2. BÖLÜM

TEST KURAMLARI

TEST THEORIES

Prof. Dr. Tahsin Oğuz BAŞOKÇU

Prof. Dr. Nuri DOĞAN

TEST KURAMLARI

Prof. Dr. Tahsin Oğuz BAŞOKÇU
Ege Üniversitesi

Prof. Dr. Nuri DOĞAN
Hacettepe Üniversitesi

Öz

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

Anahtar Kelimler

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

TEST THEORIES

Prof. Dr. Tahsin Oğuz BAŞOKÇU
Ege University

Prof. Dr. Nuri DOĞAN
Hacettepe University

Abstract

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

Keywords

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

1. GİRİŞ

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

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

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

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

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

2. TEST KURAMLARININ GELİŞİM SÜRECİ

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

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

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

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

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

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

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

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

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

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

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

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

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

3. KLASİK TEST KURAMI

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

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

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

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

$X_i = T_i + E_i$

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

Eşitlik 1

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

$E_i = X_i - T_i$ *Eşitlik 2*

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

3.1. Hatalara İlişkin Varsayımlar

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

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

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

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

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

3.2. KTK'de Güvenirliğin Hesaplanması

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

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

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

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

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

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

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

Eşitlik 4

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

3.3. KTK'nin Sınırlılıkları

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

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

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

4. MADDE TEPKİ KURAMI

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

4.1. KTK ile MTK'yi Ayıran Özellikler

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

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

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

4.2. MTK Varsayımları

4.2.1. Monotonik Artış

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

4.2.2. Tek Boyutluluk

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

4.2.3. Yerel Bağımsızlık

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

4.2.4. Değişmezlik

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

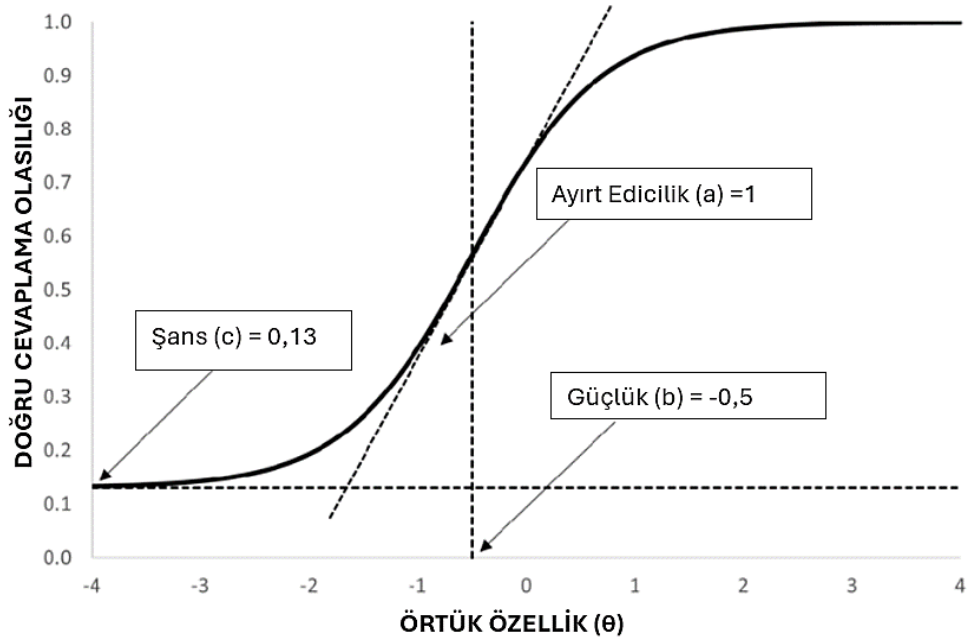
Hambleton ve arkadaşları (1991), MTK'de bir kişinin bir test maddesindeki ediminin-eyleminin, örtük özellik ya da yetenek olarak tanımlanabilecek bir faktör aracılığıyla tahmin edilebileceğini ve aynı zamanda bir kişinin bir maddedeki edimi-eylemi ile o maddeye cevap vermesini sağlayan özellik arasındaki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

4.3. Madde Karakteristik Eğrisi

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

Şekil 1

Örnek MKE

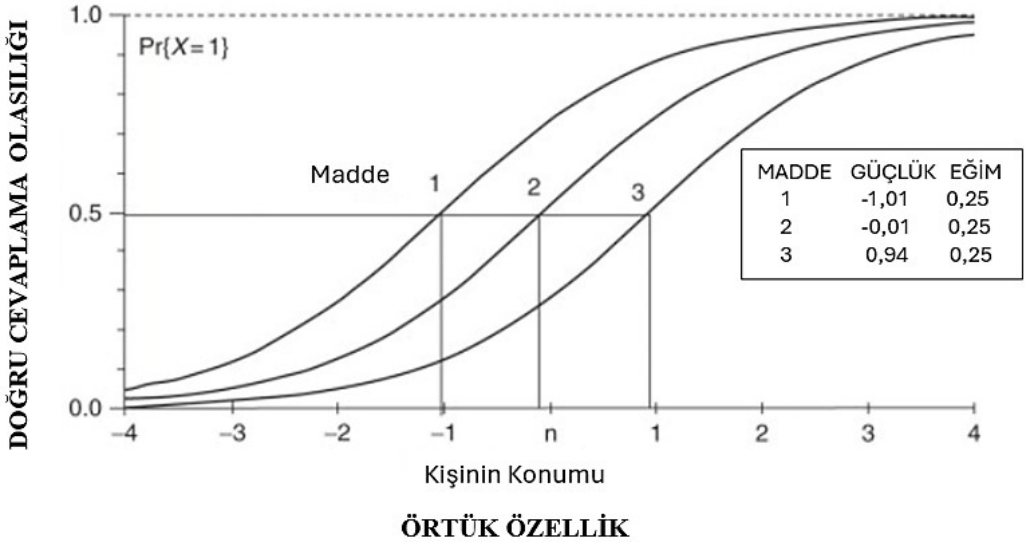


Şekil 1'deki grafik incelendiğinde madde güçlük parametresinin (*b* parametresi) –0.5 düzeyinde olduğu görülmektedir. Bu değer, fonksiyonda doğru cevaplama olasılığının 0.5'e (%50'ye) karşılık gelen örtük özellik (θ) düzeyinin değeridir.

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

Şekil 2

Farklı Güçlük Düzeyleri İçin MKE

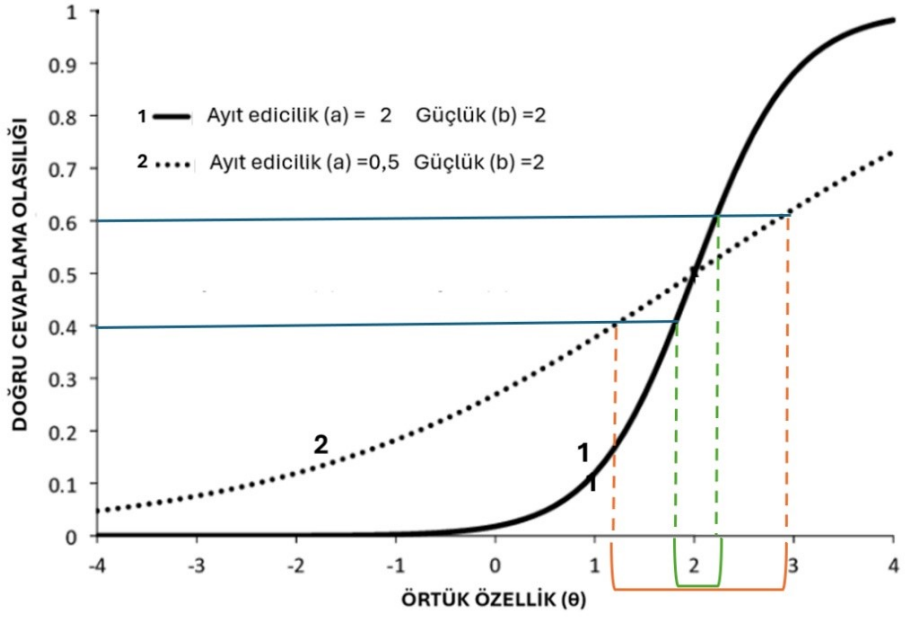


Şekil 2'deki MKE'ler hem maddelere ait farklı güçlük düzeylerini hem de farklı yetenek düzeylerindeki öğrencilerin konumlarını göstermektedir. Bu durum, MTK'de madde-yetenek etkileşiminin bir göstergesidir. Öte yandan 2. madde orta düzeyde güçlüğe sahip olup maddeler arasında en zor olanı 3., en kolay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

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

Şekil 3

Farklı Ayırt Edicilik Düzeyleri İçin MKE



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

4.4. MTK'de Güvenirlik ve Ölçme Hatası

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

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

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

4.5. MTK Modelleri

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

4.5.1. 1 Parametreli Lojistik Model

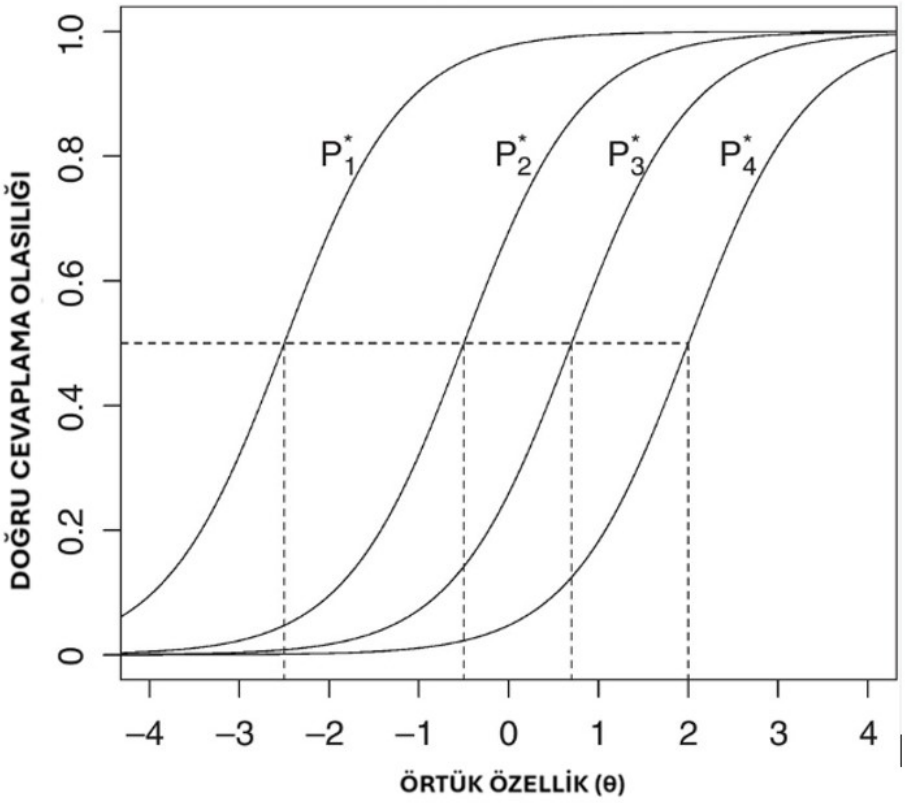
MTK modellerinin en basiti olan 1 Parametreli Lojistik Model (1PLM) maddelere yanıt veren kişinin örtük özelliğini (yetenek – θ) tanımlayan parametre ile maddenin zorluk düzeyini gösteren b (güçlük) parametresi üzerine kurulmuştur. Aşağıda 1PLM için kullanılan eşitlik verilmiştir (Baker & Kim, 2004).

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

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

Şekil 4

Farklı Madde Güçlükleri İçin 1PLM'de MKE'ler



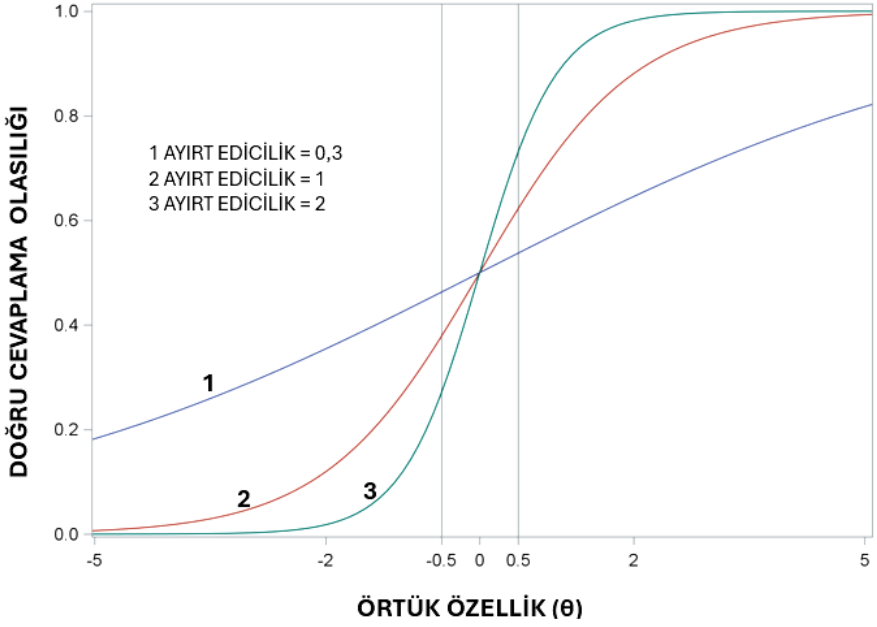
Şekil 4'te dört farklı yetenek düzeyini temsil eden MKE'ler verilmiştir. Eğimleri yani ayırt edicilikleri aynı olan bu maddelerin değişen tek özelliği maddeyi doğru yanıtlama olasılığının 0.5'ye karşılık geldiği noktadır. Bu noktalar ise farklı yetenek düzeylerine ya da o yetenek düzeylerine sahip öğrencileri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

4.5.2. 2 Parametreli Lojistik Model

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

Şekil 5

2PLM'de Farklı Ayırt Edicilikteki Maddeler İçin MKE



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

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

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

4.5.3. 3 Parametreli Lojistik Model

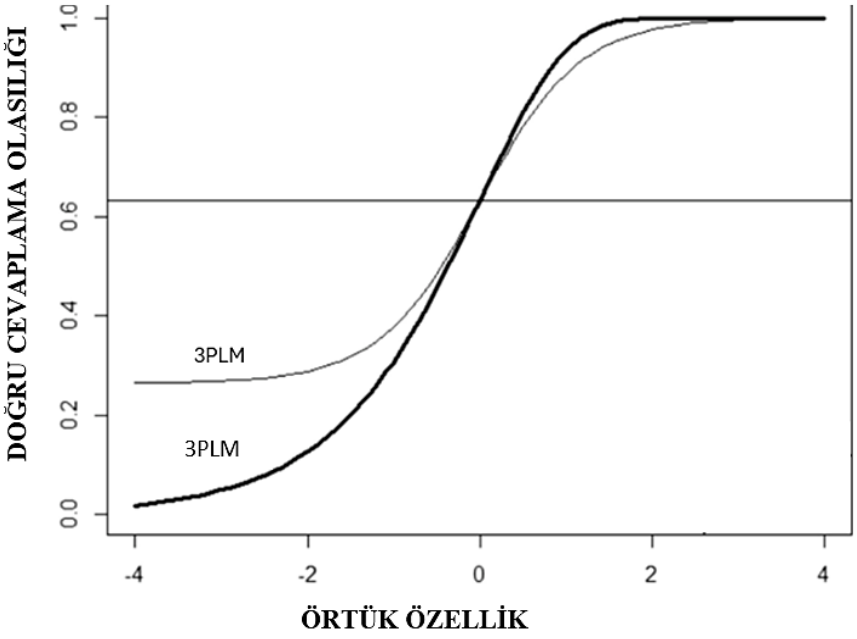
Bu model, 2PLM'ye şans (c) parametresinin eklenmesi ile oluşturulmuştur. Yani 3 Parametreli Lojistik Model'in (3PLM) hesaplamalarında güçlük (b), ayırt edicilik (a) ve şans (c) parametresi kullanılmaktadır (Baker & Kim, 2004). Modelin eşitliği aşağıdaki gibidir.

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

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

Şekil 6

3PLM İçin MKE



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

4.5.4. 4 Parametreli Lojistik Model

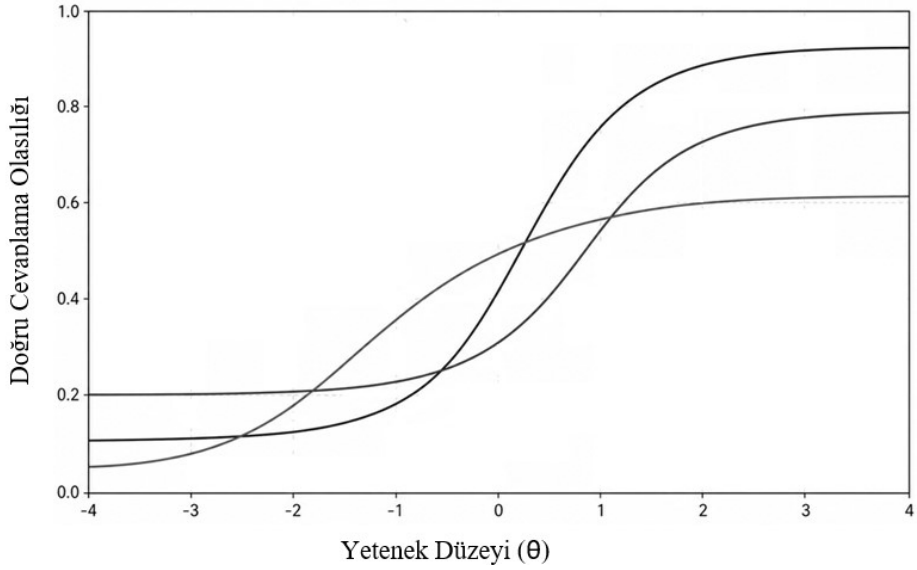
Dört parametreli lojistik model (4PLM) güçlük, ayırt edicilik ve şans parametrelerinin yanı sıra madde karakteristik eğrisinin üst asimptotuna karşılık gelen dikkatsizlik parametresini (*d*) de içermektedir. Yani bu model yüksek yetenek düzeyine sahip bireylerin dikkatsizlik ya da benzeri nedenlerle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

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

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

Şekil 7

4PLM İçin MKE



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

4.6. MTK ile KTK'nin Karşılaştırılması

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

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

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

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

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

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

5. BİLİŞSEL TANI MODELLERİ

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

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

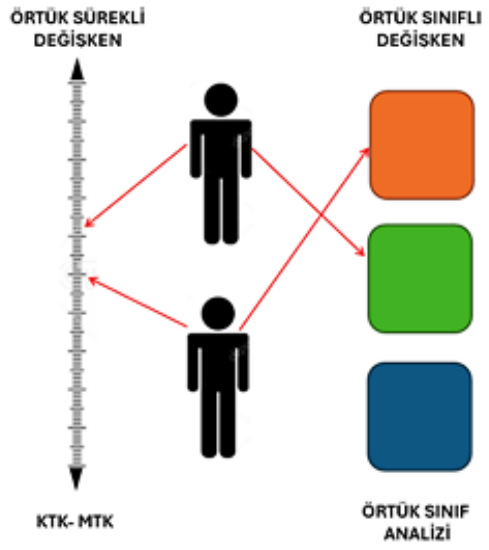
Bu noktada BTM'nin iki önemli özelliğine dikkat edilmelidir. Öncelikle BTM öğrenciyi sahip olduğu özellik bakımından sürekli bir değişken olarak raporlamayı, yani öğrencinin ölçülen özelliğine ait bir puan hesaplamayı öncelemez. Öğrencinin ilgili özelliğe sahip olup olmadığına odaklanır. Alanyazında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

5.1. Örtük Sınıf Analizi

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

Şekil 8

Sürekli ve Kategorik Ölçme Yaklaşımı



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

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

5.2. BTM'nin Genel Özellikleri

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

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

5.3. BTM Analizlerinde Q-Matris

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

Tablo 1

Örnek Q-matris

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

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

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

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

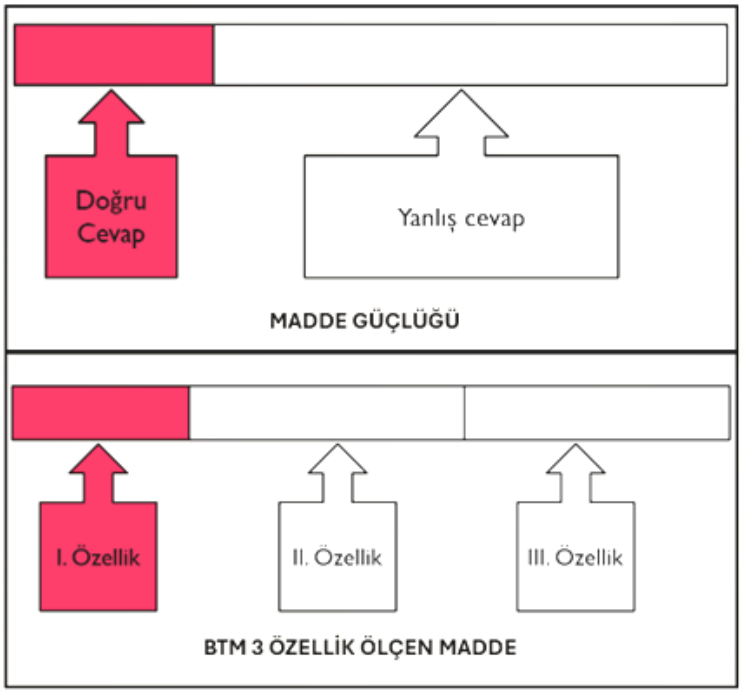
Maddelerin oluşturulması BTM analizleri için kritik önem taşımaktadır. Burada en önemli konu, Q-matrisini oluşturan özelliklerin doğru şekilde tanımlanmasıdır. Bir testte hangi becerilerin ölçüleceği, testin hangi öğrenci grubunun hangi özelliklerini ayırmak için kullanılacağı önceden belirlenmelidir.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

5.4. BTM'de Madde Özellik İlişkisi

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

Şekil 9

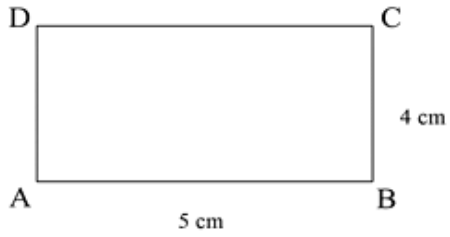
Maddeye İlişkin Geleneksel Yaklaşım İle BTM Yaklaşımına Ait Farklılık

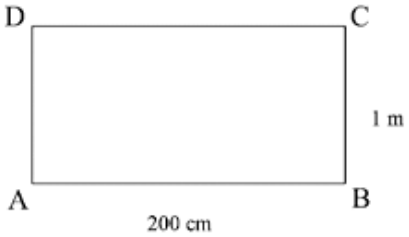


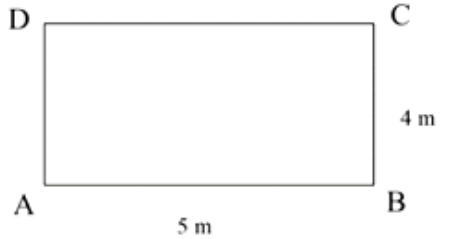
Şekil 9 incelendiğinde genel olarak kullanılan madde güçlüğü anlayışını yansıtmaktadır. KTK'de madde güçlüğü bir maddeyi doğru cevaplayanların sayısının, teste katılan toplam kişi sayısına oranı olarak tanımlanır. MTK'de ise madde güçlüğü, maddeyi doğru yanıtlama olasılığının %50 olduğu yetenek düzeyine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

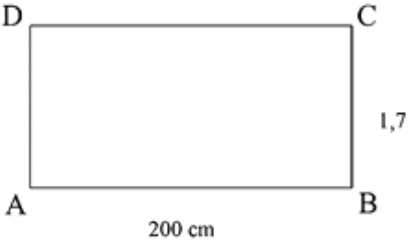
Şekil 10

Aynı Soru İfadesiyle Ölçülen Faklı Özelliklere İlişkin Örnek Maddeler

<table>
<tr>
<td>

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

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

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

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

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

5.5. BTM'de Madde Parametreleri

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

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

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

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

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

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

5.6. BTM'de Yetenek Kestirimi

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

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

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

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

Tablo 2

İki Özellik İçin Parametre Listesi

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

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

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

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

5.7. BTM Modelleri ve Türleri

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

Tablo 3

Bilişsel Tanı Modellerinde Kullanılan Teknikler

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

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

5.8. Bilişsel Tanı Modellerinin Genel Değerlendirilmesi

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

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

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

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

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

6. SONUÇ

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

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

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

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

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

KAYNAKÇA | REFERENCES

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Cambridge University Press.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

YAZARLAR HAKKINDA | ABOUT AUTHORS

PROF. DR. TAHSİN OĞUZ BAŞOKÇU

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

ORCID: 0000-0002-4821-0045

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

PROF. DR. TAHSİN OĞUZ BAŞOKÇU

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

ORCID: 0000-0002-4821-0045

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

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

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

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

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