16. BÖLÜM

YETERLİKLERİN VE NORMLARIN BELİRLENMESİ

DETERMINING PROFICIENCY LEVELS AND NORMS

Doç. Dr. Ahmet Salih ŞİMŞEK

YETERLİKLERİN VE NORMLARIN BELİRLENMESİ

Doç. Dr. Ahmet Salih ŞİMŞEK
Kırşehir Ahi Evran Üniversitesi

Öz

Bu bölüm, eğitimde ve psikolojide ölçme sonuçlarının anlamlandırılmasında temel alınan yeterlik, norm ve standart belirleme kavramlarını ele almaktadır. Ölçme aracından elde edilen ham puanların kendi başına çoğu zaman yeterli bilgi sunmadığı, bu puanların ancak bir referans çerçevesi içinde yorumlandığında karar süreçleri için anlam kazandığı vurgulanmaktadır. Bu doğrultuda yeterlik, bireyin önceden tanımlanmış bilgi ve beceri setlerini ne düzeyde karşıladığını gösteren mutlak bir çerçeve olarak açıklanırken, norm bireyin performansını ait olduğu referans grubun performansıyla karşılaştıran göreli bir bilgi kaynağı olarak ele alınmaktadır. Bölümde performans düzeyi betimleyicileri, ölçmenin standart hatası, güven aralığı, kesme puanı, test eşitleme, bağlama ve dikey ölçekleme süreçleri, puanların geçerli ve tutarlı biçimde yorumlanabilmesi açısından tartışılmaktadır. PISA okuduğunu anlama düzeyleri ile Diller İçin Ortak Avrupa Başvuru Metni dil yeterlik düzeyleri, yeterliklerin basitten karmaşığa doğru nasıl yapılandırılabileceğini gösteren örnekler olarak sunulmaktadır. Normların belirlenmesi kapsamında yaş normları, sınıf düzeyi normları, büyüme normları, yerel normlar ve koşullu normlar açıklanmakta, yüzdelik, yüzdelik sıra, standart puanlar ve norm eğrileri üzerinden normatif bilginin raporlanması değerlendirilmektedir. Norm eğrilerinin üretiminde örneklem seçimi, norm grubu, ağırlıklandırma, tasarım etkisi, hata kestirimi, düzgünleştirme, LMS yöntemi ve kantil regresyonu gibi süreçlerin rolü üzerinde durulmaktadır. Bölümün son kısmında Angoff, Modifiye Angoff, Ebel, Evet/Hayır, Nedelsky, Bookmark ve Basket yöntemleri, kesme puanlarının belirlenmesinde kullanılan test merkezli standart belirleme yaklaşımları olarak incelenmektedir. Böylece bölüm, ölçme sonuçlarının yalnızca sayısal değerler olarak değil, geçerli, güvenilir, adil ve savunulabilir kararların dayanağı olarak nasıl kullanılabileceğine ilişkin kapsamlı bir çerçeve sunmaktadır.

Anahtar Kelimeler

Yeterlik, Norm, Kesme Puanı, Norm Eğrileri, Standart Belirleme

DETERMINING PROFICIENCY LEVELS AND NORMS

Assoc. Prof. Dr. Ahmet Salih ŞİMŞEK
Kırşehir Ahi Evran University

Abstract

This chapter discusses the concepts of proficiency, norms and standard setting, which constitute a central basis for interpreting measurement results in education and psychology. It emphasizes that raw scores obtained from a measurement instrument often do not provide sufficient information on their own and that such scores become meaningful for decision-making only when they are placed within a reference framework. In this respect, proficiency is explained as an absolute framework indicating the extent to which an individual meets predefined sets of knowledge and skills, while norms are treated as a relative source of information that compares an individual's performance with the performance of a reference group. The chapter addresses performance level descriptors, standard error of measurement, confidence intervals, cut scores, test equating, linking and vertical scaling as key processes for valid and consistent score interpretation. The PISA reading proficiency levels and the Common European Framework of Reference for Languages are presented as examples showing how proficiency can be structured from basic to more complex levels. Within the context of norm development, age norms, grade-level norms, growth norms, local norms and conditional norms are explained, and the reporting of normative information through percentiles, percentile ranks, standard scores and norm curves is discussed. The chapter also considers the role of sample selection, norm groups, weighting, design effects, error estimation, smoothing, the LMS method and quantile regression in the production of norm curves. In the final part, Angoff, Modified Angoff, Ebel, Yes/No, Nedelsky, Bookmark and Basket methods are examined as test-centered standard-setting approaches used to determine cut scores. Overall, the chapter provides a comprehensive framework for understanding how measurement results can be used not merely as numerical values but as a basis for valid, reliable, fair and defensible decisions.

Keywords

Proficiency, Norms, Cut Score, Norm Curves, Standard Setting

1. GİRİŞ

Eğitimde ve psikolojide gerçekleştirilen değerlendirmeler, bireylerin belirli özelliklerine ilişkin geçerliği ve güvenirliği yüksek bilgiler elde etmeyi ve bu bilgilere dayalı isabetli kararlar vermeyi amaçlar. Ancak bir ölçme aracından elde edilen ham puan, tek başına çoğu zaman yeterli bir anlam taşımaz. Puanın neyi gösterdiği, hangi düzeye karşılık geldiği ve karar sürecinde nasıl kullanılacağı, ancak belirli bir referans çerçevesi içinde açıklık kazanır. Bu nedenle ölçme sonuçlarının yorumlanması, yalnızca sayısal değerlerin raporlanması değil, bu değerlerin kuramsal ve psikometrik açıdan anlamlandırılması sürecidir.

Bir testten elde edilen puanın anlamı, öncelikle hangi amaçla değerlendirildiğine bağlıdır. Bireyin tanımlı öğrenme çıktılarını, bilgi ve beceri setlerini ya da mesleki standartları ne düzeyde karşıladığı sorgulanıyorsa yeterlik kavramı öne çıkar. Yeterlik, bireyin bir süreklilik üzerindeki konumunu gösterir ve bu konum "Temel", "Yeterli", "İleri" gibi performans düzeyleri aracılığıyla daha anlaşılır hâle getirilir. Bu düzeylerin yalnızca etiket olarak kalmaması için Performans Düzeyi Betimleyicileri önem kazanır. Bu betimleyiciler, her yeterlik düzeyindeki bireyin tipik olarak neleri bildiğini ve neleri yapabildiğini açık biçimde tanımlar.

Norm ise puanın anlamını başka bir yönden tamamlar. Bireyin performansı, ait olduğu ya da karşılaştırıldığı referans grubun dağılımı içinde ele alındığında göreli konum bilgisi elde edilir. Böylece puan, yalnızca bireyin kendi yeterlik düzeyi hakkında değil, benzerleri arasındaki yeri hakkında da bilgi verir. Bu nedenle normların belirlenmesinde norm grubunun iyi tanımlanması, örneklemin evreni temsil etmesi ve gerektiğinde ağırlıklandırma gibi istatistiksel düzeltmelerin yapılması gerekir.

Puanların yorumlanmasında gözlenen puanın hataya açık bir gösterge olduğu da göz ardı edilmemelidir. Ölçmenin standart hatası, güven aralığı, kesme puanı, test eşitleme, bağlama ve dikey ölçekleme gibi süreçler, yeterlik düzeylerinin ve normların farklı test formları, zamanlar ve gruplar arasında tutarlı biçimde yorumlanmasını sağlar. Bu bölümde yeterlik ve norm kavramları bu temel bağlamda ele alınmakta, normların belirlenmesi ve raporlanması ile standart belirleme yöntemleri ölçme sonuçlarının daha geçerli, güvenilir ve savunulabilir kararlara dönüştürülmesi açısından incelenmektedir.

2. YETERLİK VE NORM KAVRAMLARI

Bir testten elde edilen ham puan, kendi başına çoğu zaman yetersiz bir bilgi sunar. Örneğin, bir öğrencinin matematik testinden 75 puan alması ne anlama gelir? Bu puan iyi midir, yeterli midir, yoksa gelişim mi gerektirir? İşte bu sorulara yanıt verebilmek için puanın arkasındaki anlamı ortaya koyan iki temel çerçeveye ihtiyaç duyulur: yeterlik ve

norm. Yeterlik, bireyin önceden tanımlanmış bilgi ve beceri setlerini ne düzeyde karşıladığını ifade eden mutlak bir çerçeve sunar. Norm ise bireyin performansını, ait olduğu referans grubun performansıyla karşılaştırarak göreli bir bilgi sağlar.

Bir bireyin tanımlı öğrenme çıktılarını ya da mesleki standartları hangi düzeyde karşıladığını gösteren, doğrudan gözlenemeyen (gizil) ve bir süreklilik arz eden özelliktir. Ölçme ve değerlendirmede bu süreklilik, "Temel", "Yeterli", "İleri" gibi etiketlerle tanımlanan performans düzeyleri aracılığıyla somut ve anlaşılır hâle getirilir. Bu sınıflama, basit bir "geçti/kaldı" kararının ötesine geçerek, bireyin beceri yelpazesinin neresinde durduğuna dair detaylı bilgi sunar. Yeterliğin tanımlanması yalnızca teknik bir işlem değil, aynı zamanda paydaşlarla etkili bir iletişim kurma eylemidir. Karar vericiler puanın arkasındaki anlamı, adaylar güçlü ve zayıf yönlerini, uygulayıcılar ise sınıflama kararlarının dayanağını bilmek ister (AERA, APA & NCME, 2014; Kane, 2013).

Bu iletişimin temelini, Performans Düzeyi Betimleyicileri (Performance Level Descriptors, PLDs) olarak adlandırılır. PLD'ler, her bir yeterlik düzeyindeki bir bireyin tipik olarak neleri bildiğini ve neleri yapabildiğini açık, gözlemlenebilir ve sade ifadelerle tarif eder. Bu ifadeler, ilgili yeterlik düzeyini gösteren davranışlar veya performansın işevuruk tanımları olarak da düşünülebilir. Böylece düzey etiketleri soyut birer isim olmaktan çıkar ve tüm paydaşlar için ortak bir referans çerçevesine dönüşür. İyi tasarlanmış PLD'ler, farklı zamanlarda veya farklı test formlarında uygulanan ölçümler arasında tutarlı bir anlam köprüsü kurarak programların iletişim dilini standartlaştırır (Egan ve ark., 2012; Lane ve ark., 2016).

Yeterlik, ölçme kuramında gizil bir özelliktir ve bireyin konumu bir süreklilik üzerinde yer alır. Gözlenen puan bu konumun hataya açık bir izdüşümüdür ve bu nedenle yorum yapılırken ölçmenin standart hatası ile güven aralığı dikkate alınır. Bu değerler, bireyin gözlenen puanının gerçek yeterlik düzeyinden ne kadar sapabileceğini gösterir ve gerçek yeterliğin hangi puan aralığında bulunma olasılığının yüksek olduğunu belirtir. Özellikle kesme puanı çevresinde verilen kararlar ölçme hatasından kaynaklanan belirsizlikten daha fazla etkilenir ve raporun bu durumu dürüstçe göstermesi beklenir. Bu yaklaşım, puanı tek bir sayı olmaktan çıkarır ve bireyin gerçek puanının büyük olasılıkla içinde bulunduğu bir anlam aralığına dönüştürür. Sertifikasyon gibi yüksek riskli karar bağlamlarında kesme puanına çok yakın adaylar için daha temkinli davranmayı gerektiren ihtiyatlılık ilkesi öne çıkar ve sınıflama olası hatalarının sonuçları değerlendirilir. İzleme ve raporlama odaklı programlarda ise yeterlik bilgisinin yıl içi değişimi ve yıllar arası eğiliminin izlenmesi önem kazanır. Hangi amaç söz konusu olursa olsun, yeterlik sonuçlarının raporlanmasında ölçme hatasını ve olası karar sapmalarını görünür kılması bilimsel doğruluk ve kamuya hesap verebilirlik için temel kabul edilir (AERA, APA & NCME, 2014; Kane, 2013).

Yeterlik değerlendirmesinin temelinde ise ölçek puanlarının anlamının korunması yer alır. Farklı test formlarından gelen puanlar test eşitleme süreçleriyle aynı ölçeğe taşınır ve böylece formlar arası karşılaştırma hataları azaltılır. Bu sayede farklı formları alan bireylerin puanları adil bir şekilde karşılaştırılabilir. Farklı ama benzer içerikli testlerin daha esnek varsayımlarla ilişkilendirildiği durumlarda bağlama (linking) kullanılır ve yorum kapsamı dikkatli biçimde sınırlandırılır. Bağlama, testlerin farklı zamanlardaki uygulamalarını ilişkilendirmek için de kullanılabilir. Sınıf düzeyleri boyunca ortak bir gelişim seyrini izlenmek istendiğinde dikey ölçekleme (vertical scaling) tercih edilir. Bu süreçlerin amacı, belirlenen yeterlik düzeylerinin ve kesme puanlarının anlamının yıl, form ve grup değişse bile aynı kalmasını sağlamaktır (Kolen & Brennan, 2014; Lane ve ark., 2016). Böylece, örneğin bir sonraki yıl farklı bir test formu uygulansa bile, daha önce belirlenmiş olan kesme puanı kullanılarak tutarlı ve geçerli kararlar verilebilir.

Sertifikasyon gibi mutlak bir standarda ulaşmanın kritik olduğu bağlamlarda ölçüt bilgisi kararın merkezinde yer almaktadır. Alanyazın kesme puanlarının teknik kanıtlara ve bağlamsal gerekçelere birlikte dayandırılması gerektiği ve değer yargısı içeren seçimlerin şeffafça raporlanması gerektiği vurgulanmaktadır (AERA, APA & NCME, 2014; Kolen & Brennan, 2014).

Geniş ölçekli değerlendirme uygulamaları bu amaçla açık düzey tanımlarını kullanır. Aşağıda iki yerleşik referans çerçevesi üzerinden yeterlik düzeylerinin nasıl tanımlandığı listelenmektedir. Bu örnekler, farklı alanlarda yeterliğin nasıl basitten karmaşığa doğru aşamalı bir yapıda tanımlandığını göstermektedir. Uluslararası Öğrenci Değerlendirme Programı (PISA) Okuduğunu Anlama Yeterlik Düzeyleri (OECD, 2019), 15 yaşındaki öğrencilerin modern topluma katılım için gerekli temel bilgi ve becerilerini ölçer. Okuduğunu anlama alanında belirlenen yeterlik düzeyleri ve bu düzeylere karşılık gelen puan aralıkları şöyledir (Millî Eğitim Bakanlığı, 2024):

- *Düzey 6 (698 puan ve üstü): Bu düzeydeki öğrenciler, istenilen bilginin metin içerisinde saklı olduğu uzun ve soyut metinleri anlayabilir. Bilginin nasıl kullanılacağına karar vermek için çeşitli ölçütler kullanabilir, bilgilerin benzer ve zıt yönlerini karşılaştırabilir ve bu bilgileri bir araya getirebilir. Dış ölçütler kullanarak metnin kaynağı hakkında derinlemesine düşünebilir. Bilginin kaynağı ve geçerliğiyle ilgili ipuçları aracılığıyla metinler arasındaki uyumsuzlukları belirleyebilir. Metinler arası tutarsızlıkları çözebilir, farklı metinlerdeki bilgilerin benzer ve zıt yönlerini karşılaştırabilir.*
- *Düzey 5 (626–698 puan): Bu düzeydeki öğrenciler, uzun metinlerde konuyla ilgili saklı bilgileri bularak bu metinleri kavrayabilir. Kapsamlı metinlere yönelik derin bir anlayış göstererek farklı akıl yürütme yöntemlerini kullanabilir. Çeşitli metin veya kaynaklarda yer alan bilgiler ile soru arasında ilişkiyi kurarak soruları cevaplayabilir. Belli bilgilerden yola çıkarak hipotezler*

oluşturabilir veya var olan hipotezlere ilişkin değerlendirme yapabilir. Karışık ve soyut ifadelerin olduğu durumlarda gerçek ile algı arasındaki ayrımı yapabilir. Çeşitli ipuçlarından yola çıkarak verilen bilginin kaynağının veya içeriğinin objektifliğini değerlendirebilir.

- *Düzey 4 (553–626 puan): Bu düzeydeki öğrenciler, bir veya birden çok metinde yer alan uzun paragrafları anlayabilir. Metni bir bütün olarak ele alarak dile dayalı farklılıkları yorumlayabilir. Çeşitli kaynaklardan yola çıkarak farklı bakış açılarını karşılaştırabilir ve sonuç çıkarabilir. Metne yerleştirilmiş bilgileri arayabilir, bu bilgilere ulaşabilir ve bir araya getirebilir. Bilginin uygunluğunu değerlendirmek için çıkarımlar yapabilir. Metnin dikkat çekici özelliklerinden yola çıkarak yazarların kendi fikirlerini aktarmak için kullandıkları ifadeleri belirleyebilir. Çeşitli metinlerde açıkça ifade edilen iddiaları kıyaslayabilir ve kriterlere göre bilgi kaynağının güvenirliğini değerlendirebilir.*
- *Düzey 3 (480–553 puan): Bu düzeydeki öğrenciler, açık şekilde sunulmadığı durumlarda da metnin genel anlamını ifade edilebilir. Bilgileri derleyebilir, basit ve ileri düzeyde çıkarımlar yapabilir. Çok uzun olmayan metinlerin ana düşüncesini belirlemek, ilişkileri anlamak, bir kelime veya ifadenin anlamını çıkarmak için metnin farklı bölümlerindeki bilgileri bir araya getirebilir. Çeşitli kriterlere göre elde edilen bilgiler arasındaki ilişkileri anlayabilir. Bu düzeydeki öğrenciler, bir veya birkaç metin üzerine derinlemesine düşünebilir. Sunulan bilgilerden yola çıkarak farklı yazarların bakış açılarını kıyaslayabilir.*
- *Düzey 2 (407–480 puan): Bu düzeydeki öğrenciler, orta uzunluktaki metinlerin ana düşüncesini belirleyebilir. Gerekli bilgilerin açıkça verilmediği durumlarda metnin belirli bir bölümündeki ilişkileri anlayabilir ve bir bölümünden anlam çıkarabilir. Birkaç sayfalık metin içerisinde konu ile ilgili sayfayı bulabilir ve konuyla ilgili düzeylerini artırabilir. Gerekli bilgilerin açıkça verildiği durumlarda orta uzunluktaki metinlerin genel amacı ve bu metinlerdeki belirli detaylar üzerine derinlemesine düşünebilir. İddiaları karşılaştırabilir ve bu iddiaları destekleyen düşünceleri belirleyebilir.*
- *Düzey 1a (335–407 puan): Bu düzeydeki öğrenciler, cümlelerin veya kısa paragrafların gerçek anlamını kavrayabilir. Aşina oldukları konularda yazılmış metinlerin ana fikrini ve yazarın amacını anlayabilir. Metinde verilen bilgiler arasında veya metinde verilen bilgiler ile kendi sahip olduğu bilgiler arasında basit bağlantılar kurabilir. Birkaç sayfalık metinler içinde ilgili sayfayı bulabilir ve kısa metinlerde yer alan bilgilere ulaşabilir. Bilgilerin açıkça sunulması durumunda metinlerin genel amacı ve metindeki nispeten önemli bilgiler üzerinde (metnin ana fikri gibi) derinlemesine düşünebilir.*

- *Düzey 1b (262-335 puan): Bu düzeydeki öğrenciler, basit cümlelerin gerçek anlamlarını değerlendirebilir. Soru ve/veya metindeki bilgiler arasında basit bağlantılar kurarak metinlerin gerçek anlamını yorumlayabilir. Tek bir cümle, kısa bir metin veya basit bir listede istenen bilgileri tarayabilir ve bulabilir. Açık bir şekilde istendiğinde birkaç sayfalık metin içerisinde ilgili sayfayı bulabilir.*
- *Düzey 1c (189-262 puan): Bu düzeydeki öğrenciler, kısa ve basit cümlelerin anlamını kavrayabilir. Sınırlı bir süre içinde açık, basit ve somut amaçlar için okuma yapabilir.*

Diller için Ortak Avrupa Başvuru Metni (D-OABM / CEFR) Dil Yeterlik Düzeyleri (Council of Europe, 2020a) Avrupa Konseyi tarafından geliştirilen CEFR (veya Türkçe kısaltmasıyla D-OABM), dil öğrenme, öğretme ve değerlendirme için uluslararası bir standart sağlar. Bu çerçevede tanımlanan altı temel yeterlik düzeyi şöyledir (Council of Europe, 2020b):

- *C2 (Mastery): Neredeyse tüm metin tiplerini anlayabilir. Farklı sözlü ve yazılı kaynaklardan gelen bilgileri, savları ve açıklamaları yeniden oluşturarak tutarlı bir sunumda özetleyebilir. Daha karmaşık durumlarda bile daha ince anlam ayrımlarını fark ederek kendini spontane, çok akıcı ve kesin bir şekilde ifade edebilir.*
- *C1 (Effective Operational Proficiency): Çok çeşitli zorlukta, daha uzun metinleri anlayabilir ve örtük anlamın farkına varabilir. İfadeleri çok fazla aramaksızın akıcı ve spontane bir şekilde kendisini ifade edebilir. Sosyal, akademik ve mesleki amaçlar için dili esnek ve etkili bir şekilde kullanabilir. Metin düzenleme kalıplarının, bağlayıcıların ve bağdaşıklık ögelerin kontrollü kullandığını göstererek karmaşık konularda açık, iyi yapılandırılmış, ayrıntılı metinler üretebilir.*
- *B2 (Vantage): Kendi uzmanlık alanlarındaki teknik tartışmalar dâhil olmak üzere hem somut hem de soyut konulardaki karmaşık metinlerin ana fikirlerini anlayabilir. Her iki tarafı da zorlamadan, hedef dilin kullanıcıları ile mümkün olduğunca düzenli etkileşim kuracak akıcılıkla ve spontane bir biçimde etkileşime girebilir. Çok çeşitli konularda açık, ayrıntılı metinler üretebilir ve çeşitli seçeneklerin avantaj ve dezavantajlarını vererek güncel bir konu hakkında bir bakış açısını açıklayabilir.*
- *B1 (Threshold): Düzenli olarak işte, okulda, serbest zamanlarında vs. karşılaşılan aşina olduğu meselelerdeki açık standart girdilerin ana noktalarını anlayabilir. Dilin konuşulduğu bir bölgede seyahat ederken ortaya çıkması muhtemel çoğu durumla başa çıkabilir. Aşina olduğu veya kişisel ilgi alanına giren konularda basit, bağlantılı metinler üretebilir. Deneyimleri ve olayları, hayalleri, umutları ve emelleri betimleyebilir ve kısaca görüşler ve planlar için nedenler sunabilir ve açıklamalar yapabilir.*

- *A2 (Waystage): Kendisi için en yakın alakalı (ör. çok temel kişisel ve aile bilgileri, alışveriş, yerel coğrafya, istihdam) cümleleri ve sık kullanılan ifadeleri anlayabilir. Aşina olduğu ve rutin konularda basit ve doğrudan doğruya bilgi alışverişini gerektiren basit ve rutin görevlerde iletişim kurabilir. Geçmişi, yakın çevresi ve öncelikli ihtiyaç alanlarındaki meselelerinin boyutlarını basitçe betimleyebilir.*
- *A1 (Breakthrough): Somut türden ihtiyaçların karşılanmasına yönelik aşina olduğu günlük ifadeleri ve çok temel öbekleri anlayabilir ve kullanabilir. Kendini ve başkalarını tanıtabilir ve sorular sorabilir. Nerede yaşadığı, tanıdığı insanlar ve sahip olduğu şeyler gibi kişisel ayrıntılar hakkında sorular sorabilir ve sorulan soruları yanıtlayabilir. Diğer kişinin yavaş ve açık konuşması ve yardım etmeye hazır olması koşuluyla basit bir şekilde etkileşim kurabilir.*

2.1. Norm

Yeterlik bilgisini göreli bir konuma yerleştirir ve bireyin tanımlı bir referans evreninde (örneğin; ülke genelindeki tüm 8. sınıflar) nerede durduğunu gösterir. Normun belirlenmesi, iyi tanımlanmış bir norm grubu gerektirir ve bunun vazgeçilmez koşulu örneklemin evreni temsil etme gücüdür. Evrene genelleme yapabilmek amacıyla, tabakalı örnekleme gibi olasılıklı örnekleme yöntemleri tercih edilir veya daha karmaşık örnekleme tasarımlarından yararlanılır. Örneklem ile evren arasındaki olası sistematik farklar ağırlıklandırma ile istatistiksel olarak düzeltilir.

Normlar, ölçüt-dayanaklı kararın yerine geçmez; onunla birlikte değerlendirilir. Ölçüt-dayanaklı karar bir ölçütün karşılanıp karşılanmadığını söylerken, norm-dayanaklı bilgi aynı kişinin referans grubuna göre nerede durduğu bilgisini verir. İki bilgi birlikte sunulduğunda karar verme süreci daha kapsamlı ve dengeli hâle gelir. Bir aday kesme puanını karşılamış olabilir ve yine de referans grubuna göre performansı görece düşük veya yüksek olabilir. Bu nedenle geniş ölçekli değerlendirme programları, raporlarında genellikle bu iki bilgiyi birlikte verir ve kullanıcıya tek bakışta hem standardı karşılama durumunu hem de göreli konumu sunar (AERA, APA & NCME, 2014; Lane ve ark., 2016).

3. NORMLARIN BELİRLENMESİ

Bir önceki bölümde tanımlandığı gibi normlar, bir puanın belirli bir referans grubuna göre konumunu gösteren karşılaştırma ölçüleridir. Uygulamada normlar birkaç temel eksen boyunca sınıflandırılır. Zaman ekseninde aynı anda toplanan verilere dayanan enine kesit normları ve zaman içindeki ilerlemeyi temel alan büyüme normları kullanılır. Kapsam ekseninde tüm evreni temsil etmeyi amaçlayan ulusal normlar ve belirli bir kurum ya da bölge için hesaplanan yerel normlar tercih edilir. Koşullandırma ekseninde herhangi bir değişkene bağlı olmayan koşulsuz normlar ile geçmiş puan ya da dil durumu gibi değişkenlere bağlı koşullu normlar üretilir. Raporlama ekseninde göreli konumu

doğrudan veren yüzdelikler ile ortak ölçeklere dönüştürülmüş z, T ve stanin gibi standart puanlar yer alır. Modelleme ekseninde ampirik tablolarla sunulan normlar ile dağılımı yumuşatılmış ve gelişimsel süreklilik üzerinde üreten yöntemler bulunur. Bu bütünün güvenilir çalışması üç koşula bağlıdır:

(1) Örneklemin evreni temsil etmesi,
(2) Farklı uygulama dönemleri ve test formlarından kaynaklanan farkların test eşitleme ile yönetilmesi,
(3) Uygulanan dönüşümlerin ve hesaplanan belirsizlik ölçülerinin (örneğin; standart hatalar) raporda açıkça yazılması.

Alanyazın bu ilkeleri temel kabul eder ve normların üretimi ile kullanımı bu çerçeve içinde gerekçelendirilir (AERA, APA & NCME, 2014; Angoff, 1971; Kolen & Brennan, 2014; Lane, Raymond & Haladyna, 2016).

3.1. Temel Norm Grupları ve Karşılaştırma Çerçeveleri

3.1.1. Yaş Normları

Yaş normu, aynı takvim yaşına sahip bireylerin puan dağılımını referans alır. Yorumlama, aynı yaştaki bireyler arasında konum üzerinden yapılır. Geçerli bir yaş normu için evrenin önemli alt gruplarını kapsayan tabakalı örnekleme gereklidir. Örneklem ile evren arasındaki farklılıklar ağırlıklandırma ile düzeltilir. Uygulama zaman aralığı açıkça yazılır çünkü mevsimsellik ve farklı öğrenme fırsatları aynı yaş içindeki puanları etkileyebilir. Raporlarda standart sapma ve hata ölçümleri verilerek örneklem tasarımının varyans üzerindeki etkisi olan tasarım etkisi sunulur. Sınıf temelli öğretim döngülerinin baskın olduğu bağlamlarda yaş ile sınıf bilgisi birlikte sunulur ve böylelikle kavramsal karışıklık önlenir. Bu düzen, yaşa duyarlı gelişim yorumunu anlaşılır ve denetlenebilir bir biçimde sağlayarak kullanıcıların yanlış genelleme yapmalarının önüne geçer (AERA, APA & NCME, 2014; Lane ve ark., 2016).

3.1.2. Sınıf Düzeyi Normları

Sınıf düzeyi normu, aynı sınıfta öğrenim gören öğrencilerin dağılımını temel alır. Yorumlama, aynı sınıftaki öğrenciler arasında konum üzerinden yapılır. Okul yılı içinde öğrenme fırsatları değiştiği için programlarda çoğunlukla sonbahar, kış ve bahar dönemleri için ayrı normlar hesaplanır. Yıllar arası karşılaştırma yapılacaksa test formları ve içerik farklılıkları test eşitleme ile yönetilir. Test eşitleme yapılmadan yıllar arası sınıf normu karşılaştırmaları yanlılığa neden olarak hatayı artırır. Raporlarda dönem bilgisi ve test eşitleme süreci açıkça verilir ve normun, karar verme sürecinde ölçüt-dayanaklı bilgiyi ikame etmediği net biçimde belirtilir. Sınıf normu, göreli konumu tamamlayıcı bir katman olarak yer alır ve aynı tabloda ölçüte göre erişim ve referans gruba göre konum birlikte sunulur (AERA, APA & NCME, 2014; Kolen & Brennan, 2014; Lane ve ark., 2016).

3.1.3. Büyüme Normları

Büyüme normu, bireyin iki zaman noktası arasındaki ilerlemesini, benzer geçmiş puan profiline sahip diğer bireylerin gösterdiği tipik ilerlemeyle karşılaştırır. En yaygın yaklaşım öğrenci büyüme yüzdelikleri (SGP - Student Growth Percentiles) yöntemidir. Bu yöntemde, önce bireyin geçmiş puanlarına göre benzer başlangıç noktasına sahip diğer bireylerden oluşan koşullu bir karşılaştırma grubu tanımlanır. Ardından bireyin yeni puanı bu koşullu dağılımda yüzdelik sıra olarak konumlandırılır. Farklı başlangıç seviyelerinden gelen öğrenciler bu sayede daha adil bir referansla kıyaslanır. Sonuçların güvenirliği için yeterli örneklem büyüklüğü, sağlam bir test eşitleme zinciri ve geçmiş ölçümlerden elde edilen puanların geçerliği gerekir. Raporlarda belirsizlik ve dönemler arası ölçek tutarlılığı açıkça gösterilir. Büyüme normu, ölçüt-dayanaklı yeterlik çizgilerinin yerine geçmez; ilerleme hızına ilişkin ek bir bilgi sağlar ve program hedeflerine iletişimini güçlendirir (AERA, APA & NCME, 2014; Angoff, 1971; Kolen & Brennan, 2014; Lane ve ark., 2016).

3.1.4. Yerel ve Koşullu Normlar

Yerel norm; belirli bir il, ilçe ya da kurum gibi dar bir kapsam için üretilen puan dağılımlarına dayanır. Ulusal evrenden anlamlı biçimde farklılaşan bağlamlarda daha uygun bir referans sunar. Güvenirliği, örneklem büyüklüğüne ve normların güncelleme sıklığına bağlıdır. Raporlarda normun geçerli olduğu kapsam sınırı açıkça yazılır ve standart hata değerleri verilir. Yerel ve ulusal normların birlikte sunulması yanlış genelleme riskini azaltır ve kararın kapsamını netleştirir. Kullanıcı aynı tabloda bireyin yerel bağlamdaki konumu ile ulusal düzeydeki konumu karşılaştırmalı olarak görebilir (AERA, APA & NCME, 2014; Dorans & Cook, 2016).

Koşullu norm, referans grubunu belirli değişkenlere göre daraltarak oluşturulan dağılımlar üzerinden üretilir. En yaygın kullanım, bireyin geçmiş puan profillerine göre koşullandırmadır. Dil durumu, özel eğitim gereksinimi ya da sosyoekonomik değişken gibi ek koşullar kuramsal gerekçeyle modele alınabilir. Ancak model karmaşıklaştıkça, her bir koşula uyan birey sayısı (hücre büyüklükleri) küçülür ve yapılan kestirimlerin standart hatası artar. Bu nedenle değişken seçimi dikkatle yapılır ve her bir alt grup için örneklem yeterliği sağlanır. Koşullu normlar, belirli gruplar için adil bir karşılaştırma algısını güçlendirir ve ölçüt-dayanaklı kararların yanında ayrı bir bilgilendirme kaynağı olarak kullanılır (AERA, APA & NCME, 2014; Betebenner, 2009). Örneğin, anadili Türkçe olmayan öğrenciler için hem genel sınıf normu hem de benzer dil geçmişine sahip öğrencilerden oluşan koşullu norm sunulabilir.

3.2. Normların Raporlanması: Yüzdelikler ve Standart Puanlar

Yüzdelik ve yüzdelik sıra, göreli konumu en doğrudan anlatan göstergelerden biridir. Yüzdelik, puan dağılımı belirli bir yüzdeye bölen puan değeridir. Örneğin, 75. yüzdelik, puanların %75'inin altında kaldığı puandır. Yüzdelik sıra, bireyin puanından daha düşük

puan alanların yüzdesidir. Örneğin 70. yüzdelik sırada olan bir birey, gruptakilerin %70'inden daha yüksek puan almıştır. Bu iki kavram karıştırılmamalıdır. Özellikle dağılımın uçlarında (en düşük ve en yüksek puanlar civarında) küçük puan farklarının büyük yüzdelik sıra değişimlerine yol açabileceği unutulmamalıdır. Bu nedenle raporlarda yüzdelikler çoğu zaman standart puanlarla birlikte verilir ve yüzdelik sıraların yorumlanmasında standart hata ve güven aralığı gibi belirsizlik göstergeleri eklenir (Angoff, 1971; Livingston & Zieky, 1982; Lane ve ark., 2016).

Standart puanlar; ham puanları, ortalaması ve standart sapması önceden tanımlı ortak ölçeklere taşır. z-puanı (ortalama=0, SS=1), T-puanı (ortalama=50, SS=10) ve stanin (1-9 arası dokuz kategori) en yaygın olanlarıdır. Standart puanların en sağlıklı kullanımı test eşitleme sonrasıdır. Çünkü önce farklı test formlarından kaynaklanan farklar giderilir ve sonra ortak ölçek güvenle uygulanır. Raporlarda hangi dönüşümlerin ve bu dönüşümlerde kullanılan parametrelerin (örneğin; ortalama, standart sapma) hangi örneklem grubuna dayalı olarak hesaplandığının açıkça yazılması gerekir (Kolen & Brennan, 2014).

Bu bölümde normların ne olduğu ve hangi farklı amaçlara hizmet ettiği genel bir çerçevede ele alınmıştır. Ancak normların üretimi ve yorumlanması, ölçülen özelliğin doğasına göre önemli farklılıklar gösterebilir. Ayrıca, normatif verinin etkili bir şekilde sunulması için norm eğrileri gibi görselleştirme araçları ve bu eğrilerin altında yatan modelleme teknikleri de önem taşımaktadır. Bir sonraki bölümde, öncelikle norm çalışmasının temelini oluşturan örneklem seçimi ve norm grubu oluşturma süreci ayrıntılı olarak incelenmiş, ardından norm eğrileri ve normların yaşam döngüsü ele alınmış ve son olarak normların bilişsel yeterlikler ve duyuşsal özellikler bağlamında nasıl özelleştiğini, ne gibi ek dikkat noktaları gerektirdiği incelenmiştir.

3.3. Norm Eğrileri, Modelleme ve Normların Geçerlik Süresi

Ölçme ve değerlendirmede elde edilen sayısal veriler, ham formlarıyla genellikle karmaşık ve yorumlanması zordur. Normatif veriyi, yani bir bireyin tanımlı bir referans grubuna göre konumunu, tüm paydaşlar için anlaşılır ve eyleme geçirilebilir bir formata dönüştürmenin en etkili yollarından biri görselleştirmedir. Norm eğrileri, bu görselleştirmenin en güçlü araçlarından biridir. Seçilen bir referans grubunda belirli yüzdeliklerin (örneğin; P10, P50 ve P90) zaman, yaş ya da sınıf düzeyi boyunca nasıl bir gelişim seyri izlediğini gösteren çizgilerdir.

Bu eğriler, basit birer grafikten çok daha fazlasını ifade eder; karmaşık psikometrik ve istatistiksel süreçlerin sonucudur. Güvenilir bir norm eğrisi; temsil gücü yüksek bir örnekleme, farklı test formları arasında tutarlılığı sağlayan test eşitleme süreçlerine, örneklem tasarımının etkisini düzelten ağırlıklandırma işlemlerine ve bu karmaşık yapıdan kaynaklanan hatayı doğru tahmin eden varyans kestirimi yöntemlerine dayanır. Bu nedenle AERA, APA ve NCME (2014) standartları, normları ve dolayısıyla bu

normlardan türetilen eğrileri, altında yatan tüm bu teknik altyapının şeffaf bir şekilde belgelenmesi gerektiğini vurgular. Bu alt bölüm, norm eğrilerinin üretim adımlarını, yorumlama ilkelerini, gelişmiş modelleme tekniklerini ve geçerlik süresi içindeki yönetimini ayrıntılı olarak ele almaktadır (Kolen & Brennan, 2014; Lane, Raymond & Haladyna, 2016).

3.3.1. Norm Eğrilerinin Üretim Aşamaları ve Modelleme Teknikleri

Norm eğrilerinin üretimi, titizlik gerektiren üç temel adımda gerçekleşir. Her adım, nihai görselin geçerliği ve güvenirliği üzerinde doğrudan etkiye sahiptir.

Adım 1: Verinin Hazırlanması ve Psikometrik Temellerin Atılması: Bu aşama, eğrinin üzerine inşa edileceği zemini oluşturur. İlk olarak, norm grubunun hedef evreni temsil etmesi için hesaplanan örneklem ağırlıkları veriye uygulanır. Ardından, programda farklı test formları kullanılıyorsa veya veriler farklı zaman dilimlerinde toplanmışsa, bu farklılıklardan kaynaklanan yanlılığı ortadan kaldırmak için test eşitleme işlemleri uygulanarak tüm puanlar ortak bir ölçeğe getirilir. Son olarak, tabakalı veya çok aşamalı örnekleme gibi karmaşık tasarımların standart hata hesaplamaları üzerindeki etkisini (tasarım etkisi - DEFF) hesaba katmak için çoğaltma ağırlıkları (replicate weights) gibi gelişmiş varyans kestirim araçları hazırlanır (Rust & Rao, 1996). Bu adımlar atlanarak çizilen bir eğri, teknik olarak yanıltıcı olacaktır.

Adım 2: Yüzdeliklerin Kestirimi ve Eğrilerin Düzgünleştirilmesi: Veri hazırlandıktan sonra, her bir yaş veya sınıf düzeyi için ilgili yüzdelik dilimleri (örneğin; P10, P25, P50, P75, P90) kestirilir. Ancak gözlenen (ham) yüzdelikler, örneklemden kaynaklanan rastgele dalgalanmalar içerebilir. Bu gürültüyü (rasgele değişkenliği) azaltmak ve altta yatan gerçek gelişimsel eğilimi ortaya çıkarmak için istatistiksel düzgünleştirme (smoothing) yöntemleri kullanılır. Bu yöntemler, verideki ana örüntüyü korurken küçük ve anlamsız dalgalanmaları ortadan kaldırmayı hedefler. Bu alanda iki yaklaşım öne çıkmaktadır:

1. LMS Yöntemi (Box-Cox Cole-Green): Özellikle sağlık ve gelişim alanlarında yaygın olarak kullanılan bu yöntem, bir puan dağılımını üç temel parametre ile matematiksel olarak modeller:
    - L (Lambda): Box-Cox güç dönüşümü ile dağılımın çarpıklığını düzelterek normale yaklaştırır.
    - M (Mu): Dağılımın medyanını temsil ederek gelişim çizgisinin merkezini belirler.
    - S (Sigma): Dağılımın göreli yayılımını (varyasyon katsayısını) temsil ederek puanların zamanla nasıl farklılaştığını tanımlar. Bu üç parametre, zaman veya yaş ekseni boyunca düzgün eğriler olarak modellenir ve istenen tüm yüzdelikler bu modellenmiş yapıdan matematiksel olarak türetilir. Bu sayede, P10 ve P90 gibi çizgilerin birbiriyle tutarlı ve monoton bir yapı içinde olması sağlanır (Cole & Green, 1992).

2. Kantil Regresyonu (Quantile Regression): Geleneksel regresyonun sadece ortalamayı modellemesinin aksine, kantil regresyonu doğrudan belirli yüzdeliklerin (kantillerin) kendisini bağımsız değişkenlere (örneğin; yaş, sınıf düzeyi) bağlı olarak modeller. Örneğin, P10, P50 ve P90 için ayrı ayrı regresyon eğrileri uydurur. Bu yaklaşım, bağımsız değişkenlerin dağılımın farklı bölgeleri üzerindeki etkilerinin farklı olabildiği durumlarda ve örneklem büyüklüğünün azaldığı (veri seyrekleşmesinin olduğu) uç bölgelerde daha esnek ve sağlam sonuçlar verebilir (Koenker, 2005).

Adım 3: Görselleştirme ve Belirsizliğin Raporlanması: Son aşamada, düzgünleştirilmiş yüzdelik çizgileri grafiğe aktarılır. Ancak bu çizgiler kesin değerleri göstermez. Örneklemden kaynaklanan kestirim hatası içerirler. Bu belirsizliği göstermek için eğrilerin etrafına güven bantları (confidence bands) eklenir. Bu bantlar, çoğaltma ağırlıkları gibi gelişmiş yöntemlerle hesaplanan standart hatalara dayanır ve çizilen eğrinin gerçek evren değerini (örneğin; evrendeki gerçek P50 değerini) hangi olasılıkla içerdiğini gösterir. Eksen etiketleri, ölçek birimleri ve veri toplama dönemi gibi bilgiler grafikte açıkça belirtilerek, eğrinin tam ve şeffaf bir istatistiksel özete dönüşmesi sağlanır.

3.3.2. Norm Eğrilerinin Yorumlanması: Statik Konum ve Büyüme

Norm eğrileri, doğru okunduğunda zengin bir yorumlama çerçevesi sunar. Ancak olası karar hatalarını azaltmak için belirli ilkeler izlenmelidir.

3.3.2.1. Anlık (Statik) Konumun Yorumlanması

Bir öğrencinin puanı, ilgili yaş veya sınıf eğrilerinin üzerinde bir nokta olarak işaretlenir. Bu noktanın konumu birkaç temel bilgiyi verir:

- Merkeze Göre Konum: Noktanın P50 (medyan) çizgisine göre konumu, bireyin referans grubunun ortalamasının altında mı, çevresinde mi, yoksa üstünde mi olduğunu gösterir.
- Dağılımdaki Yer: Noktanın P10 ve P90 gibi uç yüzdelik çizgilerine yakınlığı, bireyin performansının referans grubundaki diğer bireylere göre ne kadar farklılaştığını görselleştirir.
- Ölçek Etkisi: Okuyucu, yüzdelik ölçeğinin doğrusal olmadığını bilmelidir. Puan dağılımın merkezinde (örneğin P45 ile P55 arası) küçük puan farkları gözlenirken, uç bölgelerde (örneğin P90 ile P95 arası) aynı yüzdelik değişimi çok daha büyük bir puan farkına karşılık gelebilir. Bu nedenle, eğriler tek başına bireysel düzeyde kesin sınıflama kararları (örneğin "üstün yetenekli" veya "desteklenmesi gereken") için kullanılmamalıdır. Bu tür kararlar için öğrencinin puanı, ölçmenin standart hatası ve güven aralığı ile birlikte değerlendirilmelidir.

3.3.2.2. Büyüme Bağlamında Yorumlama

Norm eğrileri sadece anlık durumu değil, aynı zamanda ilerlemeyi de yorumlamaya yardımcı olabilir. Bu bağlamda iki tür kullanım vardır:

1. Enine Kesit Eğrileri ile Eğilim (Trend) Analizi: Bir öğrencinin farklı zamanlardaki (örneğin; 3, 4 ve 5. sınıf bahar dönemi) puanları aynı grafikte işaretlenerek, referans grubuna göre göreli konumunun zaman içinde nasıl değiştiği (örneğin; P50'den P65'e yükselmesi) izlenebilir.
2. Koşullu Büyüme Grafikleri: Bu daha gelişmiş yaklaşım, öğrencinin ilerlemesini "akademik akranlarına" (benzer başlangıç noktasına sahip diğer öğrenciler) göre konumlandırır. Grafiklerde genellikle bir önceki yılın puanı yatay eksene, bu yılın puanı dikey eksene yerleştirilir. Diyagonal çizgiler ise büyüme yüzdeliklerini temsil eder. Öğrencinin (geçen yılki puan, bu yılki puan) koordinatıyla belirlenen noktası, bu grafikteki yerine göre yorumlanır. Bu yorum, başlangıç düzeyi düşük bir öğrencinin gösterdiği yüksek ilerlemeyi veya başlangıç düzeyi yüksek bir öğrencinin beklenen düzeyde ilerleme gösteremediğini adil bir referansla ortaya koyar (Betebenner, 2009).

3.3.3. Sık Görülen Hatalar ve Alınabilecek Önlemler

Standart belirleme sürecinde sık görülen hatalar ve bunlara yönelik alınabilecek önlemler aşağıda sunulmuştur:

- Hata: Test eşitleme yapılmamış farklı formların veya dönemlerin eğrilerini doğrudan karşılaştırmak.

  Önlem: Yorum yapmadan önce tüm puanların ortak bir ölçekte olduğundan emin olunmalıdır.

- Hata: Örneklem ağırlıklarını ve karmaşık tasarım etkisini yok sayarak standart hataları olduğundan küçük hesaplamak.

  Önlem: Varyans kestirimi, çoğaltma ağırlıkları gibi tasarıma uygun yöntemlerle yapılmalıdır.

- Hata: Yüzdelik (bir puan) ile yüzdelik sırayı (bir yüzde) karıştırmak.

  Önlem: Raporlarda bu iki kavramın tanımını içeren kısa bir bilgi kutusu sunulmalıdır.

- Hata: Bireysel kararlarda eğrinin etrafındaki belirsizlik bantlarını veya puanın güven aralığını göz ardı etmek.

  Önlem: Bireysel raporlarda puanla birlikte mutlaka SEM veya güven aralığı bilgisi verilmelidir.

- Hata: Güncellenmiş yeni normlarla üretilen bir eğriyi, eski normlarla üretilmiş bir eğriyle doğrudan kıyaslamak.

  Önlem: Norm geçişlerinde, iki norm seti arasındaki etkiyi gösteren köprüleme tabloları sunulmalıdır.

3.3.4. Normların Geçerlik Süresi: Raporlama, Güncelleme ve Etik Kullanım

Norm setleri statik belgeler değildir. Üretildikleri, kullanıldıkları, zamanla geçerliklerini yitirdikleri ve yenilendikleri bir süreçleri vardır. Bu sürecin her aşamasında şeffaflık ve etik sorumluluk esastır. Raporlama tasarımı ve hedef kitle norm bilgilerinin ve eğrilerinin sunumu, hedef kitleye göre uyarlanmalıdır.

- Öğretmen ve Aday Odaklı Raporlar: Genellikle P10, P50 ve P90 gibi temel referans çizgilerini, bireyin puanının işaretlendiği noktayı ve sonucun ne anlama geldiğini sade bir dille açıklayan kısa bir yorum kutucuğunu içeren tek sayfalık özetler tercih edilir.
- Politika ve Araştırma Odaklı Raporlar: Bu raporlarda yıllar arası eğilim eğrileri, ölçek tutarlılığına ilişkin teknik kanıtlar ve norm değişikliklerinin etkisini analiz eden tablolar birlikte sunulur. Eğer bir programda hem ulusal hem de yerel normlar mevcutsa, iki referans çizgisinin aynı grafikte farklı renk veya stillerle gösterilmesi ve kapsam farkının açıkça belirtilmesi, yorumun zenginleşmesini sağlar.

Uluslararası İyi Uygulamalar ve Şeffaflık: TIMSS, PIRLS, NAEP ve PISA gibi geniş ölçekli uluslararası değerlendirmeler, normların ve eğrilerin nasıl belgelenmesi gerektiğine dair iyi örnekler sunar. Bu programların teknik raporları, kullanılan örnekleme ve test eşitleme yöntemlerini, ağırlıklandırma prosedürlerini ve varyans kestirimi için kullanılan çoğaltma ağırlıklarını ayrıntılı olarak yayımlar. Bu şeffaflık, araştırmacıların ve politika yapıcıların sonuçları yeniden üretebilmesini veya ikincil analizler yapabilmesini sağlar. Kullanıcı, gördüğü bir çizginin hangi karmaşık yöntem zincirinin bir ürünü olduğunu doğrudan kaynaklardan doğrulayabildiğinde, programın kamu nezdindeki güvenirliği artar (Martin, Mullis & Hooper, 2020; NCES, 2022; OECD, 2023).

Güncelleme ve Sürdürülebilirlik: Hiçbir norm seti sonsuza dek geçerli değildir. Müfredattaki değişiklikler, teknolojik gelişmeler ve öğrenci nüfusunun demografik yapısındaki değişimler ("Flynn etkisi" gibi) zamanla normların eskimesine neden olur. Bu nedenle, ölçme programları normlarını genellikle 5 ila 10 yıl gibi periyotlarla planlı bir şekilde güncellemelidir. Normlar yenilendiğinde, eski ve yeni normlar arasındaki farkın sınıflama oranları üzerindeki etkisi analiz edilmeli ve kullanıcılara bu geçişin pratik sonuçları hakkında net bilgi verilmelidir.

Tamamlayıcı bir yorumlama aracı norm eğrileri, teknik titizlikle üretilip doğru yorumlandığında, geniş ölçekli test sonuçlarını paydaşlar için güçlü bir referans aracına çevirir. Bir öğrencinin mutlak bir standardı (ölçüt-dayanaklı yorum) karşılayıp karşılamadığını gösteren bilginin yanına, referans grubuna göre göreli konumunu ve

gelişim seyrini (norm-dayanaklı yorum) ekleyerek bütüncül bir tablo sunarlar. Eğriler, ölçütlerin yerini almaz; onları tamamlarlar. Bu ikili bakış açısı, eğitimde veriye dayalı karar verme süreçlerinin hem daha adil hem de daha bilgilendirici olmasını sağlar.

3.4. Örneklem Seçimi ve Norm Grubu

Önceki bölümlerde normların, bir bireyin performansını tanımlı bir referans grubuna göre anlamlandırmamızı sağlayan güçlü bir araç olduğunu gördük. Ancak bu karşılaştırmanın geçerliği ve adaleti, referans noktası olarak kullanılan norm grubunun ne kadar bilimsel ve titiz bir süreçle oluşturulduğuna bağlıdır. İşte bu süreç, örneklem seçimi ve norm grubu oluşturma aşamasıdır. Norm grubu, bir testin sonuçlarının genelleneceği hedef evreni (örneğin; ülkedeki tüm 8. sınıf öğrencileri) temsil eden ve normların üretildiği örneklemdir ve bu grubun kalitesi, norm tabanlı tüm yorumların temelini oluşturur. Bu alt bölüm, normların doğruluğunu ve adaletini belirleyen bu temel sürecin nasıl yürütüldüğünü adım adım ele almaktadır. Temsil gücü yüksek bir örneklem tasarımı, gerekli istatistiksel düzeltmeler ve şeffaf bir raporlama zinciri olmadan, en gelişmiş norm tabloları veya eğrileri bile yanıltıcı sonuçlar üretebilir.

3.4.1. Örnekleme Tasarımının Temelleri

Güvenilir bir norm grubu oluşturma süreci üç temel adımla başlar:

1. Hedef Evrenin Tanımlanması: İlk adım, normların genelleneceği evrenin sınırlarını net bir şekilde çizmektir. Hangi yaş aralığı, sınıf düzeyi, coğrafi bölge ya da kurum türü kapsamda olacaksa bu sınır tartışmaya yer bırakmayacak şekilde yazılacaktır.
2. Örnekleme Çerçevesinin Seçimi: İkinci adım, tanımlanan hedef evrendeki örnekleme birimlerine (örneğin; okullar, öğrenciler) ulaşılabilen güncel ve eksiksiz bir liste olan örnekleme çerçevesini belirlemektir. Çerçevenin hedef evrenin bir kısmını kapsamadığı durumlar (örneğin; listeye dâhil olmayan okul türleri) kapsam hatası (coverage error) yaratacaktır. Bu hatanın boyutu ve olası etkileri teknik raporda şeffafça belirtilmelidir.
3. Örnekleme Tasarımının Belirlenmesi: Üçüncü ve en kritik adım, evrenden norm grubunu seçecek yöntemin tasarlanmasıdır. Geniş ölçekli testlerde lojistik kolaylık ve maliyet etkinliği nedeniyle standart uygulama, tabakalı ile çok aşamalı örneklemenin birleşimidir. Genellikle okullar birincil örnekleme birimi olarak seçilirken, öğrenciler ikinci aşamada bu okulların içinden genellikle basit rastgele yöntemle örneklenir.

Unutulmamalıdır ki, bir norm çalışmasının evreni temsil etme gücü, en başta yapılan bu tasarım ve tanım kararlarının kalitesine bağlıdır. Zayıf bir çerçeve veya hatalı bir tasarımla yola çıkıldığında, hiçbir ileri istatistiksel yöntem bu başlangıçtaki yanlılığı

sonradan tam olarak düzeltemez. Sağlam bir çerçeve, dikkatli bir tabakalama ve her birimin bilinen ve sıfırdan farklı bir seçilme olasılığına sahip olması, normların güvenirliğinin temelini oluşturur (AERA, APA & NCME, 2014; Kish, 1965; Lohr, 2019).

Tabakalı örnekleme, evreni coğrafya, okul türü veya sosyoekonomik düzey gibi politika açısından anlamlı alt gruplara (tabakalara) ayırarak, bu grupların örneklemde orantılı veya hedeflenen sayıda temsilini sağlar. Gerekli durumlarda, analizler için yeterli sayıda veri elde etmek amacıyla küçük ama önemli alt grupların örneklemdeki oranı artırılabilir (aşırı örnekleme - oversampling) yapılabilir. Çok aşamalı örnekleme düzeninde ise okullar genellikle öğrenci sayıları gibi büyüklükleriyle orantılı olasılıkla seçilir ve içlerinden öğrenciler belirlenen yöntemle (örneğin; basit rastgele) basit rastgele yöntemle belirlenir. Bu yapısal yaklaşım, saha uygulamalarını büyük ölçüde kolaylaştırır (Heeringa ve ark., 2017).

3.4.2. Ağırlıklandırma ve Hata Kestirimi

Örneklem seçildikten sonra bile, farklı seçilme olasılıkları veya yanıtlama oranlarındaki farklılıklar nedeniyle örneklemin evreni birebir yansıtması nadiren mümkündür. Bu noktada, istatistiksel düzeltme mekanizmaları devreye girer.

Örnekleme Ağırlıkları (Sampling Weights) Ağırlıklar, örneklemdeki her birimin evrendeki temsil payını ayarlayarak, örneklemden elde edilen sonuçların evrene doğru bir şekilde genellenmesini sağlar. Ağırlıklandırma süreci genellikle üç basamakta kurulur:

1. Taban Ağırlığı: Her birimin örnekleme dahil edilme olasılığının matematiksel tersidir. Farklı olasılıklarla seçilen birimlerin etkisini eşitler.
2. Yanıtlamama Düzeltmesi: Seçilen ancak araştırmaya katılmayan okul veya öğrencilerin yarattığı sistematik hatayı azaltmayı hedefler. Katılmayanların özelliklerine benzer katılımcıların ağırlığı artırılır.
3. Son Hizalama: Örneklemin demografik dağılımını, cinsiyet, bölge veya okul türü gibi bilinen evren toplamlarına yakınlaştırmak için yapılan son ayardır.

Ağırlıklandırma, yanlılığı azaltsa da yapılan istatistiksel kestirimlerin varyansını genellikle artırır. Bu artış, tasarım etkisi adı verilen bir katsayı ile özetlenir. Tasarım etkisi, karmaşık örneklem tasarımından (örneğin; tabakalı, aşamalı) ve ağırlıklandırmadan kaynaklanan varyans artışını, aynı büyüklükteki basit rastgele örneklemeye kıyasla gösterir. Tasarım etkisi değeri olan DEFF'in 1'den büyük olması, etkin örneklem büyüklüğünün (hesaplamalarda kullanılan istatistiksel gücü yansıtan sanal büyüklük) nominal (gerçek) büyüklükten daha düşük olmasına neden olur.

Varyans Kestirimi (Variance Estimation) Karmaşık örneklem tasarımlarında standart hata ve güven aralıkları, basit rastgele örnekleme varsayımına dayanan standart istatistik formülleriyle hesaplanamaz. Bu, hatanın olduğundan küçük tahmin edilmesine

yol açar. Doğru varyans kestirimi için iki yaygın yaklaşım kullanılır: Taylor serisi doğrusallaştırma (varyans formülünü karmaşık tasarıma uyarlayan analitik bir yöntem) veya çoğaltma ağırlıkları (*replicate weights*). Jackknife ve Dengeli Tekrarlı Çoğaltma (BRR - Balanced Repeated Replication)gibi çoğaltma yöntemleri, ana örneklemden çok sayıda alt örneklem (çoğaltma) yaratarak ilgilendiğimiz istatistiğin (örneğin; ortalama, yüzdelik) bu alt örneklemlerdeki dağılımını inceler ve bu dağılımdan doğru standart hatayı hesaplar. . PISA ve NAEP gibi büyük programlar, araştırmacıların doğru hata kestirimleri yapabilmesi için kamuya açık veri setleriyle birlikte bu çoğaltma ağırlıklarını da sunar (Rust & Rao, 1996; Lohr, 2019).

3.4.3. Normların Geçerliği, Adaleti ve Sürekliliği

Norm grubu oluşturma süreci, normlardan elde edilecek yorumların geçerlik ve adalet argümanının önemli bir parçasıdır.

Geçerlik Tehditleri: Kapsama ve yanıtlamama yanlılığı normların genellenebilirliğini tehdit eden en kritik risklerinden ikisi kapsama (coverage) ve yanıtlamama (nonresponse) yanlılığıdır. Örnekleme çerçevesi belirli okul türlerini dışarıda bırakıyorsa (kapsama hatası) veya belirli özelliklere sahip okullar ya da öğrenciler sistematik olarak katılmayı reddediyorsa (yanıtlamama yanlılığı), ağırlıklandırma bu sorunları tek başına tamamen çözemeyebilir. Bu nedenle, teknik raporlarda mutlaka bir yanıtlamama yanlılığı analizi sunulmalı ve olası kalıntı yanlılıklar (düzeltmelerden sonra hala var olabilecek sistematik hatalar) dürüstçe tartışılmalıdır (Little & Rubin, 2019).

Örnekleme ve Ölçme Süreçlerinin Bütünlüğü: Örnekleme kararları, ölçme süreçlerinden bağımsız düşünülemez. Örneğin, farklı uygulama zaman aralığı (sonbahar vs. bahar) öğrenme düzeylerindeki farklılıklar nedeniyle mevsimsellik etkisi yaratacağından, bu dönemler için ayrı normlar üretilmelidir. Farklı test formları kullanıldıysa, karşılaştırılabilirlik için önce mutlaka test eşitleme yapılmalıdır. Bu adımların bütüncül bir şekilde yönetilmesi, normların sadece göreli konum bilgisi değil, aynı zamanda zaman içinde karşılaştırılabilir ve savunulabilir bir referans noktası olmasını sağlar (Kolen & Brennan, 2014).

Adalet: Alt grup normları ve ölçme değişmezliği farklı demografik alt gruplar (örneğin; anadil durumu, cinsiyet, sosyoekonomik düzey) için ayrı normlar üretilecekse veya gruplar arası karşılaştırmalar yapılacaksa, bu karşılaştırmanın adil olması gerekir. Bunun ön koşulu, kullanılan ölçme aracının testin bu gruplarda aynı psikometrik yapıyı ölçtüğünün, yani ölçme değişmezliğinin (measurement invariance) kanıtlanmasıdır. Bu kanıt sunulmadan alt grupların normlarını doğrudan karşılaştırmakyanıltıcı sonuçlara yol açabilir (Dorans & Cook, 2016).

Normların Sürekliliği: Normlar statik değildir ve zamanla geçerliklerini yitirebilirler. Değişen müfredat, öğretim yöntemleri ve öğrenci profilleri nedeniyle normlar periyodik olarak (genellikle 5-10 yılda bir) güncellenmelidir. Yeni norm setine geçildiğinde, eski

set ile arasındaki farkı açıklayan köprüleme (bridging) çalışmaları yapılmalı ve bu değişimin yeterlik düzeylerinin sınıflama oranları üzerindeki etkisi kullanıcılara şeffaf bir şekilde raporlanmalıdır.

PISA, TIMSS ve NAEP gibi uluslararası eğitim izleme programlar, yukarıda anlatılan tüm süreçlerin (tabakalı çok aşamalı tasarım, ağırlıklandırma, çoğaltma ağırlıkları ile hata kestirimi, şeffaf teknik raporlama) en iyi uygulama örneklerini sunar. Bu programların teknik raporları ve belgeleri, norm tablolarındaki her sayının arkasında yatan bilimsel titizliği ve metodolojik zinciri izlenebilir ve doğrulanabilir kılar (OECD, 2023; Martin ve ark., 2020; NCES, 2022). Sonuç olarak, norm grubu ve örnekleme tasarımı, normların doğruluğunu ve adaletini belirleyen omurgadır. Bu çerçeve şeffaf bir şekilde belgelenip titizlikle uygulandığında, normlar yalnızca birer karşılaştırma aracı olmaktan çıkar ve eğitimde iyileştirmeyi hedefleyen açık, güvenilir ve savunulabilir bir iletişim zeminine dönüşür.

4. STANDART BELİRLEME YÖNTEMLERİ

Önceki bölümlerde, bir testten elde edilen puanların yeterlik ve norm çerçeveleriyle nasıl yorumlanabilir hâle geldiği incelenmiştir. Özellikle ölçüt-dayanaklı yorumlamanın merkezinde yer alan kesme puanları, soyut olan ham puanları "yeterli/yetersiz", "temel/orta/ileri düzey" veya "sertifika almaya hak kazandı/kazanamadı" gibi somut ve eyleme geçirilebilir kararlara dönüştürür. İşte bu kritik eşiklerin belirlendiği sürece standart belirleme (standard setting) adı verilir.

Standart belirleme, yalnızca istatistiksel bir işlem değil, aynı zamanda alan uzmanlarının yargılarını, kanıtları ve değerleri bir araya getiren sistematik bir süreçtir. Sürecin temel amacı, bir yeterlik düzeyini tanımlayan performans standardını, puan ölçeği üzerinde savunulabilir bir noktayla eşleştirmektir. Standart belirleme yöntemleri genel olarak iki gruba ayrılır:

1. Test-Merkezli (Test-Centered) Yaklaşımlar: Uzmanların (panelist) dikkatini doğrudan test maddelerine veya yeterlik düzeyi betimleyicilerine (PLD) yöneltir. Panelistler, belirli bir yeterlik düzeyinin sınırında yer alan hayali bir adayın (sınırda yetkin aday / borderline candidate) bu maddelerde nasıl bir performans göstereceğini yargılarlar. Bu alt bölümde ele alınan Angoff, Ebel, Nedelsky, Bookmark gibi yöntemler bu gruba aittir.
2. Aday-Merkezli (Examinee-Centered) Yaklaşımlar: Uzman yargılarının, test maddeleri yerine gerçek aday performanslarına, sınav kâğıtları, portfolyolar veya aday gruplarına ait puan dağılımları gibi doğrudan performans kanıtlarına dayandırıldığı yaklaşımlardır.

Bu alt bölümde, alanyazında en yaygın kullanılan test-merkezli yöntemler, tarihsel kökenleri, işlem adımları, üstün ve sınırlı yönleriyle birlikte ayrıntılı olarak incelenmiştir. Aday-merkezli yaklaşımlar ise gerçek aday performanslarının, ürün dosyalarının veya puan dağılımlarının ayrı bir kanıt temeli olarak incelenmesini gerektirdiğinden, bu bölümün test maddesi temelli odağı dışında tutulmuştur. Bu yaklaşımlar için okuyucular sınır grup, karşıt gruplar ve ürün dosyası temelli standart belirleme yöntemlerine başvurabilir. Unutulmamalıdır ki, hangi yöntem seçilirse seçilsin, sürecin kalitesini ve sonuçların savunulabilirliğini belirleyen en önemli unsurlar panelistlerin dikkatli seçimi ve eğitimi, geri bildirim turlarının yapısı ve kararların olası sonuçlarını gösteren etki verisinin şeffaf bir şekilde kullanılmasıdır (Cizek & Bunch, 2007; Cizek & Earnest, 2016).

4.1. Angoff Yöntemi

Klasik standart belirleme literatürünün temel taşı olan bu yöntem, William Angoff'un (1971) öncü çalışmasına dayanır. Yöntem, "Sınırda yetkin bir aday, bu maddeyi hangi olasılıkla doğru yanıtlar?" sorusu üzerine kuruludur. Bu yöntem, özellikle yeterliklerin önemli olduğu (tıp alanı gibi) mesleki lisanslama ve sertifikasyon sınavlarında yaygın olarak kullanılmaktadır.

İşlem Adımları:

1. Panelist Seçimi ve Eğitimi: Konu alanında uzman ve farklı bakış açılarına sahip bir grup panelist seçilir. Panele, standart belirlemenin amacı, yöntemin adımları ve en önemlisi, her bir yeterlik düzeyi için sınırda yetkin aday profilinin ne anlama geldiği konusunda ayrıntılı bir eğitim verilir.
2. 1. Tur (Bireysel Değerlendirme): Panelistler, testi madde madde inceler. Her bir madde için, sınırda yetkin adayın o maddeyi doğru yanıtlama olasılığını 0 ile 1 arasında bir değer olarak (örneğin; 0,65) bireysel olarak tahmin ederler.
3. Geri Bildirim ve Tartışma: İlk tur tahminleri toplandıktan sonra, moderatör panelistlerin tahminlerinin ortalamasını ve dağılımını grupla paylaşır. Yargıları birbirinden çok farklı olan panelistlerden kararlarının gerekçelerini açıklamaları istenir. Bu turda, ilgili maddeye ait istatistikler (örneğin; madde güçlük indeksi - p-değeri) veya örnek öğrenci performansları gibi ek kanıtlar sunulabilir.
4. Sonraki Turlar ve Yakınsama: Tartışma ve geri bildirimlerin ışığında panelistler, tahminlerini yeniden gözden geçirirler. Bu süreç, panelistlerin yargıları arasında makul bir fikir birliği (yakınsama) oluşana kadar genellikle 2-3 tur devam eder.
5. Kesme Puanının Hesaplanması: Her panelistin son turdaki olasılık tahminleri, testteki tüm maddeler için toplanarak o paneliste ait kesme puanı adayı bulunur. Panelin nihai kesme puanı ise genellikle tüm panelistlerin kesme puanlarının ortalaması veya medyanı alınarak belirlenir.

Üstün ve Sınırlı Yönleri

- Üstünlükler: İçerik uzmanlarının madde düzeyinde derinlemesine düşünmesini sağlar. Süreç şeffaftır ve farklı madde türlerini içeren (çoktan seçmeli, açık uçlu vb.) testlere kolayca uygulanabilir.
- Sınırlılıklar: Bilişsel yükü yüksek olabilir. Panelistlerin, özellikle çok zor veya çok kolay maddeler için olasılık tahminlerinde sistematik olarak sapma eğilimi gösterebildiği raporlanmıştır. Bu riskler, iyi bir eğitim ve tur bazlı geri bildirimlerle yönetilebilir. Araştırmalar, sürecin kalitesinin (eğitim, geri bildirim vb.) farklı yöntemleri karşılaştırmaktan daha önemli olduğunu göstermektedir (Brandon, 2004; Reckase, 2006).

4.2. Modifiye Angoff Yöntemi

Modifiye Angoff, klasik Angoff yönteminin temel mantığını korurken, süreci daha fazla kanıta dayalı hâle getirmek ve panelist yargılarındaki değişkenliği azaltmak için geri bildirim mekanizmalarını sistematikleştirir. Özellikle karmaşık performans görevleri içeren mesleki sınavlarda ve eğitim alanındaki geniş ölçekli değerlendirmelerde standart uygulama haline gelmiştir.

İşlem Adımları ve Farklılıklar: Modifiye Angoff, klasik sürece ek olarak aşağıdaki geri bildirim türlerini yapılandırılmış bir şekilde dâhil eder:

- Normatif Geri Bildirim: Her tur sonunda, bir panelistin kendi tahminlerinin diğer panelistlerin tahminlerine göre nerede durduğunu görmesi sağlanır.
- Empirik/Gerçeklik Geri Bildirimi (Reality Check): Panelistlere, değerlendirdikleri maddelerin gerçek aday gruplarındaki performans verileri (örneğin; madde güçlük indeksleri veya Madde Tepki Kuramı (MTK/IRT) güçlük parametreleri) sunulur. Bu, uzman yargısının gerçek veriyle hizalanmasına yardımcı olur.
- Etki Verisi Geri Bildirimi (Impact Data): Belirlenen geçici kesme puanlarının, toplamda ve önemli alt gruplarda (örneğin; cinsiyet, okul türü) ne kadar adayın "yeterli" veya "yetersiz" olarak sınıflandırılacağını gösteren veriler paylaşılır. Bu, kararın pratik sonuçları hakkında paneli bilgilendirir.

Üstün ve Sınırlı Yönleri

- Üstünlükler: Kanıta dayalı geri bildirimler, panelist yargılarının daha tutarlı ve savunulabilir olmasını sağlar, sapmaları azaltır.
- Sınırlılıklar: Aşırı veya kötü yönetilmiş bir geri bildirim süreci, panelistlerin bağımsız yargılarını bırakıp "grup düşüncesine" (groupthink) kapılmasına neden olabilir. Bu riski yönetmek için moderatörün rolü ve sürecin protokolü kritik önem taşır (Hambleton & Plake, 1995; Hambleton ve ark., 2012).

4.3. Ebel Yöntemi

Ebel yöntemi, maddeleri tek bir zorluk boyutu yerine iki boyutlu bir çerçevede değerlendirerek karar sürecine içerik geçerliği vurgusu katar. İçerik temsilinin ve program hedefleriyle uyumun ön planda olduğu sınavlarda, özellikle tıp ve sağlık alanlarındaki sınavlarda kullanılır.

İşlem Adımları:

1. Izgara Oluşturma: Panelistler, maddeleri iki temel boyuta göre bir ızgaraya yerleştirirler: Önem/Uygunluk (örneğin; Çok Önemli, Önemli, Az Önemli) ve Güçlük (örneğin, Kolay, Orta, Zor). Bu genellikle 3x3'lük dokuz hücreli bir tablo oluşturur.
2. Hücrelere Oran Atama: Panelistler, her bir hücre için (örneğin, "Çok Önemli" ve "Orta Güçlükte" olan maddeler hücresi), sınırda yetkin bir adayın bu tür maddelerin yüzde kaçını doğru yanıtlamasını beklediklerini tartışarak belirlerler.
3. Kesme Puanının Hesaplanması: Testteki her bir maddenin hangi hücreye ait olduğu belirlenir ve o hücre için atanan beklenen doğru yanıt oranıyla eşleştirilir. Test genelindeki tüm maddeler için bu oranlar bir araya getirilerek nihai kesme puanı hesaplanır.

Üstün ve Sınırlı Yönleri

- Üstünlükler: "Önem" boyutunu karar sürecine doğrudan dahil ederek, testin içerik geçerliği argümanını güçlendirir.
- Sınırlılıklar: Izgaradaki boyutların (özellikle "güçlük") tanımları soyut kalırsa, panelistlerin maddeleri hücrelere yerleştirirken tutarsızlık yaşaması en büyük risktir. Bu durum, yöntemin güvenirliğini düşürebilir (Ebel, 1972; Park, 2018).

4.4. Evet/Hayır Yöntemi

Bu yöntem, klasik Angoff'un bilişsel yükünü azaltmak için geliştirilmiş, kararı basitleştiren bir varyanttır. Karar sürecini hızlandırması nedeniyle, özellikle madde sayısının çok olduğu sınavlarda ve zamanın kısıtlı olduğu durumlarda (örneğin tıp eğitiminde) tercih edilir (Impara & Plake, 1997).

İşlem Adımları: Panelistlerden, her bir madde için "Sınırda yetkin bir aday bu maddeyi doğru yanıtlayabilir mi?" sorusuna olasılık tahmini yapmak yerine, sadece "Evet" veya "Hayır" şeklinde ikili bir yargıda bulunmaları istenir. Kesme puanı, panelistlerin "Evet" olarak işaretlediği maddelerin toplamı veya ortalaması alınarak belirlenir. Geri bildirim ve tartışma turları, Modifiye Angoff'a benzer şekilde yürütülebilir.

Üstün ve Sınırlı Yönleri

- Üstünlükler: Bilişsel yükü düşüktür ve panelistler için daha kolay ve hızlı bir karar süreci sunar.
- Sınırlılıklar: İkili karar yapısı, olasılık tahminindeki nüansları ortadan kaldırır. Özellikle çok kolay veya çok zor maddelerde panelistlerin yargılarının sistematik olarak sapma riski bulunur. Bazı uygulamalarda bu riski azaltmak için "Evet/Hayır/Emin Değilim" gibi üç kategorili bir yapı kullanılır (Yudkowsky ve ark., 2008).

4.5. Nedelsky Yöntemi

Bu yöntem, özel olarak çoktan seçmeli maddeler için tasarlanmıştır ve mantığını seçeneklerin elenmesi üzerine kurar. Yöntemin kullanımına yönelik eğilim azalmış olsa da, çeldirici kalitesinin yüksek olduğu standart testlerde hâlâ uygulanabilmektedir.

İşlem Adımları: Panelistlerden, her bir çoktan seçmeli madde için, "Sınırda yetkin bir aday, hangi seçeneklerin kesinlikle yanlış olduğunu bilir ve onları eler?" sorusunu yanıtlamaları istenir. Bir adayın eleyebileceği çeldiriciler işaretlenir. Geriye kalan seçenek sayısı (k), o maddeden şansla doğru yanıta ulaşma olasılığını (1/k) verir. Örneğin, beş seçenekli bir maddede panelist sınırda adayın üç çeldiriciyi eleyebileceğine karar verirse, geriye iki seçenek kalır ve bu maddeden beklenen doğru yanıt olasılığı 1/2 (0,50) olur. Testteki tüm maddeler için bu olasılıklar toplanarak kesme puanı hesaplanır (Nedelsky, 1954).

Üstün ve Sınırlı Yönleri

- Üstünlükler: Karar sürecini, seçeneklerin ve çeldiricilerin kalitesi gibi somut bir unsura bağlar.
- Sınırlılıklar: Yalnızca çoktan seçmeli maddelere uygulanabilir. Çeldiricilerin kalitesi düşük veya işlevsiz ise, yöntem aşırı düşük ve gerçekçi olmayan kesme puanları üretir.

4.6. Bookmark (Yer İmi) Yöntemi

Madde Tepki Kuramı'ndan (MTK/IRT) beslenen Bookmark, günümüzde en yaygın kullanılan standart belirleme yöntemlerinden biridir. ABD'deki eyalet düzeyindeki K-12 değerlendirme programlarında, üniversite yerleştirme sınavlarında (örneğin ACCUPLACER) ve diğer geniş ölçekli testlerde standart uygulama hâline gelmiştir (Mitzel ve ark., 2001).

İşlem Adımları:

1. Sıralı Madde Kitapçığı (Ordered Item Booklet - OIB): Testteki maddeler, önceden uygulanarak elde edilen IRT zorluk parametrelerine (b-değerleri) göre, en kolaydan en zora doğru sıralanarak özel bir kitapçık oluşturulur.

2. Yer İmi Koyma: Panelistler bu kitapçığı baştan sona incelerler. Her panelist, "Sınırda yetkin bir adayın, bu maddedeyi ve bundan önceki (daha kolay) maddeleri tutarlı bir şekilde doğru yanıtlama olasılığının belirli bir önceden tanımlanmış olasılık eşiğin (örneğin, 2/3 yani yaklaşık %67 veya %80) ulaştığı veya aştığı madde hangisidir?" sorusuna yanıt arar. Bu kritik maddeye fiziksel olarak veya dijital ortamda bir "yer imi" (bookmark) koyarlar.
3. Kesme Puanının Hesaplanması: Her panelistin yer imi koyduğu maddenin IRT zorluk değeri, o panelistin kesme puanı adayıdır (yetenek (θ) ölçeğinde). Panelin nihai kesme puanı, bu yetenek değerlerinin ortalaması alınarak bulunur ve daha sonra testin ham puan veya raporlama ölçeğine dönüştürülür.

Üstün ve Sınırlı Yönleri

- Üstünlükler: MTK'ya dayalı olduğu için farklı test formları arasında tutarlılığı sağlamak daha kolaydır. Süreç, yeterlik düzeyi tanımlarıyla (PLD'ler) güçlü bir bağ kurar. Panelistlerin bilişsel yükü, madde madde olasılık tahmin etmekten daha düşüktür.
- Sınırlılıklar: Yöntem, maddelerin önceden uygulanıp MTK ile kalibre edilmesini gerektirir. Sıralı madde kitapçığının kalitesi ve MTK parametrelerinin doğruluğu, sonucun geçerliği için kritiktir.

4.7. Basket (Sepet) Yöntemi

Bu yöntem, özellikle Diller için Ortak Avrupa Başvuru Metni (D-OABM / CEFR) gibi önceden tanımlanmış ve kabul görmüş yeterlik düzeylerine sahip bağlamlarda kullanılır.

İşlem Adımları: Panelistlere, her bir maddeyi, karşılık geldiği yeterlik düzeyini temsil eden kavramsal "sepetlere (baskets)" yerleştirmeleri görevi verilir. Örneğin, bir maddeyi inceler ve "Bu maddeyi doğru yanıtlamak, D-OABM'ye göre B1 düzeyi bir beceri mi gerektirir, yoksa B2 düzeyi mi?" diye karar vererek maddeyi uygun sepete atarlar. Tüm maddeler sepetlere yerleştirildikten sonra, her bir düzey için kesme puanı, o düzeydeki ve altındaki sepetlerde yer alan maddelerin sayısı veya ağırlıklandırılmasıyla elde edilen kurallara göre hesaplanır.

Üstün ve Sınırlı Yönleri

- Üstünlükler: Süreç, doğrudan yeterlik çerçevesinin diliyle yürütüldüğü için sonuçlar paydaşlar için çok anlaşılırdır.
- Sınırlılıklar: Yöntemin başarısı, panelistlerin referans alınan yeterlik çerçevesine (örneğin D-OABM) derinlemesine hakim olmalarına bağlıdır. Sepetleme yargılarındaki tutarlılığı sağlamak için yoğun bir eğitim ve kalibrasyon süreci gerektirir (Kaftandjieva, 2009).

5. SONUÇ

Yeterliklerin ve normların belirlenmesi, ölçme sonuçlarının karar süreçlerinde doğru ve tutarlı biçimde kullanılabilmesi için temel bir işleve sahiptir. Bu bölümde ele alınan kavramlar, ham puanın tek başına yeterli bir değerlendirme bilgisi sunmadığını ve puanın mutlaka bir anlam çerçevesi içinde yorumlanması gerektiğini göstermektedir. Uygulamada bu durum, her ölçme programının daha test geliştirme aşamasında hangi yeterlik düzeylerini raporlayacağını, hangi norm grubunu kullanacağını, kesme puanlarının nasıl belirleneceğini ve ölçme hatasının raporlara nasıl yansıtılacağını planlamasını gerektirir. Böylece puanlar yalnızca sayısal sonuçlar olarak değil, bireyin hangi düzeyde performans gösterdiğini ve bu performansın hangi referans içinde yorumlanacağını açıklayan değerlendirme bilgisine dönüşür.

Yeterlik düzeylerinin kullanıldığı uygulamalarda en önemli pratik gereklilik, düzey etiketlerinin açık performans düzeyi betimleyicileriyle desteklenmesidir. "Temel", "Yeterli" ve "İleri" gibi düzeyler, yalnızca puan aralıklarıyla tanımlandığında sınırlı bir açıklama gücüne sahip olur. Bu düzeylerin hangi bilgi, beceri ve davranış göstergelerine karşılık geldiği açık biçimde yazıldığında ise raporlar adaylar, öğretmenler, uzmanlar ve karar vericiler için daha anlaşılır hâle gelir. Bu nedenle sınav ve değerlendirme raporlarında yalnızca "hangi puan alındı?" sorusuna değil, "bu puan ne anlama gelir?" ve "bu düzeydeki birey tipik olarak ne yapabilir?" sorularına da yanıt verilmelidir.

Normlara dayalı yorumlarda ise öncelikli pratik çıkarım, norm grubunun niteliğinin raporlanan sonucun niteliğini doğrudan belirlediğidir. Yaş normu, sınıf düzeyi normu, büyüme normu, yerel norm ve koşullu norm farklı kullanım amaçlarına sahiptir. Bu nedenle raporlarda normun hangi evrene dayandığı, örneklemin bu evreni nasıl temsil ettiği, ağırlıklandırma işlemlerinin uygulanıp uygulanmadığı ve standart hata değerlerinin nasıl hesaplandığı açıkça belirtilmelidir. Yüzdelik, yüzdelik sıra, standart puan ve norm eğrileri kullanıcıya sade bir karşılaştırma imkânı sunar. Ancak bu göstergeler, özellikle bireysel kararlar verilirken ölçmenin standart hatası ve güven aralığı ile birlikte değerlendirilmelidir. Kesme puanı çevresindeki puanlarda bu ilke daha da önemlidir.

Standart belirleme yöntemleri açısından bölümün temel sonucu, kesme puanlarının yalnızca teknik bir hesaplama sonucunda değil, sistematik uzman yargısı ve kanıt temelli bir süreç sonunda belirlenmesi gerektiğidir. Angoff, Modifiye Angoff, Ebel, Evet/Hayır, Nedelsky, Bookmark ve Basket yöntemleri farklı test yapıları için kullanılabilir. Ancak yöntemin uygunluğu kadar panelistlerin seçimi, panelist eğitimi, sınırda yetkin aday tanımının açıklığı, geri bildirim turlarının düzeni ve etki verisinin değerlendirilmesi de kararların savunulabilirliği açısından belirleyicidir. Sonuç olarak iyi yapılandırılmış bir değerlendirme sistemi, puanı üretmekle yetinmeyen, puanın anlamını, hata payını, yeterlik düzeyini, norm içindeki konumunu ve karar gerekçesini birlikte sunan bir sistem olmalıdır.

KAYNAKÇA | REFERENCES

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

Angoff, W. H. (1971). Scales, norms, and equivalent scores. R. L. Thorndike (Ed.), *Educational measurement* (2. Baskı, s. 508–600) içinde. American Council on Education.

Betebenner, D. W. (2009). Norm- and criterion-referenced student growth. *Educational Measurement: Issues and Practice, 28*(4), 42–51. https://doi.org/10.1111/j.1745-3992.2009.00161.x

Bourque, J., Pomedli, A., Charlebois, M., Milette, K. & St-Onge, C. (2020). Performance of the Ebel standard-setting method for a high-stakes exam. *BMC Medical Education, 20*, Article 264. https://doi.org/10.3352/jeehp.2020.17.12

Brandon, P. R. (2004). Conclusions about frequently studied modified Angoff topics. *Applied Measurement in Education, 17*(1), 59–88. https://doi.org/10.1207/s15324818ame1701_4

Cizek, G. J. & Bunch, M. B. (2007). *Standard setting: A practitioner's guide to establishing and evaluating performance standards on tests*. SAGE.

Cizek, G. J. & Earnest, D. S. (2016). Setting performance standards on tests. S. Lane, M. R. Raymond & T. M. Haladyna (Ed.), *Handbook of test development* (2. Baskı, s. 212–237). Routledge.

Cole, T. J. & Green, P. J. (1992). Smoothing reference centile curves: The LMS method and penalized likelihood. *Statistics in Medicine, 11*(10), 1305–1319. https://doi.org/10.1002/sim.4780111005

Council of Europe. (2020a). *Common european framework of reference for languages: Learning, teaching, assessment (Companion volume)*. Council of Europe Publishing.

Council of Europe. (2020b). *Common European framework of reference for languages: Learning, teaching, assessment (Companion volume) [Diller İçin Avrupa Ortak Başvuru Metni: Öğrenme, Öğretme ve Değerlendirme (Tamamlayıcı Cilt)]* (Türkçeye çeviri: Millî Eğitim Bakanlığı, 2022). https://ttkb.meb.gov.tr/meb_iys_dosyalar/2022_01/04144518_cefr_tr.pdf

Council of Europe. (2020). *Common European Framework of Reference for Languages: Learning, teaching, assessment—Companion volume*. Council of Europe Publishing. https://rm.coe.int/common-european-framework-of-reference-for-languages-learning-teaching/16809ea0d4

Dorans, N. J. & Cook, L. L. (Ed.). (2016). *Fairness in educational assessment and measurement*. Routledge.

Ebel, R. L. (1972). *Essentials of educational measurement*. Prentice Hall.

Egan, K., Schneider, J. & Ferrara, S. (2012). Performance level descriptors: History, practice, and implications. G. J. Cizek (Ed.), *Setting performance standards: Foundations, methods, and innovations* (s. 279–300) içinde. Routledge.

Ferrara, S., Johnson, E. & Chen, W. L. (2005). Vertically articulated performance standards: Logic, procedures, and likely classification accuracy. *Applied Measurement in Education, 18*(1), 35–60. https://doi.org/10.1207/s15324818ame1801_3

Hambleton, R. K. & Plake, B. S. (1995). Using an extended Angoff procedure to set standards on complex performance assessments. *Applied Measurement in Education, 8*(1), 41–56. https://doi.org/10.1207/s15324818ame0801_4

Hambleton, R. K., Pitoniak, M. J. & Copella, J. M. (2012). Essential steps in setting performance standards on educational tests and strategies for assessing the reliability of results. G. J. Cizek (Ed.), *Setting performance standards: Foundations, methods, and innovations* (s. 47–76) içinde. Routledge.

Heeringa, S. G., West, B. T. & Berglund, P. A. (2017). *Applied survey data analysis* (2. baskı). CRC Press.

Impara, J. C. & Plake, B. S. (1997). Standard setting: An alternative approach. *Journal of Educational Measurement, 34*(4), 353–366. https://doi.org/10.1111/j.1745-3984.1997.tb00523.x

Kaftandjieva, F. (2009). *Methods for setting cut scores in criterion-referenced achievement tests: A comparative study*. EALTA Resources. https://ealta.eu/documents/resources/Research_Colloquium_report.pdf

Kane, M. T. (2013). *Validating the interpretations and uses of test scores*. Routledge.

Kish, L. (1965). *Survey sampling*. Wiley.

Koenker, R. (2005). *Quantile regression*. Cambridge University Press.

Kolen, M. J. & Brennan, R. L. (2014). *Test equating, scaling, and linking: Methods and practices* (3. Baskı). Springer.

Lane, S., Raymond, M. R. & Haladyna, T. M. (Ed.). (2016). *Handbook of test development* (2. Baskı). Routledge.

Lewis, D. M. & Haug, C. A. (2005). Aligning policy and methodology to achieve consistent across-grade performance standards. *Applied Measurement in Education, 18*(1), 11–34. https://doi.org/ 10.1207/s15324818ame1801_2

Lewis, D. M., Mitzel, H. C., Mercado, R. L. & Schulz, E. M. (2012). The Bookmark standard-setting procedure. G. J. Cizek (Ed.), *Setting performance standards: Foundations, methods, and innovations* (s. 225–254) içinde. Routledge.

Lin, J. (2006). The Bookmark procedure for setting cut-scores. *Alberta Journal of Educational Research, 52*(2), 143–160. https://doi.org/10.55016/ojs/ajer.v52i1.55110

Lissitz, R. W. & Huynh, H. (2003). *Vertical equating for the Arkansas ACTAAP assessments*. Arkansas Department of Education.

Little, R. J. A. & Rubin, D. B. (2019). *Statistical analysis with missing data* (3. baskı). Wiley.

Livingston, S. A. & Zieky, M. J. (1982). *Passing scores: A manual for setting standards of performance on educational and occupational tests*. Educational Testing Service.

Livingston, S. A. & Zieky, M. J. (1989). A comparative study of standard-setting methods. *Applied Measurement in Education, 2*(2), 121–141. https://doi.org/10.1207/s15324818ame0202_3

Lohr, S. L. (2019). *Sampling: Design and analysis* (3. baskı). CRC Press.

Martin, M. O., Mullis, I. V. S. & Hooper, M. (Ed.). (2020). *Methods and procedures: TIMSS 2019 technical report*. TIMSS & PIRLS International Study Center, Boston College. https://timssandpirls.bc.edu/timss2019/methods/

Millî Eğitim Bakanlığı. (2024). *PISA 2022 Türkiye raporu*. Millî Eğitim Bakanlığı. Millî Eğitim Bakanlığı. https://pisa.meb.gov.tr/meb_iys_dosyalar/2024_03/21120745_26152640_pisa2022_rapor.pdf

Messick, S. (1989). Validity. R. L. Linn (Ed.), *Educational measurement* (3. baskı, s. 13–104). American Council on Education/Macmillan.

Mitzel, H. C., Lewis, D. M., Patz, R. J. & Green, D. R. (2001). The Bookmark procedure: Psychological perspectives. G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (s. 249–281) içinde. Lawrence Erlbaum Associates.

National Center for Education Statistics. (2022). *NAEP technical documentation*. U.S. Department of Education, Institute of Education Sciences. https://nces.ed.gov/nationsreportcard/tdw/

Nedelsky, L. (1954). Absolute grading standards for objective tests. *Educational and Psychological Measurement, 14*(1), 3–19. https://doi.org/10.1177/001316445401400101

Organisation for Economic Co-operation and Development. (2019). *PISA 2018 assessment and analytical framework*. OECD Publishing.

Organisation for Economic Co-operation and Development. (2023). *PISA 2022 assessment and analytical framework*. OECD Publishing.

Park, J. (2018). Angoff, Ebel, Bookmark, and Hofstee: A comparison of the four most popular standard setting methods. *Journal of Educational Evaluation for Health Professions, 15*, Article 10. https://doi.org/10.3352/jeehp.2018.15.32

Park, J. (2022). Possibility of using the Yes/No Angoff in the Korean Medical Licensing Examination to set the pass/fail standard. *Journal of Educational Evaluation for Health Professions, 19*, Article 10. https://doi.org/ 10.3352/jeehp.2022.19.23

Reckase, M. D. (2006). A conceptual framework for a psychometric theory for standard setting with a review of the methods. S. M. Downing & T. M. Haladyna (Ed.), *Handbook of test development* (s. 415–440) içinde. Lawrence Erlbaum Associates.

Rust, K. F. & Rao, J. N. K. (1996). Variance estimation for complex surveys using replication techniques. *Statistical Methods in Medical Research, 5*(3), 283–310. https://doi.org/10.1177/096228029600500305

Yudkowsky, R., Downing, S. M. & Wirth, S. (2008). A pilot study of a three-level Angoff method for setting standards on a patient-based clinical skills examination. *Academic Medicine, 83*(10 Suppl), S8–S10. https://doi.org/ 10.1097/ACM.0b013e318183c683

YAZAR HAKKINDA | ABOUT AUTHOR

DOÇ. DR. AHMET SALİH ŞİMŞEK

Kırşehir Ahi Evran Üniversitesi | assimsek[at]ahievran.edu.tr

ORCID: 0000-0002-9764-3285

Doç. Dr. Ahmet Salih Şimşek, Eğitimde Ölçme ve Değerlendirme alanında yüksek lisans ve doktora eğitimini 2009–2017 yılları arasında Ankara Üniversitesi'nde tamamlamıştır. Doktora sonrası araştırmaları için 2019 yılında Twente Üniversitesi'nde 8 ay süreyle Öğrenme, Veri Analitikleri ve Teknoloji (LDT: Learning, Data-Analytics and Technology) bölümünde misafir araştırmacı olarak, teknoloji destekli ölçme uygulamaları için madde tepki süresi MTK modeli geliştirilmesi üzerine çalışmıştır. Çalışma hayatında akademik personel olarak 15. yılını geride bırakan Dr. Şimşek, halen Kırşehir Ahi Evran Üniversitesi'nde Eğitimde Ölçme ve Değerlendirme Anabilim Dalı Başkanı ve üniversitenin kalite süreçlerinde Veri Yönetim Koordinatörü olarak görev almaktadır. Çeşitli kurum (MEB, ÖSYM) ve organizasyonların (PIKTES) beceri temelli ölçme değerlendirme süreçleri için izleme değerlendirme sistemlerinin tasarlanmasında çalışmalara katkı sunmaktadır. Ulusal ve uluslararası 50'den fazla yayını bulunan Dr. Şimşek, teknoloji ile zenginleştirilmiş ölçme değerlendirme süreçleri, veri, istatistik ve yapay zekâ okuryazarlıkları, bireyselleştirilmiş test geliştirme ve eğitimde yapay zekâ uygulamaları konularında çalışmalarını sürdürmektedir.

ASSOC. PROF. DR. AHMET SALİH ŞİMŞEK

Kırşehir Ahi Evran University | assimsek[at]ahievran.edu.tr

ORCID: 0000-0002-9764-3285

Assoc. Prof. Dr. Ahmet Salih Şimşek completed his master's and doctoral studies in Educational Measurement and Evaluation at Ankara University between 2009 and 2017. In 2019, he conducted postdoctoral research as a visiting researcher for eight months in the Learning, Data Analytics and Technology (LDT) department at the University of Twente, where he worked on the development of response time IRT model for technology-enhanced assessment practices. Having completed his 15th year as an academic staff member, Dr. Şimşek currently serves as the Head of the Department of Educational Measurement and Evaluation at Kırşehir Ahi Evran University and as the Data Management Coordinator in the university's quality assurance processes. He contributes to the design of monitoring and evaluation systems for skill-based assessment and evaluation processes carried out by various institutions, including the Ministry of National Education (MEB) and the Assessment, Selection and Placement Center (ÖSYM), as well as organizations such as PIKTES. Dr. Şimşek has authored more than 50 national and international publications and continues his studies on technology-enhanced assessment and evaluation processes, data, statistics and artificial intelligence literacies, adaptive test development and artificial intelligence applications in education.