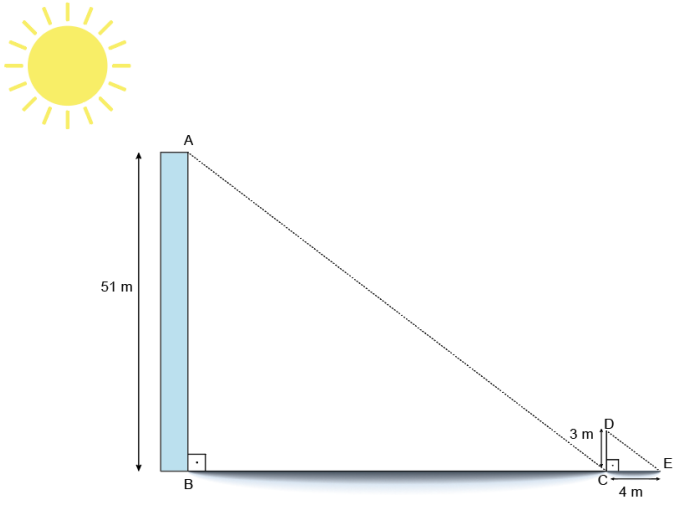
İkinci soru ğ ve h süreç bileşenlerini ölçmektedir. Bu madde, öğrencinin reklamcının kullandığı çözüm stratejisinin (benzerlik/oran temelli yaklaşımın) hangi problem türlerine doğrudan uygulanabileceğine ilişkin çıkarım yapmasını amaçlamaktadır (ğ süreç bileşeni). Ayrıca öğrenci, seçeneklerde verilen problem durumlarını değerlendirerek yaptığı çıkarımın geçerliğini matematiksel örnekler üzerinden değerlendirmek durumundadır (h süreç bileşeni). Örnek madde takımında, ç süreç bileşeni (Matematiksel temsillere dönüştürdüğü problemi kendi ifadeleri ile açıklar.) ile g süreç bileşeni (Problemin çözümü için geliştirdiği, kullandığı stratejilerdeki kısa yolları ve çözüme ulaştırmayan stratejileri belirleyerek çözüme ilişkin deneyimini gözden geçirir.) çoktan seçmeli madde ile ölçülmesi uygun olmadığından dolayı ölçülmemiştir. Bu süreç bileşenleri, öğrencinin kendi düşünce sürecini sözlü veya yazılı olarak ifade etmesini gerektiren farklı madde türleri ile ölçülebilir.

Şekil 6

9. Sınıf Matematik Dersi Öğrenme Çıktısına İlişkin Bağlama Dayalı Madde Takımı

1 ve 2. soruları verilen problem durumundan yararlanarak çözünüz.

Tanıtım filmleri çeken bir reklamcı, yerden dik olarak inşa edilmiş 51 metre yüksekliğindeki bir binanın tepesine kadar olan görüntüyü almak amacıyla bir dron uçuşu planlamaktadır. Reklamcı, binanın gölgesinin düştüğü en uç noktayı uçuşa başlangıç noktası olarak belirlemiştir. Reklamcı, uçuş mesafesini hesaplayabilmek için gölgenin uç noktasına zemine dik olacak şekilde 3 metrelik bir çubuk yerleştirmiş ve bu çubuğun gölge uzunluğu olan [CE]'nın 4 metre olduğunu ölçmüştür. (Bina zemini ile çubuğun dikildiği zemin aynı düzlemdedir.)



1. Dronun uçuş mesafesi olan [AC] nın uzunluğu kaç metredir?

A) 68 B) 72 C) 80 D) 85 E) 100

Ölçülen süreç bileşenleri: a, b, c, e ve f

2. Bu reklamcının çözüm için kullandığı strateji ile aşağıdaki problem durumlarından hangisi çözülebilir?

A) İki kişinin boylarının oranı ve gölge uzunluklarının ölçülebildiği biliniyorsa bu kişilerden herhangi birinin gölge uzunluğundan yararlanarak boyunun hesaplanması

B) Dar açılarından biri birbirine eşit iki dik üçgenin dik kenarlarının oranı biliniyorsa herhangi birinin hipotenüs uzunluğunun hesaplanması

C) Gölge uzunlukları ölçülebilen, zemine dik açıyla yan yana duran bir ağaç ve bir binadan, ağacın gölge uzunluğu, binanın yüksekliği biliniyorsa ağacın yüksekliğinin hesaplanması

D) Birbirine benzer iki üçgenden birinin hipotenüs uzunluğu, diğerinin dik kenarlarından birinin uzunluğu biliniyorsa benzerlik oranının hesaplanması

E) Gölge uzunlukları ölçülebilen ve uzunlukları bilinen, zemine dik açıyla dikilmiş doğrusal iki çubuktan birinin gölge uzunluğu biliniyorsa diğerinin gölge uzunluğunun hesaplanması

Ölçülen süreç bileşeni: ğ ve h

5. SONUÇ

Bu bölümde TYMM'de tanımlanan becerilerin ölçülmesi, beceri ve yeterlik kavramlarından başlayarak öğretim programlarının yapısı, öğrenme çıktıları, süreç bileşenleri, öğrenme kanıtları ve ölçme-değerlendirme uygulamaları çerçevesinde ele alınmıştır. Bölüm boyunca ortaya konulduğu üzere beceri, yalnızca bilginin hatırlanmasıyla sınırlı değildir. Bilginin uygun bağlamlarda işlenmesi, kullanılması ve performansa dönüştürülmesiyle görünür hâle gelen gelişimsel ve çok boyutlu bir yapıdır. Bu nedenle TYMM'de becerilerin ölçülmesi, yalnızca öğrenme çıktılarının değerlendirilmesiyle değil, bu çıktılara ulaşmayı sağlayan süreç bileşenlerinin de dikkate alınmasıyla mümkün olmaktadır.

TYMM öğretim programlarında ölçme ve değerlendirme, geleneksel sonuç odaklı anlayıştan farklı olarak öğrenme kanıtları üzerinden yapılandırılmıştır. Bu yaklaşımda değerlendirme, öğretim sürecinin sonunda yapılan bir kontrol işlemi olmaktan çıkarak öğrenme-öğretme sürecinin doğal ve yönlendirici bir parçası hâline gelmektedir. Öğrenme kanıtlarının önceden belirlenmesi, öğrencinin hedeflenen öğrenme çıktısına ulaşıp ulaşmadığını anlamayı kolaylaştırmakta; aynı zamanda öğretim sürecinin bu çıktıları destekleyecek biçimde düzenlenmesine imkân sağlamaktadır. Bu yönüyle TYMM'de ölçme ve değerlendirme, yalnızca "ne öğrenildi?" sorusuna değil, "öğrenme nasıl geliştirilebilir?" sorusuna da yanıt arayan bir yapı sunmaktadır.

Sınıf içi ölçme ve değerlendirme uygulamaları, beceri temelli yaklaşımın en güçlü biçimde hayata geçirilebileceği alanlardan biridir. Performans görevleri, dereceli puanlama anahtarları, çalışma yaprakları, açık uçlu maddeler, çıkış biletleri, öz

değerlendirme ve akran değerlendirme gibi araçlar; öğrencinin yalnızca ortaya koyduğu ürünü değil, bu ürüne ulaşma sürecini de izlemeye olanak sağlamaktadır. Ancak ölçme araçlarının çeşitlendirilmesi tek başına yeterli değildir. Kullanılan her kanıtın öğrenme çıktıları ve süreç bileşenleriyle uyumlu, öğrenci düzeyine uygun, geçerli, güvenilir ve adil biçimde yorumlanabilir olması gerekmektedir.

TYMM kapsamında öğrencilerin değerlendirme sürecine aktif katılımı da beceri temelli ölçme anlayışının önemli bir boyutunu oluşturmaktadır. Öz değerlendirme, akran değerlendirme, grup değerlendirme formları, kontrol listeleri ve öğrenme günlükleri gibi uygulamalar, öğrencilerin kendi öğrenmelerine ilişkin yargı geliştirmelerini desteklemektedir. Bu süreçte öğretmen, öğrenmeye ilişkin kanıtları toplayan, yorumlayan ve öğretimi bu kanıtlara göre uyarlayan kişi konumundadır. Öğrenci ise yanıtları, ürünleri ve yansımalarıyla değerlendirme sürecine aktif biçimde katılan özne hâline gelmektedir. Bu durum, öğrencilerin öğrenme çıktıları ile bu çıktılara ulaşma ölçütlerini daha iyi anlamalarına ve kendi öğrenmelerini daha bilinçli biçimde izlemelerine katkı sağlamaktadır.

Bununla birlikte TYMM öğretim programlarının ortaya koyduğu ideal ölçme ve değerlendirme anlayışı ile uygulama koşulları arasında bazı farklılıkların oluşabileceği göz ardı edilmemelidir. Sınıf mevcutları, öğretmenlerin ölçme ve değerlendirme yeterlikleri, zaman sınırlılıkları, okul imkânları ve merkezi sınav baskısı, beceri temelli değerlendirme uygulamalarının niteliğini doğrudan etkileyebilir. Bu nedenle TYMM'nin sürdürülebilir biçimde uygulanabilmesi için öğretmenlerin özellikle görev tasarlama, süreç bileşenlerini ölçütlere dönüştürme, dereceli puanlama anahtarı geliştirme, kanıtları yorumlama ve geri bildirim verme konularında desteklenmesi gerekmektedir. Böylece öğretim programlarında öngörülen beceri temelli yapı, sınıf içi uygulamalarda daha gerçekçi ve işlevsel biçimde karşılık bulabilir.

Yüksek riskli test uygulamaları açısından bakıldığında ise TYMM'de yer alan becerilerin ölçülmesi daha karmaşık bir yapı göstermektedir. Türkiye'de LGS ve YKS gibi sınavlarda çoktan seçmeli maddelerin yaygın olarak kullanılması, nesnel puanlama ve uygulama kolaylığı açısından anlaşılırdır. Ancak TYMM'deki bazı öğrenme çıktıları çoktan seçmeli maddelerle ölçülebilse de ürün ortaya koyma, paylaşma, yansıtma, gerekçelendirme ya da çok adımlı üst düzey düşünme gerektiren bazı çıktılar bu madde türüyle geçerli biçimde değerlendirilememektedir. Bu durum, geniş ölçekli ölçme tasarımlarında beceri temelli öğrenme çıktılarının doğasına uygun farklı madde ve görev türlerinin de dikkate alınmasını gerekli kılmaktadır. Özellikle yalnızca çoktan seçmeli maddelere dayalı yüksek riskli sınavlar, öğretim sürecinde ölçülebilen becerilerin öncelenmesine; ölçülmesi güç olan bazı beceri ve süreç bileşenlerinin ise geri planda kalmasına yol açabilir. Bu nedenle merkezi sınavlarda ölçme tasarımlarının TYMM'nin beceri temelli yapısıyla daha uyumlu hâle getirilmesi önemlidir. Açık uçlu maddeler, kısa yapılandırılmış yanıtlar, çoklu doğru cevap içeren maddeler, dijital performans görevleri, simülasyonlar ve teknoloji destekli ölçme uygulamaları bu açıdan

değerlendirilebilecek seçenekler arasındadır. Yapay zekâ destekli otomatik puanlama sistemlerindeki gelişmeler de özellikle açık uçlu yanıtların geniş ölçekli sınavlarda daha uygulanabilir hâle gelmesine katkı sunabilir. Bununla birlikte bu tür uygulamaların geçerlik, güvenirlik, puanlama tutarlılığı, adillik ve erişilebilirlik açısından güçlü kanıtlarla desteklenmesi gerekmektedir.

Sonuç olarak TYMM'de tanımlanan becerilerin ölçülmesi, yalnızca uygun soru ya da görev hazırlama süreci olarak görülmemelidir. Bu süreç; ölçülecek yapının açık biçimde tanımlanmasını, öğrenme çıktıları ile süreç bileşenlerinin birlikte ele alınmasını, gerçek yaşamla ilişkilendirilmiş anlamlı bağlamların kurulmasını ve elde edilen kanıtların geçerlik, güvenirlik ve adillik ilkeleri doğrultusunda yorumlanmasını gerektirmektedir. TYMM'nin beceri temelli yaklaşımı, ölçme ve değerlendirmeyi öğretme-öğrenme döngüsünün ayrılmaz bir bileşeni olarak konumlandırmakta, öğrencinin yalnızca ulaştığı sonucu değil, bu sonuca giden öğrenme sürecini ve beceri gelişimini de görünür kılmayı amaçlamaktadır. Bu nedenle modelin başarısı, yalnızca öğretim programlarında becerilerin tanımlanmasına değil, bu becerilerin sınıf içi ve geniş ölçekli ölçme uygulamalarında geçerli, güvenilir, adil ve sürdürülebilir biçimde değerlendirilebilmesine bağlıdır.

KAYNAKÇA | REFERENCES

Ananiadou, K. & Claro, M. (2009), *21st century skills and competences for new millennium learners in OECD countries.* OECD Education Working Papers, 41, OECD Publishing. https://doi.org/10.1787/218525261154.

Andrade, H. (2010). Students as the definitive source of formative assessment: Academic self-assessment and the self-regulation of learning. H. Andrade & G. J. Cizek (Ed.), *Handbook of formative assessment* içinde. Routledge.

Andrade, H. (2019). A critical review of research on student self-assessment. *Frontiers in Education*, *4*. https://doi.org/10.3389/feduc.2019.00087

Aşiret, S. & Su Özenir, Ö. (2025). Türkiye yüzyılı maarif modeli 2024 öğretim programlarında ölçme ve değerlendirme yaklaşımının çeşitli değişkenler açısından incelenmesi. *Ahi Evran Üniversitesi Kırşehir Eğitim Fakültesi Dergisi*, *26*(3), 895-931. https://doi.org/10.29299/kefad.1659146

Aşkar, P. & Altun, A. (2023). K-12 Beceriler çerçevesi: Türkiye bütüncül modeli üzerine bir çalışma. *Millî Eğitim Dergisi*, *52*(1), 925-940. https://doi.org/10.37669/milliegitim.1308740

Baer, J. (2012). Domain specificity and the limits of creativity theory. *The Journal of Creative Behavior*, *46*(1), 16–29. https://doi.org/10.1002/jocb.002

Baer, J. & Kaufman, J. C. (2005). Bridging generality and specificity: The amusement park theoretical (APT) model of creativity. *Roeper Review*, *27*(3), 158–163. https://doi.org/10.1080/02783190509554310

Binkley, M., Erstad, O., Herman, J., Raizen, S., Ripley, M., Miller-Ricci, M. & Rumble, M. (2012). Defining twenty-first century skills. P. Griffin, B. McGaw & E. Care (Ed.), *Assessment and teaching of 21st century skills* (s. 17–66) içinde. Springer. https://doi.org/10.1007/978-94-007-2324-5_2

Birenbaum, M. (2003). New insights into learning and teaching and their implications for assessment. Segers, M., Dochy, F. & Cascallar, E. (Ed.), *Optimising new modes of assessment: In search of qualities and standards. innovation and change in professional education*, *1* içinde. Springer. https://doi.org/10.1007/0-306-48125-1_2

Black, P. & Wiliam, D. (2009). Developing the theory of formative assessment. *Educational Assessment, Evaluation and Accountability*, *21*(1), 5-31. https://doi.org/10.1007/s11092-008-9068-5

Blömeke, S., Gustafsson, J.-E. & Shavelson, R. J. (2015). Beyond dichotomies: Competence viewed as a continuum. *Zeitschrift für Psychologie*, *223*(1), 3–13. https://doi.org/10.1027/2151-2604/a000194

Brookhart, S. M. (2018) Appropriate criteria: Key to effective rubrics. *Frontiers in Education*, *3*. https://doi.org/10.3389/feduc.2018.00022

Care, E. (2018). Twenty-first century skills: From theory to action. Care, E., Griffin, P. & Wilson, M. (Ed.), *Assessment and teaching of 21st century skills. Educational assessment in an information age*. Springer. https://doi.org/10.1007/978-3-319-65368-6_1

Care, E. & Kim, H. (2018). Assessment of twenty-first century skills: The issue of authenticity. Care, E., Griffin, P. & Wilson, M. (Ed.), *Assessment and teaching of 21st century skills. Educational assessment in an information age* içinde. Springer. https://doi.org/10.1007/978-3-319-65368-6_2

Care, E., Scoular, C. & Bui, M. (2015). Australia in the Context of the ATC21S Project. Griffin, P., Care, E. (Ed.), *Assessment and teaching of 21st century skills. Educational assessment in an information age* içinde. Springer. https://doi.org/10.1007/978-94-017-9395-7_8

Carless, D. & Boud, D. (2018). The development of student feedback literacy: enabling uptake of feedback. *Assessment & Evaluation in Higher Education*, *43*(8), 1315-1325. https://doi.org/10.1080/02602938.2018.1463354

CC2020 Task Force. (2020). *Computing curricula 2020: Paradigms for global computing education*. Association for Computing Machinery, IEEE Computer Society, Association for Information Systems. https://doi.org/10.1145/3467967

Council of the European Union. (2018). *Council Recommendation of 22 May 2018 on key competences for lifelong learning* (2018/C 189/01). *Official Journal of the European Union, C 189*, 1–13.

Darling-Hammond, L. (2017). Teacher education around the world: What can we learn from international practice? *European Journal of Teacher Education*, *40*(3), 291-309. https://doi.org/10.1080/02619768.2017.1315399

Dede, C. (2010). Comparing frameworks for 21st century skills. J. Bellance & R. Brandt (Ed.), *21st century skills: Rethinking how students learn* (s. 51-76) içinde. Solution Tree Press.

DeKeyser, R. M. (2020). Skill acquisition theory. B. VanPatten, G. D. Keating & S. Wulff (Ed.), *Theories in second language acquisition: An introduction* (3. baskı, s. 83–104) içinde. Routledge. https://doi.org/10.4324/9780429503986-5

Demirel, Ö. (2020). *Eğitimde program geliştirme: Kuramdan uygulamaya* (27. baskı). Pegem Akademi.

Demirtaşlı, N. (2010). Üst düzey düşünme becerilerinin ölçülmesinde gündelik yaşam unsuru. *CİTO Eğitim: Kuram ve Uygulama*, *7*, 9–26.

DiCerbo K. & Behrens J. (2012). *From technology enhanced assessment to assessment enhanced technology*. Paper presented at the annual meeting of the National Council on Measurement in Education.

Dreyfus, H. L. & Dreyfus, S. E. (1986). *Mind over machine: The power of human intuition and expertise in the era of the computer*. Basil Blackwell.

Ellis, C., van Haeringen, K., Harper, R., Bretag, T., Zucker, I., McBride, S., Rozenberg, P., Newton, P. & Saddiqui, S. (2020). Does authentic assessment assure academic integrity? Evidence from contract cheating data. *Higher Education Research & Development*, *39*(3), 454–469. https://doi.org/10.1080/07294360.2019.1680956

Ennis, R. H. (1989). Critical thinking and subject specificity: Clarification and needed research. *Educational Researcher, 18*(3), 4–10. https://doi.org/10.3102/0013189X018003004

Ercikan, K. & Oliveri, M. E. (2016). In search of validity evidence in support of the interpretation and use of assessments of complex constructs: Discussion of research on assessing 21st century skills. *Applied Measurement in Education*, *29*(4), 310-318. https://doi.org/10.1080/08957347.2016.1209210

Gordon Commission on the Future of Assessment in Education. (2013). *To assess, to teach, to learn: A vision for the future of assessment* (Technical report). Educational Testing Service. https://www.ets.org/Media/Research/pdf/gordon_commission_technical_report.pdf

Greiff, S. & Kyllonen, P. (2016). Contemporary assessment challenges: The measurement of 21st century skills [Editorial]. *Applied Measurement in Education*, *29*(4), 243-244. https://doi.org/10.1080/08957347.2016.1209209

Greiff, S., Wüstenberg, S., Csapó, B., Demetriou, A., Hautamäki, J., Graesser, A. C. & Martin, R. (2014). Domain-general problem solving skills and education in the 21st century. *Educational Research Review*, *13*, 74–83. https://doi.org/10.1016/j.edurev.2014.10.002

Griffin, P. & Care, E. (Ed.). (2015). *Assessment and teaching of 21st century skills: Methods and approach*. Springer. https://doi.org/10.1007/978-94-017-9395-7

Griffin, P., McGaw, B. & Care, E. (Ed.). (2012). *Assessment and teaching of 21st century skills.* Springer. https://doi.org/10.1007/978-94-007-2324-5

Griffin, P., Murray, L., Care, E., Thomas, A. & Perri, P. (2010). Developmental assessment: lifting literacy through professional learning teams. *Assessment in Education: Principles, Policy & Practice*, *17*(4), 383-397. https://doi.org/10.1080/0969594X.2010.516628

Gulikers, J., Bastiaens, T. & Kirschner, P. (2004). A five-dimensional framework for authentic assessment. *Educational Technology Research and Development*, *52*(3), 67-85. https://doi.org/10.1007/BF02504676

Heritage, M. (2010). *Formative assessment: Making it happen in the classroom*. Corwin Press.

Jonsson, A. & Svingby, G. (2007). The use of scoring rubrics: Reliability, validity and educational consequences. *Educational Research Review*, *2*, 130-144. https://doi.org/10.1016/j.edurev.2007.05.002

Kofinas, A. K., Tsay, C.-H. & Pike, D. (2025). The impact of generative AI on academic integrity of authentic assessments within a higher education context. *British Journal of Educational Technology, 56*(6), 2522–2549. https://doi.org/10.1111/bjet.13585

Kotsiou, A., Fajardo-Tovar, D. D., Cowhitt, T., Major, L. & Wegerif, R. (2022). A scoping review of future skills frameworks. *Irish Educational Studies*, *41*(1), 171-186. https://doi.org/10.1080/03323315.2021.2022522

Le Deist, F. D. & Winterton, J. (2005). What is competence? *Human Resource Development International*, *8*(1), 27–46. https://doi.org/10.1080/1367886042000338227

Leahy, S., Lyon, C., Thompson, M. & Wiliam, D. (2005). Classroom assessment: Minute-by-minute, day-by-day. *Educational Leadership*, *63*(3), 18–24.

Li, H., Xiong, Y., Hunter, C. V., Guo, X. & Tywoniw, R. (2020). Does peer assessment promote student learning? a meta-analysis. *Assessment & Evaluation in Higher Education*, *45*(2), 193-211. https://doi.org/10.1080/02602938.2019.1620679

Lippman, L., Ryberg, R., Carney, R. & Moore, K.A. (2015). *Key soft skills that foster youth workforce success; Toward a Consensus Across Fields*. Child Trends.

Lucas, B. & Claxton, G. (2009) *Wider skills for learning: what are they, how can they be cultivated, how could they be measured and why are they important for innovation?*. National Endowment for Science, Technology and the Arts (NESTA), https://hdl.voced.edu.au/10707/324716.

MEB (2024). *Öğretim programları kurul kararı* (Karar No: 20). https://mufredat.meb.gov.tr/

MEB (2025). *Türkiye Yüzyılı Maarif Modeli öğretim programları ortak metni*. https://mufredat.meb.gov.tr/ProgramDetay.aspx?PID=1945

Messick, S. (1989). Validity. R. L. Linn (Ed.), *Educational measurement* (3. baskı, s. 13-103) içinde. MacMillan.

National Research Council. (2000). *How people learn: Brain, mind, experience, and school: Expanded edition*. The National Academies Press.

Nicol, D. J. & Macfarlane-Dick, D. (2006). Formative assessment and self-regulated learning: a model and seven principles of good feedback practice. *Studies in Higher Education*, *31*(2), 199-218. https://doi.org/10.1080/03075070600572090

OECD. (2005). *The definition and selection of key competencies: Executive summary*. OECD Publishing.

OECD. (2019). *OECD learning compass 2030: A series of concept notes*. OECD Publishing.

Ornstein, A. C. & Hunkins, F. P. (2018). *Curriculum foundations, principles, and issues* (4. baskı). Allyn & Bacon.

Partnership for 21st Century Learning. (2015, March 26). *New direction for 21st century learning announced at P21 Summit.*

Panadero, E. & Jönsson, A. (2013). The use of scoring rubrics for formative assessment purposes revisited: A review. *Educational Research Review*, *9*, 129-144. https://doi.org/10.1016/j.edurev.2013.01.002

Panadero, E., Brown, G. & Strijbos, J. (2016). The future of student self-assessment: a review of known unknowns and potential directions. *Educational Psychology Review*, *28*(4), 803-830. https://doi.org/10.1007/s10648-015-9350-2

Pellegrino, J. W. & Hilton, M. L. (Eds.). (2012). *Education for life and work: Developing transferable knowledge and skills in the 21st century*. The National Academies Press. https://doi.org/10.17226/13398

Penuel, W. R. & Shepard, L. A. (2016). Assessment and teaching. D. H. Gitomer & C. A. Bell (Ed.), *Handbook of research on teaching* (5. baskı, s. 787–850) içinde. American Educational Research Association.

Perkins, D. N. & Salomon, G. (1989). Are cognitive skills context-bound? *Educational Researcher*, *18*(1), 16–25. https://doi.org/10.3102/0013189X018001016

Reddy, Y. M. & Andrade, H. (2010). A review of rubric use in higher education. *Assessment & Evaluation in Higher Education*, *35*(4), 435-448. https://doi.org/10.1080/02602930902862859

Rotherham, A. J. & Willingham, D. T. (2010). 21st-century skills: Not new, but a worthy challenge. *American Educator*, *34*(1), 17–20.

Rychen, D. S. (2004). *An overarching conceptual framework for assessing key competences in an international context. Lessons from an interdisciplinary and policy-oriented approach*. https://www.cedefop.europa.eu/files/BgR1_Rychen.pdf.

Saavedra, A. & Opfer, V. (2012). *Teaching and learning 21st century skills: Lessons from the learning sciences*. A Global Cities Education Network Report, Asia Society.

Sawyer, R. K. (2022). An introduction to the learning sciences. R. K. Sawyer (Ed.), *The Cambridge handbook of the learning sciences* (3. baskı., s. 1–24) içinde. Cambridge University Press. https://doi.org/10.1017/9781108888295.002

Scoular, C. (2025). *Essential skills for learning*. Australian Council for Educational Research. https://doi.org/10.37517/978-1-74286-748-9

Shavelson, R.J., Baxter, G.P. & Gao, X. (1993), Sampling variability of performance assessments. *Journal of Educational Measurement*, *30*, 215-232. https://doi.org/10.1111/j.1745-3984.1993.tb00424.x

Shepard, L. A. (2019). Classroom assessment to support teaching and learning. *The ANNALS of the American Academy of Political and Social Science*, *683*(1), 183-200. https://doi.org/10.1177/0002716219843818

Silva, E. (2009). Measuring skills for 21st-century learning. *Phi Delta Kappan*, *90*(9), 630-634. https://doi.org/10.1177/003172170909000905

Soland, J., Hamilton, L. S. & Stecher, B. M. (2013). *Measuring 21st century competencies: Guidance for educators*. RAND Corporation.

Soulé, H. & Warrick, T. (2015). Defining 21st century readiness for all students: What we know and how to get there. *Psychology of Aesthetics, Creativity, and the Arts*, *9*(2), 178-186. https://doi.org/10.1037/aca0000017

Sternberg, R. J. (2003). The development of creativity as a decision-making process. R. K. Sawyer, V. John-Steiner, S. Moran, R. J. Sternberg, D. H. Feldman, J. Nakamura & M. Csikszentmihalyi (Ed.), *Creativity and development* (s. 91–138) içinde. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195149005.003.0004

Teo, P. (2019). Teaching for the 21st century: A case for dialogic pedagogy. *Learning, Culture and Social Interaction*, *21*, 170-178. https://doi.org/10.1016/j.lcsi.2019.03.009

Thornhill-Miller, B., Camarda, A., Mercier, M., Burkhardt, J. M., Morisseau, T., Bourgeois-Bougrine, S., Vinchon, F., El Hayek, S., Augereau-Landais, M., Mourey, F., Feybesse, C., Sundquist, D. & Lubart, T. (2023). Creativity, critical thinking, communication, and collaboration: Assessment, certification, and promotion of 21st century skills for the future of work and education. *Journal of Intelligence*, *11*(3), 54. https://doi.org/10.3390/jintelligence11030054

Topping, K. (1998). Peer assessment between students in colleges and universities. *Review of Educational Research*, *68*(3), 249-276. https://doi.org/10.3102/00346543068003249

Tricot, A. & Sweller, J. (2014). Domain-specific knowledge and why teaching generic skills does not work. *Educational Psychology Review*, *26*, 265–283. https://doi.org/10.1007/s10648-013-9243-1

Voogt, J. & Roblin, N. P. (2012). A comparative analysis of international frameworks for 21st century competences: Implications for national curriculum policies. *Journal of Curriculum Studies*, *44*(3), 299–321. https://doi.org/10.1080/00220272.2012.668938

Wagner, T. (2010). *The global achievement gap: Why even our best schools don't teach the new survival skills our children need-and what we can do about it*. Basic Books

Weinert, F. E. (2001). Concept of competence: A conceptual clarification. D. S. Rychen & L. H. Salganik (Ed.), *Defining and selecting key competencies* (s. 45-65) içinde. Hogrefe and Huber Publishers.

Westera, W. (2001). Competences in education: A confusion of tongues. *Journal of Curriculum Studies*, *33*(1), 75–88. https://doi.org/10.1080/00220270120625

Wiggins, G. & McTighe, J. (2005). *Understanding by design* (2. baskı). Association for Supervision and Curriculum Development ASCD.

Wiliam, D. (2011). What is assessment for learning? *Studies in Educational Evaluation*, *37*(1), 3-14. https://doi.org/10.1016/j.stueduc.2011.03.001

Yan, Z., Lao, H., Panadero, E., Fernández-Castilla, B., Yang, L. & Yang, M. (2022). Effects of self-assessment and peer-assessment interventions on academic performance: a meta-analysis. *Educational Research Review*, *37*, 100484. https://doi.org/10.1016/j.edurev.2022.100484

YAZARLAR HAKKINDA | ABOUT AUTHORS

DOÇ. DR. SAMİ PEKTAŞ

Ankara Müzik ve Güzel Sanatlar Üniversitesi | pektassami[at]gmail.com

ORCID: 0000-0003-4753-6112

Doç. Dr. Sami PEKTAŞ, Ankara Müzik ve Güzel Sanatlar Üniversitesi Müzik ve Güzel Sanatlar Eğitim Fakültesi Eğitim Bilimleri Bölümü Eğitimde Ölçme ve Değerlendirme Anabilim Dalı'nda öğretim üyesidir. Lisans eğitimini Gazi Üniversitesi Kırşehir Eğitim Fakültesi Fen Bilgisi Öğretmenliği Anabilim Dalı'nda, yüksek lisans eğitimini Bolu Abant İzzet Baysal Üniversitesi'nde, doktora eğitimini ise Gazi Üniversitesi'nde Eğitimde Ölçme ve Değerlendirme alanında tamamlamıştır. Akademik kariyerinde Millî Eğitim Bakanlığı, Avrupa Birliği ve UNICEF destekli projelerde görev almış; program geliştirme süreçlerinde ölçme ve değerlendirme koordinatörlüğü yapmıştır. Ayrıca özel yetenekli öğrenciler ve eğitimde kalite odaklı projelerde akademik katkılar sunmuştur. Çalışmaları değişen madde fonksiyonu, test eşitleme, yapısal eşitlik modellemesi, ölçme değişmezliği ve genellenebilirlik kuramı üzerine yoğunlaşmaktadır.

ASSOC. PROF. DR. SAMİ PEKTAŞ

Ankara Music and Fine Arts University | pektassami[at]gmail.com

ORCID: 0000-0003-4753-6112

Assoc. Prof. Dr. Sami PEKTAŞ is a faculty member in the Department of Educational Sciences, Division of Measurement and Evaluation in Education, at the Faculty of Music and Fine Arts Education, Ankara Music and Fine Arts University. He completed his undergraduate education in Science Education at Gazi University Kırşehir Faculty of Education, his master's degree at Bolu Abant İzzet Baysal University, and his doctorate at Gazi University in Measurement and Evaluation in Education. Throughout his academic career, he has contributed to projects supported by the Ministry of National Education, the European Union and UNICEF, and has served as a measurement and evaluation coordinator in curriculum development processes. His research focuses on differential item functioning, test equating, structural equation modeling, measurement invariance and generalizability theory.

DR. ÖĞR. ÜYESİ SEMİH AŞİRET
Mersin Üniversitesi | semihasiret[at]mersin.edu.tr
ORCID: 0000-0002-0577-2603

Dr. Öğr. Üyesi Semih AŞİRET, Mersin Üniversitesi Eğitim Fakültesi Eğitim Bilimleri Bölümü Eğitimde Ölçme ve Değerlendirme Anabilim Dalı'nda öğretim üyesidir. Lisans eğitimini Orta Doğu Teknik Üniversitesi Bilgisayar ve Öğretim Teknolojileri Eğitimi Bölümü'nde, yüksek lisans ve doktora eğitimini ise Mersin Üniversitesi Eğitimde Ölçme ve Değerlendirme alanında tamamlamıştır. Uzun yıllar Millî Eğitim Bakanlığında görev yaptıktan sonra 2024 yılında Mersin Üniversitesi'nde öğretim üyesi olarak akademik çalışmalarını sürdürmeye başlamıştır. Millî Eğitim Bakanlığı projeleri ile Avrupa Birliği ve TÜBİTAK destekli projelerde araştırmacı olarak görev almış; öğretim programı geliştirme, becerilerin ölçülmesi ve değerlendirilmesi süreçlerine akademik katkılar sunmuştur. Çalışmaları bilişsel tanı modelleri, madde tepki kuramı, test eşitleme, bilgisayar ortamında bireye uyarlanmış testler, ölçme aracı geliştirme ile sınıf içi ölçme ve değerlendirme uygulamaları konularında yoğunlaşmaktadır.

ASST. PROF. DR. SEMİH AŞİRET
Mersin University | semihasiret[at]mersin.edu.tr
ORCID: 0000-0002-0577-2603

Asst. Prof. Dr. Semih AŞİRET is a faculty member in the Department of Educational Sciences, Division of Measurement and Evaluation in Education, at the Faculty of Education, Mersin University. He received his bachelor's degree from the Department of Computer Education and Instructional Technology at Middle East Technical University, and completed his master's and doctoral studies in Measurement and Evaluation in Education at Mersin University. After working for many years at the Ministry of National Education, he began working as a faculty member at Mersin University in 2024. He has served as a researcher in projects supported by the Ministry of National Education, the European Union, and TÜBİTAK, and has contributed academically to curriculum development, the measurement of skills, and assessment processes. His research focuses on cognitive diagnostic models, item response theory, test equating, computerized adaptive testing, est and scale development and classroom assessment practices.

14. BÖLÜM

ETİK, ADALET VE YASAL YÜKÜMLÜLÜKLER

ETHICS, JUSTICE AND LEGAL OBLIGATIONS

Doç. Dr. Asiye ŞENGÜL AVŞAR

ETİK, ADALET VE YASAL YÜKÜMLÜLÜKLER

Doç. Dr. Asiye ŞENGÜL AVŞAR
Recep Tayyip Erdoğan Üniversitesi

Öz

Ölçme ve değerlendirme bireylerin akademik başarıları, mesleki yeterlilikleri, gelişim düzeyleri hakkında kararlar verilmesi, kurumların eğitimsel etkililiklerinin ortaya konulması, eğitim politikalarının oluşturulması, sürdürülmesi veya iyileştirilmesi gibi birçok önemli süreçte belirleyici bir rol üstlenmektedir. Bu süreçlerde verilen kararların doğruluğu sadece geçerli ve güvenilir ölçme yapılmasıyla değil aynı zamanda güçlü bir etik anlayışla mümkün olmaktadır. Ölçme araçlarının geliştirilmesinden veya uyarlanmasından uygulanmasına, uygulanmasından puanlanmasına, puanlanmasından sonuçların duyurulmasına kadar tüm aşamalarda; ölçme ve değerlendirme uzmanları, eğitimciler, yöneticiler, sınav uygulayıcıları ve diğer paydaşların mesleki sorumluluklarının farkında olarak hareket etmeleri gerekmektedir. Bu kapsamda ölçme ve değerlendirme süreçlerinde bilimsel ilkelere bağlılık, yansızlık, hesap verebilirlik ve bireysel farklılıklara duyarlılık önemli etik ilkeler olarak öne çıkmaktadır. Etik ilkeler, ölçme ve değerlendirme süreçlerinde adalet, eşitlik, şeffaflık, gizlilik ve hesap verebilirliğin sağlanmasına rehberlik etmektedir. Ölçme ve değerlendirmede takip edilmesi gereken temel etik ilkeler; sınava giren bireylerin psikolojik, duygusal ve fiziksel güvenliğini korumayı, ilgili yasal düzenlemelere uygun hareket etmeyi, ölçme ve değerlendirme alanındaki mesleki yeterlilikleri sürdürmeyi, yalnızca yetkinlik alanı içinde hizmet sunmayı, geçerli uygulamaların anlaşılmasını desteklemeyi ve tüm mesleki sorumlulukları dürüstlük, bütünlük, özen ve adalet ilkeleri doğrultusunda yerine getirmeyi temel almaktadır. Kitabın bu bölümünde, ilk olarak ölçme ve değerlendirme süreçlerinde paydaşları bağlayıcı kılan etik sorumluluklara yer verilmiştir. Daha sonra ölçme ve değerlendirmede adalet kavramının temel bileşenleri olan yansızlık, kapsayıcılık, hesap verebilirlik ve şeffaflık unsurları açıklanmıştır. Son bölümde ise ölçme ve değerlendirme uygulamalarını düzenleyen yasal çerçeve, ilgili mevzuat hükümleri ve paydaşların yerine getirmesi gereken hukuki yükümlülükler incelenmiştir. Böylece ölçme ve değerlendirme süreçlerinin etik ve hukuki boyutlarıyla kapsamlı bütüncül olarak ele alınması amaçlanmıştır.

Anahtar Kelimeler

Etik Sorumluluk, Yansızlık, Kapsayıcılık, Hesap Verebilirlik, Yasal Yükümlülükler

ETHICS, JUSTICE AND LEGAL OBLIGATIONS

Assoc. Prof. Asiye ŞENGÜL AVŞAR
Recep Tayyip Erdoğan University

Abstract

Measurement and evaluation play a crucial role in numerous critical processes, such as making decisions regarding individuals' academic achievements, professional competencies, and developmental levels; demonstrating the educational effectiveness of institutions; and formulating, maintaining, or improving educational policies. The accuracy of decisions made during these processes depends not only on conducting valid and reliable assessments but also on a strong ethical mindset. At every stage from the development or adaptation of assessment tools to their administration, scoring, and the reporting of results measurement and evaluation specialists, educators, administrators, test administrators, and other stakeholders must act with an awareness of their professional responsibilities. In this context, adherence to scientific principles, unbiasedness, accountability, and sensitivity to individual differences emerge as key ethical principles in measurement and evaluation processes. These ethical principles guide the assurance of fairness, equity, transparency, confidentiality, and accountability within these processes. The fundamental ethical principles to be followed in measurement and evaluation are grounded in protecting the psychological, emotional, and physical well-being of examinees; complying with relevant legal regulations; maintaining professional competence in the field; providing services only within one's area of expertise; promoting an understanding of valid practices; and fulfilling all professional responsibilities in accordance with the principles of honesty, integrity, diligence, and fairness. This chapter of the book begins by addressing the ethical responsibilities that are binding upon stakeholders in measurement and evaluation processes. Subsequently, the elements constituting the core components of fairness in measurement and evaluation namely unbiasedness, inclusivity, accountability, and transparency are explained. In the final section, the legal framework governing assessment and evaluation practices, the provisions of relevant legislation, and the legal obligations to be fulfilled by stakeholders are examined. Thus, the aim is to address measurement and evaluation processes in a comprehensive and holistic manner, encompassing both their ethical and legal dimensions.

Keywords

Ethical Responsibility, Unbiasedness, Inclusivity, Accountability, Legal Obligations

1. GİRİŞ

Ölçme ve değerlendirme süreçlerinin tüm aşamalarında uluslararası kabul edilen belli standartların takip edilmesi önemlidir. Ölçme ve değerlendirme süreçleri, yalnızca takip edilmesi gereken adımları içeren teknik bir alan değil, aynı zamanda profesyonel ve etik bir sorumluluk alanıdır. Ölçme ve değerlendirmeye ilişkin standartlar; geçerlik ve güvenirlik kanıtlarının sağlanması, test geliştirme süreçlerinin paydaşları bilgilendirecek şekilde şeffaf bir biçimde planlanması, kapsayıcılığa dikkat edilerek ölçme araçlarının geliştirilmesi ve olası hataların giderilmesi şeklinde özetlenebilir. Ölçme ve değerlendirmede bir diğer konu ise kapsayıcılıktır. Kapsayıcı bir ölçme ve değerlendirme süreci için; eşit öğrenme ve eşit test fırsatları sağlanmalı, dilsel ve kültürel farklılıklara dikkat edilmeli, yanlılık içeren maddeler belirlenmeli ve bu maddelerden uzak durulmalı, özel gereksinimli bireyler için erişilebilir testler ve test düzenlemeleri yapılmalıdır. Ayrıca bilgisayar ve teknoloji tabanlı testlerin formatları ve kullanıcı arayüzleri evrensel tasarım ilkelerine uygun olarak düzenlenmelidir. Kapsayıcılığın yalnızca özel gereksinimli bireylerle ilgili bir konu olmadığı, ölçme araçlarında çeşitlilik sağlanarak bireylerin kendilerini ve yeteneklerini en iyi ifade edebilecekleri süreçler oluşturmayı da içeren çok daha geniş bir anlam taşıdığı unutulmamalıdır. Yasal yükümlülükler ölçme ve değerlendirme için önemli bir konudur. Örneğin, yüksek riskli testler, bireyler, okullar ya da kurumlar üzerinde ciddi sonuçlar doğuran ölçme araçlarıdır. Bu testlerin geliştirilmesinden uygulanmasına, puanlanmasından sonuçların duyurulmasına kadar tüm süreçlerde titizlikle çalışılması gerekmektedir. Her aşamada tüm paydaşları bağlayıcı yasal yükümlülüklerin bulunması, süreçleri güçlendirerek şeffaf ve hesap verebilir hale getirmektedir. Özetle, tüm paydaşları ilgilendiren ve önemli bir sorumluluk alanı olan yasal yükümlülükler, hesap verebilirliği ve şeffaflığı sağlayarak ölçme ve değerlendirme süreçlerini sağlamlaştırmaktadır. Kitabın bu bölümde etik bağlamında ölçme ve değerlendirme faaliyetlerinde bulunan kişilerin mesleki sorumlulukları detaylı bir şekilde açıklanmıştır. Ardından ölçme ve değerlendirmede adaletin bileşenleri olan ölçmede adalet, ölçmede yansızlık, ölçmede kapsayıcılık, ölçmede hesap verebilirlik ve şeffaflık konuları hakkında bilgi verilmiştir. Daha sonra ölçme ve değerlendirmede yasal yükümlülükler açıklanmıştır.

2. ÖLÇME VE DEĞERLENDİRMEDE ETİK

Etik; bir kişi veya grup tarafından kabul edilen ya da belli bir alanda uygun görülen ahlaken doğru davranış ilkeleri/kuralları olarak tanımlanmaktadır (*American Psychological Association [APA]*, n.d.-a). Örneğin Amerikan Eğitim Araştırmaları Derneği (*American Educational Research Association [AERA]*), Amerikan Psikoloji Derneği (*American Psychological Association*), Amerikan Sosyoloji Derneği (*American Sociological Association*), Dünya Tabipler Birliği (*World Medical Association*) gibi önemli kuruluşlar kendi mesleki etik kurallarını (*code of ethics*) yayınlamışlardır. Etik standartlar, eğitim araştırmacılarının etik davranışlarına ilişkin kurallarını belirler (AERA, 2011).

Bu bölümde Eğitimde Ölçme Ulusal Konseyi (*National Council on Measurement in Education [NCME])* ve Etik Kuralları Geliştirme Geçici Komitesi (*Ad Hoc Committee on the Development of a Code of Ethics*) tarafından 1995 yılında hazırlan "Eğitimde Ölçme Alanında Mesleki Sorumluluk Kuralları" (*Code of Professional Responsibilities in Educational Measurement*) hakkında bilgi verilmiştir. Hazırlanan bu kurallar 2023 yılında NCME Standartlar (*NCME Standards*) ve Test Kullanımı Komitesi (*Test Use Committee*) tarafından yeniden düzenlenerek kullanıma sunulmuştur (NCME, 2023).

Hazırlanan ve yeniden düzenlenen bu kurallar NCME (2023) üyelerinin;

- Psikolojik ve duygusal iyi oluşları da dâhil olmak üzere sınava giren tüm bireylerin sağlığını ve güvenliğini korumak,
- Mesleki faaliyetlerin yürütülmesiyle ilgili yerel, ulusal ve uluslararası yasalar hakkında bilgi sahibi olmak ve bunlara uygun davranmak,
- Eğitimde ölçme ve değerlendirmede mesleki yeterliliklerini korumak ve geliştirmek,
- Yalnızca kendi yeterlilik ve deneyim alanlarında ölçme ve değerlendirme hizmetleri sunmak ve mesleki yeterliliklerin tam olarak açıklanmasını sağlamak,
- Eğitimde geçerli ölçme ve değerlendirme uygulamalarının anlaşılmasını desteklemek,
- Eğitim hizmetleri sunan eğitim kurumları ve kuruluşları içinde en yüksek davranış standartlarına uymak ve mesleki açıdan sorumlu davranışı teşvik etmek,
- Tüm mesleki sorumlulukları dürüstlük, bütünlük, gerekli özen ve adaletle yerine getirmek beklentileri üzerine geliştirilmiştir.

Bu beklentilere göre *Eğitimde Ölçme ve Değerlendirme Alanında Mesleki Sorumluluk Kuralları* (NCME, 2023);

- Ölçme ve değerlendirme araçlarını geliştirmek,
- Ölçme ve değerlendirme araçlarını pazarlamak ve satışını yapmak,
- Uygun ölçme ve değerlendirme araçlarını seçmek,
- Ölçme ve değerlendirme süreçlerini yönetmek, ölçme sonuçlarını puanlamak,
- Ölçme sonuçlarını; yorumlamak, kullanmak ve açıklamak,
- Ölçme ve değerlendirme konusunda başkalarına eğitim vermek
- Programları değerlendirmek ve ölçme ve değerlendirme araçlarıyla araştırma yürütmek olmak üzere sekiz ana alanda belirlenmiştir.

Bu kuralların amacı, eğitimde ölçme ve değerlendirme faaliyetlerinde bulunan kişilerin mesleki sorumluluklarını tanımlamaktır. Kitabın bu bölümünde NCME (2023)[1] tarafından ölçme ve değerlendirme faaliyetlerinin sekiz ana alanında belirlenen mesleki sorumluluklar sırasıyla açıklanmıştır.

2.1. Ölçme ve Değerlendirme Araçlarını Geliştirenler ve Değerlendirme Hizmetlerini Verenlerin Sorumlulukları

Ölçme ve değerlendirme ürünleri ve hizmetlerini geliştirenler, en yüksek kalitede ölçme araçlarını geliştirme konusunda mesleki bir sorumluluğa sahiptir. Bu sorumluluklar sırasıyla aşağıda açıklanmıştır (NCME, 2023):

1. Ölçme ve değerlendirme ürün ve hizmetlerinin profesyonel, teknik ve yasal standartları karşılayacak şekilde geliştirilmesi sağlanmalıdır.
2. Ölçme ve değerlendirme ürün ve hizmetleri; cinsiyet, etnik köken, ırk, sosyoekonomik durum, özel gereksinimli olma, din, yaş veya ulusal köken gibi özelliklerden bağımsız olmalı, ölçülen yapılar tüm adaylar (testi alanlar) için benzerlik taşımalıdır. Diğer bir ifadeyle belirtilen bu özelliklerden bağımsız ölçme yapıldığına yönelik kanıtlar sağlanmalıdır.
3. Evrensel tasarım ilkelerini kullanarak ölçülen yapılar üzerindeki etkileri en aza indiren ve erişilebilirliği sağlayan ölçme ve değerlendirme araçları geliştirilmedir. Örneğin özel gereksinimli sınav katılımcıları için uygun düzenlemeler sağlanmalıdır. Başka bir örnek olarak ölçmenin amacı dil yeterliliğini ölçmek değilse değerlendirmede dilde sınırlı yeterliliğe sahip olanlar için gerekli düzenlemeler yapılmalıdır.
4. Test geliştiricilerin tutum, davranış ya da yargılarını etkileyebilecek gerçek veya potansiyel çıkar çatışmaları ilgili taraflara açıklanmalıdır.
5. Ölçme ve değerlendirme ürün ve hizmetlerinde yerel, ulusal ve uluslararası yasalara uygun olarak telif hakkıyla korunan materyaller kullanılmalıdır.
6. Teknik bir kılavuzda; güvenirlik, geçerlik, puanlama ve raporlama süreçleri ile ölçme ve değerlendirmenin diğer ilgili özellikleri hakkında güncel bilgiler de dâhil olmak üzere, ölçme araçlarının geliştirilmesi ve puanlanması için atılan adımlar hakkında detaylı bilgi sağlanmalıdır.
7. Teknik bilgiler, mümkün olduğunca paydaşlara uygun açıklamalarla iletilmelidir.

---

[1] Bu bölümde açıklanan tüm etik kurallar, NCME (2023) tarafından açıklanan kurallardır. NCME (2023) bu etik kuralların uygun atıf verme şartı ve eğitim amacıyla kullanılmak üzere herhangi bir izin alınmaksızın ve ücret ödemeksizin yayınlanıp çoğaltılabileceğini belirtmektedir [bkz: National Council on Measurement in Education. (2023). *Code of professional responsibilities in educational measurement* (2. baskı)].

8. Ölçme ve değerlendirme süreci boyunca; test kaydı, yönetimi ve puanların raporlaması için mekanizmalar geliştirmek de dâhil olmak kişisel olarak tanımlanabilir bilgilerin gizliliği ve güvenliği sağlanmalıdır.
9. Değerlendirme sonuçlarının uygun ve amaçlanan şekilde yorumlanmasını ve kullanılmasını destekleyecek puan raporları ve destek materyalleri geliştirilmelidir.
10. Değerlendirmenin amacı açık ve belirgin bir dille açıklanmalıdır. Değerlendirme sonuçlarının yanlış kullanımına ya da yanlış yorumlanmasına karşı test kullanıcıları uyarılmalıdır.
11. Test hazırlığı ve program destek materyallerinde, bir ölçme aracı ya da sonuçların kullanımı ve yorumlanması hakkında yanlış veya dayanağı olmayan iddialardan kaçınılmalıdır.
12. Değerlendirmelerde ya da değerlendirmeleri destekleyici materyallerde yer alan önemli yanlışlıklar, mümkün olan en kısa sürede düzeltilmelidir.

Bu standartlar özetlendiğinde ölçme ve değerlendirme süreçlerinde profesyonel bir bakış açısıyla geçerlik, güvenirlik kanıtlarının sağlanması, test geliştirme süreçlerinin şeffaf bir şekilde paydaşları da bilgilendirecek şekilde planlanması, kapsayıcılığa dikkat edilerek ölçme araçlarının geliştirilmesi ve eğer ölçme süreçlerinin herhangi bir aşamasında hata yapıldıysa gerekli düzeltmelerin gerçekleştirilmesi gerektiği vurgulanmaktadır. Türkiye'de ÖSYM başkanlığı tarafından yürütülen sınav süreçlerinde; test geliştirme, uygulama ve testin sonuçlarını değerlendirme (puanlama ve yerleştirme) aşamalarına ilişkin yukarıda belirtilen sorumluluklara bağlı kalınarak hareket edilmektedir.

2.2. Ölçme ve Değerlendirme Ürün ve Hizmetlerini Pazarlayan ve Satışını Yapanların Sorumlulukları

Testler ve diğer araçlar, puanlama, sınav hazırlık, danışmanlık ve sınav yorumlama hizmetleri gibi ölçme-değerlendirme ürün ve hizmetlerinin pazarlanması; bunları kullanmayı düşünenler için doğru, eksiksiz ve amaca uygun bilgilere dayanmalıdır. Sorumluluk ekipler arasında paylaşılmalıdır. Örneğin psikometristler değerlendirmelerin teknik özelliklerinin doğru şekilde temsil edilmesini sağlarken satış ve pazarlama çalışanları, müşterilerle etkileşime girmeden önce bu bilgiler hakkında bilgilendirilmelidir. Değerlendirme ürün ve hizmetlerini pazarlayan ve satışını yapan kişilerin mesleki sorumlulukları şunlardır (NCME, 2023):

1. Potansiyel alıcılara ölçme-değerlendirme ürünleri ve hizmetleri ile bunların önerilen kullanımları ve sınırlamaları hakkında doğru bilgiler sunmalıdır.

2. Potansiyel kullanıcılara; ölçme ve değerlendirme ürünlerinin ve hizmetlerinin bilinen yanlış kullanımları konusunda bilgi verilmeli ve yanlış kullanımdan nasıl kaçınılacağına dair önerilerde bulunulmalıdır.
3. Değerlendirme ürün ve hizmetlerini pazarlayan ve satışını yapanlar, sundukları ölçme-değerlendirme ürünleri ve hizmetleri ile bunların eğitimdeki uygun kullanımları hakkında güncel bir anlayışa sahip olmalıdır.
4. Doğru ve uygun bir seçim kararını etkileyebilecek, ölçme ve değerlendirme ürünleri ve hizmetleri hakkındaki bilgiler kasıtlı olarak saklanmamalıdır.
5. Ölçme ve değerlendirme ürünleri ve hizmetleriyle ilgili tüm iddialar, kamuya açık olarak sunulan bilgilerin geçerli yorumlarına dayalı olmalıdır.
6. Ölçme ve değerlendirme ürünleri ve hizmetleri satın alma konusunda nitelikli kullanıcılara eşit fırsat tanınmalıdır.
7. Ölçme ve değerlendirme ürünleri ve hizmetleri için makul ücretler uygulanmalıdır.
8. Ölçme ve değerlendirme ürünleri ve hizmetleriyle ilgili tüm geçerli ücretler, herhangi bir satın alma ya da kullanımdan önce, potansiyel kullanıcılara bildirilmelidir.
9. Ölçme ve değerlendirme ürün ve hizmetleri için gereken ücretleri ödeyemedikleri gerekçesiyle değerlendirme sonuçlarından doğan fırsat ya da yararlardan hiçbir bireyin mahrum bırakılmaması için çaba gösterilmelidir.
10. Ölçme ve değerlendirme ürün ve hizmetlerinin satışına ilişkin ölçütler belirlenmelidir. Örneğin bir değerlendirme ürün ve hizmetinin satışı; ölçme aracını uygun şekilde kullanma ve sonuçlarını uygun şekilde yorumlamada uzman kişilere yapılmalıdır.
11. Ölçme ve değerlendirme ürün ve hizmetlerinin kullanıcıları tarafından onaylandığını ima eden bilgiler, yalnızca kullanıcıların izniyle yayınlanmalıdır.
12. Başka bir kurum ya da kuruluştan resmi bir onay alınmadığı sürece, ölçme ve değerlendirme ürünleri ve hizmetlerinin söz konusu kuruluş tarafından onaylandığına dair iddialarda bulunulmamalıdır.
13. Bireylerin gerçek başarı düzeylerini yanlış yansıtan puanlar almasına neden olabilecek test hazırlık ürünlerinin ve hizmetlerinin pazarlaması yapılmamalıdır.

Ölçme ve değerlendirme ürün ve hizmetlerini pazarlayan ve satışını yapanların uymaları gereken bu kuralların tamamı ÖSYM tarafından dikkate alınmaktadır. ÖSYM testi alan bireylerin kişisel bilgilerini saklama konusunda son derece titiz hareket etmekte, sınavlarda görev alacak kişileri özenle seçmekte, sınavın işleyişinde sorunlara yol açan

ve/veya adayların sınavlarının iptal edilmesine yol açabilecek eylemlerde bulunan görevlilere çeşitli yaptırımlar uygulamakta ve eğitimde fırsat eşitliği ilkesini göz önünde bulundurarak sınav ücretlerini makul düzeyde tutmaya çalışmaktadır. ÖSYM kâr amacı gütmeyen bir kamu kurumudur. Dolayısıyla ÖSYM'de geliştirilen testlerin satış ve pazarlaması yapılmaz. ÖSYM yüksek standartlarda sınavlar hazırlayarak öğrencilerin ve adayların doğru ve adil şekilde değerlendirilmesine yönelik çalışmalarını sürdürmektedir.

2.3. Ölçme ve Değerlendirme Ürünlerini ve Hizmetlerini Seçenlerin Sorumlulukları

Eğitim ortamlarında kullanılmak üzere ölçme ve değerlendirme ürün ve hizmetlerini seçen ya da başkalarının bunu yapmasına yardımcı olan kişiler, ölçme ve değerlendirme araçlarının (ürün ve hizmetlerinin), kullanım amaçlarına uygun olduğundan emin olmak için önemli mesleki sorumluluklara sahiptir. Ölçme ve değerlendirme ürün ve hizmetlerini seçen kişiler aşağıda belirtilen unsurları asgari düzeyde yerine getirmeye çalışmalıdır (NCME, 2023).

1. Ölçme-değerlendirme ürünleri ve hizmetlerini değerlendirmek için gerekli kavramları anlayan paydaşlar sürece dâhil edilmelidir.
2. Amaçlanan kullanımlar için geçerli olacak değerlendirme stratejileri ve ölçme araçları hakkında kapsamlı ve titiz bir inceleme yapılmalıdır.
3. Ölçme ve değerlendirme sonuçlarını karşılaştırırken adil ve tutarlı ölçütler (örneğin dereceli puanlama anahtarları) kullanılmalıdır.
4. Ölçme araçlarını önerirken ya da seçerken teknik kalite ve fayda konusunda kanıtlanmış kamuya açık araçlara dikkate edilmelidir.
5. Satın alınması düşünülen ölçme araçlarının; yazarlarıyla, test yayıncılarıyla veya diğer kişilerle herhangi bir ilişki ya da bağlantı varsa açıklanmalıdır. Bu tür ilişkilerin ölçme araçlarının seçim sürecinin nesnelliğini etkileyebileceği durumlarda katılımdan kaçınılmalıdır.
6. Başkalarının ölçme-değerlendirme araçlarının seçimi sürecine uygunsuz etki ya da baskı girişimleri olduğunda bu durum derhal bildirilmelidir.
7. Karar vericilere ve testin gelecekteki olası kullanıcılarına; ölçme-değerlendirme aracının amaçlanan kullanım için uygunluğu, kullanımın olası sonuçları, sınava/teste girenlerin haklarının korunması, göreceli maliyetler, ölçme-değerlendirme sürecinin yürütülmesi için gereken materyaller ve hizmetler hakkında bilgi verilmelidir. Ayrıca değerlendirme sonuçlarının; bilinen sınırlılıkları, olası yanlış kullanımları ve yanlış yorumları hakkında bilgilendirme yapılmalıdır.

8. İncelenen ölçme-değerlendirme araçları ile ilgili tüm güvenlik politika ve usullerine/prosedürlerine (örneğin değerlendirme içeriğiyle ilgili gizlilik sözleşmeleri) uyulmalıdır.
9. Herhangi bir ölçme ve değerlendirme aracının; amacı dışında kullanılmasına veya yanlış uygulanmasına, puanlanmasına ya da yorumlanmasına karşı durulmalıdır.
10. Ölçme uygulamasından elde edilen puanların yapay biçimde yükselmesine neden olacak uygulamalardan uzak durulmalıdır.

Yukarıda listelenen "ölçme-değerlendirme ürün ve hizmetlerini seçenlerin eğitimde ölçme-değerlendirme okuryazarlığına sahip kişilerden oluşması, ölçme araçlarının seçiminde amaca uygunluğa dikkat edilmesi ve gizlilik kurallarına uyulması" gibi ilkelerin tamamı ÖSYM tarafından göz önünde bulundurulmaktadır. ÖSYM sınavlarda kullandığı testleri kendi öz kaynaklarıyla geliştiren bir devlet kurumudur. Test geliştirme aşamalarında soru hazırlama ihtiyacını gerçek ve tüzel kişiler ile kamu kurum ve kuruluşlarından hizmet alımı suretiyle karşılamaktadır (ÖSYM, 2011). Tüm bu süreçlerde en yüksek güvenlik koşullarını işe koşmakta, nitelikli ölçme yapma hedefine sıkı sıkıya bağlı hareket etmektedir.

2.4. Ölçme ve Değerlendirme Süreçlerini Yönetenlerin Sorumlulukları

Bireyleri ölçme ve değerlendirme süreçlerine hazırlayanlardan sınavların uygulanmasında doğrudan ya da dolaylı olarak görev alan öğretmenler ve yöneticilere, sınav evraklarını taşıyan kuryelerden sınav binalarında görev alan emniyet personeline ölçme süreçlerine dâhil olan tüm kişiler sınavların adil ve uygun bir şekilde uygulanmasını sağlamada önemli bir role sahiptir. Ölçme ve değerlendirme süreçlerini yönetenlerin sorumlulukları sırasıyla şu şekilde açıklanabilir (NCME, 2023):

1. Sınav uygulanmadan önce sınava girecek kişilere; sınavın amacı, kullanımı ve olası sonuçlarını içerek şekilde sınavın nasıl değerlendirileceği ya da puanlanacağı; sonuçların nasıl dosyalanacağı, sonuçlara kimin erişebileceği ve sonuçların nasıl ulaştırılacağı (elektronik posta, çevrimiçi ortam, posta yolu gibi) hakkında bilgi verilmelidir. Ayrıca sınavı alan kişilere; sınav öncesinde, sırasında ve sonrasında sahip oldukları haklara ilişkin bilgiler verilmelidir.
2. Ölçme ve değerlendirme süreçleri nitelikli (gerekli eğitimleri alan, lisans sahibi ya da sertifika sahibi) kişiler tarafından yürütülmelidir.
3. Ölçme-değerlendirme içeriğinin ve süreçlerinin güvenliğini sağlamak, kopya girişimlerini engellemek için ölçme-değerlendirme uygulamaları öncesinde, sırasında ve sonrasında gerekli önlemler alınmalıdır.

4. Ölçme-değerlendirme uygulamaları başlamadan önce uygulama süreçleri (kurallar/yönergeler) anlaşılmalıdır.
5. Bilgisayar tabanlı bir sınavın uygulamasından önce; sınav cihazlarının sistem gereksinimlerini karşıladığı, ağ bant genişliği gereksinimlerinin sağlandığı, gerekli tüm yazılımların yüklü olduğu ve mevcut tüm sistem hazırlık kontrollerinin yapıldığı kontrol edilmelidir.
6. Standartlaştırılmış testler, kuralları belli olan işleyiş ve koşullara göre uygulanmalıdır. Herhangi bir aykırı durum oluştuğunda ilgili kişiler bilgilendirilmelidir.
7. Uygun olan tüm adaylara sınav fırsatları sağlanmalıdır. Örneğin adaylar düşük performans göstereceği beklentisiyle sınavlardan dışlanmamalıdır.
8. Ölçme-değerlendirme uygulaması sırasında sonuçları geçersiz kılabilecek her türlü koşuldan kaçınılmalıdır.
9. Belgelenmiş özel gereksinimi ya da ihtiyaçları olan kişilere yönelik, ölçme-değerlendirme süreçlerinin yönetimi için makul ve izin verilen tüm düzenlemeler sağlanmalı ve kayıt altına alınmalıdır.
10. Ölçme-değerlendirme süreçleri bilgisayar tabanlı ölçme platformları üzerinden gerçekleştiriliyorsa bu uygulamalara aşina olmayan adaylara, ölçme-değerlendirmenin içeriğini ve biçimini tanımaları için test pratiği yapma imkânı sunulmalıdır.
11. Ölçme-değerlendirme uygulamasından önce ve uygulama sırasında belirlenen zamanlarda bireylere uygulama süreçleri ya da talimatları hakkında soru sormaları için uygun fırsatlar sağlanmalıdır.
12. Sınavı alan kişilerin mahremiyeti ve adil süreç hakları (ölçme-değerlendirme sürecinde kuralların adil işlemesi, kişinin hakkını savunabilmesi ve şeffaf, eşit bir uygulamadan geçmesinin güvence altına alması) korunmalıdır.
13. Bireylerin veya grupların gerçek başarı düzeylerini yansıtmayan puanlar almasına izin verecek ya da bunu teşvik edecek eylem ve koşullardan uzak durulmalıdır.

Sıralanan ilkelerin eksiksiz bir biçimde hayata geçirilmesinde ölçme ve değerlendirme süreçlerini yönetenlere sorumluluklar düşmektedir. Sınavı yöneten kurumlar ve kişiler ölçme-değerlendirme süreçlerini yüksek standartlarda yürütmeli, kapsayıcılığı gözeterek özel gereksinimli ya da özel ihtiyacı olan kişilere uygun ortamlar sağlamalı, bilgisayar tabanlı değerlendirmelerde bazı katılımcıların bilgisayara aşina olmamasından kaynaklanabilecek yanlılıkların önüne geçmeye dönük tedbirler almalıdır. Test süreçlerini yönetmede profesyonel bir kurum olan ÖSYM söz konusu

gereklikler doğrultusunda tüm sınav süreçlerini işletmektedir. Bütün sınav görevlileri; sınav öncesinde, sınav sırasında ve sınav sonrasında yapmaları gerekenler hakkında yönergelerle bilgilendirilmektedir. Benzer şekilde adaylara da sınavda uymaları gereken kuralların tamamı tebliği edilmektedir. Tüm bu süreçler kanun ve yönetmeliklerle güvence altına alınmaktadır (ÖSYM, 2025). ÖSYM kapsayıcılık ilkesi doğrultusunda özel gereksinimli adaylara yönelik destek hizmetleri (işaretleyici ve/veya okuyucu yardımı, ek süre ve ekipman desteği vb.) sunmaktadır.

2.5. Ölçme Sonuçlarını Puanlayanların Sorumlulukları

Sınav sonuçlarının sınava girenler, veliler, eğitimciler ve diğer paydaşlara doğru ve zamanında bildirilebilmesi için eğitsel değerlendirmeler uygun ve etkili şekilde puanlanmalıdır. Ölçme sonuçlarını puanlayan ve puanlara dayalı olarak rapor hazırlayan kişilerin mesleki sorumlulukları aşağıda sıralanmıştır (NCME, 2023):

1. Test kullanıcılarına ölçme sonuçlarının nasıl puanlanacağı hakkında (örneğin; raporlama takvimi, puanlama süreci, puanlama yaklaşımının mantıksal gerekçesi, teknik özellikler, kalite kontrol süreçleri, raporlama biçimi ve varsa bu hizmetlere ilişkin ücretler) eksiksiz ve doğru bilgi sağlanmalıdır.
2. Yapılandırılmış yanıt görevlerinde (açık uçlu maddeler gibi) puanlamada bütünlüğünü sağlamak için kalite süreçleri uygulanmalıdır. Bu kapsamda puanlamadan önce puanlayıcı nitelikleri kontrol edilmeli ve puanlayıcılara eğitim verilmeli, puanlama boyunca puanlamanın doğruluğu ve puanlayıcılar arası uyum izlenmeli ve puanlama sonrasında son bir değerlendirme yapılmalıdır.
3. Ölçme ve değerlendirme amaçlarıyla ilgisi olmayan faktörlerin, puanlama üzerindeki etkisi en aza indirilmelidir.
4. Planlanan puanlama ve raporlama hizmetinde ya da programında herhangi bir sapma meydana gelirse test kullanıcıları derhal bilgilendirilmelidir. Test kullanıcılarıyla bir çözüm üzerinde görüşme sağlanmalıdır.
5. Puanlara dayalı olarak yapılan çıkarımları etkileyebilecek hatalar belirlenirse düzeltilmiş puan sonuçları, sınav katılımcısına ya da test kullanıcısına mümkün olan en kısa sürede sağlanmalıdır.
6. Bireyleri tanımlayan bilgilerin gizliliği; yerel, ulusal ve uluslararası yasalarca belirlendiği şekilde korunmalıdır.
7. Değerlendirme sonuçlarının özeti; yalnızca yerel, ulusal veya uluslararası yasalar uyarınca bu bilgiye erişim hakkına sahip kişilere ya da puanlama hizmetlerini talep eden tarafça yetkilendirilmiş kişilere açıklanmalıdır.
8. Gerektiğinde değerlendirme sonuçlarına itiraz ve puanlamanın yeniden yapılabilmesi için adil ve uygun bir itiraz süreci oluşturulmalıdır.

Ölçme sonuçlarının puanlanması, puanlarının değerlendirilmesi ve yayınlanması ile ilgili yukarıda değinilen sorumluluklar ÖSYM'nin sınav süreçlerine de yön vermektedir. ÖSYM objektifliği ve kişisel bilgilerin gizliliğini sağlamaya azami gayret göstermekte, puanlama için hazırladığı veriler ile raporları içermesi gereken teknik detaylara göre şekillendirmektedir. ÖSYM test geliştiren, geliştirdiği testleri uygulayan, test sonuçlarını değerlendiren ve değerlendirme sonuçlarının raporlarını hazırlayan bir kurumdur. Adayların test puanları, Sınav Hizmetleri Daire Başkanlığı tarafından hesaplanmaktadır. Puanlamanın nasıl yapıldığı (düzeltme formülü uygulanması, ortaöğretim başarı puanın ağırlığı gibi) her bir sınav için test yönergesinde açıklanmaktadır. Puanlara itiraz yolu da açıktır. Adayların testlere verdikleri cevaplar bilgi işlem yöntemleriyle incelenerek gerektiğinde (kopya çekme durumu olmasında) bireysel ya da toplu olarak sınavlar iptal edilebilmektedir. Puanlamada ya da geliştirilen maddelerde hata olması halinde de itiraz yolu açıktır. Gerekirse soru iptali gerçekleştirilmekte, bu durum kamuoyu ile açık ve şeffaf olarak paylaşılmaktadır. Ayrıca ÖSYM hazırladığı detaylı sınav raporlarında sınavların istatistikleri (yıllara göre karşılaştırmalı olarak) ile sınav sonuçlarından elde edilen puanlara göre ulaşılan geçerlik ve güvenirlik kanıtlarını tüm paydaşlara sunmaktadır.

2.6. Ölçme ve Değerlendirme Sonuçlarını Yorumlayan, Kullanan ve Açıklayanların Sorumlulukları

Ölçme-değerlendirme sonuçlarının yorumlanmasında, kullanımında ve açıklanmasında geçerli çıkarımlar desteklenmeli; geçersiz çıkarımların engellenmesi sağlanmalıdır. Ölçme sonuçlarını yorumlayan, kullanan ve açıklayan kişilerin sorumlulukları aşağıda maddeler halinde açıklanmıştır (NCME, 2023):

1. Ölçme-değerlendirme sonuçları açıklanırken süreçteki sınırlılıklar göz önüne alınmalıdır. Elde edilen ölçümlerin kullanımının ortaya çıkaracağı sonuçların farkında olunmalıdır. Sonuçlar bilgiye dayalı, nesnel ve adil bir şekilde yorumlanmalı, kullanılmalı ve açıklanmalıdır.
2. Ölçme-değerlendirme sonuçlarını alan kişilere (adaylar, veliler, yasal temsilciler; öğretmenler, eğitimciler, eğitim kurumu yöneticileri, danışmanlar ve rehberlik personeli; işverenler, insan kaynakları yetkilileri [işe alım süreçlerinde]; akademik ya da profesyonel değerlendirme yapan kurumlar; yasal merciler ya da ilgili resmî kurumlar) yürütülen ölçme ve değerlendirme işleminin amaçları, sınırlılıkları, kullanımları ve puanların doğru şekilde yorumlanması için gerekli bilgiler sağlanmalıdır.
3. Puanların raporunu alan kişilere, raporlanan tüm puanların yazılı bir açıklaması yapılmalıdır. Bu açıklama doğru/uygun çıkarımları ve olası yanlış yorumları içeren anlaşılması kolay bir metin olmalıdır.
4. Ölçme ve değerlendirme sonuçları hedef kitleye zamanında açıklanmalıdır.

5. Ölçme-değerlendirme sonuçlarının yorumlanmasında kullanılan normların ya da performans standartlarının yeterliliği ile uygunluğu değerlendirilmeli ve açıklanmalıdır.
6. Ölçme-değerlendirme sürecine dâhil olan taraflar, değerlendirme sonuçlarının kendilerini nasıl etkileyebileceği konusunda bilgilendirilmelidir.
7. Eğitimle ilgili kararlar alınırken kişiler ya da programlar hakkında birden fazla kaynaktan bilgi almaya özen gösterilmeli, konuyla ilgili çeşitli bilgi türlerinden faydalanılmalıdır.
8. Ölçme-değerlendirme sonuçları hakkında; yanlış raporlar, asılsız iddialar, uygun olmayan yorumlar ya da bunların dışında, gerçeğe aykırı ve yanıltıcı diğer ifadeler kullanılmamalı; başkalarının da bu tür davranışlarda bulunması engellenmelidir.
9. Adaylar ve paydaşlar değerlendirme sonuçlarının arşivde saklanıp saklanmayacağı ya da ne kadar süreyle saklanacağı konusunda bildirilmelidir. İtiraz ve yeniden puanlama sürecine, adayların ve ilgili diğer kişilerin değerlendirme sonuçlarına erişme haklarına, ayrıca bu hakları nasıl kullanılabileceklerine ilişkin bilgiler açıklanmalıdır.
10. Ölçme-değerlendirme bilgileriyle ilgili açıkça görülen kötüye kullanımlar, süreçten sorumlu kişilere bildirilmelidir.
11. Ölçme-değerlendirme sürecine dâhil olan bireylerin ve kurumların gizlilik hakları korunmalıdır.

ÖSYM ölçme ve değerlendirme sonuçlarının yorumlanmasında, raporlanmasında ve saklanmasında dikkat edilmesi gereken yukarıdaki ilkelere bağlı kalmakta; düzenlediği sınavların sonuçlarını değerlendirirken takip ettiği yöntemleri açık bir şekilde paylaşmakta; sınav sonuçlarının arşivlenmesi, ölçme ve değerlendirme sonuçlarının zamanında açıklanması, kişisel bilgilerin korunması gibi konularda üzerine düşen sorumluluğu şeffaf bir şekilde yerine getirmektedir.

2.7. Ölçme ve Değerlendirme Konusunda Başkalarına Eğitim Verenlerin Sorumlulukları

Eğitimde ölçme-değerlendirme konusunda başkalarına eğitim verme süreci ister yükseköğrenimin, ister mesleki gelişimin, ister kamu politikası tartışmalarının, isterse de iş eğitiminin bir parçası olsun, bireyleri doğru ölçme uygulamalarını anlamaya ve testler ile test sonuçlarını seçici (eleştirel ve bilinçli) şekilde kullanmaya hazırlamalıdır. Ölçme-değerlendirme konusunda başkalarına eğitim veren ya da bilgilendiren kişiler sorumluluğunun bilincinde profesyonel kişiler olmalı, aşağıdaki ilkelere uygun davranmalıdır (NCME, 2023):

1. Eğitim verilen alanlarda yetkin/uzman olunmalı ve güncel gelişmeler takip edilmelidir. Ayrıca bunlar öğretim süreçlerine yansımalıdır.
2. Ölçme-değerlendirme hakkında eğitim verirken adil ve dengeli bir bakış açısı sağlanmalıdır.
3. Ölçme-değerlendirmenin prensipleri ve kavramları, hedef kitleye uygun olan teknik ayrıntılarla açıklanmalıdır.
4. Başkalarını herhangi bir özel ölçme-değerlendirme yöntemi, ürünü ya da hizmeti hakkında bilgilendirirken kanıtlanmış bilgiler ile kişisel görüşler açık bir biçimde ayrılmalıdır.
5. Eğitim-öğretimin konusu olan, belli bir ölçme-değerlendirme ürünü ya da hizmeti hakkında yapılacak değerlendirmeyi etkileyebileceği düşünülen herhangi bir mali çıkar söz konusu ise bu durum açıkça beyan edilmelidir.
6. İster değerlendirme isterse eğitim-öğretim amaçlı olsun, bir ölçme-değerlendirme uygulaması veya ölçme işleminden elde edilen sonuçların raporlanması, sınava giren kişiye/kişilere yönelik olumsuz etki taşıma ihtimali içeriyorsa bu duruma ilişkin riskler ilgili taraflara açıkça bildirilmelidir.
7. Eğitim-öğretim sürecinde kullanılan tescilli (yetkili kişilerce kullanılabilen ya da telif hakkı olan) ölçme-değerlendirme araçlarının ve materyallerinin güvenliği sağlanmalıdır.
8. Sorumlu ölçme ve değerlendirme uygulamaları modellenmelidir. Eğitim-öğretim, öğrencilerin eğitim ve kariyer hedefleriyle uyumlu hale getirilmelidir. Eğitim alanların, eğitimde ölçme ve değerlendirme konusundaki mesleki sorumluluklarını öğrenmelerine yardımcı olunmalıdır.
9. Politika yapıcılar, ebeveynler ve diğer vatandaşlar tarafından tartışılan ölçme ve değerlendirme konularına ilişkin, adil ve dengeli bakış açıları sağlanmalıdır.

ÖSYM'nin görev ve sorumlulukları arasında doğrudan ölçme-değerlendirme eğitimleri vermek yer almasa da ÖSYM kurumsal sorumluluğunun bir parçası olarak çeşitli kurum ve kuruluşlara ölçme-değerlendirme eğitimleri verebilmektedir (örneğin Karabük Üniversitesi Medya, 2025). Ayrıca düzenlediği hizmet içi eğitimlerle kurum çalışanlarının ölçme-değerlendirme alanındaki bilgilerinin güncel kalması sağlanmaktadır. Bunları yaparken yukarıda listelenen sorumluluklara göre eğitim verecek kişileri belirlemekte ve eğitim içeriğini şekillendirmektedir.

2.8. Programları Değerlendiren ve Değerlendirmelerle Araştırma Yürütenlerin Sorumlulukları

Değerlendirmelere dayalı araştırma yürütmek (program değerlendirme, eğitim müdahale araştırması ya da değerlendirmenin kendisini inceleme amaçlarıyla), değerlendirmelerin ve eğitim programlarının daha iyi anlaşılmasına yardımcı olan

önemli bir faaliyettir. Eğitim programlarının değerlendirilmesine katılan ya da değerlendirmelerle ilgili araştırma yürüten kişiler şu hususlara dikkat etmelidir (NCME, 2023):

1. Değerlendirme ve araştırma etkinlikleri; bilgiye dayalı, nesnel ve adil bir biçimde yürütülmelidir.
2. Gizlilik ve bilgilendirilmiş onam hakkı da dâhil olmak üzere, değerlendirme ya da araştırma çalışmasına katılanların haklarını korumak için geçerli standartlara uyulmalıdır.
3. Eğitim araştırması kapsamında insan katılımcılardan veri toplanacaksa araştırmayı yürütmeden önce Kurumsal Etik Kurul (*Institutional Review Board*) onayı ya da muafiyet belgesi alınmalıdır.
4. Yazarlar, test yayıncıları ya da değerlendirmeye dâhil olan diğer kişilerle ilişkileri varsa bu durumu açıklamalıdır. Bu tür ilişkilerin, araştırmanın ya da değerlendirmenin nesnelliğini etkilemesi ihtimal dâhilinde ise çıkar çatışması söz konusu olabilir. Çıkar çatışması oluşturabilecek bu tür durumlardan kaçınılmalıdır.
5. Değerlendirme veya müdahale araştırmalarında, programın/müdahalenin hedefleriyle uyumlu yapıları ölçen ölçme araçları seçilmeli veya geliştirilmelidir. Örneğin; bu ölçme araçları müdahalenin olası etkilerine karşı duyarlı olmalı, diğer bir ifadeyle yapılan müdahalenin muhtemel etkilerini doğru şekilde ortaya koyabilecek yeterlilikte olmalıdır.
6. Daha önce geliştirilmiş ya da yeni geliştirilen ölçme aracının kullanılması durumunda, çalışmadan elde edilen verilere dayanarak güvenirlik ve geçerlik bilgileri raporlanmalıdır.
7. Araştırma veya değerlendirme sonuçlarının, bir programın/müdahalenin etkinliğine dair geçerli göstergeleri sağlayabilmesi için tüm ölçme ve değerlendirme süreçleri güvence altına alınmalıdır.
8. Araştırma tasarımında ya da değerlendirme sonuçlarında, geçerliği düşürebilecek olası kaynakları en aza indirecek adımlar atılmalıdır. Çalışmanın sonuçlarında yanlılığa neden olabileceği bilinen faktörler (bir ölçme ve değerlendirme uygulamasının birden fazla kez yapılmasının öğrenme ya da uygulama üzerinde etkileri gibi) açıklanmalıdır.
9. Araştırmanın hem amaçlanan hem de amaçlanmayan sonuçları; adil, eksiksiz ve nesnel bir biçimde sunulmalıdır.
10. Başkalarına ait çalışma ve fikirler, eksiksiz ve uygun şekilde belirtilmelidir.
11. Araştırma sonuçları, çalışmanın sınırlılıkları dâhilinde nitelendirilmelidir.

12. Değerlendirme ve araştırma yaparken mümkünse konuyla ilgili farklı kaynaklardan toplanan bilgiler kullanılmalıdır.

Geniş kapsamlı araştırmalara hizmet edebilecek büyük bir veri havuzuna sahip olan ÖSYM yaptığı/ desteklediği araştırmalarda yukarıda açıklanan ilkelere bağlı hareket etmektedir. ÖSYM görev alanı nedeniyle belli sınırlılıklarla da olsa özellikle testlerin psikometrik özelliklerinin iyileştirilmesi üzerine çeşitli araştırmaları desteklemektedir (URL-1).

Ölçme araçlarının geliştirilmesinde, uygulanmasında, puanlanmasında değerlendirmesinde ve değerlendirme raporlarının hazırlanmasında takip edilen adımlar eğitimde ölçme ve değerlendirme süreçlerinin niteliğini doğrudan etkilemektedir. Bireyler hakkında doğru kararlar vermenin ön koşulu geçerli ve güvenilir sonuçlara ulaşmaktır. Ölçümlerin geçerliği ve güvenirliği kadar önemli bir nokta da ölçme-değerlendirme uygulamaları ile araştırma süreçlerinde etik ilkelere ne kadar uygun davranıldığıdır. Etik açıdan uygun ve nitelikli çıktılara ancak yukarıda belirtilen sorumluluklar çerçevesinde yürütülecek ölçme ve değerlendirme süreçleriyle ulaşılabilir.

3. ÖLÇME VE DEĞERLENDİRMEDE ADALETİN BİRLEŞENLERİ

Adalet kelimesi TDK'ya (t.y.-a.) göre; "*hak ve hukuka uygunluk, hakkı gözetme; yasalarla sahip olunan hakların herkes tarafından kullanılmasının sağlanması; herkese kendine uygun düşeni, kendi hakkı olanı verme*" anlamlarına gelmektedir. Eğitimde adalet kavramı, daha çok eğitim eşitliği (*equity*) olarak ifade edilmektedir. Eğitim eşitliği çocuklar ve yetişkinler için eğitime erişim ve fırsatların ne ölçüde adil ve hakkaniyetli olduğuyla ilgilidir. Bu anlamda eğitim eşitliği cinsiyet, sosyo-ekonomik durum, ikamet yeri (coğrafi konum), etnik köken, dil ve diğer özelliklere dayalı eşitsizliklerin azaltılmasına atıfta bulunan bir kavramdır (*United Nations Educational, Scientific and Cultural Organization*-UNESCO, 2008). Açıkça görüldüğü üzere eğitim eşitliğinin, birden çok bileşeni vardır. Ayrıca bu noktada adalet ve eşitlik (*equality*) kavramlarının aynı anlama gelmediğine dikkat edilmelidir. Eğitim eşitliği ile herkes kendine uygun düşeni, kendi hakkını alabilir, diğer bir ifadeyle adalet sağlanır. Oysa tek başına eşitlik kavramı imkân ve koşullarına, özel durumuna vb. faktörlere dikkat edilmeksizin herkese eşit muamelede bulunulmasını belirtir ki bu durum her zaman adaletli olmaz.

Eğitim bir sistem olarak düşünüldüğünde ölçme ve değerlendirme bu sistemin kalite-kontrol mekanizmasıdır (Turgut & Baykul, 2013). Bu başlık altında eğitimin kalite kontrol aşaması olarak tanımlanan ölçme ve değerlendirme süreçlerinde adalet ve adaletin bileşenleri olan yanlılık, kapsayıcılık, hesap verebilirlik ile şeffaflık konularında bilgiler verilmiştir.

3.1. Ölçmede Adalet

Eğitimde ölçme-değerlendirme süreçlerinde adalet, test adaleti (*test fairness*) kavramıyla açıklanabilir. Adalet/adil olma/adillik, bir ölçme aracının geliştirme/uyarlama, uygulama ve raporlaştırma olmak üzere tüm aşamalarında dikkate alınması gereken bir konudur (APA, 2004). APA (t.y.-b) adalet/adil olma (*fairness*) kavramını; farklı ırksal ya da kültürel özelliklere ve diğer çeşitli yaşam deneyimlerine sahip bireylerin test puanlarındaki sistematik varyansı (farklılığı) ortadan kaldırmak amacıyla testi alan kişilere hakkaniyetli ve adil muamele edilmesi olarak tanımlanmaktadır. Temelde teknik değil sosyokültürel bir konu olan adalet; testin tasarımında, uygulamasında ve puanlamasında kalite yönetimini ifade eder. Ölçme açısından adil bir test geliştirebilmek için içyapıya dayalı kanıtlar sağlamaya dönük yeterli çalışma yapılmalı, öğrenme fırsatlarında ve teste erişimde eşitlik sağlanmalı, hazırlanan test hedef içeriği yeterince kapsamalı ve yalnızca incelenen beceriye/yeteneğe odaklanmalıdır. Ayrıca ölçme sonuçlarında test uygulayıcısının geçmiş yaşantılarından kaynaklanan bir varyans söz konusu olmamalı ve test maddeleri ölçülmesi hedeflenen yapının dışında kalan bir özelliği ölçmemelidir (AERA, APA & NCME, 2014).

Testle yapılan ölçmelerin yanlı olmaması ile ilişkili olan ölçmede adalet kavramı, testin tüm gruplar için adil ve eşit fırsat sunmasını da içeren daha geniş bir kavramı ifade eder. Test adaleti; bireyler, gruplar ve ortamlar arasında karşılaştırılabilir içyapıya dayalı kanıtlarla elde edilir (Messick, 1998). AERA, APA ve NCME (2014) test adaletini tanımlarken; test sürecinde muamelede adillik, ölçme yanlılığının olmaması, ölçülen yapı(lar)a erişimde adillik ve bireysel test puanı yorumlarının amaçlanan kullanımlar için geçerli olması olmak üzere dört standart belirlemiştir.

Adalet/adil olma/adillik test sürecinde tüm adaylara eşit muamelede bulunulması anlamına gelir AERA, APA ve NCME (2014). Bu bakımdan ölçmede adalet, tüm adaylara yeterliliklerini gösterebilmeleri için gerekli olanaklar sunmayı ve ölçülen yapıdaki konumlarının doğru belirlenmesi için adayların karşılaştırılabilir fırsatlara sahip olmasını gerektirir. Adil muamele karşılaştırılabilir test koşulları (tüm test katılımcılarının testi eşit ve standart bir ortamda alabilmesi anlamına gelir), teste aşinalık kazandırma ve uygulama materyallerine eşit erişim, her sınav katılımcısının sınava hazırlanma fırsatına sahip olması gibi faktörleri içerir (APA, 2004; Messick, 1998). Adaletin sağlandığı bir ortamda; testler ve uygulama koşulları dikkatli bir şekilde standartlaştırılır, tüm sınav katılımcılarına ilgili alandaki performanslarını ve yetenek düzeylerini gösterebilecekleri fırsatlar sunulur (APA, 2004).

Adalet tüm ölçme-değerlendirme süreçleri için önemli olmakla birlikte yüksek riskli sınavlarda çok daha kritik hale gelmektedir. ÖSYM de yüksek riskli sınavlar yürüten bir kurum olduğundan ölçmede adalet kavramı ÖSYM'nin misyonunda kendine yer bulmuştur. ÖSYM misyonunu "*Uyguladığı sınavlarda geçerli, güvenilir ve adil ölçme, seçme ve yerleştirme yapan bir kurum olmaktır*" şeklinde belirtmektedir (URL-2).

3.2. Ölçmede Yansızlık

Eğitimde ve psikolojide temel amaç bireylere ait özelliklerin geçerli ve güvenilir ölçümlerine ulaşmaktır. Bunun için mümkün olduğunca hatalardan arınık ölçme yapılması gerekir. Bu noktada psikolojik ölçmeleri etkileyen problemleri araştıran ve bu problemleri azaltmaya çalışan bir alan olarak tanımlanan test kuramları (Crocker & Algina, 1986) önem taşır. Çeşitli test kuramlarıyla, çeşitli istatistiksel süreçler yürütülerek ölçme araçlarının farklı örneklemlere ya da farklı kültürlere uygulanmasında geçerli ve güvenilir sonuçlara ulaşılması hedeflenir.

Farklı uygulamalarda geçerli ve güvenilir sonuçlara ulaşmanın yolu ölçme araçlarının ölçmeyi hedeflediği özelliği, grubun çeşitli özelliklerinden bağımsız ölçmesinden geçer. Ölçme aracı, uygulandığı grup içinde yer alan alt grupların/örneklemlerin özelliklerine bağlı olarak bir gruba avantaj ya da dezavantaj sağlamamalıdır (Crocker & Algina, 1986). Diğer bir ifadeyle yanlı ölçme yapmamalıdır. APA (t.y.-c) yanlılığı (*bias*);

- Taraflılık (*partiality*) bir şeye karşı veya onun lehine eğilimli olma ya da yatkın olma,
- Tepki yanlılığı (*response bias*) ya da test yanlılığı (*test bias*) gibi herhangi bir eğilim veya tercih,
- Örnekleme, veri toplama ya da veri analizi sırasında ortaya çıkan sistematik hata (yanlı kestirimler, yanlı örnekleme),
- Ölçülen ya da hesaplanan bir niceliğin gerçek (doğru) değeri arasında bir sapma olması (örneğin ölçüm ya da hesaplama, ilgili unsurun temsili niteliğinde değildir) olacak şekilde tanımlamaktadır.

Ölçme aracındaki maddeler ölçülen özellik bakımından aynı yetenek düzeyinde bulunan fakat farklı alt grupların/örneklemlerin üyesi olan bir gruba avantaj ya da dezavantaj sağlamamalıdır (Zumbo, 1999). Örneğin ulusal çapta uygulanan bir genel yetenek sınavında metro istasyonlarının krokileri üzerinden uzamsal akıl yürütmeye yönelik bir madde yer alsın. Uygulama sonucunda aynı yetenek düzeyindeki öğrencilerden yaşadıkları şehirde metro olanların daha yüksek başarı gösterdiği belirlenmiş olsun. Bu örnekte, metrosu olan bir şehirde yaşayan öğrencilerin aynı yetenek düzeyinde yer alan fakat yaşadıkları şehirde metro bulunmayan öğrencilere kıyasla avantajlı olduğu görülmektedir. Burada ölçülmek istenen yapının dışında başka bir faktörün (şehirde metro olması), ölçme sonuçlarına karışarak yanlılık oluşturduğu düşünülebilir. Bir ölçme yapıldığında ölçülmek istenen yapının dışında kalan özellikler (cinsiyet, ırk, sosyoekonomik düzey gibi) aynı yetenek düzeyindeki bireylerden bir gruba avantaj ya da dezavantaj sağlamamalıdır. Aksi takdirde sistematik hata devreye girer ve yanlılık olarak nitelendirilen bu durum ölçme sonuçlarının geçerliğini zedeler; daha açık bir anlatımla ölçme sonuçlarına dayalı çıkarımların, verilecek kararların ve ulaşılacak sonuçların hatalı olmasına yol açar. Bu anlamda ölçme sonuçlarından hareketle yapılan çıkarımların doğruluğu, ölçümlerin yansız olmasına bağlıdır.

Kültürlerarası karşılaştırmaların yapıldığı test uyarlama çalışmaları ya da farklı kültürlerden çok sayıda bireyin aynı testleri aldığı ve ülkelerin akademik başarılarının göstergesi olarak değerlendirilen testlerde (örneğin PISA, TIMSS, PIRLS gibi geniş ölçekli başarı testleri), kültürler arası çalışmalar yapılması hedeflendiğinden yanlılık daha da dikkat çekici bir konu haline gelir. Yapılan ölçmenin kültürden, ölçme aracının dilinden bağımsız olması geçerli sonuçlara ulaşmada kritik bir rol üstlenir. Farklı kültürler için ölçülen yapının aynı kalması ölçme değişmezliği (*measurement invariance*) olarak tanımlanır (Şengül Avşar & Emons, 2021).

Ölçme değişmezliğini ve yanlılığı belirlemenin çeşitli istatistiksel yolları vardır. Bu yollar içinde Değişen Madde Fonksiyonu (*Differential Item Functioning*-DMF) ve Çoklu-Grup Doğrulayıcı Faktör Analizi (*Multi-Group Confirmatory Factor Analysis*-ÇGDFA) sıklıkla dikkate alınan yaklaşımlardır (Stark ve ark., 2006; Vandenberg & Lance, 2000; Zumbo ve ark., 2003). DMF bir maddenin kültürler arasında ya da farklı bir grubun üyesi olmanın bir avantaj sağlayıp sağlamadığını belirler (Ercikan ve ark., 2004). Çeşitli istatistiksel yöntemlerle maddelerin DMF içerip içermediği belirlenir. Bunun için testi alan gruplar, referans ve odak grupları olarak iki ayrılırlar ve uygun istatistiksel yöntemlerle maddelerin DMF içerip içermediği incelenir (Arıkan ve ark., 2016).

Bir maddenin DMF içermesi o maddenin yanlı ölçme yaptığının tek başına kanıtı olmaz. Yanlılık çalışmalarında ilk aşama maddelerin DMF içerip içermediğinin istatistiksel olarak test edilmesidir. İkinci aşamada DMF içerdiği belirlenen maddeler uzman görüşlerine sunulur ve bu maddelerin gerçekten yanlı olup olmadığı uzmanlar tarafından yapılan detaylı incelemeler sonrasında netleşir (Dorans & Holland, 1992). Bu noktada uzmanların niteliği de önem taşır. İnceleme yapacak uzmanların test geliştirme süreçlerine hâkim ve testin ölçtüğü konu ya da yapı hakkında bilgi sahibi olmaları gerekir (Ayan, 2024). Bir maddenin yanlı olarak etiketlenmesi (*flagged*) için madde içeriğinde testle ölçülmesi hedeflenmeyen bir özelliğin ölçülmesi gerekmektedir.

Bir maddede DMF iki farklı şekilde belirebilir. İlki DMF'nin test maddesiyle ölçülmek istenen temel yetenek bakımından gruplar arasında gerçek farklılıklar olması nedeniyle var olduğu madde etkisi (*item impact*) durumudur. Burada yanlılık söz konusu olmayıp DMF'ye kaynaklık eden ölçülen özellik bakımdan var olan gerçek farklılıktan bahsedilmektedir. Diğeri ise madde yanlılığı durumudur. Bu durum, DMF'nin test maddesinin ölçülmek istenen temel yetenek bakımından (ve dolayısıyla testin amacıyla) ilgili olmayan bir özelliği nedeniyle var olduğu durumlar olarak ifade edilmektedir (Zumbo, 2007).

Ölçme değişmezliğini belirlemenin bir diğer yolu olan ÇGDFA, ölçülen yapının farklı örneklemler arasında benzerliğini test etmeye dönük istatistiksel bir yöntemdir (Brown ve ark., 2017). Bu yöntemde iç içe geçmiş dört hiyerarşik model test edilir. Bu modeller

sırasıyla; biçimsel değişmezlik (*configural invariance*), metrik değişmezlik (*metric invariance*), skaler değişmezlik (*scalar invariance*) ve katı değişmezlik (*strict invariance*) olarak isimlendirilir (Vandenberg & Lance, 2000.) Kurulan bu modellerden elde edilen uyum iyiliği istatistikleri arasındaki farklar incelenerek ölçülen yapının gruplar arasında değişip değişmediği test edilir. Ancak katı değişmezliği sağlamanın zor olduğu, eğitim ve psikoloji araştırmalarında katı değişmezliği araştırmaya gerek olmadığı da belirtilmektedir (Bowen & Masa, 2015). Katı değişmezliğin sağlanmadığı durumda kısmi değişmezliğin (*partial invariance*) incelenmesi önerilmektedir (Bryne, 2016; Tse ve ark., 2024).

Özetle bir testi geliştirirken, uyarlarken, uygularken ve puanlarken hata kaynaklarını ve ölçmede hata türlerini dikkatli bir şekilde ele almak gerekir. Ölçme sonuçlarının yanlılık içermesi, geçerliği düşürerek yapılacak çıkarımları olumsuz etkiler. Yanlı maddeler, ölçme yapılan gruplardan birine avantaj sağlayarak test adaletini de zedeler. Bu nedenle geçerlik çalışmaları kapsamında madde bazında DMF incelemelerine ve test bazında ÇGDFA çalışmalarına mutlaka yer verilmelidir. Böylece sonuçların yanlı olmadığına yönelik kanıtlar sunulup ölçümlerin geçerliği desteklenir.

3.3. Ölçmede Kapsayıcılık

UNESCO, *Education* 2030 girişimiyle herkes için kapsayıcı, eşitlikçi ve kaliteli eğitim (*Education 2030: Inclusive, equitable, and quality education for all*-EFA) programını yürütmektedir. EFA programı, sürdürülebilir kalkınma hedefleri çerçevesinde eğitim alanında dünyayı yönlendirmektedir (URL-3). UNESCO (2015) Eğitim 2030 Incheon Deklarasyonu ve Eylem Çerçevesi programında; kapsayıcı ve eşitlikçi kaliteli eğitimin sağlanması ve herkes için yaşam boyu öğrenme fırsatlarının teşvik edilmesi için Sürdürülebilir Kalkınma Hedefi 4'ün hedeflerini düzenlenmiştir. Tüm bireylerin kapsayıcı, eşit ve kaliteli eğitim almasını ve yaşam boyu öğrenme fırsatlarına erişmesini sağlamayı amaçlamayan bu hedefler; çocukların temel eğitimi tamamlamalarını, gençlerin ve yetişkinlerin teknik, mesleki ve yükseköğretime eşit erişimlerini ve istihdam için gerekli becerileri kazanmalarını, özellikle özel gereksinimli, yerli (*indigenous*) ve kırılgan grupların eğitim haklarının güvence altına alınmasını, güvenli ve kapsayıcı öğrenme ortamlarının oluşturulmasını amaçlamaktadır. Kapsayıcı eğitimin sağlanması, eğitimin tüm süreçlerinin kapsayıcılığa uygun olarak tasarlanmasıyla mümkün olur. Bunun için;

- Politika ve planlamaların eğitimde kapsayıcılığı ve eşitliği sağlayacak şekilde düzenlenmesi,
- Öğretmen eğitiminde öğretmen adaylarının kapsayıcı eğitimin bir parçası olarak nitelikli biçimde yetiştirilmesi ve görevde olan öğretmenler için bunun hizmet içi eğitimlerle desteklenmesi,

- Eğitim programlarının ve eğitim materyallerin tüm öğrenciler için uygun, erişilebilir ve farklılıklara duyarlı bir şekilde hazırlanması,
- Ölçme ve değerlendirme uygulamalarının tüm bu aşamalara uygun olarak tasarlanması gerekir.

Ölçme ve değerlendirme süreçlerinin kapsayıcı olması için test geliştirme/uyarlama, uygulama ve raporlama süreçleri kapsayıcılığa uygun şekilde tasarlanmalıdır. Bu kapsamda testlerin farklı formatlarda uygulanması, sonuçların adil şekilde hesaplanması, kapsayıcı eğitimin hedeflerine ulaşılıp ulaşılmadığının belirlenmesine yönelik verilerin toplanması ve analizi ile eğitim süreçlerinin izlenmesi gibi işlemler yürütülebilir.

Kapsayıcı değerlendirmeler, başlangıçta özel gereksinimli öğrenciler gibi dezavantajlı öğrenci gruplarının gereksinimlerini karşılamaya odaklanmıştır. Bununla beraber kapsayıcı değerlendirme tasarımı giderek genişlemiş, daha iyi bir eğitim uygulaması olarak kabul edilmiş ve tüm öğrencilere fayda sağlamayı hedeflemiştir (Advance HE, 2019). Kapsayıcı değerlendirmeler öğrencilerin değerlendirilme biçimleri aracılığıyla dışlanma, göz ardı edilme ve/veya dezavantajlı duruma düşme olasılığını en aza indirmeye çalışmaktadır (Oxford University Centre for Teaching and Learning, 2025). Thurlow ve arkadaşları (2016) yayınladığı raporda özel gereksinimli öğrencilerin yanı sıra İngilizce öğrenenler ve özel gereksinimli İngilizce öğrenenler için tüm değerlendirme türlerinde dikkate alınabilecek bir takım kapsayıcılık ilkelerine yer vermiştir. *University of Bristol, Centre for Teaching and Learning* (2025) dört temel kapsayıcı eğitim ilkesini şöyle açıklamaktadır:

1. **Çeşitli değerlendirme yöntemleri, seçenekleri ve konuları sunmak:** Bir program genelinde yapılacak kapsayıcı değerlendirme, öğrencilerin güçlü yönlerini ortaya koymalarına olanak tanımalı ve dezavantajlarını en aza indirecek çeşitli yöntemler, seçenekler ve konular sunmalıdır. Diğer bir ifadeyle tek tipte bir ölçme süreci takip edilmemelidir. Bir program boyunca tüm öğrencilerin öğrendiklerini gösterebilmelerine imkân tanıyacak çeşitli değerlendirme yöntemleri tasarlanmalıdır. Örneğin, öğrencilerin etkili iletişim becerilerini göstermeleri gerekiyorsa öğrencilere çeşitli seçenekler (poster, kayıtlı sunum, blog gibi) sunulmalı, iletişim kurmada kendilerini en güvende ve rahat hissettikleri ortamı seçmelerine izin verilmelidir. Kapsayıcı değerlendirme yaklaşımlarında öğrencilerin geçmişleri de değerli bir kaynak olarak görülür. Öğrencilere kendi deneyimlerinden yararlanarak anlayışlarını/kavrayışlarını ortaya koyma fırsatları verilir. Konu seçme olanağı sağlamak (örneğin; proje, kompozisyon ya da sınav sorusu için) öğrencilere odak noktalarını yönetebilme fırsatı sunar. Ayrıca öğrencilerin ilgi duydukları alanlarda çalışmalarına imkân verir.

2. **Açık, adil ve şeffaf değerlendirme ölçütlerini sağlamak:** Kapsayıcı değerlendirme yaklaşımlarında; açık, adil ve şeffaf değerlendirme ölçütlerinin kullanılması gerekir. Bu ölçütler sade bir dille yazılmalı, öğrencilerin öğretim elemanları ve akranlarıyla tartışmalarına olanak tanımalıdır. Örnek olarak öğrencilerin (a) ölçütleri birbirleriyle tartışması ve bunları değerlendirme yönergesiyle ilişkilendirmesi, (b) kendi çalışmalarını ölçütlere göre gözden geçirmesi, (c) örnek çalışmaları ölçütler çerçevesinde inceleyip tartışması, (d) birbirlerinin biçimlendirici (*formative*) değerlendirmelerini ölçütlere göre gözden geçirip geri bildirim vermesi ve (e) aldıkları geri bildirimleri ölçütlerle eşleştirmesi gösterilebilir. Bu çalışmalar hem yaparak yaşayarak öğrenme süreci hem de ölçütlerin nasıl kullanılacağının gösterilmesi açısından önemlidir.
3. **Öğrencileri değerlendirmeye hazırlamak:** Öğrencileri değerlendirmeye hazırlamak, kapsayıcı tasarımın ayrılmaz bir parçasıdır. Bu hazırlık; uygulama görevlerinin hem iyi hem de yetersiz örneklerin kullanımını ve yukarıda açıklandığı üzere bu örneklerin değerlendirme ölçütleriyle karşılaştırıldığı çalışmaları içerebilir. Ayrıca öğrencilerin güçlü yönlerini vurgulayan ve gelişim yollarını destekleyen düzenli biçimlendirici geri bildirim fırsatlarını da içine almalıdır. Kapsayıcı geri bildirim; metin, ses veya ekran kaydı gibi çeşitli formatlarda sunulabilir. Temel nokta geri bildirimlerin zamanında, yapıcı ve açık bir şekilde iletilmesi, öğrencilerin özgüvenini geliştirmeye ve başarıya ulaşmalarını desteklemeye hizmet etmesidir.
4. **Öğrencinin iş yükünü göz önünde bulundurma ve değerlendirme takviminde oluşabilecek çakışmalardan kaçınma**: Değerlendirme tasarımları, öğrencilerin sıkılmadan derinlemesine öğrenmelerini sağlamalıdır. Bunun için tüm program genelindeki değerlendirme yoğunluğu hesaba katılmalıdır. Program boyunca yapılan değerlendirmeler, öğrencilerin her bir değerlendirmeye hazırlanabilmeleri için yeterli zamanı sunmalıdır. Ayrıca değerlendirme görevlerinin teslim tarihleri arasında çakışmaların önlenmesi önemlidir.

Kapsayıcı değerlendirme tasarımı için Tablo 1'de verilen kontrol listesi kullanılabilir.

Tablo 1

Kontrol Listesi (University of Guelph, Office of Teaching and Learning'ten (2025) uyarlanmıştır)

Öğrenme Çıktıları ve Öğretimle Uyum:	
☐	Değerlendirmelerin içeriği ve biçimi, derste öğretilen temel konuları ve becerileri doğru şekilde yansıtmıştır.
☐	Değerlendirmelerin içeriği ve biçimi, doğrudan dersin öğrenme çıktılarıyla ilişkilendirilmiştir.
Genel Kurs/Ders Değerlendirme Tasarımı:	
☐	Ders içinde farklı değerlendirme türleri kullanılmıştır.
☐	Yüksek riskli değerlendirmeler, birden fazla bileşene ayrılmış veya yüksek riskli sınavların denemesi niteliğinde olan düşük riskli değerlendirmeler sunularak aşamalı hâle getirilmiştir.
☐	Yüksek riskli değerlendirmelerde gereken içerik ve becerilerle ilgili açık uygulama imkânı sağlanmıştır (Bunun için notlandırmaya dayanmayan veya düşük riskli değerlendirmeler kullanılmalı, gelişimi destekleyen geri bildirimler zamanında verilmelidir).
☐	Değerlendirmelerle ilişkili iş yükü dikkate alınmıştır.
Değerlendirme Süreçleri:	
☐	Öğrencilere, öğrenmelerini nasıl gösterecekleri konusunda seçenekler sunulmuştur (Örneğin herkesin son bir makale teslim etmesi yerine; öğrencilere sunum, podcast, infografik ya da poster teslim edebilme alternatifleri sunulmalıdır).
☐	Değerlendirme beklentileri değerlendirme öncesi açık bir şekilde paylaşılmıştır.
☐	Dereceli puanlama anahtarları (rubrikler) ve örnek ödevler kullanarak mükemmel çalışmanın nasıl göründüğü açıklanmıştır.
☐	Not vermek için net ölçütler belirlenmiştir.
☐	Mümkün olduğunda öğrenci ödevleri anonim (gizli) olarak değerlendirilmiştir.
Ölçme ve Değerlendirme İçeriği:	
☐	Ölçme aracında yer alan maddeler/sorular; farklı kültürlerden ya da geçmişlerden gelen öğrenciler ile ailesinde daha önce üniversite mezunu olmayan öğrenciler için belirsiz ya da bilinmeyen kaynaklardan uzak olarak hazırlanmıştır.
☐	Ölçme ve değerlendirme içerikleri, senaryolar ve sorular aracılığıyla farklı bakış açılarının temsil edilmesini sağlamıştır.

AERA ve arkadaşları (2014) yayınladığı Eğitim ve Psikolojik Testler için Standartlarda (*Standards for Educational and Psychological Testing*) eğitimde ölçmede kapsayıcılık için çeşitli standartlar tanımlamıştır.

Standart 3.9. "Test geliştiricileri/kullanıcıları, uygun ve mümkün olduğunda, adayların hedeflenen yapılar üzerindeki yeterliklerini ortaya koymalarını engelleyebilecek, ölçülen yapıyla ilişkisiz faktörleri ortadan kaldırmaya yönelik sınav uyarlamaları geliştirmekten ve sağlamaktan sorumludur" şeklinde ifade edilmektedir. Bu standart ile öğrencilerin ölçülen yapıyla ilgili olmayan engeller nedeniyle dezavantaj yaşamalarının önüne geçilmesini sağlayacak sınav uyarlamalarının (*test accommodations*) yapılması beklenmektedir. Bunun için sınav ortamı, sunum biçimi, etkileşim/katılım, yanıt verme şekli gibi alanlarda değişiklikler yapılabilir ve gerektiğinde okuyucu veya yazıcı desteği

sağlanabilir. Uygun bir uyarlama bireysel özelliklere/ihtiyaçlara yanıt verir ama bunu yaparken testin ölçtüğü yapıyı ve puanların anlamını değiştirmez. Kapsayıcı bir uygulama için bu gereklidir. Test geliştiricileri, uyarlamanın ölçülen yapıyı değiştirmediğine yönelik kanıtlarını sunmalıdır. Uyarlamalar; bilişsel, dilsel, duyusal ya da fiziksel ihtiyaçlara yönelik olabilir ve çoğu zaman yasal zorunluluk taşır. Örneğin; dil yeterliliği olmayan öğrenciler için dilsel uyarlamalar, görme yetersizliği olan öğrenciler için metin büyütme, işaretleyici ve/veya okuyucu desteği sağlanması gibi düzenlemeler yapılabilir.

Standart 3.10. "Sınav uyarlamalarına izin verildiğinde test geliştiricileri/kullanıcıları, uyarlamaların nasıl kullanılacağına ilişkin standart düzenlemelerini belgelemekten ve uyarlamanın uygun şekilde gerçekleştirilip gerçekleştirilmediğin, izlemekten sorumludur" şeklinde ifade edilmektedir. Bu standart, uyarlamaların bir prosedüre bağlanmasını içermektedir. Test uyarlamaları yalnızca özel durumunu ve ihtiyacını belgeleyen bireylere sağlanmalı, sınav öncesinde bu ihtiyaç uzmanlarca onaylanmalı ve sürece dair bilgiler bireylere şeffaf biçimde sunulmalıdır. Uyarlamaların türü, uygulanma biçimi ve kuralları standart protokollerle belirlenmeli; test yöneticileri bu süreçte gerekli bilgi ve eğitime sahip olmalıdır. Uygulamalar standart şekilde yürütülmeli, tüm sapmalar kaydedilmeli ve özellikle büyük ölçekli sınavlarda uyarlamaların doğru ve adil biçimde uygulanıp uygulanmadığı takip edilmelidir. Böylece sınav süreci, bireysel farklılıkları gözetirken eşitlik ve kapsayıcılığı güvence altına alır. Örneğin ÖSYM görme yetersizliği olan adaylara engel durumlarını belgelemeleri ve talep etmeleri halinde okuyucu ve işaretleyici desteği sağlamakta ve okuyucu yardımı talep eden adaylarını kitapçığa bakmalarını ve soruları kendilerinin okumalarını yasaklamaktadır. Görme engelli adayların soru kitapçığını kendilerinin okuduğunun tespit edilmesi durumunda sınavlarının geçersiz sayılacağı belirtilmektedir.

Standart 3.11. "Bir testte, ölçülmek istenen yapıya erişimdeki engelleri kaldırmak için değişiklikler yapıldıysa test geliştiricileri/kullanıcıları yapılan değişiklik sonrası testin geçerliğini kanıtlayan verileri toplamak ve belgelemekle sorumludur. Geçerlik çalışmaları mümkünse yeterli büyülükte bir örneklem üzerinde yapılmalıdır" şeklinde ifade edilmektedir. Standart testlerde sınav süresinin uzatılması, dilin sadeleştirilmesi gibi değişiklikler yapıldığında pilot uygulamalar ya da saha çalışmalarıyla, gerçekleştirilen değişikliklerin geçerlik üzerindeki etkisi hedeflenen alt grupları (örneğin özel gereksinimli, dilsel/kültürel farklılıkları olan adaylar) temsil eden bireyler üzerinde test edilmelidir. Geçerlik çalışmaları hem bu değişikliğin ilgili grup için etkinliğine hem de değişiklik yapılan test ile orijinal test arasındaki puanlarının karşılaştırılabilirliğine odaklanmalıdır. Örneklem büyüklüğünün yeterli olmadığı durumlarda geçerlik kanıtlarına ulaşmak amacıyla daha küçük gruplarla nitel çalışmalar yürütülebilir ya da uzman görüşlerine başvurulabilir. Özetle kapsayıcılık adına yapılan uyarlamaların, testin psikometrik özellikleri üzerindeki etkisi araştırılmalı ve ampirik kanıtlarla destek sağlanmalıdır.

Standart 3.13. "Bir test, testin amacına en uygun olan dilde uygulanmalıdır" şeklinde ifade edilmektedir. Buna göre standart testlerde çift dilli (*bilingual*) ya da çok dilli öğrenciler göz önünde bulundurulmalı daha açık bir anlatımla test kullanıcıları adayların dilsel ve kültürel özellikleri ile dil yeterliliklerini dikkate almalıdır. Testin uygulanacağı en uygun dili belirlerken testin amacı ve bağlamı titizlikle değerlendirilmelidir. Testin amacı belirli bir dilde yeterliliği ölçmek değilse adaylar en iyi bildikleri dilde sınava alınmalıdır. Ancak bazen adayların en iyi bildiği dil, test edilen kazanımlarla ilgili eğitim aldıkları dil olmayabilir. Bu tür durumlarda test, adayın en iyi bildiği dilde değil de öğretim diliyle uygulanabilir. Adayın en iyi bildiği dilin seçilmesi, test puanlarının geçerliğinin sağlandığını garanti etmez; dil ile öğrenilen yapı arasındaki uyumsuzluklar puan yorumlarını etkileyebilir.

Kapsayıcılıkla ilgili bir diğer önemli konu da teknoloji destekli ölçme ve değerlendirmelerle ilgilidir. Uluslararası Test Komisyonu ve Test Yayıncıları Derneği (International Test Commission and Association of Test Publishers [ITC-TAP]) (2025) yayınladığı Teknoloji Tabanlı Değerlendirme Rehberi kılavuz kitabında teknoloji destekli ölçme ve değerlendirme süreçleri hakkında detaylı bilgiler vermektedir. Bu bilgilerden biri Evrensel Test Tasarımı (Universal Test Design-[ETT]) ile ilgilidir. Evrensel Tasarım, duyusal ve fiziksel engelleri olan bireyler de dâhil olmak üzere tüm bireylerin binalara ve fiziksel araçlara erişimini ve bunları kullanabilmesini sağlamak adına mimarlık alanında ortaya çıkmıştır (Story ve ark., 1998, akt.: ITC-TAP, 2025). Testlerin dijital uygulamalarıyla birlikte testlere ve testlerin uygulandığı arayüzlere erişilebilirlik üzerinde de çalışmalar başlatılmıştır. Erişilebilirlik sadece özel gereksinimli bireyler için test uyarlamaları olarak algılanmamalı, tüm test kullanıcılarını etkileyen ve ETT ile desteklenebilen bir konu olarak dikkate alınmalıdır (ITC-TAP, 2025). Erişilebilirlik; görsel, işitsel, fiziksel, konuşma, bilişsel, dil, öğrenme ve nörolojik problemler de dâhil olmak üzere oldukça çeşitli bir alanı kapsar (*Web Content Accessibility Guidelines,* 2025). Bu anlamda ETT özel gereksinimi olanlar, çok dilli öğrenenler, farklı yaş, cinsiyet ya da kültürel gruplardan gelenler gibi tüm bireylere fayda sağlar. ETT ölçülmek istenen yapıyla ilgisi olmayan engelleri, diğer bir ifadeyle ölçme hedefi dışındaki faktörlerin yol açtığı engelleri ortadan kaldırmayı hedefler. Elbette teknoloji desteği ile tasarlanan ölçme araçlarında da geleneksel yöntemlerde olduğu gibi en üst düzeyde dikkatli olmak gerekir. Teknoloji desteğiyle tasarlanmış bir testte yer alan maddeler özel gereksinimli bireylere bilişsel yük getirebilir (ITC-TAP, 2025). Bu noktada kullanıcı merkezli bir tasarım oluşturulmalı, diğer bir ifadeyle tasarım bireylere uygun hale getirilmelidir. Dijital tasarımlarda uygunluk, bireylerle fiziksel bir obje arasındaki karşılıklı ilişki olarak ifade edilir (Norman, 2013). Bilişsel yükü azaltan, testi alanlara kolaylık sağlayan, bireylere uygun olan ve bunları gerçekleştirirken psikometrik özelliklerden taviz vermeyen bir test, ETT'ye uygun kapsayıcı bir test olarak ifade edilebilir.

Teknoloji tabanlı değerlendirmelerde kapsayıcılık önemli bir avantaj olarak öne çıksa da bu değerlendirmelerde kullanılan ses ve video içerikleri, farklı özel gereksinimli bireylere uygun olmayabilir. İşitme kaybı olan bireyler için ses içerikli maddelerin mutlaka yazılı versiyonlarının olması ya da görme kaybı yaşayanlar için sesli açıklamalara yer verilmesi gerekir. Yapılabilecek bu tür düzenlemeleri planlarken özel gereksinimli bireyler için *Web*'i daha erişilebilir hale getirmeye yardımcı olacak stratejiler, standartlar ve destekleyici kaynaklar sunan *the World Wide Consortium-W3C*'nin *Web* İçeriği Erişilebilirlik Yönergeleri dikkate alınabilir (ITC-TAP, 2025).

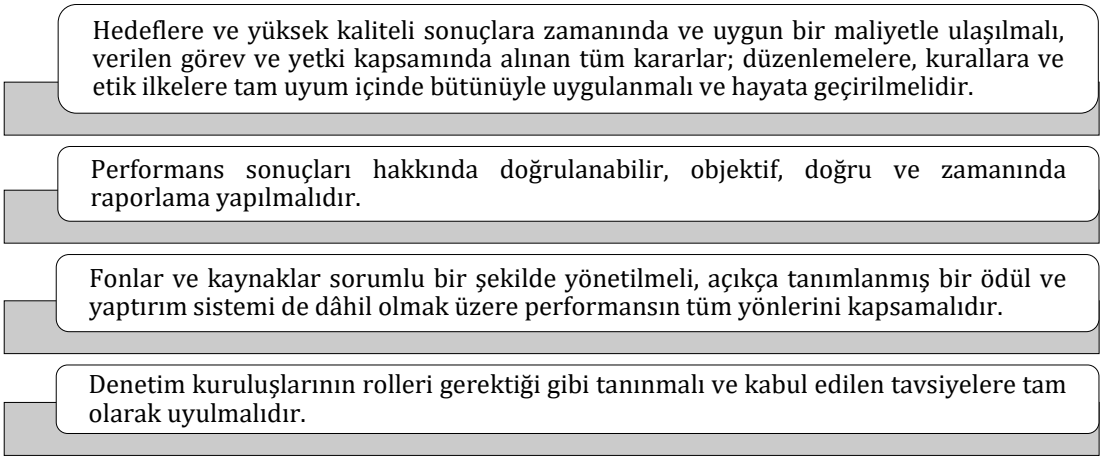
Ölçme ve değerlendirmede kapsayıcılık için yapılabilecekler özetlenecek olursa; eşit öğrenme ve eşit test fırsatları sağlanmalı, dilsel ve kültürel farklılıklara dikkat edilmeli, yanlılık içeren maddeler belirlenmeli ve bu maddelerden uzak durulmalı, özel gereksinimli bireyler için erişilebilir testler ve test düzenlemeleri yapılmalı, bilgisayar ve teknoloji tabanlı testlerin formatları ve kullanıcı arayüzleri evrensel tasarım ilkelerine uygun olarak düzenlenmelidir. Kapsayıcılığın sadece özel gereksinimli bireylerle ilgili bir konu olmadığı, ölçme ve değerlendirme süreçlerinde ölçme araçlarında çeşitlilik sağlanarak kişilerin kendilerini, yeteneklerini en iyi ifade edebilecekleri süreçler oluşturmak gibi çok daha geniş bir anlam taşıdığının farkında olunmalıdır.

3.4. Ölçmede Hesap Verebilirlik ve Şeffaflık

Hesap verebilirlik (*accountability*) bir kuruluşun/organizasyonun ve personelinin, alınan tüm kararlardan ve gerçekleştirilen tüm eylemlerden sorumlu olma ve taahhütlerini hiçbir kayıt ya da istisna olmaksızın yerine getirme yükümlülüğüdür (United Nations Development Programme [UNDP], 2025). APA (t.y.-d) hesap verebilirliği; bireylerin davranışları, kararları ya da yargıları hakkında başkalarına (örneğin yönetici, denetmen, danışman, resmi denetim organı, akran grubu) yanıt verebilme/hesap verebilme derecesi olarak tanımlamaktadır. UNDP (2025) hesap verebilirliğin içermesi gereken bileşenleri Şekil 1'deki gibi açıklamaktadır.

Şekil 1

UNDP'ye (2025) Göre Hesap Verebilirliğin Bileşenleri



UNESCO (2017) eğitimde hesap verebilirliği; öğretmenlerin, okulların, üniversitelerin, hükümetlerin ve diğer aktörlerin eğitimin sonuçları konusunda sorumluluk taşıması ve bunu şeffaflıkla raporlaması olarak ifade etmekte; ayrıca hesap verebilirliğin açık biçimde tanımlanmış sorumluluklar, sorumlulukların nasıl yerine getirildiğine dair hesap verme yükümlülüğü ve hesap verme yükümlülüğünün yasal, politik, sosyal ya da ahlaki gerekçesi olmak üzere üç temel unsuru olduğunu belirtmektedir. Eğitimde; düzenlemelere (bazen aynı anda uygulanan) uyum, mesleki normlara bağlılık ve sonuç odaklılık olmak üzere üç tür hesap verebilirlik türü olduğunu bildirmektedir (Anderson, 2005). Tüm bu tanımlar ve açıklamalar birlikte değerlendirildiğinde hesap verebilirliğin, sorumluluğa ve sonuçların şeffaf bir şekilde açıklanmasına odaklanan bir kavram olduğu ifade edilebilir.

Eğitimde testler ve test sonuçları hesap verebilirliğin bir ölçüsü olarak kullanılabilir. AERA ve arkadaşları (2014) eğitimde hesap verebilirlik amacıyla kullanılan testlere ilişkin belgelerin tüm potansiyel kullanıcıların/ paydaşların (okul personeli, veliler, öğrenciler) erişimine açık olacak biçimde, erişilebilir bir formatta ve anlaşılır bir dilde yayımlanması gerektiğini belirtmektedir. Hesap verebilirlik, eğitimde ölçme ve değerlendirme sürecinde paydaşların sonuçların doğruluğu, güvenirliği ve adilliği hakkında bilgi sahibi olmalarını sağlar. AERA (2014) hesap verebilirlikle ilgili bazı standartlar tanımlamıştır.

Standart 13.3. "Hesap verebilirlik indeksleri, program değerlendirmelerinde ve politika çalışmalarında etkililik göstergeleri ya da katma-değer modelleri gibi diğer istatistiksel modeller olarak kullanıldığında söz konusu indekslerin, göstergelerin veya modellerin hangi yöntemle oluşturulduğu açıklanmalı, gerekçelendirilmeli ve teknik özellikleri rapor edilmelidir." şeklinde ifade edilmektedir. Katma değerli değerlendirme yaklaşımı,

öğrenci ve okullara özgü farklılıkları kontrol altına alarak her bir öğrencinin sınav sonuçlarını inceleyerek öğrencinin ne kadar gelişim gösterdiğini belirleyen ve bu gelişimi okullarla ya da öğretmenlerle ilişkilendirmede objektif bir değerlendirme sağlayan sistem olarak ifade edilebilir (Şen ve ark., 2020). Test puanlarının birleştirilmesiyle oluşturulan indeksler, dayanak noktaları olan test puanlarında olduğu gibi geçerlik, güvenirlik ve adalet bakımından incelenip şeffaf biçimde raporlanmalıdır. Bu tür indekslerin oluşturulmasına ilişkin yöntem ve kurallar, teknik özelliklerinin dokümantasyonu ile birlikte kullanıcılara sunulmalıdır. Puanları birleştirmeye yönelik çeşitli yaklaşımların güçlü ve zayıf yönleri değerlendirilmeli; indekslerin, göstergelerin ya da modellerin bilgileri bağımsız araştırmacıların süreci tekrar edebilmesine imkân verecek şekilde açıkça paylaşılmalıdır.

Standart 13.4. "Program değerlendirmede, politika çalışmasında ya da hesap verebilirlik sisteminde kullanılan bir teste ilişkin geçerlik, güvenilirlik ve adalet kanıtları toplanmalı, bu kanıtlar ulaşılabilir olmalıdır" şeklinde ifade edilmektedir. Bir testin program değerlendirmede, politika çalışmalarında ya da hesap verebilirlik sistemlerinde kullanıma ve hedef örnekleme uygunluğu kanıtlarla ortaya konulmalıdır. Test sonuçlarının yayımlanmasından ya da raporlanmasından sorumlu kişiler, verilerin yanlış yorumlanmasını ve kötüye kullanımını önleyecek ek bilgileri açıklamalıdır. Özellikle bir değerlendirme ya da hesap verebilirlik sistemi; bir program, kurum ya da hizmet sağlayıcısının etkililiğine ilişkin yorumları desteklemek için tasarlanmışsa bu yorumların amaçlanan kullanımlar açısından geçerliği araştırılmalı ve belgelenmelidir. Bir test birden fazla amaç için kullanılıyorsa (örneğin; öğrencilerin bireysel başarısının izlenmesi, bireyler ya da öğrenci grupları için öğretim planlamasına yardımcı olacak bilgi sağlanması, bölgelerin, okulların veya öğretmenlerin değerlendirilmesi), bu kullanımların her biri için yorumların geçerliğine dayalı kanıtlar toplanmalı ve kullanıcılara sunulmalıdır. Personelin performansını değerlendirmek için kullanılan bir test söz konusu ise bu testin farklı personel grupları (örneğin öğretmenler, özel eğitim öğretmenleri, müdürler) için uygunluğu incelenmelidir.

Standart 13.5. "Testlerin değerlendirme ya da hesap verebilirlik amaçlarıyla geliştirilmesinden ve kullanılmasından sorumlu olan kişiler, test sonuçlarının uygulanacağı tüm gruplar için doğru yorumları ve uygun kullanım biçimlerini teşvik edecek adımlar atmalıdır" şeklinde ifade edilmektedir. Ölçme işleminden sorumlu kişiler, program ya da politika değerlendirmesinde yer alan farklı alt gruplardaki test katılımcıları için erişimi teşvik edecek ve yorumların geçerliğini en üst düzeye çıkaracak (örneğin uygun düzenlemeler sağlayarak) bir test süreci tasarlamalıdır. Diğer bir ifadeyle, test kullanıcıları, hesap verebilirlik amaçları doğrultusunda ilgili tüm alt grupları test programına dâhil etmek için her türlü çabayı göstermelidir. Ayrıca varsa dışlama kuralları, test düzenlemeleri (*modifications*) veya test/uygulama koşullarındaki diğer değişikliklerle ilgili belgeleri sağlamalı ve alt gruplar için test puanlarına dayalı olarak yapılan yorumlarının geçerliğine ilişkin kanıtlar sunmalıdır. Test puanlarının özetleri alt gruplara göre (örneğin ırk/etnik grup) ayrı ayrı raporlandığında test

kullanıcıları bu gruplar için puanların güvenilirliğini/kesinliğini ve puanlardan hareketle yapılan yorumların geçerliğini değerlendirmek üzere analizler yapmalı, sonuçların özetlerini yayınlarken bu bilgileri paylaşmalıdır. Hesap verebilirlik ya da uygulanan programın etkililiğini ölçmek için kullanılan karmaşık indekslerin analizleri, belirli alt gruplara ya da bu alt gruplara hizmet veren programlara/kurumlara karşı yanlılık olasılığını ele almalıdır. Örneğin indeksteki puanların, sınava giren kişilerin ırk/etnik köken gibi özellikleriyle ilgili bir yanlılık içerdiği belirlenirse bu puanlar yanlılık giderilmeden kullanılmamalıdır.

Standart 13.6. "Test performansındaki grup farklılıklarına ilişkin raporlar, farklılıkların anlamlı bir şekilde yorumlanabilmesi için mümkünse ilgili bağlamsal bilgilerle birlikte sunulmalıdır. Uygun bağlamsal bilgi mevcut değilse kullanıcılar yanlış yorumlamaya karşı uyarılmalıdır" şeklinde ifade edilmektedir. Grupların (örneğin cinsiyet, ırk/etnik köken, özel gereksinimli olma, dil yeterliliği, sosyoekonomik durum ya da coğrafi bölge) ortalama test puanları arasında gözlemlenen farklar; öğrenme fırsatı, eğitim deneyimi, çaba, eğitmen kalitesi, ebeveyn desteğinin düzeyi ve türü gibi faktörlerdeki değişkenliklerden kaynaklanabilmektedir. Eğitimde zaman içinde grup performansında ortaya çıkan farklılıklar, test edilen kişilerin yer aldığı örneklemdeki değişikliklerden (örneklem büyüklüğündeki değişiklikler dâhil) ya da bu bireylerin deneyimlerindeki değişikliklerden etkilenebilmektedir. Kullanıcılara, bu grup farklılıklarını yorumlarken ve bu farklılıklara dönük politikalar/uygulamalar tasarlarken uygun bağlamsal bilgileri dikkate almaları tavsiye edilmelidir. Ayrıca değerlendirmeler ulusal sınırlar içinde test puanlarının karşılaştırılmasını içeriyorsa puanların karşılaştırılabilirliğine ilişkin kanıtlar sunulmalıdır.

Standart 13.7. "Bir test değerlendirme ya da hesap verebilirlik uygulamalarında kullanılmak üzere seçildiğinde testin kullanım amacı ile beklenen sonuçları açıkça belirtilmeli ve uygunsuz kullanımlara karşı uyarılarda bulunulmalıdır." şeklinde ifade edilmektedir. Bazı durumlarda testin amacı, belli bir müfredat programını değerlendirmek gibi daha sınırlı olabilir ya da test değerlendirme hakkında bilgilendirme yapmak dışında belli sonuçları desteklemek için tasarlanmamış olabilir. Özellikle test temelli hesap verebilirlik sistemlerinde, karar vericilere işlevsel bilgiler sağladığı ve eğitimciler ile öğrencilerin daha iyi performans göstermelerini destekleyecek teşvikler oluşturarak eğitimin kalitesini artırdığı gerekçeleriyle testlerin kullanımı desteklenmektedir. Eğer böyle bir bakış açısı benimseniyor ya da testlerin kullanımı zorunlu hale getiriliyorsa mutlaka bu iddiaları destekleyen geçerlik kanıtları sunulmalıdır. Belirli bir geçerlik iddiasına ilişkin kanıtların toplanması ve raporlanması, program tasarımına dâhil edilmesi gereken işlerdir. Öğrencilerin başarısını artırmak gibi test kullanımının faydalarına ilişkin belirli bir iddia, ampirik verinin yanı sıra teorik ya da mantıksal argümanlarla da desteklenebilir. Belirtilen iddiayla çelişkili olan bilimsel literatürdeki bulgulara da dikkat edilmelidir.

Standart 13.8. "Testlerin politika, değerlendirme ve hesap verebilirlik bağlamlarında kullanılmasını zorunlu kılanlar ve testleri bu bağlamlara uygun olarak kullananlar, testlerin etkilerini izlemeli ve olumsuz sonuçları belirleyip en aza indirmelidir." şeklinde ifade edilmektedir. Testlerin politika, değerlendirme ve hesap verebilirlik amaçlarıyla kullanımı bazı durumlarda beklenmedik sonuçlara yol açabilir. Özellikle testin taşıdığı riskler yüksek olduğunda testleri zorunlu kılanlar ile test sonuçlarını kullananlar olası beklenmedik sonuçların belirlenmesine yönelik adımlar atmalıdır. Beklenmeyen olumsuz sonuçlar arasında; test maddelerinin önceden öğretilmesi, test uygulama prosedürlerinin değiştirilmesi, belirli katılımcıların teste girmesinin engellenmesi ya da test dışında tutulması yer alabilmektedir. Bu uygulamalar, testle ölçülmek istenen yapının gerçek performansı yansıtmamasına yol açabilmekte ve sahte yüksek puanlara ulaşılmasına neden olabilmektedir. Ayrıca bu uygulamaların kanunen yasaklanması söz konusu olabilir. Test prosedürleri bu tür sonuçların meydana gelme olasılığını en aza indirecek şekilde tasarlanmalı, kullanıcılara doğru olmayan test hazırlık uygulamalarından kaçınmaları için rehberlik ve teşvik sağlanmalıdır.

Standart 13.9. "Değerlendirme ya da hesap verebilirlik uygulamalarında, ek bilgi kullanımı genel yorumun geçerliğine katkıda bulunuyorsa test sonuçları diğer kaynaklardan gelen bilgilerle birlikte kullanılmalıdır" şeklinde ifade edilmektedir. Test dışındaki performans göstergeleri çoğu durumda gerekli ve neredeyse her zaman fayda sağlayıcıdır. İncelenen programın ya da politikanın kapsamlı bir resmini sunmak ve test sonuçlarının yorumlanmasına yardımcı olmak için genellikle hizmet alanların/yararlanıcıların özellikleri ile seçim ölçütleri, sunulan hizmetler, ortam ve kaynaklar gibi değişkenleri betimlemek ya da analiz etmek gerekir. Hesap verebilirlik bağlamında, bir öğretmen veya sağlık hizmeti sağlayıcısı gibi bireyler ya da kuruluşlar üzerinde büyük etkiye sahip olacak bir karar alırken test puanlarına yanında ilgili diğer bilgiler de göz önünde bulundurulmalıdır. Değerlendirmelere ya da hesap verebilirlik sistemlerine dâhil edilebilecek diğer bilgilere; eğitimcilerin ya da sağlık hizmeti sağlayıcılarının uygulamalarına ait ölçümler (örneğin sınıf gözlemleri, kontrol listeleri) ile öğrenci başarısının test dışı ölçümleri (ders alma, üniversiteye devam) örnek olarak verilebilir. Katma değerli modelleme durumunda, bazı araştırmacılar öğrencilerin demografik özelliklerini (örneğin ırk/etnik köken ve sosyoekonomik durum) kontrol değişkeni olarak ele almak gerektiğini savunurken diğer bazı araştırmacılar bu tür değişkenlerin ölçümlerin performansını iyileştirmediğini ve istenmeyen sonuçlara (bazı öğrenciler için diğerlerinden daha düşük standartlar belirlenmesi gibi) yol açabileceğini öne sürmektedir. Bu tür modellere hangi değişkenleri dahil etmek gerektiğine ilişkin dair kararlar, söz konusu değişkenleri dahil etmenin ya da hariç tutmanın ne tür etkilerinin olabileceğine ilişkin ampirik kanıtlara dayanmalıdır. Politika düzenlemelerinde test sonuçlarının yorumlanmasıyla ilgili ek bir bilgi türü, sınava girenlerin/test uygulanan kişilerin motivasyon derecesidir. Özellikle bireysel puanların test katılımcılarına bildirilmediği ya da puanların test katılımcıları için sonuçlarla ilişkilendirilmediği

durumlarda katılımcıların sınav deneyimini ciddiye alıp almadıklarını belirlemek önemli hale gelmektedir. Şüpheli motivasyona sahip bireylerin puanlarının dahil edilip edilmeyeceğine ilişkin karar kriterleri açıkça belgelenmelidir.

Tüm bu standartlar özetlendiğinde eğitimde hesap verebilirliğin; süreçlerin objektif bir şekilde yürütülmesi, sonuçların ve raporların herkes için anlaşılır ve şeffaf bir şekilde oluşturulması, geçerlik, güvenirlik ve adalet kanıtlarının sunulması ve bu süreçlerin farklı zamanlarda bağımsız araştırmacılar tarafından tekrar edilebilirliğini sağlayacak bilgilerin paylaşımı ile gerçekleştirilebileceği anlaşılmaktadır.

Hesap verebilirliğin bir parçası olarak şeffaflık (*transparency*) ölçme ve değerlendirme süreçlerinin tüm paydaşlar için açık, anlaşılır, erişilebilir olmasını gerektirir. Şeffaf(lık), açıkça görünen ve belirgin olan anlamına gelmektedir (TDK, n.d.-b). Ölçme ve değerlendirme süreçlerinde; testlerin geliştirilmesi/uyarlanması, uygulanması, puanlanması ve puanların yorumlaması ve tüm bunların raporlaştırılması aşamalarında şeffaf olunması gerekmektedir. Şeffaflık, hesap verebilirlik için önemli bir unsurdur. Öğrencilerin kullanılacak ölçme aracı ile hangi performans düzeyine hangi puanın karşılık geleceği hakkında bilgilendirilmesi, öğretmenin puanlama yaptığı aracı (dereceli puanlama anahtarı, cevap kâğıdı gibi) öğrencilerle paylaşması ve ölçümlerle ilgili olarak zamanında ve anlamlı geri bildirimlerde bulunması eğitimde ölçme-değerlendirme uygulamalarında şeffaflığın sağlanmasına katkı getirecek örneklerdir. Şeffaflığın sağlandığı sistemlerde hesap verebilirlik süreçleri hızlı ve verimli bir şekilde gerçekleşmektedir.

4. ÖLÇME VE DEĞERLENDİRMEDE YASAL YÜKÜMLÜLÜKLER

Yasal yükümlülük; bireylerin veya kurumların kanunlar ve ilgili mevzuat çerçevesinde yerine getirmek zorunda oldukları görevler ve sorumluluklar anlamına gelir (Dworkin, 2007). Ölçme ve değerlendirme süreçlerinde yasal yükümlülük; ölçmeyi yapan öğretmenler, eğitim kurumları ve kuruluşları ile sınava giren adayların uymaları gereken yükümlülükleri ifade eder. Hukuki bağlayıcılığı olan bu yükümlülükler hesap verebilirlik ve şeffaflık sağlar.

Millî Eğitim Bakanlığı (MEB) ilk kez 2023 yılında Ölçme ve Değerlendirme Yönetmeliği'ni (URL-4) yayınlamıştır. Bu yönetmeliğin amacı Madde 1'de "*Millî Eğitim Bakanlığı tarafından yapılan merkezî sistem sınavları, ulusal/uluslararası izleme araştırmaları, okul öncesi eğitim kurumları ve ilkokullarda akademik ve sosyal gelişimin takibi, ortaokul ve ortaöğretim kurumlarındaki ortak yazılı sınav faaliyetlerinin usul ve esasları ile ölçme değerlendirme merkezi müdürlüklerinin görev, yetki ve sorumluluklarını düzenlemektir.*" şeklinde açıklanmaktadır. İlgili yönetmelik "*Başlangıç Hükümleri; Ölçme ve Değerlendirme İlke ve Esasları; Sınav Uygulama Esasları ile Ulusal ve Uluslararası İzleme Araştırmaları; Merkezî Sistem Sınavlarına İlişkin Usul ve Esaslar; Ölçme Değerlendirme Merkezi Müdürlüklerinin (ÖDM) Kuruluşu, Açılması, Kapatılması ile Görev, Yetki ve*

Sorumlulukları ve *Çeşitli ve Son Hükümler*" olmak üzere altı bölümden oluşmaktadır. Yönetmeliğin içeriği özetlenecek olursa ulusal ve sınıf içi ölçmeler başta olmak üzere uluslararası ölçme ve değerlendirme süreçleri hakkında bilgiler verildiği, Türkçe dil becerisinin dört boyutta ölçüleceği, öğrencilerin akademik ve sosyal gelişimlerinin takibinin yapılacağı, ölçme ve değerlendirme süreçlerinin güçlendirilmesi, ÖDM'lerin etkililiklerinin artırılması, ülke çapındaki izleme çalışmalarıyla (ABİDE, okul tabanlı izleme) öğrenciler hakkında ulusal düzeyde geri bildirim sağlanması konuları üzerinde durulduğu görülmektedir (Şengül Avşar, 2023).

MEB Ortaöğretim Kurumları Yönetmeliği'nde MEB'e bağlı resmî ve özel örgün ortaöğretim kurumlarında eğitim, öğretim, yönetim ve işleyişe ilişkin usul ve esasları düzenlemiştir (URL-5). Bu yönetmeliğin dördüncü kısmı "Öğrenci Başarısının Değerlendirilmesi" olup "*Ölçme ve Değerlendirme*; *Sınavlar*; *Sınıf Geçme*; *Telafi Programı, Yoğunlaştırılmış Eğitim, Tamamlayıcı Eğitim ve Önceki Öğrenmelerin Tanınması*; *Okul Birincilerinin Tespiti"* şeklinde beş bölümden oluşmaktadır. Bu bölümler incelendiğinde kazanımlara ve ölçülen özelliğe uygun ölçme, üst düzey zihinsel becerilerin ölçülmesi, psikometrik özellikleri belirlenen nitelikli ölçme araçlarının kullanımı, sınavların, performans çalışmalarının, proje ve uygulamaların nasıl puanlanacağı, sınavların uygulanış biçimi (sınav sayısı, sınav süresi), mazeret sınavları, sınav sonuçlarının nasıl duyurulacağı, öğrencilerin sınav sonuçlarına itirazı, sınava katılmayan öğrencilerin değerlendirilmesi, sınıf geçme ölçütü, mezuniyet puanı gibi konularda yapılacaklar hakkında bilgiler verildiği görülmektedir (Şengül Avşar, 2023).

ÖSYM'ye ilişkin; "*6114 sayılı Ölçme, Seçme ve Yerleştirme Merkezi Hizmetleri Hakkında Kanun, Ölçme, Seçme ve Yerleştirme Merkezi Başkanlığı Personeli Görevde Yükselme ve Unvan Değişikliği Yönetmeliği; Ölçme, Seçme ve Yerleştirme Merkezi Başkanlığı Disiplin Amirleri Yönetmeliği; ÖSYM Uzman ve Uzman Yardımcılığı Sınav, Atama, Yetiştirilme, Görev ve Çalışma Usul ve Esasları Hakkında Yönetmelik; Ölçme, Seçme ve Yerleştirme Merkezi Başkanlığında Çalıştırılacak Yerli ve Yabancı Sözleşmeli Personel Hakkında Yönetmelik; Ölçme, Seçme ve Yerleştirme Merkezi Başkanlığının Sınav Hizmetlerine Yönelik Mal ve Hizmet Alımları Yönetmeliği; Adayların ve Sınav Görevlilerinin Sınav Binalarına Giriş Koşullarına İlişkin Yönetmelik; Soruların Hazırlanması, Soru Havuzunun Oluşturulması ve Soruların Güvenliğinin Sağlanmasına İlişkin Yönetmelik; Ölçme, Seçme ve Yerleştirme Merkezi Başkanlığı Sınav Koordinatörlükleri ve Başvuru Merkezlerinin Çalışma Usul ve Esasları Hakkında Yönetmelik*" olmak üzere çeşitli kanun ve yönetmeliklere ulaşılabilmektedir (ÖSYM, 2025). Bu kanun ve yönetmelikler genel olarak kurum içi düzenlemeler, aday ve sınav görevlileri ile sınav başvuru merkezleri üzerinde durmaktadır. İlgili kanun ve yönetmelikler detaylı incelendiğinde hem kurumun hem de paydaşların yerine getirmek zorunda olduğu görev ve sorumluluklara odaklandığı görülür. Örneğin *Adayların ve Sınav Görevlilerinin Sınav Binalarına Giriş Koşullarına İlişkin Yönetmelik'*te adaylar ile sınav görevlilerinin; sınav binasına ve

salonlarına ne zaman alınacağı, yanlarında neleri getirip getiremeyecekleri, kimliklerini beyan ederken kullanabilecekleri kimlik kartları, adaylar için sınav giriş ve görevliler için görevlendirme belgesi taşıma zorunluluğu, sınav görevlilerinin adayın kimliğinden şüphelendiği durumlarda neler yapması gerektiği gibi hususlar detayları ile açıklanmıştır.

Yasal yükümlükler, bireyler üzerinde önemli etkileri olan yüksek riskli testler için ayrı bir önem taşımaktadır. Bu testler sonuçlarının bireyler, okullar ya da kurumlar üzerinde önemli ve ciddi sonuçlar doğurduğu ölçme araçlarıdır. Yüksek riskli testler; bir kişinin gelecekteki akademik kariyerini ve iş fırsatlarını belirleyen, öğrencilerin bir okulu bitirmek, bir eğitim programına girmek, bir üniversiteye devam etmek, burs alabilmek ya da bir başvuru için yeterlilik elde etmek için geçmeleri gereken testleri ifade etmektedir (Cizek, 2001; Jones ve ark., 2003). Bu testler sadece bireyleri değil öğretmenleri, aileleri, sosyal çevreyi de etkilemektedir (Johnson, 2017). Yüksek riskli testler; *öğrenciler* için lise mezuniyeti ya da sınıf yükseltme, *eğitimciler* için iş süreçleri/terfiler, maaşlar/promosyonlar vb. kararlar alma, *okullar ve bölgeler* için bütçe belirleme ve okulları yeniden yapılandırma gibi çalışmalara yön verebilmektedir (Blazer, 2011). ÖSYM tarafından düzenlenen üniversiteye giriş sınavları, kamu personeli seçme sınavları bu testlere örnek olarak gösterilebilir. Böylesine önemli testlerin geliştirilmesinden uygulanmasına, uygulanmasından puanlanmasına, puanlanmasından sonuçların duyurulmasına kadar tüm süreçlerde titiz çalışmak gerekmektedir. Her aşamada tüm paydaşları bağlayıcı yasal yükümlülüklerin olması, süreçleri güçlendirerek şeffaf ve hesap verebilir hale getirmektedir. Özetle tüm paydaşları ilgilendiren ve önemli bir sorumluluk alanı olan yasal yükümlülükler, hesap verebilirliği ve şeffaflığı sağlayarak ölçme ve değerlendirme süreçlerini sağlamlaştırmaktadır.

5. SONUÇ

Ölçme ve değerlendirme süreçleri çoklu paydaşlarıyla birlikte ciddi bir sorumluluk gerektirmektedir. Testleri geliştiren ve değerlendirme hizmeti verenler; öncelikle teknik anlamda psikometrik özellikleri üzerinde çalışılmış nitelikli ölçme araçlarını geliştirmeli/uyarlamalı, kapsayıcılığa dikkat etmeli, gerekli gizliliği koruyarak süreçleri şeffaf bir şekilde yürütmelidir. Eğitimde kullanılacak ölçme ve değerlendirme ürün ve hizmetlerini seçen veya bu seçime rehberlik edenlerin de sahip olması gereken sorumluluklar vardır. Bu kişiler seçtikleri ölçme araçlarının ölçme amacına uygunluğundan emin olmalıdırlar. Bu da bu kişilerin alan uzmanı ve ölçme ve değerlendirme okuryazarlığına sahip kişiler olmasını gerektirir. Ölçme ve değerlendirme süreçlerini yönetenler de sınavların adil ve uygun bir şekilde uygulanmasını sağlamalıdırlar. Bunu yaparken yüksek standartlara sahip olunmalı, kapsayıcılığa dikkat edilmeli ve yanlılığa neden olabilecek unsurlar ortadan kaldırılmalıdır. Ölçme sonuçlarının uygun puanlama yöntemleriyle tüm paydaşlara zamanında ve doğru bir şekilde bildirilmesi esastır. Gerekli durumlarda sonuçlara itiraz hakkı verilmesi, bu

itirazlara bağlı olarak soru iptali veya cevap anahtarı değişikliği gibi süreçler objektif bir şekilde yürütülmelidir. Ölçme ve değerlendirme süreç içinde değişen, gelişen ya da tekrarlı eğitimlerle bilgilerin güncel tutulması gereken bir alandır. Bunun için ölçme ve değerlendirme alanında eğitim veren kişilerin alanında uzman profesyonel kişiler olması, farklı alan uzmanlarıyla uyumlu bir şekilde çalışması gerekir. Ayrıca ölçme ve değerlendirme sonuçları sistematik bir şekilde arşivlenip üzerinde gerekli araştırmalar yapılarak eğitim programlarının ve değerlendirmelerinin niteliği artırılabilir.

Adalet ölçme ve değerlendirmede üzerinde durulması gereken önemli bir kavramdır. En temel haliyle ölçmede adaletin sağlanabilmesi için testlerin tüm gruplara eşit fırsat sunması gerekir. Ölçme ve değerlendirmede temel standart geçerli ve güvenilir sonuçlara ulaşmaktır. Farklı uygulamalarda geçerli ve güvenilir sonuçlara ulaşmanın yolu ölçme araçlarının ölçmeyi hedeflediği özelliği, grubun özelliklerinden bağımsız ölçmesinden geçer. Bu da ölçmede yansızlığı gerektirir. Adil ölçme yapmanın bir diğer koşulu da yanlı ölçme yapmamaktır.

Ölçme ve değerlendirme süreçlerinde dikkat edilmesi gereken bir diğer konu kapsayıcılıktır. Kapsayıcılığın sağlanabilmesi için ölçme ve değerlendirme sürecinin planlama aşamasından başlanarak uygulama ve puanların duyurulmasına kadar tüm aşamalarda kapsayıcılık ilkeleri gözetilmelidir. Eğitim ortamlarında çeşitli değerlendirme yöntemleri, seçenekleri ve konuları sunma, açık, adil ve şeffaf değerlendirme ölçütlerini sağlama gibi çabalar kapsayıcılığı güçlendirir. Ayrıca kapsayıcılığın sadece özel gereksinimli bireylerle ilgili bir konu olmadığı, ölçme araçlarında çeşitlilik sağlanarak kişilerin kendilerini, yeteneklerini en iyi ifade edebilecekleri süreçler oluşturmak gibi çok daha geniş bir anlam taşıdığının farkında olunmalıdır.

Ölçme ve değerlendirmede etik sorumluluklar, adalet ve kapsayıcılık beraberinde bir diğer önemli kavram olan hesap verebilirliği gündeme getirir. Eğitimde testler ve test sonuçları hesap verebilirliğin bir ölçüsü olarak kullanılabilirken bu testlere ilişkin belgelerin de tüm paydaşlar için anlaşılabilir bir formatta erişime açık olması gerekir. Hesap verebilirlikle birlikte paydaşlar; sonuçların doğruluğu, güvenirliği ve adilliği hakkında bilgi sahibi olur. Son olarak eğitimciler, eğitim kurumları ve kuruluşları ile sınava giren adayların uymaları gereken yükümlülükleri ifade eden yasal yükümlülükler, hukuki bağlayıcılığıyla hesap verebilirlik ve şeffaflığı sağlar.

KAYNAKÇA | REFERENCES

Advance HE. (2019, 24 Ocak). *Inclusive assessment: Where next?* https://www.advance-he.ac.uk/news-and-views/inclusive-assessment-where-next

American Educational Research Association. (2011). *Code of ethics*. American Educational Research Association.

American Educational Research Association, American Psychological Association & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. American Educational Research Association.

American Psychological Association. (2004). *Code of fair testing practices in education*. Joint Committee on Testing Practices. https://www.apa.org/science/programs/testing/fair-testing.pdf

American Psychological Association. (t.y.-a). Ethics. *APA dictionary of psychology* içinde. https://dictionary.apa.org/ethics adresinden 7 Ekim 2025 tarihinde alınmıştır.

American Psychological Association. (t.y.-b). Fairness. *APA dictionary of psychology* içinde. https://dictionary.apa.org/fairness adresinden 7 Ekim 2025 tarihinde alınmıştır.

American Psychological Association. (t.y.-c). Bias. *APA dictionary of psychology* içinde. https://dictionary.apa.org/bias adresinden 7 Ekim 2025 tarihinde alınmıştır.

American Psychological Association. (t.y.-d). Accountability. *APA dictionary of psychology i*çinde. https://dictionary.apa.org/accountability adresinden 7 Ekim 2025 tarihinde alınmıştır.

Anderson, J. A. (2005). *Accountability in education*. International Institute for Educational Planning.

Arıkan, Ç. A., Uğurlu, S. & Atar, B. (2016). Mimic, sibtest, lojistik regresyon ve Mantel-Haenszel yöntemleriyle gerçekleştirilen DMF ve yanlılık çalışması. *Hacettepe University Journal of Education, 31*(1), 34–52. https://doi.org/10.16986/HUJE.2015014226

Ayan, C. (2024). Değişen madde fonksiyonu analizleri. A. Şengül Avşar (Ed.), *Jamovi ile kolay istatistik* (3. baskı, s. 277–290) içinde. Nobel Akademik Yayıncılık.

Blazer, C. (2011). *Unintended consequences of high-stakes testing* (Information Capsule No. 1008). Research Services, Miami-Dade County Public Schools.

Bowen, N. K. & Masa, R. D. (2015). Conducting measurement invariance tests with ordinal data: A guide for social work researchers. *Journal of the Society for Social Work and Research, 6*(2), 229–249. https://doi.org/10.1086/681607

Brown, G. T. L., Harris, L. R., O'Quin, C. & Lane, K. E. (2017). Using multi-group confirmatory factor analysis to evaluate cross-cultural research: Identifying and understanding non-invariance. *International Journal of Research & Method in Education, 40*(1), 66–90. https://doi.org/10.1080/1743727X.2015.1070823

Byrne, B. M. (2016). *Structural equation modeling with AMOS: Basic concepts, applications, and programming* (3. baskı). Routledge.

Cizek, G. J. (2001). More unintended consequences of high-stakes testing. *Educational Measurement: Issues and Practice, 20*(4), 19–27. https://doi.org/10.1111/j.1745-3992.2001.tb00072.x

Crocker, L. & Algina, J. (1986). *Introduction to classical and modern test theory*. Holt, Rinehart and Winston.

Dorans, N. J. & Holland, P. W. (1992). DIF detection and description: Mantel–Haenszel and standardization. *ETS Research Report Series, 1992*(1), i–40. https://doi.org/10.1002/j.2333-8504.1992.tb01440.x

Dworkin, R. (2007). The model of rules. L. B. Nielsen (Ed.), *Theoretical and empirical studies of rights* (s. 14–46) içinde. Routledge.

Ercikan, K., Gierl, M. J., McCreith, T., Puhan, G. & Koh, K. (2004). Comparability of bilingual versions of assessments. *Applied Measurement in Education, 17*(3), 301–321. https://doi.org/10.1207/s15324818ame1703_4

International Test Commission & Association of Test Publishers. (2025). *Technology-based assessment: Guidance for test developers and users*. https://www.intestcom.org/upload/media-library/tba-guidelines-ver-11-july-2025-1754044782Z3vV9.pdf

Johnson, P. B. (2017). High-stakes testing and NCLB. *What is authentic educational reform?* (s. 55–69) içinde. Routledge.

Jones, G. M., Jones, B. D. & Hargrove, T. Y. (2003). *The unintended consequences of high-stakes testing*. Bloomsbury.

Karabük Üniversitesi Medya. (2025, 21 Haziran). *ÖSYM'den KBÜ'lü akademisyenlere soru hazırlama ve test geliştirme eğitimi.* https://kbumedya.karabuk.edu.tr/osymden-kbulu-akademisyenlere-soru-hazirlama-ve-test-gelistirme-egitimi

Messick, S. (1998). Consequences of test interpretation and use. *Educational Measurement: Issues and Practice, 17*(2), 35–39. https://doi.org/10.1002/j.2333-8504.1998.tb01797.x

National Council on Measurement in Education. (2023). *Code of professional responsibilities in educational measurement* (2. baskı). https://ncme.org/wp-content/uploads/2025/10/NCME_STUC_2023_CodeofProfResp_Final_Updated.pdf

Norman, D. A. (2013). *The design of everyday things*. MIT Press.

Oxford University Centre for Teaching and Learning. (2025). *IncludED: Designing inclusive assessment*. https://www.ctl.ox.ac.uk/included-designing-inclusive-assessment

Ölçme, Seçme ve Yerleştirme Merkezi Başkanlığı. (2011). *Ölçme, Seçme ve Yerleştirme Merkezi Başkanlığının sınav hizmetlerine yönelik mal ve hizmet alımları yönetmeliği*. https://www.mevzuat.gov.tr/mevzuat?MevzuatNo=15562&MevzuatTur=7&MevzuatTertip=5

Ölçme, Seçme ve Yerleştirme Merkezi Başkanlığı. (2025). *Kanun ve yönetmelikler*. https://www.osym.gov.tr/TR,12849/kanun-ve-yonetmelikler.html

Stark, S., Chernyshenko, O. S. & Drasgow, F. (2006). Detecting differential item functioning with confirmatory factor analysis and item response theory: Toward a unified strategy. *Journal of Applied Psychology, 91*(6), 1292–1306. https://doi.org/10.1037/0021-9010.91.6.1292

Şen, S., Yıldırım, İ. & Karacabey, M. F. (2020). The implementation of value-added assessment. *Education and Science, 45*(202), 79–91. https://doi.org/10.15390/EB.2020.8023

Şengül Avşar, A. & Emons, W. H. M. (2021). A cross-cultural comparison of non-cognitive outputs. *Studies in Educational Evaluation, 70*, 101053. https://doi.org/10.1016/j.stueduc.2021.101053

Şengül Avşar, A. (2023). Ortaöğretim ve yükseköğretimde ölçme ve değerlendirme. H. Tabak & V. Özdemir (Ed.), *Sınıf içi öğrenmeleri ölçme ve değerlendirme* (s. 379–417) içinde. Pegem Akademi.

Türk Dil Kurumu. (n.d.-a). Adalet. *Güncel Türkçe Sözlük* içinde. https://sozluk.gov.tr/?q=adalet adresinden 7 Ekim 2025 tarihinde alınmıştır.

Türk Dil Kurumu. (n.d.-b). Şeffaflık. İçinde *Güncel Türkçe Sözlük* içinde. https://sozluk.gov.tr/?q=seffaflik adresinden 7 Ekim 2025 tarihinde alınmıştır.

Thurlow, M. L., Lazarus, S. S., Christensen, L. L. & Shyyan, V. (2016). *Principles and characteristics of inclusive assessment systems* (NCEO Report 400). https://nceo.umn.edu/docs/OnlinePubs/Report400/NCEOReport400.pdf

Tse, W. W. Y., Lai, M. H. C. & Zhang, Y. (2024). Does strict invariance matter? *Behavior Research Methods, 56*(4), 3117–3139. https://doi.org/10.3758/s13428-023-02247-6

Turgut, M. F. & Baykul, Y. (2013). *Eğitimde ölçme ve değerlendirme* (5. baskı). Pegem Akademi.

United Nations Development Programme. (2025). *Accountability*. https://www.undp.org/sites/g/files/zskgke326/files/2025-04/accountability-system-policy-popp-final-1.pdf

UNESCO. (2008). *Education for all by 2015: Will we make it?* https://unesdoc.unesco.org/ark:/48223/pf0000154743

UNESCO. (2015). *Education 2030: Incheon declaration and framework for action*. https://unesdoc.unesco.org/ark:/48223/pf0000245656

UNESCO. (2017). *Accountability in education: Meeting our commitments*. https://unesdoc.unesco.org/ark:/48223/pf0000259338

University of Bristol. (2025). *Principles for inclusive assessment*. https://www.bristol.ac.uk/bilt/sharing-practice/guides/inclusive-teaching-and-assessment/principles-for-inclusive-assessment/

University of Guelph. (t.y.). *Assessing student learning in in-person courses*. https://otl.uoguelph.ca/teaching-assessment-resources/assessing-student-learning-person-courses

URL-1. T.C. Ölçme, Seçme ve Yerleştirme Merkezi Başkanlığı. https://www.osym.gov.tr/TR,6548/arastirma-hakkinda.html

URL-2. T.C. Ölçme, Seçme ve Yerleştirme Merkezi Başkanlığı. https://www.osym.gov.tr/TR,8789/hakkinda.html

URL-3. UNESCO. *Global Education Monitoring Report.* https://www.unesco.org/gem-report/en/monitoring-sdg4

URL-4. Millî Eğitim Bakanliği Ölçme ve Değerlendirme Yönetmeliği. *Resmî Gazete,* 32304. T.C. Cumhurbaşkanlığı Mevzuat Bilgi Sistemi. https://www.mevzuat.gov.tr/mevzuat?MevzuatNo=40317&MevzuatTur=7&MevzuatTertip=5

URL-5. Millî Eğitim Bakanlığı Ortaöğretim Kurumları Yönetmeliği. *Resmi Gazete*, 28758. T.C. Cumhurbaşkanlığı Mevzuat Bilgi Sistemi. https://mevzuat.gov.tr/mevzuat?MevzuatNo=18812&MevzuatTur=7&MevzuatTertip=5

Vandenberg, R. J. & Lance, C. E. (2000). A review of MI literature. *Organizational Research Methods, 3*, 4–69. https://doi.org/10.1177/109442810031002

World Wide Web Consortium. (2025). *Web content accessibility guidelines (WCAG) 2.3*. https://www.w3.org/WAI/standards-guidelines/wcag/

Zumbo, B. D. (1999). *A handbook on the theory and methods of DIF*. Directorate of Human Resources Research and Evaluation.

Zumbo, B. D. (2007). Three generations of DIF analyses: Considering where it has been, where it is now, and where ıt ıs going. *Language Assessment Quarterly, 4*(2), 223–233. https://doi.org/10.1080/15434300701375832

Zumbo, B. D., Sireci, S. G. & Hambleton, R. K. (2003, Nisan). Re-visiting exploratory methods for construct comparability: Is there something to be gained from the ways of old?. Paper presented at the National Council on Measurement in Education, Chicago, IL.

YAZAR HAKKINDA | ABOUT AUTHOR

DOÇ. DR. ASİYE ŞENGÜL AVŞAR

Recep Tayyip Erdoğan Üniversitesi | asiye.sengul[at]erdogan.edu.tr

ORCID: 0000-0001-5522-2514

Doç. Dr. Asiye Şengül Avşar, Recep Tayyip Erdoğan Üniversitesi Eğitim Fakültesi, Eğitimde Ölçme ve Değerlendirme Anabilim Dalında öğretim üyesidir. Lisansüstü eğitimini Ankara Üniversitesi Eğitimde Ölçme ve Değerlendirme alanında tamamlayan Şengül Avşar, doktora sonrası araştırmacı olarak 2017-2018 yılları arasında Hollanda Tilburg Üniversitesi, Metodoloji ve İstatistik bölümünde, 2025-2026 yılları arasında Amerika Birleşik Devletleri'nde Georgia Institute of Technology Üniversitesi'nde psikoloji bölümünde bulunmuştur. Uzmanlık alanları arasında; veri bilimi, sosyal bilimlerde istatistik uygulamaları, öğrenci başarısının ölçülmesi, ölçme aracı geliştirme ve uyarlama, madde tepki kuramı modelleri ve uygulamaları, bireysel test puanlarının geçerliğinin belirlenmesi konuları yer almaktadır.

ASSOC. PROF. DR. ASİYE ŞENGÜL AVŞAR

Recep Tayyip Erdoğan University | asiye.sengul[at]erdogan.edu.tr

ORCID: 0000-0001-5522-2514

Assoc. Prof. Dr. Asiye Şengül Avşar is a faculty member in the Department of Measurement and Evaluation in Education at the Faculty of Education, Recep Tayyip Erdoğan University. Having completed her graduate studies in Measurement and Evaluation in Education at Ankara University, Şengül Avşar conducted postdoctoral research at Tilburg University (Netherlands) in the Department of Methodology and Statistics (2017–2018) and at the Georgia Institute of Technology (USA) in the Department of Psychology (2025–2026). Her areas of expertise include data science, statistical applications in the social sciences, the measurement of student achievement, the development and adaptation of measurement instruments, item response theory models and applications, and the determination of the validity of individual test scores.

15. BÖLÜM

ÖLÇME VE DEĞERLENDİRMEDE İLERİ YAKLAŞIMLAR

ADVANCED APPROACHES TO MEASUREMENT AND ASSESSMENT

Prof. Dr. Ayfer SAYIN

Doç. Dr. Selma ŞENEL

ÖLÇME VE DEĞERLENDİRMEDE İLERİ YAKLAŞIMLAR

Prof. Dr. Ayfer SAYIN
Gazi Üniversitesi

Doç. Dr. Selma ŞENEL
Balıkkesir Üniversitesi

Öz

Yapay zekâ teknolojisi ve dijitalleşme alanındaki gelişmeler, eğitimde ölçme ve değerlendirme süreçlerinde ileri yaklaşımların gelişip yaygınlaşmasını sağlamıştır. Bu doğrultuda bu bölümde ilk olarak BOBUT ve çok aşamalı testlere (ÇAT) ilişkin kuramsal temeller; giriş kuralı, madde seçme yöntemi, yetenek kestirimi, olabilirlik ve klasik Bayesçi yaklaşımlar, sağlam kestirimciler (robust estimator) ile sonlandırma kuralları gibi bileşenler üzerinden ayrıntılı biçimde ele alınmış, ayrıca panel deseni, modül uzunluğu ve yönlendirme koşulları gibi ÇAT'a özgü tasarım unsurlarına da yer verilmiştir. GRE, PISA ve SAT gibi uluslararası ölçeklerden sunulan tasarım ve test sistemi örnekleriyle, bu kuramsal tasarım kararlarının gerçek sınav sistemlerinde nasıl hayata geçirildiği somut biçimde gösterilmiştir. Bölümde ayrıca geleneksel madde yazımının zaman, maliyet ve öznellik sınırlılıklarına çözüm olarak geliştirilen otomatik madde üretimi (OMÜ) incelenmiştir. Model tabanlı, yapay zekâ tabanlı ve hibrit (köprü) yaklaşımlar; aşamaları, güçlü ve sınırlı yönleri ile birlikte ele alınmış, yaklaşımlar Türkçe uygulama örnekleri üzerinden somutlaştırılmıştır. Yapay zekâ ve büyük dil modellerinin sunduğu ölçeklenebilirlik olanaklarının yanında halüsinasyon, tutarsızlık, yanlılık, telif ve veri güvenliği gibi etik riskler tartışılmış; insan–yapay zekâ iş birliğine dayalı "artırılmış zekâ" yaklaşımının önemi vurgulanmıştır. Bu bölümde son olarak aynı yapıyı ölçen farklı test formlarından elde edilen puanların karşılaştırılabilir hâle getirilmesini amaçlayan test eşitleme konu edilmiştir. Eşitlemenin temel kavramları, koşulları (simetriklik, aynı yapıyı ölçme, eşit güvenirlik, eşitlik, gözlenen puan eşitliği ve grup değişmezliği) ile veri toplama desenleri (seçkisiz gruplar, tek grup, dengelenmiş grup, eş olmayan gruplarda ortak madde/ortak değişken) açıklanmıştır. Klasik Test Kuramı ve Madde Tepki Kuramı temelli eşitleme yöntemleri ve eşitleme hatası türleri; uluslararası ile ulusal uygulama örnekleriyle birlikte sunulmuştur.

Anahtar Kelimeler

Bilgisayar Ortamında Bireye Uyarlanmış Test (BOBUT), Çok Aşamalı Test (ÇAT), Otomatik Madde Üretimi, Yapay Zekâ Tabanlı Madde Üretimi, Test Eşitleme

ADVANCED APPROACHES TO MEASUREMENT AND ASSESSMENT

Prof. Dr. Ayfer SAYIN
Gazi University

Assoc. Prof. Dr. Selma ŞENEL
Balıkkesir University

Abstract

Developments in artificial intelligence and digitalization have helped new and advanced approaches grow and become more common in educational measurement and evaluation. In this chapter, we first explain the theoretical basis of Computerized Adaptive Testing (CAT) and Multistage Testing (MST). This includes components such as the entry rule, the item selection method, ability estimation, likelihood and classical Bayesian approaches, robust estimators, and termination rules. We also cover MST-specific design elements such as panel design, module length, and routing conditions. Using examples from international tests such as GRE, PISA, and SAT, we show clearly how these theoretical design choices are put into practice in real testing systems. The chapter also examines Automatic Item Generation (AIG), which was developed to solve the time, cost, and subjectivity limitations of traditional item writing. Model-based, artificial intelligence–based, and hybrid (bridge) approaches are presented together with their stages and their strengths and limitations, and they are illustrated through examples of applications in Turkish. In addition to the scalability that artificial intelligence and large language models offer, ethical risks such as hallucination, inconsistency, bias, copyright, and data security are discussed, and the importance of the "augmented intelligence" approach based on human–artificial intelligence collaboration is highlighted. Finally, the chapter deals with test equating, which aims to make scores from different test forms that measure the same construct comparable. It explains the basic concepts of equating, its conditions (symmetry, measuring the same construct, equal reliability, equity, observed score equity, and group invariance), and the data collection designs (random groups, single group, counterbalanced group, and common item/common variable in non-equivalent groups). Equating methods based on Classical Test Theory and Item Response Theory, as well as the types of equating error, are presented together with international and national application examples.

Keywords

Computerized Adaptive Testing (CAT), Multistage Testing (MST), Automatic İtem Generation (AIG), AI-based Item Generation, Test Equating

1. GİRİŞ

Geleneksel ölçme ve değerlendirme uygulamaları tüm adaylara aynı maddelerin aynı sırayla sunulduğu, kâğıt-kalem temelli ve puanlaması uygulama sonrasında yapılan doğrusal testlere dayanmaktadır. Bu testlerde her aday, yetenek düzeyinden bağımsız olarak kendisi için çok kolay ya da çok zor maddelerle karşılaşabilmekte; bu da hem test süresini uzatmakta hem de kestirimin belirli yetenek aralıkları dışında kesinliğini düşürmektedir. Dijitalleşme, büyük veri analitiği, yapay zekâ ve bilgisayar destekli sistemler; öğretim süreçleri gibi öğrenci başarısının belirlenmesi, izlenmesi ve yorumlanması süreçlerini de değiştirmiştir. Dijitalleşme ile testlerin bilgisayar ortamına taşınması, bu sınırlılıkların aşılmasına imkân tanımıştır. Madde Tepki Kuramı'nın da yaygınlaşmasıyla madde parametrelerinin önceden kestirilmesi, adayın yanıtlarına göre madde seçiminin anlık olarak yapılması ve sonuçların uygulama sırasında hesaplanabilmesi, ölçme sürecini adaya uyarlanabilen ve geri bildirim üretebilen bir yapıya dönüşmüştür.

Ölçme ve değerlendirme alanında yaşanan gelişmeler, ileri düzey yaklaşımların da yaygınlaşmasını sağlamıştır. Bunlardan biri; adaya yetenek düzeyine uygun maddeler ya da alt testler sunarak daha az madde ile kabul edilebilir kesinlikte kestirim elde etmeyi amaçlayan bireye uyarlanmış testlerdir. Bu yaklaşım, uyarlamanın madde düzeyinde yapıldığı Bilgisayar Ortamında Bireye Uyarlanmış Testler ile aşama düzeyinde yapıldığı Çok Aşamalı Testler biçiminde uygulanmaktadır. Bu tür uygulamaların gerektirdiği geniş madde havuzlarının oluşturulmasında karşılaşılan zaman, maliyet ve uzman yargısına bağlı öznellik sorunlarına yönelik geliştirilen otomatik madde üretimidir. Bu yöntem, tek tek madde yazımı yerine önceden tanımlanmış modeller ya da eğitilmiş dil modelleri aracılığıyla çok sayıda maddenin üretilmesine dayanır. Günümüzde ayrıca aynı amaç için uygulanan farklı test formları bulunmaktadır. Farklı test formlarından elde edilen puanların birbirinin yerine kullanılabilmesi için de test eşitleme işlemlerinin gerçekleştirilmesi beklenmektedir. Güvenlik gerekçesiyle bir sınavın çok sayıda alternatif formunun kullanılması, formlar arasında güçlük farklarının ortaya çıkmasına yol açmaktadır. Test eşitleme de bu farkların puanlar üzerindeki etkisini gidererek adayların yalnızca yetenek düzeyleri temelinde karşılaştırılmasını amaçlamaktadır.

Bireye uyarlanmış testler için geniş ve nitelikli madde havuzlarına ihtiyaç duyulmaktadır; otomatik madde üretimi bu havuzların oluşturulmasını hızlandırmakta, ürettiği madde ve form sayısının artması da farklı formlardan alınan puanların eşitlenmesi gereğini belirginleştirmektedir. Bu doğrultuda 15. bölümdeki konular Bilgisayar Ortamında Bireye Uyarlanmış Testler ile Çok Aşamalı Testler, Otomatik Madde Üretimi ve Test Eşitleme olmak üzere üç başlıkta ele alınmıştır. Her başlıkta konunun kuramsal temelleri açıklanmış, güçlü ve sınırlı yönleri tartışılmış ve ulusal ile uluslararası uygulama örnekleriyle desteklenmiştir. Böylece teknolojik gelişmelerin ölçme süreçlerine sağladığı olanakların yanında bu uygulamalarda geçerlik, güvenirlik ve adillik ölçütlerinin nasıl korunabileceği de ele alınmıştır.

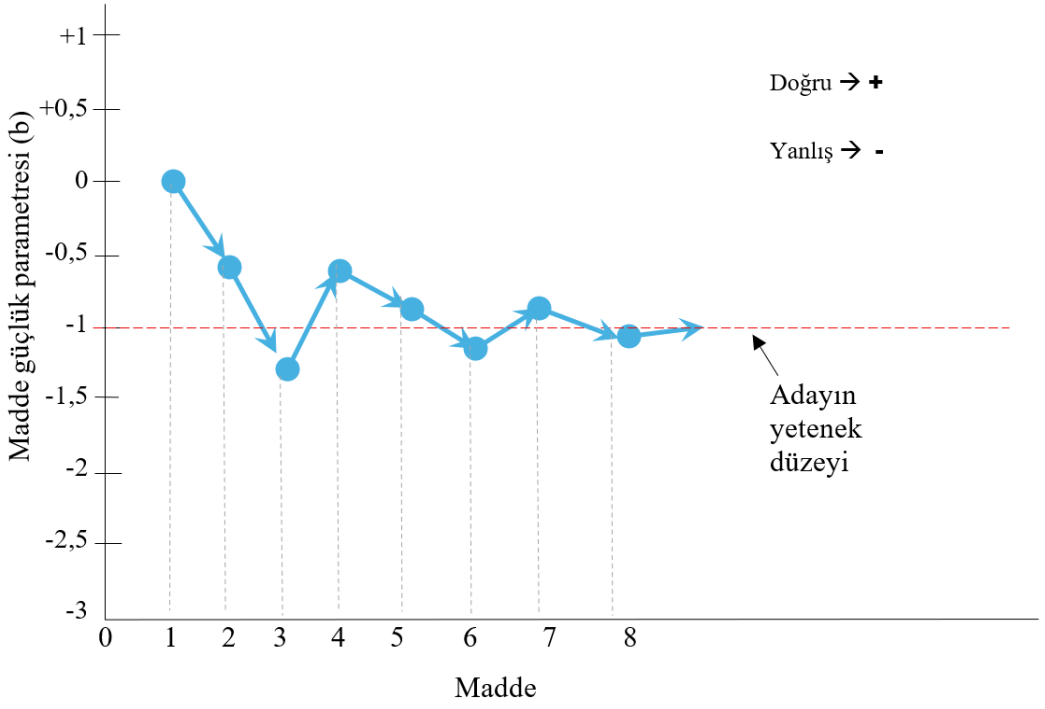
2. BİLGİSAYAR ORTAMINDA BİREYE UYARLANMIŞ TESTLER VE ÇOK AŞAMALI TESTLER

Bireye uyarlanmış testlerde, yetenek düzeyini daha yüksek kesinlikte tahmin edebilmek için bireye yetenek düzeyine uygun olan test maddeleri ya da alt testler yöneltilir. Bu yöntemde aday, yeteneğinin çok altına hitap eden, yani kendisi için çok kolay olan maddelerle veya yeteneğinin çok üstüne karşılık gelen, diğer bir deyişle kendisi için çok zor olan maddelerle karşılaşmaz. Buna bağlı olarak bireye uyarlanmış testler geleneksel doğrusal testlere kıyasla çok daha kısa olmakta, bu da daha kısa test süresi anlamına gelmektedir.

Bireye uyarlanmış testlerde uyarlama alt test ya da madde temelinde olabilir. Madde temelinde uyarlamada, adayın maddeye verdiği yanıtın doğruluğuna göre karşılacağı bir sonraki madde belirlenir. Birey maddeye doğru yanıt verdiğinde daha zor bir madde, yanıt yanlış verdiğinde ise daha kolay bir madde yöneltilir. Öte yandan alt test tabanlı uyarlamada adayın bir alt testten elde ettiği puan belirlenen yetenek puanının altındaysa daha kolay bir alt test, yetenek puanı yüksekse daha zor bir alt test uygulanır. Klasik bir madde tabanlı uyarlamanın ikili puanlanan maddelerdeki işleyişi Şekil 1'de, temel düzey bir test tabanlı uyarlamanın simgesel gösterimi ise Şekil 2'de sunulmuştur.

Şekil 1

Madde Tabanlı Uyarlamanın İkili Puanlanan Maddelerde İşleyişi

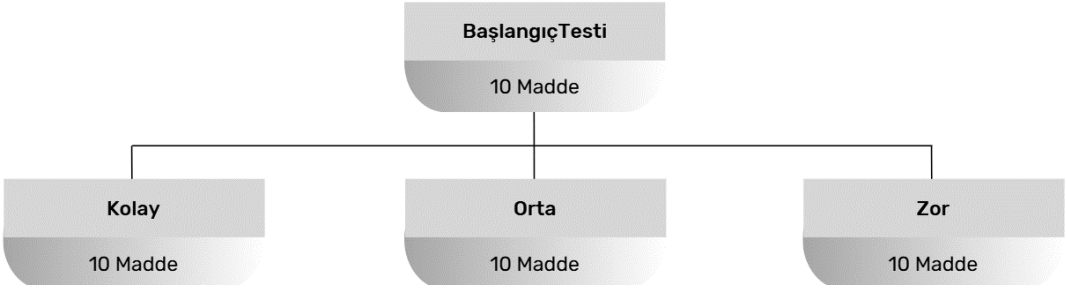


Şekil 1'de görüldüğü üzere adaya öncelikle orta güçlükte (b_1=0) bir madde uygulanır. Örnekte, aday birinci maddeyi yanlış yanıtladığı için ikinci madde olarak daha kolay bir madde yöneltilmiştir. Bundan sonra bireyin karşılacağın her yeni madde önceki yanıtınının doğru ya da yanlışlığına göre gelir. Bu işlem devam ettikçe uygulanan maddenin güçlük düzeyleri bireyin yeteneğini kestirmek üzere belli bir güçlüğe doğru yönelir ve grafikte bir düzleşme oluşmaya başlar. Düzleşme noktası bir anlamda bireyin yeteneğini işaret eder. Bireyin yeteneğine uygun maddelerin nasıl getirildiği ve yeteneğinin çok altında ya da üstünde soru görmediği bu şekil araclılığıyla görselleştirilmiştir.

Şekil 2'de ise basit bir test tabanlı uyarlama resmedilmiştir. Aday öncelikle, diğer aşamalara yönlendirmesinde önemli ve ilk test olan *başlangıç (yönlendirme) testi*ni alır. Başlangıç testinden aldığı puana göre bir sonraki aşamadaki olası alt testlerden, kendisi hakkında daha fazla bilgi sunacak alt teste yönlendirilir. Bu alt-test, adayın yeteneği yüksekse daha zor, düşükse daha kolay bir test olmaktadır. Şekil 2'deki örnekte her bir testte 10 madde bulunmaktadır. Test tabanlı uyarlamanın, madde tabanlı uyarlamadan daha az esnek bir yapı sunduğu buradan görülebilmektedir.

Şekil 2

Test Tabanlı Uyarlamanın Klasik İşleyişi



Bu noktaya kadar bireye uyarlanmış test yaklaşımın temellerinin anlaşılması için test tabanlı ve madde tabanlı uyarlama ifadeleri kullanılmıştır. Ancak bu uyarlama yaklaşımları; madde tabanlı ise "Bilgisayar Ortamında Bireye Uyarlanmış Test (Computerized Adaptive Testing [BOBUT; CAT])" ve test tabanlı uyarlama ise "Çok Aşamalı Test (Multi Stage Testing [ÇAT; MST])" olarak anılmaktadır.

Bu başlıkta vurgulanan temel özelleri bireye uyarlanmış testleri geleneksel testlere göre çeşitli açılardan daha güçlü kılmakta ama bazı açılardan da daha zorlayıcı hale getirmektedir. Bireye uyarlanmış testlerin güçlü yönleri ile zorlayıcı tarafları Şekil 3'te özetlenmiştir (Şenel, 2023).

Şekil 3

Bireye Uyarlanmış Testlerin Güçlü ve Zorlayıcı Yönleri (Şenel, 2023)

Güçlü Yönler

Psikometrik açıdan güçlü yönleri

- Bireysel güvenirlik kestirimi
- Güvenirliği yüksek yetenek kestirimi
- Kısa test ve test süresi
- Test eşitlemeye ihtiyaç duymadan puanların karşılaştırılabilirliği
- Bireysel test zamanı ve programı
- Yüksek test güvenliği

Olumlu psikolojik etkiler

- Bireye özel program seçeneği ile test kaygısını azaltma
- Yeterliğine uygun maddelerle test alma motivasyonunu canlı tutma ve cesaretlendirici olma

Bilgisayar tabanlı olması açısından güçlü yönleri

- Test düzenlemelerinde kolaylık
- Anında geri bildirim sağlama
- Yanıt süresi vb. ek veri sağlama
- Sürdürülebilirlik

Zorlayıcı Yönler

- Geniş ve nitelikli madde havuzu gerekliliğini sağlamanın zorluğu
- Toplumun olumsuz algısı (BOBUT algoritmasının karmaşıklığı ve anlaşılmaması nedeniyle)

Bilgisayar tabanlı olması açısından sınırlı yönleri

- Bilgisayar teknolojisi maliyetinin yüksek olması
- Teknolojik olanakların eşdeğerliğinin olmaması
- Teknoloji ve bilgisayar okuryazarlığının test puanını etkileyebilmesi
- Bilgisayar kaygısının test puanını etkileyebilmesi

Şekil 3'te özetlendiği üzere bireye uyarlanmış testlerin çok sayıda güçlü yönü bulunmaktadır. Ancak burada özellikle vurgulanması gereken güçlü yönleri öncelikle testlerin çok daha kısa olmasıdır. Aynı özelliği geleneksel testlerle ölçmek için 100 madde gerekiyorsa bireye uyarlanmış testlerde 15 madde yeterli olabilmektedir. Bu durum özellikle uzun oturumlar halinde yapılan testlerdeki yorulma etkisini ortadan kaldırmak için önemli bir güçtür. Bunun yanında, farklı dönemlerdeki puanların birbiriyle karşılaştırılabilirliği önemli olan testlerde (örneğin; Akademik Personel ve Lisansüstü Eğitimi Giriş Sınavı, Yabancı Dil Bilgisi Seviye Tespit Sınavı), daha karşılaştırılabilir puanlar sunmaktadır. Özellikle yılda birçok kez uygulanan sınavlarda maddelerin ifşası sorunu nedeniyle çok sayıda madde atıl hale gelebilmektedir. Buna karşın geniş madde havuzu koşulu sağlandığı taktirde BOBUT tek bir madde setinin binlerce kişiye uygulanması durumunu ortadan kaldırmaktadır. BOBUT sayesinde farklı bireylere geniş madde havuzundaki farklı maddeler yöneltilmekte, böylece sınav güvenliği artmaktadır. Madde havuzu kullanımı; maddelerin âtıl hale gelmesinin önüne geçmektedir. Ek olarak, her geçen yıl genişletilebilmektedir.

Bu güçlü yönlerinin yanında bireylere uyarlanmış testlerin bazı sınırlılıkları da vardır. Bireye uyarlanmış testlerde, özellikle BOBUT'ta, deneme uygulaması yapılmış ve madde parametreleri kestirilmiş geniş bir madde havuzuna ihtiyaç duyulur. Madde havuzunun ne kadar geniş olması gerektiğine ilişkin farklı tartışmalar bulunmaktadır. Psikolojik testlerde veya görece düşük riskli yetenek testlerinde daha sınırlı sayıda madde ile çalışıldığına yönelik örnekler bulunmaktadır (Aybek & Çıkrıkçı, 2018; Şenel & Kutlu, 2018). Bu bağlamda 100-200 maddeden oluşan madde havuzlarının kullanıldığı sıklıkla görülmektedir. Ancak özellikle ulusal çapta uygulanan ve binlerce bireyin gireceği bir sınavlarda madde havuzun çok geniş olması şarttır. Ulusal ölçekte uygulanan sınavlarda katılımcıların sosyo-ekonomik, kültürel ve eğitimsel geçmişleri çok çeşitlidir. Yeterince geniş bir madde havuzu, bu farklılıkları daha iyi kapsayarak sınavın adil ve temsil gücü yüksek olmasını sağlar. Ulusal sınavlarda, çok geniş gruplara uygulama yapıldığından havuzda farklı zorlukta ve içerikte yeterli sayıda madde olmalıdır. Geniş ve yeterli bir madde havuzu, farklı bireylere farklı maddeler yöneltmeyi mümkün kılarak sınav güvenliğini artırır. Aksi hâlde kopya çekme veya madde sızdırma riski daha yüksek olan ulusal sınavlarda; aynı maddelerin farklı sınavlarda tekrar edilmesi, ölçme güvenirliğini ve gizliliğini zedeler. Haliyle BOBUT'un güçlü yönlerinden istifade edilemez.

Madde parametreleri kestirilmiş madde havuzu, daha önce bir grupta denenmiş ve istatistikleri elde edilmiş maddelerden oluşan bir havuzu ifade eder. Fakat maddeleri bir grup üzerinde önceden denemek maddelerin ifşasıyla sonuçlanabilir. Bireye uyarlanmış testler, *ifşa edilen maddelerin kullanılması yönünde kamudan* olumsuz eleştiriler alabilmekte ve bahse konu eleştiriler nedeniyle daha az tercih edilebilmektedir. Bu kaygıların önüne geçmede deneme uygulamasını sınırlayacak bazı destekleyici yöntemler yol gösterici olabilmektedir.

- Testteki madde yapısı, zorluğu vb. benzer olan geçmiş sınavların verisinden, aynı yapıdaki maddeler için parametreler kestirilebilir.
- Her sınavda, adayın puanlamasına dahil olmayan deneme maddeleri adaylara uygulanır ve havuza parametreleri oluşturularak eklenir.
- Uzman yargısına dayalı olarak madde ayırt edicilik ve güçlük parametreleri oluşturulup ön parametreler olarak kullanılır. Gerçek veri ile parametreler güncellenir.
- Simüle edilmiş birey yanıtları üzerinden parametre kestirimi yapılabilir. Bu yöntem, testin başlangıç aşamasında fikir vermesi için kullanılır.

Özellikle geniş ölçekli ve yüksek riskli testlerde bireye uyarlanmış testler kullanımı için zorlayıcı olabilen bir başka faktör de şudur: Her bireyin farklı maddeler alması ancak puanlarının aynı yarışta karşı karşıya getirilmesi kamuoyu tarafından anlaşılmayabilir. "Farklı testleri alan adaylar nasıl karşılaştırılabilir puanlar elde ederler?" konusu

kamuoyuna iyi anlatılmalıdır. BOBUT algoritmasının karmaşıklığı, test sonuçlarına güven duyulmasının önünde r engel teşkil edebilir. Bu noktada özellikle bilgilendirmeler yapılması, hesap verebilirlik açısından elzemdir.

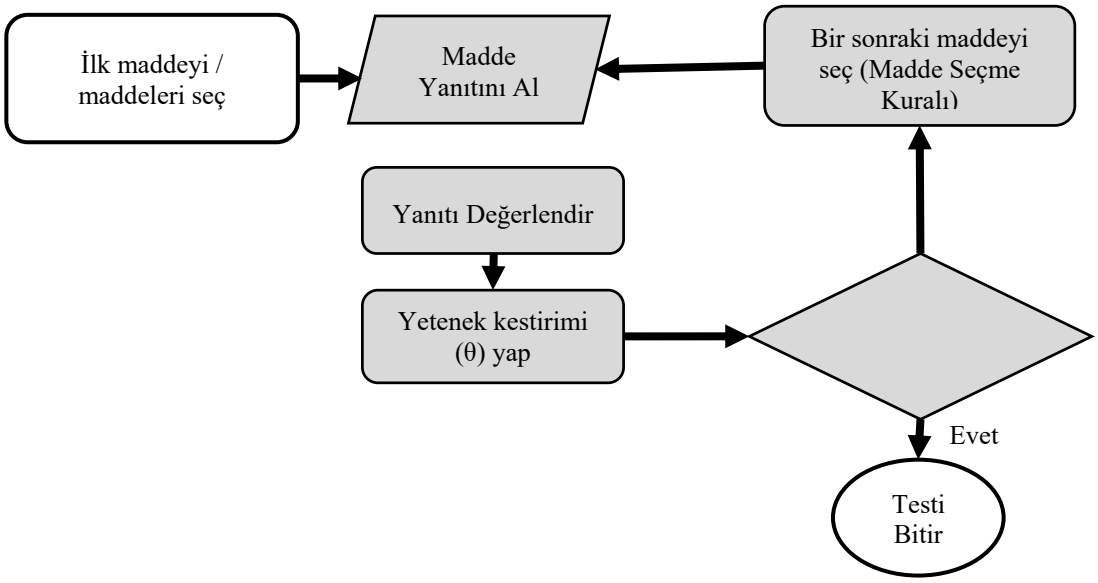
2.1. Bilgisayar Ortamında Bireye Uyarlanmış Testler

Madde tabanlı uyarlama yöntemi olan BOBUT, Madde Tepki Kuramı'nın (MTK) temel bir uygulamasıdır. BOBUT, bireyin yetenek düzeyine ilişkin anlık kestirimler yaparak sonraki adımda en fazla bilgi sağlayacak maddeyi seçme prensibine dayanır. Bu nedenle, bireyin performansına göre madde seçimini optimize eden MTK, BOBUT'un işleyişi için teorik ve pratik bir temel sunmaktadır.

BOBUT, Şekil 4'te sunulan dört aşamalı algoritma doğrultusunda işlemektedir. Bu algoritma; testin hangi maddeyle başlayacağını belirleyen *Giriş Kuralı*, her yetenek kestirimi sonrasında bir sonraki maddenin hangi yöntemle seçileceğini belirleyen *Madde Seçme Kuralı*, her madde uygulamasının ardından bireyin yetenek düzeyinin hangi yöntemle kestirileceğini belirleyen *Yetenek Kestirim Yöntemi* ve testin hangi koşullarda sonlandırılacağını tanımlayan *Sonlandırma Kuralı* aşamalarından oluşmaktadır.

Şekil 4

Temel BOBUT Algoritması



Şekil 4'te özetlenen BOBUT'un bu dört temel adımına ilişkin temel yöntemler aşağıda açıklanmıştır.

2.1.1. Giriş Kuralı

Giriş kuralı, BOBUT uygulamasında ilk madde ya da ilk birkaç maddeyi belirleyen kuraldır. Giriş kuralı olarak kullanılan yöntemler beş maddede özetlenebilir.

- Hedef grubun ***ortalama yetenek düzeyi***ne (normal dağılımda θ=0) yönelik yüksek bilgi sağlayan madde seçilmesi (Magis ve ark., 2018) en sık tercih edilen yöntemdir. Birey hakkında bilgi sahibi olmadığımız test başlangıcında, ortalama yetenek düzeyi olduğunu varsaymak iyi bir başlangıç noktası olacaktır.
- ***Orta güçlük düzeyi***nde (b_i=0) veya orta güçlüğe yakın [Örneğin, -0,5 < b_i < 0,5]) madde seçilmesi de giriş kuralı olarak iyi bir yaklaşımdır (Segall, 2004). Eğer, uygulanan grup normal dağılım gösteriyorsa bu yaklaşım ilk yaklaşıma benzer şekilde çalışır.
- Adayın testte ölçülen özelliği ne düzeyde sahip olduğu hakkında bir ön bilgi varsa giriş maddesi seçilirken bu bilgiden faydalanılabilir (Embretson & Reise, 2000). Örneğin adayın diploma notu, bir derse ait başarı puanı, akademik not ortalaması vb. biliniyorsa; ilgili yetenek düzeyi için en çok bilgi veren madde bireye yöneltilir.
- Farklı yetenek düzeylerine hitap eden birden çok madde uygulamak bir başka yöntemdir. Ancak bu yöntemde testi bir miktar uzatabilir.
- Madde havuzundan rastgele madde seçimi yaygın olmayan, testi uzatabilse de tercih edilen bir diğer yöntemdir.

2.1.2. Madde Seçme Yöntemi

BOBUT'un güçlü yönleri olan daha kısa testler sağlamada ve etkinliğinde *madde seçme yöntemi*nin önemi büyüktür. MTK çerçevesinde bir bireyin yetenek düzeyi (θ) bilindiğinde bu bireyin yüksek olasılıkla doğru ya da yanlış yanıtlayacağı önceden tahmin edilebilen bir maddenin uygulanması, birey hakkında ancak sınırlı bilgi sağlayacaktır. Bunun temel nedeni, kullanılan MTK modelinin, bireyin söz konusu maddeye vereceği yanıtı önceden yüksek doğrulukla öngörebilmesidir. Buna karşılık, bireyin yaklaşık %50 olasılıkla doğru yanıtlayabileceği bir madde, bireyin yetenek düzeyi hakkında daha fazla bilgi sunar. Bu nedenle, madde seçimine dayalı uyarlamalı test uygulamalarında, bireyin bilgi düzeyine en uygun maddelerin seçilmesi esastır. Bu yaklaşım, testin ölçme gücünü artırmakta ve birey hakkında daha isabetli sonuçlar elde edilmesini sağlamaktadır. Bu başlıkta, yaygın kullanılan bazı madde seçme yöntemleri ile bu yöntemlerin güçlü ve ayrışan yönlerine dair bir özet sunulmuştur.

En Yüksek Fisher Bilgisi (Maximum Fisher Information [MFI]) yöntemi, bireyin yetenek düzeyinde en fazla Fisher bilgisi sağlayan maddeyi seçen, tamamen bilginin hesaplanmasına dayanan ve en yaygın tercih edilen yöntemdir (van der Linden & Glas, 2000). En Yüksek Olabilirlik Ağırlıklandırılmış Bilgisi (Maximum Likelihood Weighted Information [MLWI]) ise MFI'dan farklı olarak adayın yetenek düzeyine ilişkin olabilirlik dağılımını dikkate alır. Başka bir ifadeyle, adayın yeteneği hakkındaki olasılıkları kullanarak madde seçimi yapar. Bu nedenle adayın yetenek düzeyi hakkında henüz çok az bilgiye sahip olduğumuz test başlangıcında, MFI'ya göre daha dengeli ve

güvenilir madde seçimi sağlar (Sarı & Raborn, 2018). Bunun yanında çok yüksek veya düşük yetenek düzeyine sahip adaylarda MLWI daha iyi kestirim yaparken orta yetenek düzeylerinde MFI daha iyi kestirimde bulunur (Jatoba, 2018; Jatobá ve ark., 2020).

Diğer yaygın yöntem grubu ise bireyin yetenek düzeyine ilişkin olasılık dağılımını (*posterior distribution*) dikkate alan Bayesçi madde seçme yöntemleridir. Bayes yaklaşımı madde seçiminde, bireyin yeteneği hakkındaki belirsizliği dikkate alır. Özellikle testin başlangıcında ya da daha az madde içeren testlerde daha dengeli ve güvenilir sonuçlar vermektedir (van der Linden, 1999; Thissen & Mislevy, 2000*). Sonsal Ağırlıklandırılmış Bilgi (Posterior Weighted Information[PWI])* ve *Beklenen Sonsal Bilgi (Expected Posterior Information [EPI])* en yaygın kullanılan Bayesçi yaklaşımlardır.

Kullback-Leibler Sapma Kriteri (Kullback-Leibler Deviation Criterion [KL]) hesaplama yükü daha fazla olan ancak en etkili yöntem olarak sıklıkla bulgulanan (Chang & Ying, 1996; Şenel, 2024) tercih edilen bir diğer yöntem olarak görülmektedir. Bir diğer grup olarak, hesaplama açısından daha basit olan yaklaşımlar basit uygulamalar için tercih edilebilir. Bunlardan ilki *En Uygun b Değeri (bOpt- b Optimum)* yöntemi, adayın ara yetenek kestirimine en yakın güçlük parametresine (b_i) sahip maddeyi seçer. Bir diğeri ise *En Uygun Theta (θ) (thOpt-theta Optimum) yöntemidir.* Havuzdaki her bir maddenin, en çok bilgi verdiği θ düzeyini hesaplar. BOBUT madde seçerken adayın θ'sı ile en uyumlu maddeye yönelir (Magis ve ark., 2018). thOpt ve bOpt yöntemleri hesaplaması kolay ve hızlı olan ancak sağlanan bilgi açısından sınırlı kalabilen yöntemlerdir.

2.1.3. Yetenek Kestirim Yöntemi

BOBUT algoritmasında yanıtladığı her maddeden sonra adayın yetenek kestirimi yapılır. Yapılan bu ara kestirime göre bir sonraki maddeye karar verilir. MTK'da yetenek kestiriminde; En Yüksek Olabilirlik (Maximum Likelihood[ML]), Ağırlıklandırılmış Olabilirlik (Weighted Likelihood [WL]), Bayesçi yöntemler, Sağlam Kestirimciler ve farklı yöntemleri birleştiren alternatif yaklaşımlar kullanılabilmektedir (Liao ve ark, 2012; Warm, 1989; Wang ve ark., 2020). Temel yetenek kestirim yöntemleri, güçlü ve ayrışan yönleriyle aşağıda özetlenmiştir.

2.1.4. Olabilirlik Yöntemleri

ML, en yaygın kullanılan yetenek kestirim yöntemdir. Tamamı doğru ya da tamamı yanlış yanıt örüntüleründe hesaplama yapamaz. Yazılımlar bu mükemmel yanıt örüntüleri devam ettiği sürece aday için uç yetenek puanı (Örneğin, tümü doğru yanıtlar için +4, tümü yanlış yanıtlar için -4) vererek ara kestirime devam eder. Hesaplanması kolaydır ve madde sayısı arttıkça daha güvenilir sonuçlar sunar (Wang & Vispoel,1998).

WL ise her bir gözlemin güvenirliğine göre katkısını farklı derecelerde hesaba katan bir ağırlıklandırma ile olabilirlik fonksiyonu kestirimi yapar (Warm, 1989). Bu ağırlıklandırmayla özellikle uçdeğerlerde ML'ye göre daha dengeli bir kestirim yapabilir. Önsel bilgiyi kullanma açısından Bayesçi yöntemlere benzese de daha fazla hesaplama gerektiren bir yöntemdir. ML yöntemi ve Bayesçi yöntemler arasında bir köprü gibidir.

2.1.5. Klasik Bayesçi Yaklaşımlar

En Yüksek Sonsal Kestirim (Maximum a Posteriori [MAP]) yöntemi, adayın yeteneği hakkındaki önsel bilgiyi (*prior*) ve gözlem sonuçlarını birlikte kullanır. Bayesçi yaklaşımların bilinen özelliği olarak, bireyin yetenek kestirimindeki belirsizliği dikkate aldığından kararlı tahminler sağlar. MAP, sonsal dağılımın tepe noktasını (modunu) bularak en olası parametre değerini yetenek kestirimi olarak verir. Bayesçi Sonsal Beklenti Kestirimi (Expected A Posteriori[EAP]) yöntemi ise gözlem verisinden gelen bilgiyi birlikte kullanır. EAP sonsal dağılımın ortalamasını alır. Yani dağılımın tüm şekli göz önünde bulundurur, MAP gibi sadece tepe değerini kullanmaz (van der Linden & Glas, 2000). Bayesçi yaklaşımlara yönelik temel eleştirilerden biri, analiz sürecinde kullanılan önsel dağılımların sonuçlar üzerinde öznellikten kaynaklanan bir yanlılık oluşturma riski taşımasıdır.

2.1.6. Sağlam Kestirimciler (Robust Estimator [RE])

Aykırı yani uç değerlerin yeteneğin yüksek hatayla kestirilmesine engel olan bir dizi yöntem için kullanılan yaklaşımın genel adıdır. Farklı modeller kurarak özellikle veri kalitesinin düşük olduğu durumlarda sağlam kestirim yapmayı sağlar (Liao ve ark., 2012). Gerçek verilerle çalışıldığında aşırı değerler daha mümkün olacağından daha güvenilir kestirimler sağlayabilir. Hesaplama karmaşıklığından dolayı uygulamada kullanımı sınırlandırabilmektedir.

2.1.7. Sonlandırma Kuralları

BOBUT algoritmasında sonlandırma kurallarını sabit uzunluklu ve değişken uzunluklu yaklaşımlar olarak ikiye ayırmak mümkündür. Sabit uzunluk yaklaşımı oldukça basit olup testin uyarlanabilme yeteneğini kısıtlar. Bu yaklaşımda test belirli madde sayısı ya da test süresi ile sınırlanır (Wainer, 2000). Daha açık bir anlatımla, test, her adaya önceden belirlenmiş sabit sayıda madde sunar ve o sayıya ulaşıldığında sonlanır. Her adaya "20" madde sunulması buna örnek verilebilir. Her adayın farkı sayıda ve farklı maddeler aldığı tam uyarlama sağlayan değişken uzunluklu BOBUT uygulaması kamuoyunda test sonuçlarının güvenirliğine ilişkin endişelere sebep olurken; adayların aynı sayıda madde aldığı sabit uzunluklu sonlandırma kamuoyunun teste bakışı açısından destekleyici olabilmektedir. Ancak bu durumda bazı adaylar için test gereksiz yere uzayabilmektedir. BOBUT'un çıkış noktası ve amacı için daha uyumlu olan yöntem ise değişken uzunluklu sonlandırma yaklaşımıdır. Bu yöntemlerde bireyin yetenek kestirimi kabul edilebilir bir kesinliğe ulaştığı nokada test sonlanır.

*Standart hata (SH) yöntem*i en temel değişken uzunluklu yöntemdir. Adaya uygulanan son maddeden sonra yapılan yetenek kestirimine ilişkin standart hata değeri, belirlenen eşik hata değerinden düşükse test sonlanır. Başka bir ifadeyle hatanın belli bir değerin altına düşmesi beklenir. Bazı durumlarda test bilgi fonksiyonundan faydalanılıp test hakkında belli bir bilgi sağlanınca test sonlandırılabilir. Test bilgi fonksiyonu $I(\theta)$ ile standart hata $SH(\theta)$ arasındaki ilişki dikkate alındığında bu yöntemler iki ayrı yaklaşım olarak değerlendirilmemiştir (Embretson & Reise, 2000).

*θ farkı yöntem*inde ise bir önceki maddeyi yanıtladıktan sonra kestirilen θ ile son θ kestirimi arasındaki fark, belirlenen kriter değerin altına düştüğünde testin sonlandırılır. Burada, θ değerinin belirli bir aralıkta seyredip ciddi değişiklik göstermemesi, güvenilir bir kestirim olarak değerlendirilir.

Sonlandırma kuralı olarak tek bir yöntem yerine birden çok yöntem kullanılması daha sık tercih edilen bir yaklaşımdır. Örneğin, değişken uzunluklu sonlandırma kuralına maksimum test uzunluğu kuralı eklenerek özellikle tutarsız yanıtlarla testin uzamayıp sonlanması sağlanır. Yani SH eşiği ya da maksimum test uzunluğu koşulundan herhangi biri sağlandığında test sonlanır. Bununla birlikte standart hata belirlenen eşik değerin altına inse bile testin belirli bir sayıda madde uygulanmadan sonlandırılması engellenebilir. Bu *minimum test uzunluğu* kuralı, çok kısa bir test alması durumunda adayın dikkatsiz yanıtlarındaki hata olasılığını tolere eder. Bunun yanında adayın çok kısa test alması dolayısıyla test sonucuna duyabileceği şüpheyi de azaltabilir.

2.1.8. BOBUT için Güncel Yaklaşımlar

2.1.8.1. Çok Boyutlu BOBUT (Multidimensional Computerized Adaptive Testing [MCAT])

Klasik BOBUT uygulamaları, tek boyutlu MTK modelleri üzerine kurulan ve tek bir yetenek alanının kestirimine olanak tanıyan biçimde tasarlanmaktadır. Örneğin bir yetenek testinde; sözel yetenek kestirimi için bir madde havuzu geliştirilerek BOBUT olarak uygulanabilir. Bir taraftan, sayısal yetenek de ayrı bir BOBUT olarak uygulanıp her bir boyuta ilişkin kestirim yapılabilir. Bu örnekte; sözel yetenek için oluşturulmuş BOBUT uygulamasını aday *n* sayıda maddeyle tamamlar ve yeteneği kestirilir. Ardından sayısal yeteneği ölçmek üzere geliştirilmiş madde havuzundan madde seçen başka bir BOBUT sunulur ve adayın sayısal yetenek puanı kestirilir. Çok Boyutlu Bilgisayar Ortamında Bireye Uyarlanmış Test (MCAT) ise aynı anda farklı yetenek boyutlarını tek bir BOBUT algoritması içinde kestirmeyi amaçlar. Bu nedenle genellikle Çok Boyutlu MTK modellerine dayalı olarak geliştirilir. Bununla beraber; her maddenin yalnızca tek bir boyutu ölçtüğü modeller (Between-Item Multidimensionality), birden fazla boyutu aynı anda ölçen maddelerle kurgulanan modeller (Within-Item Multidimensionality) ve her maddenin hem genel bir faktöre (örneğin genel yetenek) hem de özgül bir alt faktöre (örneğin sözel ya da sayısal yetenek) yüklendiği bifaktör modeller gibi farklı modelleme yaklaşımları ile de uygulanabilmektedir (Braeken & Paap, 2020; İnce Aracı & Tan, 2022).

2.1.8.2. Yapay Zekâ Destekli BOBUT

Yapay zekâ teknolojisi her geçen gün gelişmekte, kestiremediğimiz yenilikler sunmaktadır. Bu teknolojinin adayın yetenek kestirimini yapması ve yetenek kestirimine göre yeni maddeler getirmesi rahatlıkla geliştirilebilir sistemlerdir. Nitekim yapay zekâ destekli öğrenme içerikleri ya da içeriklere ilişkin madde seçimlerini uygulayan sistemler geliştirilmektedir (Settles ve ark., 2020).

Hâlihazırda *yapay zekâ destekli BOBUT yaklaşımlarına baktığımızda, genel olarak;* Makine Öğrenmesi, Doğal Dil İşleme ve Takviye Öğrenmesi (Reinforcement Learning) yaklaşımlarının vurgulandığı görülmektedir. Ancak bu yaklaşımların MTK odaklı ve literatürde tanımlanan anlamında klasik BOBUT olmadığı unutulmamalıdır. Doğal Dil İşleme ve Takviye Öğrenmesi, bir bakıma Makine Öğrenmesinin altında sınıflandırılabilir. Bu bölümde detaylı olarak ele alınamayan yapay zekâ teknolojileri; farklı yönlerden BOBUT amacını güçlendirici çok önemli gelişmelerdir. Yapay zekâ destekli BOBUT teknolojileri ile adayın önceki yanıtları ve davranışları analiz edilerek sonraki maddenin hangi güçlük düzeyinde olması gerektiği tahmin edilir. Bu tahmin, MTK temelli bilgi fonksiyonu yerine öğrenilmiş örüntülere dayanır. Öğrencinin yeterli olmadığı konu ve kazanımları belirlemek amacıyla bireye uyarlama yapılabilir. En kısa sürede en doğru yetenek kestirimini yapmak için her soru seçiminden sonra aldığı geri bildirimle öğrenmeye devam eden takviye öğrenmesi sistemleri de tercih edilmektedir. Öğrencinin öğrenme hedefleri, ilgi alanları veya geçmiş performansına göre maddeler filtrelenerek kişiselleştirilmiş madde seçimi olanağı tanınabilir.

2.1.8.3. Otomatik Madde Üretimi (AIG) Entegrasyonu

BOBUT, bir yandan her adaya uygun maddeler seçebilmek bir yandan da her adaya farklı maddeler yöneltip test güvenliğini koruyabilmek için geniş bir madde havuzuna gereksinim duyar. Bu sayede madde tekrarı azaltılabilir ve test güvenliği yükselir. AIG, bir madde modeli üzerinden binlerce varyasyon üretebilir. Gerçekten de Gierl ve Lai (2013) yaptıkları çalışmada tek bir madde modeliyle binlerce madde üretilebildiği çıkarımına varmıştır. BOBUT'un geniş madde havuzu gereksinimi en büyük sınırlılıklarından iken (Şenel, 2023), AIG teknolojileri bu süreci otomatikleştirerek söz konusu sınırlılığı bertaraf etme potansiyeli taşımaktadır (Sayın & Gierl, 2023). Özellikle büyük ölçekli test uygulamalarında, klasik yöntemlerle yeterli sayıda madde üretmek oldukça zor olduğundan AIG büyük bir güç kazandırmaktadır. AIG ile üretilen maddeler, önceden tanımlanmış parametre aralıklarına göre yapılandırılabilir. Bu sayede BOBUT'un madde seçimini bilgi fonksiyonuna göre optimize etmesini kolaylaştırır. Ayrıca, madde havuzunu dinamik biçimde genişletmek de mümkündür.

2.1.8.4. İçerik-Kısıtlı (Content-Restricted / Content-Balanced) BOBUT

Algoritmasında, içerik kontrolünü de sağlayan BOBUT uygulamaları içerik kısıtlı BOBUT olarak adlandırılır. Klasik bir BOBUT uygulamasında, algoritmanın odağı aday için en uygun maddeyi seçmektir. Maddenin ölçtüğü konu ağılıkları, kapsam odakta değildir. BOBUT algoritmasına, sınavın farklı alt konularından veya içerik alanlarından, önceden belirlenen oranlarda madde gelmesini sağlayan bir içerik kontrolü eklenebilir (Şenel, 2023). Burada amaç, bir yetenek kestiriminin yanında, sınavın belirli içerik alanlarını yeterince temsil etmesini sağlamaktır. Örneğin, matematiksel düşünme becerilerine yönelik bir BOBUT'ta alt konuların dağılımı; *aritmetik %40, cebir %20; geometri %20 ve*

%10 olasılık olarak kararlaştırılabilir. Belirlenen sayıda maddeye ulaşıldığında algoritma diğer içerik alanlarından madde seçmeye devam eder. Bu özelliği ile kapsam geçerliği destekleyici bir algoritma eklentisi gibi düşünülebilir.

2.1.8.5. Çok Yöntemli (Multimodal CAT) BOBUT

Bireye uyarlama ya da yetenek kestiriminde çok çeşitli ölçümlerin ve yanıtların dikkate alındığı BOBUT uygulamaları çok yöntemli BOBUT olarak adlandırılmaktadır. Farklı türde (metin, resim, video, ses vb.) madde ya da yanıtlar, adayın yanıt süresi, yaptığı hatalar gibi veriler de değerlendirmeye alınır. Çok kanallı ölçümler uyarlama sürecinde değerlendirildiği için daha karmaşık ama gerçek yaşama daha uygun sonuçlar sunabilir.

2.2. Çok Aşamalı Testler

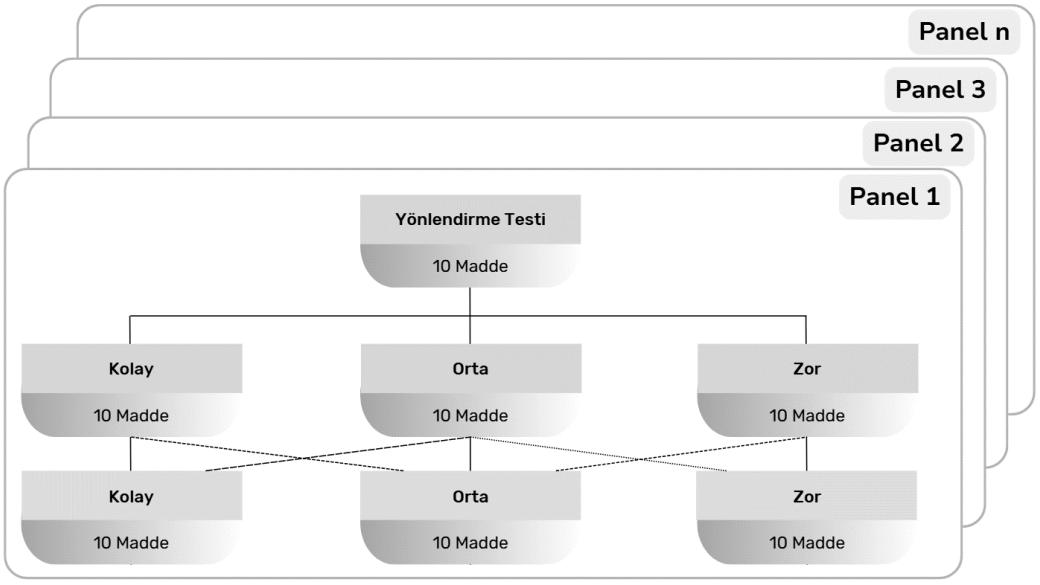
Bireye uyarlamanın madde tabanlı değil test tabanlı yapıldığı bireye uyarlanmış testler çok aşamalı testler (ÇAT) olarak adlandırılmaktadır. Başka bir ifadeyle testin adaya uygun ilerlemesi için her aşama sonunda bir uyarlama yapılır. Aday bir testle başlar, testten sonra elde edilen yetenek kestirimine göre bir sonraki test belirlenir. Klasik bir Çok Aşamalı Test gösterimi, iki aşamalı olarak Şekil 2'de sunulmuştur. Çok aşamalı testlerde her bir test bir modül olarak adlandırılır. Aday her aşamada bir modül alıp sonraki aşamaya geçerek aşamalar arasında ilerler. Bu bölümde *Panel Deseni, Modül Uzunluğu* ve *Yönlendirme Koşulları* çok aşamalı testlerin yapısını ve verimliliğini etkileyen temel unsurlar olarak ele alınmıştır.

2.2.1. Panel Deseni

Çok aşamalı testlerde, her bir aşama ve modül kombinasyonu bir "panel" olarak isimlendirilir. Panel yapısı, testin kapsam ve ölçme kesinliğini etkiler. Bir testte bir ya daha fazla panel kullanılabilir. Panel sayısının artması, madde havuzunun genişlemesi anlamına da geleceğinden, testin madde teşhir oranını düşürecektir. Şekil 5'te *n* panelli bir testin grafiksel gösterimi sunulmuştur.

Şekil 5

n Panelli Bir ÇAT Gösterimi



Şekil 5'te sunulan örnekte; Panel 1'in 1. aşamasında bir modül, 2. ve 3. aşamasında üçer modül bulunmaktadır. Bu panel yapısının literatürde 1-3-3 deseni olarak bilinir. Bu desendeki sayılardan her biri, her aşamadaki modülü göstermektedir. Panel desenleri çok çeşitli şekillerde yapılandırılabilir (Yiğiter & Doğan, 2023; Zenisky ve ark., 2010). Şekil 6'da bazı panel deseni örnekleri ve açıklamalarına yer verilmiştir:

Şekil 6

Çeşitli Panel Deseni Örnekleri

1-3 Panel Deseni: 1. aşamada bulunan bir başlangıç testinin arkasından, 2. aşamada kolay, orta, zor olmak üzere üç modülden oluşur. Kısa bir test isteniyorsa 1-3 deseni temel düzey bir uyarlama sağlar.

1-2-3 Panel Deseni: 3 aşamadan oluşur. 1. aşamada bir başlangıç testinin arkasından, 2. aşamada kolay ve zor olmak üzere iki; 3. aşamada ise kolay, orta, zor olmak üzere üç modülden meydana gelir. Toplam 6 modüllü bir yapı söz konusudur.

1-4-4 Panel Deseni: 3 aşamadan oluşur. 1. aşamada bir başlangıç testinin arkasından, 2. ve 3. aşamalarda dörder modül olmak üzere, toplamda 9 modüllü bir yapı vardır.

Elbette aşama sayısı ya da modül sayısı artırılarak daha uzun çok aşamalı testler yapılandırılabilir. Ancak bu durumda bireye uyarlamanın kısa test ve geçerliği yüksek sonuçlar avantajından uzaklaşılır. Bu doğrultuda alanyazında en verimli panel desenleri ile ilgili araştırmalar yürütülmektedir. Yapılan araştırmalar 1-2-3 ve 1-3-3 desenlerinde daha düşük ölçme hatası gözlendiğine işaret etmektedir. Üç aşamalı desenlerin iki aşamalı testlere göre daha verimli olduğu vurgulanmaktadır (Büyükkıdık ve ark., 2024; Kim ve ark., 2013; Yiğiter & Doğan, 2023).

Şekil 5'te görüldüğü üzere aday bir panelden teste başladığında bu panel üzerinde her bir aşamada başarısına göre bir modüle yönlendirilecek ve testi tamamlayacaktır. Öncelikle başlangıç testini alan adayın testi tamamladıktan sonra yeteneği kestirilecektir. Kestirim sonucundan hareketle, daha önceden belirlenen yönlendirme koşuluna göre (bkz. 15.3.3.3. Yönlendirme Koşulları) başarısı düşükse kolay, yüksekse daha zor bir modüle yönlendirilecektir. İkinci aşamadaki bu testi de tamamladıktan sonra, yenilenen yetenek kestirimine göre üçüncü aşama modülü belirlenecektir. Aday her aşamada bir modül tamamlayarak bu örneğe göre 30 maddede paneli tamamlamış olacaktır.

2.2.2. Modül Uzunluğu

Çok aşamalı testlerde, her aşamadaki modüllerin uzunluğu, diğer ifadeyle modüldeki madde sayısı ölçme doğruluğunu ve yanlılığı doğrudan etkileyen kritik bir özelliktir. Literatürdeki simülasyon çalışmalarına göre modül ne kadar uzun olursa ölçme hatası (RMSE) o kadar düşmektedir (Doğruöz & Kelecioğlu, 2024; Min & Bishop, 2024; Şahin, 2020). Ancak testin uzunluğundaki artış bireye uyarlanmış testlerin kısa test varsayımı ile çeliştiğinden ideal test uzunluğunun ya da minimum test uzunluğunun ne olabileceği üzerinde önemle durulması gereken bir konu olmaktadır.

Başlangıç modülleri teste başlangıç yapılan aşama olup test verimini doğrudan etkileyen bir noktadadır. Başlangıç modülü genellikle orta güçlükte bir test ile başlar. Berger ve arkadaşlarının (2019) yaptığı çalışmaya göre düşük ve yüksek yetenekli bireylerde çok uzun başlangıç modülleri, kısa başlangıç modüllerine göre daha verimsiz olmaktadır. Buna göre başlangıç modülü, sonraki aşamalarda yapılacak yönlendirme için yeterli bilgiyi sağlamalı ama öğrenciyi yormayacak kadar kısa olmalıdır.

Büyükkıdık ve Ayva Yörücü (2024) yaptıkları simülasyon çalışmasında 6, 12 ve 18 soruluk modülleri karşılaştırmış ve en az 12 soruluk modüllerin hem düşük hata oranı hem de yüksek ölçme doğruluğu sağladığını rapor etmişlerdir. Şahin (2020) de 10, 15 ve 20 maddelik modelleri karşılaştırmış, madde sayısının artmasıyla ölçme duyarlığının ve doğruluğunun arttığını belirtmiştir. Doğruöz ve Kelecioğlu da (2024) 6 ve 12 maddelik modülleri karşılaştırmış ve uzun modüllerin daha yüksek kesinlikte kestirimler sağladığını belirtmiştir. Kısaca literatürde başlangıç modülü için daha kısa testler olabilse de ara aşamalardaki her bir modül için 12 ve üstü madde sayısı önerilmektedir.

2.2.3. Yönlendirme Koşulları

Çok aşamalı testlerde, adayların bir aşamadan sonraki aşamaya yönlendirmede işletilen kurallar yönlendirme koşulları olarak adlandırılmaktadır. Yönlendirme kurallarının seçimi testin kapsamı, hedef kitlesi ve teknik altyapısı gibi faktörlere bağlı olarak değişkenlik gösterebilir. Bu bağlamda, test geliştiricilerinin hem ölçme duyarlılığı hem de uygulama verimliliğini göz önünde bulundurarak uygun yönlendirme stratejisini belirlemeleri önerilmektedir.

Çok aşamalı testlerde, adayların bir aşamadan sonraki aşamaya yönlendirilmesi statik ve dinamik yönlendirme olarak iki ayrı yaklaşımla uygulanmaktadır (Luecht, 2003; van der Linden & Glas, 2010). Statik yönlendirme yaklaşımında aday kestirilen yetenek düzeyine göre önceden belirlenen, diğer bir ifadeyle yönlendirme sürecinde değişmeyen bir kurala doğrultusunda bir sonraki modüle yönlendirilir. 1-3 panel deseni olan bir testte başlangıç testi sonrasında yetenek kestirimi -1'in altında olanların kolay, +1'in üstü olanların zor modüle yönlendirilmesi için bu kriter değerlerin test algoritmasına önceden girilmesi bu yaklaşıma örnek teşkil edebilir. Statik yönlendirmede kurallar önceden tanımlandığı için tüm adaylara aynı şekilde uygulanır ve test sırasında test sistemi açısından ek işlem gücü gerektirmez (Zenisky ve ark., 2010).

Dinamik yönlendirme ise daha karmaşık bir yapıya sahiptir. Bu yöntemde, adayın uygulanan maddelere verdiği yanıtlar, yetenek kestirimi ve potansiyel modüllerdeki madde bilgileri birlikte değerlendirilerek aday için en uygun modüle yönlendirme yapılır. Dolayısıyla her birey için uygulama anında bir yönlendirme hesaplaması yapılması gerekir. Bu yaklaşım, özellikle büyük ölçekli test uygulamalarında işlem yükü ve operasyonel karmaşıklık nedeniyle zorlayıcı olabilmektedir (Luecht, 2003; van der Linden & Glas, 2010; Luo & Wang, 2019).

2.3. ÇAT için Güncel Yaklaşımlar

2.3.1. Bireyselleştirilmiş Çok Aşamalı Test (BÇAT)(Targeted Multistage Testing)

Bireyselleştirilmiş çok aşamalı testler, çok aşamalı testleri uygulamadan önce bireyin daha önceki yetenek düzeyinden ya da test başarısını yordadığı bilinen başka bir değişkenden faydalanarak testi bireye uyarlamayı artıran bir yaklaşımdır. Test başarısıyla ilişkili olabilecek bu başlangıç testinin düzeyini etkileyen değişken sayesinde daha az madde ile yüksek güvenirlikte kestirim amaçlanır. Bireyselleştirilmiş çok aşamalı testlerin, özellikle üst ya da alt yetenek düzeyindeki adaylar için etkili sonuçlar ürettiği bilinmektedir (Berger ve ark., 2019). Bu yaklaşımda, ön değişken ile testte ölçülen özellik arasındaki yüksek bir korelasyon olması gerekir. Bu iki değişken arasındaki ilişkinin düşük olması, yanlı sonuçlara veya geçersiz erken sonlandırma kararlarına yol açabilir. Bu nedenle yönlendirmeyi sağlayacak ön değişken ya da değişkenleri istatistiksel hesaplamalarla doğrulamak önemlidir.

2.3.2. Değişken Uzunluklu Çok Aşamalı Test

Bu başlığa kadar değinilen çok aşamalı testler klasik anlamdaki çok aşamalı testleri, başka bir anlatımla sabit uzunluklu çok aşamalı testleri kapsıyordu. Her adayın aynı sayıda aşamadan geçerek aynı sayıda modülde, aynı sayıda madde alarak testi tamamladığı klasik yaklaşım sabit uzunluklu çok aşamalı test olarak adlandırılmaktadır. Değişken uzunluklu çok aşamalı testler ise aşama/modül sayısının veya uzunluğunun katılımcı yanıtlarına göre değişebildiği testlerdir. Bu yaklaşımla gereksiz yönlendirme aşamaları azaltılıp testin verimi ve ölçme duyarlılığı artırılmaya çalışılır. Bunu sağlamak için çeşitli algoritma ya da regresyona dayalı modeller kullanılabilmektedir (Han, 2019). Değişken uzunluklu Çok aşamalı testler için BOBUT uygulamalarının sonlandırma kuralları da ayrıca kullanılabilmekte; standart hatanın veya yetenek düzeyi değişiminin belirli bir değerin altına düşmesi sonlandırma kuralı olarak işe koşulabilmektedir (Hsu, 2024).

2.3.3. Yapay Zekâ Destekli Yönlendirme (AI-enhanced Routing)

Yapay zekâ desteğinden çok aşamalı testlerde, temel olarak yönlendirme amacıyla yararlanılmaktadır. Ancak yapay zeka başlangıç ataması, modül içi uyarlama ve testin sonlandırılması gibi süreçlerde de giderek daha yaygın kullanılmaktadır. Test sırasında toplanan adayın zaman kullanımı, tıklama işlemleri, göz hareketi, yanıt örüntüleri, yanıt süresi gibi davranışsal verileri kullanılarak yapay zekâdan testin tüm süreçlerinde faydalanılabilmektedir (Yan ve ark., 2024).

2.3.4. Bilişsel Tanılayıcı Çok Aşamalı Test (Cognitive Diagnostic Multistage Testing)

Bilişsel Tanılayıcı Çok Aşamalı Testler, bilişsel tanı modelleri ile çok aşamalı testleri birleştiren bir yaklaşımdır. Her aşama bir veya daha fazla bilişsel özelliği ölçmek üzere tasarlanır. Adayın modül sonuçları, ilgili özelliğin düzeyi hakkında bilgi verir. Sonuç olarak, *Bilişsel Tanılayıcı Çok Aşamalı Test* adayın ölçülen özelliklere ilişkin profiline ulaşmak için en bilgilendirici modülleri seçmeye odaklanır. Klasik çok aşamalı testteki gibi en yüksek bilgi için uygun testi seçmek odakta değildir. Test sonucunda, çok aşamalı testlere göre daha ayrıntılı tanılama bilgisi sunulur (Kim & Yoo, 2024).

2.4. Uluslararası BOBUT ve ÇAT Uygulamalarından Örnekler

Uluslararası büyük ölçekli testlerde BOBUT ve çok aşamalı testler yaygın olarak kullanılmaktadır. Farklı yapıları ve desenleri ile birçok yönden ayrışsa ve her geçen yıl farklı psikometrik ve teknolojik özelliklerle zenginleşse de temel olarak bireye uyarlanmış testler çatısında toplanabilen bazı testler bu başlıkta kısaca tanıtılmıştır.

2.4.1. GRE

Lisansüstü Eğitim Giriş Sınavı (Graduate Record Examination - GRE), lisansüstü programlara başvuran adayların sözel akıl yürütme, sayısal akıl yürütme ve analitik yazma becerilerini ölçen, uluslararası bir sınavdır. Testin yapısı ile ilgili Tablo 1'da özetlenmiştir.

Tablo 1

GRE Yapısına İlişkin Özet Tablo (Educational Test Service, 2023a)

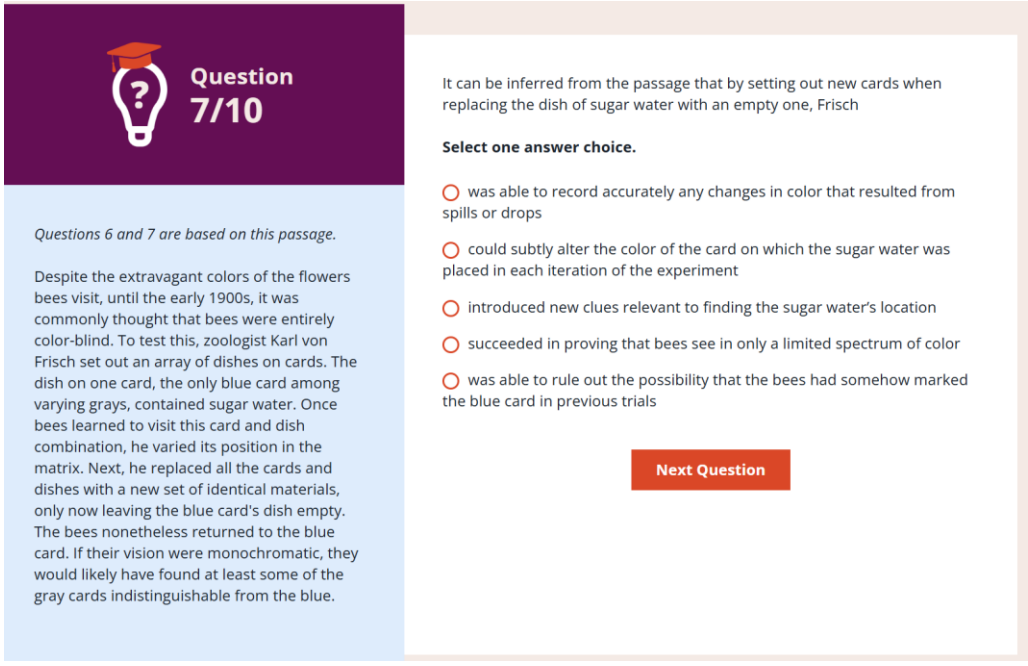
Ölçülen Özellik	Soru Sayısı	Süre
Analitik Yazma (Bir bölüm)	Bir "Bir Sorunu Analiz Et" görevi	30 dakika
Sözel Muhakeme (İki bölüm)	Bölüm 1: 12 soru Bölüm 2: 15 soru	Bölüm 1: 18 dakika Bölüm 2: 23 dakika
Nicel Muhakeme (İki bölüm)	Bölüm 1: 12 soru Bölüm 2: 15 soru	Bölüm 1: 21 dakika Bölüm 2: 26 dakika

GRE'deki Sözel Muhakeme (Verbal Reasoning) ve Sayısal Muhakeme (Quantitative Reasoning) bölümleri bölüm düzeyinde uyarlanabilir (section-level adaptive) olarak tasarlanmıştır. Örneğin, Sözel muhakeme için ilk bölüm ortalama zorlukta sorular içerir. Bu bölümdeki performansa göre Sözel muhakemenin ikinci bölümünün zorluk düzeyi belirlenir (Educational Test Service, 2023a).

Test tasarım özelliklerini görebilmekeri ve testi deneyimleyebilmekeri için GRE adaylara deneme uygulaması olanağı sunmaktadır (Educational Test Service, 2023b). Deneme ekranına ilişkin bir örnek Şekil 7'de sunulmuştur.

Şekil 7

GRE Deneme Uygulaması Ekranı



Şekil 7'de görüleceği üzere test arayüzü oldukça sade bir yapıya sahiptir. Bu sadelik, bilinçli bir tercihtir; çünkü ölçme araçlarında görsel yoğunluk veya tasarımsal karmaşa, dikkati dağıtabilmektedir. Özellikle bireye uyarlanmış test tasarımlarında, ölçülmek istenen özelliğin dışındaki herhangi bir dikkat çekici öge, sonuçlara hata karışmasına neden olabilir. Sonuç olarak, test arayüzlerinde yalnızca işlevsel ve ölçme amacına hizmet eden ögelere yer verilmesi önemlidir. Tasarım sürecinde, görsel sadelik ve bilişsel yükün minimize edilmesi ilkeleri gözetilmelidir.

2.4.2. PISA

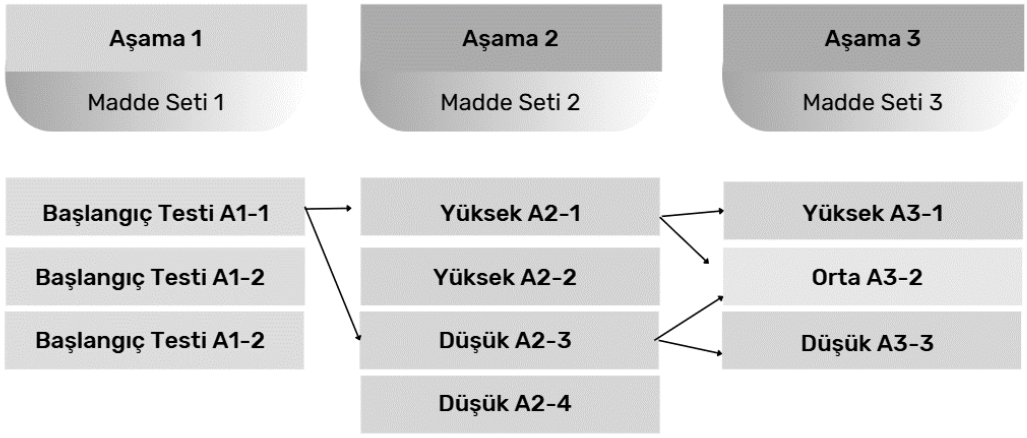
Uluslararası Öğrenci Değerlendirme Programı (Programme for International Student Assessment [PISA]), Ekonomik İşbirliği ve Kalkınma Örgütü (OECD) tarafından yürütülen ve 15 yaş grubundaki öğrencilerin okuma, matematik ve fen alanlarındaki bilgi ve becerilerini, gerçek yaşam problemlerini çözme bağlamında değerlendiren uluslararası bir ölçme programıdır (OECD, 2023). PISA, ölçme sürecinin verimliliğini ve duyarlılığını artırmak amacıyla 2018 yılında ilk olarak okuma becerileri alanında çok aşamalı test yöntemini uygulamaya geçmiştir. Ardından, 2022 yılında matematik ve fen bilimleri alanlarında da aynı yaklaşım benimsenerek uygulanmaya başlanmıştır. Bu geçiş, hem ölçme araçlarının bireysel düzeyde daha hassas sonuçlar üretmesini sağlamakta hem de test sürecinin daha etkin yönetilmesine olanak tanımaktadır.

PISA 2025 teknik raporuna göre (OECD, 2025_a), çok aşamalı test uygulamasında aday önce bir başlangıç testi almaktadır. Bu testten sonra elde edilen kestirime göre belirlenen zorluk düzeyine sahip modüller sunulmaktadır. Modüller, belirli yetenek aralıklarını hedefleyecek şekilde oluşturulmuştur. Adayın başlangıç testindeki performansı, önceden belirlenmiş kesme puanlarına göre değerlendirildiği için sabit yönlendirme koşulu uygulandığı belirtilebilir.

PISA Okuma ÇAT örneği üzerinden PISA'daki çok aşamalı testin yapısı anlaşılabilir. Üç aşamalı bir test olarak yapılandırılan bu testte, adaylar ilk aşamaya rastgele bir başlangıç testi ile başlar. Ardından aldıkları puana göre ikinci aşamada düşük veya yüksek zorlukta modüllere yönlendirilirler. İkinci aşamadaki performanslarına göre ise üçüncü aşamada düşük, orta veya yüksek zorlukta modüllere atanırlar. Ancak, testin içerik kapsamını dengelemek ve madde maruziyetini eşit dağıtmak amacıyla gruplar (A, B ve C) ve madde setleri ile biraz daha karmaşık bir yapı sunmaktadır. Bu yapı Şekil 8'de resmedilmiştir.

Şekil 8

PISA ÇAT Okuma Test Deseninde Grup A İçin Dört Olası Yol (OECD, 2025b)



PISA ÇAT yapısında yer alan her madde setinde; birinci aşama için 3 başlangıç modülü; ikinci aşama için 2 düşük zorlukta modül, 2 yüksek zorlukta modül ve üçüncü aşama 3 için 1 düşük, 1 orta, 1 yüksek zorlukta modül bulunmaktadır. Yani her madde setinde, farklı aşamalarda kullanılmak üzere toplamda 9 modül bulunur. Öğrencilerin hangi madde seti ile hangi aşamada karşılaşacaklarını belirlemek için A, B ve C şeklinde üç grup oluşturulmuştur. Farklı madde setleri olarak adlandırılan, ayrı panellerle her grup farklı aşamalarda karşılaşmaktadır. Tablo 2'de bu aşamalardaki madde seti değişimi özetlenmiştir. Üç grup (A, B, C) için toplamda 36 farklı yol tanımlanmıştır.

Tablo 2

Madde Setlerinin Döndürülmüş Tasarımı (OECD, 2025b)

Grup	Aşama 1	Aşama 2	Aşama 3
A	Madde Seti 1	Madde Seti 2	Madde Seti 3
B	Madde Seti 2	Madde Seti 3	Madde Seti 1
C	Madde Seti 3	Madde Seti 1	Madde Seti 2

Tablo 2'de özetlenen yapı sayesinde rastgele olarak Grup A, B veya C'ye atanan adaylar için madde setlerinin Aşama 1, 2 ve 3 boyunca nasıl ilerleyeceği belirlenmiş olur.

2.4.3. SAT

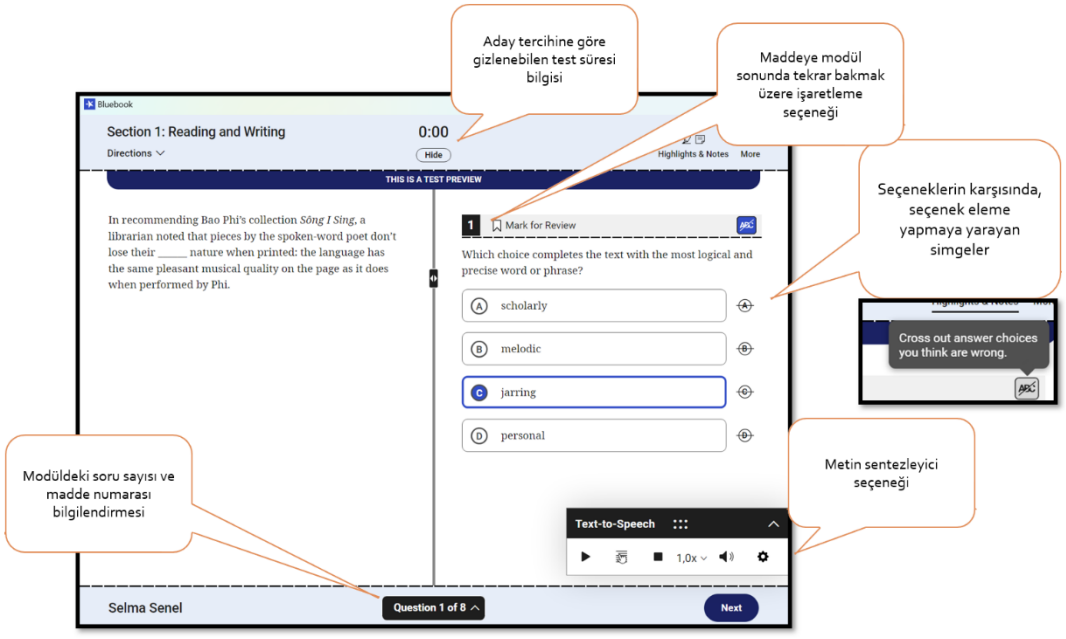
SAT (Scholastic Assessment Test), Amerika Birleşik Devletleri'nde üniversiteye girişte kullanılan ve öğrencilerin okuma, yazma, dil ve matematik alanlarındaki akademik hazırlık düzeyini ölçen bir testtir (College Board, 2025a). SAT'ın 2024'ten itibaren hayata

geçirilen dijital formu, çok aşamalı test olarak uygulanmaktadır. *Okuma ve yazma* ile *Matematik* bölümleri iki aşamadan oluşan iki ayrı test olarak uygulanmaktadır. Adayın ilk aşamadaki performansına göre ikinci aşamada güçlük düzeyi farklı olan modüller belirlenmektedir. Her aşamada yaklaşık 22 madde yer almaktadır. Her bölümde toplamda yaklaşık 44 madde, tüm testte ise yaklaşık 88 madde bulunmaktadır.

SAT, adaylara online deneme uygulaması seçeneği sunmaktadır. College Board tarafından geliştirilen SAT deneme uygulaması için Bluebook sisteminin öncelikle bilgisayar ortamına kurulması gerekmektedir (College Board, 2025b). Bluebook test sistemde ek süre, metin sentezleyici, ekstra mola gibi test düzenlemesi seçenekleri de sunulmaktadır. Bu özellikler, farklı ihtiyaçlara sahip adayların sınav deneyimini daha erişilebilir ve adil hale getirmeyi amaçlamaktadır. Şekil 9'da SAT deneme uygulamasına ait bir ekran görüntüsü ile birlikte, arayüzün temel bileşenlerine ilişkin tanıtıcı bilgiler sunulmuştur.

Şekil 9

SAT Deneme Uygulaması Tanıtımı



Şekil 9'a göre SAT'ın dijital test arayüzü, kullanıcıya yönlendirme, izleme ve etkileşim açısından zengin bir deneyim sunacak şekilde tasarlanmıştır. Testin başlangıcında ilgili modülün yönergesi ekranda belirir ve test süresince erişilebilir kalır. Katılımcı, modülde yer alan toplam soru sayısını ve yanıtladığı soru miktarını anlık olarak takip edebilmektedir. Süre göstergesi, kullanıcı tercihine bağlı olarak gizlenebilir; bu özellik, zaman baskısını azaltmaya yönelik bir esneklik sunar. Sorular, "daha sonra bakılacak"

şeklinde işaretlenerek modül sonunda tekrar gözden geçirilmek üzere hatırlatılabilir. Seçeneklerin yanında yer alan butonlar aracılığıyla katılımcı seçenek eleme işlemi gerçekleştirebilir ki bu karar verme sürecini kolaylaştıran bir işlevdir. Ayrıca metin sentezleyici özelliği sayesinde metinler baştan sona ya da üzerine gelindiğinde sesli olarak okunabilir. Bu özellikler, özellikle erişilebilirlik ve kullanıcı odaklılık açısından çağdaş dijital ölçme araçlarının gelişmişliğini yansıtmaktadır.

3. OTOMATİK MADDE ÜRETİMİ

Eğitim sürecinde öğrencilerin başarı, yetenek, kişilik, ilgi ve diğer özellikleri belirlenerek bireysel, akademik ve mesleki gelişimine yönelik kararlar verilmektedir (Tekin, 2019). Bu kararların doğru olabilmesi için AERA (Amerika Eğitim Araştırmaları Derneği - American Educational Research Association), APA (Amerika Psikoloji Derneği – American Psychological Association) ve NCME (Eğitimde Ölçme Ulusal Kurulu - National Council on Measurement in Education) (2014) tarafından yayınlanan 2014 APA Standartları'nda belirtildiği üzere kullanılan ölçme aracının ve bu ölçme aracından elde edilen sonuçların "geçerlik, güvenirlik ve adillik (fairness)" kanıtlarına sahip olması gerekir. Bu nedenle test geliştirme sürecinde ölçülecek özellikler doğrultusunda madde türlerine karar verilip nitelikli maddelerin geliştirilmesi beklenir. Maddeler geliştirildikten sonra uzman görüşüne dayalı olarak maddelerin konuyu uygun şekilde yansıtıp yansıtmadığının değerlendirdiği (kapsam geçerliği kanıtları) ve maddelerin başlangıçta belirlenen kuramsal çerçeve ve istenilen bilgi veya beceriye ne ölçüde uyumlu olduğu (yapı geçerlik kanıtı) incelenir. İnceleme sonucunda maddelerde gerekli düzenlemeler uzman görüşlerine ve ön uygulama sonuçlarına dayalı olarak gerçekleştirilir (Jaroslawska ve ark., 2020). Madde geliştirme ve düzenleme sürecinde ayrıca test maddelerinin belli bir grup için avantaj ya da dezavantaj oluşturmamasına, kültürel olarak uygunluğuna dikkat edilir (Campoy ve ark., 2015). Daha sonra ön uygulama sonuçları kullanılarak maddeler üzerinde gerekli düzenlemeler yapılır ve bir madde havuzu oluşturulur. Testin kalite standardını korumak için madde havuzunun sürekli incelenmesi ve güncellenmesi gerekir. Tüm bu süreçler uzun bir zaman ve yoğun bir çaba gerektirir. Ayrıca madde yazımı ve düzenleme süreci titizlikle yürütülse dahi kimi zaman hatalı maddeler geliştirilmiş olabilir. Madde yazım süreçleri için detaylı yönergeler yayınlanmış olsa da (Haladyna & Rodriguez, 2013) aynı öğrenme çıktısına yönelik farklı uzmanlar tarafından geliştirilen maddelerden birinin psikometrik özellikleri çok iyi iken bazılarının daha düşük nitelikte olabilir. Bu durumda madde geliştirme sürecindeki işlemler yenilenir. Geleneksel yöntemde maddelerin tek tek geliştirilmesine dayalı bu süreç için günümüzde otomatik madde üretimi sistemi kullanılmaktadır.

Otomatik Madde Üretimi (OMÜ), test maddelerinin bilgisayar teknolojisi yardımıyla hızlı ve etkili bir şekilde üretilmesini sağlayan yenilikçi bir yöntemdir. Bu yöntem, test maddelerinin uzmanlar tarafından tek tek yazılması yerine önceden tanımlanmış üretim

çerçeveleri (modeller, istemler vb.) ile çok sayıda madde üretilmesine olanak tanır (Gierl ve ark., 2021; Sommer & Arendasy, 2025; von Davier, 2018). Böylece öğretmenlerin ya da test geliştiricilerin sınırlı sürede oluşturabileceklerinden çok daha fazla sayıda madde kısa sürede üretilebilir. Ayrıca maddeler belirli standartlara uygun biçimde üretildiği için psikometrik özellikleri yüksek düzeyde olacak şekilde yapılandırılır (Choi & Zhang, 2019). Özellikle doğal dil işleme (NLP) teknolojisi ile üretilen maddelerin dil bilgisel doğruluğu, anlam bütünlüğü ve içerik geçerliği daha yüksek düzeyde sağlanabilmektedir. Gorgun ve Bulut (2024) çalışmasında NLP ile otomatik madde üretiminde daha kaliteli maddeler elde edildiğini ifade etmektedir. İşlem süreci kullanılacak OMÜ yaklaşımına göre farklılık göstermekle birlikte OMÜ süreci genel olarak iki temel aşamadan oluşur: Üretim çerçevesi ve madde üretimi. Madde üretim çerçevesi aşamasında ölçülmek istenen öğrenme çıktılarına uygun biçimde madde şablonları, prototipler, algoritmalar veya eğitilmiş modeller hazırlanır. Daha sonra bilgisayar algoritmaları bu çerçeveleri kullanarak yüzlerce yeni madde üretir. Bu maddeler, farklı güçlük düzeylerine uyarlanabilir, farklı öğrenci gruplarına göre çeşitlendirilebilir ve farklı bağlamlarda kullanılabilir (Sun ve ark., 2019).

OMÜ'nün ilk sistematikleşme girişimleri Guttman'ın (1957) yüzey teorisiyle başlamış, ardından Bormuth'un (1970) madde yazımında öznelliği azaltmayı hedefleyen kural tabanlı yaklaşımıyla kuramsal temelleri atılmıştır. Bormuth'un "wh-dönüşümü" metinleri madde cümlelerine dönüştürmeye dayanırken Finn (1975) ile Roid ve Haladyna'nın (1978) çalışmaları bu yaklaşımı sadeleştirmiştir. 1980'lerden itibaren paralel test ihtiyacı ve test güvenliği, OMÜ'ye olan ilgiyi artırmış, bu süreçte Roid ve Haladyna (1982) "madde kabuğu (item shell)" yaklaşımını geliştirmiştir. Ardından Bejar'ın (2002) "madde modeli" ve Embretson'un (2002) bilişsel karmaşıklığı içeren modeli, farklı zorluk düzeylerinde maddeler üretilmesine olanak sağlayarak OMÜ'nün kapsamını genişletmiştir. Irvine ve Kyllonen'un (2002) çalışmalarıyla OMÜ ile geliştirilen maddeler için kalibrasyon çalışmaları yapılmaya başlanmıştır. Kuramsal temellerine yönelik çalışmalar gerçekleştirilirken 2000'li yıllardan itibaren bilgisayar teknolojilerindeki gelişimler, OMÜ'nün gelişimini hızlandırmıştır. Gierl (2013) ve Lai'nin (2015) çalışmalarıyla OMÜ için bilişsel ve madde modelleri geliştirilmiştir. Doğal dil işleme teknikleri, yapay zekâ ve özellikle büyük dil modelleri ile birlikte OMÜ ile psikometrik açıdan daha farklı nitelikte maddeler üretilmeye başlanmış ve kısa sürede alanda yaygınlık kazanmıştır.

OMÜ'nün gelişiminin temelinde daha önce sözü edildiği gibi geleneksel madde yazım süreçlerinin sınırlılıkları yer almaktadır. Geleneksel yöntemle her madde, uzmanlar tarafından tek tek yazılır, ardından gözden geçirilir, düzeltilir, mümkünse ön uygulama yapılır ve gerekli düzenlemeler sonucunda testlere dâhil edilir. Bu yaklaşım zaman alıcıdır, maliyetlidir ve yoğun çaba gerektirir (Kosh ve ark., 2018). Özellikle geniş ölçekli testler için çok fazla maddeye ihtiyaç duyulduğundan geleneksel yaklaşım hem insan

kaynağı açısından hem de maliyet açısından oldukça maliyetli bir hal alır. Ancak OMÜ ile kısa sürede çok sayıda madde üretilebilir ve üretilen maddeler güncellenebilir. Maddeler genellikle oluşturulduğu dönemin bilgilerini yansıtır ve zamanla geçerliliğini kaybedebilir. Bu maddelerin madde havuzundan çıkarılarak düzenlenmesi, gerekirse yeniden geliştirilmesi gerekir ki bu da hem zaman hem de maliyet açısından büyük zorluklar doğurur. Oysa OMÜ'de üretim çerçevesinde yapılan bir güncelleme ile daha önce üretilen maddeler hızlı şekilde güncellenebilir (Gierl & Lai, 2017). Böylece bilgi güncellendikçe test havuzları kolayca yenilenebilir. Bu esneklik, özellikle hızla değişen bilgi alanlarındaki maddelerin hızlı bir şekilde güncellenmesinde avantaj sağlar. OMÜ'nün gelişimini hızlandıran bir diğer önemli faktör ise bireyselleştirilmiş ölçme ve öğrenme ihtiyacıdır. Öğrencilerin farklı öğrenme hızları, stilleri ve bilgi düzeyleri, geleneksel testlerin sınırlı kalmasına neden olur. Oysa OMÜ ile farklı zorluk düzeylerinde ve farklı bilişsel becerileri ölçen maddeler hızlı bir şekilde üretilebilir. Bu durum, bireyselleştirilmiş test uygulamalarını da kolaylaştırır (Falcão ve ark., 2022). Test geliştirme sürecinde ayrıca öğrencilerin farklı zamanlarda aynı kapsamda bilgi ve becerilerin ölçülmesi istenildiğinde eş değer testlerin üretilmesi gerekir. Bununla birlikte aynı öğrenme çıktısı için farklı uzmanlar tarafından yazılan maddeler, psikometrik özellikler açısından önemli farklılıklar gösterebilir. OMÜ ise standart çerçeveler üzerinden üretim yaptığı için bu öznelliği azaltma ve maddeler arasında daha fazla tutarlılık sağlama potansiyeline sahiptir. Bu sayede birbirinin yerine kullanılabilecek eş değer maddelerin üretilmesine olanak sağlar (Sayın & Gierl, 2024).

3.1. Otomatik Madde Üretimi Yöntemleri

Otomatik madde oluşturma süreci üretim çerçevesinin tasarımı ve bu tasarımlara dayalı olarak maddelerin algoritmik olarak oluşturulması gibi birkaç aşamayı içerir (Sun ve ark., 2019). Bu aşamada üretim çerçevesinin oluşturulmasına ve kullanılan algoritmaların niteliğine bağlı olarak farklı otomatik madde üretimi yaklaşımları benimsenebilmektedir. Literatürde otomatik madde üretim yöntemleri genel olarak model tabanlı yöntemler, yapay zekâ tabanlı yöntemler ve köprü yaklaşımı (hibrit) modeller olmak üzere üç başlık altında sınıflandırılmaktadır (Shin, 2021).

3.1.1. Model Tabanlı Otomatik Madde Üretim Yöntemi

Otomatik madde üretimi fikrinin ortaya çıkışından bu yana en yaygın kullanılan yaklaşımlardan biri model tabanlı OMÜ yöntemi olmuştur. Bu yöntemde madde üretimi, bireysel maddelerin tek tek yazılmasından çok, bu maddeleri doğuracak modellerin geliştirilmesine dayanır. Model tabanlı otomatik madde üretiminin temel amacı, madde üretim sürecini sistematik hâle getirerek üretilen her maddenin güçlü teorik bir temele dayanmasını sağlamaktır. Bu yaklaşımda rastgele veya sezgisel madde üretiminin aksine ölçülmek istenen özelliğin net tanımından başlanarak sistematik bir süreç izlenerek maddeler üretilir.

3.1.1.1. Model Tabanlı Otomatik Madde Üretiminin Aşamaları

Model tabanlı otomatik madde üretiminde (OMÜ) sürece genellikle geçerlik ve güvenirlik kanıtları elde edilmiş, hedef öğrenme çıktısını yüksek nitelikte ölçen bir temel madde (parent item) ile başlanır. Temel maddenin özelliklerinden hareketle üretilecek maddelerin iskeleti oluşturularak madde bileşenleri belirlenir. Bu temel bileşenler ve özellikler tanımlandıktan sonra maddelerin ölçmek istediği bilgi ve beceriler için yeni maddeler üretmek için değiştirilebilir özellikler (features) sistematik biçimde manipüle edilir. Bu yaklaşım sayesinde aynı amaç (özellik, güçlük vb.) için çok fazla sayıda yeni madde üretilebilir (Bejar, 2002; Singley & Bennett, 2002;).

Model tabanlı otomatik madde üretiminin ilk dönemlerinde madde kabuğu adı verilen bir yaklaşımla madde içinde değiştirilebilecek özellikler belirlenmiştir. Örneğin Singley ve Bennett (2002) madde kökündeki değiştirilebilir değişkenleri ("cisim türü", "bina yüksekliği", "gezegen") tanımlayarak her bir değişken için olası değerleri belirlemiştir. Bu değerler sistematik şekilde bir araya getirilmiş ve temel maddenin birçok farklı versiyonu üretilmiştir. Böylece aynı özelliği ölçmek için farklı maddeler tek tek yazmaya gerek kalmadan hızlı bir şekilde üretilebilmiştir. Bununla birlikte model tabanlı ilk dönem üretimlerde maddelerin klonlama şeklinde birbirinin benzeri niteliğinde olduğu görülmüştür. Bu erken dönem uygulamaların sınırlılıklarını aşmak için Gierl ve Lai (2013) tarafından OMÜ için daha gelişmiş ve sistematik bir çerçeve önerilmiştir. Bu çerçeve üç aşamadan oluşmaktadır: *(i)* bilişsel modelin geliştirilmesi, *(ii)* madde modelinin oluşturulması ve *(iii)* bilgisayar algoritması ile maddelerin üretilmesi. Üç aşamalı sistematik yapılandırma süreci, her aşamasında kalite kontrolünün sağlanmasına ve nihai ürünlerin eğitim standartlarına uygun olmasına olanak tanımaktadır.

Model tabanlı OMÜ yönteminin ilk aşaması olan bilişsel model geliştirme, yöntemin başlangıcını ve teorik temelini oluşturur. Bu aşamada ölçülmek istenen özellik detaylı olarak analiz edilir ve maddenin çözümü için gerekli bilgi, beceriler ve içerik; bilişsel model adı verilen bir şekilde modellenir. Bu sayede bilişsel yapı ile uygulamalı yapı bir araya getirilir. Ayrıca bilişsel modelde her maddenin ölçmeyi amaçladığı özelliğin açık ve detaylı tanımı yapılır. Bu sayede öğrencilerin hangi becerilere sahip olduğu veya hangi bilişsel süreçlerde eksiklik gösterdiği sistematik olarak belirlenebilir. Böylece bilişsel model sadece madde üretimi için değil aynı zamanda üretilen maddelerin kalitesinin değerlendirilmesine ve etkili geri bildirimler için gerekçeler oluşturulmasına kaynaklık eder (Sayın & Gierl, 2023a). Model tabanlı yaklaşımda ölçülecek özelliğin niteliğine göre farklı bilişsel model türleri geliştirilmiştir. Bu çeşitlilik, yaklaşımın farklı konu alanları ve madde türleri için uyarlanabilirliğini göstermektedir. Anahtar özellikler bilişsel modeli, özellikler arasındaki sistematik kombinasyonlara dayanır. Bu model türü özellikle sağlık bilimleri, sosyal bilimler ve dil eğitimi gibi alanlarda kullanılmaktadır. Bir maddenin çözümü için gerekli bilgi, beceri ve içerikler; belirli özellikler çerçevesinde ele

alınabiliyorsa anahtar özellikler bilişsel modeli kullanılmaktadır. Bir diğer bilişsel model olan mantıksal yapılar bilişsel modeline ise genellikle fen ve matematik alanlarında başvurulmaktadır. Bu modelin ayırt edici özelliği, içeriğin değişebilmesi ancak bu içeriği işlemek için gereken formül, algoritma veya mantıksal sürecin sabit kalması, başka bir anlatımla maddenin mantıksal yapısına göre maddelerin üretilmesidir.

Model tabanlı OMÜ'nün ikinci aşamasında bilişsel modelde tanımlanan özellikler somut bir madde formatına dönüştürülür. Madde modeli, bilişsel modeldeki tüm özelliklerin madde formatlarında nasıl temsil edileceğini belirler. Bu aşamada maddenin hangi bölümlerinin değiştirileceği, hangi unsurların sabit kalacağı ve nasıl bir manipülasyon stratejisi izleneceği belirlenir. Bu aşamada kullanılan madde modelleri tek katmanlı ve çok katmanlı modeller olarak geliştirilmiştir. Tek katmanlı madde modelleri, basit manipülasyonlar içeren ve görece az sayıda madde üretimine olanak tanıyan yapıları ifade etmektedir. Bu modeller, bilişsel modeldeki temel değişkenleri manipüle ederek madde üretimi sağlar ama üretilen maddelerin benzerlik oranı yüksektir. Çok katmanlı modeller, dilin sözdizimsel imkanlarını kullanarak hiyerarşik yapılandırma sağlar. Bu yaklaşımda madde gövdesi, madde cümlesi ve seçenekler iç içe geçmiş yapılar hâlinde organize edilir. Bu sayede aynı bilişsel modelden çok daha fazla sayıda ve birbirinden daha farklı maddeler üretilebilir. Çok katmanlı modellerin geliştirilmesi, tek katmanlı modellere göre teknik açıdan daha karmaşık olsa da sonuçta elde edilen madde çeşitliliği ve kalitesi açısından tek katmanlı modellere göre önemli avantajlar sağlar.

Model tabanlı OMÜ'nün üçüncü ve son aşamasında bilgisayar teknolojisi kullanılarak otomatik üretim gerçekleştirilir. Bu aşamada bilişsel model ve madde modelinde tanımlanan kural ve kısıtlamalar doğrultusunda sistematik olarak çok sayıda madde üretilir. Model tabanlı otomatik madde üretimi için çeşitli yazılım araçları geliştirilmiştir. IGOR (Gierl ve ark., 2018), Java SE 6.0 tabanlı kapsamlı bir sistem olarak model oluşturma, kanıt toplama, madde üretimi ve madde havuzu oluşturma süreçlerini desteklemektedir. Bu yazılım, kalibrasyon çalışmalarının da gerçekleştirilebilmesine olanak tanımaktadır. Math Test Creation Assistant (Singley & Bennett, 2002), ModelCreator (Higgins ve ark., 2005), Item Distiller (Higgins, 2007) gibi özel amaçlı yazılımlar, farklı konu alanları için özelleştirilmiş çözümler sunmaktadır. R programlama dilinde QAIG ve IMak gibi paketler de bulunmaktadır. Ayrıca Python tabanlı uygulamalarla da otomatik üretimi hızlı bir şekilde gerçekleştirilebilmektedir.

3.1.1.2. Model Tabanlı Otomatik Madde Üretimi Uygulama Örnekleri

Model tabanlı OMÜ yaklaşımıyla İngilizce başta olmak üzere farklı dillerde ve madde türlerinde yapılan birçok çalışma örneği bulunmaktadır. Türkçede de OMÜ için model tabanlı yöntemin kullanıldığı çalışmalar olduğu görülmektedir. Örneğin Kıyak (2023) Türkçede klinik muhakeme becerilerini değerlendiren vaka temelli çoktan seçmeli maddeler oluşturmak için model tabanlı yaklaşımı benimsemişlerdir. Çalışmada önce

bilişsel model, ardından madde modeli uzmanlar tarafından geliştirilerek 1.73 saniyede hipertansiyon konusuna yönelik 1600 çoktan seçmeli madde üretilmiştir. Uzman görüşüne dayalı olarak üretilen maddelerin kabul edilebilir niteliklere sahip olduğu belirlenmiştir. Sayın (2024) tarafından model tabanlı OMÜ kullanılarak görsel içerikli okuduğunu anlama maddeleri geliştirilmiştir. Mantıksal bilişsel modeli ve çok katmanlı madde modeli geliştirilerek 325 madde üretilmiştir. Maddelerin psikometrik özellikleri ön uygulama çalışması ile değerlendirilmiş; ön denemesi yapılan beş maddenin güçlük indekslerinin 0.58 ile 0.61; ayırt edicilik indeksinin ise 0.48 ile 0.69 arasında değiştiği tespit edilmiştir. Bozdağ (2024) tarafından model tabanlı OMÜ yöntemi ile sözel olmayan akıl yürütme maddeleri otomatik olarak üretilmiştir. Bilim ve Sanat Merkezi (BİLSEM) giriş sınavlarında kullanılmak üzere Cattell-Horn-Carroll zekâ modeli temelinde öğrencilerin görsel bilgileri anlama, analiz etme ve yeni duruma uygulama becerilerini ölçmek için öncelikle mantıksal yapılar bilişsel modeli geliştirilmiştir. Model görsel akıl yürütme alanını döndürme, perspektif, sıralama, eşleştirme, birleştirme ve şifre çözme alt boyutlarına ayırmış ve örnek olarak döndürme işlemleri ele alınmıştır. Ardından çoktan seçmeli madde formatı için madde modeli geliştirilmiş ve çalışma sonucunda her bir madde modeli için 1000 madde üretilmiştir. Üretilen maddeler geliştirilen testlere yerleştirilmiş ve otomatik üretilen maddelerin normal gelişim gösteren ve özel yetenekli öğrencileri başarılı şekilde ayırt edebildiği saptanmıştır. Model tabanlı OMÜ'nün kullanıldığı bir diğer çalışmada Emekli ve arkadaşları (2025), psikiyatri eğitiminde klinik akıl yürütmeyi değerlendirmek için otomatik madde üretimi gerçekleştirmiştir. Çalışmada ilk olarak belirlenen bilişsel modeller, ardından çoktan seçmeli maddeler için çok katmanlı madde modeli geliştirilmiştir. Bilişsel model ile madde modeli Python tabanlı bir yazılımda bir araya getirilerek toplam 1189 madde üretimi gerçekleştirilmiştir. Ardından kuramsal olarak belirlenen tanıları örneklendiren maddeler seçilerek uzman görüşü alınmıştır. Uzmanlar üretilen maddelerin psikometrik özelliklerinin yeterli olduğunu ifade etmiştir.

Model tabanlı otomatik madde üretimi yönteminin aşamaları Sayın ve Gierl (2023b) tarafından gerçekleştirilen çalışma üzerinden örneklendirilmiştir. Model tabanlı otomatik madde üretimi yaklaşımının Türk dili ve edebiyatı alanında nasıl uygulanabileceğini göstermek amacıyla tasarlanan çalışmada 2018 yılında üniversiteye giriş sınavında yer alan bir Türk dili ve edebiyatı maddesi temel madde olarak ele alınmıştır (bkz. Şekil 10).

Şekil 10

Model Tabanlı OMÜ İçin Temel Madde (parent item)

12. XVIII. yüzyılda yaşamış Klasik Türk edebiyatı şairidir. Sebk-i Hindî'nin önemli temsilcileri arasında gösterilir. Genç yaşta divan sahibi olmuştur. Divan şiiri geleneğinin son büyük temsilcilerinden biri olarak kabul edilir. Mesnevi yazma gücünü alegorik bir eser olan Hüsn ü Aşk adlı eseriyle ispat etmiştir.

Bu parçada söz edilen şair aşağıdakilerden hangisidir? (2018, AYT)

A) Nâbî B) Nedîm C) Şeyh Gâlip* D) Şeyhülislam Yahyâ E) Hayâlî

Daha önce de söz edildiği gibi model tabanlı OMÜ için psikometrik özelliklerinin iyi olduğu bilenen bir madde, temel madde olarak ele alınmaktadır. Daha sonra maddenin çözülebilmesi için gerekli bilgi, beceri ve içerikler belirlenmektedir. Sayın ve Gierl'in (2023b) araştırmasında da Şekil 10'daki temel maddenin açıklamasında yer alan şairin yaşadığı yüzyıl, edebî akımın temsilcilerinden biri olma, genç yaşta divan sahibi olma, eser ile eserin özellikleri vb. bilgi kaynakları kullanılarak şairler belirlenmiştir. Bu doğrultuda birinci aşama olan bilişsel model oluşturulmaya başlanmış, temel madde özellikleri doğrultusunda anahtar özellikler bilişsel model seçilmiştir. Bunun için bir öğrencinin divan edebiyatı şairlerini tespit edebilmesi için gereken bilgi kaynakları (yaşadığı dönem, edebî anlayışı, çalışmaları ve edebiyattaki önemi) belirlenmiştir. Ardından her bir bilgi kaynağı içinde yer alan özellikler ve elementler listelenmiştir. Örneğin araştırmada bir bilgi kaynağı olarak şairlerin edebî anlayışı seçilmiştir. Edebî anlayışı bilgi kaynağı içinde edebî akımı özellik olarak belirlenmiş, edebî akım özelliği içinde de Mahallileşme akımı bir element olarak listelenmiştir. Sonra her bir elementin ait olduğu şair eşleştirmesi gerçekleştirilmiş, gerekli ise sınırlandırmalar kullanılmıştır. Örneğin Nef'î, Sebk-i Hindî akımının bir temsilcisi iken Nâilî Mahallileşme akımını benimsemiştir. Bu sınırlamalar, madde üretiminde doğru cevabın üretilmesinde kullanılarak hem seçeneklerde birden fazla doğru cevabın olmasını engellemekte hem de madde kökünün ilgili şaire göre oluşturulmasını sağlamaktadır.

Problem ve senaryo, bilgi kaynakları, özellik, element ve sınırlama ilişkilerinin listelendiği model, bilişsel model olarak adlandırılmaktadır. Model uzmanlar tarafından geliştirilmiş olup Türkiye Diyanet Vakfı İslam Ansiklopedisi temel kaynak olarak kullanılmıştır. Bilişsel modelin geliştirilmesinin ardından ikinci aşamada madde modeli geliştirme süreci gerçekleştirilmiştir. Madde modeli, bilişsel modeldeki anahtar özelliklerin maddenin hangi bölümlerine yerleştirileceğini belirleyen şablondur. Bu model; madde gövdesi, madde cümlesi ve seçeneklerde yer alabilecek temel bileşenleri içermektedir. Araştırmacılar hem tek katmanlı hem de çok katmanlı madde modeli geliştirmişlerdir. Tek katmanlı madde modelinde bir metin içinde şairin yaşadığı dönem, eseri vb. özelliklerin manipüle edileceği alanlar belirlenerek maddeler üretilmiştir. Çok

katmanlı madde modelinde ise üretilen maddelerin birbirine çok fazla benzemesini engellemek amacıyla madde kökündeki metin akışı değiştirilmektedir. Örneğin önce şairin yaşadığı dönemden söz edilerek edebî kişiliğine geçilirken başka bir katmanda önce şairin edebî kişiliğinden bahsedilerek eserleriyle devam edilebilmektedir. Dolayısıyla çok katmanlı modelde dönem, üslup ve konu özellikleri ile eser, divan ve edebî özellikler farklı kombinasyonlarda ve hiyerarşik yapılarda sunulmuştur. Bu yaklaşımda aynı bilgi farklı cümle yapıları ve ifade biçimleri kullanılarak maddeler oluşturulabilmektedir. Üçüncü aşamada bilgisayar teknolojisi kullanarak otomatik madde üretimi gerçekleştirilmiştir. Araştırmacılar Python dilinde yazılan betikleri (scriptler) kullanarak bilişsel model içeriğini madde modeline yerleştirmiş ve otomatik madde üretimini gerçekleştirmişlerdir. Bu süreç sonucunda tek katmanlı modelden 320 madde, çok katmanlı modelden ise 2.560 madde otomatik olarak üretilmiştir. Üretilen maddelerin ne düzeyde benzer olduğunu, aynı kelimeleri içerip içermediğini belirlemek için hesaplanan değerler bulunmaktadır. Bunlardan biri de kosinüs benzerlik indeksidir. Sözü edilen araştırmada tek katmanlı modelle üretilen maddelerin benzerlik değerinin 0.66 ile 0.99 arasında, çok katmanlı modelde ise 0.27 ile 0.99 arasında hesaplandığı belirlenmiştir. Ayrıca üretilen maddelere ilişkin Türk dili ve edebiyatı eğitimi alanından dört uzmanın görüşleri alınmıştır. Uzmanlar, otomatik üretilen maddelerden rastgele seçilen 60 maddeyi inceleyerek maddelerde dil doğruluğu açısından hata bulunmadığını, maddeleri cevaplamak için verilen bilgilerin yeterli ve açık olduğunu belirtmişlerdir. Üretilen madde örnekleri Şekil 11'de gösterilmiştir.

Şekil 11

Model Tabanlı Otomatik Madde Üretimi İle Üretilen Örnek Maddeler

138. IV. Mehmet'ten III. Ahmet'e dört padişah döneminde yaşamış Klasik Türk edebiyatı şairidir. Şiirlerinde dinî tasavvufî konuları ele alan şair, şiirlerinde öğüt ifadelerini ustalıkla kullanan önemli şairler arasında gösterilir. Yaşadığı dönemde ve divan edebiyatında hikemî şairi olarak ün yapmıştır. Şairin Türkçe ve Farsça divanı vardır. Şairlik maharetini çocuklar için yazdığı mesnevisi ile göstermiştir. **Bu parçada söz edilen şair aşağıdakilerden hangisidir?** A) Nâilî B) Nâbî* C) Şeyh Galip D) Nef'î E) Neşâtî
302. XVII. yüzyılda yaşamış Klasik Türk edebiyatı şairidir. Şiirlerinde dinî tasavvufî konuları ele alan şair, şiirlerinde rindâne söyleyişleri ustalıkla kullanan önemli şairler arasında gösterilir. Şairin şiirlerini topladığı divanı bulunmaktadır. Pek çok şair tarafından şiirlerine nazireler yazılmıştır. Divan şiiri geleneğinde tasavvufî gazel şairi olarak kabul edilir. Yazarlık gücünü sosyal yaşamdaki gerçekliği konu alan suriyye kasidesi ile ispat etmiştir. **Bu parçada söz edilen şair aşağıdakilerden hangisidir?** A) Nâili* B) Şeyh Galip C) Neşâtî D) Nef'î E) Nâbî
396. Şiirlerinde dinî tasavvufî konuları ele alan şair, yerli ve mahalli söyleyişleri ustalıkla kullanan şairler arasında yer alır. XVII. yüzyılda yaşamış divan edebiyatı şairlerinden biridir. Pek çok şair tarafından şiirlerine nazireler yazılmıştır. Yazmadaki ustalığını çocuklar için yazdığı mesnevisi ile göstermiştir. Şairin Türkçe ve Farsça divanı vardır. Divan edebiyatında hikemî şairi olarak bilinmektedir. **Bu parçada söz edilen şair aşağıdakilerden hangisidir?** A) Şeyh Galip B) Nef'î C) Nâilî D) Nâbî* E) Neşâtî

3.1.1.3. Model Tabanlı Otomatik Madde Üretimi Yönteminin Değerlendirilmesi

Model tabanlı OMÜ, hem ölçeklenebilirlik hem de üretilen maddelerin ayırt ediciliği gibi psikometrik özellikleri açısından avantajlara sahiptir. Çünkü madde üretimi ardındaki bilişsel modele odaklanır. Bu sayede ne ölçtüğü ve nasıl ölçtüğü bilinen maddeler hızlı bir şekilde üretilebilir. Başka bir anlatımla bilişsel modelde tanımlanan özellikler, üretilen tüm maddelerde tutarlı şekilde yer aldığı için madde havuzunun özellikleri bilinebilir. Ayrıca bu yöntemle doğru cevap ve çeldiricilerin mantıksal kurallar aracılığıyla üretilebilmesi, sınava katılanların verdiği cevaplardan kavram yanılgılarının belirlenmesine, bu yolla etkili bir geri bildirim sisteminin kurulmasına olanak tanır. Söz gelimi verilen örnekte farklı dönemlerde yaşamış ancak aynı edebî kişilik özelliklerine

sahip iki şairden doğru ile çeldirici arasındaki fark bilinmektedir. Modellerin geliştirme aşaması görece maliyetli olsa da geliştirilen modeller sürekli kullanım için uygun olduğundan madde başına düşen maliyet önemli ölçüde azalmaktadır (Kosh ve ark., 2018).

Model tabanlı OMÜ'nün önemli sınırlılıkları da bulunmaktadır. En önemli sınırlılık, sürecin başarısının büyük ölçüde uzman deneyimine bağlı olmasıdır. Uzmanlar artık madde yazarı özelliklerinden çıkarak model geliştiren kişilere dönüşmektedir. Bilişsel ve madde modeli geliştirme süreçleri ise hem konu alanı uzmanlığı hem de model bilgisi gerektirmektedir. Bu durum, nitelikli uzman sayısının sınırlı olduğu alanlarda uygulamayı zorlaştırmaktadır (Gierl & Lai, 2013). Ayrıca içerik özelliklerinin açıkça tanımlanamadığı alanlarda (örneğin okuduğunu anlama ve çıkarım gerektiren maddelerde) bilişsel model oluşturmak daha zordur. Çünkü bu tür maddelerde doğru cevabın dayandığı metinsel kanıtlar (ana fikir, tema, duygu vb.) arasında karmaşık anlamsal ilişkiler bulunur ve bunların sistematik olarak modellere dönüştürülmesi uzmanlar için oldukça zordur. Bilişsel ve madde modeli uzmanlar tarafından geliştirildikten sonra bunların bir araya getirilmesi ve üretilen maddelerin yönetimi için bilgisayar teknolojisi bilgisine ihtiyaç duyulur. Bu nedenle de süreçte farklı disiplinden kişilerin bir arada çalışması gerekir. Model tabanlı otomatik madde üretiminin bir diğer dezavantajı üretilen maddelerin aynı bilgi ve beceriye odaklı bir şekilde, aynı içerikte üretilmesidir. Bundan dolayı maddeler içerik ve yapı açısından genel itibariyle birbirine yüksek düzeyde benzerlik göstermektedir. Ayrıca başlangıç maliyeti yüksekliği - modellerin geliştirilme aşaması- özellikle tek seferlik kullanımlar için dezavantaj oluşturmaktadır. Model geliştirme süreci ve uzman eğitimi maliyetleri göz önüne alındığında küçük ölçekli uygulamalar için uygun olmayabilir.

Son yıllarda bilgisayar teknolojilerindeki özellikle doğal dil işleme ile büyük dil modellerindeki gelişimlere bağlı olarak model tabanlı otomatik madde üretimi yönteminde değişiklikler meydana gelmiştir. Büyük dil modelleri, otomatik madde üretim süreçlerinde hem model geliştirme aşamasında uzman desteği sağlama hem de doğrudan madde üretimi gerçekleştirme konusunda yeni olanaklar sunmuş ve yapay zekâ tabanlı otomatik madde üretimi yöntemi geliştirilmiştir.

3.1.2. Yapay Zekâ Tabanlı Otomatik Madde Üretimi

Bilgisayar teknolojisindeki gelişmeler birçok alanda olduğu gibi otomatik madde üretimi alanında da etkili olmuştur. Önceleri model tabanlı olmayan otomatik madde üretimi olarak adlandırılan yöntemde geleneksel model tabanlı yaklaşımdan farklı olarak sözdizimsel (syntax-based) ve anlamsal (semantic-based) özelliklerini kullanarak metinlerden doğrudan madde üretilmesi amaçlanmıştır. Bu süreçte bilgisayar programları, dilin yapısını çözümleyerek bir metinden kavramları, ilişkileri veya kritik noktaları çıkararak bunları madde kökleri ile seçeneklere dönüştürmüştür (Choi ve ark.,

2018). Örneğin Brown ve arkadaşları (2020) tanımlama, eş anlamlı, zıt anlamlı gibi kelime anlamlarına dayalı boşluk doldurma maddeleri üretmiştir. Anlamsal temelli yaklaşımı benimseyen Mazidi (2017) anahtar kelimeler, ana fikir cümleleri, özet bilgileri üzerinden maddeler üretmiştir. İnsan değerlendiriciler bu sistemin önceki semantik yaklaşımlara göre daha kaliteli maddeler ürettiğini belirtmiştir. Bununla birlikte bir dokümandan ya da dil özelliklerinden madde üretilmesine dayanan bu süreçteki ilk örneklerde dil özelliklerinin yüzeysel kullanıldığı ve doğrudan alıntılar yapılarak daha alt bilişsel düzeyde maddeler üretildiği gözlenmiştir. Daha sonra derin öğrenme yöntemlerinin doğal dil işleme (NLP) alanındaki gelişimi daha doğal ve çeşitli maddelerin üretilmesine olanak sağlamıştır. Word2Vec, GloVe gibi kelime gömme (word embeddings) teknikleri sayesinde kelimeler arasındaki anlam ilişkileri daha iyi temsil edilmiştir. RNN (Recurrent Neural Networks-Yinelemeli Sinir Ağı), LSTM (Long Short-Term Memory-Uzun Kısa Süreli Bellek) ve GRU (Gated Recurrent Unit-Kapı Yinelemeli Birimler) gibi mimariler, dizisel verileri işleyerek makine çevirisi, okuduğunu anlama ve otomatik madde üretimi gibi alanlarda büyük ilerlemeler kaydedilmesine yardımcı olmuştur. Örneğin sinir ağı tabanlı yöntemlerin (neural seq2seq modelleri) madde üretiminde kullanılması için büyük miktarda eğitim verisi üzerinden öğretimler gerçekleştirilmiş, yeni metinler veya çeşitli içeriklere sahip maddeler üretilmeye başlanmıştır.

Konuyla ilgili von Davier (2018), derin öğrenmenin otomatik madde üretiminde kullanılabilmesi için ilk adımın kapsamlı bir eğitim verisi seti oluşturmak olduğunu ifade etmektedir. Bu eğitim verileri genellikle madde–cevap çiftlerinden oluşur ve her bir çift, bir metin (kaynak metin), bu metinden seçilmiş bir yanıt ve ona karşılık gelen madde kökünü içermektedir. Böylece model, "hangi tür içeriklerin cevap olarak seçildiğini" ve "bu cevaba uygun maddenin nasıl biçimlendirildiğini" öğrenmeye başlar. Model gördüğü örneklerden genelleme yaptığı için eğitim verisinin çeşitliliği ve niteliği, bu aşamada önemli görülmektedir. Eğitim verisi hazırlandıktan sonra metinler ön işleme sürecinden geçirilir. Bu aşamada metinler küçük parçalara ayrılır (tokenization) ve her parça sayısal temsillere, başka bir anlatımla kelime gömme yöntemine dönüştürülür. Bu yöntem, kelimeler arasındaki semantik benzerlikleri modele aktarmayı sağlar. Verilerin ön işlemesi tamamlandıktan sonra RNN mimarisi sırayla her kelimeyi işlerken aynı zamanda önceki adımlardan gelen bilgiyi belleğinde tutar. Bu "hidden state" adı verilen mekanizma sayesinde model kelimeleri tek tek öğrenmekle kalmaz, bütün cümlenin bağlamını da öğrenir. Böylelikle metin içindeki bir ifadenin diğer kelimelerle olan ilişkisini ve cümlenin genel anlamını dikkate alarak çalışır. Ardından süreç madde üretimi aşamasına geçer. Burada encoder–decoder yapısı kullanılır: Encoder RNN, metni ve seçilmiş cevabı işler, bunlardan bir gizli temsil oluşturur. Decoder RNN ise bu temsil üzerinden doğal bir madde cümlesi üretir (von Davier, 2018).

Otomatik madde üretiminde derin öğrenmeden sonra önceden eğitilmiş dil modelleri (Pre-trained Language Models) kullanılmaya başlanmıştır. 2018'de Google tarafından geliştirilen BERT (Bidirectional Encoder Representations from Transformers-Dönüştürücülerden Çift Yönlü Kodlayıcı Temsilleri), bağlamı iki yönlü (hem önceki hem sonraki kelimelerden) öğrenebilen ilk büyük modellerden birini oluşturmaktadır (Devlin ve ark., 2019). Bu model, okuduğunu anlama ve metin sınıflandırma gibi görevlerde olumlu sonuçlar üretmiştir. Aynı dönemde GPT (Generative Pre-trained Transformer - Üretken Ön-Eğitimli Dönüştürücü) serisi dili anlama ve üretim konusunda başarılı örnekler vermiştir. Günümüzde büyük dil modelleri "Large Language Models-Büyük Dil Modelleri (LLMs)" adı verilen modellerle; büyük ölçekli veriler üzerinde eğitilmiş, yüz milyarlarca parametreye sahip ve çok çeşitli görevlerde kullanılabilen modeller çeşitli dil işleme görevlerinde kullanılmaktadır. Bu görevlerden biri de otomatik madde üretimidir.

3.1.2.1. Yapay Zekâ Tabanlı Otomatik Madde Üretiminin Aşamaları

Yapay zekâ alanındaki gelişmelerle birlikte büyük dil modellerinin göreve uygun şekilde yeniden eğitilebilmesi ve ince ayar (fine-tuned) edilebilmesi sayesinde gelişmiş otomatik metin ve madde üretimleri gerçekleştirilebilmekte ve yaygın olarak kullanılmaktadır (Vu & Van Nguyen, 2022). Yapay zekâ tabanlı otomatik madde üretimi çalışmalarının incelendiği bir araştırmada (Tan ve ark., 2025) sırasıyla T5, BERT ve GPT serisinin otomatik madde üretiminde en sık kullanılan modeller olduğu belirlenmiştir. Modellerin her biri aynı dönüştürücü (transformer) mimariyi paylaşmakla birlikte farklı eğitim stratejileri ve veri kaynakları nedeniyle özgün avantajlar sunmaktadır. Bu modeller genel dil yetkinliklerini koruyarak madde üretimi doğrultusunda yeniden eğitildiklerinde ve ince ayar işlemleri gerçekleştirildiğinde tutarlılık ve ölçeklenebilirliği artmaktadır (Ackerman & Balyan, 2023; Tan ve ark., 2025).

Yapay zekâ tabanlı OMÜ'nün ilk adımını kullanılacak veri setlerinin hazırlanması oluşturmaktadır. Bu veri setleri genellikle ders kitapları, açık erişimli makaleler, çevrim içi öğrenme materyalleri veya mevcut test maddelerinden oluşur. Öncelikle bu metinlerin temizlenmesi, gereksiz ifadelerin çıkarılması, dil bilgisi hataların düzeltilmesi ve maddelerin tutarlı bir biçimde yapılandırılması gerekir. Bu aşamada veriler LLM'lerin anlayabileceği formata dönüştürülür ve modelin anlamlı, işlenebilir ve öğretim amaçlı olarak uygun verilerle eğitilmesi sağlanır (VM ve ark., 2024). Genellikle binlerce örnek madde içeren bu veri setleri farklı zorluk seviyelerini, konu alanlarını ve madde türlerini kapsayacak şekilde düzenlenir. Veri setlerinin düzenlenmesini takiben eğitim süreci başlatılır. Bu aşamada farklı ince ayar yaklaşımlarından amaca uygun olan seçilir. Örneğin denetimli ince ayar (supervised fine-tuning) yaklaşımında model belirli girdi-çıktı çiftleri kullanılarak eğitilir. Bu süreçte öğrenme oranı, belirli hiperparametreler dikkatli bir şekilde ayarlanmaktadır. Parametre verimli ince ayar (parameter-efficient fine-tuning) yöntemleri olan LoRA (Low-Rank Adaptation-Düşük Sıralı Uyarlama) ve

QLoRA'da (Quantized Low-Rank Adaptation-Nicemlenmiş Düşük Sıralı Uyarlama) ise büyük modellerin tamamını yeniden eğitmek yerine sadece küçük katmanlar eklenerek eğitimler gerçekleştirilir. İnce ayar eğitimleri tamamlandıktan sonra model artık otomatik madde üretimi için hazır hâle gelmektedir. Bu aşamada model değerlendirmesi ve test süreci başlatılır. Eğitilmiş model, ayrı bir test veri seti kullanılarak değerlendirilir ve ürettiği maddelerin kalitesi, pedagojik uygunluğu ve teknik doğruluğu analiz edilir. Model değerlendirme sürecinde üretilen maddeler uzman değerlendiriciler tarafından incelenerek madde güçlük düzeyi, ayırt edicilik gücü, çeldirici etkinliği ve kapsam geçerliği açısından değerlendirilir. Bu değerlendirme sonuçlarına göre modelin performansı belirlenir ve gerekli durumlarda ayrıca ince-ayar işlemleri uygulanır. Model hazır olduğunda üretim aşaması başlar. Bu aşamada belirlenen parametreler (konu alanı, zorluk seviyesi, madde sayısı, madde türü) doğrultusunda otomatik madde üretimi gerçekleştirilir. Yapay zekâ tabanlı OMÜ, oldukça uzun ve karmaşık bir süreci ifade etmektedir. Bu nedenle bazı çalışmalarda sürecin ayrıştığı ve önce otomatik metin üretimleri yapıldığı ve bu metinlere dayalı maddeler geliştirildiği görülmektedir. Örneğin Bulut ve Yıldırım-Erbaşlı (2022) otomatik madde üretimi sürecinde GPT-2 ve Google's T5 modellerini kullanarak önce metin üretimi daha sonra madde üretimi gerçekleştirmişlerdir. Benzer şekilde Bezirhan ve von Davier (2023) okuduğunu anlama maddeleri için metin üretimi gerçekleştirmişler ve metinlerin genel olarak tutarlılık, uygunluk ve hedef yaş grubu için okunabilirlik özelliklerinin uygun olduğunu belirlemişlerdir.

LLM tabanlı otomatik madde üretiminde öne çıkan yaklaşımlardan biri de istemler (prompts) aracılığıyla yapılan üretimdir. Bu yaklaşımda model, önceden büyük veri setleri üzerinde eğitilmiş hâliyle kullanılır; ek bir eğitim (training) veya ince ayar (fine-tuning) yapılmadan yalnızca dikkatle tasarlanmış istemler aracılığıyla yönlendirilir. Böylece modelden belirli formatlarda, belirli bilişsel düzeylere uygun veya belirli alanlara özgü test maddeleri üretmesi istenir. Bu yöntem sayesinde hızlı ve çok sayıda madde üretmek mümkündür. Ancak LLM'lerin madde üretim görevi doğrultusunda yeniden eğitilmeden kullanıldığı bu yaklaşımın bazı sınırlılıkları bulunmaktadır. Öncelikle üretilen maddeler ölçeklenebilir değildir, başka bir anlatımla kullanılan istem ile farklı zamanlarda benzer sonuçlar alınamamaktadır. Ayrıca üretilen maddelerin tutarlılığı, ölçme amacıyla uyumluluğu ve kavramsal doğruluğu her zaman garanti edilememektedir. LLM'lerin "halüsinasyon" eğilimi nedeniyle yanlış bilgi üretebilmesi, yanıltıcı veya pedagojik açıdan yetersiz/uygunsuz maddeler üretilebilmektedir. Aynı içeriğe yönelik çok fazla madde üretilmek istendiğinde model zamanla tekrarlayıcı hâle gelmekte ve yaratıcılığını yitirmektedir. Bu durum, belirli bir konu alanında yüzlerce madde üretmek gerektiğinde daha belirgin hale gelmektedir. Sözü edilen durumlarla başa çıkmak için sohbet ekranları yerine LLM'ler belirli programlama dilleri (Python, JavaScript vb.) üzerinden API (Application Programming Interface-Uygulama Programlama Arayüzü) çağrılarıyla kullanılmaktadır. API entegrasyonu sayesinde istem

parametreleri dinamik olarak tanımlanabilmektedir. Tüm bu süreçler dolayısıyla "istem mühendisliği" (prompt engineering) teknikleri geliştirilmeye başlanmıştır. İstem mühendisliği, LLM'den istenen kalitede ve formatta çıktı elde etmek için talimatların sistematik olarak tasarlanması ve optimize edilmesi sürecini ifade etmektedir. Bu süreç, basit bir "madde yaz" isteminin arzu edilen sonucu elde etmede yetersiz kaldığını, istemlerin belirli özellikler bağlamında şekillenmesi gerektiğini ortaya koymaktadır. Etkili istem tasarımı, öncelikle detaylı bağlam bilgilerinin sağlanmasıyla başlamaktadır. LLM'ye sadece konu başlığı vermek yerine hedef kitle (yaş grubu, eğitim seviyesi), bilişsel beceri düzeyi (örneğin Bloom taksonomisi'ne göre), madde formatı (çoktan seçmeli, açık uçlu, eşleştirme), zorluk seviyesi ve pedagojik hedefler gibi parametreler açık bir şekilde tanımlanmaktadır. Ayrıca rol tanımlamaları "Sen çok iyi bir asistansın." ile başlayan örnekle gelişmiş rollerin tanımlanması önemlidir. Bu sayede daha gerçekçi ve amaca uygun içerikler üretilebilmektedir (Zheng ve ark., 2017).

İstem tabanlı üretimlerde veri seti ile LLM eğitimleri gerçekleştirilmese de isteme eklenen örneklerle süreç sistematik olarak yönlendirilmeye çalışılmaktadır. Bu yaklaşımda kullanılan örnek sayısı, üretim kalitesini doğrudan etkilemektedir. Yaklaşımlar, verilen örnek sayısına bağlı olarak sınıflandırılmıştır. Örneğin sıfır-atım istem (zero-shot prompting) yaklaşımında herhangi bir örnek vermeden yalnızca görev tanımı ve bağlam bilgileri kullanılarak üretim sağlandığı için bu yaklaşımın tutarlılık açısından sınırlılıkları bulunmaktadır. Tek-atım istem (one-shot prompting) ise tek bir örnek madde üzerinden modelin istenen formatı kavramasını sağlamakta fakat bu yaklaşım da çeşitlilik açısından yetersiz kalabilmektedir. Araştırmalar, birkaç-atım istem (a few-shot prompting) (2–5 örnek) tekniğinin modelin hem format tutarlılığını hem de içerik çeşitliliğini yakalamasına imkân tanıdığını ve özellikle karmaşık madde türlerinde daha iyi performans sergilediğini göstermektedir. Daha ileri bir aşama olan çok-atım istem (many-shot prompting) ise çok sayıda örneğin verilmesiyle modelin daha derin anlayış geliştirmesini ve çok daha çeşitli üretimler yapmasını mümkün kılmakta, böylece en yüksek kaliteli sonuçları vermektedir (He ve ark., 2025). Ancak bu yaklaşımlar uzun istem gerektirdiği için aynı anda gönderilen parça limitleri ve maliyet açısından sınırlı olabilmektedir. Ayrıca yapılan araştırmalar LLM'lerin muhakeme performanslarının girdi uzunluğu arttıkça belirgin şekilde düştüğünü göstermektedir. Bu durum, istem tabanlı otomatik madde üretiminde optimal bir denge kurulması gerekliliğini ortaya koymaktadır. Etkili sonuçlar elde etmek için yeterli bağlam ve örnek sağlanırken aynı zamanda modelin performansını olumsuz etkileyecek kadar uzun olmayan istem ve örneklerin sunulmasının önem arz etmektedir.

3.1.2.2. Yapay Zekâ Tabanlı Otomatik Madde Üretimi Uygulama Örnekleri

Yapay zekâ tabanlı otomatik madde üretimi alanındaki çalışmaların İngilizcede yoğunlaştığı görülmektedir. Örneğin Chughtai ve arkadaşları (2022), T5 modelini bağlam ve paragrafları kullanarak madde-cevap çiftleri üretmek üzere

özelleştirmişlerdir. Benzer şekilde Agrawal ve Shukla (2023), T5 modelini metin ve cevapları birlikte girdi olarak kullanarak madde üretimi için eğitmişlerdir. Türkçede yapılan çalışmalarda ise üretken yapay zekâ araçlarıyla yürütülen çalışmaların yoğunlukta olduğu görülmektedir. Yeşilyurt (2024) yapay zekâ ile C1 seviyesi Türkçe okuma sınavı oluşturmayı amaçladığı çalışmasında otomatik madde üretimi için ChatGPT sohbet arayüzünü kullanmıştır. Çalışma sonucunda ChatGPT'nin verilen konularla ilgili metinleri ve maddeleri çok kısa sürede hazırladığı ancak maddelerin seviyeye uygun olmadığı, seçeneklerin uzunluk-kısalık, zorluk-kolaylık gibi temel ölçütleri karşılamadığı saptanmıştır. Karakaya ve arkadaşları (2024) üst düzey düşünme becerilerinin ölçülmesi için açık uçlu madde üretimini ChatGPT ve ChatGPT temelli bir Özelleştirilmiş GPT aracılığıyla gerçekleştirmiştir. Her iki araçla da üretilen maddelerin açıklık, anlatım, içerik, yapı ve benzeri durumlar bağlamında geliştirilmeye ihtiyaç duyduğunu; özelleştirilmiş GPT ile üretilen maddelerin ChatGPT ile üretilen maddelerden daha iyi nitelikte olduğunu tespit etmişlerdir.

3.1.2.3. Yapay Zekâ Tabanlı Otomatik Madde Üretimi Yönteminin Değerlendirilmesi

Yapay zekâ tabanlı OMÜ'nün en önemli avantajı, çok kısa sürede büyük miktarda madde üretebilme kapasitesidir. Bu durum, madde yazımının maliyetli ve zaman alıcı olduğu alanlarda araştırmacılar ve öğretmenler için büyük kolaylık sağlamaktadır (Jiao ve ark., 2023). Ayrıca yapay zekâ destekli sistemler farklı zorluk düzeylerinde, çok dilli ve farklı formatlarda maddeler üretebilmektedir. Model tabanlı OMÜ yönteminin aksine bilişsel model ve madde modeli gibi üretim çerçevesi için yoğun bir insan çabası gerektirmemektedir. Benzer şekilde yapay zekâ tabanlı OMÜ ile birbirinden çok farklı içerikte ve zorluk düzeylerinde maddeler üretebilmektedir. Gelişmiş istem mühendisliği veya ek veriyle ince ayar sayesinde düşük, orta ve yüksek zorluk seviyelerine uygun madde bankaları oluşturulabilmektedir (Kohnke ve ark., 2023).

Bununla birlikte yapay zekâ tabanlı OMÜ için yüksek hesaplama kaynağı, uzun eğitim süreleri ve büyük miktarda özelleştirilmiş veri gerekmektedir. Ayrıca model her güncellendiğinde ince-ayar sürecinin yeniden başlatılması gerekir ki bu da zaman ve maliyet açısından sürdürülebilirlik sorunları yaratmaktadır. Ayrıca LLM'lerdeki hızlı değişimler yapılan çalışmaların da hızlı değişmesine neden olmaktadır. Örneğin GPT-3 modelinde metin üretimi için text-davinci modeli bulunurken GPT-4 ile birlikte role-based modele geçilmiştir. LLM'lerin ince-ayar işlemi gerçekleştirilmeden önce istemler aracılığıyla otomatik madde üretimi yaklaşımında ise "en iyi istem" kavramının göreceli olması temel bir sorundur. Aynı görev için farklı modeller farklı istem yapıları gerektirmekte hatta aynı LLM'in farklı versiyonları bile farklı istem stratejileri ile daha iyi sonuçlar verebilmektedir. Benzer şekilde ince-ayar işlemi yapılmadan kullanılan LLM'lerin halüsinasyon eğilimi otomatik madde üretiminde kalite sorunları oluşturmaktadır. Modeller gerçek olmayan bilgiler, yanlış kavramlar veya var olmayan referanslar içeren maddeler üretebilmektedir. Ayrıca aynı istem ile farklı zamanlarda

yapılan üretimler arasında tutarsızlıklar gözlenmekte, bu da sistemin tekrarlanabilirliğini ve üretilen maddelerin ölçeklenebilirliğini olumsuz etkilemektedir. Belirli bir bilişsel model veya ölçme kuramına dayalı sistematik bir süreç işletilmediği için maddeler yüzeysel kalabilmekte, tekrara düşebilmekte veya ölçmek istediği yapıyı yansıtmayabilmektedir. Bu nedenle üretilen maddelere ilişkin detaylı uzman görüşlerinin alınması şarttır. Ancak yapılan araştırmalar üretilen maddelerin geçerlik ve güvenirlik kanıtlarının yeterince toplanmadığını göstermektedir (Tan ve ark., 2025). Sözü edilen sınırlılıkların üstesinden gelinmesi amacıyla model tabanlı ve yapay zekâ tabanlı otomatik madde üretimi yaklaşımlarının güçlü yönlerinin bir araya getirildiği köprü (hibrit) yöntem geliştirilmiştir.

3.1.3. Hibrit (Köprü Yaklaşımı) Tabanlı Otomatik Madde Üretimi

Otomatik madde üretimi yöntemleri incelendiğinde model tabanlı yöntemde insan uzmanlığına çok fazla ihtiyaç duyduğu ancak kontrollü bir üretim yaptığı için nitelikli üretimler yaptığı görülmektedir. Yapay zekâ tabanlı yöntemde ise insan uzmanlığına daha az ihtiyaç duysa da üretilen maddelerde ölçeklenebilirlik sorunu yaşanmaktadır. Bu doğrultuda "artırılmış zekâ (augmented intelligence)" adı verilen, insan ve yapay zekânın güçlü yanlarının bir araya getirildiği bir yaklaşımla otomatik madde üretimi gerçekleştirilmektedir. Bu yaklaşım hibrit yaklaşım ya da köprü yaklaşımı olarak adlandırılmaktadır (Sayin & Gierl, 2024). Hibrit yaklaşımın temel amacı, insan ve yapay zekâ yeteneklerinin birbirini tamamlayıcı özelliklerinden yararlanmaktır. Yapay zekâ sistemleri büyük miktarda veri işleme, hızlı üretim ve tutarlı format sağlama konularında üstünlük sergilerken insan uzmanları pedagojik uygunluk değerlendirmesi, yaratıcılık konularında avantaj sağlamaktadır (Zheng ve ark., 2017). Bu sayede hem hızlı ve ekonomik bir şekilde geniş madde havuzları elde edilebilmekte hem de belirli modellere dayalı üretimler gerçekleştirildiği için üretilen maddelerin niteliği daha yüksek olmaktadır.

3.1.3.1. Hibrit Tabanlı Otomatik Madde Üretiminin Aşamaları

Hibrit tabanlı otomatik madde üretimi; model tabanlı yöntem ile yapay zekâ tabanlı yöntemin birleşmesiyle geliştirilmiştir. Bu yöntem genel olarak üç aşamadan oluşmaktadır: *(i)* model tabanlı yöntemde olduğu gibi bilişsel ve madde modeli ve içeriğinde olması gerekenler uzmanlar tarafından belirlenir, *(ii)* bilişsel ve madde modeli yapay zekâ tarafından geliştirilir, *(iii)* bilgisayar algoritmaları ile modeller bir araya getirilerek madde üretimi tamamlanır (Sayin & Gierl, 2024; Tan ve ark., 2025).

Hibrit tabanlı otomatik madde üretimi yönteminin ilk aşamasında modeller geliştirilir. Öncelikle bir testin hangi bilgi, beceri ve süreçleri ölçmeyi amaçladığının tanımlandığı bilişsel model, ardından da maddenin üretileceği madde türündeki özellikler tanımlanır. Sözü edilen modellere ilişkin detaylar model tabanlı OMÜ kapsamında ele alınmıştır. Bu aşamada geçerlik ve güvenirlik kanıtları elde edilmiş temel bir madde ile sürece

başlanabilir. Burada model tabanlı OMÜ'den farklı olarak uzmanlar modellerde olması gereken değişkenlerin sadece ne olduğuna karar verir. Uzmanların bu aşamadaki katkısı maddenin çözümü için gerekli bilgi, beceri ve içeriklerin listelenerek üretilmesi istenen madde özelliklerinin belirlenmesidir.

Uzmanların belirlediği model çerçevesi doğrultusunda modellerin geliştirilmesi yapay zekâ tarafından gerçekleştirilir. Bu aşamada istem mühendisliği özelliğinden yararlanılarak temel madde bileşenleri, birkaç-atım yaklaşım örnekleri olarak tanımlanabilir. Ayrıca modellerin ve üretilecek maddelerin özelliklerine göre üretim yöntemi (kural tabanlı, zincirli düşünsel tabanlı vb.) seçilerek modeller için gerekli bileşenler üretilerek modeller geliştirilmiş olur. Bu aşamada yapay zekâ tarafından geliştirilen bilişsel ve madde modeli uzmanlar tarafından tekrar incelenerek modellere son şekli verilir. Bu sayede yapay zekâ ile olan üretim bir kara kutu olmaktan çıkarak insan denetiminde sürdürülür.

Modellerin geliştirilmesinin ardından bilişsel modeller bilgisayar algoritmaları ile birlikte madde modellerine yerleştirilerek otomatik madde üretimi gerçekleştirilir. Üretilen maddelerde düzenleme yapılması durumunda modeller güncellenerek üretim hızlıca yenilenebilir. Bu aşamada Python gibi yazılımlar/araçlar kullanılmaktadır.

3.1.3.2. Hibrit Tabanlı Otomatik Madde Üretimi Uygulama Örnekleri

Hibrit yaklaşıma yönelik Türkçede çalışma örnekleri bulunmaktadır. Sayin ve Gierl (2024) tarafından Türkçe okuduğunu anlama maddelerinin geliştirilme sürecinde uzmanlar bilişsel model ve madde modelindeki bilgi kaynakları ve özellikleri belirlemişlerdir. Ardından GPT 3.5 yapay zekâ modeli ile modellerdeki elementler üretilmiştir. Üretimler bir araya getirilerek metinde anlam akışını bozan cümleyi bulmaya yönelik 12.500 madde üretilmiştir. Üretilen maddelerden rastgele seçilen 6'sı için uzman görüşü alınmış, uzmanlar maddelerin kabul edilebilir olduğunu ya da minör revizyon gerektiğini ifade etmiştir. Ardından maddeler altı farklı kitapçığa yerleştirilerek 854 öğrenciye uygulanmıştır. Uygulama sonucunda üretilen maddelerin güçlük düzeylerinin benzer olduğu, ayırt ediciliklerinin 0.30'u aştığı tespit edilmiştir. Benzer şekilde Kıyak ve arkadaşları (2025) tıp eğitimi alanında hibrit otomatik madde üretimi yaklaşımını uygulamışlardır. Araştırmacıların geliştirdiği hibrit yöntem yedi adımdan oluşmaktadır. Bu adımlar sırasıyla ana maddenin sağlanması, manipülasyon için elementlerin tanımlanması, seçeneklerin belirlenmesi ve elementlere değer atanması ile bilişsel modelin üretilmesini içermektedir. Altıncı adımda uzmanın son incelemesi yapılmakta, yedinci adımda ise şablondaki içerik geleneksel yazılım aracılığıyla madde üretimi için kullanılmaktadır. Çalışmada erişkinlerde solunum enfeksiyonları ve pediatrik akut alerjik reaksiyonlar konularında iki tip çoktan seçmeli soru kullanılarak yöntem örneklendirilmiştir. Uygulama sonuçları, yapay zekânın insan uzman kontrolü altında sadece 10 dakika içinde otomatik madde üretimi için madde modelleri

geliştirebileceğini göstermiştir. Modelleri geliştirmede yapay zekâdan yararlanılması süreci daha kolay kılmıştır. Aynı şekilde Sayin ve Gierl (2025) tarafından hibrit yaklaşımla otomatik madde üretiminin gerçekleştirildiği çalışmada temel bir okuduğunu anlama maddesi için sadece problem kaynakları belirlenmiş, diğer tüm aşamalar yapay zekâ tarafından üretilerek modeller geliştirilmiştir. 1212 öğrenci ile gerçekleştirilen saha çalışması sonucunda üretilen maddelerin benzer güçlükte ve yüksek ayırt edicilikte olduğu saptanmıştır.

3.1.3.3. Hibrit Tabanlı Otomatik Madde Üretimi Yönteminin Değerlendirilmesi

Hibrit yaklaşımın en önemli avantajı, model tabanlı ve yapay zekâ tabanlı yöntemlerin güçlü yanlarını bir araya getirmesidir. Model tabanlı yöntemde olduğu gibi bilişsel ve madde modeli insan uzmanları tarafından çerçevelendiği için geçerlik ve güvenirlik temelli bir yapı korunmaktadır. Aynı zamanda yapay zekânın büyük miktarda veri işleme ve hızlı üretim kapasitesi sayesinde ölçeklenebilirlik sağlanmaktadır. Böylece hem kısa sürede binlerce madde üretilebilmekte hem de maddeler belirli bilişsel süreçlere dayandırıldığı için pedagojik uygunluk güçlendirilmektedir. Ayrıca yapay zekâ tarafından önerilen madde bileşenlerinin uzmanlarca revize edilmesi, süreci bir "kara kutu" olmaktan çıkararak insan denetimiyle şeffaf ve kontrol edilebilir hale getirmektedir. Farklı disiplinlerde yapılan uygulamalar (ör. okuduğunu anlama, tıp eğitimi) hibrit yaklaşımın alanlar arası esnekliğe sahip olduğunu göstermektedir (Kıyak ve ark., 2025; Sayin & Gierl, 2025; Tan ve ark., 2025).

Bununla birlikte hibrit yaklaşımın da bazı sınırlılıkları vardır. Öncelikle süreç, hem uzmanların katkısını hem de yapay zekâ desteğini gerektirdiğinden çok aşamalı ve karmaşıktır. Uzmanların bilişsel ve madde modellerini oluşturma sürecinde yüksek düzeyde alan bilgisi gereklidir, bu da insan kaynağı ihtiyacını tamamen ortadan kaldırmaz. Dahası yapay zekâ modelleri sürekli güncellendiğinden, kullanılan istem stratejileri zamanla geçersiz hale gelebilir. Sürecin sonunda üretilen maddelerin kalitesi büyük ölçüde uzmanların yaptığı incelemelere bağlıdır. Ayrıca hibrit yöntemin teknik boyutları (örneğin API entegrasyonları, python tabanlı yazılımlar) ek teknolojik altyapı ve beceri gerektirmektedir.

3.1.4. Otomatik Madde Üretiminin Gelecek Yönelimleri ve Etik

Otomatik madde üretimi, eğitimde ölçme ve değerlendirme süreçlerinde kullanılan yenilikçi teknolojilerden biridir. Gelecekte çok daha hızlı bir gelişim göstermesi beklenen bu alan, tartışılması gereken bazı etik sorunları beraberinde getirmektedir.

3.1.4.1. Otomatik Madde Üretiminin Gelecek Yönelimleri

İlk olarak model tabanlı modellerle başlayan bu süreç, günümüzde yapay zekâ teknolojisiyle farklı bir boyut kazanmıştır. Otomatik madde üretimi sayesinde ihtiyaca uygun kaliteli maddeler hızlı bir şekilde üretilerek özellikle öğretmenlerin iş yükü

azaltılabilmekte, öğrencilere bireyselleştirilmiş test uygulamaları için geniş madde havuzları oluşturulabilmekte, eş değer maddelerle farklı zamanlarda uygulamalar yapılabilmektedir. Mevcut büyük dil modellerinin sürekli olarak iyileştirilmesi, otomatik madde üretimi için yeni olanaklar sunmaktadır. İnce-ayar işlemi sonucunda istenilen sonuca uygun dil modellerinin ilerleyen süreçlerde amaca uygun özelleştirilmiş versiyonları yayınlanabilir. Ayrıca modellerin sürekli öğrenme yetenekleri gittikçe gelişmektedir. Böylece modellerin daha az halüsinasyon görme, daha fazla gerçekçi sonuç üretme potansiyelleri artış göstermektedir.

Öğrenciden beklenen özelliklerin günümüzde değişiklik göstermesiyle birlikte test maddelerinde de değişiklikler olacaktır. Yapay zekâ okuryazarlığı gittikçe önem kazanmaktadır. Nitekim PISA 2029 uygulamasında bu alanda maddelerin olacağı ifade edilmiştir. Öğrencilerin yapay zekâ ile etkileşimlerinin ölçüldüğü çok boyutlu madde türlerinin (soru soran, cevabı sorgulayan, görsel, işitsel ve etkileşimli elementleri içeren) kullanımı ve üretimi söz konusu olacaktır. Aynı zamanda bilgisayar ortamında uygulanabilecek sanal ve artırılmış gerçeklik teknolojileriyle kurgulanan madde üretimleri de yakın gelecekte tartışılır olacaktır. Gelecekte otomatik madde üretimi sayesinde her öğrenci için öğrenciye uygun içerikte ve düzeyde madde üretimleri gerçekleştirilerek standart testlerin tek-boyutlu değerlendirme yaklaşımdan farklı olarak çok boyutlu bir yaklaşımla ölçme-değerlendirme süreçlerinin şekillenmesi beklenmektedir.

3.1.4.2. Otomatik Madde Üretimi ve Etik

Yapay zekâ modelleri, eğitildikleri veri setlerinin özelliklerini taşımaktadır. Bu nedenle bazı modeller Türkçe dilinde, eğitildiği İngilizce dili kadar iyi sonuç vermemektedir. Aynı zamanda cinsiyet, etnik köken, kültürel bağlam gibi hassas değişkenler üretilen maddelerde yer alabilir. Benzer şekilde maddeler kültürel bağlama uygun olmayan ve öğrenciler için karmaşık içerikler barındırabilir ve bu durum madde yanlılığına neden olabilir. Ayrıca özellikle üretken yapay zekâ araçları içerikleri kopyalayarak küçük değişikliklerle yeniden ve farklı kullanıcılara çok benzer üretimler sunabilir. Bu da özellikle geniş ölçekli testlerde kullanılan maddeler için telif ve orijinallik sorunu yaratabilir, yanı sıra madde çeşitliliğini ve güvenliğini azaltabilir. Öğrenci verilerinin yapay zekâ tabanlı platformlarda kullanılması, gizlilik ve güvenlik sorunlarını beraberinde getirmektedir. Öğrencilerin kişisel bilgileri, yanıtları ve performans verilerinin yetkisiz erişimlere karşı korunması, OMÜ uygulamalarında etik sorumluluk olarak ön plana çıkmaktadır. OMÜ sistemleri, kısa sürede binlerce madde üretebilse de bu maddelerin pedagojik uygunluğu ve psikometrik özellikleri uygun olmayabilir. Nitekim alanyazınında LLM tabanlı sistemlerin ürettiği maddelerin ölçme amacıyla kullanıma uygun olmadığı, çoğu zaman yoğun revizyon gerektirdiği vurgulanmaktadır. Ne yazık ki bazen yapay zekâ çıktılarına aşırı güven oluştuğu da görülmekte, kullanım sıklığının yarattığı bağımlılık kişilerin yaratıcılık gibi becerilerinin zayıflamasına yol açabilmektedir.

Sözün özü otomatik madde üretiminin eğitimde bireyselleştirilmiş ölçme, çok boyutlu değerlendirme ve yeni nesil teknolojilerle (ör. artırılmış gerçeklik, yapay zekâ okuryazarlığına yönelik maddeler) birleşerek ileride ölçme-değerlendirme süreçlerinde daha kapsamlı bir rol üstleneceği öngörülmektedir. Ancak etik açıdan yanlılık, telif, veri güvenliği ve otomasyona aşırı bağımlılık gibi riskler de bulunmaktadır. Bu nedenle insan–yapay zekâ iş birliğiyle geliştirilen artırılmış zekâ modelleri üzerine odaklanılması önerilmektedir.

4. TEST EŞİTLEME

Eğitimde, psikolojide ve insan kaynaklarında kullanılan testler, bireylerin belirli özelliklerini ölçmek ve bu ölçümler üzerinden kararlar almak amacıyla geliştirilir. Bu nedenle de aynı amacı taşıyan fakat farklı sorulardan oluşan alternatif test formları üretilir çünkü tek bir test formunun sınırsız sayıda kullanılması güvenlik açısından sorun oluşturur. Ancak farklı test formları hazırlanırken benzerlik ve eşdeğerlik korunmaya çalışılsa da bu formlar arasında belirli düzeyde güçlük farkları ortaya çıkabilir. Bu durum, daha zor formu cevaplandıran adayların formun güçlüğünden kaynaklı olarak daha düşük puanlar almasına yol açabilir. Dolayısıyla aynı yetenek düzeyindeki adaylar, farklı formları cevaplandırdıklar için farklı puanlar alabilir ve bu durum da bir adalet sorununa neden olur. Ölçme ve değerlendirme uygulamalarında test eşitleme, bu adalet sorununa çözüm getirilmesi amacıyla geliştirilmiş istatistiksel bir yaklaşımdır (Dorans ve ark., 2010).

Test eşitleme, aynı yapıyı ölçen iki ya da daha fazla test formundan elde edilen puanlar arasında istatistiksel bir ilişki kurma süreci olarak tanımlanabilir. Amaç, farklı formlardan elde edilen puanların birbiri yerine kullanılabilir hâle getirilmesidir. Başka bir anlatımla test eşitleme, bireylerin aldıkları puanlar arasındaki farkların sadece gerçek başarı veya yetenek farklılıklarını yansıtmasını, test formunun kolay ya da zor olmasından kaynaklanmamasını sağlamaktır. Bu süreç, bireyler hakkında doğru ve adil kararlar verilebilmesi için ölçümlerin güvenilirliğini ve geçerliliğini artırılmasına kanıt sağlar (Alordiah & Oji, 2024; von Davier, 2011). Test eşitlemenin temel amacı, test formlarından kaynaklı zorluk farkını düzenlemeyi amaçlamaktadır. Örneğin burs programı için başvuran ve farklı formlardaki testleri alan iki adayın puanları farklı olabilir. Bu sonuç iki durumdan kaynaklıdır: (i) Ya adaylardan biri daha başarılıdır ya da (ii) adayın aldığı test daha kolaydır. Birinci durum ölçme açısından istenilen bir sonucu ifade ederken, ikinci durum test formları arasındaki güçlük farkından kaynaklanan bir adaletsizliğe işaret eder. Bu nedenle test formları başarılı bir şekilde eşitlenerek adaylara ait eşitlenmiş puanlar arasındaki farklılığın yalnızca gerçek yetenek düzeylerinden kaynaklanması, aldıkları test formunun daha kolay veya zor olmasının etkisinin ortadan kaldırılması amaçlanır (Kolen & Brennan, 2014).

Test eşitlemenin amacı, farklı formlardan elde edilen puanların karşılaştırılabilir hâle getirilmesidir. Bu süreci öncelikle günlük yaşamdan bir örnekle ele alalım: İki farklı terazi düşünelim. Terazilerden birinde hasar meydana geldiği için daha fazla ölçüm yapmaktadır. Dolayısıyla da satın alınan ürün aynı ya da aynı ağırlıkta olsa da iki teraziden elde edilen sonuçlar birbirinden farklı olacaktır. Terazilerin hata payları tespit edilip düzeltildiğinde ise ürünün hangi terazide tartılmış olduğuna bakılmaksızın gerçek ağırlığı elde edilebilir. Test eşitlemede de benzer bir süreç uygulanır: Aynı yetenek düzeyindeki bir aday, farklı güçlük düzeylerindeki test formlarında farklı ham puanlar alabilir. Bu doğrultuda adayın A testinden aldığı 80 puanın B testinde hangi puana karşılık geleceği belli formülleri içeren eşitleme süreci ile belirlenir. Ancak buradaki önemli nokta, testlerin aynı amaç için geliştirilmiş formlar olmasıdır. Başka bir anlatımla üniversiteye giriş sınavı testi ile yabancı dil sınavı testi eşitlenmesi beklenmemektedir.

Test eşitleme için aynı içerikte ve benzer istatistiksel özelliklerde geliştirilmiş en az iki farklı test formuna ve bu formlardan alınan puanlara ihtiyaç duyulduğu söylenebilir. Test formlarının başarılı bir şekilde eşitlenmesi sonucunda da farklı formlardaki puanlar birbirinin yerine kullanılabilir hâle gelir. Bu durumda aday hangi test formunu cevaplandırırsa cevaplandırsın yeteneği doğrultusunda bir puan alır (Zhu, 1998). Eşitleme, içerik farklılıklarını değil, güçlük açısından farklılıkları birbirine uydurmayı amaçlayan bir süreci ifade eder. Örneğin, başarılı bir eşitleme sonucunda A testinden 27 puan alan bir adayla B testinden 26 puan alan bir adayın başarı düzeyleri aynı kabul edilir (burada 26, A testindeki 27 puanın B testindeki eşitlenmiş karşılığıdır). Ancak eşitlemenin bazı temel varsayımları ve sınırlılıkları vardır:

- Gruptaki kişilerin yetenek dağılımları birbirinden çok farklıysa,
- Test formları arasında istatistiksel açıdan büyük farklılıklar bulunuyorsa,
- Test formları arasında kapsam bakımından büyük farklılıklar söz konusu ise

eşitleme yapmak ve puanları birbirinin yerine kullanmak doğru bir yaklaşım değildir. İlerleyen bölümlerde de söz edileceği gibi bu durumlarda yapılacak eşitleme hatası yüksek hesaplanacaktır (Kolen & Brennan, 2014).

4.1. Test Eşitlemedeki Temel Kavramlar

Yatay Eşitleme: Bireyler veya gruplar arasındaki karşılaştırma süreci, temel olarak iki faktöre dayanır: Testin güçlük düzeyi ve uygulandığı grubun yetenek dağılımı. Bu iki faktörün durumuna göre eşitleme süreci farklı şekillerde adlandırılır ve farklı amaçlara hizmet eder. Eğer uygulanan test formlarının güçlük düzeyleri birbirine benzer ve bu testleri alan grupların yetenek dağılımları da benzerlik gösteriyorsa bu durum "yatay eşitleme" olarak adlandırılır. Yatay eşitlemede amaç, benzer özelliklere sahip formlar arasındaki küçük güçlük farklılıklarını gidererek puanları birbirinin yerine kullanılabilir

hâle getirmektir. Bu tür eşitleme, genellikle aynı sınav sisteminde farklı dönemlerde uygulanan testlerde, aynı yaş grubundaki öğrencilere yönelik eş değer formlarda veya güvenlik nedeniyle hazırlanıp uygulanan alternatif formlarda kullanılır.

Dikey Ölçekleme: Test formlarının güçlük düzeyleri veya bu testleri alan grupların yetenek dağılımları önemli farklılıklar gösterdiği durumda yapılan dönüşüm, "dikey ölçekleme" olarak adlandırılır. Dikey ölçekleme, yatay eşitlemeden hem amaç hem de uygulama bakımından farklılıklar gösterir. Bu süreçteki amaç belirli bir özellik açısından davranıştaki gelişimin izlenmesidir. Örneğin TIMSS (Trends in International Mathematics and Science Study-Uluslararası Matematik ve Fen Bilimleri Eğilimleri Araştırması) uygulamasında 4. ve 8. sınıf öğrencilerinin matematik becerilerindeki değişimin incelenmesi için dikey ölçekleme çalışması yapılmaktadır. Dikey ölçeklemede yatay eşitlemeden farklı olarak bir testten elde edilen puanlar diğerinin yerine kullanılmaz. Bunun nedeni, bu testlerin farklı güçlük düzeylerinde olması ve farklı yetenek gruplarına uygulanmasıdır. Buradaki amaç, her iki puanın aynı ölçek üzerinde gösterilmesiyle gelişimin incelenmesidir.

Yatay eşitleme ile dikey ölçekleme arasındaki farkı, dil yeterlik testleri üzerinden de açıklayabiliriz. Örneğin aynı amaçla farklı zamanlarda uygulanan dil yeterlik testlerinin farklı formları için yatay eşitleme yapılır. Ancak dil öğrenmeye yeni başlayan ve farklı seviyeler sonrasında o seviyeye ait bitirme testlerini cevaplandıran bir öğrencinin gelişimini izlemek içinse dikey ölçekleme çalışması yapılır. Bu kitap kapsamında yatay eşitlemeden söz edilecek olup test eşitleme kavramı, yatay eşitleme sürecini ifade etmek için kullanılmaktadır. Ayrıca bölümün okuyucular tarafından daha iyi anlaşılması için bölümde geçen bazı kavramların tanımları aşağıda verilmiştir:

Ham puan: Puanlayıcılar tarafından yapılan derecelerin toplamı ya da çoktan seçmeli testlerde genellikle toplam doğru cevap sayısını ifade etmektedir.

Ölçek puanı: Ham puanların yorumlanabilirliğini artırmak için dönüştürülmüş puanlardır. Örneğin PISA (Programme for International Student Assessment-Uluslararası Öğrenci Değerlendirme Programı)

puanları ortalaması 500, standart sapması 100 olan puanlara dönüştürülmektedir.

Eşitlenmiş puan: Farklı test formlarından elde edilen birbirinin yerine kullanılabilecek puanları ifade etmektedir.

Referans Form: Eşitleme sürecinde standart olarak kabul edilen ve diğer formların ona göre eşitlendiği test formudur. Genellikle daha önce uygulanmış, istatistiksel özellikleri bilinen formlardır.

Yeni Form: Referans forma eşitlenmek üzere geliştirilmiş olan test formudur. Bu formun puanları, referans formun ölçeğine dönüştürülür.

Ortak Madde (Çapa madde - ankor madde): Farklı test formları arasında ortak olan ve eşitleme sürecinde formlar arasında bağlantı kurmak için kullanılan madde grubudur.

Tesadüfi eşitleme hatası: Evrenden seçilen örneklemin bir eşitlik ilişkisi kestirmeyi içeren parametrelerin olduğu durumda ortaya çıkan hatadır. Tesadüfi hata, genellikle «eşitlemenin standart hatası» olarak adlandırılır. Eşitleme sürecinden sonra puanların eş değerlerinin (eşitlenmiş puanların dağılımının) standart sapmasıdır.

Sistematik eşitleme hatası: Test eşitleme varsayımların veya eşitleme şartlarının ihlal edilmesi sonucunda ortaya çıkan hatadır.

4.2. Test Eşitleme Koşulları

Test eşitleme sürecinin geçerli ve güvenilir bir şekilde gerçekleşebilmesi için simetriklik, aynı yapıyı ölçme, eşitlik, eşit güvenirlik, gözlenen puan eşitliği ve grup değişmezliği koşullarının sağlanması gerekir (Kolen & Brennan, 2014). Bu varsayımların genel özellikleri şu şekildedir:

4.2.1. Simetriklik

Eşitleme işleminde hangi test formunun referans, hangisinin yeni form olarak seçildiğinden bağımsız olarak sonucu değiştirmemesini ifade eder. Örneğin X formundaki 26 puanın Y formundaki eşitlenmiş puanı 27 ise Y formundaki 27 puanın da X formundaki eşitlenmiş puanının 26 puan olması gerekir. Bu doğrultuda da aslında hangi test formunun diğerine eşitlendiğinin bir önemi yoktur.

4.2.2. Aynı Yapıyı Ölçme

Eşitlenecek test formlarının aynı amaçta ve dolayısıyla aynı içeriğe sahip olması koşulunu ifade eder. Daha önce de söz edildiği gibi içerik ve istatistiksel olarak benzerlik gösteren formlar eşitlenebilir. Ancak bu durumda puanlar birbirinin yerine kullanılabilir.

4.2.3. Eşit Güvenirlik

Eşitlenecek testlerin benzer yapıyı ölçmesinin yanı sıra güvenirliklerinin de eş olması gerektiği koşuludur. Test formları arasında güvenirlik farklılıkları yüksekse formların eşitlenmesi uygun değildir (Dorans & Holland, 2000).

4.2.4. Eşitlik

Lord'un eşitlik varsayımına göre eğer iki test formu eşitlenirse bireylere X ya da Y test formunun uygulanması bir fark oluşturmamalıdır. Katılımcıların X ve Y formundaki gözlenen puan ortalamalarının, standart sapmalarının, X ve Y formundaki dönüştürülmüş puanların dağılım şeklinin aynı olduğu, ölçmenin koşullu standart hatasının iki formda da eşit olduğu anlamına gelir. Örneğin eğer X formu, yüksek puanlarda Y formunu daha kesin ölçüyorsa Lord'un eşitlik özelliği sağlanmamış olur. Lord'un eşitlik özelliği incelendiğinde bu eşitlik özelliğinin sağlanmasının çok zor olduğu

görülür. Bu nedenle Morris (1982) tarafından önerilen «birinci dereceden eşitlik özelliği» sağlanması yeterli görülmektedir. Buna göre belirli bir gerçek puana sahip cevaplayıcıların eşitlenmiş puan ortalamalarının aynı olması beklenir.

4.2.5. Gözlenen Puan Eşitleme Özelliği

Eşitlemenin gözlenen puanlar üzerinden gerçekleştirildiği durumda gerçek puan kestirimleri kullanılmamasını ifade eder. Gözlenen puan eşitleme yöntemleri daha az karmaşık oldukları için araştırmalarda daha sık kullanılmakta ancak her zaman testlerin aynı özelliği ölçmesi gerekliliğini taşımamaktadır.

4.2.6. Grup Değişmezliği

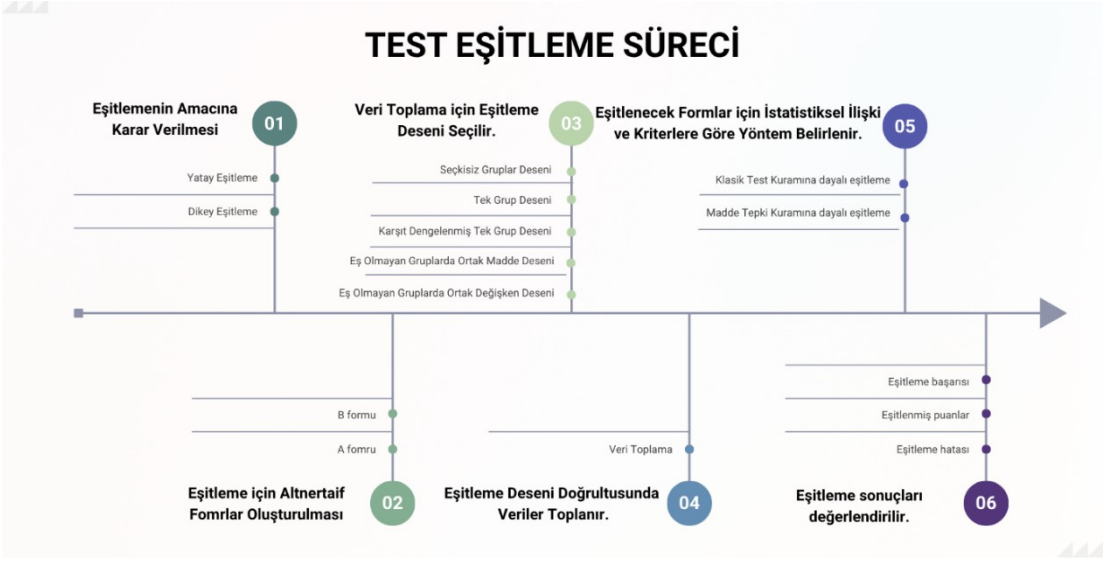
Belirlenen eşitleme ilişkisinin farklı demografik gruplar (cinsiyet, yaş, etnik köken vb.) arasında tutarlı olmasını ve örneklem özelliklerinden etkilenmemesini ifade eder.

4.3. Test Eşitleme İşlem Süreci

Test eşitleme; desenler, tanımlar, istatiksel teknikler ve değerlendirme süreçleri arasından uygun olanları seçilmesi işlemlerini içermektedir ve eşitleme sürecinin başarısı, bu aşamaların her birinin doğru bir şekilde yürütülmesine ve aşamalar arasındaki tutarlılığın sağlanmasına bağlıdır. Bu işlem sıraları bazen uygulamada değişiklik göstermekle birlikte genellikle Şekil 12'de yer alan aşamalardan oluşur ve her aşama bir sonrakinin temelini oluşturur (Kolen & Brennan, 2014; Zhu, 1998).

Şekil 12

Test Eşitleme Aşamaları



Test eşitleme ile farklı test formlarının veya uygulamalarının karşılaştırılabilir hâle gelmesini sağlamak amacıyla istatistiki bir süreç gerçekleştirilir. Bunun için de aynı zamanda ya da farklı zamanlarda uygulanmak üzere test formları geliştirilerek test eşitlemenin birinci ve ikinci aşaması tamamlanmış olur.

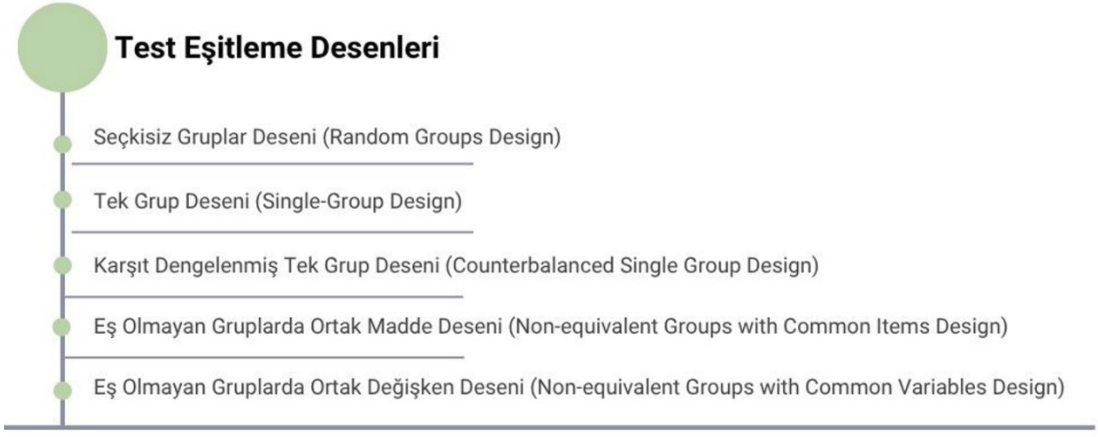
4.4. Test Eşitleme Desenleri

Test eşitleme sürecinde veri toplamak için uygulama koşullarına bağlı olarak veri toplama desenleri kullanılır. Veri toplama desenleri, verilerin toplandığı grubun özelliklerine ve araştırmanın amacına bağlı olarak değişkenlik gösterir. Test eşitleme desenine bağlı olarak eşitleme yöntemleri de farklılık gösterir çünkü her desenin kendine özgü avantajları, sınırlılıkları ve varsayımları bulunur. Test eşitleme desenleri, verilerin nasıl toplanacağını da belirlediği için "veri toplama desenleri" olarak da adlandırılır.

Test eşitleme desenleri, temel olarak iki şekilde sınıflandırılır. Birinci sınıflandırma test eşitleme için gerekli ortaklık üzerine kuruludur. Bu nedenle de ortak/denk bireyleri kullanan desenler (common examinees designs) ile ortak maddeleri kullanan desenler (common items design) olarak ifade edilir. Buradaki her bir desen içindeki uygulamaları da temsil eden daha yaygın bir sınıflama ise Şekil 13'te gösterilmiştir (von Davier, 2011; Kolen & Brennan, 2014; Sansivieri ve ark., 2017).

Şekil 13

Test Eşitleme Desenleri

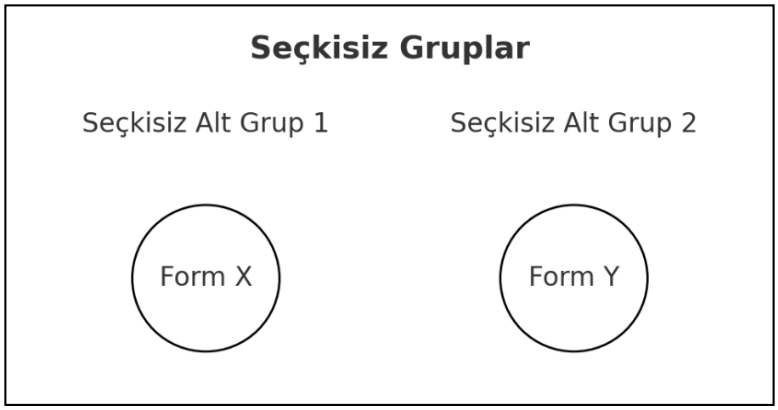


4.4.1. Seçkisiz Gruplar Deseni

Seçkisiz gruplar deseni, test eşitleme alanında en yaygın kullanılan desenlerden biridir ve "eş değer gruplar deseni" olarak da bilinmektedir. Bu desende bireylerin oluşturduğu ana evrenden (P evreni) seçkisiz ve bağımsız olarak iki örneklem belirlenir ve bu örneklemlerden birine X formu, diğerine Y formu uygulanır (Şekil 14).

Şekil 14

Seçkisiz Gruplar Deseni



Dolayısıyla bu yaklaşım, formların uygulandığı grupların eş değer olduğu varsayımına, desenin işleyişi de büyük sayılar yasasına dayanır: Örneklem büyüklüğü yeterli olduğunda ve seçim gerçekten rastgele yapıldığında iki grubun ortalama yetenek düzeylerinin birbirine yakın olması beklenir. Bu varsayım karşılandığında seçkisiz gruplar deseninde iki formdaki grup düzeyindeki performans farklılığı, formlar arasındaki güçlük farkının doğrudan bir göstergesi olarak kabul edilir. Tablo 3'te görüldüğü gibi X formunu sadece birinci grup Y formunu da sadece ikinci grup cevaplandırmaktadır. Burada X formunu alan grubun ortalama ham puanının 72, Y formunu alan grubun ortalama ham puanının da 77 olarak hesaplandığını varsayalım. Grupların rastgele oluşturulduğu ve yeterli büyüklükte olduğu varsayımı karşılandığında sonuçlar, Y formunun X formundan ortalama 5 puan daha kolay olduğunu göstermektedir (González & Wiberg, 2017; Kolen & Brennan, 2014).

Tablo 3

Seçkisiz Grup Deseni Örneği

Gruplar	**Formlar**	
	X formu	**Y formu**
I. grup	72	---
II. grup	---	77

Seçkisiz gruplar deseninin uygulanmasında en yaygın yaklaşım, sarmal süreç yöntemidir. Bu yöntemde aynı test uygulama ortamında katılımcılara sırayla X ve Y formları dağıtılarak uygulama yapılır. Başka bir anlatımla art arda sıralarda oturan ilk katılımcıya X formu, ikincisine Y formu, üçüncüsüne tekrar X formu, dördüncüsüne Y formu şeklinde devam eder. Bu sayede testi cevaplandıran kişiler rastgele iki gruba

ayrılmış olur. Buradaki önemli nokta sarmal süreç yönteminde katılımcıların sıralara seçkisiz şekilde yerleştirilmiş olmasıdır. Aksi takdirde sistematik yanlılıklar ortaya çıkabilir. Örneğin katılımcılar bir kadın bir erkek şeklinde oturmuşsa ve seçkisiz grup deseni için sarmal süreç uygulanırsa X formu sadece belirli bir cinsiyet grubundaki kişilere uygulanmış olur.

Desenin Avantajları ve Sınırlılıkları

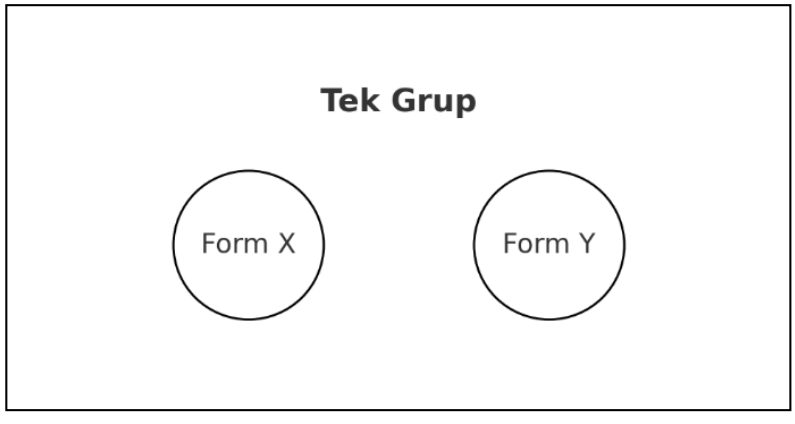
Seçkisiz gruplar deseninin en önemli avantajı, uygulamanın pratikliğidir. Bu yöntemde bir kişi sadece bir formdaki maddeleri cevaplandırır, bu da zamandan tasarruf edilmesini sağlar. Ayrıca desende katılımcılar sadece bir form cevaplandırdığı için sıra etkisi oluşmaz. Seçkisiz grup deseni kullanılarak ikiden fazla form da eşitlenebilir. Ancak form sayısı artıkça grubun örneklem büyüklüğü azalır ve örneklemler evreni yeterince temsil etmezse eşitleme hatası artar. Seçkisiz gruplar deseninde farklı formları cevaplandıracak grupların yeteneklerinin eş olduğu varsayımı bulunur. Bu varsayımın sağlanıp sağlanmadığının belirlenmesi için yeteneklerin değişmezliği incelenebilir, bunun için de büyük örneklemlere ihtiyaç duyulur.

4.4.2. Tek Grup Deseni

Tek grup deseninde aynı gruba hem X hem de Y formu uygulanır. Böylece her bir katılımcı her iki test formunu da cevaplandırmış olur (Şekil 15).

Şekil 15

Tek Grup Deseni



Test formlarına aynı bireyler cevap verdikleri için seçkisiz grup deseninde olan grubun yetenek düzeyi varsayımının incelenmesine gerek kalmamaktadır. Test formlarına aynı bireyler cevap verdikleri için bireylerdeki yetenek değişimi kontrol edilerek test formlarındaki güçlük düzeyindeki değişimler, yetenek düzeyindeki değişimle karışmaz. Başka bir ifadeyle eşitleme sürecindeki "Ortalamalar arasındaki farklılık grup farklılığından mı yoksa form farklılığından mı kaynaklanmaktadır?" sorusunu ortadan kaldırır. Çünkü aynı bireyler her iki formu da cevaplandırdığı için farklılıkların kaynağı

sadece test formlarından gelir. Örneğin (Tablo 4) bir grubun X formundaki ortalaması 72, Y formundaki ortalaması 75 ise ortalamalar arasındaki 3 puanlık farklılığın formlardan kaynaklandığı söylenebilir (von Davier, 2011; Kolen & Brennan, 2014; Woldbeck, 1998).

Tablo 4

Tek Grup Deseni Örneği

Gruplar	**Formlar**	
	X formu	**Y formu**
I. grup	72	75
II. grup	----	---

Desenin Avantajları ve Sınırlılıkları

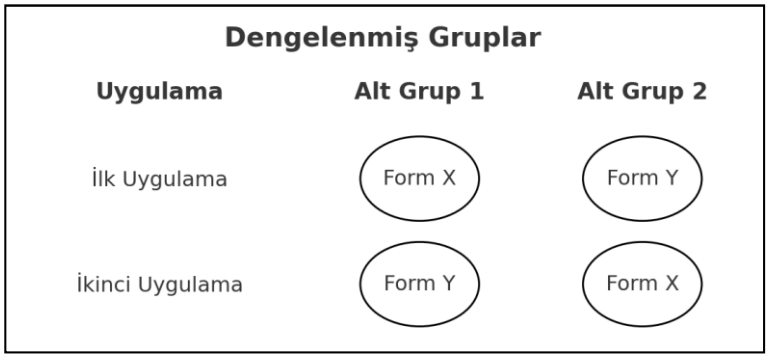
Tek grup deseninde her katılımcı, hem X hem de Y formunu aldığı için bireysel yetenek farklılıkları her iki form için de aynıdır. Ayrıca tek bir grup olduğu için seçkisiz grup desenine göre daha küçük örneklem gerektirir. Bununla birlikte aynı katılımcıya iki test formu uygulanacağı ve bu uygulamaların benzer koşullarda olması gerektiği için kullanılması hem uygulayıcı hem de katılımcılar için ekonomik değildir. Tek grup deseninin en önemli sınırlılığı sıra etkisi oluşmasıdır. Bu desenin temel olarak sıra etkisi olmadığı, katılımcıların ikinci formdaki cevaplarının birinci formdan etkilenmediği varsayımına dayanır. Ancak bu varsayımı karşılamak zordur çünkü katılımcılar benzer içerik ve özellikte hazırlanan iki formu cevaplandırır. Dolayısıyla birinci formdaki cevaplarından etkilenebilir, teste aşinalık oluşabilir. Ayrıca yorgunluk ve isteksizlik gibi durumlar nedeniyle katılımcıların performansları desenden kaynaklı değişebilir. Benzer şekilde iki uygulama koşullarının da benzer olması, arada geçen sürenin de ideal olması gerekir. İki form uygulaması arasındaki sürenin kısa olması; sözü edilen yorgunluk, aşinalık durumlarının oluşmasına, uzun olması da öğrenmeye dayalı gelişimin gerçekleşmesine neden olur. Küçük örneklem gruplarının olduğu durumlarda tek gruplu eşitleme deseni kullanılsa da sıra etkisi başta olmak üzere diğer belirtilen durumlar, parametre tahminlerine dolayısıyla da eşitleme sonuçlarına etki eder.

4.4.3. Dengelenmiş Grup Deseni

Dengelenmiş grup deseninde katılımcılar seçkisiz şekilde ikiye bölünerek alt gruplar oluşturulur. Eşitlenmek istenen testin iki formu, her iki gruba da uygulanır. Bu süreçte birinci grup önce X sonra Y formunu, ikinci grup da önce Y sonra X formunu cevaplandırır (Şekil 16).

Şekil 16

Dengelenmiş Grup Deseni



Dengelenmiş grup deseni; seçkisiz grup deseni ve tek grup deseninin birleşimi gibidir. Önce katılımcılar eş iki gruba ayrılır daha sonra da her iki formu da cevaplandırır. Bununla birlikte tek grup deseninde oluşan sıra etkisi ortadan kaldırılmış olur çünkü uygulamaya gruplar farklı formlarla başlarlar. Uygulama sonucunda da her grubun her form için bir puanı hesaplanır. Tablo 5 incelendiğinde birinci grubun Y formundaki ham puan ortalamasının 75, X formundaki ham puan ortalamasının 72 olduğu görülmektedir. II. grupta ise Y formunun ham puan ortalaması 77, X formunun ham puan ortalaması ise 71 olarak hesaplanmıştır. Hesaplanan değerler Y formunun her iki grupta da X formundan daha yüksek ortalamaya sahip olduğunu göstermektedir. Ayrıca gruplar eş ise X formunu ilk alan grubun ham puan ortalaması 72, ikinci alan grubun 71; Y formunu birinci sırada alan grubun ham puan ortalaması 77, ikinci sırada alan grubun ortalaması 75 olduğu için sıra etkisinden söz edilebilir. Başka bir anlatıma birinci sırada alınan formlarda grubun performansı ikinci sırada aldıkları formlardan daha yüksektir. Ayrıca birinci uygulamada Y formu lehine 5 puanlık, ikinci uygulamada 4 puanlık bir farklılık bulunmaktadır. Bu durum Y formunun X formundan daha kolay olduğu göstermektedir (Kolen & Brennan, 2014).

Tablo 5

Dengelenmiş Grup Deseni Örneği

Uygulama	**Gruplar**	
	I. uygulama	**II. uygulama**
I. grup	**Form X**	**Form Y**
	72	75
II. grup	**Form Y**	**Form X**
	77	71

Desenin Avantajları ve Sınırlılıkları

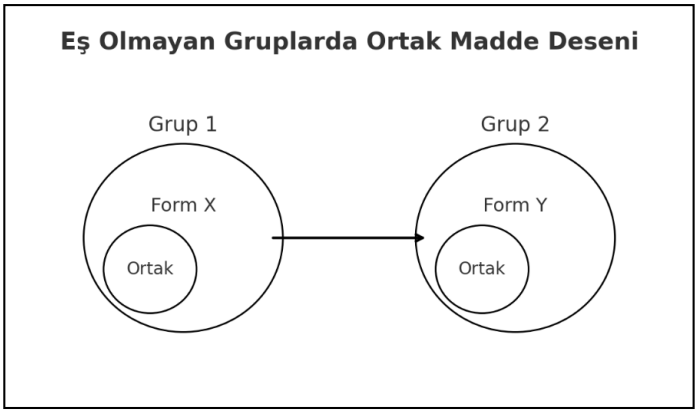
Dengelenmiş grup deseni, tek grup deseninin sıra etkisi sorununa çözüm getirmek için geliştirilmiş bir yaklaşımdır. Bu desende katılımcılar rastgele olarak iki alt gruba bölünür ve katılımcıların yarısı önce X formunu sonra Y formunu alırken diğer yarısı önce Y formunu sonra X formunu alır. Bu yaklaşım sayesinde sıra etkisinin eşitleme sonuçları üzerindeki sistematik etkisi dengelenebilir. Bununla birlikte farklı formların aynı örneklem grubuna uygulanması, her formun iki kez uygulanması nedeniyle desenin kullanışlılığı düşüktür. Tek grup deseninde olduğu gibi katılımcılarda yorgunluk oluşabilir ve iki form uygulaması arasında geçen zamanın iyi ayarlanması gerekir. Bu doğrultuda cevaplayıcılara iki form uygulama işlevsel olarak mümkün olduğunda, değişen sıra etkisinin ortaya çıkması beklenmediğinde, seçkisiz grup deseni kullanan bir eşitleme çalışmasında yeterli cevaplayıcı sayısını elde etmek zor olduğunda dengelenmiş grup deseni kullanılabilir.

4.4.4. Eş Olmayan Gruplarda Ortak Madde Deseni

Test uygulamalarında güvenlik endişeleri ve uygulama zorluğu nedeniyle aynı test oturumunda birden fazla form kullanılması ya da aynı gruba birden fazla form uygulanması çoğu zaman mümkün değildir. Eşitlenecek test formları katılımcılara genellikle farklı dönemlerde uygulanır. Örneğin bir sınav kurumu, aynı amaç için hazırladığı bir testin bahar döneminde X formunu, güz döneminde Y formunu kullanabilir. Bu koşullarda daha önce sözü edilen seçkisiz gruplar, tek grup veya dengelenmiş grup desenleri uygulanamaz. Çünkü bu desenlerin tamamı ya aynı kişilerin birden fazla form almasını ya da formların eş zamanlı uygulanmasını gerektirir. Bu durumda farklı test formları arasında bazı maddelerin ortak olarak kullanılmasına dayanan bir desen geliştirilir. Eş olmayan gruplarda ortak madde deseni olarak adlandırılan bu desende her iki formda da aynı maddeler yer alır (Şekil 17).

Şekil 17

Eş Olmayan Gruplarda Ortak Madde Deseni



Eş olmayan gruplarda ortak madde deseni, test eşitleme sürecinde sıklıkla kullanılmaktadır. Birtakım ortak maddeler içeren X ve Y formları farklı gruplara uygulanır. Her bir grup sadece tek bir formu cevaplandırır. Bu formlar farklı zamanlarda hatta farklı yıllarda bile uygulanabilir ancak bu durumun dikey ölçekleme ile karıştırılmaması gerekir. Çünkü buradaki amaç bireylerin gelişimini görmek değil, farklı yıllarda aynı amaç için geliştirilen aynı içerik ve istatistiksel özelliklere sahip test formu sonuçlarının birbirinin yerine kullanılmasını sağlamaktır. Bu desende katılımcılara eşitlenecek test formlarından sadece biri uygulanır ve iki grup arasındaki eşitleme ilişkisini belirlemek için ortak maddeler üzerinden grupların performansı karşılaştırılır. X ve Y formlarının ortalamaları arasındaki farklılıklar, cevaplayıcı grup farklılığının ve test formu farklılığı karışımının bir sonucudur. Form farklılıkları, grup farklılıklarından ayrılır. Bunun için de ortak maddeler kullanılır. Tablo 6'da X formunun ham puanlarının ortalaması 72, Y puanlarının ham puanlarının ortalaması ise 77 olarak hesaplanmıştır. Seçkisiz grup deseninde bu sonuçlara dayalı Y formunun X formundan 5 puan daha kolay olduğunu belirtmiştik. Ancak ortak madde sonuçlarına bakıldığında I. grup 20 madde ortak maddenin 13'ünü, II. grup 20 ortak maddenin 15'ini doğru cevaplandırmıştır. Bu durum gruplar arasında ölçülen özellik bakımından yetenek farklılığı olduğunu göstermektedir. Başka bir anlatımla X-Y formları arasında görülen farklılığın bir kısmı formların güçlüğünden bir kısmı ise gruptaki yetenek farklılığından kaynaklanmaktadır (Kolen & Brennan, 2014; Sansivieri ve ark., 2017).

Tablo 6

Eş Olmayan Gruplarda Ortak Madde Deseni Örneği

Gruplar	**Formlar**		
	X formu	**Y formu**	**V (ortak maddeler)**
I. grup	72	---	13
II. grup	---	77	15

Desenin Avantajları ve Sınırlılıkları

Test eşitleme sürecinde eş değer olmayan gruplarda ortak madde deseni sıklıkla kullanılmaktadır. Bir grup için tek bir uygulama yapılması, cevaplayıcıların seçkisiz alt gruplara ayrılma zorunluluğu ve varsayımının bulunmaması nedeniyle avantajlı bir desendir. Bununla birlikte gruplar arasında yetenek farklılığın çok yüksek olduğu durumda kullanılması uygun değildir. Ayrıca kullanılan ortak maddelerin özellikleri de desenden kaynaklı eşitleme hatası oluşmasına neden olur. Bu nedenle bu bölümde ortak maddelerle ilgili ayrı bir başlıklandırma yapılmıştır.

Test Eşitlemede Ortak Maddeler

Bir testin farklı formları kullanıldığında puan dağılımlarında gözlenen farklılıklar; formların güçlüğünden veya test alan bireylerin yeteneklerinden kaynaklanabilir. Form ve grup farklılığı ortak maddelerle ayarlanır. Ortak maddeler ankor test maddesi olarak da adlandırılır. İki ortak madde tasarımı bulunmaktadır: İç ve dış ortak maddeler. Ortak maddelerin puanı, katılımcının toplam test puanına katkıda bulunuyorsa iç; toplam puana dâhil edilmiyorsa dış ortak madde olarak adlandırılır. Ayrıca iç ortak maddeler, testin kendisinin parçasıdır ve diğer maddelerle birlikte uygulanır. Dış ortak maddeler ise formlarla aynı uygulama koşullarında ancak farklı bir bölüm olarak uygulanır.

Ortak maddeler, test formlarının hem içeriğini hem de güçlük düzeyini yansıtmalıdır. Bu sayede toplam test puanı ile ortak madde puanları arasında yüksek bir pozitif korelasyon hesaplanabilir. Bunun için de ortak madde setinin testin ölçtüğü bilgi ve becerileri içerik açısından temsil etmesi gerekir. Başka bir anlatımla eşitlemede kullanılacak ortak maddeler testin ölçtüğü aynı özelliklerin küçük bir versiyonu olmalıdır. Ayrıca ortak maddeler formların güçlük düzeylerini de temsil etmelidir. Bu nedenle ortak maddelerden oluşan test, "mini test" olarak adlandırılır. Ancak test çok sayıda alt konuyu kapsıyor ve geniş bir güçlük aralığına sahipse bu durumda testin tümünü temsil eden ortak maddeler geliştirmek zor olabilir. Bu tür durumlarda "midi test" olarak adlandırılan orta test yaklaşımı önerilir. Testin genel özelliğine uygun maddelerle ortalama yeteneğe sahip katılımcılar için ortak maddeler kullanılır (Sinharay & Holland, 2006a).

Ortak madde sayısının genellikle test maddelerinin en az %20'sinin ortak madde olması gerektiği kabul edilir. Ortak madde sayısı arttıkça maddeler, testi daha iyi temsil eder ve hatayı azaltabilir ancak çok fazla ortak madde kullanılması da test süresini uzatır ve formların özgün madde özelliklerini gölgede bırakabilir. Bu nedenle asıl önemli olan ortak madde setinin testi yeterince temsil edecek büyüklükte olmasıdır. Örneğin beş maddelik bir testte yalnızca bir ortak madde yeterli temsil gücü sağlayamaz. Ayrıca testin boyut sayısına göre de kullanılacak ortak madde sayısı farklılık gösterir. Çok boyutlu testlerde tek boyutlu testlere göre daha fazla ortak madde kullanılması gerekir (Bramley, 2018; Kolen & Brennan, 2014; Sinharay & Holland, 2006b).

Ortak maddeler belirlendikten sonra bunların içsel mi yoksa dışsal mı kullanılacağına karar verilmelidir. İçsel ortak maddeler testin içine gömülüdür ve formlar arasında aynı sırada yer almalı, tamamen aynı şekilde hazırlanmalı, ifade veya cevap seçeneklerinde farklılık olmamalıdır. Ayrıca bu maddeler test boyunca dengeli dağıtılmalıdır. Başka bir anlatımla yalnızca testin başında veya sonunda yoğunlaşmamalıdır. Buna karşılık dışsal ortak maddeler ayrı bir test bölümü olarak uygulanır ve aynı koşullar altında gerçekleştirilmelidir. Ancak bu yöntem ek süre ve katılımcı kaybı riski yaratabileceğinden çoğunlukla içsel ortak maddeler tercih edilir (Kolen & Brennan, 2014; Sinharay & Holland, 2006a).

Uygulanan testin karma formatta, farklı madde türlerini içermesi durumunda ortak maddelerin de temsiliyet açısından karma formatta olması beklenir. Bununla birlikte çoktan seçmeli maddelerin daha fazla olduğu, kısa cevaplı ve/veya açık uçlu maddelerin kısıtlı olduğu bir test senaryosunda ortak maddeler yalnızca çoktan seçmeli sorulardan oluşabilir veya çoktan seçmeli ve yapılandırılmış yanıt maddelerinin bir birleşimini içerebilir. Burada (i) Kısa cevaplı sorular testin önemli bir bölümünü oluşturuyor mu? (ii) Bu sorular çoktan seçmeli sorulardan farklı becerileri mi ölçüyor? (iii) Eğer öyleyse gruplar bu becerilerde farklılık gösterebilir mi? sorularının sorulması, bu sorulara eğer evet cevabı alınıyor ise kısa cevaplı maddelerin de ortak madde setine dâhil edilmesi, ortak maddelerin de karma formatta olması gerekir. Bununla birlikte bu tür maddelerin puanlaması zor olduğu ve puanlamasına hatalar da karışabildiğinden bazı uygulamalarda ortak madde seti, genellikle çoktan seçmeli maddelerle sınırlı kalır. Bu durumda gruplar arası yetenek farklılıklarının çoktan seçmeli bölümde tüm testi yeterince temsil ettiği varsayılır (Sinharay & Holland, 2006a).

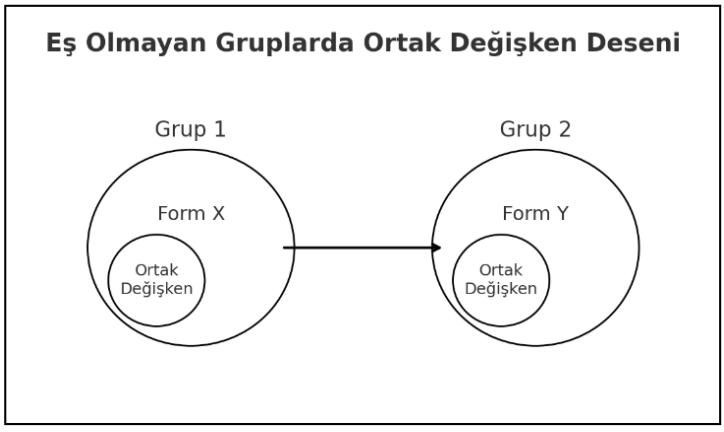
Test formları bazen ortak köke dayalı maddelerden oluşabilir. Madde takımı (testlet) adı verilen bu durumda yerel bağımsızlık varsayımını ihlal edebilir. Başka bir anlatımla bir maddenin doğru cevaplanması başka bir maddeden etkileniyor olabilir. Bu durum güvenirliği ve korelasyonu etkileyeceği için ortak madde setinde olabildiğince bağımsız maddelerin olması önerilir (Cao ve ark., 2014; Sansivieri ve ark., 2017).

4.4.5. Eş Olmayan Gruplarda Ortak Değişken Deseni

Eş olmayan gruplarda ortak madde kullanımı deseni oldukça yaygın kullanılsa da bazı durumlarda (sürenin kısıtlı olması, ortak maddelerin bilinmesi vb.) ortak maddeler kullanılamamaktadır. Bu durumda test formları ve grup farklılıkları ayrışması için gruplara ait ortak değişkenler kullanılır (Şekil 18).

Şekil 18

Eş Olmayan Gruplarda Ortak Değişken Deseni



Eş olmayan gruplarda ortak değişken deseninde ortak değişken olarak cinsiyet, sosyo-ekonomik düzey, başarı gibi test performansıyla ilişkili değişkenler seçilebilir. Farklı evrenlerden gelen ve eş olmayan iki gruptan biri X formunu, diğeri Y formunu cevaplandırır. Ortalama puanlar arasındaki farklılık ortak değişken ile ayarlanır (Sansivieri ve ark., 2017).

Desenin Avantajları ve Sınırlılıkları

Eş olmayan gruplarda ortak değişken deseninde her bir test formu sadece bir katılımcı grubuna uygulandığı için maliyeti düşük ve avantajlı bir desendir. Ayrıca grupların eş değer olduğu varsayımı ile form uygulanmasından kaynaklı bir sıra etkisi söz konusu değildir. Test formlarında ortak madde olmadığı için ortak maddenin geliştirilmesi, test sürecinin karmaşıklaşması ve cevaplandırılma süresinin uzaması gibi durumlar söz konusu değildir. Ayrıca ortak maddelerin açığa çıkma gibi bir sorunla karşılaşılmaz. Bununla birlikte desenin en büyük sınırlılığı test performansını açıklayan ortak değişken belirleme zorluğudur. Ortak değişken ile test puanları arasındaki ilişki zayıfsa eşitleme sonuçları güvenilir olmaz. Bu durum özellikle ortak değişkenin test ile ölçülen becerileri tam olarak yansıtmadığı durumlarda sorun yaratır.

4.5. Test Eşitleme Yöntemleri

Test eşitleme aşamalarında da söz edildiği gibi test eşitleme deseninin seçimine bağlı olarak test eşitleme yöntemi seçilmektedir. Örneğin seçkisiz grup deseninin kullanıldığı bir eşitleme sürecinde ortak madde olmadığı için ortak maddeye dayalı bir eşitleme yöntemi kullanılamaz. Benzer şekilde eş değer olmayan gruplar yöntemlerinde grup ve form farklılıklarını ayrıştırmaya yönelik yöntemlerin kullanılması gerekir. Bu da istatistiki olarak daha karmaşık bir süreci ifade eder. Bu nedenle de alan yazınında test eşitleme yöntemleri, Klasik Test Kuramı ve Madde Tepki Kuramına dayalı yöntemler olarak da sınıflandırılır. Sınıflandırma Tablo 7'de yer almaktadır (González & Wiberg, 2017; Kolen & Brennan, 2014).

Tablo 7

Test Eşitleme Yöntemleri

Test Eşitleme Deseni	Test Eşitleme Yöntemleri
Klasik Test Kuramına Dayalı Eşitleme (Seçkisiz grup deseni)	Ortalama eşitleme
	Doğrusal eşitleme
	Eşit yüzdelikli eşitleme
Klasik Test Kuramına Dayalı Eşitleme (Eş olmayan gruplar deseni)	Doğrusal eşitleme Tucker yöntemi Braun-Holland yöntemi Levine-gözlenen yöntemi Levine-gerçek yöntemi Zincirleme doğrusal yöntemi
	Eşit yüzdelikli eşitleme Frekans kestirim yöntemi Düzeltilmiş frekans yöntemi Zincirleme eşit yüzdelikli yöntemi
Madde Tepki Kuramına Dayalı Eşitleme	Momentler eşitlemesi Ortalama-ortalama yöntemi Ortalama-sigma yöntemi
	Karakteristik eğri eşitlemesi Haebara yöntemi Stocking-Lord yöntemi

4.5.1. Ortalama Eşitleme Yöntemi

Seçkisiz grup deseninde kullanılan ortalama eşitleme yöntemi grupların eş olduğu varsayımından hareketle ortalamalar arasındaki farklılığın sadece formlardan kaynaklandığını varsayar. Ayrıca bu farklılığın tüm katılımcı grubunda eş olduğu düşüncesine dayanır (Eşitlik 1) (Kolen & Brennan, 2014).

$$x - \mu(X) = y - \mu(Y)$$
$$m_Y(x) = y = x - \mu(X) + \mu(Y)$$

- x: X formundaki ham puan (eşitlenecek puan)
- μ(X): X formunun ortalaması
- y: Eşitlenmiş puan
- μ(Y): Y formunun ortalaması

Eşitlik 1

Tablo 4'te seçkisiz grup desenindeki uygulama sonucunu hatırlayalım. X formunu alan grubun ham puan ortalaması 72, Y formunu alan grubun 77 olarak hesaplanmıştı. Bu durumda X formu ile Y formu arasında hesaplanan 5 puan farkın tüm katılımcılar için düşük yetenek-yüksek yetenek fark etmeksizin aynı şekilde olduğu belirlenir. Test eşitleme sürecinde de X formunu cevaplandıran herkese 5 puan eklenerek X formunun Y formundaki eşitlenmiş puanı hesaplanmış olur. Başka bir anlatımla X formundan 40 alan bir katılımcının Y formundaki eşitlenmiş puanı 45; X formundan 90 alan bir katılımcının Y formundaki eşitlenmiş puanı da 95 olarak hesaplanır.

Ortalama eşitleme yöntemi oldukça basit bir eşitleme süreci içermektedir. Test formlarından tam puan alınması durumunda ranjın dışında eşitlenmiş puanlar üretilmektedir. Ayrıca formların güçlük farklılıklarının tüm yetenek düzeylerindeki katılımcıları aynı şekilde etkilediği varsayımını karşılamak güçtür. Bu anlamda bakıldığında oldukça sınırlı bir eşitleme yöntemi olsa da bazı önemli eşitleme kavramlarını göstermesi ve küçük örneklemlerde kullanılması bakımından önemlidir.

4.5.2. Doğrusal Eşitleme Yöntemi

Doğrusal eşitleme yöntemi de ortalama eşitleme yönteminde olduğu gibi grupların eş değer olduğu ve farklılıkların sadece formdan kaynaklandığı varsayımını temele alır. Bununla birlikte ortalama eşitleme yönteminden farklı olarak formlar arasındaki farklılığın tüm katılımcı grubunu aynı şekilde etkilemediğini belirtir. Bu doğrultuda da eşitleme sürecine ortalamanın yanı sıra standart sapma değerini de ekler (Eşitlik 2).

$$l_Y(x) = y = \sigma(Y)/\sigma(X)\, x + [\mu(Y) - \sigma(Y)/\sigma(X)\, \mu(X)]$$

- x: X formundaki ham puan (eşitlenecek puan)
- y: Y formundaki eşitlenmiş puan (bulunacak değer)
- μ(X): X formunun ortalaması
- μ(Y): Y formunun ortalaması
- σ(X): X formunun standart sapması
- σ(Y): Y formunun standart sapması
- $l_Y(x)$: Doğrusal eşitleme fonksiyonu (linear equating)

Eşitlik 2

Eşitlik 2'de görüldüğü gibi eğim ve kesişim değerlerinden oluşan doğrusal eşitleme yönteminde güçlük farklılıkları ölçek puanları boyunca değişir. Tablo 3'teki örnekte X formunu alan grubun ham puan ortalaması 72, Y formunu alan grubun 77 olarak hesaplanmıştı. X formunun puanlarının standart sapmasının 10, Y formunun puanlarının standart sapmasının 9 olduğunu varsayalım. Bu durumda X formundan 40 alan bir katılımcının Y formundaki eşitlenmiş puanı 48.2 iken 90 puan alan birinin Y formundaki eşitlenmiş puanı 93.2'dir.

Doğrusal eşitlemede X formundaki puanlar, ölçek boyunca formların farklı güçlükte olmasına olanak sağlayan bir doğrusal eşitlik ile düzeltilir. Bununla birlikte doğrusal eşitleme yönteminde de eşitlenmiş puanların olası gözlenen puan aralığı dışına çıkması mümkündür.

4.5.3. Eşit Yüzdelikli Eşitleme Yöntemi

Eşit yüzdelikli eşitleme, seçkisiz gruplar deseninde iki farklı test formunda aynı yüzdelik sırasına denk gelen puanların birbirine eşitlenmesine dayalı bir yöntemdir. Bu doğrultuda eşit yüzdelik eşitlemede ilk olarak her form için yüzdelik sıralar hesaplanır. Daha sonra aynı yüzdelik sıraya sahip katılımcıların puanları birbirine eşitlenir (Eşitlik 3).

$$e_Y(x) = G^{-1}[F(x)]$$
$$e_X^{-1}(x) = e_Y(x) \text{ ve } e_Y^{-1}(y) = e_X(y)$$
$$e_X(y) = F^{-1}[G(y)]$$

- x: X formundaki ham puan
- $e_Y(x)$: X'ten Y'ye eşitleme fonksiyonu
- F(x): X formunda x puanının kümülatif dağılım fonksiyonu
- G^{-1}: G'nin ters fonksiyonu
- e_x^{-1}: X'in ters eşitleme fonksiyonu
- e_y^{-1}: Y'nin ters eşitleme fonksiyonu
- y: Y formundaki ham puan
- $e_X(y)$: Y'den X'e eşitleme fonksiyonu
- G(y): Y formunda y puanının kümülatif dağılım fonksiyonu
- F^{-1}: F'nin ters fonksiyonu

Eşitlik 3

Eşitleme sürecinde (i) belirli bir puanın X formundaki yüzdeliği altında puan alan bireylerin yüzdesi bulunur, (ii) belirlenen yüzdeliğin Y formunda karşılık geldiği puan bulunur, (iii) belirlenen puan eş değer olarak belirlenir. Tablo 3'teki örnekte X formunu alan grubun ham puan ortalaması 72, Y formunu alan grubun 77 olarak hesaplanmıştı. X formundan 40 puan alan bir kişinin puanı eşitlenmek istendiğinde X formunda katılımcıların %20'sinin 40 ve altında puan aldığı belirlenir. Y formunda katılımcıların %20'sinin 44 ve altında puan aldığı tespit edilir. Bu doğrultuda X formundaki 40 puanın Y formundaki eşitlenmiş puanı 44 olur. Ancak Y formunda katılımcıların %20'sinin aldığı bir puan da olmayabilirdi, bu nedenle eşit yüzdelikli eşitleme, ortalama ve doğrusal eşitleme yöntemlerine göre daha karmaşık bir matematiksel süreci ifade eder.

Ortalama ve doğrusal eşitleme yöntemlerinin aksine eşit yüzdelik eşitleme sonucunda elde edilen eşitlenmiş puanlar ranjın dışında hesaplanmaz. Eşit yüzdelikli eşitleme sonucunda Y formundaki puanların dağılımının, X formundan eşitlenmiş puanların dağılımı ile benzerlik göstermesi istenir. Bu nedenle de eşitleme sürecine ortalama, standart sapma, çarpıklık ve basıklık değerleri dâhil edilir. Eşit yüzdelikli eşitlemede eşitleme örneklem yüzdeleri ve yüzdelik sıralar kullanılarak gerçekleştirildiği için daha geniş örneklemlere ihtiyaç duyulur. Ayrıca örneklem dağılımından yapılan kestirimlerle evrende düzgünlük özelliği olan eşit yüzdelikli ilişkiler için düzgünleştirme yöntemleri kullanılır.

4.5.4. Eş Olmayan Gruplarda Doğrusal Eşitleme Yöntemi

Eş olmayan gruplar deseninde katılımcılara eşitlenecek test formlarından sadece biri uygulanır. Ancak gruplar arasında ortak madde veya ortak değişkene bağlı bir bağlantı belirlenerek grupların performansı bu ortaklık üzerinden karşılaştırılır. X ve Y formlarının ortalamaları arasındaki farklılıklar, katılımcı grup farklılığının ve test formu farklılığı karışımının bir sonucudur. Bu nedenle ortak madde ya da değişkenden hareketle grup farklılıkları, form farklılıklarından ayrılır. Bu farklılıkları ayırma sürecinde doğrusal yöntemler içinde Tucker Yöntemi, Levine Gözlenen Puan Yöntemi, Levine Gerçek Puan Yöntemi ve Zincirleme Doğrusal Eşitleme Yöntemine dayalı olarak bilinmeyen parametreler kestirilir. Tablo 6'da yer alan örneği hatırlarsak birinci grup X formunu cevaplandırmış, ortalaması 72, standart sapması 10 olarak hesaplanmıştır. Ayrıca ortak maddelerdeki ortalama puanı 13'tür. İkinci grup ise sadece Y formunu cevaplandırmış, ortalaması 77, standart sapması 9 olarak hesaplanmıştır. Ortak maddelerdeki ortalaması 15 olarak hesaplanmıştır. Bu doğrultuda birinci grup Y formunu alsaydı ortalaması ve standart sapması kaç olurdu ile ikinci grup X formunu alsaydı ortalaması ve standart sapmasının kaç olacağı eşitleme sürecinde bilinmeyen parametreleri oluşturmaktadır (Eşitlik 4). Ortak maddelerin cevaplarına dayalı olarak bilinmeyen parametre kestirimleri gerçekleştirilir.

$$\mu_s(X) = w_1\mu_1(X) + w_2\mu_2(X)$$
$$\mu_s(Y) = w_1\mu_1(Y) + w_2\mu_2(Y)$$
$$\sigma^2{}_s(X) = w_1\sigma^2{}_1(X) + w_2\sigma^2{}_2(X) + w_1w_2[\mu_1(X) - \mu_2(X)]^2$$
$$\sigma^2{}_s(Y) = w_1\sigma^2{}_1(Y) + w_2\sigma^2{}_2(Y) + w_1w_2[\mu_1(Y) - \mu_2(Y)]^2$$

- $\mu_s(X)$: Sentetik evrende X formunun ortalaması
- $\mu_s(Y)$: Sentetik evrende Y formunun ortalaması
- w_1: 1. grubun ağırlığı
- w_2: 2. grubun ağırlığı
- $\mu_1(X)$: 1. grupta X formunun ortalaması
- $\mu_2(X)$: 2. grupta X formunun ortalaması
- $\mu_1(Y)$: 1. grupta Y formunun ortalaması
- $\mu_2(Y)$: 2. grupta Y formunun ortalaması
- $\sigma^2{}_s(X)$: Sentetik evrende X formunun varyansı
- $\sigma^2{}_s(Y)$: Sentetik evrende Y formunun varyansı
- $\sigma^2{}_1(X)$: 1. grupta X formunun varyansı
- $\sigma^2{}_2(X)$: 2. grupta X formunun varyansı
- $\sigma^2{}_1(Y)$: 1. grupta Y formunun varyansı
- $\sigma^2{}_2(Y)$: 2. grupta Y formunun varyansı

Eşitlik 4

Doğrusal yöntemler içinde Tucker yöntemi, eşitleme sürecinde bilinmeyen parametreleri tahmin edebilmek için iki temel varsayıma dayanır. İlk varsayım, toplam test puanları ile ortak madde puanları arasındaki regresyon ilişkisidir; bu ilişki farklı gruplarda benzer kabul edilir. İkinci varsayım ise ortak madde puanları verildiğinde toplam puanların koşullu varyanslarının gruplar arasında tutarlı olmasıdır. Yöntem,

ortak maddelerin gruplar arasındaki performans farklılıklarını doğru biçimde temsil ettiği varsayımına dayanır. Bu çerçevede Tucker yöntemi, gözlenmeyen parametreleri ortak madde verilerinden yola çıkarak tahmin eder ve test formları arasında puanların karşılaştırılabilirliğini sağlar.

Braun–Holland yöntemi, Tucker eşitlemesine yakın bir mantıkla çalışır ancak özellikle frekans kestirim yönteminden elde edilen ortalama ve varyans değerlerini kullanır. Bu çerçevede X formunun ortalama ve varyansı sentetik evren üzerinden tahmin edilerek Y formuyla uyumlu hâle getirilir. Ardından regresyon katsayıları ve sabitler hesaplanır; bu hesaplamada Y formunun standart sapmasının X formunun standart sapmasına oranı eğim değerini verirken, ortalamalar arasındaki fark da sabiti oluşturur. Böylece elde edilen parametreler eşitleme denklemine yerleştirilerek X formu Y formuna dönüştürülür.

Levine gözlenen yöntemde sentetik evren kavramı doğrudan ele alınmaz, bunun yerine X formundan elde edilen gözlenen puanların Y formunun gözlenen puan ölçeğine aktarılması hedeflenir. Eğer dönüşüm işlemi gerçek puanlar üzerinden yapılıyorsa yöntem Levine gerçek yöntem olarak adlandırılır. Farklı test formları gerçek puanlar üzerinden ortak bir ölçeğe aktarılır.

Zincirleme doğrusal yöntemde önce X formu, ortak madde testi V testine dönüştürülür; ardından Y formu aynı ortak test üzerinden dönüştürülür; son aşamada ise X'ten V'ye ve V'den Y'ye yapılan bu eşitlemeler birleştirilerek X formu doğrudan Y formuna aktarılır. Bu yöntem, X ve Y formlarının V ile ilişkili olması varsayımına dayanır.

4.5.5. Eş Olmayan Gruplarda Eşit Yüzdelikli Eşitleme Yöntemi

Eş olmayan gruplarda eşit yüzdelikli eşitleme yönteminde gözlenen puanlara dayalı olarak bir eşitleme süreci yürütülür. Tucker yönteminde olduğu gibi burada da sentetik bir evren oluşturulur ancak yalnızca ortalama, standart sapma ve kovaryans değerleri dikkate alınmaz aynı zamanda toplam ve ortak puanların dağılımları da sürece dâhil edilir. Eşitleme işlemi, kümülatif dağılımların yüzdelik sıraları üzerinden gerçekleştirilir. Bu yaklaşımda ortak maddeler aracılığıyla X ve Y formlarının puan dağılımları birbiriyle uyumlu hâle getirilerek farklı gruplardan elde edilen puanların karşılaştırılabilirliği sağlanır. Bu süreçte yine bilinmeyen parametreler frekans kestirim yöntemi, düzeltilmiş frekans yöntemi ve zincirleme eşit yüzdelikli eşitleme yöntemi ile gerçekleştirilir.

Frekans kestirim yönteminde her bir puanın frekansı dikkate alınır ve sentetik evren üzerinden ortalama ile varyans değerleri hesaplanır. Böylece farklı formlardan elde edilen dağılımlar birbirine uyarlanır. Küçük örneklem büyüklüklerinden kaynaklanabilecek dengesizlikleri azaltmak için özellikle uç değerlerdeki frekansların etkisini azaltarak eşitleme sürecinin daha kararlı sonuçlar vermesini sağlamak için düzeltilmiş frekans yöntemi kullanılır. Doğrudan X ile Y formlarını eşitlemek yerine ortak bir test veya değişken üzerinden dolaylı bir ilişki kurarak yapılan eşitleme yöntemi de zincirleme eşit yüzdelikli eşitleme yöntemi olarak adlandırılır.

4.5.6. Madde Tepki Kuramına Dayalı Momentler Eşitleme Yöntemleri

Madde tepki kuramına (MTK) dayalı olarak test eşitleme sürecinde MTK'nin tek boyutluluk, yerel bağımsızlık, parametre değişmezliği ve madde karakteristik eğrisinin doğası varsayımlarının karşılanması gerekir. Ardından parametre kestirimleri yapılarak eşitleme sürecine geçilir. X formunun MTK parametreleri Y formunun parametrelerinden ayrı ayrı kestirilebilir. Eğer yetenek kestirimi aynı ölçekleme kuralı (ort 0, ss 1) ile kestirilirse iki form için parametre tahminleri daha fazla dönüşüm olmadan aynı ölçekte varsayılır. Seçkisiz grup deseninde gruplar eş olduğundan yetenekler her iki grupta da aynı dağılıma sahip olacağından teta için ortalama ve standart sapma değerleri aynı/benzer olacaktır. Ancak iki formu alan bireyler eş olmadığında yetenek ve parametre tahminleri aynı ölçek düzeyinde değildir. Bu durumda da iki formda da yer alan ortak madde seti ya da değişkenden hareketle parametre dönüşümleri gerçekleştirilir.

Ortak maddeli eş değer olmayan grup desenindeki ölçekleri dönüştürmenin en basit yolu ortak maddelerdeki madde parametrelerinin ortalama ve standart sapmaları birbirinin yerine kullanmaktır. Dönüştürmeden sonra madde parametreleri, genellikle «kalibre edilmiş» olarak adlandırılır. Loyd ve Hoover (1980) tarafından önerilen yöntemde A sabitini tahmin etmek için a parametresinin ortalaması kullanılır (Eşitlik 5).

$$A = \mu(aI) / \mu(aJ)$$
$$B = \mu(b_J) - A\mu(b_I)$$

- A: Eğim parametresi (slope/ölçek dönüşüm katsayısı)
- b: Madde güçlük parametresi
- a: Madde ayırt edicilik parametresi
- I: Form I (veya 1. grup)
- J: Form J (veya 2. grup)
- B: Kesişim parametresi (intercept)

Eşitlik 5

$$A = \sigma(bJ) / \sigma(bI)$$
$$B = \mu(bJ) - A\mu(bI)$$

- σ : Parametrelerin standart sapması
- μ: Parametrelerin ortalaması

Ortalama-sigma yönteminde ise sabitlerin kestirilmesi için b parametrelerinin ortalaması ile standart sapması kullanılır. Ortalama-ortalama ve ortalama-sigma yöntemleri genellikle farklı sonuçlar verir. b parametresi a parametresine göre daha kararlı kestirilen bir parametredir. Bu, ortalama-sigma yönteminin bir avantajıdır. Ancak ortalama da standart sapmadan daha kararlı kestirilen bir değer olduğundan ortalama-ortalama yönteminin kullanılması da uygundur.

4.5.7. Madde Tepki Kuramına Dayalı Karakteristik Eğrisi Dönüşüm Yöntemleri

Madde Tepki Kuramı ile kestirilen a, b, c ve bazı durumlarda kestirilen d parametre tahminleri sonucunda maddelerin madde karakteristik eğrileri genel olarak benzerlik gösterse de farklı parametre kestirimleri nedeniyle ortalama-ortalama ve ortalama-sigma yöntemleri bu durumdan olumsuz etkilenebilir. Bu nedenle Haebara (1980) tüm madde parametrelerini aynı anda dikkate alan bir yöntem önermiştir (Eşitlik 6).

$$Hdiff(\theta_i) = \Sigma\,[p_{ij}(\theta_i;\, \hat{a}_j,\, \hat{b}_j,\, \hat{c}_j) - p_{ij}(\theta_i;\, \hat{a}_j/A,\, A\hat{b}_j + B,\, \hat{c}_j)]^2$$

- $Hdiff(\theta_i)$: Haebara fark fonksiyonu
- θ_i: i. yetenek parametresi (ability/theta)
- Σ j:V: Ortak maddeler (V) üzerinden toplam
- p_{ij}: Madde tepki fonksiyonu - j. maddeye i. yetenek seviyesinde doğru yanıt olasılığı
- $\hat{a}_j$: j. maddenin ayırt edicilik parametresi tahmini

Eşitlik 6

- $\hat{b}_j$: j. maddenin güçlük parametresi tahmini
- $\hat{c}_j$: j. maddenin şans parametresi tahmini
- A: Eğim dönüşüm parametresi
- B: Kesişim dönüşüm parametresi

Eşitlik 6'da görüldüğü gibi Haebara yönteminde gruplara ilişkin madde karakteristik eğrileri çizilerek bunlar arasındaki farklılık hesaplanır. Sonra da belirli bir yetenekteki bireylerin tüm maddeler için madde karakteristik eğrileri arasındaki farklılığın kareleri toplanmaktadır. Başka bir anlatımla iki ölçekteki tüm maddelerin MKE arasındaki farklılığın kareleri her bir cevaplayıcı için hesaplanmış ve toplanmıştır. Tüm cevaplayıcıların sonuçları da toplanmaktadır. Stocking-Lord yönteminde Haebara yaklaşımından farklı olarak tüm maddeler üzerinde toplamların farkının karesi alınır. Bu hâli ile bakıldığında yöntem, test karakteristik eğrisi arasındaki farkın karesi olarak ifade edilebilir (Stocking & Lord, 1983).

4.6. Test Eşitleme Hatası

Test eşitleme sürecine karışan iki türlü hata vardır: Sistematik ve tesadüfi hata. Sistematik hata; eşitlemenin koşullarının, varsayımların ihlal edilmesinden ve yanlılıktan kaynaklanan hatayı ifade ederken tesadüfi hata, katılımcı örnekleminden kaynaklanan hatayı ifade etmektedir. Örneğin;

- test formlarının içerik olarak uygun yapılandırılmaması
- ortak maddeli desende ortak maddelerin X ve Y formundaki yerlerinin farklılaşması
- eşitleme yapmak için uygulama yapılan birey ve grubun önemli derecede farklılaşması
- seçkisiz grup deseninde grupların eş olmaması
- desene uygun kestirim yöntemi ya da düzgünleştirme yönteminin seçilmemesi
- eşitleme yöntemlerinin varsayımlarının ihlal edilmesi
- Tucker yönteminde X'in ortak madde formu olan V formundaki regresyonunun Evren 1 ve Evren 2'de farklılaşması

gibi durumlar sistematik eşitleme hatasının oluşmasına neden olur.

Eşitlemenin standart hatası olarak adlandırılan tesadüfi hata ise eşitlenmiş puanların bir birey evreninden elde edilen örneklemlerdeki bir eşitleme yönteminin hipotetik replikasyonları üzerindeki standart sapmasıdır. Başka bir anlatımla her bir puan düzeyindeki eşitlemenin standart hatası X formundaki her bir puan düzeyi için Y formu karşılıklarının standart hatasıdır. Örneğin;

- Evrenden 1000 kişilik seçkisiz bir örneklem seçelim.
- Eşitleme yöntemi ve verileri kullanarak X formundaki puanın Y formundaki eş değerini bulalım.
- Birinci ve ikinci basamağı tekrarlayarak X formundaki 75 puanın eş değeri olan birçok Y formu puanı kestirelim.
- Bu kestirimlerin standart sapması, X formunun 75 puanı için eşitlemenin standart hatasının kestirimidir.

Başka bir anlatımla eşitleme hatası örneklemden X formunun Y formundaki eş değeri ile beklenen eş değeri arasındaki farktır. Evrenden rastgele seçilen örneklemler üzerindeki beklentidir.

4.7. Test Eşitleme Uygulama Örnekleri

Test eşitleme, uluslararası standartlaştırılmış testlerde genel bir prosedür olarak uygulanmaktadır. Örneğin Amerika Birleşik Devletleri'nde üniversiteye giriş sınavında uygulanan SAT (Scholastic Assessment Test-Akademik Değerlendirme Sınavı), 1926'dan beri uygulanmaktadır. 1995 yılından itibaren her yeni formun önceki formlara eşitlenmesiyle yıllar arasında karşılaştırılabilir puanlar üretmektedir. Böylece 1995'te 1200 puan alan bir öğrenci ile 2023'te aynı puanı alan öğrencinin benzer yetenek düzeyinde olduğu kabul edilmektedir. Benzer şekilde GRE (Graduate Record Examination- Lisansüstü Giriş Sınavı) bilgisayara uyarlanmış test formatında uygulanmakta ve her aday için farklı bir test formu oluşturulmaktadır. Madde Tepki Kuramı tabanlı yöntemlerle farklı madde kombinasyonlarından elde edilen puanların aynı ölçek üzerinde raporlanmasını sağlar. Avrupa'da Hollanda merkezli CITO (Centraal Instituut voor Toetsontwikkeling - Merkezi Test Geliştirme Enstitüsü) testleri, ilkokul düzeyindeki öğrenci performanslarını yıllar boyunca izlemek amacıyla her yeni yıl geliştirilen formlar önceki yıllara eşitlenerek öğrenci başarısındaki değişimler izlenmektedir. OECD (Organisation for Economic Co-operation and Development-Ekonomik İş Birliği ve Kalkınma Örgütü) tarafından uygulanan PISA uygulamasında da farklı formlardan alınan test puanları birbirine eşitlenir (Cao ve ark., 2014; Dorans ve ark., 2010; Issayeva, 2021; Sinharay & Holland, 2006a; 2006b).

Ulusal alanyazında da test eşitlemeye yönelik yapılan birçok çalışma olduğu görülmektedir. Öztürk ve Anıl (2012) Akademik Personel ve Lisansüstü Eğitimi Giriş Sınavı (ALES) puanlarını eşitlemeye yönelik bir çalışma yapmışlardır. Öztürk Gübeş ve

Kelecioğlu (2015) da farklı test eşitleme yöntemlerinin eşitlik özelliği ölçütüne göre karşılaştırılmışlardır. Mutluer (2021) çalışmasında Klasik Test Kuramına ve Madde Tepki Kuramına Dayalı test eşitleme yöntemlerini karşılaştırmıştır. Demir ve Güler (2014) eş olmayan gruplarda ortak madde kullanımını incelemişlerdir. Pektaş ve Kılınç (2016), gözlenen puan eşitleme sonuçlarını karşılaştırmıştır. Aksekioğlu (2017), madde tepki kuramına dayalı test eşitleme yöntemlerini karşılaştırmıştır. Akın Arıkan (2017) Kernel eşitleme ve Madde Tepki Kuramına dayalı eşitleme yöntemlerini karşılaştırmıştır. Atar ve Yeşiltaş (2017) çok boyutlu eşitleme yöntemlerinin sonuçlarını karşılaştırmışlardır. Uyar ve arkadaşları (2018), farklı ölçekleme ve eşitleme yöntemlerine ilişkin sonuçları karşılaştırmışlardır. Yıldırım ve arkadaşları. (2026), ölçek dönüştürme yöntemlerinin, ortak madde oranının ve madde ayırt ediciliğinin eşitleme hatasına etkisini araştırmıştır. Koğar ve Sayın (2018) tarafından denk olmayan gruplarda Kernel eşitleme yönteminden elde edilen sonuçlar karşılaştırılmıştır. Çörtük ve Sinan (2023) çok kategorili puanlanan test eşitleme yöntemlerini karşılaştırmıştır. Sezer Başaran (2023) farklı ortak değişkenlerle test eşitlemenin ortak maddeli test eşitlemeyle karşılaştırılmıştır. Demiröz ve Doğan (2025) çok boyutlu MTK ile test eşitleme sonuçlarını karşılaştırmıştır. Test eşitleme alanındaki birçok çalışma yöntemler üzerinde etkili olan değişkenleri bulmaya da odaklanmıştır. Örneğin Bozdağ ve Kan (2010) şans başarısının test eşitlemeye etkisini, Demirus ve Gelbal (2016) tarafından ortak maddelerin değişen madde fonksiyonu gösterip göstermemesi durumunu, Karagül (2020) örneklem sayısının eşitleme üzerindeki etkisini, Uysal ve Doğan (2021) açık uçlu madde puanlamasının test eşitleme üzerindeki etkisini Doğuyurt ve Tan (2023) yerel bağımsızlık varsayımın test eşitleme üzerindeki etkisini incelemiştir.

4.8. Test Eşitlemenin Gelecek Yönelimleri

Test eşitleme çalışmaları, test formlarından alınan puanların birbirinin yerine kullanıldığı durumlarda uygulanması gereken bir prosedürü ifade etmektedir. Uluslararası test uygulamalarında test geliştirme sürecinin bir prosedürü olarak kullanılmakla birlikte ülkemizde özellikle geniş ölçekli testlerde kullanım alanı sınırlıdır. Bu durumun başlıca nedeni test eşitleme desenleri içinde seçkisiz grup deseni varsayımının karşılanamaması ve eş değer olmayan gruplar için de ortak madde kullanımının güvenlik nedeniyle risk taşımasıdır. Ayrıca test formlarını eşitledikten sonra eşitlenmiş test puanlarının kullanımını da açıklamak zordur. Örneğin bir uygulamada X formunu alarak 75 maddeyi doğru cevaplandırılmış bir adayın puanını önceki forma eşitleyerek eşitlenmiş puanının 70 ya da 80 olduğunu açıklamak bir güvenlik endişesi yaratabilir. Bu bağlamda test eşitlemenin ülkemizde daha etkin kullanılabilmesi için hem yöntemsel hem de uygulamaya yönelik çözümlere ihtiyaç vardır. Özellikle madde tepki kuramına dayalı yöntemlerin kullanımı yaygınlaştırılabilir, güvenlik risklerini azaltacak ortak madde tasarımları geliştirilebilir ve eşitlenmiş puanların raporlanmasına ilişkin şeffaf standartları oluşturulabilir.

Özellikle bilgisayar teknolojisindeki gelişmelerle birlikte bilgisayar tabanlı test uygulamaları yaygınlaşmaktadır. Dijital test platformlarının yaygınlaşmasıyla birlikte gerçek zamanlı eşitleme sistemlerinin geliştirileceği düşünülmektedir. Özellikle bilgisayar uyarlamalı testlerde bu yaklaşım, her katılımcının performansına göre anlık ayarlamalar yapılmasını sağlayabilir.

5. SONUÇ

Bu bölüm, ölçme ve değelendirmede teknoloji entegrasyonu ve yapay zeka uygulamaları ile birlikte gelişin uygulamalar etrafında şekillenmiştir. BOBUT'un verimliliğinin, madde bilgi fonksiyonlarının ve yetenek kestiriminin doğru işletilmesine bağlı olduğu görülmektedir. BOBUT'un verimliliği, öncelikle testin başlangıç noktasını belirleyen giriş kuralının, ardından işletilen madde bilgi fonksiyonlarının ve yetenek kestiriminin doğru biçimde işletilmesine bağlıdır. Giriş kuralı, adayın ilk maddesinin ya da başlangıç θ değerinin, önceden tanımlı bir sabit değere göre mi yoksa adaya ilişkin ön bilgiye (ör. önceki test sonuçları) göre mi belirleneceğini ortaya koyar ve testin ilerleyen aşamalarındaki kestirim isabetini doğrudan etkiler. Madde seçiminde MFI, olabilirliği ağırlıklandıran MLWI, hesaplama yükü yüksek olsa da etkili bulunan Kullback-Leibler ölçütü ve önsel bilgiyi dikkate alan PWI, EPI gibi Bayesçi yaklaşımlar, adayın her yanıtından sonra en fazla bilgi sağlayacak maddeyi belirlemek üzere kullanılır. Benzer biçimde yetenek kestiriminde ML, uç örüntülerde daha dengeli sonuç veren WL, önsel dağılımı işe koşan MAP ve EAP ile aykırı değerlere karşı dirençli sağlam kestirimciler tercih edilebilmektedir. Testin nerede sonlanacağı ise sabit uzunluk yerine standart hata eşiği ya da ardışık θ kestirimleri arasındaki farkın belirli bir değerin altına inmesi gibi ölçütlerle belirlenmekte; bu ölçütler çoğu zaman minimum ve maksimum test uzunluğu kurallarıyla birlikte kullanılmaktadır. Çok aşamalı testlerde ise uyarlama madde yerine modül düzeyinde gerçekleşmekte ve tasarım kararları ölçme kesinliğini doğrudan etkilemektedir. Panel deseni (örneğin 1-2-3 veya 1-3-3), modül uzunluğu ve yönlendirme koşulları; ölçme hatasını (RMSE) ve madde teşhir oranını belirleyen temel değişkenlerdir. Alanyazındaki simülasyon çalışmaları, modül uzunluğu arttıkça hatanın azaldığını, ancak bunun kısa test avantajıyla dengelenmesi gerektiğini göstermektedir. Yönlendirmenin önceden tanımlı kesme puanlarına dayalı statik biçimi işlem yükü açısından ekonomikken, yanıtları ve modül bilgisini anlık değerlendiren dinamik yönlendirme daha yüksek uyum sağlamakta fakat uygulama karmaşıklığı artırmaktadır.

Bu bölümde otomatik madde üretimine yönelik üç temel yaklaşımdan söz edilmiştir. Bu yaklaşımlardan ilki olan model tabanlı yaklaşım, psikometrik nitelikleri bilinen bir temel maddeden yola çıkarak üç aşamalı bir süreç izlemektedir: Önce maddenin çözümü için gerekli bilgi, beceri ve içeriğin tanımlandığı bilişsel model; ardından bu özelliklerin madde biçimine nasıl yerleştirileceğini belirleyen madde modeli; son olarak da bu iki yapıyı birleştirerek üretimi gerçekleştiren bilgisayar algoritması oluşturulur. Bilişsel modelin, içeriğin belirli özellikler çerçevesinde ele alındığı anahtar özellikler modeli ile içeriğin değişip işlem mantığının sabit kaldığı mantıksal yapılar modeli gibi türleri;

madde modelinin ise sınırlı sayıda ve birbirine benzer madde üreten tek katmanlı yapılar ile dilin sözdizimsel olanaklarını kullanarak daha çok sayıda ve daha çeşitli madde üreten çok katmanlı yapılar biçiminde geliştirilmesi, bu yaklaşımın esnekliğini belirler. Nitekim aynı bilişsel modelden tek katmanlı bir madde modeliyle görece az, çok katmanlı bir modelle çok daha fazla madde üretilebilmekte; bu üretim IGOR, ModelCreator, Item Distiller gibi özel yazılımlar ya da R ve Python tabanlı araçlarla gerçekleştirilebilmektedir. Model tabanlı yaklaşımın en güçlü yönü, üretilen tüm maddelerin tanımlı bir bilişsel çerçeveye dayanması sayesinde ne ölçtüğünün ve nasıl ölçtüğünün önceden bilinebilmesidir; başlıca sınırlılığı ise sürecin büyük ölçüde uzman deneyimine bağlı, maliyetli ve yoğun bir emek gerektirmesidir. Otomatik madde üretimindeki bir diğer yaklaşım olan yapay zekâ tabanlı yaklaşımı metnin sözdizimsel ve anlamsal özelliklerinden doğrudan madde üretmeyi amaçlayan erken uygulamalardan; kelime gömme teknikleri (Word2Vec, GloVe) ve dizisel verileri işleyen RNN, LSTM, GRU gibi derin öğrenme mimarileriyle güçlenen kodlayıcı-çözücü (encoder-decoder) modellerine, oradan da BERT, GPT ve T5 gibi önceden eğitilmiş dönüştürücü modellere doğru gelişmiştir. Bu modeller, madde üretimi doğrultusunda yeniden eğitildiğinde ya da ince ayar (fine-tuning) işlemine tabi tutulduğunda tutarlılık ve ölçeklenebilirlik kazanır; denetimli ince ayar ile modelin tümü yerine yalnızca küçük katmanların güncellendiği LoRA ve QLoRA gibi parametre-verimli yöntemler bu süreci daha erişilebilir kılmaktadır. Eğitim sürecinde modelin tahmini ile doğru çıktı arasındaki farkın "kayıp (loss)" değeriyle izlenmesi ve bu değere göre modelin adım adım güncellenmesi, yaklaşımın hesaplamaya dayalı doğasını yansıtır. Yeniden eğitim yapılmadan, yalnızca istem mühendisliği yoluyla üretim de yaygındır; burada sağlanan örnek sayısına göre sıfır-atım, tek-atım, birkaç-atım ve çok-atım istem stratejileri bulunmaktadır. Bununla birlikte modellerin halüsinasyon eğilimi, aynı istemle farklı zamanlarda tutarsız sonuçlar alınabilmesi ve üretimin ölçeklenebilirliğinin garanti edilememesi, bu yaklaşımın başlıca sınırlılıkları arasında yer almaktadır. Üretilen maddelerin niteliği, hem içerik hem psikometrik açıdan çeşitli ölçütlerle değerlendirilmektedir. Model tabanlı yaklaşımın kontrollü ama emek yoğun üretimi ile yapay zekâ tabanlı yaklaşımın hızlı ama denetimi güç üretimi arasındaki bu durum, hibrit (artırılmış zekâ) yaklaşımının benimsenmesini sağlamıştır. Bu yaklaşımda insan uzmanı bilişsel ve madde modelinde yer alması gereken bileşenlerin çerçevesini belirlemekte, yapay zekâ bu çerçeve doğrultusunda model bileşenlerini üretmekte, üretilen modeller yeniden uzman denetiminden geçtikten sonra bilgisayar algoritmalarıyla birleştirilerek maddeler oluşturulmaktadır. Böylece süreç, geçerlik ve güvenirlik temelli bir yapıyı koruyarak ölçeklenebilir hâle gelmekte ve yapay zekâ üretimi bir "kara kutu" olmaktan çıkarak insan denetiminde, şeffaf ve hesap verebilir bir nitelik kazanmaktadır.

Bölüm içinde son olarak test eşitleme sürecinden söz edilmektedir. Test eşitleme, aynı amaç için geliştirilen formlardan elde edilen sonuçların karşılaştırılabilmesini sağlayan istatistiksel yöntemdir. Buradaki temel amaç içerik farklılıklarını değil, aynı yapıyı ölçen formlar arasındaki güçlük farklarını gidermek; böylece adayın hangi formu

yanıtladığından bağımsız olarak yalnızca yetenek düzeyini yansıtan bir puan elde etmektir. Bu nedenle eşitlemenin başarısı, öncelikle uygun bir veri toplama deseninin seçimine bağlıdır. Grupların eş değer kabul edildiği seçkisiz gruplar deseni, aynı bireylerin iki formu da yanıtladığı tek grup ve sıra etkisini dengeleyen dengelenmiş grup desenleri; formların eş zamanlı ya da aynı kişilere uygulanmasını gerektirdiğinden, güvenlik nedeniyle formların farklı zamanlarda uygulandığı büyük ölçekli sınavlarda çoğu kez kullanılamaz. Bu durumda formlar arasında bağlantıyı kuran ortak maddelere (çapa/ankor maddeler) ya da gruba ait ortak değişkenlere dayanan desenler tercih edilir. Ortak maddelerin testin hem içeriğini hem güçlük dağılımını temsil etmesi beklenir; bu nedenle testin küçültülmüş bir örneği olarak "mini test", geniş güçlük aralıklarında ise "midi test" biçiminde tasarlanmaları önerilir. Genel olarak maddelerin en az beşte birinin ortak madde olması, bu maddelerin form boyunca dengeli dağıtılması ve yerel bağımsızlığı zedeleyecek madde takımlarından (testlet) kaçınılması beklenmektedir. Test eşitleme deseninin seçiminden sonra eşitleme yöntemine karar verilir. Klasik Test Kuramı çerçevesinde seçkisiz gruplarda yalnızca ortalamalar arasındaki farkı tüm puanlara sabit biçimde ekleyen ortalama eşitleme; buna standart sapmayı da katarak güçlük farkının yetenek düzeyine göre değişmesine izin veren doğrusal eşitleme ve puan dağılımının tümünü (çarpıklık ve basıklık dâhil) dikkate alan eşit yüzdelikli eşitleme yer alır. Eş olmayan gruplarda ise form farklılıklarının grup farklılıklarından ayrıştırılmasını ele alan Tucker, Braun-Holland, Levine gözlenen puan, Levine gerçek puan ve zincirleme doğrusal yöntemler ile eşit yüzdelikli tarafta frekans kestirim, düzeltilmiş frekans ve zincirleme yöntemler kullanılır. Madde Tepki Kuramı temelli yöntemlerde ise iki formun ayrı ayrı kestirilen ölçekleri, ortak maddeler aracılığıyla doğrusal bir dönüşümle ($\theta_J = A \cdot \theta_I + B$) aynı ölçeğe taşınır. Bu dönüşümün eğim (A) ve kesişim (B) katsayıları, ayırt edicilik parametrelerinin ortalamasına dayanan ortalama-ortalama ile güçlük parametrelerinin standart sapmasına dayanan ortalama-sigma gibi moment temelli yöntemlerle ya da tüm madde karakteristik eğrilerini birlikte dikkate alan Haebara ve test karakteristik eğrisi üzerinden hesaplama gerçekleştiren Stocking-Lord gibi karakteristik eğri yöntemleriyle kestirilir. Simetriklik, aynı yapıyı ölçme, eşit güvenirlik, Lord'un eşitlik özelliği, gözlenen puan eşitliği ve grup değişmezliği koşullarının karşılanması, test eşitlemenin temel varsayımlarıdır. Eşitleme sürecine iki türlü hata karışmaktadır. Varsayımların ya da eşitleme koşullarının ihlalinden kaynaklanan sistematik hata ile örneklemden kaynaklanan ve genellikle "eşitlemenin standart hatası" olarak adlandırılan tesadüfi hata. Ortak maddelerin yerinin formlar arasında değişmesi, grupların yetenek dağılımlarının belirgin biçimde farklılaşması ya da desene uygun olmayan bir kestirim yönteminin seçilmesi gibi durumlar sistematik hatayı; örneklem büyüklüğünün yetersizliği ise tesadüfi hatayı artırır. Bu hataların kabul edilebilir düzeyde tutulması; GRE, SAT, CITO ve PISA gibi uluslararası uygulamalarda olduğu gibi farklı dönemlerde ve farklı formlarla elde edilen puanların karşılaştırılabilmesini mümkün kılmaktadır. Dolayısıyla test eşitleme, bireye uyarlanmış testler ve otomatik madde üretimiyle artan form çeşitliliğinin adil bir şekilde karşılaştırılmasına olanak tanımaktadır.

KAYNAKÇA | REFERENCES

Ackerman, R. & Balyan, R. (2023). Automatic multilingual question generation for health data using LLMs. *AI-generated content* (s. 1–11) içinde. Springer Nature Singapore. https://doi.org/10.1007/978-981-99-7587-7_1

Agrawal, A. & Shukla, P. (2023). Context aware automatic subjective and objective question generation using fast text to text transfer learning. *International Journal of Advanced Computer Science and Applications, 14*(4), 456–463.

Aksekioğlu, B. (2017). *Madde tepki kuramına dayalı test eşitleme yöntemlerinin karşılaştırılması: PISA 2012 fen testi örneği* [Yayımlanmamış Yüksek Lisans Tezi]. Akdeniz Üniversitesi.

Akın Arıkan, Ç. (2017). *Kernel eşitleme ve madde tepki kuramına dayalı eşitleme yöntemlerinin karşılaştırılması* [Yayımlanmamış Doktora Tezi]. Hacettepe Üniversitesi.

Alordiah, C. & Oji, J. (2024). Test equating in educational assessment: A comprehensive framework for promoting fairness, validity, and cross-cultural equity. *Asian Journal of Assessment in Teaching and Learning, 14*(1), 70–84.

American Educational Research Association (EARA), American Psychological Association (APA) & National Council on Measurement in Education (NCME). (2014). *Standards for educational and psychological testing*. American Educational Research Association.

Atar, B. & Yeşiltaş, G. (2017). Çok boyutlu eşitleme yöntemlerinin eşdeğer olmayan gruplarda ortak madde deseni için performanslarının incelenmesi. *Journal of Measurement and Evaluation in Education and Psychology, 8*(4), 421–434. https://doi.org/10.21031/epod.335284

Aybek, E. C. & Çıkrıkçı, R. N. (2018). Kendini Değerlendirme Envanteri'nin bilgisayar ortamında bireye uyarlanmış test olarak uygulanabilirliği. *Turkish Psychological Counseling and Guidance Journal, 8*(50), 117–141.

Bejar, I. I. (2002). Adaptive generative testing: From conception to implementation. S. Irvine & P. Kyllonen (Ed.), *Item generation for test development* (s. 199–218) içinde. Lawrence Erlbaum.

Berger, S., Verschoor, A., Eggen, T. & Moser, U. (2019). Improvement of measurement efficiency in multistage tests by targeted assignment. *Frontiers in Education, 4*. https://doi.org/10.3389/feduc.2019.00001

Bezirhan, U. & von Davier, M. (2023). Automated reading passage generation with OpenAI's large language model. *Computers and Education: Artificial Intelligence, 5*, 100161. https://doi.org/10.1016/j.caeai.2023.100161

Bormuth, J. R. (1970). *On a theory of achievement test items*. University of Chicago Press.

Bozdağ, S. (2024). *Otomatik madde üretimi ile geliştirilmiş sözel olmayan akıl yürütme maddelerinin psikometrik özelliklerinin incelenmesi* [Yayımlanmamış Yüksek Lisans Tezi]. Gazi Üniversitesi.

Bozdağ, S. & Kan, A. (2010). Şans başarısının test eşitlemeye etkisi. *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi, 39*(39), 91–108.

Braeken, J. & Paap, M. C. S. (2020). Making fixed-precision between-item multidimensional computerized adaptive tests even shorter by reducing the asymmetry between selection and stopping rules. *Applied Psychological Measurement, 44*(7–8), 531–547. https://doi.org/10.1177/0146621620932666

Bramley, T. (2018). *Evaluating the 'similar items method' for standard maintaining* [Konferans bildirisi]. Cambridge Assessment.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., … & Amodei, D. (2020). Language models are few-shot learners. *Advances in neural information processing systems* (s. 1877–1901) içinde.

Bulut, O. & Yıldırım-Erbaşlı, S. N. (2022). Automatic story and item generation for reading comprehension assessments with transformers. *International Journal of Assessment Tools in Education, 9*(Special Issue), 72–87. https://doi.org/10.21449/ijate.1124382

Büyükkıdık, S. & Ayva Yörücü, F. G. (2024). Çok aşamalı testlerin panel deseni, modül uzunluğu, örneklem büyüklüğü ve yetenek parametresi kestirim yöntemleri açısından farklı koşullar altında karşılaştırılması. *Boğaziçi Üniversitesi Eğitim Dergisi, 41*(3). https://doi.org/10.52597/buje.1329338

Campoy, G., Castellà, J., Provencio, V., Hitch, G. J. & Baddeley, A. D. (2015). Automatic semantic encoding in verbal short-term memory: Evidence from the concreteness effect. *Quarterly Journal of Experimental Psychology, 68*(4), 759–778. https://doi.org/10.1080/17470218.2014.966248

Cao, Y., Lu, R. & Tao, W. (2014). Effect of item response theory (IRT) model selection on testlet-based test equating. *ETS Research Report Series, 2014*(2), 1–13. https://doi.org/10.1002/ets2.12017

Chang, H. H. & Ying, Z. (1996). A global information approach to computerized adaptive testing. *Applied Psychological Measurement, 20*(3), 213–229. https://doi.org/10.1177/014662169602000303

Choi, J., Kim, H. & Pak, S. (2018). Evaluation of automatic item generation utilities in formative assessment application for Korean high school students. *Journal of Educational Issues, 4*(1), 68–89.

Choi, J. & Zhang, X. (2019). Computerized item modeling practices using computer adaptive formative assessment automatic item generation system: A tutorial. *The Quantitative Methods for Psychology, 15*(3), 214–225. https://doi.org/10.20982/tqmp.15.3.p214

Chughtai, R., Azam, F., Anwar, M. W., Haider But, W. & Farooq, M. U. (2022). A lecture centric automated distractor generation for post-graduate software engineering courses. *2022 International Conference on Frontiers of Information Technology (FIT)* içinde. IEEE. https://doi.org/10.1109/fit57066.2022.00028

College Board. (2025a). *About the SAT*. https://satsuite.collegeboard.org/sat

College Board. (2025b). *Bluebook: The official digital testing app for the SAT*. https://satsuite.collegeboard.org/sat/practice-preparation/bluebook

Çörtük, M. & Sinan, A. (2023). Çok kategorili puanlanan maddelerden oluşan testlerde klasik test kuramı ve madde tepki kuramı'na dayalı test eşitleme yöntemlerinin karşılaştırılması. *Turkish Academic Research Review, 8*(4), 1429–1439. https://doi.org/10.30622/tarr.1371179

Demir, S. & Güler, N. (2014). Ortak maddeli denk olmayan gruplar desenine ilişkin test eşitleme çalışması. *Journal of Human Sciences, 11*(2), 190–208. https://doi.org/10.14687/ijhs.v11i2.2870

Demiröz, B. & Doğan, N. (2025). Çok boyutlu madde tepki kuramı test eşitleme yöntemlerinin karşılaştırılması. *Manisa Celal Bayar Üniversitesi Eğitim Fakültesi Dergisi, 13*(1), 156–176. https://doi.org/10.52826/mcbuefd.1587864

Demirus, K. B. & Gelbal, S. (2016). Ortak maddelerin değişen madde fonksiyonu gösterip göstermemesi durumunda test eşitlemeye etkisinin farklı yöntemlerle incelenmesi. *Journal of Measurement and Evaluation in Education and Psychology, 7*(1), 182–201. https://doi.org/10.21031/epod.56218

Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics* (Cilt. 1, s. 4171–4186). Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1423

Doğruöz, E. & Kelecioğlu, H. (2024). Investigation of a multistage adaptive test based on test assembly methods. *International Journal of Assessment Tools in Education, 11*(2), 270–287. https://doi.org/10.21449/ijate.1268614

Doğuyurt, M. F. & Tan, Ş. (2023). Test eşitlemede yerel bağımsızlık varsayımının ihlalinin delta ve bootstrap eşitleme hatalarına etkisinin çeşitli değişkenlere göre incelenmesi. *Uludağ Üniversitesi Eğitim Fakültesi Dergisi, 36*(1), 23–50. https://doi.org/10.19171/uefad.1231069

Dorans, N. J. & Holland, P. W. (2000). Population invariance and the equatability of tests: Basic theory and the linear case. *Journal of Educational Measurement, 37*(4), 281–306. https://doi.org/10.1111/j.1745-3984.2000.tb01088.x

Dorans, N. J., Moses, T. P. & Eignor, D. R. (2010). Principles and practices of test score equating. *ETS Research Report Series, 2010*(2), i–41. https://doi.org/10.1002/j.2333-8504.2010.tb02236.x

Educational Testing Service. (2023a). *GRE General Test structure*. https://www.ets.org/gre/test-takers/general-test/prepare/test-structure.html

Educational Testing Service. (2023b). *POWERPREP online practice tests for the GRE General Test*. https://www.ets.org/gre/test-takers/general-test/prepare/powerprep.html

Embretson, S. & Reise, S. P. (2000). *Item response theory for psychologists*. Lawrence Erlbaum Associates.

Embretson, S. E. (2002). Generating abstract reasoning items with cognitive theory. S. Irvine & P. Kyllonen (Ed.), *Generating items for cognitive tests: Theory and practice* içinde. Erlbaum.

Emekli, E., Emekli, E., Kıyak, Y. S., Hoşgören Alıcı, Y., Coşkun, Ö. & Budakoğlu, I. İ. (2025). Psikiyatri eğitiminde klinik akıl yürütmenin değerlendirilmesi: Türkçe otomatik soru üretimi ile çoktan seçmeli soruların geliştirilmesi. *Türk Psikiyatri Dergisi, 36*(1), 336–343. https://doi.org/10.5080/u27540

Falcão, F., Costa, P. & Pêgo, J. M. (2022). Feasibility assurance: A review of automatic item generation in medical assessment. *Advances in Health Sciences Education, 27*, 405–425. https://doi.org/10.1007/s10459-022-10092-z

Finn, P. J. (1975). A question writing algorithm: The value of explicitly specified processes in test construction. *Journal of Reading Behavior, 7*(4), 341–367. https://doi.org/10.1080/10862967509547153

Gierl, M. J. & Lai, H. (2013). Evaluating the quality of medical multiple-choice items created with automated processes. *Medical Education, 47*(7), 726–733. https://doi.org/10.1111/medu.12202

Gierl, M. J. & Lai, H. (2015). Automatic item generation. *Handbook of test development* (s. 410–429) içinde. Routledge.

Gierl, M. J. & Lai, H. (2017). The role of cognitive models in automatic item generation. A. Rupp & J. Leighton (Ed.), *The Wiley handbook of cognition and assessment: Frameworks, methodologies, and applications* (s. 124–145) içinde. Wiley.

Gierl, M. J., Lai, H. & Tanygin, V. (2021). *Advanced methods in automatic item generation*. Routledge. https://doi.org/10.4324/9781003025634

Gierl, M. J., Lai, H. & Zhang, X. (2018). Automatic item generation. M. Khosrow-Pour (Ed.), *Encyclopedia of information science and technology* (s. 2369–2379) içinde. IGI Global.

González, J. & Wiberg, M. (2017). *Applying test equating methods: Using R*. Springer. https://doi.org/10.1007/978-3-319-51824-4

Gorgun, G. & Bulut, O. (2024). Exploring quality criteria and evaluation methods in automated question generation: A comprehensive survey. *Education and Information Technologies*. https://doi.org/10.1007/s10639-024-12771-3

Guttman, L. (1957). A necessary and sufficient formula for matric factoring. *Psychometrika, 22*(1), 79–81. https://doi.org/10.1007/bf02289212

Haebara, T. (1980). Equating logistic ability scales by a weighted least squares method. *Japanese Psychological Research, 22*(3), 144–149. https://doi.org/10.4992/psycholres1954.22.144

Haladyna, T. M. & Rodriguez, M. C. (2013). *Developing and validating test items*. Routledge. https://doi.org/10.4324/9780203850381

Han, K. (2019). Framework for developing multistage testing with intersectional routing for short-length tests. *Applied Psychological Measurement, 44*(2), 87–102. https://doi.org/10.1177/0146621619837226

He, W., Peng, L., Jiang, Z. & Go, A. (2025). You Only Fine-tune Once: Many-Shot In-Context Fine-Tuning for Large Language Model. *arXiv preprint arXiv:2506.11103*.

Higgins, D. (2007). *Item distiller: Text retrieval for computer-assisted test item creation*. Educational Testing Service.

Higgins, D., Futagi, Y. & Deane, P. (2005). Multilingual generalization of the ModelCreator software for math item generation. *ETS Research Report Series, 2005*(1), i–38. https://doi.org/10.1002/j.2333-8504.2005.tb01979.x

Hsu, C. L. (2024, Eylül). *Variable-length multistage testing: Precision-based termination rules in adaptive assessments* [Konferans sunumu]. International Association for Computerized Adaptive Testing (IACAT) Conference, Yonsei University, Seoul, Güney Kore.

Irvine, S. & Kyllonen, P. (2002). *Generating items for cognitive tests: Theory and practice*. Erlbaum.

İnce Aracı, F. G. & Tan, Ş. (2022). Multidimensional computerized adaptive testing simulations in R. *International Journal of Assessment Tools in Education, 9*(1), 118–137. https://doi.org/10.21449/ijate.909616

Issayeva, L. (2021). *Test score equating and linking*. ASC.

Jaroslawska, A. J., Rhodes, S., Belletier, C., Doherty, J. M., Cowan, N., Naveh-Benjamin, M., Barrouillet, P., Camos, V. & Logie, R. H. (2020). What affects the magnitude of age-related dual-task costs in working memory? The role of stimulus domain and access to semantic representations. *Quarterly Journal of Experimental Psychology, 73*(12), 1–17. https://doi.org/10.1177/1747021820970744

Jatobá, V. (2018). *Uma abordagem personalizada no processo de seleção de itens em Testes Adaptativos Computadorizados* [Yüksek Lisans Tezi, Universidade de São Paulo]. https://doi.org/10.11606/D.100.2019.tde-27012019-110739

Jatobá, V., Farias, J., Freire, V., Ruela, A. & Delgado, K. (2020). ALICAT: A customized approach to item selection process in computerized adaptive testing. *Journal of the Brazilian Computer Society, 26*. https://doi.org/10.1186/s13173-020-00098-z

Jiao, Y., Shridhar, K., Cui, P., Zhou, W. & Sachan, M. (2023). Automatic educational question generation with difficulty level controls. *International Conference on Artificial Intelligence in Education* (s. 476–488) içinde. Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-36272-9_39

Karagül, A. E. (2020). *Küçük örneklemlerde çok kategorili puanlanan maddelerden oluşan testlerde klasik test eşitleme yöntemlerinin karşılaştırılması* [Yayımlanmamış Yüksek Lisans Tezi]. Ankara Üniversitesi.

Karakaya, İ., Umur, Ö. Ç. & Gümüş, G. (2024). Üst düzey zihinsel becerilerin ölçülmesinde açık uçlu maddelerin yapay zekâ destekli uygulamalarda hazırlanması. *Journal of Applied Measurement and Assessment, 1*(2), 59–70.

Kıyak, Y. S. (2023). A ChatGPT prompt for writing case-based multiple-choice questions. *Revista Española de Educación Médica, 4*(3). https://doi.org/10.6018/edumed.587451

Kıyak, Y. S., Budakoğlu, I. İ., Coşkun, Ö. & Koyun, E. (2023). The first automatic item generation in Turkish for assessment of clinical reasoning in medical education. *Tıp Eğitimi Dünyası, 22*(66), 72–90. https://doi.org/10.25282/ted.1225814

Kıyak, Y. S. & Kononowicz, A. A. (2025). Using a hybrid of AI and template-based method in automatic item generation to create multiple-choice questions in medical education: Hybrid AIG. *JMIR Formative Research, 9*, e65726. https://doi.org/10.2196/65726

Kim, J., Chung, H., Park, R. & Dodd, B. G. (2013). A comparison of panel designs with routing methods in the multistage test with the partial credit model. *Behavior Research Methods, 45*(4), 1087–1098. https://doi.org/10.3758/s13428-013-0316-3

Kim, Y. & Yoo, H. (2024). *Multistage testing for cognitive diagnosis based on skill-space partitioning*. Advance. https://doi.org/10.31124/advance.171272698.83636340/v1

Koğar, H. & Sayın, A. (2018, Eylül). *Denk olmayan gruplarda Kernel eşitleme yönteminden elde edilen sonuçların karşılaştırılması* [Konferans sunumu]. 6th International Congress on Measurement and Evaluation in Education and Psychology, Prizren, Kosova.

Kohnke, L., Moorhouse, B. L. & Zou, D. (2023). Exploring generative artificial intelligence preparedness among university language instructors: A case study. *Computers and Education: Artificial Intelligence, 5*, 100156.

Kolen, M. J. & Brennan, R. L. (2014). *Test equating, scaling, and linking: Methods and practices* (3. baskı). Springer. https://doi.org/10.1007/978-1-4939-0317-7

Kosh, A. E., Simpson, M. A., Bickel, L., Kellogg, M. & Sanford-Moore, E. (2018). A cost–benefit analysis of automatic item generation. *Educational Measurement: Issues and Practice, 38*(1), 48–53. https://doi.org/10.1111/emip.12237

Liao, W., Ho, R., Yen, Y. & Cheng, H. (2012). The four-parameter logistic item response theory model as a robust method of estimating ability despite aberrant responses. *Social Behavior and Personality, 40*, 1679–1694. https://doi.org/10.2224/sbp.2012.40.10.1679

Loyd, B. H. & Hoover, H. D. (1980). Vertical equating using the Rasch model. *Journal of Educational Measurement, 17*(3), 179–193. https://doi.org/10.1111/j.1745-3984.1980.tb00825.x

Luecht, R. M. (2003). Multistage tests: A framework for design and analysis. *Applied Psychological Measurement, 27*(5), 341–355. https://doi.org/10.1177/014662503247343

Luecht, R. M. (2012). Automatic item generation for computerized adaptive testing. D. J. Weiss (Ed.), *Computerized adaptive testing: From theory to application* (s. 217–238) içinde. CRC Press.

Luecht, R., Brumfield, T. & Breithaupt, K. (2006). A testlet assembly design for adaptive multistage tests. *Applied Measurement in Education, 19*(3), 189–202. https://doi.org/10.1207/s15324818ame1903_2

Luo, X. & Wang, X. (2019). Dynamic multistage testing: A highly efficient and regulated adaptive testing method. *International Journal of Testing, 19*, 227–247. https://doi.org/10.1080/15305058.2019.1621871

Magis, D., Yan, D. & von Davier, A. A. (2018). Computerized adaptive and multistage testing with R: Using packages catR and mstR. *Measurement: Interdisciplinary Research and Perspectives, 16*, 1–18. https://doi.org/10.1080/15366367.2018.1520560

Mazidi, K. (2017). Automatic question generation from passages. *International Conference on Computational Linguistics and Intelligent Text Processing* (s. 655–665) içinde. Springer International Publishing.

Min, S. & Bishop, K. (2024). A shortened test is feasible: Evaluating a large-scale multistage adaptive English language assessment. *Language Testing, 41*, 627–648. https://doi.org/10.1177/02655322231225426

Morris, C. N. (1982). On the foundations of test equating. P. W. Holland & D. B. Rubin (Ed.), *Test equating* (s. 169–191) içinde. Academic Press.

Mutluer, C. (2021). *Klasik test kuramına ve madde tepki kuramına dayalı test eşitleme yöntemlerinin karşılaştırması: Uluslararası Öğrenci Değerlendirme Programı (PISA) 2012 matematik testi örneği* (Yayın No. 658052) [Doktora tezi, Gazi Üniversitesi]. YÖK Ulusal Tez Merkezi.

OECD. (2023). *PISA: Programme for International Student Assessment*. https://www.oecd.org/en/about/programmes/pisa.html

OECD. (2025a). *PISA 2025 technical standards* (EDU/PISA/GB(2022)24/REV1/FINAL). Organisation for Economic Co-operation and Development. https://one.oecd.org/document/EDU/PISA/GB%282022%2924/REV1/FINAL/en/pdf

OECD. (2025b). *PISA 2025 integrated design*. Organisation for Economic Co-operation and Development. https://one.oecd.org/document/EDU/PISA/GB%282022%2924/REV1/FINAL/en/pdf

Öztürk, N. & Anıl, D. (2012). Akademik personel ve lisansüstü eğitimi giriş sınavı puanlarının eşitlenmesi üzerine bir çalışma. *Eğitim ve Bilim, 37*(165), 180–193.

Öztürk Gübeş, N. & Kelecioğlu, H. (2015). Farklı test eşitleme yöntemlerinin eşitlik özelliği ölçütüne göre karşılaştırılması. *Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi, 6*(1), 117–125. https://doi.org/10.21031/epod.65039

Pektaş, S. & Kılınç, M. (2016). PISA 2012 matematik testlerinden iki kitapçığın gözlenen puan eşitleme yöntemleri ile eşitlenmesi. *Mehmet Akif Ersoy Üniversitesi Eğitim Fakültesi Dergisi*, *40*, 432–444. https://doi.org/10.21764/efd.49376

Roid, G. H. & Haladyna, T. M. (1978). A comparison of objective-based and modified-Bormuth item writing techniques. *Educational and Psychological Measurement, 38*(1), 19–28. https://doi.org/10.1177/001316447803800104

Roid, G. H. & Haladyna, T. M. (1982). *A technology for test-item writing*. Academic Press.

Sansivieri, V., Wiberg, M. & Matteucci, M. (2017). A review of test equating methods with a special focus on IRT-based approaches. *Statistica, 77*(4), 329–352. https://doi.org/10.6092/issn.1973-2201/7066

Sari, H. I. & Raborn, A. (2018). What information works best? A comparison of routing methods. *Applied Psychological Measurement, 42*(6), 499–515. https://doi.org/10.1177/0146621617752990

Sayın, A. (2024). Evaluating the psychometric characteristics of generated visual reading comprehension items. *Bartın University Journal of Faculty of Education, 13*(2), 380–395. https://doi.org/10.14686/buefad.1424213

Sayın, A. & Gierl, M. J. (2023a). Automatic item generation for online measurement and evaluation: Turkish literature items. *International Journal of Assessment Tools in Education, 10*(2), 218–231. https://doi.org/10.21449/ijate.1249297

Sayın, A. & Gierl, M. J. (2023b, Nisan 12–15). *Using automated item generation to provide individualized feedback in formative tests* [Konferans sunumu]. National Council on Measurement in Education (NCME) Annual Meeting, Chicago, IL, ABD.

Sayın, A. & Gierl, M. J. (2024). Using OpenAI GPT to generate reading comprehension items. *Educational Measurement: Issues and Practice, 43*(1), 5–18. https://doi.org/10.1111/emip.12590

Sayın, A. & Gierl, M. (2025, Nisan 23–26). *Generating reading comprehension items using GPT and the text analysis cognitive model* [Konferans sunumu]. National Council on Measurement in Education (NCME) Annual Meeting, Denver, CO, ABD.

Segall, D. O. (2004). Computerized adaptive testing. *Encyclopedia of social measurement* (s. 429–438) içinde. https://doi.org/10.1016/B0-12-369398-5/00444-8

Settles, B., LaFlair, G. T. & Hagiwara, M. (2020). Machine learning–driven language assessment. *Transactions of the Association for Computational Linguistics, 8*, 247–263. https://doi.org/10.1162/tacl_a_00310

Sezer Başaran, E. (2023). *Farklı ortak değişkenlerle test eşitlemenin ortak maddeli test eşitlemeyle karşılaştırılması* (Yayın No. 788504) [Doktora Tezi, Gazi Üniversitesi]. YÖK Ulusal Tez Merkezi.

Shin, E. (2021). *Automated item generation by combining the non-template and template-based approaches to generate reading inference test items* [Doktora Tezi, University of Alberta]. https://era.library.ualberta.ca/items/690b2288-07f9-4d66-adcf-59d6c4af1b7b

Sinharay, S. & Holland, P. W. (2006a). The correlation between the scores of a test and an anchor test. *ETS Research Report Series, 2006*(1), i–28.

Sinharay, S. & Holland, P. W. (2006b). Choice of anchor test in equating. *ETS Research Report Series, 2006*(2), i–43.

Singley, M. K. & Bennett, R. E. (2002). Item generation and beyond: Applications of schema theory to mathematics assessment. S. H. Irvine & P. C. Kyllonen (Ed.), *Item generation for test development* (s. 361–384) içinde. Erlbaum.

Sommer, M. & Arendasy, M. (2025). Automatic- and transformer-based automatic item generation: A critical review. *Journal of Intelligence, 13*(8), 102. https://doi.org/10.3390/jintelligence13080102

Stocking, M. L. & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement, 7*(2), 201–210. https://doi.org/10.1177/014662168300700208

Sun, L., Liu, Y. & Luo, F. (2019). Automatic generation of number series reasoning items of high difficulty. *Frontiers in Psychology, 10*. https://doi.org/10.3389/fpsyg.2019.00884

Şahin, M. (2020). Analyzing different module characteristics in computer adaptive multistage testing. *International Journal of Assessment Tools in Education*. https://doi.org/10.21449/ijate.676947

Şenel, S. (2023). *Bilgisayar ortamında bireye uyarlanmış testler* (3. baskı). Pegem Akademi.

Şenel, S. (2024). Comparison of CAT procedures at low ability levels. *Bartın University Journal of Faculty of Education, 13*(3), 547–559.

Şenel, S. & Kutlu, Ö. (2018). Comparison of two test methods for VIS: Paper-pencil test and CAT. *European Journal of Special Needs Education, 33*(5), 631–645. https://doi.org/10.1080/08856257.2017.1391014

Tan, B., Armoush, N., Mazzullo, E., Bulut, O. & Gierl, M. (2025). A review of automatic item generation techniques leveraging large language models. *International Journal of Assessment Tools in Education, 12*(2), 317–340. https://doi.org/10.21449/ijate.1602294

Tekin, H. (2019). *Eğitimde ölçme ve değerlendirme*. Yargı Yayınevi.

Thissen, D. & Mislevy, R. J. (2000). Testing algorithms. H. Wainer (Ed.), *Computerized adaptive testing: A primer* (2. baskı, s. 101–133) içinde. Lawrence Erlbaum Associates.

Uyar, Ş., Aksekioğlu, B. & Öztürk Gübeş, N. (2018). PISA 2012 matematik okuryazarlığı testinde farklı ölçek dönüştürme yöntemlerinin karşılaştırılması. *Mehmet Akif Ersoy Üniversitesi Eğitim Fakültesi Dergisi*, *46*, 121–148. https://doi.org/10.21764/maeuefd.330613

Uysal, İ. & Doğan, N. (2021). Automated essay scoring effect on test equating errors in mixed-format test. *International Journal of Assessment Tools in Education, 8*(2), 222–238. https://doi.org/10.21449/ijate.815961

van der Linden, W. J. (1999). Empirical Bayes estimation in adaptive testing. *Psychometrika, 64*(1), 41–58. https://doi.org/10.1007/BF02294531

van der Linden, W. J. & Glas, C. A. W. (2000). *Computerized adaptive testing: Theory and practice*. Springer.

van der Linden, W. J. & Glas, C. A. W. (Ed.). (2010). *Elements of adaptive testing*. Springer. https://doi.org/10.1007/978-0-387-85461-8

Vm, K., Warrier, H. & Gupta, Y. (2024). *Fine tuning LLM for enterprise: Practical guidelines and recommendations*. arXiv.

von Davier, A. A. (Ed.). (2011). *Statistical models for test equating, scaling, and linking*. Springer.

von Davier, M. (2018). Automated item generation with recurrent neural networks. *Psychometrika, 83*(4), 847–857. https://doi.org/10.1007/s11336-018-9608-y

Vu, N. & Van Nguyen, K. (2022). Enhancing Vietnamese question generation with reinforcement learning. *Intelligent information and database systems* (s. 559–570) içinde. Springer International Publishing. https://doi.org/10.1007/978-3-031-21743-2_45

Wainer, H. (Ed.). (2000). *Computerized adaptive testing: A primer* (2. baskı). Lawrence Erlbaum Associates.

Wang, S., Xiao, H. & Cohen, A. (2020). Adaptive weight estimation of latent ability: Application to computerized adaptive testing with response revision. *Journal of Educational and Behavioral Statistics, 46*, 560–591. https://doi.org/10.3102/1076998620972800

Wang, T. & Vispoel, W. (1998). Properties of ability estimation methods in computerized adaptive testing. *Journal of Educational Measurement, 35*, 109–135. https://doi.org/10.1111/j.1745-3984.1998.tb00530.x

Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika, 54*(3), 427–450. https://doi.org/10.1007/BF02294627

Weiss, D. J. (1982). Improving measurement quality and efficiency with adaptive testing. *Applied Psychological Measurement, 6*(4), 473–492. https://doi.org/10.1177/014662168200600405

Weissman, A. (2014). IRT-based multistage testing. D. Yan, A. A. von Davier & C. Lewis (Eds.), *Computerized multistage testing: Theory and applications* (s. 121–140) içinde. CRC Press.

Woldbeck, T. (1998, Nisan). *Basic concepts in modern methods of test equating* [Konferans bildirisi]. Annual Meeting of the Southwest Psychological Association, New Orleans, LA, ABD.

Yeşilyurt, Ş. (2024). *Yapay zekâ ile C1 seviyesi Türkçe okuma sınavı oluşturma ve uygulama*. Dizgi ve Grafik Tasarım Gökhan Çakır.

Yıldırım, Y., Gündüz, T. & İnce-Aracı, F. G. (2026). Madde tepki kuramına dayalı test eşitlemede ölçek dönüştürme yöntemlerinin, ortak madde oranının ve madde ayırt ediciliğinin eşitleme hatasına etkisi. *Mehmet Akif Ersoy University Journal of Education Faculty, 78*, 1-17. https://doi.org/10.21764/maeuefd.1366213

Yiğiter, M. S. & Doğan, N. (2023). Computerized multistage testing: Principles, designs and practices with R. *Measurement: Interdisciplinary Research and Perspectives, 21*(4), 254–277. https://doi.org/10.1080/15366367.2022.2158017

Zenisky, A. L., Hambleton, R. K. & Luecht, R. M. (2010). Multistage testing: Issues, designs, and research. W. J. van der Linden & C. A. W. Glas (Ed.), *Elements of adaptive testing* (s. 355–372) içinde. Springer. https://doi.org/10.1007/978-0-387-85461-8_18

Zheng, N.-n., Liu, Z.-y., Ren, P.-j., Ma, Y.-q., Chen, S.-t., Yu, S.-y., Xue, J.-r., Chen, B.-d. & Wang, F.-y. (2017). Hybrid-augmented intelligence: Collaboration and cognition. *Frontiers of Information Technology & Electronic Engineering, 18*(2), 153–179. https://doi.org/10.1631/fitee.1700053

Zhu, W. (1998). Test equating: What, why, how? *Research Quarterly for Exercise and Sport, 69*(1), 11–23. https://doi.org/10.1080/02701367.1998.10607662

YAZARLAR HAKKINDA | ABOUT AUTHORS

PROF. DR. AYFER SAYIN

Gazi Üniversitesi | ayfersayin[at]gazi.edu.tr

ORCID: 0000-0003-1357-5674

Prof. Dr. Ayfer Sayın, Gazi Üniversitesi Gazi Eğitim Fakültesi Ölçme ve Değerlendirme Ana Bilim Dalında öğretim üyesi olarak görev yapmaktadır. Hacettepe Üniversitesi Eğitimde Ölçme ve Değerlendirme Ana Bilim Dalında yüksek lisans ve doktora eğitimlerini tamamlayan Dr. Sayın; 2019 yılında doçent, 2024 yılında profesör ünvanını almıştır. 2022-2024 yılları arasında ve 2025 yılında Kanada'da bulunan Alberta Üniversitesinde misafir öğretim üyesi olarak yapay zekâ destekli projelerde görev almıştır. Dr. Sayın'ın araştırmaları; yapay zekâ tabanlı ölçme ve değerlendirme, otomatik madde üretimi, otomatik puanlama, test eşitleme, Rasch analizi ve dil becerilerinin ölçülmesi alanlarında yoğunlaşmaktadır. Bu alanlarda TÜBİTAK, ÖSYM, Dünya Bankası başta olmak üzere birçok ulusal ve uluslararası kurumda proje yürütücüsü olarak çalışmalar yürütmektedir. Ayrıca Gazi Teknopark bünyesinde faaliyet gösteren IGENAS Yapay Zekâ Yazılım LTD. ŞTİ'nin Yönetim Kurulu Başkanlığı görevini sürdürmektedir.

PROF. DR. AYFER SAYIN

Gazi University | ayfersayin[at]gazi.edu.tr

ORCID: 0000-0003-1357-5674

Prof. Dr. Ayfer Sayın is a faculty member in the Department of Educational Measurement and Evaluation at Gazi Faculty of Education, Gazi University. She completed her master's and doctoral degrees in the Department of Educational Measurement and Evaluation at Hacettepe University. She was awarded the title of Associate Professor in 2019 and promoted to Full Professor in 2024. Between 2022 and 2024, and again in 2025, she served as a visiting professor at the University of Alberta, Canada, where she participated in artificial intelligence-supported research projects. Prof. Dr. Sayın's research focuses on AI-based measurement and evaluation, automatic item generation, automated scoring, test equating, Rasch analysis, and the language assessment. She has several national and international research projects as the project manager, funded by and conducted in collaboration with organizations including the Turkish Measurement, Selection and Placement Center (known as OSYM), TUBITAK, and the World Bank. She also serves as the Chair of the Board of Directors of IGENAS Artificial Intelligence Software Ltd., a company operating within Gazi Technopark.

DOÇ. DR. SELMA ŞENEL
Balıkesir Üniversitesi | selmasenel[at]balikesir.edu.tr
ORCID: 0000-0002-5803-0793

Doç. Dr. Selma ŞENEL, Balıkesir Üniversitesi Necatibey Eğitim Fakültesi, Eğitimde Ölçme ve Değerlendirme Anabilim Dalı'nda görev yapmaktadır. Akademik çalışmalarını; bilgisayar ortamında bireye uyarlanmış test sistemleri, test düzenlemeleri, özel gereksinimli bireyler için ölçme araçlarının geliştirilmesi, ölçmede teknoloji kullanımı gibi alanlarda yoğunlaştırmaktadır. Bu kapsamda Ölçme, Seçme ve Yerleştirme Merkezi Başkanlığı başta olmak üzere farklı kurumlarla iş birliği içinde bireye uyarlanmış test sistemlerinin tasarımına yönelik projelerde yürütücülük yapmaktadır. Özel gereksinimli bireylere yönelik test uygulamaları gibi özel alanlarda da aktif olarak çalışmaktadır. Balıkesir Üniversitesi'nde çeşitli idari görevler üstlenmiştir. Ulusal ve uluslararası düzeyde yayımlanmış akademik çalışmalarıyla eğitimde kapsayıcılığı ve erişilebilirliği artırmayı hedeflemekte; eğitimcilere yönelik seminer, çalıştay ve dijital içeriklerle bilgi ve deneyimlerini paylaşmayı sürdürmektedir.

ASSOC. PROF. DR. SELMA ŞENEL
Balıkesir University | selmasenel[at]balikesir.edu.tr
ORCID: 0000-0002-5803-0793

Assoc. Prof. Dr. Selma ŞENEL works in the Department of Educational Measurement and Evaluation at Balıkesir University, Necatibey Faculty of Education. Her academic work focuses on computerized adaptive testing systems, test administration, the development of assessment tools for individuals with special needs, and the use of technology in measurement. Within this scope, she serves as project coordinator for studies on the design of adaptive testing systems, conducted in collaboration with various institutions, primarily the Measurement, Selection and Placement Center. She is also actively engaged in specialized areas such as test applications for individuals with special needs. She has held various administrative positions at Balıkesir University. Through her nationally and internationally published academic work, she aims to enhance inclusiveness and accessibility in education, and continues to share her knowledge and experience through seminars, workshops, and digital content designed for educators.

16. BÖLÜM

YETERLİKLERİN VE NORMLARIN BELİRLENMESİ

DETERMINING PROFICIENCY LEVELS AND NORMS

Doç. Dr. Ahmet Salih ŞİMŞEK

YETERLİKLERİN VE NORMLARIN BELİRLENMESİ

Doç. Dr. Ahmet Salih ŞİMŞEK
Kırşehir Ahi Evran Üniversitesi

Öz

Bu bölüm, eğitimde ve psikolojide ölçme sonuçlarının anlamlandırılmasında temel alınan yeterlik, norm ve standart belirleme kavramlarını ele almaktadır. Ölçme aracından elde edilen ham puanların kendi başına çoğu zaman yeterli bilgi sunmadığı, bu puanların ancak bir referans çerçevesi içinde yorumlandığında karar süreçleri için anlam kazandığı vurgulanmaktadır. Bu doğrultuda yeterlik, bireyin önceden tanımlanmış bilgi ve beceri setlerini ne düzeyde karşıladığını gösteren mutlak bir çerçeve olarak açıklanırken, norm bireyin performansını ait olduğu referans grubun performansıyla karşılaştıran göreli bir bilgi kaynağı olarak ele alınmaktadır. Bölümde performans düzeyi betimleyicileri, ölçmenin standart hatası, güven aralığı, kesme puanı, test eşitleme, bağlama ve dikey ölçekleme süreçleri, puanların geçerli ve tutarlı biçimde yorumlanabilmesi açısından tartışılmaktadır. PISA okuduğunu anlama düzeyleri ile Diller İçin Ortak Avrupa Başvuru Metni dil yeterlik düzeyleri, yeterliklerin basitten karmaşığa doğru nasıl yapılandırılabileceğini gösteren örnekler olarak sunulmaktadır. Normların belirlenmesi kapsamında yaş normları, sınıf düzeyi normları, büyüme normları, yerel normlar ve koşullu normlar açıklanmakta, yüzdelik, yüzdelik sıra, standart puanlar ve norm eğrileri üzerinden normatif bilginin raporlanması değerlendirilmektedir. Norm eğrilerinin üretiminde örneklem seçimi, norm grubu, ağırlıklandırma, tasarım etkisi, hata kestirimi, düzgünleştirme, LMS yöntemi ve kantil regresyonu gibi süreçlerin rolü üzerinde durulmaktadır. Bölümün son kısmında Angoff, Modifiye Angoff, Ebel, Evet/Hayır, Nedelsky, Bookmark ve Basket yöntemleri, kesme puanlarının belirlenmesinde kullanılan test merkezli standart belirleme yaklaşımları olarak incelenmektedir. Böylece bölüm, ölçme sonuçlarının yalnızca sayısal değerler olarak değil, geçerli, güvenilir, adil ve savunulabilir kararların dayanağı olarak nasıl kullanılabileceğine ilişkin kapsamlı bir çerçeve sunmaktadır.

Anahtar Kelimeler

Yeterlik, Norm, Kesme Puanı, Norm Eğrileri, Standart Belirleme

DETERMINING PROFICIENCY LEVELS AND NORMS

Assoc. Prof. Dr. Ahmet Salih ŞİMŞEK
Kırşehir Ahi Evran University

Abstract

This chapter discusses the concepts of proficiency, norms and standard setting, which constitute a central basis for interpreting measurement results in education and psychology. It emphasizes that raw scores obtained from a measurement instrument often do not provide sufficient information on their own and that such scores become meaningful for decision-making only when they are placed within a reference framework. In this respect, proficiency is explained as an absolute framework indicating the extent to which an individual meets predefined sets of knowledge and skills, while norms are treated as a relative source of information that compares an individual's performance with the performance of a reference group. The chapter addresses performance level descriptors, standard error of measurement, confidence intervals, cut scores, test equating, linking and vertical scaling as key processes for valid and consistent score interpretation. The PISA reading proficiency levels and the Common European Framework of Reference for Languages are presented as examples showing how proficiency can be structured from basic to more complex levels. Within the context of norm development, age norms, grade-level norms, growth norms, local norms and conditional norms are explained, and the reporting of normative information through percentiles, percentile ranks, standard scores and norm curves is discussed. The chapter also considers the role of sample selection, norm groups, weighting, design effects, error estimation, smoothing, the LMS method and quantile regression in the production of norm curves. In the final part, Angoff, Modified Angoff, Ebel, Yes/No, Nedelsky, Bookmark and Basket methods are examined as test-centered standard-setting approaches used to determine cut scores. Overall, the chapter provides a comprehensive framework for understanding how measurement results can be used not merely as numerical values but as a basis for valid, reliable, fair and defensible decisions.

Keywords

Proficiency, Norms, Cut Score, Norm Curves, Standard Setting

1. GİRİŞ

Eğitimde ve psikolojide gerçekleştirilen değerlendirmeler, bireylerin belirli özelliklerine ilişkin geçerliği ve güvenirliği yüksek bilgiler elde etmeyi ve bu bilgilere dayalı isabetli kararlar vermeyi amaçlar. Ancak bir ölçme aracından elde edilen ham puan, tek başına çoğu zaman yeterli bir anlam taşımaz. Puanın neyi gösterdiği, hangi düzeye karşılık geldiği ve karar sürecinde nasıl kullanılacağı, ancak belirli bir referans çerçevesi içinde açıklık kazanır. Bu nedenle ölçme sonuçlarının yorumlanması, yalnızca sayısal değerlerin raporlanması değil, bu değerlerin kuramsal ve psikometrik açıdan anlamlandırılması sürecidir.

Bir testten elde edilen puanın anlamı, öncelikle hangi amaçla değerlendirildiğine bağlıdır. Bireyin tanımlı öğrenme çıktılarını, bilgi ve beceri setlerini ya da mesleki standartları ne düzeyde karşıladığı sorgulanıyorsa yeterlik kavramı öne çıkar. Yeterlik, bireyin bir süreklilik üzerindeki konumunu gösterir ve bu konum "Temel", "Yeterli", "İleri" gibi performans düzeyleri aracılığıyla daha anlaşılır hâle getirilir. Bu düzeylerin yalnızca etiket olarak kalmaması için Performans Düzeyi Betimleyicileri önem kazanır. Bu betimleyiciler, her yeterlik düzeyindeki bireyin tipik olarak neleri bildiğini ve neleri yapabildiğini açık biçimde tanımlar.

Norm ise puanın anlamını başka bir yönden tamamlar. Bireyin performansı, ait olduğu ya da karşılaştırıldığı referans grubun dağılımı içinde ele alındığında göreli konum bilgisi elde edilir. Böylece puan, yalnızca bireyin kendi yeterlik düzeyi hakkında değil, benzerleri arasındaki yeri hakkında da bilgi verir. Bu nedenle normların belirlenmesinde norm grubunun iyi tanımlanması, örneklemin evreni temsil etmesi ve gerektiğinde ağırlıklandırma gibi istatistiksel düzeltmelerin yapılması gerekir.

Puanların yorumlanmasında gözlenen puanın hataya açık bir gösterge olduğu da göz ardı edilmemelidir. Ölçmenin standart hatası, güven aralığı, kesme puanı, test eşitleme, bağlama ve dikey ölçekleme gibi süreçler, yeterlik düzeylerinin ve normların farklı test formları, zamanlar ve gruplar arasında tutarlı biçimde yorumlanmasını sağlar. Bu bölümde yeterlik ve norm kavramları bu temel bağlamda ele alınmakta, normların belirlenmesi ve raporlanması ile standart belirleme yöntemleri ölçme sonuçlarının daha geçerli, güvenilir ve savunulabilir kararlara dönüştürülmesi açısından incelenmektedir.

2. YETERLİK VE NORM KAVRAMLARI

Bir testten elde edilen ham puan, kendi başına çoğu zaman yetersiz bir bilgi sunar. Örneğin, bir öğrencinin matematik testinden 75 puan alması ne anlama gelir? Bu puan iyi midir, yeterli midir, yoksa gelişim mi gerektirir? İşte bu sorulara yanıt verebilmek için puanın arkasındaki anlamı ortaya koyan iki temel çerçeveye ihtiyaç duyulur: yeterlik ve

norm. Yeterlik, bireyin önceden tanımlanmış bilgi ve beceri setlerini ne düzeyde karşıladığını ifade eden mutlak bir çerçeve sunar. Norm ise bireyin performansını, ait olduğu referans grubun performansıyla karşılaştırarak göreli bir bilgi sağlar.

Bir bireyin tanımlı öğrenme çıktılarını ya da mesleki standartları hangi düzeyde karşıladığını gösteren, doğrudan gözlenemeyen (gizil) ve bir süreklilik arz eden özelliktir. Ölçme ve değerlendirmede bu süreklilik, "Temel", "Yeterli", "İleri" gibi etiketlerle tanımlanan performans düzeyleri aracılığıyla somut ve anlaşılır hâle getirilir. Bu sınıflama, basit bir "geçti/kaldı" kararının ötesine geçerek, bireyin beceri yelpazesinin neresinde durduğuna dair detaylı bilgi sunar. Yeterliğin tanımlanması yalnızca teknik bir işlem değil, aynı zamanda paydaşlarla etkili bir iletişim kurma eylemidir. Karar vericiler puanın arkasındaki anlamı, adaylar güçlü ve zayıf yönlerini, uygulayıcılar ise sınıflama kararlarının dayanağını bilmek ister (AERA, APA & NCME, 2014; Kane, 2013).

Bu iletişimin temelini, Performans Düzeyi Betimleyicileri (Performance Level Descriptors, PLDs) olarak adlandırılır. PLD'ler, her bir yeterlik düzeyindeki bir bireyin tipik olarak neleri bildiğini ve neleri yapabildiğini açık, gözlemlenebilir ve sade ifadelerle tarif eder. Bu ifadeler, ilgili yeterlik düzeyini gösteren davranışlar veya performansın işevuruk tanımları olarak da düşünülebilir. Böylece düzey etiketleri soyut birer isim olmaktan çıkar ve tüm paydaşlar için ortak bir referans çerçevesine dönüşür. İyi tasarlanmış PLD'ler, farklı zamanlarda veya farklı test formlarında uygulanan ölçümler arasında tutarlı bir anlam köprüsü kurarak programların iletişim dilini standartlaştırır (Egan ve ark., 2012; Lane ve ark., 2016).

Yeterlik, ölçme kuramında gizil bir özelliktir ve bireyin konumu bir süreklilik üzerinde yer alır. Gözlenen puan bu konumun hataya açık bir izdüşümüdür ve bu nedenle yorum yapılırken ölçmenin standart hatası ile güven aralığı dikkate alınır. Bu değerler, bireyin gözlenen puanının gerçek yeterlik düzeyinden ne kadar sapabileceğini gösterir ve gerçek yeterliğin hangi puan aralığında bulunma olasılığının yüksek olduğunu belirtir. Özellikle kesme puanı çevresinde verilen kararlar ölçme hatasından kaynaklanan belirsizlikten daha fazla etkilenir ve raporun bu durumu dürüstçe göstermesi beklenir. Bu yaklaşım, puanı tek bir sayı olmaktan çıkarır ve bireyin gerçek puanının büyük olasılıkla içinde bulunduğu bir anlam aralığına dönüştürür. Sertifikasyon gibi yüksek riskli karar bağlamlarında kesme puanına çok yakın adaylar için daha temkinli davranmayı gerektiren ihtiyatlılık ilkesi öne çıkar ve sınıflama olası hatalarının sonuçları değerlendirilir. İzleme ve raporlama odaklı programlarda ise yeterlik bilgisinin yıl içi değişimi ve yıllar arası eğiliminin izlenmesi önem kazanır. Hangi amaç söz konusu olursa olsun, yeterlik sonuçlarının raporlanmasında ölçme hatasını ve olası karar sapmalarını görünür kılması bilimsel doğruluk ve kamuya hesap verebilirlik için temel kabul edilir (AERA, APA & NCME, 2014; Kane, 2013).

Yeterlik değerlendirmesinin temelinde ise ölçek puanlarının anlamının korunması yer alır. Farklı test formlarından gelen puanlar test eşitleme süreçleriyle aynı ölçeğe taşınır ve böylece formlar arası karşılaştırma hataları azaltılır. Bu sayede farklı formları alan bireylerin puanları adil bir şekilde karşılaştırılabilir. Farklı ama benzer içerikli testlerin daha esnek varsayımlarla ilişkilendirildiği durumlarda bağlama (linking) kullanılır ve yorum kapsamı dikkatli biçimde sınırlandırılır. Bağlama, testlerin farklı zamanlardaki uygulamalarını ilişkilendirmek için de kullanılabilir. Sınıf düzeyleri boyunca ortak bir gelişim seyrini izlenmek istendiğinde dikey ölçekleme (vertical scaling) tercih edilir. Bu süreçlerin amacı, belirlenen yeterlik düzeylerinin ve kesme puanlarının anlamının yıl, form ve grup değişse bile aynı kalmasını sağlamaktır (Kolen & Brennan, 2014; Lane ve ark., 2016). Böylece, örneğin bir sonraki yıl farklı bir test formu uygulansa bile, daha önce belirlenmiş olan kesme puanı kullanılarak tutarlı ve geçerli kararlar verilebilir.

Sertifikasyon gibi mutlak bir standarda ulaşmanın kritik olduğu bağlamlarda ölçüt bilgisi kararın merkezinde yer almaktadır. Alanyazın kesme puanlarının teknik kanıtlara ve bağlamsal gerekçelere birlikte dayandırılması gerektiği ve değer yargısı içeren seçimlerin şeffafça raporlanması gerektiği vurgulanmaktadır (AERA, APA & NCME, 2014; Kolen & Brennan, 2014).

Geniş ölçekli değerlendirme uygulamaları bu amaçla açık düzey tanımlarını kullanır. Aşağıda iki yerleşik referans çerçevesi üzerinden yeterlik düzeylerinin nasıl tanımlandığı listelenmektedir. Bu örnekler, farklı alanlarda yeterliğin nasıl basitten karmaşığa doğru aşamalı bir yapıda tanımlandığını göstermektedir. Uluslararası Öğrenci Değerlendirme Programı (PISA) Okuduğunu Anlama Yeterlik Düzeyleri (OECD, 2019), 15 yaşındaki öğrencilerin modern topluma katılım için gerekli temel bilgi ve becerilerini ölçer. Okuduğunu anlama alanında belirlenen yeterlik düzeyleri ve bu düzeylere karşılık gelen puan aralıkları şöyledir (Millî Eğitim Bakanlığı, 2024):

- *Düzey 6 (698 puan ve üstü): Bu düzeydeki öğrenciler, istenilen bilginin metin içerisinde saklı olduğu uzun ve soyut metinleri anlayabilir. Bilginin nasıl kullanılacağına karar vermek için çeşitli ölçütler kullanabilir, bilgilerin benzer ve zıt yönlerini karşılaştırabilir ve bu bilgileri bir araya getirebilir. Dış ölçütler kullanarak metnin kaynağı hakkında derinlemesine düşünebilir. Bilginin kaynağı ve geçerliğiyle ilgili ipuçları aracılığıyla metinler arasındaki uyumsuzlukları belirleyebilir. Metinler arası tutarsızlıkları çözebilir, farklı metinlerdeki bilgilerin benzer ve zıt yönlerini karşılaştırabilir.*
- *Düzey 5 (626–698 puan): Bu düzeydeki öğrenciler, uzun metinlerde konuyla ilgili saklı bilgileri bularak bu metinleri kavrayabilir. Kapsamlı metinlere yönelik derin bir anlayış göstererek farklı akıl yürütme yöntemlerini kullanabilir. Çeşitli metin veya kaynaklarda yer alan bilgiler ile soru arasında ilişkiyi kurarak soruları cevaplayabilir. Belli bilgilerden yola çıkarak hipotezler*

oluşturabilir veya var olan hipotezlere ilişkin değerlendirme yapabilir. Karışık ve soyut ifadelerin olduğu durumlarda gerçek ile algı arasındaki ayrımı yapabilir. Çeşitli ipuçlarından yola çıkarak verilen bilginin kaynağının veya içeriğinin objektifliğini değerlendirebilir.

- *Düzey 4 (553–626 puan): Bu düzeydeki öğrenciler, bir veya birden çok metinde yer alan uzun paragrafları anlayabilir. Metni bir bütün olarak ele alarak dile dayalı farklılıkları yorumlayabilir. Çeşitli kaynaklardan yola çıkarak farklı bakış açılarını karşılaştırabilir ve sonuç çıkarabilir. Metne yerleştirilmiş bilgileri arayabilir, bu bilgilere ulaşabilir ve bir araya getirebilir. Bilginin uygunluğunu değerlendirmek için çıkarımlar yapabilir. Metnin dikkat çekici özelliklerinden yola çıkarak yazarların kendi fikirlerini aktarmak için kullandıkları ifadeleri belirleyebilir. Çeşitli metinlerde açıkça ifade edilen iddiaları kıyaslayabilir ve kriterlere göre bilgi kaynağının güvenirliğini değerlendirebilir.*
- *Düzey 3 (480–553 puan): Bu düzeydeki öğrenciler, açık şekilde sunulmadığı durumlarda da metnin genel anlamını ifade edilebilir. Bilgileri derleyebilir, basit ve ileri düzeyde çıkarımlar yapabilir. Çok uzun olmayan metinlerin ana düşüncesini belirlemek, ilişkileri anlamak, bir kelime veya ifadenin anlamını çıkarmak için metnin farklı bölümlerindeki bilgileri bir araya getirebilir. Çeşitli kriterlere göre elde edilen bilgiler arasındaki ilişkileri anlayabilir. Bu düzeydeki öğrenciler, bir veya birkaç metin üzerine derinlemesine düşünebilir. Sunulan bilgilerden yola çıkarak farklı yazarların bakış açılarını kıyaslayabilir.*
- *Düzey 2 (407–480 puan): Bu düzeydeki öğrenciler, orta uzunluktaki metinlerin ana düşüncesini belirleyebilir. Gerekli bilgilerin açıkça verilmediği durumlarda metnin belirli bir bölümündeki ilişkileri anlayabilir ve bir bölümünden anlam çıkarabilir. Birkaç sayfalık metin içerisinde konu ile ilgili sayfayı bulabilir ve konuyla ilgili düzeylerini artırabilir. Gerekli bilgilerin açıkça verildiği durumlarda orta uzunluktaki metinlerin genel amacı ve bu metinlerdeki belirli detaylar üzerine derinlemesine düşünebilir. İddiaları karşılaştırabilir ve bu iddiaları destekleyen düşünceleri belirleyebilir.*
- *Düzey 1a (335–407 puan): Bu düzeydeki öğrenciler, cümlelerin veya kısa paragrafların gerçek anlamını kavrayabilir. Aşina oldukları konularda yazılmış metinlerin ana fikrini ve yazarın amacını anlayabilir. Metinde verilen bilgiler arasında veya metinde verilen bilgiler ile kendi sahip olduğu bilgiler arasında basit bağlantılar kurabilir. Birkaç sayfalık metinler içinde ilgili sayfayı bulabilir ve kısa metinlerde yer alan bilgilere ulaşabilir. Bilgilerin açıkça sunulması durumunda metinlerin genel amacı ve metindeki nispeten önemli bilgiler üzerinde (metnin ana fikri gibi) derinlemesine düşünebilir.*

- *Düzey 1b (262-335 puan): Bu düzeydeki öğrenciler, basit cümlelerin gerçek anlamlarını değerlendirebilir. Soru ve/veya metindeki bilgiler arasında basit bağlantılar kurarak metinlerin gerçek anlamını yorumlayabilir. Tek bir cümle, kısa bir metin veya basit bir listede istenen bilgileri tarayabilir ve bulabilir. Açık bir şekilde istendiğinde birkaç sayfalık metin içerisinde ilgili sayfayı bulabilir.*
- *Düzey 1c (189-262 puan): Bu düzeydeki öğrenciler, kısa ve basit cümlelerin anlamını kavrayabilir. Sınırlı bir süre içinde açık, basit ve somut amaçlar için okuma yapabilir.*

Diller için Ortak Avrupa Başvuru Metni (D-OABM / CEFR) Dil Yeterlik Düzeyleri (Council of Europe, 2020a) Avrupa Konseyi tarafından geliştirilen CEFR (veya Türkçe kısaltmasıyla D-OABM), dil öğrenme, öğretme ve değerlendirme için uluslararası bir standart sağlar. Bu çerçevede tanımlanan altı temel yeterlik düzeyi şöyledir (Council of Europe, 2020b):

- *C2 (Mastery): Neredeyse tüm metin tiplerini anlayabilir. Farklı sözlü ve yazılı kaynaklardan gelen bilgileri, savları ve açıklamaları yeniden oluşturarak tutarlı bir sunumda özetleyebilir. Daha karmaşık durumlarda bile daha ince anlam ayrımlarını fark ederek kendini spontane, çok akıcı ve kesin bir şekilde ifade edebilir.*
- *C1 (Effective Operational Proficiency): Çok çeşitli zorlukta, daha uzun metinleri anlayabilir ve örtük anlamın farkına varabilir. İfadeleri çok fazla aramaksızın akıcı ve spontane bir şekilde kendisini ifade edebilir. Sosyal, akademik ve mesleki amaçlar için dili esnek ve etkili bir şekilde kullanabilir. Metin düzenleme kalıplarının, bağlayıcıların ve bağdaşıklık ögelerin kontrollü kullandığını göstererek karmaşık konularda açık, iyi yapılandırılmış, ayrıntılı metinler üretebilir.*
- *B2 (Vantage): Kendi uzmanlık alanlarındaki teknik tartışmalar dâhil olmak üzere hem somut hem de soyut konulardaki karmaşık metinlerin ana fikirlerini anlayabilir. Her iki tarafı da zorlamadan, hedef dilin kullanıcıları ile mümkün olduğunca düzenli etkileşim kuracak akıcılıkla ve spontane bir biçimde etkileşime girebilir. Çok çeşitli konularda açık, ayrıntılı metinler üretebilir ve çeşitli seçeneklerin avantaj ve dezavantajlarını vererek güncel bir konu hakkında bir bakış açısını açıklayabilir.*
- *B1 (Threshold): Düzenli olarak işte, okulda, serbest zamanlarında vs. karşılaşılan aşina olduğu meselelerdeki açık standart girdilerin ana noktalarını anlayabilir. Dilin konuşulduğu bir bölgede seyahat ederken ortaya çıkması muhtemel çoğu durumla başa çıkabilir. Aşina olduğu veya kişisel ilgi alanına giren konularda basit, bağlantılı metinler üretebilir. Deneyimleri ve olayları, hayalleri, umutları ve emelleri betimleyebilir ve kısaca görüşler ve planlar için nedenler sunabilir ve açıklamalar yapabilir.*

- *A2 (Waystage): Kendisi için en yakın alakalı (ör. çok temel kişisel ve aile bilgileri, alışveriş, yerel coğrafya, istihdam) cümleleri ve sık kullanılan ifadeleri anlayabilir. Aşina olduğu ve rutin konularda basit ve doğrudan doğruya bilgi alışverişini gerektiren basit ve rutin görevlerde iletişim kurabilir. Geçmişi, yakın çevresi ve öncelikli ihtiyaç alanlarındaki meselelerinin boyutlarını basitçe betimleyebilir.*
- *A1 (Breakthrough): Somut türden ihtiyaçların karşılanmasına yönelik aşina olduğu günlük ifadeleri ve çok temel öbekleri anlayabilir ve kullanabilir. Kendini ve başkalarını tanıtabilir ve sorular sorabilir. Nerede yaşadığı, tanıdığı insanlar ve sahip olduğu şeyler gibi kişisel ayrıntılar hakkında sorular sorabilir ve sorulan soruları yanıtlayabilir. Diğer kişinin yavaş ve açık konuşması ve yardım etmeye hazır olması koşuluyla basit bir şekilde etkileşim kurabilir.*

2.1. Norm

Yeterlik bilgisini göreli bir konuma yerleştirir ve bireyin tanımlı bir referans evreninde (örneğin; ülke genelindeki tüm 8. sınıflar) nerede durduğunu gösterir. Normun belirlenmesi, iyi tanımlanmış bir norm grubu gerektirir ve bunun vazgeçilmez koşulu örneklemin evreni temsil etme gücüdür. Evrene genelleme yapabilmek amacıyla, tabakalı örnekleme gibi olasılıklı örnekleme yöntemleri tercih edilir veya daha karmaşık örnekleme tasarımlarından yararlanılır. Örneklem ile evren arasındaki olası sistematik farklar ağırlıklandırma ile istatistiksel olarak düzeltilir.

Normlar, ölçüt-dayanaklı kararın yerine geçmez; onunla birlikte değerlendirilir. Ölçüt-dayanaklı karar bir ölçütün karşılanıp karşılanmadığını söylerken, norm-dayanaklı bilgi aynı kişinin referans grubuna göre nerede durduğu bilgisini verir. İki bilgi birlikte sunulduğunda karar verme süreci daha kapsamlı ve dengeli hâle gelir. Bir aday kesme puanını karşılamış olabilir ve yine de referans grubuna göre performansı görece düşük veya yüksek olabilir. Bu nedenle geniş ölçekli değerlendirme programları, raporlarında genellikle bu iki bilgiyi birlikte verir ve kullanıcıya tek bakışta hem standardı karşılama durumunu hem de göreli konumu sunar (AERA, APA & NCME, 2014; Lane ve ark., 2016).

3. NORMLARIN BELİRLENMESİ

Bir önceki bölümde tanımlandığı gibi normlar, bir puanın belirli bir referans grubuna göre konumunu gösteren karşılaştırma ölçüleridir. Uygulamada normlar birkaç temel eksen boyunca sınıflandırılır. Zaman ekseninde aynı anda toplanan verilere dayanan enine kesit normları ve zaman içindeki ilerlemeyi temel alan büyüme normları kullanılır. Kapsam ekseninde tüm evreni temsil etmeyi amaçlayan ulusal normlar ve belirli bir kurum ya da bölge için hesaplanan yerel normlar tercih edilir. Koşullandırma ekseninde herhangi bir değişkene bağlı olmayan koşulsuz normlar ile geçmiş puan ya da dil durumu gibi değişkenlere bağlı koşullu normlar üretilir. Raporlama ekseninde göreli konumu

doğrudan veren yüzdelikler ile ortak ölçeklere dönüştürülmüş z, T ve stanin gibi standart puanlar yer alır. Modelleme ekseninde ampirik tablolarla sunulan normlar ile dağılımı yumuşatılmış ve gelişimsel süreklilik üzerinde üreten yöntemler bulunur. Bu bütünün güvenilir çalışması üç koşula bağlıdır:

(1) Örneklemin evreni temsil etmesi,
(2) Farklı uygulama dönemleri ve test formlarından kaynaklanan farkların test eşitleme ile yönetilmesi,
(3) Uygulanan dönüşümlerin ve hesaplanan belirsizlik ölçülerinin (örneğin; standart hatalar) raporda açıkça yazılması.

Alanyazın bu ilkeleri temel kabul eder ve normların üretimi ile kullanımı bu çerçeve içinde gerekçelendirilir (AERA, APA & NCME, 2014; Angoff, 1971; Kolen & Brennan, 2014; Lane, Raymond & Haladyna, 2016).

3.1. Temel Norm Grupları ve Karşılaştırma Çerçeveleri

3.1.1. Yaş Normları

Yaş normu, aynı takvim yaşına sahip bireylerin puan dağılımını referans alır. Yorumlama, aynı yaştaki bireyler arasında konum üzerinden yapılır. Geçerli bir yaş normu için evrenin önemli alt gruplarını kapsayan tabakalı örnekleme gereklidir. Örneklem ile evren arasındaki farklılıklar ağırlıklandırma ile düzeltilir. Uygulama zaman aralığı açıkça yazılır çünkü mevsimsellik ve farklı öğrenme fırsatları aynı yaş içindeki puanları etkileyebilir. Raporlarda standart sapma ve hata ölçümleri verilerek örneklem tasarımının varyans üzerindeki etkisi olan tasarım etkisi sunulur. Sınıf temelli öğretim döngülerinin baskın olduğu bağlamlarda yaş ile sınıf bilgisi birlikte sunulur ve böylelikle kavramsal karışıklık önlenir. Bu düzen, yaşa duyarlı gelişim yorumunu anlaşılır ve denetlenebilir bir biçimde sağlayarak kullanıcıların yanlış genelleme yapmalarının önüne geçer (AERA, APA & NCME, 2014; Lane ve ark., 2016).

3.1.2. Sınıf Düzeyi Normları

Sınıf düzeyi normu, aynı sınıfta öğrenim gören öğrencilerin dağılımını temel alır. Yorumlama, aynı sınıftaki öğrenciler arasında konum üzerinden yapılır. Okul yılı içinde öğrenme fırsatları değiştiği için programlarda çoğunlukla sonbahar, kış ve bahar dönemleri için ayrı normlar hesaplanır. Yıllar arası karşılaştırma yapılacaksa test formları ve içerik farklılıkları test eşitleme ile yönetilir. Test eşitleme yapılmadan yıllar arası sınıf normu karşılaştırmaları yanlılığa neden olarak hatayı artırır. Raporlarda dönem bilgisi ve test eşitleme süreci açıkça verilir ve normun, karar verme sürecinde ölçüt-dayanaklı bilgiyi ikame etmediği net biçimde belirtilir. Sınıf normu, göreli konumu tamamlayıcı bir katman olarak yer alır ve aynı tabloda ölçüte göre erişim ve referans gruba göre konum birlikte sunulur (AERA, APA & NCME, 2014; Kolen & Brennan, 2014; Lane ve ark., 2016).

3.1.3. Büyüme Normları

Büyüme normu, bireyin iki zaman noktası arasındaki ilerlemesini, benzer geçmiş puan profiline sahip diğer bireylerin gösterdiği tipik ilerlemeyle karşılaştırır. En yaygın yaklaşım öğrenci büyüme yüzdelikleri (SGP - Student Growth Percentiles) yöntemidir. Bu yöntemde, önce bireyin geçmiş puanlarına göre benzer başlangıç noktasına sahip diğer bireylerden oluşan koşullu bir karşılaştırma grubu tanımlanır. Ardından bireyin yeni puanı bu koşullu dağılımda yüzdelik sıra olarak konumlandırılır. Farklı başlangıç seviyelerinden gelen öğrenciler bu sayede daha adil bir referansla kıyaslanır. Sonuçların güvenirliği için yeterli örneklem büyüklüğü, sağlam bir test eşitleme zinciri ve geçmiş ölçümlerden elde edilen puanların geçerliği gerekir. Raporlarda belirsizlik ve dönemler arası ölçek tutarlılığı açıkça gösterilir. Büyüme normu, ölçüt-dayanaklı yeterlik çizgilerinin yerine geçmez; ilerleme hızına ilişkin ek bir bilgi sağlar ve program hedeflerine iletişimini güçlendirir (AERA, APA & NCME, 2014; Angoff, 1971; Kolen & Brennan, 2014; Lane ve ark., 2016).

3.1.4. Yerel ve Koşullu Normlar

Yerel norm; belirli bir il, ilçe ya da kurum gibi dar bir kapsam için üretilen puan dağılımlarına dayanır. Ulusal evrenden anlamlı biçimde farklılaşan bağlamlarda daha uygun bir referans sunar. Güvenirliği, örneklem büyüklüğüne ve normların güncelleme sıklığına bağlıdır. Raporlarda normun geçerli olduğu kapsam sınırı açıkça yazılır ve standart hata değerleri verilir. Yerel ve ulusal normların birlikte sunulması yanlış genelleme riskini azaltır ve kararın kapsamını netleştirir. Kullanıcı aynı tabloda bireyin yerel bağlamdaki konumu ile ulusal düzeydeki konumu karşılaştırmalı olarak görebilir (AERA, APA & NCME, 2014; Dorans & Cook, 2016).

Koşullu norm, referans grubunu belirli değişkenlere göre daraltarak oluşturulan dağılımlar üzerinden üretilir. En yaygın kullanım, bireyin geçmiş puan profillerine göre koşullandırmadır. Dil durumu, özel eğitim gereksinimi ya da sosyoekonomik değişken gibi ek koşullar kuramsal gerekçeyle modele alınabilir. Ancak model karmaşıklaştıkça, her bir koşula uyan birey sayısı (hücre büyüklükleri) küçülür ve yapılan kestirimlerin standart hatası artar. Bu nedenle değişken seçimi dikkatle yapılır ve her bir alt grup için örneklem yeterliği sağlanır. Koşullu normlar, belirli gruplar için adil bir karşılaştırma algısını güçlendirir ve ölçüt-dayanaklı kararların yanında ayrı bir bilgilendirme kaynağı olarak kullanılır (AERA, APA & NCME, 2014; Betebenner, 2009). Örneğin, anadili Türkçe olmayan öğrenciler için hem genel sınıf normu hem de benzer dil geçmişine sahip öğrencilerden oluşan koşullu norm sunulabilir.

3.2. Normların Raporlanması: Yüzdelikler ve Standart Puanlar

Yüzdelik ve yüzdelik sıra, göreli konumu en doğrudan anlatan göstergelerden biridir. Yüzdelik, puan dağılımı belirli bir yüzdeye bölen puan değeridir. Örneğin, 75. yüzdelik, puanların %75'inin altında kaldığı puandır. Yüzdelik sıra, bireyin puanından daha düşük

puan alanların yüzdesidir. Örneğin 70. yüzdelik sırada olan bir birey, gruptakilerin %70'inden daha yüksek puan almıştır. Bu iki kavram karıştırılmamalıdır. Özellikle dağılımın uçlarında (en düşük ve en yüksek puanlar civarında) küçük puan farklarının büyük yüzdelik sıra değişimlerine yol açabileceği unutulmamalıdır. Bu nedenle raporlarda yüzdelikler çoğu zaman standart puanlarla birlikte verilir ve yüzdelik sıraların yorumlanmasında standart hata ve güven aralığı gibi belirsizlik göstergeleri eklenir (Angoff, 1971; Livingston & Zieky, 1982; Lane ve ark., 2016).

Standart puanlar; ham puanları, ortalaması ve standart sapması önceden tanımlı ortak ölçeklere taşır. z-puanı (ortalama=0, SS=1), T-puanı (ortalama=50, SS=10) ve stanin (1-9 arası dokuz kategori) en yaygın olanlarıdır. Standart puanların en sağlıklı kullanımı test eşitleme sonrasıdır. Çünkü önce farklı test formlarından kaynaklanan farklar giderilir ve sonra ortak ölçek güvenle uygulanır. Raporlarda hangi dönüşümlerin ve bu dönüşümlerde kullanılan parametrelerin (örneğin; ortalama, standart sapma) hangi örneklem grubuna dayalı olarak hesaplandığının açıkça yazılması gerekir (Kolen & Brennan, 2014).

Bu bölümde normların ne olduğu ve hangi farklı amaçlara hizmet ettiği genel bir çerçevede ele alınmıştır. Ancak normların üretimi ve yorumlanması, ölçülen özelliğin doğasına göre önemli farklılıklar gösterebilir. Ayrıca, normatif verinin etkili bir şekilde sunulması için norm eğrileri gibi görselleştirme araçları ve bu eğrilerin altında yatan modelleme teknikleri de önem taşımaktadır. Bir sonraki bölümde, öncelikle norm çalışmasının temelini oluşturan örneklem seçimi ve norm grubu oluşturma süreci ayrıntılı olarak incelenmiş, ardından norm eğrileri ve normların yaşam döngüsü ele alınmış ve son olarak normların bilişsel yeterlikler ve duyuşsal özellikler bağlamında nasıl özelleştiğini, ne gibi ek dikkat noktaları gerektirdiği incelenmiştir.

3.3. Norm Eğrileri, Modelleme ve Normların Geçerlik Süresi

Ölçme ve değerlendirmede elde edilen sayısal veriler, ham formlarıyla genellikle karmaşık ve yorumlanması zordur. Normatif veriyi, yani bir bireyin tanımlı bir referans grubuna göre konumunu, tüm paydaşlar için anlaşılır ve eyleme geçirilebilir bir formata dönüştürmenin en etkili yollarından biri görselleştirmedir. Norm eğrileri, bu görselleştirmenin en güçlü araçlarından biridir. Seçilen bir referans grubunda belirli yüzdeliklerin (örneğin; P10, P50 ve P90) zaman, yaş ya da sınıf düzeyi boyunca nasıl bir gelişim seyri izlediğini gösteren çizgilerdir.

Bu eğriler, basit birer grafikten çok daha fazlasını ifade eder; karmaşık psikometrik ve istatistiksel süreçlerin sonucudur. Güvenilir bir norm eğrisi; temsil gücü yüksek bir örnekleme, farklı test formları arasında tutarlılığı sağlayan test eşitleme süreçlerine, örneklem tasarımının etkisini düzelten ağırlıklandırma işlemlerine ve bu karmaşık yapıdan kaynaklanan hatayı doğru tahmin eden varyans kestirimi yöntemlerine dayanır. Bu nedenle AERA, APA ve NCME (2014) standartları, normları ve dolayısıyla bu

normlardan türetilen eğrileri, altında yatan tüm bu teknik altyapının şeffaf bir şekilde belgelenmesi gerektiğini vurgular. Bu alt bölüm, norm eğrilerinin üretim adımlarını, yorumlama ilkelerini, gelişmiş modelleme tekniklerini ve geçerlik süresi içindeki yönetimini ayrıntılı olarak ele almaktadır (Kolen & Brennan, 2014; Lane, Raymond & Haladyna, 2016).

3.3.1. Norm Eğrilerinin Üretim Aşamaları ve Modelleme Teknikleri

Norm eğrilerinin üretimi, titizlik gerektiren üç temel adımda gerçekleşir. Her adım, nihai görselin geçerliği ve güvenirliği üzerinde doğrudan etkiye sahiptir.

Adım 1: Verinin Hazırlanması ve Psikometrik Temellerin Atılması: Bu aşama, eğrinin üzerine inşa edileceği zemini oluşturur. İlk olarak, norm grubunun hedef evreni temsil etmesi için hesaplanan örneklem ağırlıkları veriye uygulanır. Ardından, programda farklı test formları kullanılıyorsa veya veriler farklı zaman dilimlerinde toplanmışsa, bu farklılıklardan kaynaklanan yanlılığı ortadan kaldırmak için test eşitleme işlemleri uygulanarak tüm puanlar ortak bir ölçeğe getirilir. Son olarak, tabakalı veya çok aşamalı örnekleme gibi karmaşık tasarımların standart hata hesaplamaları üzerindeki etkisini (tasarım etkisi - DEFF) hesaba katmak için çoğaltma ağırlıkları (replicate weights) gibi gelişmiş varyans kestirim araçları hazırlanır (Rust & Rao, 1996). Bu adımlar atlanarak çizilen bir eğri, teknik olarak yanıltıcı olacaktır.

Adım 2: Yüzdeliklerin Kestirimi ve Eğrilerin Düzgünleştirilmesi: Veri hazırlandıktan sonra, her bir yaş veya sınıf düzeyi için ilgili yüzdelik dilimleri (örneğin; P10, P25, P50, P75, P90) kestirilir. Ancak gözlenen (ham) yüzdelikler, örneklemden kaynaklanan rastgele dalgalanmalar içerebilir. Bu gürültüyü (rasgele değişkenliği) azaltmak ve altta yatan gerçek gelişimsel eğilimi ortaya çıkarmak için istatistiksel düzgünleştirme (smoothing) yöntemleri kullanılır. Bu yöntemler, verideki ana örüntüyü korurken küçük ve anlamsız dalgalanmaları ortadan kaldırmayı hedefler. Bu alanda iki yaklaşım öne çıkmaktadır:

1. LMS Yöntemi (Box-Cox Cole-Green): Özellikle sağlık ve gelişim alanlarında yaygın olarak kullanılan bu yöntem, bir puan dağılımını üç temel parametre ile matematiksel olarak modeller:
    - L (Lambda): Box-Cox güç dönüşümü ile dağılımın çarpıklığını düzelterek normale yaklaştırır.
    - M (Mu): Dağılımın medyanını temsil ederek gelişim çizgisinin merkezini belirler.
    - S (Sigma): Dağılımın göreli yayılımını (varyasyon katsayısını) temsil ederek puanların zamanla nasıl farklılaştığını tanımlar. Bu üç parametre, zaman veya yaş ekseni boyunca düzgün eğriler olarak modellenir ve istenen tüm yüzdelikler bu modellenmiş yapıdan matematiksel olarak türetilir. Bu sayede, P10 ve P90 gibi çizgilerin birbiriyle tutarlı ve monoton bir yapı içinde olması sağlanır (Cole & Green, 1992).

2. Kantil Regresyonu (Quantile Regression): Geleneksel regresyonun sadece ortalamayı modellemesinin aksine, kantil regresyonu doğrudan belirli yüzdeliklerin (kantillerin) kendisini bağımsız değişkenlere (örneğin; yaş, sınıf düzeyi) bağlı olarak modeller. Örneğin, P10, P50 ve P90 için ayrı ayrı regresyon eğrileri uydurur. Bu yaklaşım, bağımsız değişkenlerin dağılımın farklı bölgeleri üzerindeki etkilerinin farklı olabildiği durumlarda ve örneklem büyüklüğünün azaldığı (veri seyrekleşmesinin olduğu) uç bölgelerde daha esnek ve sağlam sonuçlar verebilir (Koenker, 2005).

Adım 3: Görselleştirme ve Belirsizliğin Raporlanması: Son aşamada, düzgünleştirilmiş yüzdelik çizgileri grafiğe aktarılır. Ancak bu çizgiler kesin değerleri göstermez. Örneklemden kaynaklanan kestirim hatası içerirler. Bu belirsizliği göstermek için eğrilerin etrafına güven bantları (confidence bands) eklenir. Bu bantlar, çoğaltma ağırlıkları gibi gelişmiş yöntemlerle hesaplanan standart hatalara dayanır ve çizilen eğrinin gerçek evren değerini (örneğin; evrendeki gerçek P50 değerini) hangi olasılıkla içerdiğini gösterir. Eksen etiketleri, ölçek birimleri ve veri toplama dönemi gibi bilgiler grafikte açıkça belirtilerek, eğrinin tam ve şeffaf bir istatistiksel özete dönüşmesi sağlanır.

3.3.2. Norm Eğrilerinin Yorumlanması: Statik Konum ve Büyüme

Norm eğrileri, doğru okunduğunda zengin bir yorumlama çerçevesi sunar. Ancak olası karar hatalarını azaltmak için belirli ilkeler izlenmelidir.

3.3.2.1. Anlık (Statik) Konumun Yorumlanması

Bir öğrencinin puanı, ilgili yaş veya sınıf eğrilerinin üzerinde bir nokta olarak işaretlenir. Bu noktanın konumu birkaç temel bilgiyi verir:

- Merkeze Göre Konum: Noktanın P50 (medyan) çizgisine göre konumu, bireyin referans grubunun ortalamasının altında mı, çevresinde mi, yoksa üstünde mi olduğunu gösterir.
- Dağılımdaki Yer: Noktanın P10 ve P90 gibi uç yüzdelik çizgilerine yakınlığı, bireyin performansının referans grubundaki diğer bireylere göre ne kadar farklılaştığını görselleştirir.
- Ölçek Etkisi: Okuyucu, yüzdelik ölçeğinin doğrusal olmadığını bilmelidir. Puan dağılımın merkezinde (örneğin P45 ile P55 arası) küçük puan farkları gözlenirken, uç bölgelerde (örneğin P90 ile P95 arası) aynı yüzdelik değişimi çok daha büyük bir puan farkına karşılık gelebilir. Bu nedenle, eğriler tek başına bireysel düzeyde kesin sınıflama kararları (örneğin "üstün yetenekli" veya "desteklenmesi gereken") için kullanılmamalıdır. Bu tür kararlar için öğrencinin puanı, ölçmenin standart hatası ve güven aralığı ile birlikte değerlendirilmelidir.

3.3.2.2. Büyüme Bağlamında Yorumlama

Norm eğrileri sadece anlık durumu değil, aynı zamanda ilerlemeyi de yorumlamaya yardımcı olabilir. Bu bağlamda iki tür kullanım vardır:

1. Enine Kesit Eğrileri ile Eğilim (Trend) Analizi: Bir öğrencinin farklı zamanlardaki (örneğin; 3, 4 ve 5. sınıf bahar dönemi) puanları aynı grafikte işaretlenerek, referans grubuna göre göreli konumunun zaman içinde nasıl değiştiği (örneğin; P50'den P65'e yükselmesi) izlenebilir.
2. Koşullu Büyüme Grafikleri: Bu daha gelişmiş yaklaşım, öğrencinin ilerlemesini "akademik akranlarına" (benzer başlangıç noktasına sahip diğer öğrenciler) göre konumlandırır. Grafiklerde genellikle bir önceki yılın puanı yatay eksene, bu yılın puanı dikey eksene yerleştirilir. Diyagonal çizgiler ise büyüme yüzdeliklerini temsil eder. Öğrencinin (geçen yılki puan, bu yılki puan) koordinatıyla belirlenen noktası, bu grafikteki yerine göre yorumlanır. Bu yorum, başlangıç düzeyi düşük bir öğrencinin gösterdiği yüksek ilerlemeyi veya başlangıç düzeyi yüksek bir öğrencinin beklenen düzeyde ilerleme gösteremediğini adil bir referansla ortaya koyar (Betebenner, 2009).

3.3.3. Sık Görülen Hatalar ve Alınabilecek Önlemler

Standart belirleme sürecinde sık görülen hatalar ve bunlara yönelik alınabilecek önlemler aşağıda sunulmuştur:

- Hata: Test eşitleme yapılmamış farklı formların veya dönemlerin eğrilerini doğrudan karşılaştırmak.

  Önlem: Yorum yapmadan önce tüm puanların ortak bir ölçekte olduğundan emin olunmalıdır.

- Hata: Örneklem ağırlıklarını ve karmaşık tasarım etkisini yok sayarak standart hataları olduğundan küçük hesaplamak.

  Önlem: Varyans kestirimi, çoğaltma ağırlıkları gibi tasarıma uygun yöntemlerle yapılmalıdır.

- Hata: Yüzdelik (bir puan) ile yüzdelik sırayı (bir yüzde) karıştırmak.

  Önlem: Raporlarda bu iki kavramın tanımını içeren kısa bir bilgi kutusu sunulmalıdır.

- Hata: Bireysel kararlarda eğrinin etrafındaki belirsizlik bantlarını veya puanın güven aralığını göz ardı etmek.

  Önlem: Bireysel raporlarda puanla birlikte mutlaka SEM veya güven aralığı bilgisi verilmelidir.

- Hata: Güncellenmiş yeni normlarla üretilen bir eğriyi, eski normlarla üretilmiş bir eğriyle doğrudan kıyaslamak.

  Önlem: Norm geçişlerinde, iki norm seti arasındaki etkiyi gösteren köprüleme tabloları sunulmalıdır.

3.3.4. Normların Geçerlik Süresi: Raporlama, Güncelleme ve Etik Kullanım

Norm setleri statik belgeler değildir. Üretildikleri, kullanıldıkları, zamanla geçerliklerini yitirdikleri ve yenilendikleri bir süreçleri vardır. Bu sürecin her aşamasında şeffaflık ve etik sorumluluk esastır. Raporlama tasarımı ve hedef kitle norm bilgilerinin ve eğrilerinin sunumu, hedef kitleye göre uyarlanmalıdır.

- Öğretmen ve Aday Odaklı Raporlar: Genellikle P10, P50 ve P90 gibi temel referans çizgilerini, bireyin puanının işaretlendiği noktayı ve sonucun ne anlama geldiğini sade bir dille açıklayan kısa bir yorum kutucuğunu içeren tek sayfalık özetler tercih edilir.
- Politika ve Araştırma Odaklı Raporlar: Bu raporlarda yıllar arası eğilim eğrileri, ölçek tutarlılığına ilişkin teknik kanıtlar ve norm değişikliklerinin etkisini analiz eden tablolar birlikte sunulur. Eğer bir programda hem ulusal hem de yerel normlar mevcutsa, iki referans çizgisinin aynı grafikte farklı renk veya stillerle gösterilmesi ve kapsam farkının açıkça belirtilmesi, yorumun zenginleşmesini sağlar.

Uluslararası İyi Uygulamalar ve Şeffaflık: TIMSS, PIRLS, NAEP ve PISA gibi geniş ölçekli uluslararası değerlendirmeler, normların ve eğrilerin nasıl belgelenmesi gerektiğine dair iyi örnekler sunar. Bu programların teknik raporları, kullanılan örnekleme ve test eşitleme yöntemlerini, ağırlıklandırma prosedürlerini ve varyans kestirimi için kullanılan çoğaltma ağırlıklarını ayrıntılı olarak yayımlar. Bu şeffaflık, araştırmacıların ve politika yapıcıların sonuçları yeniden üretebilmesini veya ikincil analizler yapabilmesini sağlar. Kullanıcı, gördüğü bir çizginin hangi karmaşık yöntem zincirinin bir ürünü olduğunu doğrudan kaynaklardan doğrulayabildiğinde, programın kamu nezdindeki güvenirliği artar (Martin, Mullis & Hooper, 2020; NCES, 2022; OECD, 2023).

Güncelleme ve Sürdürülebilirlik: Hiçbir norm seti sonsuza dek geçerli değildir. Müfredattaki değişiklikler, teknolojik gelişmeler ve öğrenci nüfusunun demografik yapısındaki değişimler ("Flynn etkisi" gibi) zamanla normların eskimesine neden olur. Bu nedenle, ölçme programları normlarını genellikle 5 ila 10 yıl gibi periyotlarla planlı bir şekilde güncellemelidir. Normlar yenilendiğinde, eski ve yeni normlar arasındaki farkın sınıflama oranları üzerindeki etkisi analiz edilmeli ve kullanıcılara bu geçişin pratik sonuçları hakkında net bilgi verilmelidir.

Tamamlayıcı bir yorumlama aracı norm eğrileri, teknik titizlikle üretilip doğru yorumlandığında, geniş ölçekli test sonuçlarını paydaşlar için güçlü bir referans aracına çevirir. Bir öğrencinin mutlak bir standardı (ölçüt-dayanaklı yorum) karşılayıp karşılamadığını gösteren bilginin yanına, referans grubuna göre göreli konumunu ve

gelişim seyrini (norm-dayanaklı yorum) ekleyerek bütüncül bir tablo sunarlar. Eğriler, ölçütlerin yerini almaz; onları tamamlarlar. Bu ikili bakış açısı, eğitimde veriye dayalı karar verme süreçlerinin hem daha adil hem de daha bilgilendirici olmasını sağlar.

3.4. Örneklem Seçimi ve Norm Grubu

Önceki bölümlerde normların, bir bireyin performansını tanımlı bir referans grubuna göre anlamlandırmamızı sağlayan güçlü bir araç olduğunu gördük. Ancak bu karşılaştırmanın geçerliği ve adaleti, referans noktası olarak kullanılan norm grubunun ne kadar bilimsel ve titiz bir süreçle oluşturulduğuna bağlıdır. İşte bu süreç, örneklem seçimi ve norm grubu oluşturma aşamasıdır. Norm grubu, bir testin sonuçlarının genelleneceği hedef evreni (örneğin; ülkedeki tüm 8. sınıf öğrencileri) temsil eden ve normların üretildiği örneklemdir ve bu grubun kalitesi, norm tabanlı tüm yorumların temelini oluşturur. Bu alt bölüm, normların doğruluğunu ve adaletini belirleyen bu temel sürecin nasıl yürütüldüğünü adım adım ele almaktadır. Temsil gücü yüksek bir örneklem tasarımı, gerekli istatistiksel düzeltmeler ve şeffaf bir raporlama zinciri olmadan, en gelişmiş norm tabloları veya eğrileri bile yanıltıcı sonuçlar üretebilir.

3.4.1. Örnekleme Tasarımının Temelleri

Güvenilir bir norm grubu oluşturma süreci üç temel adımla başlar:

1. Hedef Evrenin Tanımlanması: İlk adım, normların genelleneceği evrenin sınırlarını net bir şekilde çizmektir. Hangi yaş aralığı, sınıf düzeyi, coğrafi bölge ya da kurum türü kapsamda olacaksa bu sınır tartışmaya yer bırakmayacak şekilde yazılacaktır.
2. Örnekleme Çerçevesinin Seçimi: İkinci adım, tanımlanan hedef evrendeki örnekleme birimlerine (örneğin; okullar, öğrenciler) ulaşılabilen güncel ve eksiksiz bir liste olan örnekleme çerçevesini belirlemektir. Çerçevenin hedef evrenin bir kısmını kapsamadığı durumlar (örneğin; listeye dâhil olmayan okul türleri) kapsam hatası (coverage error) yaratacaktır. Bu hatanın boyutu ve olası etkileri teknik raporda şeffafça belirtilmelidir.
3. Örnekleme Tasarımının Belirlenmesi: Üçüncü ve en kritik adım, evrenden norm grubunu seçecek yöntemin tasarlanmasıdır. Geniş ölçekli testlerde lojistik kolaylık ve maliyet etkinliği nedeniyle standart uygulama, tabakalı ile çok aşamalı örneklemenin birleşimidir. Genellikle okullar birincil örnekleme birimi olarak seçilirken, öğrenciler ikinci aşamada bu okulların içinden genellikle basit rastgele yöntemle örneklenir.

Unutulmamalıdır ki, bir norm çalışmasının evreni temsil etme gücü, en başta yapılan bu tasarım ve tanım kararlarının kalitesine bağlıdır. Zayıf bir çerçeve veya hatalı bir tasarımla yola çıkıldığında, hiçbir ileri istatistiksel yöntem bu başlangıçtaki yanlılığı

sonradan tam olarak düzeltemez. Sağlam bir çerçeve, dikkatli bir tabakalama ve her birimin bilinen ve sıfırdan farklı bir seçilme olasılığına sahip olması, normların güvenirliğinin temelini oluşturur (AERA, APA & NCME, 2014; Kish, 1965; Lohr, 2019).

Tabakalı örnekleme, evreni coğrafya, okul türü veya sosyoekonomik düzey gibi politika açısından anlamlı alt gruplara (tabakalara) ayırarak, bu grupların örneklemde orantılı veya hedeflenen sayıda temsilini sağlar. Gerekli durumlarda, analizler için yeterli sayıda veri elde etmek amacıyla küçük ama önemli alt grupların örneklemdeki oranı artırılabilir (aşırı örnekleme - oversampling) yapılabilir. Çok aşamalı örnekleme düzeninde ise okullar genellikle öğrenci sayıları gibi büyüklükleriyle orantılı olasılıkla seçilir ve içlerinden öğrenciler belirlenen yöntemle (örneğin; basit rastgele) basit rastgele yöntemle belirlenir. Bu yapısal yaklaşım, saha uygulamalarını büyük ölçüde kolaylaştırır (Heeringa ve ark., 2017).

3.4.2. Ağırlıklandırma ve Hata Kestirimi

Örneklem seçildikten sonra bile, farklı seçilme olasılıkları veya yanıtlama oranlarındaki farklılıklar nedeniyle örneklemin evreni birebir yansıtması nadiren mümkündür. Bu noktada, istatistiksel düzeltme mekanizmaları devreye girer.

Örnekleme Ağırlıkları (Sampling Weights) Ağırlıklar, örneklemdeki her birimin evrendeki temsil payını ayarlayarak, örneklemden elde edilen sonuçların evrene doğru bir şekilde genellenmesini sağlar. Ağırlıklandırma süreci genellikle üç basamakta kurulur:

1. Taban Ağırlığı: Her birimin örnekleme dahil edilme olasılığının matematiksel tersidir. Farklı olasılıklarla seçilen birimlerin etkisini eşitler.
2. Yanıtlamama Düzeltmesi: Seçilen ancak araştırmaya katılmayan okul veya öğrencilerin yarattığı sistematik hatayı azaltmayı hedefler. Katılmayanların özelliklerine benzer katılımcıların ağırlığı artırılır.
3. Son Hizalama: Örneklemin demografik dağılımını, cinsiyet, bölge veya okul türü gibi bilinen evren toplamlarına yakınlaştırmak için yapılan son ayardır.

Ağırlıklandırma, yanlılığı azaltsa da yapılan istatistiksel kestirimlerin varyansını genellikle artırır. Bu artış, tasarım etkisi adı verilen bir katsayı ile özetlenir. Tasarım etkisi, karmaşık örneklem tasarımından (örneğin; tabakalı, aşamalı) ve ağırlıklandırmadan kaynaklanan varyans artışını, aynı büyüklükteki basit rastgele örneklemeye kıyasla gösterir. Tasarım etkisi değeri olan DEFF'in 1'den büyük olması, etkin örneklem büyüklüğünün (hesaplamalarda kullanılan istatistiksel gücü yansıtan sanal büyüklük) nominal (gerçek) büyüklükten daha düşük olmasına neden olur.

Varyans Kestirimi (Variance Estimation) Karmaşık örneklem tasarımlarında standart hata ve güven aralıkları, basit rastgele örnekleme varsayımına dayanan standart istatistik formülleriyle hesaplanamaz. Bu, hatanın olduğundan küçük tahmin edilmesine

yol açar. Doğru varyans kestirimi için iki yaygın yaklaşım kullanılır: Taylor serisi doğrusallaştırma (varyans formülünü karmaşık tasarıma uyarlayan analitik bir yöntem) veya çoğaltma ağırlıkları (*replicate weights*). Jackknife ve Dengeli Tekrarlı Çoğaltma (BRR - Balanced Repeated Replication)gibi çoğaltma yöntemleri, ana örneklemden çok sayıda alt örneklem (çoğaltma) yaratarak ilgilendiğimiz istatistiğin (örneğin; ortalama, yüzdelik) bu alt örneklemlerdeki dağılımını inceler ve bu dağılımdan doğru standart hatayı hesaplar. . PISA ve NAEP gibi büyük programlar, araştırmacıların doğru hata kestirimleri yapabilmesi için kamuya açık veri setleriyle birlikte bu çoğaltma ağırlıklarını da sunar (Rust & Rao, 1996; Lohr, 2019).

3.4.3. Normların Geçerliği, Adaleti ve Sürekliliği

Norm grubu oluşturma süreci, normlardan elde edilecek yorumların geçerlik ve adalet argümanının önemli bir parçasıdır.

Geçerlik Tehditleri: Kapsama ve yanıtlamama yanlılığı normların genellenebilirliğini tehdit eden en kritik risklerinden ikisi kapsama (coverage) ve yanıtlamama (nonresponse) yanlılığıdır. Örnekleme çerçevesi belirli okul türlerini dışarıda bırakıyorsa (kapsama hatası) veya belirli özelliklere sahip okullar ya da öğrenciler sistematik olarak katılmayı reddediyorsa (yanıtlamama yanlılığı), ağırlıklandırma bu sorunları tek başına tamamen çözemeyebilir. Bu nedenle, teknik raporlarda mutlaka bir yanıtlamama yanlılığı analizi sunulmalı ve olası kalıntı yanlılıklar (düzeltmelerden sonra hala var olabilecek sistematik hatalar) dürüstçe tartışılmalıdır (Little & Rubin, 2019).

Örnekleme ve Ölçme Süreçlerinin Bütünlüğü: Örnekleme kararları, ölçme süreçlerinden bağımsız düşünülemez. Örneğin, farklı uygulama zaman aralığı (sonbahar vs. bahar) öğrenme düzeylerindeki farklılıklar nedeniyle mevsimsellik etkisi yaratacağından, bu dönemler için ayrı normlar üretilmelidir. Farklı test formları kullanıldıysa, karşılaştırılabilirlik için önce mutlaka test eşitleme yapılmalıdır. Bu adımların bütüncül bir şekilde yönetilmesi, normların sadece göreli konum bilgisi değil, aynı zamanda zaman içinde karşılaştırılabilir ve savunulabilir bir referans noktası olmasını sağlar (Kolen & Brennan, 2014).

Adalet: Alt grup normları ve ölçme değişmezliği farklı demografik alt gruplar (örneğin; anadil durumu, cinsiyet, sosyoekonomik düzey) için ayrı normlar üretilecekse veya gruplar arası karşılaştırmalar yapılacaksa, bu karşılaştırmanın adil olması gerekir. Bunun ön koşulu, kullanılan ölçme aracının testin bu gruplarda aynı psikometrik yapıyı ölçtüğünün, yani ölçme değişmezliğinin (measurement invariance) kanıtlanmasıdır. Bu kanıt sunulmadan alt grupların normlarını doğrudan karşılaştırmakyanıltıcı sonuçlara yol açabilir (Dorans & Cook, 2016).

Normların Sürekliliği: Normlar statik değildir ve zamanla geçerliklerini yitirebilirler. Değişen müfredat, öğretim yöntemleri ve öğrenci profilleri nedeniyle normlar periyodik olarak (genellikle 5-10 yılda bir) güncellenmelidir. Yeni norm setine geçildiğinde, eski

set ile arasındaki farkı açıklayan köprüleme (bridging) çalışmaları yapılmalı ve bu değişimin yeterlik düzeylerinin sınıflama oranları üzerindeki etkisi kullanıcılara şeffaf bir şekilde raporlanmalıdır.

PISA, TIMSS ve NAEP gibi uluslararası eğitim izleme programlar, yukarıda anlatılan tüm süreçlerin (tabakalı çok aşamalı tasarım, ağırlıklandırma, çoğaltma ağırlıkları ile hata kestirimi, şeffaf teknik raporlama) en iyi uygulama örneklerini sunar. Bu programların teknik raporları ve belgeleri, norm tablolarındaki her sayının arkasında yatan bilimsel titizliği ve metodolojik zinciri izlenebilir ve doğrulanabilir kılar (OECD, 2023; Martin ve ark., 2020; NCES, 2022). Sonuç olarak, norm grubu ve örnekleme tasarımı, normların doğruluğunu ve adaletini belirleyen omurgadır. Bu çerçeve şeffaf bir şekilde belgelenip titizlikle uygulandığında, normlar yalnızca birer karşılaştırma aracı olmaktan çıkar ve eğitimde iyileştirmeyi hedefleyen açık, güvenilir ve savunulabilir bir iletişim zeminine dönüşür.

4. STANDART BELİRLEME YÖNTEMLERİ

Önceki bölümlerde, bir testten elde edilen puanların yeterlik ve norm çerçeveleriyle nasıl yorumlanabilir hâle geldiği incelenmiştir. Özellikle ölçüt-dayanaklı yorumlamanın merkezinde yer alan kesme puanları, soyut olan ham puanları "yeterli/yetersiz", "temel/orta/ileri düzey" veya "sertifika almaya hak kazandı/kazanamadı" gibi somut ve eyleme geçirilebilir kararlara dönüştürür. İşte bu kritik eşiklerin belirlendiği sürece standart belirleme (standard setting) adı verilir.

Standart belirleme, yalnızca istatistiksel bir işlem değil, aynı zamanda alan uzmanlarının yargılarını, kanıtları ve değerleri bir araya getiren sistematik bir süreçtir. Sürecin temel amacı, bir yeterlik düzeyini tanımlayan performans standardını, puan ölçeği üzerinde savunulabilir bir noktayla eşleştirmektir. Standart belirleme yöntemleri genel olarak iki gruba ayrılır:

1. Test-Merkezli (Test-Centered) Yaklaşımlar: Uzmanların (panelist) dikkatini doğrudan test maddelerine veya yeterlik düzeyi betimleyicilerine (PLD) yöneltir. Panelistler, belirli bir yeterlik düzeyinin sınırında yer alan hayali bir adayın (sınırda yetkin aday / borderline candidate) bu maddelerde nasıl bir performans göstereceğini yargılarlar. Bu alt bölümde ele alınan Angoff, Ebel, Nedelsky, Bookmark gibi yöntemler bu gruba aittir.
2. Aday-Merkezli (Examinee-Centered) Yaklaşımlar: Uzman yargılarının, test maddeleri yerine gerçek aday performanslarına, sınav kâğıtları, portfolyolar veya aday gruplarına ait puan dağılımları gibi doğrudan performans kanıtlarına dayandırıldığı yaklaşımlardır.

Bu alt bölümde, alanyazında en yaygın kullanılan test-merkezli yöntemler, tarihsel kökenleri, işlem adımları, üstün ve sınırlı yönleriyle birlikte ayrıntılı olarak incelenmiştir. Aday-merkezli yaklaşımlar ise gerçek aday performanslarının, ürün dosyalarının veya puan dağılımlarının ayrı bir kanıt temeli olarak incelenmesini gerektirdiğinden, bu bölümün test maddesi temelli odağı dışında tutulmuştur. Bu yaklaşımlar için okuyucular sınır grup, karşıt gruplar ve ürün dosyası temelli standart belirleme yöntemlerine başvurabilir. Unutulmamalıdır ki, hangi yöntem seçilirse seçilsin, sürecin kalitesini ve sonuçların savunulabilirliğini belirleyen en önemli unsurlar panelistlerin dikkatli seçimi ve eğitimi, geri bildirim turlarının yapısı ve kararların olası sonuçlarını gösteren etki verisinin şeffaf bir şekilde kullanılmasıdır (Cizek & Bunch, 2007; Cizek & Earnest, 2016).

4.1. Angoff Yöntemi

Klasik standart belirleme literatürünün temel taşı olan bu yöntem, William Angoff'un (1971) öncü çalışmasına dayanır. Yöntem, "Sınırda yetkin bir aday, bu maddeyi hangi olasılıkla doğru yanıtlar?" sorusu üzerine kuruludur. Bu yöntem, özellikle yeterliklerin önemli olduğu (tıp alanı gibi) mesleki lisanslama ve sertifikasyon sınavlarında yaygın olarak kullanılmaktadır.

İşlem Adımları:

1. Panelist Seçimi ve Eğitimi: Konu alanında uzman ve farklı bakış açılarına sahip bir grup panelist seçilir. Panele, standart belirlemenin amacı, yöntemin adımları ve en önemlisi, her bir yeterlik düzeyi için sınırda yetkin aday profilinin ne anlama geldiği konusunda ayrıntılı bir eğitim verilir.
2. 1. Tur (Bireysel Değerlendirme): Panelistler, testi madde madde inceler. Her bir madde için, sınırda yetkin adayın o maddeyi doğru yanıtlama olasılığını 0 ile 1 arasında bir değer olarak (örneğin; 0,65) bireysel olarak tahmin ederler.
3. Geri Bildirim ve Tartışma: İlk tur tahminleri toplandıktan sonra, moderatör panelistlerin tahminlerinin ortalamasını ve dağılımını grupla paylaşır. Yargıları birbirinden çok farklı olan panelistlerden kararlarının gerekçelerini açıklamaları istenir. Bu turda, ilgili maddeye ait istatistikler (örneğin; madde güçlük indeksi - p-değeri) veya örnek öğrenci performansları gibi ek kanıtlar sunulabilir.
4. Sonraki Turlar ve Yakınsama: Tartışma ve geri bildirimlerin ışığında panelistler, tahminlerini yeniden gözden geçirirler. Bu süreç, panelistlerin yargıları arasında makul bir fikir birliği (yakınsama) oluşana kadar genellikle 2-3 tur devam eder.
5. Kesme Puanının Hesaplanması: Her panelistin son turdaki olasılık tahminleri, testteki tüm maddeler için toplanarak o paneliste ait kesme puanı adayı bulunur. Panelin nihai kesme puanı ise genellikle tüm panelistlerin kesme puanlarının ortalaması veya medyanı alınarak belirlenir.

Üstün ve Sınırlı Yönleri

- Üstünlükler: İçerik uzmanlarının madde düzeyinde derinlemesine düşünmesini sağlar. Süreç şeffaftır ve farklı madde türlerini içeren (çoktan seçmeli, açık uçlu vb.) testlere kolayca uygulanabilir.
- Sınırlılıklar: Bilişsel yükü yüksek olabilir. Panelistlerin, özellikle çok zor veya çok kolay maddeler için olasılık tahminlerinde sistematik olarak sapma eğilimi gösterebildiği raporlanmıştır. Bu riskler, iyi bir eğitim ve tur bazlı geri bildirimlerle yönetilebilir. Araştırmalar, sürecin kalitesinin (eğitim, geri bildirim vb.) farklı yöntemleri karşılaştırmaktan daha önemli olduğunu göstermektedir (Brandon, 2004; Reckase, 2006).

4.2. Modifiye Angoff Yöntemi

Modifiye Angoff, klasik Angoff yönteminin temel mantığını korurken, süreci daha fazla kanıta dayalı hâle getirmek ve panelist yargılarındaki değişkenliği azaltmak için geri bildirim mekanizmalarını sistematikleştirir. Özellikle karmaşık performans görevleri içeren mesleki sınavlarda ve eğitim alanındaki geniş ölçekli değerlendirmelerde standart uygulama haline gelmiştir.

İşlem Adımları ve Farklılıklar: Modifiye Angoff, klasik sürece ek olarak aşağıdaki geri bildirim türlerini yapılandırılmış bir şekilde dâhil eder:

- Normatif Geri Bildirim: Her tur sonunda, bir panelistin kendi tahminlerinin diğer panelistlerin tahminlerine göre nerede durduğunu görmesi sağlanır.
- Empirik/Gerçeklik Geri Bildirimi (Reality Check): Panelistlere, değerlendirdikleri maddelerin gerçek aday gruplarındaki performans verileri (örneğin; madde güçlük indeksleri veya Madde Tepki Kuramı (MTK/IRT) güçlük parametreleri) sunulur. Bu, uzman yargısının gerçek veriyle hizalanmasına yardımcı olur.
- Etki Verisi Geri Bildirimi (Impact Data): Belirlenen geçici kesme puanlarının, toplamda ve önemli alt gruplarda (örneğin; cinsiyet, okul türü) ne kadar adayın "yeterli" veya "yetersiz" olarak sınıflandırılacağını gösteren veriler paylaşılır. Bu, kararın pratik sonuçları hakkında paneli bilgilendirir.

Üstün ve Sınırlı Yönleri

- Üstünlükler: Kanıta dayalı geri bildirimler, panelist yargılarının daha tutarlı ve savunulabilir olmasını sağlar, sapmaları azaltır.
- Sınırlılıklar: Aşırı veya kötü yönetilmiş bir geri bildirim süreci, panelistlerin bağımsız yargılarını bırakıp "grup düşüncesine" (groupthink) kapılmasına neden olabilir. Bu riski yönetmek için moderatörün rolü ve sürecin protokolü kritik önem taşır (Hambleton & Plake, 1995; Hambleton ve ark., 2012).

4.3. Ebel Yöntemi

Ebel yöntemi, maddeleri tek bir zorluk boyutu yerine iki boyutlu bir çerçevede değerlendirerek karar sürecine içerik geçerliği vurgusu katar. İçerik temsilinin ve program hedefleriyle uyumun ön planda olduğu sınavlarda, özellikle tıp ve sağlık alanlarındaki sınavlarda kullanılır.

İşlem Adımları:

1. Izgara Oluşturma: Panelistler, maddeleri iki temel boyuta göre bir ızgaraya yerleştirirler: Önem/Uygunluk (örneğin; Çok Önemli, Önemli, Az Önemli) ve Güçlük (örneğin, Kolay, Orta, Zor). Bu genellikle 3x3'lük dokuz hücreli bir tablo oluşturur.
2. Hücrelere Oran Atama: Panelistler, her bir hücre için (örneğin, "Çok Önemli" ve "Orta Güçlükte" olan maddeler hücresi), sınırda yetkin bir adayın bu tür maddelerin yüzde kaçını doğru yanıtlamasını beklediklerini tartışarak belirlerler.
3. Kesme Puanının Hesaplanması: Testteki her bir maddenin hangi hücreye ait olduğu belirlenir ve o hücre için atanan beklenen doğru yanıt oranıyla eşleştirilir. Test genelindeki tüm maddeler için bu oranlar bir araya getirilerek nihai kesme puanı hesaplanır.

Üstün ve Sınırlı Yönleri

- Üstünlükler: "Önem" boyutunu karar sürecine doğrudan dahil ederek, testin içerik geçerliği argümanını güçlendirir.
- Sınırlılıklar: Izgaradaki boyutların (özellikle "güçlük") tanımları soyut kalırsa, panelistlerin maddeleri hücrelere yerleştirirken tutarsızlık yaşaması en büyük risktir. Bu durum, yöntemin güvenirliğini düşürebilir (Ebel, 1972; Park, 2018).

4.4. Evet/Hayır Yöntemi

Bu yöntem, klasik Angoff'un bilişsel yükünü azaltmak için geliştirilmiş, kararı basitleştiren bir varyanttır. Karar sürecini hızlandırması nedeniyle, özellikle madde sayısının çok olduğu sınavlarda ve zamanın kısıtlı olduğu durumlarda (örneğin tıp eğitiminde) tercih edilir (Impara & Plake, 1997).

İşlem Adımları: Panelistlerden, her bir madde için "Sınırda yetkin bir aday bu maddeyi doğru yanıtlayabilir mi?" sorusuna olasılık tahmini yapmak yerine, sadece "Evet" veya "Hayır" şeklinde ikili bir yargıda bulunmaları istenir. Kesme puanı, panelistlerin "Evet" olarak işaretlediği maddelerin toplamı veya ortalaması alınarak belirlenir. Geri bildirim ve tartışma turları, Modifiye Angoff'a benzer şekilde yürütülebilir.

Üstün ve Sınırlı Yönleri

- Üstünlükler: Bilişsel yükü düşüktür ve panelistler için daha kolay ve hızlı bir karar süreci sunar.
- Sınırlılıklar: İkili karar yapısı, olasılık tahminindeki nüansları ortadan kaldırır. Özellikle çok kolay veya çok zor maddelerde panelistlerin yargılarının sistematik olarak sapma riski bulunur. Bazı uygulamalarda bu riski azaltmak için "Evet/Hayır/Emin Değilim" gibi üç kategorili bir yapı kullanılır (Yudkowsky ve ark., 2008).

4.5. Nedelsky Yöntemi

Bu yöntem, özel olarak çoktan seçmeli maddeler için tasarlanmıştır ve mantığını seçeneklerin elenmesi üzerine kurar. Yöntemin kullanımına yönelik eğilim azalmış olsa da, çeldirici kalitesinin yüksek olduğu standart testlerde hâlâ uygulanabilmektedir.

İşlem Adımları: Panelistlerden, her bir çoktan seçmeli madde için, "Sınırda yetkin bir aday, hangi seçeneklerin kesinlikle yanlış olduğunu bilir ve onları eler?" sorusunu yanıtlamaları istenir. Bir adayın eleyebileceği çeldiriciler işaretlenir. Geriye kalan seçenek sayısı (k), o maddeden şansla doğru yanıta ulaşma olasılığını (1/k) verir. Örneğin, beş seçenekli bir maddede panelist sınırda adayın üç çeldiriciyi eleyebileceğine karar verirse, geriye iki seçenek kalır ve bu maddeden beklenen doğru yanıt olasılığı 1/2 (0,50) olur. Testteki tüm maddeler için bu olasılıklar toplanarak kesme puanı hesaplanır (Nedelsky, 1954).

Üstün ve Sınırlı Yönleri

- Üstünlükler: Karar sürecini, seçeneklerin ve çeldiricilerin kalitesi gibi somut bir unsura bağlar.
- Sınırlılıklar: Yalnızca çoktan seçmeli maddelere uygulanabilir. Çeldiricilerin kalitesi düşük veya işlevsiz ise, yöntem aşırı düşük ve gerçekçi olmayan kesme puanları üretir.

4.6. Bookmark (Yer İmi) Yöntemi

Madde Tepki Kuramı'ndan (MTK/IRT) beslenen Bookmark, günümüzde en yaygın kullanılan standart belirleme yöntemlerinden biridir. ABD'deki eyalet düzeyindeki K-12 değerlendirme programlarında, üniversite yerleştirme sınavlarında (örneğin ACCUPLACER) ve diğer geniş ölçekli testlerde standart uygulama hâline gelmiştir (Mitzel ve ark., 2001).

İşlem Adımları:

1. Sıralı Madde Kitapçığı (Ordered Item Booklet - OIB): Testteki maddeler, önceden uygulanarak elde edilen IRT zorluk parametrelerine (b-değerleri) göre, en kolaydan en zora doğru sıralanarak özel bir kitapçık oluşturulur.

2. Yer İmi Koyma: Panelistler bu kitapçığı baştan sona incelerler. Her panelist, "Sınırda yetkin bir adayın, bu maddedeyi ve bundan önceki (daha kolay) maddeleri tutarlı bir şekilde doğru yanıtlama olasılığının belirli bir önceden tanımlanmış olasılık eşiğin (örneğin, 2/3 yani yaklaşık %67 veya %80) ulaştığı veya aştığı madde hangisidir?" sorusuna yanıt arar. Bu kritik maddeye fiziksel olarak veya dijital ortamda bir "yer imi" (bookmark) koyarlar.
3. Kesme Puanının Hesaplanması: Her panelistin yer imi koyduğu maddenin IRT zorluk değeri, o panelistin kesme puanı adayıdır (yetenek (θ) ölçeğinde). Panelin nihai kesme puanı, bu yetenek değerlerinin ortalaması alınarak bulunur ve daha sonra testin ham puan veya raporlama ölçeğine dönüştürülür.

Üstün ve Sınırlı Yönleri

- Üstünlükler: MTK'ya dayalı olduğu için farklı test formları arasında tutarlılığı sağlamak daha kolaydır. Süreç, yeterlik düzeyi tanımlarıyla (PLD'ler) güçlü bir bağ kurar. Panelistlerin bilişsel yükü, madde madde olasılık tahmin etmekten daha düşüktür.
- Sınırlılıklar: Yöntem, maddelerin önceden uygulanıp MTK ile kalibre edilmesini gerektirir. Sıralı madde kitapçığının kalitesi ve MTK parametrelerinin doğruluğu, sonucun geçerliği için kritiktir.

4.7. Basket (Sepet) Yöntemi

Bu yöntem, özellikle Diller için Ortak Avrupa Başvuru Metni (D-OABM / CEFR) gibi önceden tanımlanmış ve kabul görmüş yeterlik düzeylerine sahip bağlamlarda kullanılır.

İşlem Adımları: Panelistlere, her bir maddeyi, karşılık geldiği yeterlik düzeyini temsil eden kavramsal "sepetlere (baskets)" yerleştirmeleri görevi verilir. Örneğin, bir maddeyi inceler ve "Bu maddeyi doğru yanıtlamak, D-OABM'ye göre B1 düzeyi bir beceri mi gerektirir, yoksa B2 düzeyi mi?" diye karar vererek maddeyi uygun sepete atarlar. Tüm maddeler sepetlere yerleştirildikten sonra, her bir düzey için kesme puanı, o düzeydeki ve altındaki sepetlerde yer alan maddelerin sayısı veya ağırlıklandırılmasıyla elde edilen kurallara göre hesaplanır.

Üstün ve Sınırlı Yönleri

- Üstünlükler: Süreç, doğrudan yeterlik çerçevesinin diliyle yürütüldüğü için sonuçlar paydaşlar için çok anlaşılırdır.
- Sınırlılıklar: Yöntemin başarısı, panelistlerin referans alınan yeterlik çerçevesine (örneğin D-OABM) derinlemesine hakim olmalarına bağlıdır. Sepetleme yargılarındaki tutarlılığı sağlamak için yoğun bir eğitim ve kalibrasyon süreci gerektirir (Kaftandjieva, 2009).

5. SONUÇ

Yeterliklerin ve normların belirlenmesi, ölçme sonuçlarının karar süreçlerinde doğru ve tutarlı biçimde kullanılabilmesi için temel bir işleve sahiptir. Bu bölümde ele alınan kavramlar, ham puanın tek başına yeterli bir değerlendirme bilgisi sunmadığını ve puanın mutlaka bir anlam çerçevesi içinde yorumlanması gerektiğini göstermektedir. Uygulamada bu durum, her ölçme programının daha test geliştirme aşamasında hangi yeterlik düzeylerini raporlayacağını, hangi norm grubunu kullanacağını, kesme puanlarının nasıl belirleneceğini ve ölçme hatasının raporlara nasıl yansıtılacağını planlamasını gerektirir. Böylece puanlar yalnızca sayısal sonuçlar olarak değil, bireyin hangi düzeyde performans gösterdiğini ve bu performansın hangi referans içinde yorumlanacağını açıklayan değerlendirme bilgisine dönüşür.

Yeterlik düzeylerinin kullanıldığı uygulamalarda en önemli pratik gereklilik, düzey etiketlerinin açık performans düzeyi betimleyicileriyle desteklenmesidir. "Temel", "Yeterli" ve "İleri" gibi düzeyler, yalnızca puan aralıklarıyla tanımlandığında sınırlı bir açıklama gücüne sahip olur. Bu düzeylerin hangi bilgi, beceri ve davranış göstergelerine karşılık geldiği açık biçimde yazıldığında ise raporlar adaylar, öğretmenler, uzmanlar ve karar vericiler için daha anlaşılır hâle gelir. Bu nedenle sınav ve değerlendirme raporlarında yalnızca "hangi puan alındı?" sorusuna değil, "bu puan ne anlama gelir?" ve "bu düzeydeki birey tipik olarak ne yapabilir?" sorularına da yanıt verilmelidir.

Normlara dayalı yorumlarda ise öncelikli pratik çıkarım, norm grubunun niteliğinin raporlanan sonucun niteliğini doğrudan belirlediğidir. Yaş normu, sınıf düzeyi normu, büyüme normu, yerel norm ve koşullu norm farklı kullanım amaçlarına sahiptir. Bu nedenle raporlarda normun hangi evrene dayandığı, örneklemin bu evreni nasıl temsil ettiği, ağırlıklandırma işlemlerinin uygulanıp uygulanmadığı ve standart hata değerlerinin nasıl hesaplandığı açıkça belirtilmelidir. Yüzdelik, yüzdelik sıra, standart puan ve norm eğrileri kullanıcıya sade bir karşılaştırma imkânı sunar. Ancak bu göstergeler, özellikle bireysel kararlar verilirken ölçmenin standart hatası ve güven aralığı ile birlikte değerlendirilmelidir. Kesme puanı çevresindeki puanlarda bu ilke daha da önemlidir.

Standart belirleme yöntemleri açısından bölümün temel sonucu, kesme puanlarının yalnızca teknik bir hesaplama sonucunda değil, sistematik uzman yargısı ve kanıt temelli bir süreç sonunda belirlenmesi gerektiğidir. Angoff, Modifiye Angoff, Ebel, Evet/Hayır, Nedelsky, Bookmark ve Basket yöntemleri farklı test yapıları için kullanılabilir. Ancak yöntemin uygunluğu kadar panelistlerin seçimi, panelist eğitimi, sınırda yetkin aday tanımının açıklığı, geri bildirim turlarının düzeni ve etki verisinin değerlendirilmesi de kararların savunulabilirliği açısından belirleyicidir. Sonuç olarak iyi yapılandırılmış bir değerlendirme sistemi, puanı üretmekle yetinmeyen, puanın anlamını, hata payını, yeterlik düzeyini, norm içindeki konumunu ve karar gerekçesini birlikte sunan bir sistem olmalıdır.

KAYNAKÇA | REFERENCES

American Educational Research Association (AERA), American Psychological Association (APA) & National Council on Measurement in Education (NCME). (2014). *Standards for educational and psychological testing*. American Educational Research Association.

Angoff, W. H. (1971). Scales, norms, and equivalent scores. R. L. Thorndike (Ed.), *Educational measurement* (2. Baskı, s. 508–600) içinde. American Council on Education.

Betebenner, D. W. (2009). Norm- and criterion-referenced student growth. *Educational Measurement: Issues and Practice, 28*(4), 42–51. https://doi.org/10.1111/j.1745-3992.2009.00161.x

Bourque, J., Pomedli, A., Charlebois, M., Milette, K. & St-Onge, C. (2020). Performance of the Ebel standard-setting method for a high-stakes exam. *BMC Medical Education, 20*, Article 264. https://doi.org/10.3352/jeehp.2020.17.12

Brandon, P. R. (2004). Conclusions about frequently studied modified Angoff topics. *Applied Measurement in Education, 17*(1), 59–88. https://doi.org/10.1207/s15324818ame1701_4

Cizek, G. J. & Bunch, M. B. (2007). *Standard setting: A practitioner's guide to establishing and evaluating performance standards on tests*. SAGE.

Cizek, G. J. & Earnest, D. S. (2016). Setting performance standards on tests. S. Lane, M. R. Raymond & T. M. Haladyna (Ed.), *Handbook of test development* (2. Baskı, s. 212–237). Routledge.

Cole, T. J. & Green, P. J. (1992). Smoothing reference centile curves: The LMS method and penalized likelihood. *Statistics in Medicine, 11*(10), 1305–1319. https://doi.org/10.1002/sim.4780111005

Council of Europe. (2020a). *Common european framework of reference for languages: Learning, teaching, assessment (Companion volume)*. Council of Europe Publishing.

Council of Europe. (2020b). *Common European framework of reference for languages: Learning, teaching, assessment (Companion volume) [Diller İçin Avrupa Ortak Başvuru Metni: Öğrenme, Öğretme ve Değerlendirme (Tamamlayıcı Cilt)]* (Türkçeye çeviri: Millî Eğitim Bakanlığı, 2022). https://ttkb.meb.gov.tr/meb_iys_dosyalar/2022_01/04144518_cefr_tr.pdf

Council of Europe. (2020). *Common European Framework of Reference for Languages: Learning, teaching, assessment—Companion volume*. Council of Europe Publishing. https://rm.coe.int/common-european-framework-of-reference-for-languages-learning-teaching/16809ea0d4

Dorans, N. J. & Cook, L. L. (Ed.). (2016). *Fairness in educational assessment and measurement*. Routledge.

Ebel, R. L. (1972). *Essentials of educational measurement*. Prentice Hall.

Egan, K., Schneider, J. & Ferrara, S. (2012). Performance level descriptors: History, practice, and implications. G. J. Cizek (Ed.), *Setting performance standards: Foundations, methods, and innovations* (s. 279–300) içinde. Routledge.

Ferrara, S., Johnson, E. & Chen, W. L. (2005). Vertically articulated performance standards: Logic, procedures, and likely classification accuracy. *Applied Measurement in Education, 18*(1), 35–60. https://doi.org/10.1207/s15324818ame1801_3

Hambleton, R. K. & Plake, B. S. (1995). Using an extended Angoff procedure to set standards on complex performance assessments. *Applied Measurement in Education, 8*(1), 41–56. https://doi.org/10.1207/s15324818ame0801_4

Hambleton, R. K., Pitoniak, M. J. & Copella, J. M. (2012). Essential steps in setting performance standards on educational tests and strategies for assessing the reliability of results. G. J. Cizek (Ed.), *Setting performance standards: Foundations, methods, and innovations* (s. 47–76) içinde. Routledge.

Heeringa, S. G., West, B. T. & Berglund, P. A. (2017). *Applied survey data analysis* (2. baskı). CRC Press.

Impara, J. C. & Plake, B. S. (1997). Standard setting: An alternative approach. *Journal of Educational Measurement, 34*(4), 353–366. https://doi.org/10.1111/j.1745-3984.1997.tb00523.x

Kaftandjieva, F. (2009). *Methods for setting cut scores in criterion-referenced achievement tests: A comparative study*. EALTA Resources. https://ealta.eu/documents/resources/Research_Colloquium_report.pdf

Kane, M. T. (2013). *Validating the interpretations and uses of test scores*. Routledge.

Kish, L. (1965). *Survey sampling*. Wiley.

Koenker, R. (2005). *Quantile regression*. Cambridge University Press.

Kolen, M. J. & Brennan, R. L. (2014). *Test equating, scaling, and linking: Methods and practices* (3. Baskı). Springer.

Lane, S., Raymond, M. R. & Haladyna, T. M. (Ed.). (2016). *Handbook of test development* (2. Baskı). Routledge.

Lewis, D. M. & Haug, C. A. (2005). Aligning policy and methodology to achieve consistent across-grade performance standards. *Applied Measurement in Education, 18*(1), 11–34. https://doi.org/ 10.1207/s15324818ame1801_2

Lewis, D. M., Mitzel, H. C., Mercado, R. L. & Schulz, E. M. (2012). The Bookmark standard-setting procedure. G. J. Cizek (Ed.), *Setting performance standards: Foundations, methods, and innovations* (s. 225–254) içinde. Routledge.

Lin, J. (2006). The Bookmark procedure for setting cut-scores. *Alberta Journal of Educational Research, 52*(2), 143–160. https://doi.org/10.55016/ojs/ajer.v52i1.55110

Lissitz, R. W. & Huynh, H. (2003). *Vertical equating for the Arkansas ACTAAP assessments*. Arkansas Department of Education.

Little, R. J. A. & Rubin, D. B. (2019). *Statistical analysis with missing data* (3. baskı). Wiley.

Livingston, S. A. & Zieky, M. J. (1982). *Passing scores: A manual for setting standards of performance on educational and occupational tests*. Educational Testing Service.

Livingston, S. A. & Zieky, M. J. (1989). A comparative study of standard-setting methods. *Applied Measurement in Education, 2*(2), 121–141. https://doi.org/10.1207/s15324818ame0202_3

Lohr, S. L. (2019). *Sampling: Design and analysis* (3. baskı). CRC Press.

Martin, M. O., Mullis, I. V. S. & Hooper, M. (Ed.). (2020). *Methods and procedures: TIMSS 2019 technical report*. TIMSS & PIRLS International Study Center, Boston College. https://timssandpirls.bc.edu/timss2019/methods/

Millî Eğitim Bakanlığı. (2024). *PISA 2022 Türkiye raporu*. Millî Eğitim Bakanlığı. Millî Eğitim Bakanlığı. https://pisa.meb.gov.tr/meb_iys_dosyalar/2024_03/21120745_26152640_pisa2022_rapor.pdf

Messick, S. (1989). Validity. R. L. Linn (Ed.), *Educational measurement* (3. baskı, s. 13–104). American Council on Education/Macmillan.

Mitzel, H. C., Lewis, D. M., Patz, R. J. & Green, D. R. (2001). The Bookmark procedure: Psychological perspectives. G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (s. 249–281) içinde. Lawrence Erlbaum Associates.

National Center for Education Statistics. (2022). *NAEP technical documentation*. U.S. Department of Education, Institute of Education Sciences. https://nces.ed.gov/nationsreportcard/tdw/

Nedelsky, L. (1954). Absolute grading standards for objective tests. *Educational and Psychological Measurement, 14*(1), 3–19. https://doi.org/10.1177/001316445401400101

Organisation for Economic Co-operation and Development. (2019). *PISA 2018 assessment and analytical framework*. OECD Publishing.

Organisation for Economic Co-operation and Development. (2023). *PISA 2022 assessment and analytical framework*. OECD Publishing.

Park, J. (2018). Angoff, Ebel, Bookmark, and Hofstee: A comparison of the four most popular standard setting methods. *Journal of Educational Evaluation for Health Professions, 15*, Article 10. https://doi.org/10.3352/jeehp.2018.15.32

Park, J. (2022). Possibility of using the Yes/No Angoff in the Korean Medical Licensing Examination to set the pass/fail standard. *Journal of Educational Evaluation for Health Professions, 19*, Article 10. https://doi.org/ 10.3352/jeehp.2022.19.23

Reckase, M. D. (2006). A conceptual framework for a psychometric theory for standard setting with a review of the methods. S. M. Downing & T. M. Haladyna (Ed.), *Handbook of test development* (s. 415–440) içinde. Lawrence Erlbaum Associates.

Rust, K. F. & Rao, J. N. K. (1996). Variance estimation for complex surveys using replication techniques. *Statistical Methods in Medical Research, 5*(3), 283–310. https://doi.org/10.1177/096228029600500305

Yudkowsky, R., Downing, S. M. & Wirth, S. (2008). A pilot study of a three-level Angoff method for setting standards on a patient-based clinical skills examination. *Academic Medicine, 83*(10 Suppl), S8–S10. https://doi.org/ 10.1097/ACM.0b013e318183c683

YAZAR HAKKINDA | ABOUT AUTHOR

DOÇ. DR. AHMET SALİH ŞİMŞEK

Kırşehir Ahi Evran Üniversitesi | assimsek[at]ahievran.edu.tr

ORCID: 0000-0002-9764-3285

Doç. Dr. Ahmet Salih Şimşek, Eğitimde Ölçme ve Değerlendirme alanında yüksek lisans ve doktora eğitimini 2009–2017 yılları arasında Ankara Üniversitesi'nde tamamlamıştır. Doktora sonrası araştırmaları için 2019 yılında Twente Üniversitesi'nde 8 ay süreyle Öğrenme, Veri Analitikleri ve Teknoloji (LDT: Learning, Data-Analytics and Technology) bölümünde misafir araştırmacı olarak, teknoloji destekli ölçme uygulamaları için madde tepki süresi MTK modeli geliştirilmesi üzerine çalışmıştır. Çalışma hayatında akademik personel olarak 15. yılını geride bırakan Dr. Şimşek, halen Kırşehir Ahi Evran Üniversitesi'nde Eğitimde Ölçme ve Değerlendirme Anabilim Dalı Başkanı ve üniversitenin kalite süreçlerinde Veri Yönetim Koordinatörü olarak görev almaktadır. Çeşitli kurum (MEB, ÖSYM) ve organizasyonların (PIKTES) beceri temelli ölçme değerlendirme süreçleri için izleme değerlendirme sistemlerinin tasarlanmasında çalışmalara katkı sunmaktadır. Ulusal ve uluslararası 50'den fazla yayını bulunan Dr. Şimşek, teknoloji ile zenginleştirilmiş ölçme değerlendirme süreçleri, veri, istatistik ve yapay zekâ okuryazarlıkları, bireyselleştirilmiş test geliştirme ve eğitimde yapay zekâ uygulamaları konularında çalışmalarını sürdürmektedir.

ASSOC. PROF. DR. AHMET SALİH ŞİMŞEK

Kırşehir Ahi Evran University | assimsek[at]ahievran.edu.tr

ORCID: 0000-0002-9764-3285

Assoc. Prof. Dr. Ahmet Salih Şimşek completed his master's and doctoral studies in Educational Measurement and Evaluation at Ankara University between 2009 and 2017. In 2019, he conducted postdoctoral research as a visiting researcher for eight months in the Learning, Data Analytics and Technology (LDT) department at the University of Twente, where he worked on the development of response time IRT model for technology-enhanced assessment practices. Having completed his 15th year as an academic staff member, Dr. Şimşek currently serves as the Head of the Department of Educational Measurement and Evaluation at Kırşehir Ahi Evran University and as the Data Management Coordinator in the university's quality assurance processes. He contributes to the design of monitoring and evaluation systems for skill-based assessment and evaluation processes carried out by various institutions, including the Ministry of National Education (MEB) and the Assessment, Selection and Placement Center (ÖSYM), as well as organizations such as PIKTES. Dr. Şimşek has authored more than 50 national and international publications and continues his studies on technology-enhanced assessment and evaluation processes, data, statistics and artificial intelligence literacies, adaptive test development and artificial intelligence applications in education.

DİZİN

A

Açık uçlu madde, 86, 91, 95, 99, 100, 121, 123, 125, 131, 139, 183, 185, 186, 207, 212, 217, 232, 252, 253, 264, 265, 267, 268, 270, 271, 272, 349, 398, 400, 402, 416, 432, 448, 508, 510, 521, 522, 543, 611, 630, 640, 648

Adalet, 135, 367, 531, 533, 535, 548, 549, 560, 563, 566, 569, 617, 674

Ağırlıklı puanlama, 393, 400

Akademik Personel ve Lisansüstü Eğitimi Giriş Sınavı (ALES), 431, 580, 639

Ayırt edicilik, 43, 44, 47, 48, 49, 50, 51, 52, 53, 54, 86, 212, 226, 227, 228, 245, 401, 402, 421, 435, 436, 437, 438, 440, 449, 450, 455, 581, 602, 609, 614, 637, 638, 643

Ayırt edicilik parametresi (a parametresi), 47, 49, 52, 435, 438, 439, 457, 637, 638

B

Bağıl değerlendirme, 23, 24, 336, 469, 473, 474, 483

Belirtke tablosu, 106, 107, 109, 110, 144, 145, 162, 207, 208, 209, 210, 214, 216, 228

Bilgisayar Ortamında Bireye Uyarlanmış Test (BOBUT), 49, 54, 137, 198, 265, 266, 379, 380, 389, 453, 463, 530, 575, 577, 578, 579, 580, 581, 582, 583, 584, 585, 586, 587, 588, 592, 641, 644, 654

Bilişsel alan, 126, 138, 139, 149, 178, 182, 187, 237, 239, 241, 252, 271

Bilişsel beceriler, 239, 345, 494, 506, 599

Bilişsel tanı modelleri, 31, 54, 73, 137, 275, 530, 592

Ç

Çok aşamalı test (ÇAT), 74, 236, 265, 266, 380, 575, 578, 579, 588, 589, 590, 591, 592, 594, 595, 596, 641, 645

Çoktan seçmeli madde, 100, 120, 124, 151, 217, 221, 242, 243, 244, 246, 247, 248, 250, 251, 252, 256, 257, 267, 269, 271, 274, 332, 333, 334, 335, 397, 398, 399, 400, 433, 475, 491, 508, 512, 513, 515, 520, 522, 601, 630, 679

Cronbach alfa, 48, 88, 91, 92, 93, 94, 97, 105, 447, 450

D

Değerlendirme araçları, 210, 289, 294, 295, 302, 303, 304, 329, 369, 375, 510, 536, 537, 540, 541, 546

Değişen Madde Fonksiyonu (DMF), 74, 201, 225, 227, 275, 441, 442, 445, 448, 529, 551, 552, 567, 640, 646

Değişken, 3, 5, 6, 7, 8, 9, 10, 17, 18, 19, 20, 21, 35, 49, 55, 56, 57, 62, 66, 95, 116, 117, 119, 123, 124, 436, 439, 444, 445, 455, 575, 585, 586, 591, 592, 600, 630, 631, 636, 666

Dereceli puanlama anahtarı (Rubrik), 157, 159, 213, 217, 267, 279, 289, 301, 307, 309, 310, 311, 312, 313, 314, 315, 316, 321, 341, 348, 399, 410, 509, 522, 563

Doğru-yanlış maddesi, 90, 109, 182, 184, 207, 215, 217, 241, 242, 251, 252, 258, 259, 267, 268, 269, 271, 284, 331, 335, 337, 370, 371, 398,

Duyuşsal alan, 37, 167, 169, 178, 241

Duyuşsal özellikler, 67, 135, 162, 166, 167, 169, 189, 265, 667

E

Elektronik sınav, 100, 379

Eşleştirmeli madde, 211, 217, 250, 251, 259, 260, 261, 267, 269, 271, 367

Etik, 135, 169, 185, 189, 496, 531, 533, 535, 536, 537, 547, 548, 566, 575, 615, 616, 671

G

Geçerlik, 37, 70, 77, 79, 81, 103, 104, 105, 106, 110, 112, 114, 115, 116, 117, 119, 120, 121, 122, 123, 125, 126, 128, 131, 135, 175, 177, 180, 189, 206, 211, 217, 222, 223, 229, 231, 232, 239, 241, 246, 258, 260, 263, 267, 272, 287, 304, 306, 308, 359, 370, 374, 377, 380, 386, 399, 413, 442, 447, 448, 450, 475, 505, 507, 508, 512, 523, 535, 537, 538, 544, 547, 552, 556, 560, 561, 563, 577, 597, 600, 612, 613, 615, 642, 667, 668, 671, 674

Geçerlik kanıtları, 77, 105, 115, 117, 121, 122, 206, 211, 442, 448, 508, 512, 556, 561

Genellenebilirlik kuramı, 69, 70, 82, 211, 529

Geniş ölçekli sınavlar, 209, 214, 219, 268, 270, 397, 475, 508, 523

Geri bildirim, 25, 121, 206, 329, 332, 335, 346, 348, 356, 369, 396, 467, 470, 471, 472, 506, 508, 509, 510, 511, 522, 554, 555, 563, 564, 577, 587, 600, 605, 676, 677, 678

Güçlük parametresi (b parametresi), 46, 49, 431, 434, 435, 438, 439, 584, 637, 638

Güvenirlik, 38, 40, 41, 42, 48, 70, 77, 79, 82, 83, 84, 86, 87, 88, 89, 91, 94, 95, 96, 97, 98, 99, 100, 101, 102, 104, 105, 118, 123, 125, 126, 128, 135, 175, 177, 180, 189, 212, 231, 232, 239, 241, 255, 260, 262, 263, 268, 304, 305, 306, 308, 346, 347, 370, 374, 396, 399, 400, 401, 412, 413, 441, 442, 447, 475, 505, 523, 535, 537, 538, 544, 547, 560, 563, 575, 577, 597, 600, 612, 613, 615, 620, 642, 643

Güvenirlik katsayısı, 38, 42, 89, 91, 97, 101, 102, 447

H

Hata, 10, 36, 39, 40, 41, 42, 58, 67, 77, 79, 80, 81, 82, 83, 91, 92, 97, 100, 101, 102, 105, 157, 201, 215, 223, 224, 228, 229, 230, 232, 243, 260, 263, 267, 268, 272, 384, 400, 413, 440, 457, 458, 514, 538, 544, 550, 552, 585, 586, 590, 594, 604, 611, 617, 619, 638, 639, 643, 657, 665, 670, 671, 673, 674, 675

İ

İç tutarlılık, 77, 83, 87, 88, 89, 90, 91, 92, 93, 95, 97, 105, 118, 125, 442, 447

K

Kamu Personeli Seçme Sınavı (KPSS), 24, 26, 269, 374, 376, 379

Kararlılık, 83, 105, 138, 175, 186, 258, 477

Katkı gerektiren madde türleri, 213, 217, 241, 252, 254, 266, 267, 270, 271, 272

Kısa cevaplı madde, 247, 252, 253, 262, 263, 265, 267, 270, 271, 272, 284, 398, 630

Klasik test kuramı, 31, 38, 69, 74, 77, 79, 82, 225, 227, 268, 645, 649

Kontrol listesi, 213, 297, 311, 315, 316, 320, 321, 322, 554

Korelasyon, 3, 5, 17, 18, 19, 20, 21, 26, 35, 43, 82, 83, 84, 85, 86, 87, 88, 89, 92, 96, 97, 99, 115, 116, 117, 118, 436, 437, 442, 447, 451, 458, 591, 629

KR-20, 88, 90, 91, 92, 94, 448

KR-21, 88, 90, 91, 94

Kullanışlılık, 3, 12, 77, 79, 124, 125, 126, 128, 257, 258, 264, 399, 426, 511

M

Madde istatistikleri, 38, 67, 201, 212, 219, 225, 228, 232, 246, 371, 421, 431, 441, 460

Madde tepki kuramı, 31, 43, 69, 74, 82, 212, 227, 268, 275, 276, 362, 530, 571, 637, 640, 644, 645, 649, 652

Millî Eğitim Bakanlığı (MEB), 24, 27, 43, 145, 193, 196, 197, 209, 210, 234, 235, 244, 274, 371, 373, 377, 379, 380, 385, 387, 388, 470, 472, 473, 475, 476, 484, 497, 498, 499, 500, 502, 503, 526, 529, 530, 563, 564, 570, 661

Mutlak değerlendirme, 23, 96, 269, 411, 467, 473, 483

N

Norm, 4, 36, 37, 67, 178, 232, 370, 474, 657, 658, 659, 660, 664, 665, 666, 667, 668, 669, 670, 671, 672, 674, 675

O

Otomatik madde üretimi, 126, 217, 575, 577, 597, 598, 599, 600, 601, 602, 604, 606, 607, 608, 609, 610, 612, 613, 614, 615, 616, 641, 643, 644, 653

Otomatik puanlama, 124, 267, 270, 382, 393, 395, 402, 403, 404, 405, 406, 407, 408, 409, 410, 411, 413, 463, 523, 653

Ö

Ölçme, Seçme ve Yerleştirme Merkezi Başkanlığı (ÖSYM), 23, 24, 43, 69, 196, 197, 231, 233, 234, 257, 265, 266, 273, 275, 276, 370, 376, 377, 379, 381, 382, 384, 388, 431, 473, 476, 477, 483, 485, 538, 539, 541, 542, 544, 545, 546, 548, 549, 556, 564, 565, 568, 570, 653

Ölçmede hata, 77, 79, 552

Örneklem, 38, 41, 44, 67, 84, 85, 86, 119, 123, 372, 449, 475, 556, 561, 621, 622, 623, 624, 625, 627, 634, 636, 639, 640, 643, 645, 657, 664, 665, 666, 667, 668, 669, 670, 672, 673

Öz değerlendirme, 169, 187, 188, 296, 297, 299, 301, 302, 329, 346, 347, 352, 353, 354, 355, 472, 510, 511, 522

P

Performans, 90, 95, 99, 100, 109, 145, 157, 158, 159, 161, 164, 165, 169, 178, 179, 181, 183, 186, 187, 188, 212, 213, 214, 217, 277, 279, 281, 282, 283, 284, 285, 286, 287, 288, 289, 290, 291, 292, 293, 294, 295, 298, 300, 301, 302, 304, 305, 309, 310, 311, 312, 313, 315, 316, 317, 323, 324, 353, 355, 356, 372, 374, 377, 379, 396, 398, 399, 400, 405, 407, 410, 411, 412, 472, 504, 506, 507, 508, 509, 510, 521, 522, 542, 545, 561, 562, 563, 564, 610, 616, 623, 636, 657, 660

Performans değerlendirme, 277, 279, 281, 282, 283, 284, 285, 286, 287, 293, 298, 300, 301, 309, 311, 323, 374, 396, 507, 509

Portfolyo, 188, 279, 281, 295, 296, 297, 298, 299, 300, 301, 302, 303, 304, 305, 306, 307, 308, 311, 355

Programme for International Student Assessment (PISA), 49, 54, 71, 167, 182, 194, 363, 373, 375, 379, 380, 387, 467, 468, 474, 484, 485, 551, 575, 576, 594, 595, 616, 619, 639, 643, 644, 649, 651, 657, 658, 661, 671, 674, 675

Progress in International Reading Literacy Study (PIRLS), 54, 71, 182, 373, 375, 387, 474, 551, 671

Proje, 108, 185, 186, 188, 279, 281, 283, 286, 287, 293, 294, 295, 311, 354, 496, 553, 564, 611, 653

Psikomotor beceri, 138, 177, 178, 180, 181, 284

Puanlayıcı güvenirliği, 77, 83, 95, 96, 217, 267, 287, 487

R

Rasch modeli, 31, 38, 74, 211, 412

S

Seçme gerektiren madde türleri, 241, 266, 267, 268, 269

Sınav güvenliği, 201, 224, 225, 232, 257, 367, 369, 380, 381, 382, 386, 580, 581

Standart belirleme, 22, 96, 229, 230, 377, 657, 659, 670, 675, 676, 679

Standart hata, 41, 42, 49, 77, 101, 102, 103, 260, 261, 413, 415, 456, 458, 585, 586, 592, 619, 620, 639, 641, 643, 657, 659, 660, 665, 666, 667, 668, 669, 670, 673

Standart test, 43, 87, 370, 371, 440, 556, 557, 616, 679

Ş

Şans başarısı, 52, 74, 81, 100, 211, 221, 246, 250, 257, 258, 266, 267, 268, 271, 332, 398, 401, 433, 640, 644

Şans parametresi (c parametresi), 52, 62, 439, 458, 638

T

Test eşitleme, 198, 275, 276, 529, 530, 575, 577, 617, 618, 619, 620, 621, 622, 628, 629, 631, 633, 637, 638, 639, 640, 642, 644, 645, 646, 648, 649, 650, 652, 653, 657, 659, 661, 665, 666, 667, 668, 670, 671, 674

Test geliştirme, 33, 37, 59, 66, 73, 74, 88, 109, 197, 199, 201, 203, 204, 206, 207, 212, 216, 224, 225, 227, 229, 231, 232, 233, 239, 276, 362, 370, 371, 386, 433, 438, 535, 538, 541, 551, 553, 568, 597, 599, 640, 650

Test istatistikleri, 43, 212, 273, 421, 442, 460

T-puanı, 430, 431, 667

Trends in International Mathematics and Science Study (TIMSS), 49, 54, 71, 182, 373, 375, 379, 380, 387, 474, 485, 551, 618, 671, 675

Türkiye Yüzyılı Maarif Modeli (TYMM), vi, xi, xvii, 208, 470, 472, 484, 489, 491, 494, 497, 498, 499, 500, 501, 502, 503, 504, 505, 506, 507, 508, 509, 510, 511, 512, 514, 521, 522, 523, 526

Tutarlılık, 83, 87, 88, 89, 92, 95, 96, 97, 105, 125, 308, 314, 319, 348, 406, 411, 599, 608, 609, 610, 642

Ü

Üst düzey düşünme becerileri, 73, 149, 150, 151, 162, 165, 182, 242, 251, 253, 271, 272, 274, 353, 371, 396, 498, 499, 525, 611

Y

Yabancı Dil Bilgisi Seviye Tespit Sınavı (YDS), 269, 379, 485, 580

Yanlılık, 81, 218, 227, 255, 411, 442, 445, 448, 535, 548, 550, 551, 552, 558, 561, 567, 575, 585, 616

Yapay zekâ, 31, 125, 126, 131, 135, 137, 189, 203, 217, 275, 353, 393, 396, 402, 403, 405, 407, 408, 410, 413, 418, 493, 508, 523, 575, 577, 586, 587, 592, 598, 599, 606, 608, 610, 612, 613, 614, 615, 616, 642, 648, 650, 652, 653

Yerleştirme, 24, 56, 164, 206, 212, 213, 231, 269, 276, 367, 370, 376, 378, 388, 409, 424, 444, 465, 467, 469, 476, 477, 478, 480, 481, 483, 485, 538, 549, 564, 568, 569, 611, 650, 654, 679

Yüksek riskli sınavlar, 211, 219, 220, 221, 224, 229, 257, 374, 399, 469, 491, 494, 522, 549, 555

Yükseköğretim Kurumları Sınavı (YKS), 24, 234, 269, 373, 374, 376, 377, 379, 388, 470, 485, 491, 492, 511, 522

Z

z-puanı, 429, 430, 667

Ölçme ve Değerlendirme: Kuramlar, Süreçler ve İleri Uygulamalar, Ölçme, Seçme ve Yerleştirme Merkezi Başkanlığı (ÖSYM) ile Türkiye Bilimler Akademisi (TÜBA) iş birliğiyle hazırlanmıştır. Alanında uzman 23 yazarın katkılarıyla oluşturulan bu eser, 16 bölümde, ölçme ve değerlendirmenin kavramsal ve kuramsal temellerinden ileri psikometrik yöntemlere ve güncel uygulamalara uzanan kapsamlı bir çerçeve sunmaktadır. Kuramsal bilgiyi uygulamayla bütünleştiren bu eser; örnek uygulamalar ve bilimsel yaklaşımlar aracılığıyla ölçme ve değerlendirme süreçlerinin planlanması, geliştirilmesi, uygulanması, analiz edilmesi ve yorumlanmasına yönelik bütüncül bir bakış sunmaktadır. ÖSYM uzmanları, akademisyenler, öğretmenler, lisans ve lisansüstü öğrenciler, alan uzmanları ve politika yapıcılar başta olmak üzere geniş bir okuyucu kitlesine hitap eden bu eser, ölçme ve değerlendirme alanında başvurulabilecek kapsamlı bir kaynak niteliğindedir.

TÜBA

ÖSYM

Türkiye Bilimler Akademisi
Turkish Academy of Sciences

Vedat Dalokay Caddesi No: 112 Çankaya 06670 ANKARA
Telefon : +90 (312) 442 29 03 **Faks :** +90 (312) 442 72 36

www.tuba.gov.tr
TUBAakademi
TUBAakademi
TUBAGencakademi
Turkiyebilimlerakademisi

ISBN 978-625-6110-89-2