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.

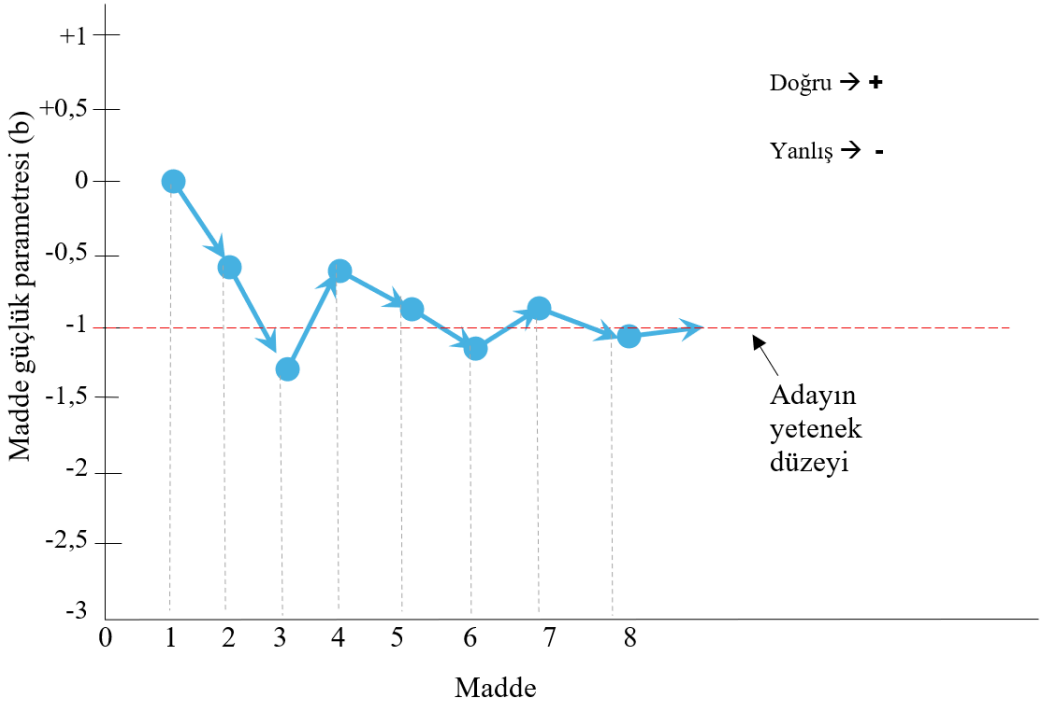
2. BİLGİSAYAR ORTAMINDA BİREYE UYARLANMIŞ TESTLER VE ÇOK AŞAMALI TESTLER

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

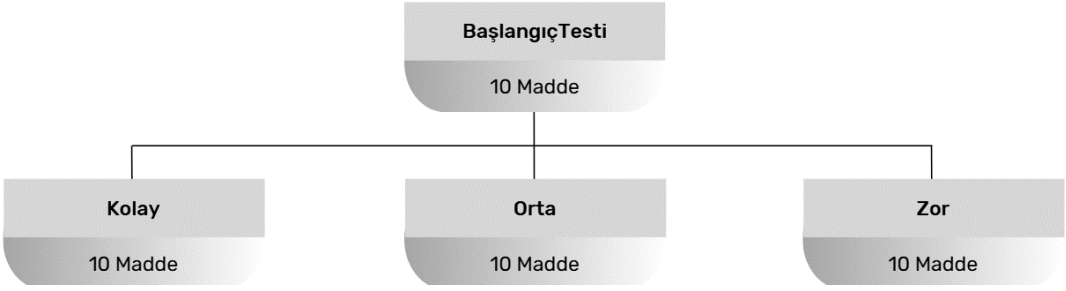


Şekil 1'de görüldüğü üzere adaya öncelikle orta güçlükte (b_1=0) bir madde uygulanır. Örnekte, aday birinci maddeyi yanlış yanıtladığı için ikinci madde olarak daha kolay bir madde yöneltilmiştir. Bundan sonra bireyin karşılacağın her yeni madde önceki yanıtınının doğru ya da yanlışlığına göre gelir. 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.

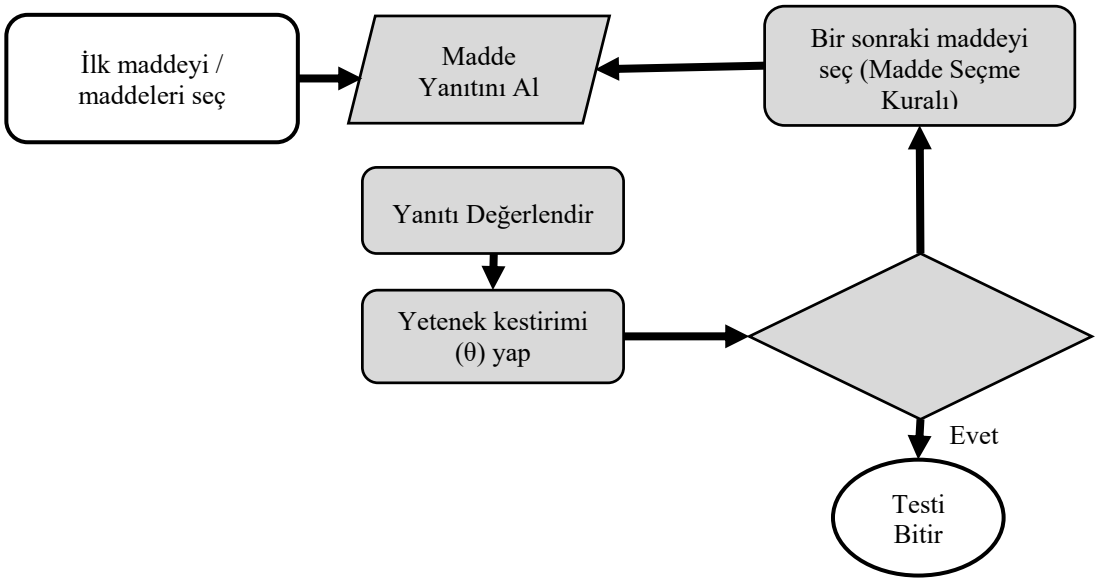
2.1. Bilgisayar Ortamında Bireye Uyarlanmış Testler

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

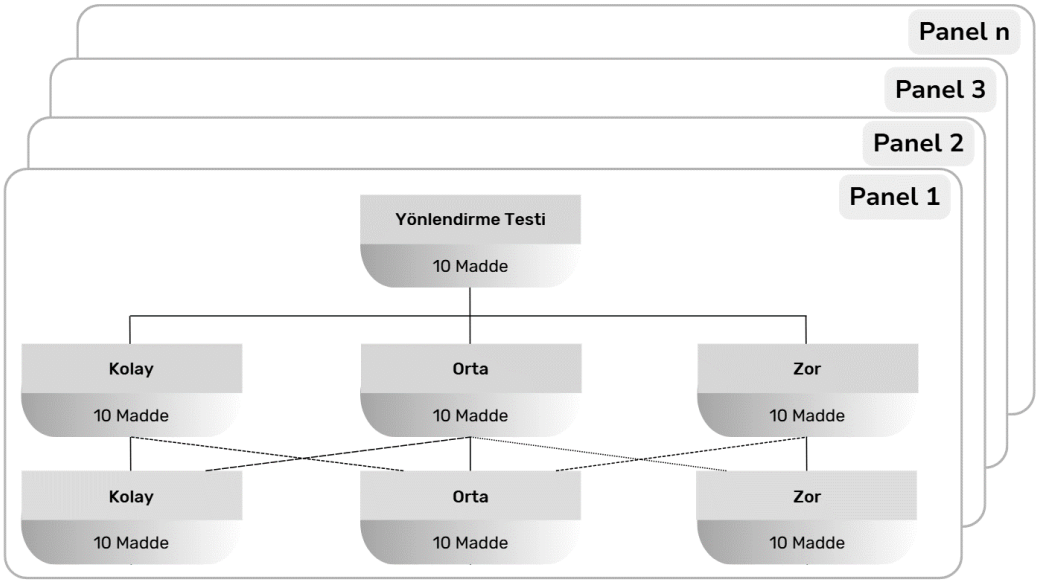
Bireye uyarlamanın madde tabanlı değil test tabanlı yapıldığı bireye uyarlanmış testler çok aşamalı testler (ÇAT) olarak adlandırılmaktadır. Başka bir ifadeyle testin adaya uygun ilerlemesi için her aşama sonunda bir uyarlama yapılır. Aday bir testle başlar, testten sonra elde edilen yetenek kestirimine 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)

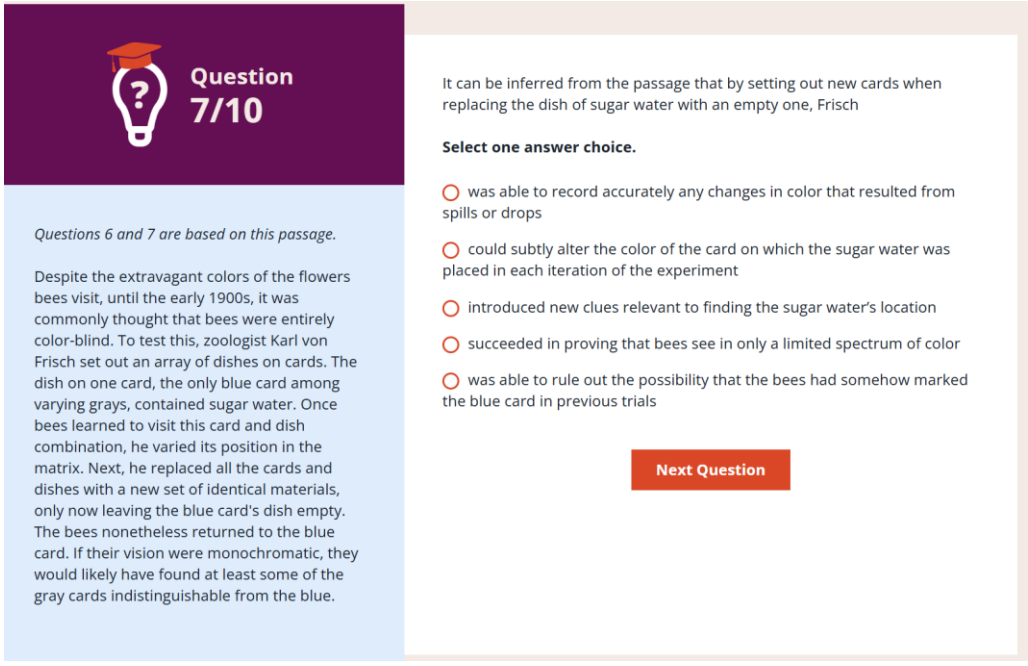
Ölçülen Özellik	Soru Sayısı	Süre
Analitik Yazma (Bir bölüm)	Bir "Bir Sorunu Analiz Et" görevi	30 dakika
Sözel Muhakeme (İki bölüm)	Bölüm 1: 12 soru Bölüm 2: 15 soru	Bölüm 1: 18 dakika Bölüm 2: 23 dakika
Nicel Muhakeme (İki bölüm)	Bölüm 1: 12 soru Bölüm 2: 15 soru	Bölüm 1: 21 dakika Bölüm 2: 26 dakika

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

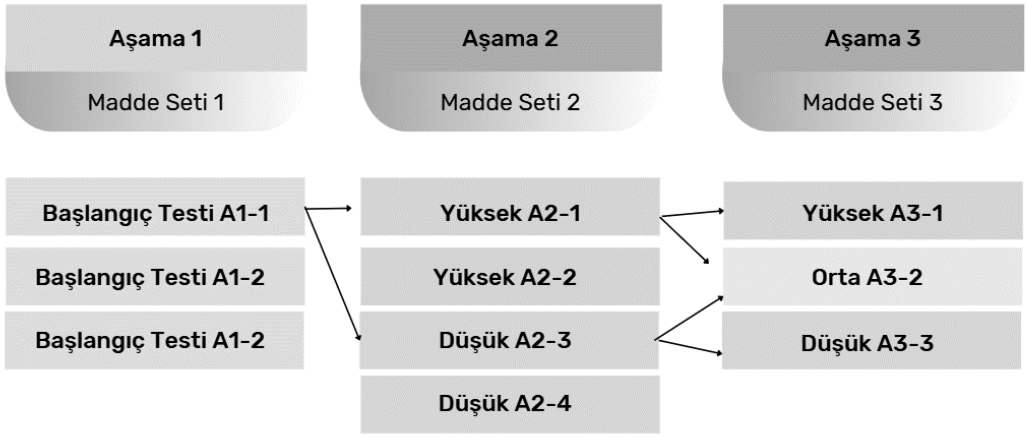
Uluslararası Öğrenci Değerlendirme Programı (Programme for International Student Assessment [PISA]), Ekonomik İşbirliği ve Kalkınma Örgütü (OECD) tarafından yürütülen ve 15 yaş grubundaki öğrencilerin okuma, matematik ve fen alanlarındaki bilgi ve becerilerini, gerçek yaşam problemlerini çözme bağlamında 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

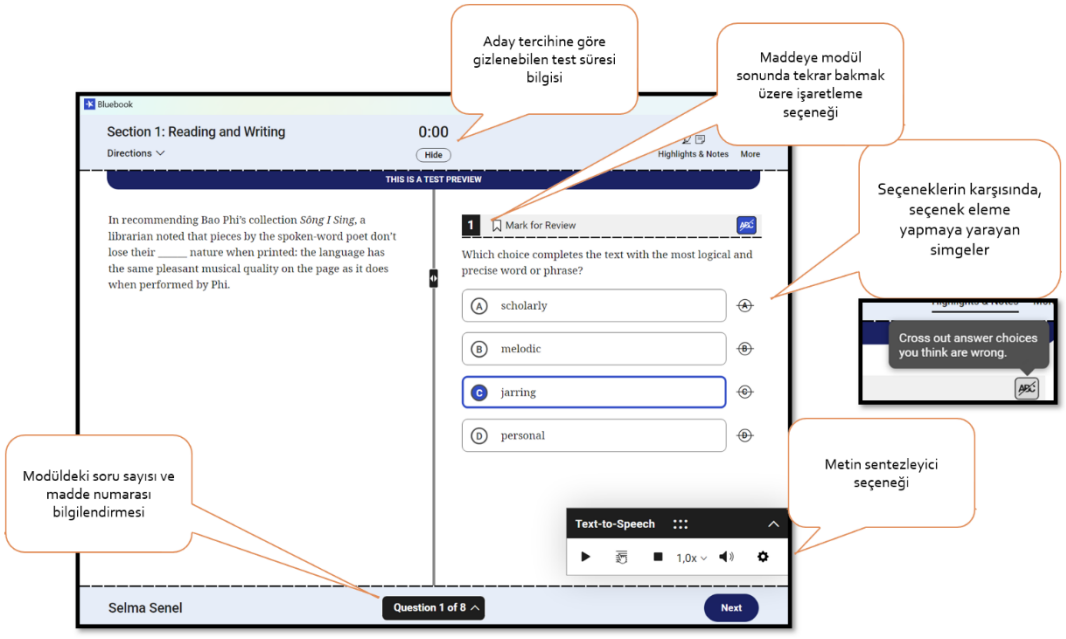
SAT (Scholastic Assessment Test), Amerika Birleşik Devletleri'nde üniversiteye girişte kullanılan ve öğrencilerin okuma, yazma, dil ve matematik alanlarındaki akademik hazırlık düzeyini ölçen bir testtir (College Board, 2025a). SAT'ın 2024'ten itibaren hayata

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

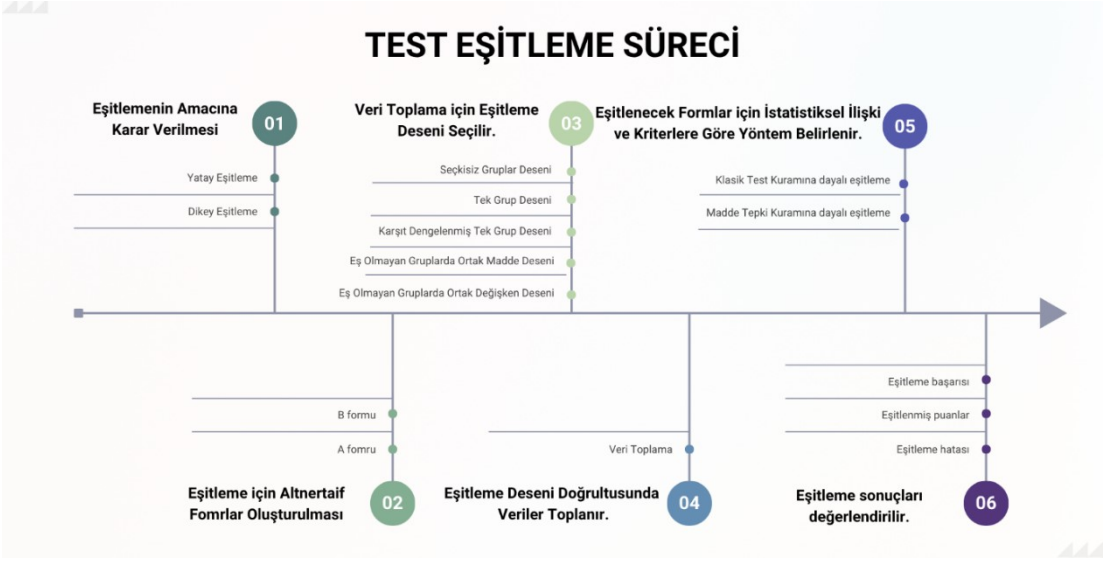
Belirlenen eşitleme ilişkisinin farklı demografik gruplar (cinsiyet, yaş, etnik köken vb.) arasında tutarlı olmasını ve örneklem özelliklerinden etkilenmemesini ifade eder.

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.

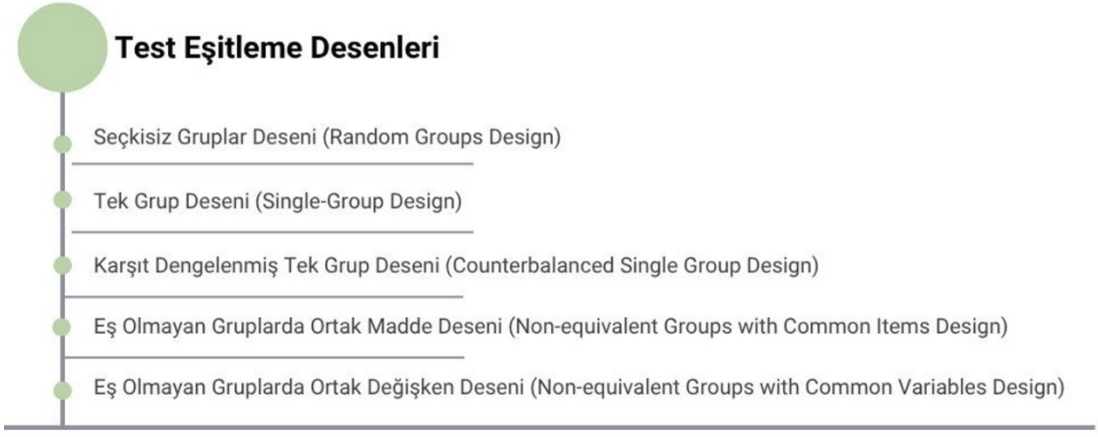
4.4. Test Eşitleme Desenleri

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

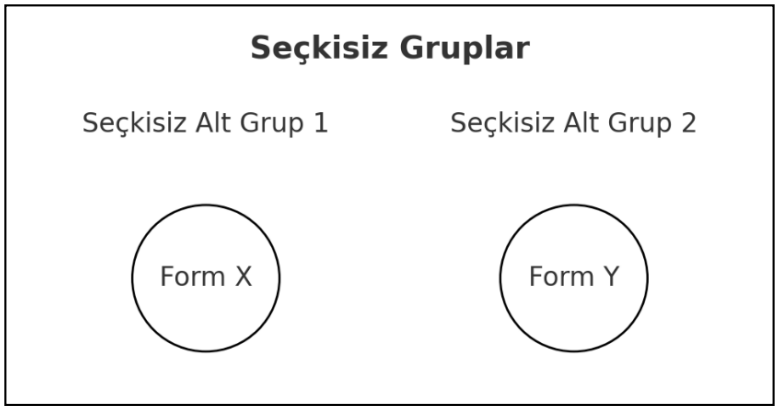


4.4.1. Seçkisiz Gruplar Deseni

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

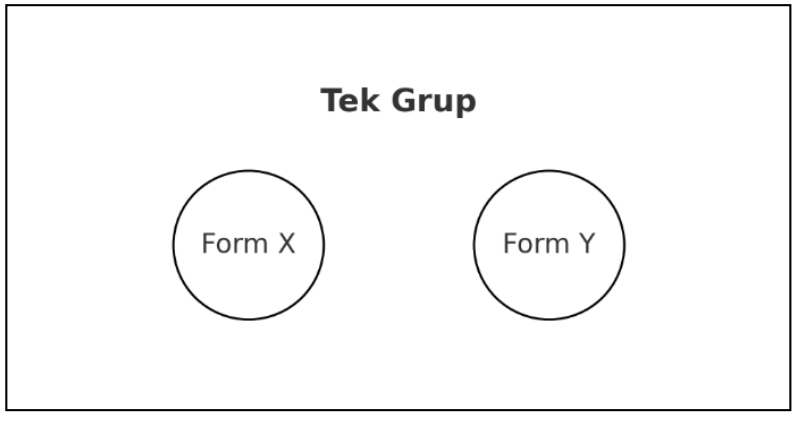
Seçkisiz gruplar deseninin en önemli avantajı, uygulamanın pratikliğidir. Bu yöntemde bir kişi sadece bir formdaki maddeleri cevaplandırır, bu da zamandan tasarruf edilmesini sağlar. Ayrıca desende katılımcılar sadece bir form cevaplandırdığı için sıra etkisi oluşmaz. Seçkisiz grup deseni kullanılarak 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

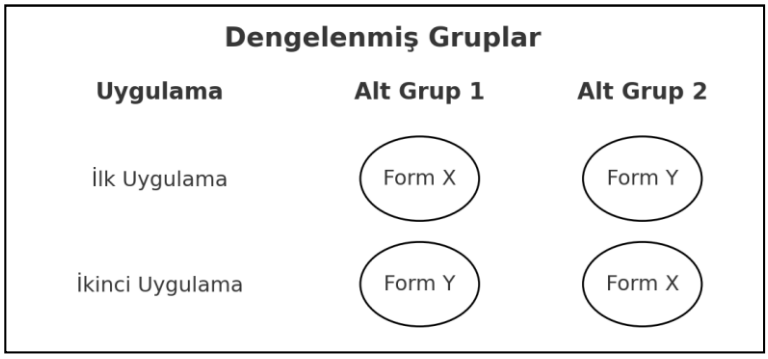
Tek grup deseninde her katılımcı, hem X hem de Y formunu aldığı için bireysel yetenek farklılıkları her iki form için de aynıdır. Ayrıca tek bir grup olduğu için seçkisiz grup desenine göre daha küçük örneklem gerektirir. Bununla birlikte aynı katılımcıya iki test formu uygulanacağı ve bu uygulamaların 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

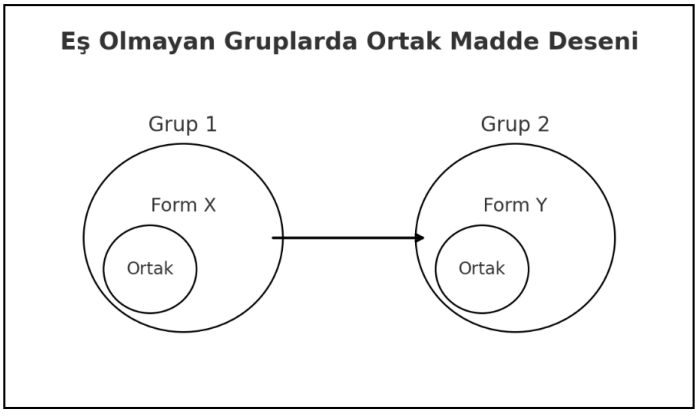
Dengelenmiş grup deseni, tek grup deseninin sıra etkisi sorununa çözüm getirmek için geliştirilmiş bir yaklaşımdır. Bu desende katılımcılar rastgele olarak iki alt gruba bölünür ve katılımcıların yarısı önce X formunu sonra Y formunu alırken diğer yarısı önce Y formunu sonra X formunu alır. Bu yaklaşım 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).

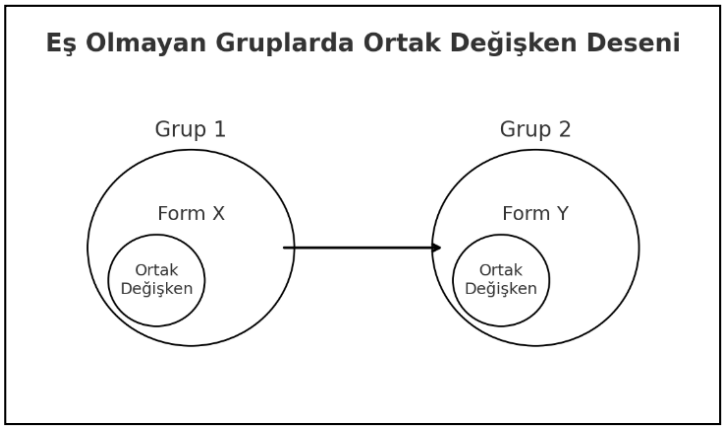
Test formları bazen ortak köke dayalı maddelerden oluşabilir. Madde takımı (testlet) adı verilen bu durumda yerel bağımsızlık varsayımını ihlal edebilir. Başka bir anlatımla bir maddenin doğru cevaplanması başka bir maddeden etkileniyor olabilir. Bu durum güvenirliği ve korelasyonu etkileyeceği için 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.