21. Yüzyılda Eğitim ve Toplum / Education And Society In The 21st Century
Cilt / Volume 14, Sayı / Issue 42, Kış / Winter 2025
Araştırma Makalesi / Research Article

Sayfa/Page : 593 - 624

Test Geliştirmenin Dünü, Bugünü ve Yarını: Tarihsel Bir İnceleme

The Past, Present, and Future of Test Development: A Historical Review

Eray SELÇUK*

Öz

Test geliştirme süreci, psikometri alanının köklü ve önemli bir işidir. Tarihsel süreçte büyük bir gelişme gösteren bu alan, ölçme kuramlarından, standartların oluşturulmasından ve teknolojik ilerlemelerden önemli ölçüde etkilenmiştir. Bu çalışma, eğitimde ve psikolojide test geliştirme sürecine kapsamlı bir tarihsel bakış sunmaktadır. Test uygulamalarının tarihsel kökenlerinden günümüzdeki uygulamalarına kadar olan gelişimi incelenmiş ve bu alandaki potansiyel trendler ve yenilikler ortaya konmuştur. Buna göre, eğitimde ve psikolojide testlerin gelişimini keşfetmek için ilgili alanyazının kapsamlı bir incelemesi yapılmıştır. Kaynaklar arasında akademik dergiler, endüstri raporları ve test oluşturmanın tarihsel gelişimi, mevcut kuramlar ve gelecekteki eğilimleri kapsayan bilimsel makaleler yer almıştır. Bu derleme, test geliştirmenin izini sürerek, değerlendirmelerin eğitim ve psikoloji ortamını nasıl şekillendirdiği ve şekillendirmeye devam ettiğinin anlaşılmasını sağlamayı amaçlamaktadır. Bu amaçla test geliştirmenin dünü; test kuramları, test geliştirme süreci olarak, test geliştirmenin bugünü; geniş ölçekli ve öğretmen yapımı testler, test kuramlarına dayalı uygulamalar olarak, test geliştirmenin yarını; BOBUT, otomatik madde üretimi ve yapay zekâ destekli madde yazma olarak alt başlılarda incelenmiştir. Test geliştirme sürecinde etik ve mahremiyet konuları, YZ'nın dahi olduğu süreçte önyargıların test sonuçlarına yansıması büyük bir risk taşımaktadır. Test geliştirme alanı, kapsayıcı,

* erayselcuk84@gmail.com, ORCID: 0000-0003-4033-4219, Millî Eğitim Bakanlığı

Makale Bilgisi / Article Info
Geliş / Received : 15.09.2025 - Kabul / Accepted : 17.10.2025

verimli ve çeşitli ihtiyaçlara cevap verebilen ölçme ve değerlendirme sonuçlarına ulaşmak için yapay zekâ entegrasyonu, uyarlanabilir yöntemler ve etik durumları konu edinmeli ve bu yönde ilerlemelidir. Test geliştirme ve madde yazma uzmanlık gerektiren bir iştir. Bu nedenle bu alanda çalışacak araştırmacılara, küresel çapta kullanılan güvenilir ve geçerli kaynaklara başvurmaları ve teknolojik gelişmeleri takip etmeleri önerilmektedir.

Anahtar Kelimeler: Test geliştirme, Madde Yazma, Otomatik Madde Üretimi, BOBUT, Yapay Zekâ

Abstract

The test development procedure is a crucial part of the field of psychometrics. This field, which has undergone substantial development over time, has been significantly influenced by measurement theories, the standardization of procedures, and technological improvements. This research provides a comprehensive historical review of the test development procedure in education and psychology. The development of test administrations from their historical origins to their current applications has been examined, and potential trends and innovations in this field have been identified. Accordingly, a comprehensive review of the relevant literature has been conducted to investigate the development of tests in education and psychology. The sources included scholarly journals, industry reports, and articles covering the historical development of test construction, current theories, and future trends. This review aims to provide an understanding of how assessments have shaped and continue to shape the educational and psychological environment by tracing the history of test development. To this end, the past of test development (test theories and the test development procedure), the present of test development (large-scale and teacher-made tests, practices based on test theories), and the future of test development (CAT, automatic item generation, and AI-assisted item writing) were examined under subheadings. Ethical and privacy issues in the test development procedure, as well as the reflection of biases in test results, even in AI-based procedures, pose a significant risk. The field of test development should address AI integration, adaptive methods, and ethical issues in order to achieve inclusive, efficient, and diverse measurement and evaluation results, and should progress in this direction. Test development and item writing are specialized tasks. Therefore, researchers working in this field are advised to refer to reliable and valid sources used globally and to keep abreast of technological developments.

Keywords: Test development, Item writing, Automatic item generation, CAT, Artificial Intelligence

Giriş

Eğitim, bireyin bilgi, beceri, değer ve tutumlarında planlı yaşantılar yoluyla kalıcı değişiklikler meydana getirmeyi amaçlayan ve birey ile çevresi arasındaki etkileşimler sonucunda sürekli işleyen ve yaşam boyu devam eden bir süreçtir. Bu sürecin tüm adımları dikkatle izlenmeli, değerlendirilmeli ve dönütler alınmalıdır. Bu dönüt alma işi, ölçme ve değerlendirme uygulamalarıyla gerçekleştirilmektedir. Ölçme ve değerlendirme, eğitim sürecini izleme açısından en önemli bileşenlerden biridir. Eğitim yoluyla kazandırılmak istenen davranışların beklenen düzeyde öğrenilip öğrenilmediğine karar verebilmek için bu davranışların geçerlik ve güvenirliği yeterince yüksek ölçme araçlarıyla ölçülmesi gerekmektedir (Kutlu ve Altıntaş, 2021). Ölçme ve değerlendirme işlemiyle öğrencilerin öğrenmeleri değerlendirilmekte, öğretim uygulamaları biçimlendirilmekte ve eğitim politikaları yönlendirilmektedir (Wiliam, 2004). Ancak, ölçme ve değerlendirme uygulamalarının uzmanlık gerektirmesi ve bunların öğretim ve öğrenim üzerindeki uzun vadedeki etkileri daha derin bir incelemeyi gerektirmektedir (DeLuca vd., 2019).

Ölçme ve değerlendirme süreci, ölçmenin gerçekleştirilmesiyle başlar, değerlendirme ise bu ölçüm sonuçlarının anlamlandırılmasıdır. Ölçme işlemi, nesnel gözlemlerle nitel veya nicel düzeyde betimlemeler yapmaktır (Çelen, 2008). Ancak ölçme, sadece bir betimleme değil, belirli bir özelliğe uygun ölçütlere dayalı olarak değer atama sürecidir. Ölçme; ölçen kişi, ölçülen özellik, kullanılan araç ve bu öğeler arasında geçerli, güvenilir ve nesnel bir ilişki kurulmasına olanak tanıyan kuralları içerir (Erkuş, 2012). Ölçme işleminden sonra yapılan değerlendirme, ölçümlerin bir ölçüt/norm ya da bir hedef ile karşılaştırılarak karar verme sürecidir (Turgut ve Baykul, 2010). Ölçme ve değerlendirme, eğitim-öğretim süreçlerinin etkililiğini izlemeye, geri bildirim sağlamaya ve sürekli iyileştirmeye yönelik işlevsel bir araçtır. Ölçme ve değerlendirme, eğitimde ve psikolojide testler ve alternatif ölçme araçlarıyla yapılmaktadır.

Eğitim ve psikoloji alanındaki testler, bireylerin yetenekleri, bilgi düzeyleri ve gelişim potansiyellerinin ölçülmesi ve yorumlanmasında önemli bir araçlardır. Psikolojik ve eğitimsel testler, bireyler hakkında sistematik veri toplamak amacıyla araştırma ve profesyonel uygulamalarda yaygın biçimde kullanılan ölçme araçlarıdır (Muñiz ve Fonseca-Pedrero, 2019). Test, bireylerin belirli bir psikolojik ya da davranışsal özelliğini nicel ya da nitel olarak değerlendirmek amacıyla geliştirilen geçerli ve güvenilir ölçme aracıdır (Anastasi ve Urbina, 1997). Bu testlerin gelişimi, eğitim sistemlerinin değişen ihtiyaçlarını, bilimsel teorilerdeki ilerlemeleri ve teknolojik yeniliklerin entegrasyonunu yansıtan dinamik ve çok boyutlu bir süreci temsil etmektedir (Zhang ve Chang, 2016).

Eğitimde teknolojinin kullanımı, ölçme ve değerlendirme uygulamalarını önemli ölçüde etkilemiş ve eğitim sistemleri genelinde daha dengeli ve tutarlı iç içe geçmiş ölçme ve değerlendirme sistemlerinin geliştirilmesine olanak sağlamıştır. Bu yaklaşım, öğrenci öğrenmesine bütünsel bir bakış açısı sağlayan ve çeşitli düzeylerdeki öğretim uygulamalarını bilgilendiren daha kapsamlı ve bütünleşik bir değerlendirme sistemi oluşturmayı hedeflemektedir. Yapay zekâ (YZ) ve diğer teknolojik gelişmelerin test geliştirmeye dahil edilmesi, araştırmacılar ve uygulayıcılar arasında endişelere yol açmıştır. Bu yenilikler, değerlendirme sürecini kolaylaştırma ve kişiselleştirme

potansiyeline sahiptir, ancak bu tür yaklaşımların etik sonuçları ve potansiyel sınırlamaları hakkında geçerli endişeler vardır (Tippins, 2015).

Temel kaygılardan biri, yapay zekâ tarafından oluşturulmuş testlerden elde edilen verilerde ve bu süreçte kullanılan algoritmalarda gömülü önyargıları sürdürme eğilimidir. Elde edilen veriler veya modelleme süreçleri toplumsal önyargıları yansıtıyorsa, ortaya çıkan ölçme ve değerlendirme sonuçları belirli bireyleri haksız yere cezalandırabilir ve eşitlikçi ve kapsayıcı test hedeflerini baltalayabilir. Ayrıca, yapay zekâ karar verme mekanizmalarının şeffaf olmaması, adalet ve değişen madde fonksiyonu sorunlarının tespit edilmesini ve azaltılmasını zorlaştırabilir (Ferrara, 2023). Dahası, bilgisayar ortamında bireye uyarlanabilir testlere ve bireyselleştirilmiş ölçme ve değerlendirme yaklaşımlarına giderek daha fazla vurgu yapılması, sonuçların teste katılanlar arasında karşılaştırılabilirliği ve standardizasyondan ödün verme potansiyeli konusunda endişelere yol açmaktadır. Değerlendirme deneyimi giderek bireysel öğrencilere göre uyarlandıkça, anlamlı karşılaştırmalar yapma ve test puanlarına dayalı bilinçli kararlar alma becerisi daha zor hale gelebilir (Dadey vd.,2018).

Yapay zekanın gelişim süreci ilerledikçe, araştırmacılar ve uygulayıcılar teknolojik yeniliklerin etik ve pratik sonuçlarını dikkatle değerlendirmelidir. Adillik, şeffaflık ve karşılaştırılabilirlik gibi önemli standartlar korunurken inovasyonun faydalarından yararlanan dengeli bir yaklaşım, eğitimde ve psikolojide kullanılan testlerin pozitif yönde gelişmesini sağlayacaktır. Leach ve Oakland (2007), çalışmalarında 35 ülkeyi test geliştirme süreçlerinde kullanılan 31 etik kod üzerinden karşılaştırmışlardır. Bu ülkeler Avustralya, Kanada, Şili, Çin, Kosta Rika, Hırvatistan, Çek-Moravya, Dominik Cumhuriyeti, Estonya, Almanya, Birleşik Krallık, Hong Kong, Macaristan, İrlanda, İsrail, İtalya, Letonya, Litvanya, Malta, Hollanda, Yeni Zelanda, Filipinler, Polonya, Rusya, İskandinavya (beş ülkeyi temsilen), Singapur, Slovenya, Güney Afrika, İspanya, İsviçre ve Türkiye'dir. Çalışmanın sonucunda, test geliştirme ve kullanımı ile ilgili meslek etiği kurallarındaki benzerlikler, kuralları incelenen 20 ülkenin çoğunda bulunduğu görülmüştür. Bu benzerlikler, test uygulamalarını etkileyen konulara yönelik yaygın, hatta küresel ilgiyi yansıtmaktadır. Test geliştirme ve kullanımı yaygınlaştıkça ve etik kurallar yenilendikçe benzerliklerin devam etmesi beklenmektedir. Ulusal psikoloji derneklerinin, özellikle de Avustralya, Güney Afrika, Türkiye, Singapur, Şili ve Amerika Birleşik Devletleri'nin çeşitli etik kurallarının yapısı ve içeriği yüksek derecede benzerlik gösterdiği belirtilmiştir.

Test geliştirme süreci çeşitli adımlardan oluşan bir süreçtir. Teknolojik gelişmeler ve kuramsal araştırmalar bu adımlar üzerinde fazla bir değişiklik yaratmamıştır. Ancak test geliştirmenin güvenirliği ve geçerliği ile ilgili sorulara çeşitli çözümler üretmiştir. Tiffin-Richards ve Pant (2017), test geliştirme süreçlerinde geçerlik kavramını incelemişlerdir. Özellikle test puanlarının kullanılması konusunun testin geçerliğine nasıl etki edeceğini tartışmışlardır. Benzer şekilde Kelecioğlu ve Göçer Şahin (2014), geçmişten günümüze geçerlik kavramının ne anlama geldiğine yanıt bulmaya çalışmışlardır. Sonuç olarak geçerliğin bir kanıt toplama süreci olarak ele alınması konusunda uzlaşmaya varıldığını belirtmişlerdir.

Test geliştirme süreçlerindeki benzer adımlar uzun yıllardan beri kullanılmaktadır. Bu adımlar gerçekleştirilirken testin güvenirliğinden ve geçerliğinden taviz

verilmemesi gerekir. Geçmişten günümüze test geliştirme süreci önemli ölçüde değişiklik göstermese de birçok kuramsal ve teknolojik yeniliklerden etkilenmiştir. Bu çalışmadaki amaç, tarihsel süreçte test geliştirmenin nasıl etkilendiğinin incelenmesidir. Bu etkilenme sürecinde meydana gelen değişimlerin test geliştirme adımlarında yarattığı değişiklikler ve testin güvenirliği ve geçerliği adına geliştirilen yöntemler detaylı olarak incelenmiştir. Bu inceleme yapılırken alanyazındaki akademik dizinlerdeki dergiler (Web of Science, Scopus, TR-Dizin), endüstri raporları ve test oluşturmanın tarihsel gelişimi, mevcut kuramları ve gelecekteki eğilimlerini kapsayan bilimsel makaleler incelenmiştir. Çalışma bu açıdan klasik bir derleme türünde olup tarihsel bir akışa sahiptir.

Test Geliştirmenin Dünü

Modern test geliştirmenin temelleri, ilk psikologların ve eğitimcilerin öncü çalışmalarına kadar uzanmaktadır. Zekâ testleri gibi standartlaştırılmış ölçme araçlarının geliştirilmesi, bireysel farklılıkları ölçmeyi ve karşılaştırmayı amaçlamış, değerlendirmelerin standartlaştırılması ve ölçülen temel yapıların anlaşılmasındaki ilk zorlukların önünü açmıştır (Curtis ve Glaser, 1984).

Sir Francis Galton, Alfred Binet ve Charles Spearman gibi ilk isimlerin çalışmaları da dahil olmak üzere psikometrik teorilerin gelişimi, test tasarımı ve yorumunun istatistiksel ve teorik temelleri için zemin hazırlamıştır. Alan geliştikçe, araştırmacılar adalet, yanlılık ve ilk ölçme araçlarının sınırlılıkları gibi konulara eğilim göstermişlerdir. Bununla birlikte test ve madde yanlılığı, adalet, etik, geçerlik ve güvenirlik, zaman ve uygulamayla ilgili sorunları ele almaya ve test geliştirme ve uygulama için standartlar oluşturmaya devam etmişlerdir. Bu durum, test yapısının iyileştirilmesine ve daha detaylı psikometrik modellerin ortaya çıkmasına yol açmıştır (Hambleton vd., 2011). Buna göre ilk incelenmesi gereken test geliştirme sürecinin arka planında yer alan test kuramlarıdır.

Test Kuramları

Test kuramları, gözlenen değişkenleri ile gizil değişkenler arasında bir fonksiyon oluşturur. Daha ayrıntılı bir tanımla; test kuramlarının amacı, tanımlanmış psikolojik yapılar için ölçüm sonuçlarının hatalarından ne ölçüde etkilendiğini belirlemek ve bu hataları en aza indirebilecek yöntem ve uygulamaları keşfetmektir. Test kuramlarının amaçlarından biri uzmanların test kullanımı ve geliştirilmesindeki standart uygulamaların altındaki mantık ve matematiksel modelleri kavramalarını sağlamaktır (Crocker ve Algina, 1986). Test geliştirme süreçleri ve bunların psikometrik özellikleriyle ilgili konular üzerinde çalışmalar yapan psikometri biliminin tarihsel sürecinde yaygın olarak kullanılan iki ölçme kuramı bulunmaktadır. Bunlar, ilk olarak ortaya çıkan Klasik Test Kuramı (KTK), günümüzde giderek yaygınlaşan Madde Tepki Kuramı (MTK)'dır. Ancak, KTK'yı kapsayan ve onun uzantısı olarak kabul edilen Genellenebilirlik Kuramı ve temelleri uzak doğuya dayanan Sato Test Kuramı, test geliştirme süreçlerine önemli katkılar sunmuştur.

İlk olarak «gerçek puan teorisi" olarak da bilinen KTK, ölçme sonuçlarını anlamlandırmamıza, yorumlamamıza ve değiştirmemize imkân tanıyan bir ölçme çerçevesidir. KTK'nın temel varsayımlarından biri, her ölçümün hata payı içerdiği ve gözlemlerin mükemmel olmadığıdır (Desjardins ve Bulut, 2018). KTK, uzun yıllar boyunca ölçme ve test geliştirme alanında uzmanlara rehberlik etmesine rağmen, test eşitleme, değişen madde fonksiyonu analizleri, birey ve madde parametrelerinin gruptan bağımsız olarak tahmin edilmesi gibi bazı konularda yetersiz kalmıştır (Hambleton ve Swaminathan, 1985; Hambleton vd., 1991; Hambleton ve Jones, 1993). KTK ile ilgili ayrıntılı bilgiler ve matematiksel eşitlikler için Gulliksen (1950) tarafından yazılan "Theory of Mental Test" kitabının incelenmesi önerilmektedir.

Bu tür sınırlamaların üstesinden gelmek amacıyla 1930'lu yıllarda geliştirilmeye başlanan ve bilgisayar teknolojisinin ilerlemesiyle ivme kazanan Madde Tepki Kuramı (MTK), ölçme ve değerlendirme uzmanlarına ve psikometristlere yeni çözüm yolları sunmaya devam etmektedir. Pratik çözümler sunması nedeniyle günümüzde araştırmacılar tarafından MTK, daha fazla tercih edilmektedir (Embretson ve Reise, 2000). MTK, bireylerin temel özellikleri ile test maddeleri üzerindeki performansları arasındaki ilişkiyi modelleyen bir diğer etkili test kuramıdır. Bu kuram, test maddelerini kalibre etmek ve standart bir ölçekte bireylerin yeteneklerini tahmin etmek için bir çerçeve sunar. MTK, test geliştiricilere bireylerin yeterlilik seviyelerine uyum sağlayan özelleştirilmiş değerlendirmeler oluşturma yetkisi vererek daha kesin ve verimli ölçümler yapılmasını sağlar (Baker ve Kim, 2017). Başka bir ifadeyle, MTK, testteki maddelerin ölçtüğü yeteneklerle incelenen madde ve test performansı arasındaki ilişkiyi açıklayan kapsamlı bir istatistiksel teoridir (Hambleton ve Jones, 1993). MTK ile ilgili ayrıntılı bilgiler ve matematiksel eşitlikler için Baker (2001), De Ayala (2009), De Mars (2010), Embretson ve Reise (2000), Lord (1968, 1980, 1983) tarafından yazılan kaynakların incelenmesi önerilmektedir.

Genellenebilirlik kuramı, ölçme sonuçlarının farklı hata kaynaklarını analiz ederek güvenilirliğini inceleyen bir psikometrik kuramdır. KTK'nın ötesine geçerek, ölçüm hatalarını daha detaylı bir şekilde incelemeye olanak tanır ve bu hataların kaynağını ayrıştırır. Genellenebilirlik kuramı, test içindeki farklı değişkenlik kaynaklarından oluşan ölçüm hatalarını analiz ederek bir testin güvenilirliğini değerlendirmek için kullanılan bir test kuramıdır. Bu kuram, test puanlarının nasıl genellenebileceğini ve çeşitli koşullar altında nasıl değişebileceğini açıklar. Genellenebilirlik kuramı, test geliştiricilerinin potansiyel ölçüm hatası kaynaklarını belirlemelerine ve azaltmalarına olanak tanıyarak daha güvenilir ve geçerli ölçümler elde edilmesini kolaylaştırır (Brennan, 1992). Genellenebilirlik kuramıyla ilgili ayrıntılı bilgiler ve matematiksel eşitlikler için Brennan (1992, 2001), Shavelson ve Webb (1991) tarafından yazılan kitapların incelenmesi önerilmektedir.

Sato Test Kuramı, Japon psikolog Tatsuoka Sato tarafından 1970'lerde geliştirilen bir ölçme ve değerlendirme teorisidir. Bu kuram, özellikle öğrenme süreçlerinin analizine odaklanır ve geleneksel test kuramlarından farklı olarak öğrenci hatalarına dayalı bir ölçüm modelini temel alır. Sato'nun kuramı, öğrencilerin hatalarından yola çıkarak, öğrenme süreçlerini ve bilgi düzeylerini anlamaya yönelik bir yaklaşımdır (Çüm, 2019). Sato Test Kuramıyla ilgili ayrıntılı bilgiler ve matematiksel eşitlikler için Ohta vd. (1987) tarafından yazılan makalenin incelenmesi önerilmektedir.

Sato'nunki gibi test kuramlarının birincil amacı, ölçüm hatalarının tanımlanmış psikolojik yapılar için ölçüm sonuçlarını ne ölçüde etkilediğini belirlemek ve bu hataları en aza indirecek yöntem ve uygulamaları keşfetmektir. Test teorilerinin bir diğer amacı da uzmanların test kullanımı ve geliştirilmesindeki standart uygulamaların altında yatan mantık ve matematiksel modelleri anlamalarına yardımcı olmaktır. Bu nedenle test kuramlarıyla birlikte standart bir test geliştirme sürecinin olması da kaçınılmazdır. Hangi test kuramı kullanılırsa kullanılsın geçerli ve güvenilir bir test geliştirmenin en önemli adımı bir standart oluşturmaktır.

Test Geliştirme Süreci

Testler, eğitim ihtiyaçlarının ve öğretimin kalitesinin belirlenmesi, belli öğretim kademelerine öğrenci seçilmesi ve yerleştirilmesi, rehberlik ve klinik hizmetlerin yürütülmesi amaçlarıyla kullanılmaktadır. Buna göre testler, bireylerin kendi içinde ve kendi aralarındaki farkları hassas bir şekilde ortaya çıkarma amacını gütmektedir. Psikolojik testler, bireyin davranış örneklemini ölçmektedir. Bu davranış örneklemleri, bireyler arasındaki farkları gözlemlememizi sağlar (Price, 2016). Buna göre testleri, eğitimde ve psikolojide bireyin özelliklerini standart koşullar altında belirlemeye yarayan ölçme araçları şeklinde tanımlayabiliriz (Anastasi, 1988).

Öğretme-öğrenme sürecinin kontrol edilmesi ve sağlıklı işletilebilmesi için güvenilir ve geçerli ölçme değerlendirme yöntemlerinin kullanılması gerekmektedir (Miller vd., 2009). Güvenilir ve geçerli ölçme, ancak doğru geliştirilmiş ve belirli adımlar izlenerek işletilmiş bir ölçme aracıyla sağlanabilir. Bireylerin ölçülecek özelliği iki farklı türden ele alınabilir. Bunlar, maksimum performans ve tipik özelliklerdir. Ölçme araçlarını sınıflandırma kriteri olarak bu özellikler kullanılabilmektedir. Buna göre ölçme araçları, maksimum performansı ölçen testler, tipik tepkileri ölçen ölçekler olarak iki başlık altında toplanabilmektedir (Cronbach, 1990).

Güvenilir ve geçerli bir test geliştirmenin belirli adımları vardır. Test geliştirme, test puanlarını yorumların ve bu yorumlara dayalı kullanımların geçerliliğini sağlamak için sistematik bir yaklaşım gerektirir (Downing ve Haladyna, 2006; Thorndike, 1982). Bu sistematik yaklaşımın adımları şu şekildedir (Crocker ve Algina, 1986):

1. Test sonuçlarının ne amaçla kullanılacağının belirlenmesi,
2. Ölçülmek istenen alanı ya da yapıyı temsil eden davranışların belirlenmesi,
3. Test yönergelerinin hazırlanması ve ikinci adımda belirtilen her bir davranış için madde oranlarının (belirtke tablosunun) oluşturulması,
4. Madde havuzunun (madde bankasının) oluşturulması,
5. Maddelerin alansal ve dilsel iyileştirmelerinin yapılması,
6. İyileştirme amaçlı ön denemelerin gerçekleştirilmesi,
7. Testin uygulanacağı örnekleme benzer geniş bir grubun oluşturulması,
8. Madde analizlerinin yapılması ve kriterleri karşılamayan maddelerin testten çıkarılması,

9. Son formun güvenirlik ve geçerlik durumlarına kanıtlar toplanması ve raporlanması,
10. Testin uygulanması, puanlanması ve yorumlanmasıyla (kesme puanları, performans standartları) ilgili kılavuzların hazırlanmasıdır.

Lane vd. (2012), bu adımlara genel planlama ve test güvenliği adımlarını ekleyerek 12 bileşene kadar çıkarmaktadır. Bu adımlara göre;

- Denemelik testin belirli bir amaç doğrultusunda kullanılabilecek güvenirlikte ve geçerlikte bir test olup olmadığı,
- Bu testin güvenirliği ve geçerliği artırılmış bir test durumuna getirilip getirilemeyeceği ve
- Testin güvenilirliği ve geçerliği artırılabilecekse bunun için neler yapılması gerektiği ortaya konabilir.

Yukarıda sözü edilen gerekliliklerin karşılanmasına test geliştirme denilmektedir (Özçelik, 1981). Özellikle testin ölçmek istediği odak özelliğin miktarının belirlenmesinde ve bu özelliğin ortaya çıkarılmasında etken olan kısım maddeler boyutudur. Bir testin tek başına puanlanabilen en küçük birimi madde olarak ifade edilir (Turgut ve Baykul, 2010). Madde, ilgili özelliğin bir parçasını uyaran, kendi başına bir anlam ifade eden ve bir puan (değer) atfedilen bir ölçeğin en küçük parçasıdır; ancak birimi değildir (Erkuş, 2012). Zihinsel özelliklerin incelenmesinde kullanılan bir test maddesi, uyarıcı ve cevaplama için kuralcı bir biçimi olan bir ölçüm birimidir ve sınava giren kişiden bazı psikolojik yapılardaki (bilgi, yetenek, yatkınlık veya özellik gibi) performansın çıkarılabileceği bir yanıt elde etmeyi amaçlar (Osterlind, 1990). Maddelerin amaca uygun ve ölçülmek istenen davranış evrenini doğru bir şekilde örneklemesi testin geçerliği ile doğrudan ilişkilidir. Geçerlik, testleri değerlendirmedeki en önemli psikometrik özelliklerden biridir. Bu özellik, test puanlarına dayalı spesifik çıkarsamaların uygunluğuna, anlamlılığına ve kullanışlığına dikkat çeker. Bir testin geçerli bir hal alması, bu tür çıkarsamaları desteklemek için kanıt toplama işlemidir. Test ve madde puanlarından birçok çıkarsamalar yapılabilir ve belirli bir çıkarsamayı destekleyen kanıtların toplanması için çeşitli yollar vardır. Bununla birlikte geçerlik, tek bir kavramdır. Kanıtların çeşitli yollarla elde edilebilmesine karşılık; geçerlik, daima verilerin test ve madde puanlarına dayalı çıkarsamaları ne ölçüde desteklediğini gösterir. Testin kendisinin değil, testin özel kullanımlarına ait çıkarsamaların geçerliği ortaya konur (AREA, APA ve NCME, 1997).

Test geliştirme süreci belirli bir zaman dilimiyle kısıtlı bir süreç ya da belirli adımların izlenmesiyle tamamlanan bir süreç değildir. Eğitimde ve psikolojide ölçme standartlarının (1997, s.44) 4.9. maddesinde; "Zaman içinde ortak ve genel bir ölçek/test kullanımını sürdürmek amacını güden sürekli değerlendirme programlarında kullanılan ölçeğin/testin tutarlığına, kararlılığına ve geçerliğine yönelik periyodik kontroller yapılmalıdır" şeklinde belirtilmektedir. Bu ifadeyle, testin ya da ölçeğin uygulandığı her grup için ortaya çıkan yeni bulguların rapor edilmesi gerektiği vurgulanmıştır. Bu şekilde, testin ölçtüğü yapıdaki ve madde parametrelerindeki değişimlerin gözlenmesi gerekliliği ortaya çıkmaktadır.

Test geliştirme bir amaçla başlar. Bu amaç, testin ölçmek istediği davranışı örnekleme işidir. Test amacının belirlenmesi, test geliştiriciye testin özelleştirilmesi, maddelerin geliştirilmesi, denenmesi ve gözden geçirilmesi için bütüncül bir çerçeve sağlamaktadır (Millman ve Greene, 1989). Bununla birlikte, testin amacı elde edilen puanların nasıl kullanılacağıyla ilgilidir. Buna göre test sonucundan elde edilen puanlar tanılama, biçimlendirme ve değer biçme gibi eğitimsel kararlar, rehberlik ve danışmanlığın meslekî, eğitimsel ve kişisel alanlardaki kararları, seçme, sınıflama ve öğretim programı planlama gibi politik kararlar ve araştırmalarla ilgili kararlar almak amacıyla kullanılmaktadır (Mehrens ve Lehman, 1991).

Yeni bir test programı için madde geliştirmenin başlangıcında, madde bankası boştur. Bu noktada, madde bankasını doldurmak için bir plan oluşturulmalıdır. Mevcut maddeleri olan test programları için bile bir plan yapmak her zaman yararlıdır. Plan, ihtiyaç duyulan madde sayısı, kullanılacak madde formatlarının türleri, bu madde formatlarının kullanım gerekçeleri, amaçlanan içerik ve bilişsel talep türü, madde geliştirme ve doğrulamadan sorumlu personel ve en önemlisi bir zamanlama gibi birçok prosedür ve konuyu içermektedir (Haladyna ve Rodriguez, 2013).

İlk başlarda madde yazma işi bir sanat olarak görülmüştür (Ebel, 1951). Ancak, hangi test kuramı kullanılırsa kullanılsın kaliteli maddelerden oluşan yeterince geniş bir madde havuzundan seçilen maddelerle oluşturulmuş bir test daha hatasız, dolayısıyla daha geçerli ölçmeler yapılmasına olanak vermektedir. Etkili madde yazmak hem bilimi hem de sanatı yansıtan bir beceridir. Öğrenilmesi ve uygulanması gerekmektedir. Bu nedenle madde yazarları, bir atölye çalışmasından geçmelidirler. Çoğu tecrübesiz madde yazarı, temel olgu ve kavramları tanıma becerisini ölçen kusurlu maddeler oluşturur (Lane vd., 2012). Bunun nedenleri arasında, alan bilgisi eksikliği, hedef davranışla ilgili bilgi eksikliği, testin geliştirildiği dilin yapısıyla ilgili eksiklikler ve ölçme ve değerlendirme bilgisinin eksikliği gösterilebilir. Başlangıçta kusurlu yazılan maddeler ölçülen yapıya bir varyans katmamaktadır (Kane, 2012; Selçuk vd., 2024). Standartlaştırılmış testlerin yapı taşları olan test maddeleri, test puanlarının amaçlanan yorumlarının ve kullanımlarının geçerliliğini desteklemek için dikkatli ve sistematik bir şekilde geliştirilmelidir (Reckase, 2010).

Maddeler yazıldıktan sonra yapılması gereken adım maddelerin iyileştirilmesidir. Redaksiyon kavramı tam anlamıyla madde iyileştirmenin karşılığı değildir. Redaksiyon, "yazılmış bir metin üzerinde gereken düzeltmeleri yaparak yazıyı yayıma hazır duruma getirme" olarak tanımlanmaktadır (TDK, 2024). Halbuki madde iyileştirme süreci şu şekilde sıralanabilir (Baykul, 2015):

- Her bir test maddesinin ölçülmek istenen yapıyı ölçebilme kalitesinin olup olmadığının,
- Bilimsel ve alansal doğruluğu sağlayıp sağlamadığının,
- Hazırlandığı dil açısından sade ve anlaşılır olup olmadığının,
- Testin ve maddelerin ölçme ve değerlendirme açısından hatalı olup olmadığının kontrolleri yapılmalıdır.

Madde iyileştirme aşamasından sonra maddeler deneme uygulamasına sokulmalı ve madde analizleri gerçekleştirilmelidir. Denemelik formda yer alması istenen madde sayısı son testte bulunması istenen madde sayısının en az üç katı kadar hazırlanmalıdır (Crocker ve Algina, 1986). Pilot uygulamasının yapılacağı grubun büyüklüğü test geliştirme aşamasında kullanılan kuramsal temele bağlıdır. Örneğin, KTK kullanılarak geliştirilen bir test için madde sayısının beş katı kadar büyüklük yeterli olabilecek iken MTK kullanılarak geliştirilen bir test için en az 500 ve üzeri büyüklüğünde bir örnekleme ulaşılması gerektiği tavsiye edilmiştir (De Mars, 2010; Hambleton ve Jones, 1993).

Testin deneme uygulamasından sonra elde edilen verilerden madde analizleri yapılmalıdır. Madde analizinde bireylerin maddelere verdikleri yanıtlara ve test puanlarına dayanarak öncelikle bireyler hakkında değil maddeler hakkında bilgi sağlanması amaçlanmalıdır (Livingston, 2006). Hesaplanan klasik madde analizlerinin başında, madde güçlük ve madde ayırt edicilik parametreleri gelmektedir. Klasik madde analizi, madde performansları ile test puanı arasındaki korelasyona dayanmaktadır (Lord ve Novic, 1968). Madde analizleri sonucunda istenilen güçlük ve ayırt edicilik düzeyine sahip maddeler son forma yerleştirilir.

Son formun uygulanmasından sonra madde ve test istatistikleri hesaplanır.

Özellikle son kısımda geliştirilen teste ait güvenirlik ve geçerlik kanıtları aranır. Bir testin güvenirliği, aynı koşullar altında bireylerin aldıkları puanlarda tutarlık göstermesidir (Carmines ve Zeller, 1979). Geçerlik, bir ölçme aracının ölçmeyi amaçladığı yapıyı, başka yapılarla karıştırmadan, titiz ölçebilme kabiliyetidir (Tekin, 1977). Testlerin sahip olması gereken en önemli iki geçerlik türü; kapsam ve yapı geçerliğidir. Güvenilir ve geçerli ölçme gerçekleştiren testler için bundan sonraki adım testin yönergesinin ve normlarının oluşturulmasıdır. Eğer gerekiyorsa testin uygulayıcıları bir eğitimden geçirilmelidir.

Test geliştirme ile ilgili alanyazın incelendiğinde uluslararası ve ulusal düzeyde birçok kaynak olduğu görülmektedir. Uluslararası alanyazında test geliştirme ve madde yazma üzerine Millman ve Greene (1989), Irvine ve Kyllonen (2002), Osterlind (1990; 2002), Haladyna (2004), Gierl ve Haladyna (2012), Lane vd. (2012), Haladyna ve Rodriguez (2013)'in yayınları mevcuttur. Bununla birlikte Crocker ve Algina (1986), Ebel ve Frisbie (1991), Mehrens ve Lehman (1991), Leighton ve Gierl (2007), Miller vd. (2009), Cohen ve Swerdlik, 2010, Thorndike ve Hagen (1964), Price (2016) gibi birçok ölçme ve değerlendirme kitabında test geliştirme ve madde yazma bölümleri bulunmaktadır.

Türkiye'de ilk test geliştirme madde yazma kitabı Özçelik (1981) tarafından yazılmıştır. Aynı zamanda ulusal alanyazında Erkuş (2012), Turgut ve Baykul (2010), Tekin (1977) gibi öncü yazarlar tarafından yazılan ölçme ve değerlendirme kitaplarında test geliştirme ve madde yazma bölümleri bulunmaktadır. Ayrıca ulusal alanyazında yazılan diğer ölçme ve değerlendirme kitaplarında da test geliştirme süreci ve farklı madde türlerinin hazırlanmasıyla ilgili bölümler yer almaktadır.

Ülkemizde ve yurtdışında hemen hemen tüm pozitif (fen bilimleri, mühendislik, sağlık vb.) ve sosyal bilimler (psikoloji, sosyoloji, eğitim, tarih vb.) alanlarında test geliştirme çalışmaları yapılmaktadır. Test geliştirme sürecinin doğru yürütülmesiyle

ilgili olarak alanyazında Uluman (2023), çalışması bulunmaktadır. Bu çalışmada Uluman (2023), 2020 ve 2023 yılları arasında başarı testi geliştirilen 40 çalışmayı, Turgut ve Baykul'un (2010) test geliştirme adımlarına göre incelemiştir. Sonuç olarak incelenen makalelerin birçoğunda test geliştirme adımlarının doğru yürütülmediği belirtilmiştir. Benzer şekilde Şahin vd. (2023), 2015-2020 yılları arasında matematik alanında başarı testi geliştiren 39 makaleyi test geliştirme süreci açısından beş uzmanla birlikte değerlendirmişlerdir. Sonuç olarak birçok çalışmada, madde havuzu oluşturma, pilot uygulama, madde analizi gibi konularda eksiklikler olduğunu tespit etmişlerdir. Test geliştirme çalışmalarının ortaya çıkış psikometri bilimiyle paralellik göstermektedir. Psikometrik kuramlar test geliştirmenin arka planındaki bakışı farklılaştırmakta ve sürecin adımlarında değişikliklere neden olmaktadır.

Test Geliştirmenin Bugünü

Teknolojik gelişmelerin ölçme ve değerlendirme süreçlerine entegrasyonu ve modern psikometrik teorilerin uygulanması, test geliştirmenin günümüzdeki durumunu etkilemektedir. MTK'nın ve BOBUT'un yükselişi, testlerin tasarlanma, uygulanma ve puanlanma şekillerinde devrim yaratarak daha bireyselleştirilmiş ve verimli ölçme ve değerlendirme deneyimlerine olanak sağlamaktadır. Bununla birlikte, test geliştirmenin artan karmaşıklığı, adalet, yanlılık ve değişen madde fonksiyonu konularının ele alınması gibi yeni zorlukları da beraberinde getirmiştir. Araştırmacılar ve uygulayıcılar, değerlendirmelerin kapsayıcı ve temsili olmasını ve toplumun çeşitli ihtiyaçlarını karşılamasını sağlamak için çalışmaktadır.

Test geliştiricilerin, test geliştirme süreçlerini sistematik bir şekilde yürüttüğü ve eğitimde karar verici durumunda olan ölçme işleminin gerçekleştiği iki test türü mevcuttur. Bunlar, standart testler olarak da adlandırılan geniş ölçekli testler ve daha dar uygulama alanına sahip öğretmen yapımı testlerdir.

Geniş Ölçekli (Standart) Testler ve Öğretmen Yapımı Testler

Geniş ölçekli testler, büyük öğrenci gruplarına uygulanan ve okul öğrenmelerinin izlenmesi ya da öğrencilerin bir üst eğitim kurumuna seçme ve yerleştirme amacıyla yapılan çalışmalarda çeşitli derslere yönelik öğrenme çıktılarının ölçülmesinde kullanılan araçlardır (Çakan, 2003). Geniş ölçekli değerlendirmeler, eğitim ve psikolojik olgular hakkında kapsamlı ve temsili veriler elde etmek için çok önemlidir. Çeşitli katılımcılardan oluşan bir havuza dayanan bu yapılandırılmış testler, bulguların daha geniş gruplara genellemesine olanak tanır.

Elde edilen veriler, politika yapıcılar, eğitimciler ve araştırmacılar için paha biçilemez bir kaynaktır. Karar alma süreçlerini bilgilendirir ve etkili müdahale ve politikaların geliştirilmesine rehberlik eder. Aynı zamanda, bu büyük ölçekli değerlendirmelerin oluşturulması ve uygulanması büyük bir özen ve dikkatle ele alınmalıdır. Değerlendirmelerin temsili, tarafsız ve ilgili birey ve toplumların refahı ile uyumlu olmasını sağlamak için adalet, nesnellik ve etik sonuçlara ilişkin endişeler göz önüne alınmalıdır. Bu endişeleri gidermek için devam eden çabalar, geniş ölçekli eğitimsel

ve psikolojik ölçümlerin bütünlüğünü ve güvenilirliğini korumak için çok önemlidir (Duque vd., 2023).

Öğretmen yapımı testler, sınıf içi değerlendirme ve öğretim amaçlarına yönelik olarak öğretmenler tarafından geliştirilen testlerdir. Öğretmenler, öğrencilerinin öğrenme düzeylerini tespit etmek ve öğretim yöntemlerini uyarlamak için bu önemli değerlendirme araçlarını kullanırlar. Bu türden testlerin kalitesi öğretmen bilgi ve becerisine bağlıdır (Kilmen, 2024).

Bununla birlikte, mevcut alanyazın, öğretmen yapımı testlerin genellikle yeterli psikometrik özelliklere sahip olmadığını ortaya koymaktadır. Örneğin, test maddelerinin güçlük ve ayırt ediciliği yeterince dikkate alınmayabilir, madde havuzu sınırlı olabilir ve test geçerliği ile güvenirliği yeterince sağlanamayabilir. Bu nedenle kademeler arası geçiş sistemlerinin olduğu ülkelerde daha çok geniş ölçekli testler tercih edilmektedir. Çelen ve Aybek (2013), çalışmalarında 40 maddelik öğretmen yapımı bir testin KTK ve MTK'ya göre puan sonuçlarını karşılaştırmışlardır. Sonuç olarak, iki kuramın sonuçları arasında birbiriyle pozitif ve yüksek ilişkili bulunmuştur. Benzer şekilde, Çağlar ve Kılıç (2019), merkezi sınav ile öğretmen yapımı testlerin kapsam geçerlikleri, sınavlardan alınan puanlar çeşitli değişkenler açısından incelenmiştir. Çalışmanın sonucunda öğretmen yapımı testler ile merkezi sınavların kapsam geçerliklerinin aynı düzeyde olduğu, öğrencilerin iki testten aldıkları puanların arasında yüksek düzeyde ilişki gözlendiği belirlenmiştir.

Türkiye'de eğitim alanında kullanılan standart testleri geliştiren iki kurum bulunmaktadır. Bunlar, Millî Eğitim Bakanlığı Ölçme, Değerlendirme ve Sınav Hizmetleri Genel Müdürlüğü (MEB-ÖDSGM) ve Ölçme, Seçme ve Yerleştirme Merkezi (ÖSYM)'dir. Bu iki kurum yüksek beklentili ve geniş ölçekli testlerin uzmanlar tarafından hazırlanıp dağıtımının yapıldığı yerlerdir.

Test Kuramlarına Dayalı Uygulamalar

Test kuramlarının uygulama alanları incelendiğinde küçük gruplarda halen KTK'nın kullanımı elverişliyken, daha geniş gruplarda MTK'nın kullanımı daha çok tercih edilmektedir. Çünkü MTK, daha önce bahsedildiği gibi KTK'nın bazı sınırlılıklarına çözüm önerileri getirmekle beraber test geliştirme süreçlerinde ve sonrasında kullanılan yanlılık, değişen madde fonksiyonu (DMF), ölçme değişmezliği, test eşitleme ve BOBUT uygulamaları gibi konularda daha etkin çözümler sunmaktadır (Embretson ve Reise, 2000).

Test eşitleme yöntemleri (Livingston, 2004; von Davier, 2010), değişen madde fonksiyonu uygulamaları (Camilli ve Shephard, 1994; Holland ve Wainer, 1993), ölçme değişmezliği (Horn ve McArdle, 1992; Meredith, 1993) ve BOBUT uygulamaları (Weiss ve Kingsbury, 1984) günümüz test geliştirme süreçlerinde ve sonrasında geçerliğe kanıt sağlayan tekniklerdir. Özellikle teknolojik ve kuramsal gelişmelerle birlikte derinlemesine çalışmalar beraberinde gelmektedir.

Test geliştirmede karşılaşılan kritik zorluklardan biri adaleti sağlamak ve yanlılığı azaltmaktır. DMF, test maddeleri çeşitli alt gruplar için farklı işlev gördüğünde ortaya

çıkar ve değerlendirmede potansiyel adaletsizliğe yol açar. Bu durum, gruplar eşdeğer becerilere sahip olsa bile belirli test maddelerinin farklı alt gruplar arasında sistematik, açıklanamayan bir performans farkı göstermesi halinde ortaya çıkabilir. Araştırmacılar, daha adil ve kapsayıcı değerlendirmeler oluşturmayı amaçlayan bu DMF sorunlarını belirlemek ve ele almak için MTK gibi ileri istatistiksel teknikler geliştirmiştir. MTK modelleri, madde düzeyindeki performansın daha iyi anlaşılmasını ve olası yanlılıkların tespit edilmesini sağlayarak test geliştiricilerin farklı gruplar arasında adil ve geçerli ölçüm sağlamak için test maddelerini iyileştirmesine ve geliştirmesine olanak tanır (Paek, 2018; Wells, 2021). Test eşitleme konusunda ulusal alan yazında Kelecioğlu (2024) tarafından yazılan kaynak ayrıntılı olarak incelenebilir.

Test Geliştirmenin Yarını

Eğitim ve psikoloji dünyası hızla dönüşmeye devam ederken, test geliştirmenin geleceği de önemli değişimlere hazırlanmaktadır. Yapay zekanın test geliştirme sürecine dahil edilmesi, bilgisayar ortamında bireye uyarlanabilir testlerin geliştirilmesi ve etik hususların araştırılması gibi potansiyel yenilikler, insan yeteneklerini nasıl ölçtüğümüzü, değerlendirdiğimizi ve anladığımızı şekillendirme konusunda umut vaat etmektedir. Özellikle, test geliştirmede yapay zekanın rolü, madde üretimini otomatikleştirmeyi, test içeriğini kişiselleştirmeyi ve daha verimli ve bireyselleştirilmiş değerlendirme deneyimleri oluşturmak için test yönetimini optimize etmeyi kapsayabilir. Ayrıca, değerlendirme deneyimini öğrencilerin ihtiyaçlarını ve yeterlilik seviyelerini karşılayacak şekilde uyarlayan bilgisayar ortamında uyarlanabilir testlerde devam eden ilerleme, eğitim ve psikolojik alanlardaki ilerlemeyi ölçme ve izleme şeklimizde devrim yaratabilir. Bu teknolojik ilerlemelerin yanı sıra, test geliştirmenin geleceği, ölçme ve değerlendirmelerin adil, kapsayıcı ve bireylerin iyiliğine uygun kalmasını sağlamak için bu yeniliklerin etik sonuçlarıyla da boğuşmalıdır. Yapay zekâ, uyarlanabilir test ve etik hususları titiz bir şekilde entegre ederek, test geliştirmenin geleceği, insan yeteneklerini anlamlı ve etkili yollarla değerlendirme ve anlama şeklimizi dönüştürme potansiyeline sahiptir (Baqar ve Khanda, 2024; Bulut vd., 2024). Ayrıca bu tür teknolojik gelişmeler öğretme-öğrenme süreçlerini de hızlandırabilir. Öğretim programlarında ölçme ve değerlendirme için ayrılan süre yeterli değildir. Hatta bazen programların en zayıf kısmı olduğu konusunda fikirler mevcuttur. Stiggins (2014), öğretmenlerin mesleki sürelerinin üçte birini ölçme ve değerlendirmeye ayırdıklarını belirtmiştir. Buna göre yapay zekâ ölçme ve değerlendirme işlemini önemli ölçüde kolaylaştırabilir.

Bilgisayar Ortamında Bireye Uyarlanmış Test (BOBUT) Uygulamaları

Bilgisayar ortamında bireye uyarlanabilir testler (BOBUT), test geliştirme alanında önemli bir gelişme olarak ortaya çıkmıştır. BOBUT, her bir katılımcı için en bilgi verici soruları dinamik olarak seçmek üzere MTK'nın ilkelerinden yararlanarak daha verimli ve bireyselleştirilmiş bir ölçme ve değerlendirme deneyimi sağlar (Liu vd., 2024). Yani, yetenek düzeyleri önceden belirli olan madde havuzundan, bireylerin yetenek düzeylerine uygun maddeler seçilerek ve her birey için aynı maddelerden oluşmayan

test olarak tanımlanabilir (Weiss, 2004). Test içeriğini sınava giren kişinin yeterlilik seviyesine göre uyarlayan BOBUT, test süresini ve yanıtlayıcı yükünü azaltırken daha doğru yetenek kestirimleri elde edilmesini sağlar. Bu uyarlanabilir yaklaşımın, seçme, yerleştirme, başarı ve tanılama testleri gibi çeşitli eğitimde ve psikolojide kullanılan test bağlamlarında güçlü bir araç olduğu görülmüştür. Test içeriğini bireyin beceri ve bilgisine göre uyarlayan BOBUT, daha ilgi çekici ve bilgilendirici bir değerlendirme deneyimi sunarak sınava giren kişinin yeteneklerinin daha kesin bir şekilde ölçülmesini sağlar (Thompson ve Weiss, 2011).

BOBUT, özellikle MTK ile paralel bir şekilde gelişmektedir. Çünkü BOBUT, MTK temelli uygulamaları kullanmaktadır. MTK'nın avantajlarını kullanarak test geliştirmede ve uygulamada kolaylıklar sağlamaktadır (Özbaşı ve Demirtaşlı, 2016). BOBUT'un geçmişi 60 yıla dayansa da bilgisayar uygulamalarının gelişmesiyle son 30 yılda daha hızlı bir ivme kazanmıştır (Wainer, 2000). BOBUT, ABD'de birçok sınav uygulamasında kullanılmaktadır (Tian vd., 2007). Son yıllarda araştırmacılara canlı BOBUT uygulama fırsatı sunan uygulamalar üretilmiştir. Bunlar; IRT-CAT, OSCATS, Concerto ve OpenTest uygulamalarıdır (Aybek, 2016). IRT-CAT, MTK temelli, açık kaynaklı ve web tabanlı bir adaptif test yazılımıdır. 1PL, 2PL ve 3PL modellerini desteklemektedir. Akademik Free License ile lisanslanmıştır. Özellikle araştırma ve eğitim amaçlı geliştirilmiştir (Aybek, 2016). OSCATS, GPL lisanslı, açık kaynaklı bir CAT yazılımıdır. 1PL, 2PL, 3PL, Nominal Response Model, Partial Credit Model ve Graded Response Model gibi birçok MTK modelini desteklemektedir. Hem simülasyon hem de canlı uygulamalar için uygundur. R ve MATLAB entegrasyonu planlanmıştır (Aybek, 2016). Concerto, Cambridge Üniversitesi Psikometri Merkezi tarafından geliştirilmiş, R tabanlı açık kaynaklı bir CAT platformudur. catR gibi R paketlerini kullanarak test algoritmalarını özelleştirme olanağı sağlamaktadır. MySQL veri tabanı ile çalışır, esnek ve web tabanlıdır. Hem araştırma hem de kurumsal uygulamalar için yaygın kullanılmaktadır (Scalise ve Allen, 2015). Genel olarak bu dört uygulama değerlendirildiğinde, IRT-CAT ve OSCATS, daha çok akademik ve araştırma amaçlı geliştirilmiş, teorik modelleri test etmeye uygundur. Concerto hem araştırmada hem de gerçek uygulamada en yaygın kullanılan esnek ve güçlü platformdur. OpenTest, daha çok kurumsal sınav uygulamaları için geliştirilen, adaptif test modülleri eklenebilen sistemdir.

Türkiye'de ilk BOBUT uygulamasının Kaptan (1993) tarafından yapıldığı bilinmektedir. Bunula birlikte, Akyıldız (2015), Arslan vd. (2022), Aybek (2016), Çoban (2020), Gür ve Karabay (2018), Kalender (2009), Kezer (2013), Özbaşı ve Demirtaşlı (2016), Okul ve Uygun (2024) ve Tercan (2010)'ın BOBUT'la ilgili çalışmaları mevcuttur. Bu çalışmalar daha çok yükseköğretim, bilgisayar okuryazarlığı, yabancı dil sınavları, ehliyet sınavları ve uzaktan eğitim bağlamında yürütülmüştür. Ayrıca adaptif test uygulamaları üzerine tezler, makaleler ve pilot uygulamalar yapılmış olsa da henüz ÖSYM ve MEB gibi kurumlarda merkezi sınavlarda yaygın bir uygulama alanı bulmamıştır.

BOBUT, yerleştirme testi, başarı testi ve tanı testi dahil olmak üzere çeşitli eğitimsel ve psikolojik test bağlamlarında uygulanmaktadır. Son araştırmalar, madde seçme algoritmaları, madde ifşa kontrolü ve test sonlandırma kriterleri gibi BOBUT'un çeşitli yönlerini optimize etmek için makine öğrenimi tekniklerinin entegrasyonunu

araştırmaktadır. Psikometri ve makine öğreniminin kesişimi sayesinde, BOBUT'un geleceği, katılımcıların çeşitli ihtiyaçlarını karşılayan daha güvenilir, adil ve etkili test çözümleri sunma potansiyeline dayanmaktadır (Han, 2018).

Köklü psikometrik prensipleri son teknoloji makine öğrenimi yaklaşımlarıyla birleştirmek, daha kesin ve güvenilir yetenek tahminleri geliştirmek için umut verici yollar açmaktadır. Birbirinden farklı ancak birbirini tamamlayan bu alanların sinerjisinden yararlanarak, BOBUT'un geleceği, daha adil, objektif ve ölçüm hassasiyeti yüksek bireyselleştirilmiş değerlendirmeler sunmayı vaat etmektedir.

Otomatik Madde Üretimi

Otomatik madde üretimi, eğitimde ve psikolojide kullanılan testlerin geliştirilmesinde oldukça yenilikçi bir yaklaşımdır. Bu süreçte, bilgisayar tekniklerinden yararlanılarak test maddeleri algoritmik olarak üretilebilmekte, madde yazma süreci kolaylaşmakta ve konu uzmanlarının üzerindeki yük azalmaktadır. Bu artan verimlilik, ölçme ve değerlendirme sürecini önemli ölçüde güçlendirmektedir (Embretson ve Yang, 2006; Gierl vd., 2012).

Geleneksel madde üretiminin zorluklarına bir çözüm önerisi olarak geliştirilen otomatik madde üretiminin başlangıcı Guttman (1959) tarafından ortaya konan sistematik madde yazımı kabul edilebilir. Daha sonra Bormuth (1970)'un başarı testi maddelerini sistematikleştirme çalışmaları, Roid ve Haladyna'nın (1982) otomatik madde üretimi üzerine yayınladıkları kitap bu alanın gelişimine katkı sağlamıştır. Otomatik madde üretiminin teknik özellikleri üzerine son dönemlerde Gierl ve Lai (2012) ve Gierl ve Haladyna (2012) tarafından yayınlanan çalışmalar bu alana farklı modeller kazandırmıştır. Türkiye'de Kucam (2023) ve Sayın ve Bozdağ'ın (2024) hazırladıkları Türkçe kaynaklar bu alandaki araştırmacılara kapsamlı bilgiler sunmaktadır. Otomatik madde üretimiyle ilgili eğitim dışı ilk çalışma Kıyak vd. (2023) tarafından tıp alanında yapılmıştır. Bu çalışmada hipertansiyon konusunda bir bilgisayar yazılımıyla 1.73 saniye 1600 adet çoktan seçmeli soru hazırlanmıştır. Ayrıca, Emekli vd. (2023), çalışmalarında psikiyatri eğitiminde klinik akıl yürütmeyi ölçmek için Türkçe otomatik madde üretimi tabanlı çoktan seçmeli sorular geliştirilmiştir. Python tabanlı bir yazılım ile sorular üretilmiştir. Toplamda 1189 soru üretilmiş ve her bir tanı için 11 soru örneklendirilmiştir. Sonuç olarak otomatik madde üretim yöntemi, tıp eğitiminin ihtiyaç duyduğu yüksek kaliteli soruların hızlı ve etkili bir şekilde üretilmesine olanak sağlamaktadır. Sayın (2024) çalışmasında görsel okuma maddelerinin oluşturulmasında otomatik madde üretim sürecini kullanmıştır. Üretilen görsel okuduğunu anlama maddelerinin psikometrik özelliklerini değerlendirmek için 8. sınıfta öğrenim gören 1380 öğrenci ile ön uygulama çalışması gerçekleştirilmiştir. Buna göre otomatik madde üretimi sonucunda madde güçlüğü ve ayırt edicilik açısından verimli sonuçlar elde edilmiştir. Yapılan çalışmalar genel olarak değerlendirildiğinde Türkiye'de otomatik madde üretimi uygulamaları özellikle tıp eğitimi, psikoloji, yabancı dil sınavları ve görsel okuduğunu anlama testleri gibi alanlarda hızla yayılmaktadır. Son 3 yılda özellikle GPT ve ChatGPT tabanlı otomatik madde üretimi üzerine yapılan çalışmalar öne çıkmaktadır. Madde türü olarak hem çoktan seçmeli hem de açık uçlu maddelerin otomatik üretilmesi üzerine denemeler mevcuttur.

Doğal dil işleme (NLP), makine öğrenimi (ML) ve diğer gelişmiş bilgisayar araçlarını kullanan otomatik madde üretimi, büyük ve çeşitli test maddesi havuzları üretebilir. Bu maddeler belirli değerlendirme hedeflerine, öğretim programı standartlarına veya hedef bilişsel yapılara göre uyarlanabilir ve böylece sınav katılımcılarının bilgi ve becerilerinin daha kapsamlı ve özelleştirilmiş bir şekilde değerlendirilmesine olanak sağlar (Embretson ve Kingston, 2018).

Madde üretimine yönelik bu otomatik yöntem, gerekli psikometrik niteliği ve titizliği korurken test içeriğinin çeşitliliğini ve genişliğini artırabilmektedir. Test geliştirme süreci, otomatik madde üretimini MTK gibi yerleşik test kuramlarıyla bütünleştirerek, katılımcıların farklı ihtiyaçlarını ve yeterlilik düzeylerini daha iyi karşılayan daha güvenilir, uyarlanabilir ve bireyselleştirilmiş ölçme ve değerlendirmelerin oluşturulmasının önünü açabilmektedir (Goldhammer vd., 2020).

Otomatik madde üretimi ve BOBUT birbirine entegre olabilecek uygulamalardır. Çünkü BOBUT, geniş ve psikometrik olarak kalibre edilmiş madde havuzlarına ihtiyaç duymaktadır (van der Linden, 2018, Weiss, 2011). Otomatik madde üretimi, bilişsel modeller veya yapay zekâ ile sınırsız sayıda yeni madde üreterek CAT'in madde havuzu problemini çözebilecek güçtedir (Gierl vd., 2012). Gierl ve Lai (2012), çalışmalarında otomatik madde üretimiyle BOBUT için madde bankasına entegre edilmesinin testin kapsam geçerliğini artırdığını göstermişlerdir. Ancak bu durumun dezavantajları da mevcuttur. Bunlar;

Otomatik olarak üretilmiş maddelerin BOBUT'ta kullanılmadan önce MTK parametrelerinin kestirilmesi ve kalibrasyonlarının yapılması gerekmektedir (van der Linden, 2018).

Otomatik olarak üretilen her madde kaliteli bakımından sıkıntılı olabilir. Bu nedenle uzman denetiminden geçirilmelidir (Gierl ve Lai, 2012).

Otomatik madde üretimi, iyi tanımlanmış bilişsel modeller gerektirmektedir. Aksi halde düşük nitelikli madde üretimi olabilmektedir (Huff ve Sireci, 2001).

Otomatik madde üretme tekniklerinin stratejik ve bilinçli kullanımı yaygınlaştıkça test geliştirme de dönüşümsel ilerlemeler kaydetmeye hazırlanmaktadır. Bu yaklaşım, eğitimsel ve psikolojik ölçme ve değerlendirmelerin oluşturulması ve sunulmasında potansiyel olarak devrim yaratabilir ve ölçme ve değerlendirme ortamını yeniden şekillendirebilir.

Yapay Zekâ Destekli Madde Yazma

Test geliştirme ve madde yazımında yapay zekanın uygulanması, eğitimde ve psikolojide ölçme ve değerlendirme alanında önemli bir potansiyele sahiptir. Doğal dil işleme (NLP) ve makine öğrenimi (ML) gibi gelişmiş bilgisayar yöntemlerinden yararlanarak test maddelerinin otomatik olarak oluşturulması, madde yazma sürecini kolaylaştırabilir ve çeşitli ve kişiye özel değerlendirmelerin oluşturulmasını önemli ölçüde geliştirebilir.

Ayrıca, yapay zekâ (YZ) destekli test geliştirme ve madde yazma süreçleri, ölçme değerlendirme alanında önemli yenilikler sunmaktadır. Özellikle madde yazımı ve test geliştirme aşamalarında YZ, madde üretimi, dil kontrolü, bilişsel zorluk seviyelerinin ayarlanması gibi konularda katkı sağlamaktadır (Chalapathy ve Chawla, 2019). Ayrıca, YZ ile yapılan analizler, test kapsamının artırılması ve madde geçerliğinin sağlanması açısından önem taşımaktadır (Van der Linden, 2017). YZ, farklı gruplar arasında farklı işlev gösteren maddeleri tespit etmek için değişen madde fonksiyonu (DMF) analizlerinde de kullanılmaktadır (Zumbo, 1999). YZ'nin büyük veri analizi kapasitesi, testlerin daha hızlı ve etkili şekilde geliştirilmesini sağlar ve BOBUT gibi bireyselleştirilmiş ölçme araçları sunar (Joshi, 2023). Bu teknolojinin sunduğu veri işleme ve geri bildirim sağlama özellikleri, eğitimde daha adil ve geçerli testlerin geliştirilmesine olanak tanımaktadır (De Ayala, 2009). Çavuş (2024), eğitimde yapay zekâ tabanlı ölçme ve değerlendirme üzerine kapsamlı bir derleme hazırlamıştır. Bu çalışmada yapay zekâ tabanlı ölçme ve değerlendirme araçlarının eğitim alanında sağladığı faydaların, süreçte yaşanan zorlukların ve bu süreçte öğretmenlerin rolünün incelenmesi amaçlanmıştır.

Bununla birlikte, bu YZ destekli teknolojiler büyük bir potansiyele sahip olsa da dikkatli bir değerlendirme ve dengeli bir şekilde uygulanmalıdır. Otomatik madde üretimine çok fazla güvenmek, konu uzmanlarının sürece getirdiği uzmanlığı ve ayrıntılı anlayış biçimini potansiyel olarak zayıflatabilmektedir. YZ güdümlü yaklaşımların önemli içeriksel faktörleri gözden kaçırma veya değerlendirilen yapıların uygun derinliğini ve karmaşıklığını yakalayamama riski bulunmaktadır. Ayrıca, YZ sistemlerinde kullanılan veri ve algoritmaların doğasında bulunan potansiyel önyargılar, hatalı bir şekilde oluşturulan test maddelerine yansıyarak değerlendirmelerin adilliğini ve geçerliliğini zedeleyebilir. Test geliştirmenin titizliğini, uygunluğunu ve etik bütünlüğünü korumasını sağlamak için hem insan hem de makine odaklı süreçlerin güçlü yönlerini bütünleştiren dengeli bir yaklaşım gerekli olabilmektedir (Zhang ve Chang, 2016).

Türkiye'de yapay zekâ kullanılarak test maddelerinin üretildiği çalışmalar mevcuttur. Karakaya vd. (2024) çalışmalarında üst düzey zihinsel becerilerin ölçülmesinde açık uçlu maddelerin yapay zekâ destekli hazırlanma sürecini incelemişlerdir. İlk aşamada ChatGPT-4o kullanılarak açık uçlu maddeler üretilmiştir. Daha sonra özelleştirilmiş GPT modeli geliştirilmiş ve bu model yalnızca açık uçlu madde üretimi üzerine eğitilmiştir. Üretilen maddeler uzman görüşü ve literatürdeki ölçütler doğrultusunda incelenmiştir. Sonuç olarak, yapay zekâ tabanlı madde üretimi, üst düzey zihinsel becerilerin ölçülmesi için potansiyel bir yenilikçi yöntem sunmaktadır. Özelleştirilmiş GPT modelleri, ChatGPT gibi genel modellerden daha yüksek performans göstermektedir. Gelecekte, otomatik puanlama sistemleriyle entegrasyon sağlanması önerilmektedir. Boduroğlu vd. (2023) çalışmalarında, ChatGPT'nin çoktan seçmeli test maddelerinin güçlük düzeylerini tahmin etme performansını incelemişlerdir. Araştırma, 4930 lise öğrencisinin katıldığı 20 soruluk çoktan seçmeli Matematik testi üzerinden yürütülmüştür. Test maddelerinin gerçek güçlük düzeyleri öğrenci yanıtlarıyla hesaplanmıştır. Aynı maddeler hem ChatGPT 3.5 hem de üç matematik uzmanına verilmiş; "çok kolay, kolay, orta, zor, çok zor" kategorilerine sınıflandırmaları istenmiştir. Analizlerde korelasyon ve sınıf içi korelasyon katsayıları kullanılmıştır.

Sonuç olarak, ChatGPT'nin, madde güçlüğü tahmininde uzmanlara yakın ama daha düşük performans gösterdiğini belirlemişlerdir. Ayrıca deneme uygulamasının yapılamadığı veya uzmanlara ulaşılamadığı durumlarda yardımcı bir araç olarak kullanılabileceğini belirtmişlerdir. Boduroğlu ve Yiğiter (2024), çalışmalarında yapay zekâ (YZ) araçlarının öğretmen yapımı testlere sunduğu geribildirimler incelenmişlerdir. Araştırmada 8. sınıf İngilizce dersi için hazırlanmış 20 maddelik çoktan seçmeli başarı testi kullanılmıştır. Test, 5115 öğrenciye uygulanmıştır. Ardından, ChatGPT-4o, Google Gemini ve Microsoft Copilot modelleri test maddeleri üzerinde analiz yaparak geribildirim üretmiştir. Sonuç olarak, yapay zekâ araçlarının, öğretmen yapımı testlerin psikometrik kalitesini artırmada destek olabileceği, öğretmenlere kapsam geçerliği, madde güçlüğü ve süre öngörüsü gibi önemli veriler sağladığı, ancak farklı YZ araçlarının ürettiği tahminler tam uyumlu olmadığı, uzman doğrulaması hâlâ gerekli olduğu bulunmuştur. Tüm çalışmalar genel olarak değerlendirildiğinde yapay zekâ, test geliştirme sürecine üç ana katkı sağlamaktadır:

- Madde üretiminde hız ve çeşitlilik (AIG ile nitelikli madde havuzları).
- Psikometrik analizlerde destek (madde güçlüğü, kapsam geçerliği tahminleri).
- Öğretmen ve araştırmacılara pratik kolaylık (süre tahmini, madde kalitesi geri bildirimi).

Ancak yine de YZ'nin bu süreçte uzman denetimi ve psikometrik doğrulama olmadan doğrudan kullanılması risklidir. Bu nedenle YZ, test geliştirme süreçlerinde tamamlayıcı ve destekleyici bir araç olarak görülmelidir.

Sonuç

Test geliştirme, psikometri alanının ortaya çıkışıyla beraber var olan bir süreçtir. Tarihsel süreç içerisinde klasik ölçme kuramlarından başlayarak günümüzde yapay zekâ (YZ) destekli uygulamalara kadar önemli bir dönüşüm geçirmiştir. Sir Francis Galton, Alfred Binet ve Charles Spearman gibi öncülerin bireysel farklılıkları ölçmeye yönelik çalışmaları, test geliştirme disiplininin ilk adımlarını oluşturmuştur (Curtis ve Glaser, 1984). Zamanla, Klasik Test Kuramı (KTK), Madde Tepki Kuramı (MTK), Genellenebilirlik Kuramı ve Sato Test Kuramı gibi psikometrik yaklaşımlar, ölçme ve değerlendirme süreçlerine daha sistematik ve güvenilir bir çerçeve sunmuştur (Brennan, 1992; Crocker ve Algina, 1986; Çüm, 2019; Hambleton ve Swaminathan, 1985). Bu kuramlar, testlerin geçerlik ve güvenirliğinin sağlanmasında kritik rol oynamış, aynı zamanda test geliştirme adımlarının bilimsel temeller üzerine oturtulmasına katkı sağlamıştır.

Günümüzde ise test geliştirme süreci, teknolojik yenilikler ve YZ tabanlı yaklaşımların etkisiyle yeniden şekillenmektedir. Özellikle otomatik madde üretimi, test geliştiricilere çok kısa sürede büyük madde havuzları oluşturma olanağı tanımaktadır (Embretson ve Yang, 2006; Gierl vd., 2017). Türkiye'de yapılan son çalışmalar da bu alandaki ilerlemeyi destekler niteliktedir: Karakaya vd. (2024) açık uçlu maddelerin yapay zekâ destekli üretiminin üst düzey zihinsel becerilerin ölçülmesinde yenilikçi bir yaklaşım sunduğunu göstermiştir. Benzer şekilde, Boduroğlu vd. (2023) ile Bo-

duroğlu ve Yiğiter (2024), ChatGPT ve benzeri YZ modellerinin madde güçlüğü tahmininde ve öğretmen yapımı testlerde kapsam geçerliği analizinde uzman görüşlerine yakın sonuçlar verdiğini ortaya koymuştur. Bu bulgular, YZ'nin test geliştirme sürecinde tamamlayıcı bir araç olarak kullanılabileceğini kanıtlamaktadır.

Tarihsel sürece baktığımızda 1920'ler test geliştirmede bir patlama dönemi olarak kabul edilmektedir. Walter S. Monroe 1928 yılında 1300 farklı standart testin mevcut olduğunu ve bir yıl içinde 30.000.000 test puanı elde edildiğini tahmin etmiştir. 1944 yılına gelindiğinde, yaklaşık 50000 testin mevcut olduğu ve 60.000.000 maddenin uygulandığı düşünülmektedir (Dragositz, 1963).

Test geliştirme ve madde yazmayla ilgili çalışmalara bakıldığında 2000'li yıllardan itibaren alanyazında hızlı bir artış olduğu görülmüştür. Bu durumun nedenleri arasında birçok test geliştirme kaynağının 2000'lerden sonra artış gösterdiği ve 1990'lı yıllarda internetin hayatımıza girmesiyle bilgisayar teknolojiler sayesinde kaynaklara ulaşmanın kolaylaştığı gösterilebilir (Hambleton ve Swaminathan, 1985).

Test geliştirme alanında, alanyazında birçok çalışma yapılmaktadır. Sınav odaklı bir eğitim sisteminde ders başarısını belirlemede testler ön plana çıkmaktadır. Türkiye, eğitim ve öğretimde başarı düzeyinin belirlenmesinde testlerle sarmalanmış bir ülkedir. Sadece yerli alanyazında ölçme araçlarının dizinlendiği TOAD (bkz. https://toad.halileksi.net/) platformunda 15333 yazara ait 13439 ölçek veya test bulunmaktadır. Bu kadar ölçme aracı geliştirilen bir alanyazında araştırmacıların tercih ettiği kaynakların belirlenmesi yeni test geliştiricilere bir yol göstermesi açısından önemlidir.

Eğitimde ve psikolojide kullanılan testlerin geliştirilmesi, farklı durumların değişen ihtiyaçlarını ve bilimsel kuramlar ile teknolojik yeniliklerdeki ilerlemeleri yansıtan dinamik ve çok yönlü bir çaba olmuştur. Test oluşturmanın tarihsel temelleri, psikometrik modellerin ortaya çıkışı ve değerlendirmelerin standartlaştırılmasında karşılaşılan zorluklar, test geliştirmenin mevcut manzarasının yolunu açmıştır. Alan büyümeye devam ettikçe, değerlendirmelerin kapsayıcı, verimli ve öğrencilerin ve uygulayıcıların çeşitli ihtiyaçlarına cevap verebilmesini sağlamak için yapay zekayı entegre etmek, uyarlanabilir test yöntemlerini geliştirmek ve etik hususları araştırmak kritik önem taşımaktadır. Bu çok yönlü yaklaşım, test geliştirmenin gelecekteki yörüngesini şekillendirmede etkili olacak ve değerlendirmelerin daha kişiselleştirilmiş, uyarlanabilir ve sürekli değişen eğitim ve psikoloji ortamıyla uyumlu hale gelmesini sağlayacaktır. Etik ilkelere bağlı kalarak teknolojik gelişmeleri kucaklayan bu alan, yalnızca kapsamlı ve doğru değil aynı zamanda tüm bireyler için eşitlikçi ve güçlendirici değerlendirmeler oluşturmak için çaba gösterebilir.

Gelecekte, YZ'nın test yapımına dahil edilmesi, bilgisayar ortamında bireye uyarlanabilir testlerin geliştirilmesi ve etik konuların araştırılması, insan yeteneklerini değerlendirme ve anlama biçimimizi şekillendirme potansiyeli taşımaktadır. Tüm bu gelişmelere rağmen, etik sorunlar ve veri mahremiyeti konuları test geliştirmenin geleceği açısından ciddi bir tartışma alanı oluşturmaktadır. YZ tabanlı sistemler, büyük veri setlerinden beslendiği için, kullanılan verilerdeki gizli önyargıların testlere yansıma riski bulunmaktadır (Ferrara, 2023). Bu durum, belirli sosyo-demografik grupların haksız yere dezavantajlı duruma düşmesine yol açabilir. Ayrıca, YZ algoritmalarının "kara kutu" niteliği, yani karar süreçlerinin şeffaf olmaması, adalet, değişen

madde fonksiyonu (DMF) ve ölçme değişmezliği gibi psikometrik sorunların tespitini güçleştirmektedir (Paek, 2018; Wells, 2021). Veri güvenliği ve mahremiyetin korunması da başka bir kritik noktadır. Öğrencilerin performans verilerinin kötüye kullanımı hem etik açıdan hem de hukuki açıdan riskler barındırmaktadır (Tippins, 2015).

Sonuç olarak, test geliştirmenin geleceği çift yönlü bir stratejiye ihtiyaç duymaktadır. Bir yanda YZ'nin sunduğu hız, çeşitlilik ve bireyselleştirme potansiyelinden yararlanmak, diğer yanda adalet, şeffaflık, etik standartlar ve veri mahremiyetini korumak gerekir. YZ tabanlı test geliştirme süreçlerinde, uzman görüşleri ve psikometrik analizlerle birlikte kullanıldığında ölçme ve değerlendirmeye büyük katkılar sunacağı açıktır. Ancak bu süreçte insan uzmanlığının yerini tamamen alması mümkün ve doğru görünmemektedir. Aksine, en sağlıklı yaklaşım, insan uzmanlığının titizliği ile YZ'nin hız ve hesaplama gücünün dengeli bir şekilde bütünleştirilmesidir. Böylece gelecekte test geliştirme süreçleri, hem psikometrik açıdan geçerli ve güvenilir, hem de etik açıdan adil ve kapsayıcı bir yapıya kavuşabilecektir. Ayrıca, test geliştirme ve madde yazma uzmanlık gerektiren bir iştir. Buna göre bu alanda çalışacak araştırmacılara, küresel çapta kullanılan güvenilir ve geçerli kaynaklara başvurmaları ve teknolojik gelişmeleri takip etmeleri önerilmektedir.

Kaynakça

AKYILDIZ, M. (2015, Şubat, 4-6), ***"Online Ölçme ve Değerlendirme, Bireye Uyarlanmış Testlerde İmkanlar, Ufuklar",*** XVII. Akademik Bilişim Konferansı, Anadolu Üniversitesi, Eskişehir.

ANASTASI, A. (1988), ***Psychological Testing,*** New York: MacMillan Pub. Co. Inc.

ANASTASI, A. ve URBINA, S. (1997), ***Psychological Testing,*** Prentice Hall/Pearson Education.

AREA, APA ve NMCE. (1997), ***Standards For Educational and Psychological Testing,*** Washington, DC: American Educational Research Association.

ARSLAN, Y. K., ÇOLAK, M. ve BİLGİN, U. (2021), ***"Determining ability levels by using computerized adaptive testing and Paper-And-Pencil testing in the distance Education: a Multi-Centred Real Data Study"***, Tıp Eğitimi Dünyası, XXI, 63: 95-103. https://doi.org/10.25282/ted.1003962

AYBEK, E. C. (2016), ***"Concerto: Bilgisayar Ortamında Bireye Uyarlanmış Test Uygulamaları İçin Bir Platform",*** Journal of Measurement and Evaluation in Education and Psychology, VII, 2: 251-271.

BAKER, F. B. (2001), ***The Basics of Item Response Theory,*** College Park, (MD): ERIC Clearinghouse on Assessment and Evaluation.

BAKER, F. B. ve KIM, S. (2001), ***Item Response Theory: Parameter Estimation Techniques,*** New York: Marcel Dekker.

BAKER, F. B. ve KIM, S. (2017), ***Test Calibration,*** Springer Nature.

BAQAR, M. ve KHANDA, R. (2024), ***The Future of Software Testing: AI-Powered Test Case Generation and Validation,*** Cornell University. https://doi.org/10.48550/arxiv.2409.05808

BAYKUL, Y. (2015), ***Eğitimde ve Psikolojide Ölçme: Klasik Test Teorisi ve Uygulaması,*** Ankara: Pegem Akademi.

BODUROĞLU, E., KOÇ, O. ve YİĞİTER, M. S. (2023), ***"Madde Güçlüklerinin Tahmin Edilmesinde Uzman Görüşleri ve ChatGPT Performansının Karşılaştırılması",*** Disiplinlerarası Eğitim Araştırmaları Dergisi, VII, 15: 202-210. https://doi.org/10.57135/jier.1296255

BODUROĞLU, E. ve YİĞİTER, M. S. (2024), ***"Öğretmen Yapımı Testler İçin Yapay Zekâ Destekli Geribildirim",*** Journal of Applied Measurement and Evaluation, XIV, 3: 121-138.

BORMUTH, J. R. (1970), ***On The Theory of Achievement Test Items,*** Chicago: University Chicago Press.

BRENNAN, R. L. (1992), ***"Generalizability theory"***, Educational Measurement Issues and Practice, XI, 4: 27-34. https://doi.org/10.1111/j.1745-3992.1992.tb00260.x

BRENNAN, R. L. (2001), ***Generalizability Theory,*** Iowa: ACT Publications.

BULUT, O., BEITING-PARRISH, M., CASABIANCA, J. M., SLATER, S., JIAO, H., SONG, D., ORMEROD, C.M., FABIYI, D.G., IVAN, R., WALSH, C., RIOS, O., WILSON, J. M., YILDIRIM-ERBASLI, S. N., WONGVORACHAN, T., LIU, J. X., TAN, B. ve MORİLOVA, P. (2024), ***"The Rise of Artificial Intelligence in Educational Measurement: Opportunities and Ethical Challenges"***, Cornell University. https://doi.org/10.48550/arxiv.2406.18900

CAMILLI, G. ve SHEPARD, L. A. (1994), ***Methods for Identifying Biased Test Items,*** California: Sage.

CARMINES, E. G. ve ZELLER, R. A. (1979), ***Reliability and Validity Assessment,*** Thousand Oaks, CA: Sage. https://doi.org/10.4135/9781412985642

CHALAPATHY, R. ve CHAWLA, S. (2019), ***"Deep Learning for Anomaly Detection: A Survey"***, ACM Computing Surveys (CSUR), LIV, 2:1-34.

COHEN, R. J. ve SWERDLIK, M. E. (2010), ***Psychological Testing and Assessment An Introduction to Tests and Measurement,*** NY: McGraw-Hill Higher Education.

CROCKER, L. ve ALGINA, L. (1986), ***Introduction to Classical and Modern Test Theory,*** New York: Holt, Rinehart and Winston.

CRONBACH, L. J. (1990), ***Essentials of Psychological Testing,*** New York: Harper & Row.

CURTIS, M. E. ve GLASER, R. (1984), ***"Intelligence Testing, Cognition, and Instruction",*** International Journal of Psychology, XIX:475-497. https://doi.org/10.1080/00207598408247542

ÇAĞLAR, M. ve KILIÇ, A. (2019), ***"Merkezi Sınav ve Öğretmen Yapımı Sınavların Bazı Değişkenler Açısından İncelenmesi: Ortaöğretime Geçiş Sınavı Örneği"***, Bolu Abant İzzet Baysal Üniversitesi Eğitim Fakültesi Dergisi, XIX, 4: 1288:1305. https://doi.org/10.17240/aibuefd.2020..-416761

ÇAKAN, M. (2003), ***"Geniş Ölçekli Başarı Testlerinin Eğitimindeki Yeri ve Önemi",*** Eğitim ve Bilim, XXVIII, 128: 19-26.

ÇAVUŞ, M. N. (2024), ***"Eğitimde Yapay Zekâ Tabanlı Ölçme ve Değerlendirme Üzerine Bir Derleme",*** International Journal of English for Specific Purposes (JOINESP), II,1: 39-54.

ÇELEN, Ü. (2008), ***"Klasik Test Kuramı ve Madde Tepki Kuramı Yöntemleriyle Geliştirilen İki Testin Geçerlilik ve Güvenilirliğinin Karşılaştırılması"***, İlköğretim Online, VII, 3: 758-768.

ÇELEN, Ü. ve AYBEK, E. C. (2013), ***"Öğrenci Başarısının Öğretmen Yapımı Bir Testle Klasik Test Kuramı ve Madde Tepki Kuramı Yöntemleriyle Elde Edilen Puanlara Göre Karşılaştırılması"***, Journal of Measurement and Evaluation in Education and Psychology, IV, 2: 64-75. https://doi.org/10.21031/epod.77213

ÇOBAN, E. (2020), ***"Bilgisayar Temelli Bireyselleştirilmiş Test Yaklaşımlarının Türkiye'deki Merkezi Dil Sınavlarında Uygulanabilirliğinin Araştırılması"***, Yayımlanmamış doktora tezi, Ankara Üniversitesi, Eğitim Bilimleri Enstitüsü, Ankara.

ÇÜM, S. (2019), ***Sato Test Kuramı'nın Klasik Test Kuramı ve Madde Tepki Kuramı ile Psikometrik Açıdan Karşılaştırılması,*** Yayımlanmamış doktora tezi, Hacettepe Üniversitesi, Eğitim Bilimleri Enstitüsü, Ankara.

DADEY, N., LYONS, S. ve DE PASCALE, C. A. (2018), ***"The Comparability of Scores from Different Digital Devices: A Literature Review and Synthesis with Recommendations for Practice"***, Taylor & Francis, XXXI, 1: 30-50. https://doi.org/10.1080/08957347.2017.1391262

DE AYALA, R. J. (2009), ***The Theory and Practice of Item Response Theory,*** Guilford Press.

DELUCA, C., COOMBS, A., MACGREGOR, S. ve RASOOLI, A. (2019), ***"Toward a Differential and Situated View of Assessment Literacy: Studying Teachers'*** Responses to Classroom Assessment Scenarios", Frontiers Media, IV: 1-10. https://doi.org/10.3389/feduc.2019.00094

DE MARS, C. (2010), ***Item Response Theory: Understanding Statistics Measurement,*** New York: Oxford University Press.

DESJARDINS, C. D. ve BULUT, O. (2018), ***Handbook of Educational Measurement and Psychometrics Using R,*** Boca Raton, (FL): CRC Press.

DOWNING, S. M. ve HALADYNA, T. M. (2006), ***Handbook of Test Development,*** Mahwah, NJ: Lawrence Erlbaum.

DRAGOSITZ, A. (1963), ***"The National Council On Measurement in Education: Its History, Purposes And Activities"***, The Yearbook of the National Council on Measurement in Education, XX: 170-172.

DUQUE, S. R., TAL, E. ve BARBIC, S. (2023), ***"The Role of Ethical and Social Values in Psychosocial Measurement",*** Measurement, CCXXV, 113993. https://doi.org/10.1016/j.measurement.2023.113993

EBEL, R. L. (1951), ***"Writing The Test Item",*** (Ed. E. F. Lindquist), Educational measurement (s. 185-249) içinde. Washington, DC: American Council on Education.

EBEL, R. L. ve FRISBIE, D. A. (1991), ***Essentials of Educational Measurement,*** NJ: Englewood Cliffs, Prentice-Hall.

ERKUŞ, A. (2012), ***Psikolojide Ölçme ve Ölçek Geliştirme-I Temel Kavramlar ve İşlemler,*** Ankara: Pegem Akademi.

EMBRETSON, S. E. ve KINGSTON, N. M. (2018)***, "Automatic Item Generation: A More Efficient Process for Developing Mathematics Achievement Items?"***, Journal of Educational Measurement, LV, 1: 112-131. https://doi.org/10.1111/jedm.12166

EMBRESTON, S. E. ve REISE, S. P. (2000), ***Item Response Theory for Psychologists,*** Lawrence: Erlbaum Ass., Inc.

EMBRETSON, S. E. ve YANG, X. (2006), ***"Automatic Item Generation and Cognitive Psychology"***, (Ed. C. R. Rao ve S. Sinharay), Handbook of Statistics (s.747-768) içinde. Elsevier BV. https://doi.org/10.1016/s0169-7161(06)26023-1

EMEKLİ, E., KIYAK, Y. S., ALICI, Y. H., COŞKUN, Ö. ve BUDAKOĞLU, I. İ. (2025). ***"Clinical Reasoning in Psychiatric Education: Development of Multiple-Choice Questions with Automatic Item Generation in Turkish"***, Turkish Journal of Psychiatry, XXXVI, 336:343. https://doi.org/10.5080/u27540

FERRARA, E. (2023), ***"Fairness and Bias In Artificial Intelligence: A Brief Survey of Sources, Impacts, and Mitigation Strategies"***, Sci, VI, 1: 3-15. https://doi.org/10.3390/sci6010003

GIERL, M. J. ve HALADYNA, T. M. (2012), ***Automatic Item Generation***, Theory and Practice, New York, NY: Routledge Chapman & Hall. https://doi.org/10.4324/9780203803912

GIERL, M. J. ve LAI, H. (2012), ***"The Role of Item Models in Automatic Item Generation"***, International journal of testing, XII, 3: 273-298.

GIERL, M. J., LAI, H. ve TURNER, S. R. (2012), ***"Using Automatic Item Generation to Create Multiple-Choice Test Items",*** Medical Education, XXXXVI, 8: 757-765. https://doi.org/10.1111/j.1365-2923.2012.04289.x

GOLDHAMMER, F., SCHERER, R. ve GREIFF, S. (2020), ***"Editorial: Advancements in Technology-Based Assessment: Emerging Item Formats, Test Designs, and Data Sources"***, Frontiers in Psychology, 10:3047. https://doi.org/10.3389/fpsyg.2019.03047

GULLIKSEN, H. (1950), ***Theory of Mental Tests,*** New York: Wiley.

GUTTMAN, L. (1959), ***"Introduction to Facet Design and Analysis"***, Acta Psychologica, XV: 130-138.

GÜR, R. ve KARABAY, E. (2018), ***"Motorlu Taşıtlar Sürücü Kursiyerleri Sınavının simülatif bilgisayar ortamında bireye uyarlanmış test uygulaması",*** Kuramsal Eğitimbilim, XI, 2: 201-228. https://doi.org/10.30831/akukeg.395252

HALADYNA, T. M. (2004), ***Developing and Validating Multiple-Choice Test Items,*** Mahwah, NJ: Lawrence Erlbaum Associates, Publishers. https://doi.org/10.4324/9780203825945

HALADYNA, T. M. ve RODRIGUEZ, M. C. (2013), ***Developing and Validating Test Items,*** New York: Routledge. https://doi.org/10.4324/9780203850381

HAN, K. T. (2018), ***"Components of The Item Selection Algorithm in Computerized Adaptive Testing"***, Korea Health Insurance Licensing Examination Institute, XV, 7: 1-13. https://doi.org/10.3352/jeehp.2018.15.7

HAMBLETON, R. K., BARTRAM, D. ve OAKLAND, T. (2011), ***"Technical Advances and Guidelines for Improving Testing Practices"***, (Ed. P. R. Martin, F. M. Cheung, M. C. Knowles, M. Kyrios, J. B. Overmier ve J. M. Prieto), IAAP Handbook of Applied Psychology, (s.338-361) içinde. John Wiley & Sons. https://doi.org/10.1002/9781444395150.ch14

HAMBLETON, R. K. ve SWAMINATHAN H. (1985), ***Item Response Theory: Principals and Applications,*** Norwell, (MA): Kluwer Academic Publishers.

HAMBLETON, R. K., SWAMINATHAN, H. ve ROGERS, H. J. (1991), ***Fundamentals of Item Response Theory,*** California: Sage Publications Inc.

HAMBLETON, R. K. ve JONES R. W. (1993), ***"An NCME Instructional Module on Comparison of Classical Test Theory and Item Response Theory and Their Applications to Test Development***", Educational Measurement: Issues and Practice, XII, 3: 38-47. https://doi.org/10.1111/j.1745-3992.1993.tb00543.x

HOLLAND, P. W. ve WAINER, H. (1993), ***Differantial Item Functioning,*** Educational Testing Service.

HORN, J. L. ve MCARDLE, J. J. (1992), ***"A Practical and Theoretical Guide to Measurement Invariance in Aging Research"***, Experimental Aging Research, XVIII, 3: 117-144. https://doi.org/10.1080/03610739208253916

HUFF, K. L. ve SIRECI, S. G. (2001), ***"Validity issues in Computer-Based testing",*** Educational Measurement Issues and Practice, XX, 3: 16-25. https://doi.org/10.1111/j.1745-3992.2001.tb00066.x

IRVINE, S. H. ve KYLLONEN, P. C. (2002), ***Item Generation for Test Development,*** Mahwah, NJ: Lawrence Erlbaum Associates, Publishers.

JOSHI, M. (2023), ***"Adaptive Learning through Artificial Intelligence"***, SSRN Electronic Journal, IV, 4: 1-2. https://doi.org/10.2139/ssrn.4514887

KALENDER, İ. (2009), ***"Başarı ve yetenek kestirimlerinde yeni bir yaklaşım: Bilgisayar ortamında bireyselleştirilmiş testler*** (Computerized adaptive tests-CAT)", CITO Egitim Kuram ve Uygulama, V: 39-48.

KANE, M. (2012), ***"Validation Strategies: Delineating and Validating Proposed Interpretations and Uses of Test Scores"***, (Ed. T. M. Haladyna), Handbook of Test Development, (c.1, s.64-80) içinde. New York: Routledge.

KAPTAN, F. (1993), ***Yetenek Kestiriminde Adaptive (Bireyselleştirilmiş) Test Uygulaması ile Geleneksel Kâğıt-Kalem Testi Uygulamasının Karşılaştırılması,*** Yayımlanmamış doktora tezi, Hacettepe Üniversitesi, Sosyal Bilimler Enstitüsü, Ankara.

KARAKAYA, İ., UMUR, Ö. Ç. ve GÜMÜŞ, G. (2024), ***"Üst Düzey Zihinsel Becerilerin Ölçülmesinde Açık Uçlu Maddelerin Yapay Zekâ Destekli Hazırlanması"***, Journal of Applied Measurement and Evaluation, XV, 2: 55-72.

KELECİOĞLU, H. (2024), ***Test eşitleme,*** Ankara: EPODDER.

KELECİOĞLU, H. ve GÖÇER ŞAHİN, S. (2014), ***"Geçmişten Günümüze Geçerlik",*** Journal of Measurement and Evaluation in Education and Psychology, V, 2: 1-11. https://doi.org/10.21031/epod.41706

KEZER, F. (2013), ***"Bilgisayar Ortamında Bireye Uyarlanmış Test Yöntemi",*** Journal of Educational Sciences & Practices, XII, 24: 177-194.

KIYAK, Y. S., BUDAKOĞLU, I. İ., COŞKUN, Ö. ve KOYUN, E. (2023), ***"The First Automatic Item Generation in Turkish for Assessment of Clinical Reasoning in Medical Education",*** World of Medical Education, XXII, 66: 72-90. https://doi.org/10.25282/ted.1225814

KİLMEN, S. (2024), ***"Ölçme ve Değerlendirmede Temel Kavramlar"***, (Ed. N. Çıkrıkçı-Demirtaşlı), Eğitimde Ölçme ve Değerlendirme, (s.34–68) içinde. Ankara: Anı yayıncılık.

KUCAM, E. (2023), ***Eğitimde Otomatik Madde Üretimine Giriş.*** Ankara: EPODDER.

KUTLU, Ö. ve ALTINTAŞ, Ö. (2021), ***"Psikolojik Ölçmelerin Kısa Tarihi ve 21. Yüzyılda Sınıf İçi Durum Belirleme Anlayışı"***, Trakya Eğitim Dergisi, XI, 3: 1599-1620. https://doi.org/10.24315/tred.896121

LANE, S., RAYMOND, M. R., HALADYNA, T. M. ve DOWNING, S. M. (2012), ***"Test Development Process"***, (Ed. Downing, S. M. ve Haladyna, T. M.), Handbook of Test Development, (c.1, s.3-18) içinde. New York: Routledge.

LEACH, M. M. ve OAKLAND, T. (2007), ***"Ethics Standards Impacting Test Development and Use: A Review of 31 Ethics Codes Impacting Practices in 35 Countries"***, International Journal of Testing, VII, 1: 71-88. https://doi.org/10.1080/15305050709336859

LEIGHTON, J. P. ve GIERL, M. J. (2007), ***Cognitive Diagnostic Assessment for Education Theory and Applications,*** Cambridge: Cambridge University Press.

LIU, Q., ZHUANG, Y., BI, H., HUANG, Z., HUANG, W., LI, J., YU, J., LIU, Z., HU, Z., HONG, Y., PARDOS, Z. A., MA, H., ZHU, M., WANG, S. ve CHEN, E. (2024), ***Survey of Computerized Adaptive Testing: A Machine Learning Perspective,*** Cornell University. https://doi.org/10.48550/arxiv.2404.00712

LIVINGSTON, S. A. (2004), ***Equating Test Scores (Without IRT),*** Educational Testing Service.

LIVINGSTON, S. A. (2006), ***"Item Analysis",*** (Ed. S. M. Downing ve T. M. Haladyna), Handbook of Test Development (s.421-441) içinde. New Jersey: Lawrence Erlbaum Publishing.

LORD, F. M. (1968), ***"An Analysis of The Verbal Scholastic Aptitude Test Using Birnbaum's Three-Parameter Logistic Model",*** Educational and Psychological Measurement, XXVIII: 989-1020. https://doi.org/10.1177/001316446802800401

LORD, F. M. (1980), ***Applications of Item Response Theory to Practical Testing Problems***, Hillsdale, (NJ): Lawrence Erlbaum Associates.

LORD, F. M. (1983), ***"Unbiased Estimators of Ability Parameters, of Their Variance, and of Their Parallel Forms Reliability"***, Psychometrika, XXXXVIII: 233-245. https://doi.org/10.1007/BF02294018

LORD, F. M. ve NOVIC, M. R. (1968), ***Statistical Theories of Mental Testing Scores,*** Addison-Canada: Wesley Publishing Company.

MEHRENS, W. A. ve LEHMAN, I. J. (1991), ***Measurement and Evaluation in Education and Psychology,*** Canada: Wadsworth/Thomson Learning.

MEREDITH, W. (1993), ***"Measurement Invariance, Factor Analysis, and Factorial Invariance"***, Pyschometrika, LVIII: 525-543. https://doi.org/10.1007/bf02294825

MILLER, M. D., LINN, R. L. ve GRONLUND, N. E. (2009), ***Measurement and assessment in teaching,*** NJ: Upper Saddle River, Pearson Education.

MILLMAN, J. ve GREENE, J. (1989), ***"The Specification and Development of Tests of Achievement and Ability",*** (Ed. R.L. Linn), Educational Measurement (s.335-366) içinde. New York: McMillan.

MUÑIZ, J. ve FONSECA-PEDRERO, E. (2019), ***"Ten Steps for Test Development"***, Psicothema, XXXI, 1: 7-16. https://doi.org/10.7334/psicothema2018.291

OHTA, Y., SATSUMA, J., TAKAHASHI, D. ve TOKIHIRO, T. (1987), ***"An Elementary Introduction to Sato Theory"***, Progress of Theoretical Physics Supplement, LXXXXIV: 210-241. https://doi.org/10.1143/PTPS.94.210

OKUL, Ö. ve UYGUN, D. Ç. (2024, Mayıs 21-24), ***"Bilgisayar Ortamında Bireye Uyarlanmış Test Konulu Tezlerin Tematik ve Metodolojik Açıdan İncelenmesi",*** International Eurasian Educational Research Congress, Kocaeli Üniversitesi, Kocaeli.

OSTERLIND, S. J. (1990), ***"Toward A Uniform Definition of A Test Item"***, Educational Research Quarterly, XIV, 4: 2-5.

OSTERLIND, S. J. (2002), ***Constructing Test Items: Multiple-Choice,*** Constructed-Response, Performance, and Other Formats, New York: Kluwer Academic Publishers.

ÖZBAŞI, D. ve DEMİRTAŞLI, N. (2016), ***"Bilgisayar Okuryazarlığı Testinin Bilgisayar Ortamında Bireye Uyarlanmış Test Olarak Geliştirilmesi"***, Journal of Measurement and Evaluation in Education and Psychology, VI, 2: 218-237. https://doi.org/10.21031/epod.79491

ÖZÇELİK, D. A. (1981), ***Test Hazırlama Kılavuzu,*** ÜSYM-TAB Eğitim Yayınları.

PAEK, I. (2018), ***"Understanding Differential Item Functioning and Item Bias in Psychological Instruments"***, Psychology and Psychotherapy: Research Study, I, 3: 1-2. https://doi.org/10.31031/pprs.2018.01.000514

PRICE, L. R. (2016), ***Psychometric Methods Theory Into Practise,*** New York: Guilford Publications..

RECKASE, M. D. (2010), ***"NCME 2009 Presidential Address: What I Think I Know",*** Educational Measurement: Issues and Practice, XXIX, 3: 3-7. https://doi.org/10.1111/j.1745-3992.2010.00178.x

ROID, G. H. ve HALADYNA, T. M. (1982), ***A Technology For Test-Item Writing,*** Academic Press. https://ci.nii.ac.jp/ncid/BA19238136

SAYIN, A. (2024), ***"Evaluating the psychometric characteristics of generated visual reading comprehension items"***, Bartın University Journal of Faculty of Education, XXIII, 2: 380-395. https://doi.org/10.14686/buefad.1424213

SAYIN, A. ve BOZDAĞ, S. (2024), ***Çeşitli Yönleri ile Otomatik Madde Üretimi,*** Ankara: EPODDER.

SCALISE, K. ve ALLEN, D. D. (2015). ***"Use Of Open-Source Software For Adaptive Measurement: Concerto As An R-Based Computer Adaptive Development and Delivery Platform"***, British Journal of Mathematical and Statistical Psychology, LXVIII, 3: 478-496. https://doi.org/10.1111/bmsp.12057

SELÇUK, KARAKAYA ve YILMAZ (2024), ***"Examination of Teachers' Opinions On The Concept of New Generation Item According to Some Variables"***, e-Kafkas Journal of Educational Research, XI: 33-47. https://doi.org/10.30900/kafkasegt.1435336

SHAVELSON, R. J. ve WEBB, N. M. (1991), ***Generalizability Theory: A Primer,*** USA: Sage Publications. https://doi.org/10.5860/choice.29-4192

STIGGINS, R. (2014), ***"Improve Assessment Literacy Outside of Schools too"***, Phi Delta Kappan, LXXXXVI, 2: 67-72. https://doi.org/10.1177/0031721714553413

ŞAHİN, M. G., YILDIRIM, Y. ve ÖZTÜRK, N. B. (2023), ***"Examining the Achievement Test Development Process in the Educational Studies",*** Participatory Educational Research, X, 1: 251-274. http://dx.doi.org/10.17275/per.23.14.10.1

TDK. (2024), ***Redaksiyon, Türk dil kurumu sözlükleri içinde.*** Erişim adresi: https://sozluk.gov.tr/

TEKİN, H. (1977), ***Eğitimde Ölçme ve Değerlendirme,*** Hacettepe Üniversitesi Eğitim Fakültesi Yayınları.

TERCAN, S. S. (2010), ***"Çevrimiçi Uyarlamalı Bir Test Aracı Geliştirilmesi, Uygulanması ve Öğrenci Görüşlerinin Belirlenmesi"***, Yayımlanmamış Yüksek Lisans tezi, Sakarya Üniversitesi, Sosyal Bilimler Enstitüsü, Sakarya.

TIAN, J., MIAO, D., ZHU, X. ve GONG, J. (2007), ***"An Introduction to The Computerized Adaptive Testing",*** US-China Education Rewiew, IV, 1: 72-81. http://files.eric.ed.gov/fulltext/ED497385.pdf

TIFFIN-RICHARDS, S. P. ve PANT, H. A. (2017), ***"Arguing Validity in Educational Assessment"***, (Ed. Leutner, D., Fleischer, J., Grünkorn, J., Klieme, E.), Competence Assessment in Education. Methodology of Educational Measurement and Assessment (s.469–485) içinde. Springer, Cham. https://doi.org/10.1007/978-3-319-50030-0_27

TIPPINS, N. T. (2015), ***"Technology and Assessment in Selection",*** Annual Reviews, II, 1: 551-582. https://doi.org/10.1146/annurev-orgpsych-031413-091317

THORNDIKE, R. L. (1982), ***Applied Psychometrics,*** Boston: Houghton Mifflin.

THORNDIKE, R. M. ve HAGEN, E. P. (1964), ***Measurement and Evaluation in Psychology and Education,*** NY: John Wiley & Sons, Inc.

THOMPSON, N. A. ve WEISS, D. J. (2011), ***"A Framework for the Development of Computerized Adaptive Tests",*** Practical Assessment, Research, and Evaluation, XVI, 1: 1-9. https://doi.org/10.7275/wqzt-9427

TURGUT, F. ve BAYKUL, Y. (2010), ***Eğitimde Ölçme ve Değerlendirme,*** Ankara: Pegem Akademi.

ULUMAN, M. (2023), ***"A Review of Achievement Test Development in Türkiye Regarding The Achievement Test Development Process"***, Journal of Human and Social Sciences, VI, 2: 377-405.

VAN DER LINDEN, W. J. (2017), ***Handbook of Item Response Theory,*** Volume One: Models, Chapman and Hall/CRC.

VAN DER LINDEN, W. J. (2018), ***Handbook of Item Response Theory,*** Volume 3: Applications, CRC Press.

VON DAVIER, A. A. (2010), ***Statistical Models for Test Equating,*** Scaling, and Linking, Springer Science & Business Media.

WAINER, H. (Ed.) (2000), ***Computerized Adaptive Testing: A Primer,*** Hillsdale, NJ: Lawrence Erlbaum Associates.

WEISS, D. J. (2004), ***"Computerized Adaptive Testing For Effective and Efficient Measurement İn Counseling And Education"***, Measurement and Evaluation in Counseling and Development, XXXVII, 2: 70-84. https://doi.org/10.1080/0748 1756.2004.11909751

WEISS, D. J. (2011), ***"Better data from better measurements using computerized adaptive testing"***, Journal of Methods and Measurement in the Social Sciences, II, 1: 1-27. https://doi.org/10.2458/v2i1.12351

WEISS, D. J. ve KİNGSBURY, G. G. (1984), ***"Application of Computer Adaptive Testing to Educational Problems"***, Journal of Educational Measurement, XXI, 4: 361-375. https://doi.org/10.1111/j.1745-3984.1984.tb01040.x

WELLS, C. S. (2021), ***"Methods Based on Item Response Theory"***, (Ed. Craig S. Wells), Assessing Measurement Invariance for Applied Research (s.161-244) içinde. Cambridge University Press. https://doi.org/10.1017/9781108750561.005

WILIAM, D. (2004), ***Keeping Learning On Track: Integrating Assessment With Instruction,*** 30th annual conference of the International Association for Educational Assessment (IAEA) held in June 2004, Philadelphia, PA.

ZHANG, S. ve CHANG, H. (2016), ***"From Smart Testing to Smart Learning: How Testing Technology Can Assist The New Generation of Education",*** International Journal of Smart Technology and Learning (IJSMARTTL), I, 1: 67-92. https://doi.org/10.1504/ijsmarttl.2016.078162

ZUMBO, B. D. (1999), ***A Handbook on the Theory and Methods of Differential Item Functioning (DIF): Logistic Regression Modeling as a Unitary Framework for Binary and Likert-Type (Ordinal) Item Scores,*** Ottawa, ON: Directorate of Human Resources Research and Evaluation, Department of National Defense.

Extended abstract

Education is a continuous enterprise whose every stage should be systematically monitored and informed by feedback. This feedback is operationalized through measurement and evaluation practices, which together constitute a governance system for instructional quality. In education and psychology, such practices are typically enacted via tests. The roots of modern test development can be traced to early pioneers in psychology and education. Standardized assessments-exemplified by intelligence tests-sought to quantify and compare individual differences, thereby foregrounding early challenges of standardization and the need to define and validate the constructs under measurement (Curtis & Glaser, 1984).

Historically, psychometric inquiry into test development and its attendant quality issues has been anchored in two principal theoretical frameworks: Classical Test Theory (CTT) and Item Response Theory (IRT), also known as Latent Trait Theory (LTT). CTT emerged first; IRT gained prominence with its explicit modeling of item and person parameters and its capacity to support modern test applications. Generalizability Theory-commonly regarded as an extension of CTT-decomposes error into multiple sources (facets), providing a powerful design-based approach to reliability that has materially advanced test development. Sato Test Theory, with origins in the Far East, offers additional, context-sensitive contributions within specific testing scenarios.

Tests serve multiple functions: diagnosing educational needs, monitoring instructional quality, selecting and placing students into academic programs, and supporting counseling and clinical decision making. Their overarching objective is to detect intra- and inter-individual differences with precision. Psychological tests, by design, measure samples of behavior, enabling inferences about individual variation (Price, 2017). In this sense, tests can be defined as standardized instruments used under uniform conditions to ascertain individual characteristics (Anastasi, 1988).

Within a test, items constitute the elements most directly responsible for capturing the focal construct. An item is the smallest scorable unit (Turgut & Baykul, 2021): a minimal, meaningful segment designed to elicit a response indicative of a slice of the target trait and to which a score is assigned-though it is not, in itself, the complete unit of analysis (Erkuş, 2014). In assessments of mental attributes, the item functions as a stimulus-response measurement unit intended to provoke behavior from which performance on a psychological construct (e.g., knowledge, ability, disposition, trait) can be inferred (Osterlind, 1990). Test validity, therefore, depends critically on items that are purpose-aligned and that adequately sample the relevant behavioral domain; failures here risk construct underrepresentation or construct-irrelevant variance.

Item writing was initially viewed as an art (Ebel, 1951). Irrespective of the guiding measurement theory, however, tests assembled from a sufficiently large pool of high-quality items are typically more accurate and hence more reliable. Effective item writing is a learned craft that integrates scientific principles with practical artistry; it requires structured training and deliberate practice. Novice item writers frequently produce flawed items that, at best, tap recognition of basic facts rather than higher-order cognition. Common root causes include insufficient domain expertise, limited

command of the linguistic structure of the test's language, and gaps in measurement and evaluation knowledge (Lane et al., 2012).

The literature on test development and item writing is substantial. Influential international references include Millman and Greene (1989), Irvine and Kyllonen (2002), Osterlind (1990, 2002), Haladyna (2004), Gierl and Haladyna (2012), Lane et al. (2012), and Haladyna and Rodriguez (2013). Comprehensive measurement texts that contain extensive treatments of test development and item writing include Crocker and Algina (1986), Ebel and Frisbie (1991), Mehrens and Lehman (1991), Leighton and Gierl (2007), Miller et al. (2009), Cohen and Swerdlik (2010), Thorndike and Thorndike-Christ (2014), and Price (2017). In Turkish, the earliest dedicated monograph on test development and item writing is Özçelik (2013), while substantial coverage is also present in works by Erkuş (2014), Turgut and Baykul (2021), and Tekin (2023), among others.

The contemporary landscape is defined by the convergence of technological innovation and modern psychometric theory. IRT and computer adaptive testing (CAT) have transformed test design, administration, and scoring, enabling personalized and efficient assessment pathways. Looking ahead, significant innovations likely include integrating artificial intelligence (AI) into test construction, further scaling CAT, and addressing ethical considerations with rigor. Natural language processing (NLP) and machine learning (ML) hold particular promise for automating item generation, diversifying content, and tailoring assessments to learner profiles—potentially improving scalability and responsiveness.

These opportunities are accompanied by nontrivial risks. Over-reliance on AI may obscure the indispensable, context-sensitive expertise of subject-matter experts, weakening construct explication and fidelity. Data and algorithmic biases can seep into automatically generated items, jeopardizing fairness and validity. To maintain psychometric rigor, relevance, and ethical integrity, a balanced human–machine workflow is essential: harness AI for efficiency and breadth, while retaining expert oversight to ensure construct coherence, content representativeness, and equity (Zhang & Chang, 2016).

In examination-oriented systems, tests frequently dominate determinations of course achievement. In Turkey, tests are widely employed to gauge educational attainment. The TOAD platform—which indexes domestically developed instruments-lists 13,439 scales or tests authored by 15,333 individuals. In such a prolific environment, it becomes crucial to identify the sources and references that practitioners most frequently consult, both to codify best practices and to guide new test developers toward high-quality, broadly accepted resources.

As the field advances, three interlocking priorities stand out: (1) ethically grounded AI integration (with safeguards against bias and misuse), (2) continued refinement and broad adoption of adaptive testing methodologies, and (3) explicit attention to inclusivity, accessibility, and responsiveness to diverse learner and practitioner needs. Pursued jointly, these priorities can support a future in which assessments are more personalized, adaptive, and context-aware, while remaining psychometrically sound and equitable. Ultimately, test development and item writing are specialized undertakings.

Researchers should engage with globally validated, reliable sources and track technological developments closely-embracing innovations that demonstrably enhance validity, reliability, fairness, and utility, while preserving the expert judgment and ethical standards that underpin high-quality assessment.

Etik Kurul İzni
Bu çalışma için etik kurul izni gerekmemektedir. Yaşayan hiçbir canlı (insan ve hayvan) üzerinde araştırma yapılmamıştır. Makale edebiyat sahasına aittir.

Çatışma Beyanı
Makalenin yazarı, bu çalışma ile ilgili herhangi bir kurum, kuruluş, kişi ile arasında mali çıkar çatışması bulunmadığını beyan eder.

Destek ve Teşekkür
Çalışmada herhangi bir kurum ya da kuruluştan destek alınmamıştır.

Ethics Committee Permission
Ethics committee permission is not required for this study. No research has been conducted on any living creature (human and animal). The article belongs to the field of literature.

Deconfliction Statement
The author of the article declares that there is no conflict of financial interest between him and any institution, organization, person related to this study.

Support and Thanks
Support was not received from any institution or organization in the study.