ASOS JOURNAL
The Journal of Academic Social Science

Akademik Sosyal Araştırmalar Dergisi, Yıl: 14, Sayı: 179, Ağustos 2026, s. 400-416

ISSN: 2148-2489 Doi Number: http://dx.doi.org/10.29228/ASOS.94940

Yayın Geliş Tarihi / Article Arrival Date
15.06.2026

Yayımlanma Tarihi / The Publication Date
29.08.2026

Emir KAYHAN
Kırıkkale Üniversitesi Sosyal Bilimler Enstitüsü, Eğitimde Ölçme ve Değerlendirme YL
rehber.403842@gmail.com
Orcid: 0000-0001-8227-2752

Hasan ATAK
Kırıkkale Üniversitesi, Eğitimde Ölçme ve Değerlendirme ABD
hasanatak@kku.edu.tr
Orcid: 0000-0001-5637-2186

PERFORMANS DEĞERLENDİRMEDE ÇOK YÜZEYLİ RASCH ÖLÇME MODELİ: KURAMSAL TEMELLER, UYGULAMALAR VE EĞİTİM ARAŞTIRMALARINA YANSIMALARI

Öz

Bu çalışmanın amacı, performans değerlendirmelerinde öğrenci yeterliğiyle birlikte puanlara karışabilen puanlayıcı katılığı, görev güçlüğü, ölçüt işleyişi ve kategori kullanımını eş zamanlı biçimde modelleyen Çok Yüzeyli Rasch Ölçme Modeli'ni (ÇYRÖM) kuramsal ve uygulamalı boyutlarıyla incelemektir. Çalışma; temel Rasch ölçme kaynakları, ÇYRÖM ve FACETS yazını, puanlayıcı etkileri araştırmaları ve Genellenebilirlik Kuramı çalışmalarının kavramsal temalar altında çözümlendiği bir anlatı derlemesi olarak tasarlanmıştır. İnceleme; modelin matematiksel yapısı, logit ölçeği, yüzey parametreleri, infit ve outfit uyum istatistikleri, ayırma ve güvenirlik değerleri, kategori işleyişi ile analiz ve raporlama sürecine odaklanmıştır. Sentez sonuçları, ÇYRÖM'ün katılık-cömertlik, halo ve horn etkileri, merkeze yönelme, tutarsız puanlama ve farklılaşan puanlayıcı işleyişini bireysel parametreler üzerinden tanılayabildiğini göstermektedir. Modelin, puanlayıcı ve görev etkilerinden arındırılmış öğrenci ölçüleri üreterek ölçme adaletini güçlendirdiği; buna karşılık bağlantılı veri deseni, model varsayımları ve uzmanlık gerektiren yorumlama

süreçleri bakımından sınırlılıklar taşıdığı belirlenmiştir. Genellenebilirlik Kuramı ile yapılan karşılaştırma, iki yaklaşımın rakip olmaktan çok tamamlayıcı nitelikte olduğunu ortaya koymaktadır. Sonuç olarak ÇYRÖM, insan yargısına dayalı performans değerlendirmelerinde geçerlik, güvenirlik ve adil puanlama kanıtlarını güçlendiren kapsamlı bir psikometrik çerçeve sunmaktadır.

Anahtar kelimeler: Çok Yüzeyli Rasch Modeli, Puanlayıcı Güvenirliği, Performans Değerlendirme, FACETS, Puanlayıcı Yanlılığı, Adil Puanlama.

MANY-FACET RASCH MEASUREMENT IN PERFORMANCE ASSESSMENT: THEORETICAL FOUNDATIONS, APPLICATIONS AND IMPLICATIONS FOR EDUCATIONAL RESEARCH

Abstract

This study aims to examine the Many-Facet Rasch Measurement model (MFRM), which simultaneously models rater severity, task difficulty, criterion functioning, and category use, in addition to examinee proficiency, from theoretical and applied perspectives. The study was designed as a narrative review in which foundational Rasch measurement sources, MFRM and FACETS literature, research on rater effects, and studies on Generalizability Theory were synthesized under conceptual themes. The review focused on the mathematical structure of the model, the logit scale, facet parameters, infit and outfit fit statistics, separation and reliability indices, category functioning, and the analysis and reporting process. The synthesis indicates that MFRM can diagnose severity and leniency, halo and horn effects, central tendency, inconsistent scoring, and differential rater functioning through individual parameters. The model strengthens measurement fairness by producing examinee measures adjusted for rater and task effects; however, it also has limitations related to connected data designs, model assumptions, and the expertise required for interpretation. Comparison with Generalizability Theory suggests that the two approaches are complementary rather than competing. In conclusion, MFRM provides a comprehensive psychometric framework for strengthening validity, reliability, and fair-scoring evidence in performance assessments based on human judgment.

Keywords: Many-Facet Rasch Measurement, Rater Reliability, Performance Assessment, FACETS, Rater Bias, Fair Scoring.

GİRİŞ

Eğitimde ölçme ve değerlendirme, öğrenci başarısının belirlenmesinin çok ötesinde öğretimi yönlendiren, eğitim politikasına dayanak oluşturan ve bireyler hakkında alınan kararları meşrulaştıran temel bir süreçtir. Bu sürecin bilimsel niteliği; elde edilen ölçme sonuçlarının geçerliğine, güvenirliğine ve adilliğine bağlıdır (American Educational Research Association [AERA], American Psychological Association [APA] ve National Council on Measurement in Education [NCME], 2014). Geleneksel ölçme araçları —çoktan seçmeli testler, kısa yanıt maddeleri, doğru-yanlış soruları— bu standartları görece nesnel koşullarda karşılayabilmektedir. Ancak yazma, konuşma, sunum, müzik yorumu, sanat ürünü ya da proje çalışması gibi performans gerektiren alanlarda değerlendirme kaçınılmaz olarak insan yargısına dayanmaktadır (Engelhard, 1994; Weigle, 2002).

İnsan yargısına dayanan değerlendirmelerde öğrencinin aldığı puan yalnızca gerçek yeterliğini yansıtmaz; aynı zamanda değerlendirmeyi yapan puanlayıcının bireysel eğilimleri, kullanılan görevin zorluğu, rubrik ölçütlerinin yapısı ve puanlama kategorilerinin işleyişi gibi sistematik etkenlerden de etkilenir (Linacre, 1989; Myford ve Wolfe, 2003). Klasik istatistiksel yaklaşımlar —korelasyon katsayıları, yüzde uyum, kappa— puanlayıcılar arası genel uyumu betimleyebilir; ancak her bir puanlayıcının ne derece katı ya da cömert davrandığını, hangi görevin sistematik olarak daha zor olduğunu veya rubriğin hangi ölçütünde puanlayıcıların zorlandığını ayrıştırmada yetersiz kalmaktadır (Brennan, 2001; Stemler, 2004). Bu sınırlılıklar, pek çok araştırmacının dikkatini daha gelişmiş ölçme çerçevelerine yöneltmiştir (Baird vd., 2017; Hoyt, 2000).

Çok Yüzeyli Rasch Ölçme Modeli (ÇYRÖM; İng. Many-Facet Rasch Measurement, MFRM), tam olarak bu boşluğu doldurmak amacıyla Linacre (1989) tarafından geliştirilmiştir. Georg Rasch'ın (1960) tek parametreli madde tepki kuramı modelini temel alan bu yaklaşım, ölçme sürecini etkileyen birden fazla "yüzeyi" —birey, görev, puanlayıcı, ölçüt ve puanlama kategorisi— aynı ortak logit ölçeğinde modelleyerek her birinin etkisini bağımsız parametreler olarak tahmin etmektedir. Bu sayede öğrencilerin yeterlik düzeyleri, puanlayıcı katılığından ve görev zorluğundan arındırılmış biçimde kestirilmekte; değerlendirme sürecinin hangi bileşenlerinin sorunlu işlediği açıkça görülebilmektedir.

Model, dil değerlendirmesi (Eckes, 2005, 2008; Weigle, 2002), eğitimde performans ölçmesi (Engelhard, 1994; Engelhard ve Wind, 2018), öğretmen değerlendirmesi ve ölçme aracı geliştirme (Myford ve Wolfe, 2003, 2004) gibi pek çok alanda uygulama bulmuş; FACETS yazılımının geliştirilmesiyle araştırmacılar için güçlü ve pratik bir analiz ortamı oluşmuştur (Linacre, 2024). Türkiye'de de son yıllarda performans değerlendirme araçlarının psikometrik incelenmesinde bu modele yönelim belirgin biçimde artmaktadır.

Bu derleme makalenin amacı, ÇYRÖM'ün kuramsal temellerini, matematiksel yapısını, temel parametrelerini ve analiz sürecini sistematik biçimde aktarmak; modelin performans değerlendirmede sağladığı katkıları ve sınırlılıkları tartışmak; Genellenebilirlik Kuramı ile karşılaştırmalı bir değerlendirme yapmak ve Türkiye'deki araştırmalar için yönelimler önermektir. Makale önce klasik ölçme yaklaşımlarının sınırlılıklarını ortaya koymakta, ardından Rasch modelinin mantığından ÇYRÖM'e geçişi açıklamakta, temel kavramları ve analiz çıktılarını ayrıntılandırmakta ve son olarak modelin güçlü yanları ile sınırlılıklarını tartışmaktadır.

YÖNTEM

Araştırma Deseni

Bu çalışma, Çok Yüzeyli Rasch Ölçme Modeli'ne ilişkin kuramsal ve uygulamalı alanyazını bütüncül biçimde değerlendirmeyi amaçlayan bir anlatı derlemesidir. Çalışmada birincil nicel veri toplanmamış; modelin kuramsal temellerini, puanlayıcı aracılı değerlendirmelerdeki kullanımını ve analiz sürecini açıklayan temel kaynaklar kavramsal bir çerçevede incelenmiştir. Bu nedenle çalışma, sistematik derleme ya da meta-analizden farklı olarak, alandaki temel kavramları ve yöntemsel tartışmaları açıklayıcı bir sentez içinde sunmaktadır.

Kaynakların Belirlenmesi

Kaynak seçiminde; (a) Rasch ve çok yüzeyli Rasch modellerinin kuramsal temellerini açıklama, (b) puanlayıcı etkileri, model-veri uyumu ve kategori işleyişine ilişkin yöntemsel katkı

sunma, (c) performans değerlendirme uygulamalarıyla doğrudan ilişkili olma ve (d) ÇYRÖM ile Genellenebilirlik Kuramı arasındaki benzerlik ve farklılıkları tartışma ölçütleri esas alınmıştır. Temel kitaplar, hakemli dergi makaleleri, yazılım belgeleri ve güncel öğretim programı dokümanları birlikte değerlendirilmiştir.

Verilerin Analizi

İncelenen kaynaklar; klasik ölçme yaklaşımlarının sınırlılıkları, Rasch modelinden ÇYRÖM'e geçiş, temel model parametreleri, puanlayıcı etkileri, FACETS analiz süreci, Genellenebilirlik Kuramı ile karşılaştırma, modelin güçlü ve sınırlı yönleri ile Türkiye'deki araştırma yönelimleri temaları altında betimsel olarak çözümlenmiştir. Benzer görüşler bir araya getirilmiş, farklılaşan yaklaşımlar karşılaştırılmış ve elde edilen çıkarımlar ölçme adaleti, geçerlik ve güvenirlik ekseninde yorumlanmıştır.

BULGULAR VE YORUM

Klasik Ölçme Yaklaşımlarının Sınırlılıkları

Performans değerlendirmelerde puanlayıcı güvenirliği, geleneksel olarak iki ana yaklaşımla incelenmiştir: uyum istatistikleri ve varyans bileşenleri analizi. Her iki yaklaşım da belirli bağlamlarda değerli bilgiler sunsa da özellikle analitik rubriklerle yürütülen çok boyutlu performans değerlendirmelerinin gerektirdiği psikometrik derinliği tam olarak karşılayamamaktadır.

Uyum İstatistikleri Yaklaşımı

Yüzde uyum, Cohen'in kappa katsayısı (Cohen, 1960) ve ağırlıklı kappa (Cohen, 1968) gibi uyum istatistikleri, iki ya da daha fazla puanlayıcının ne sıklıkla aynı ya da bitişik puanı verdiğini ölçmektedir. Pearson ve Spearman korelasyon katsayıları puanlayıcı sıralamaları arasındaki örtüşmeyi; sınıf içi korelasyon katsayısı (ICC) ise hem uyumu hem de tutarlılığı bir arada inceleme olanağı sunmaktadır (Stemler, 2004).

Bu yaklaşımların kritik sınırlılığı, puanlayıcıların birbirinden sistematik olarak ne kadar katı ya da cömert olduğunu belirleyememeleridir. İki puanlayıcı arasında yüksek korelasyon, birinin sürekli diğerinden 1 puan yüksek puan verdiği durumda bile elde edilebilir. Öte yandan bu istatistikler, puanlayıcı güvenirliğinin düşük olmasının tutarsızlıktan mı yoksa sistematik bir yönelimden mi kaynaklandığını ayırt etmemekte; ayrıca birden fazla hata kaynağını eş zamanlı olarak modelleyememektedir (Crocker ve Algina, 2008; Hoyt, 2000).

Genellenebilirlik Kuramı

Cronbach ve arkadaşları (1972) tarafından geliştirilen Genellenebilirlik Kuramı (G Kuramı), varyans bileşenleri analizi aracılığıyla ölçme hatalarını öğrenci, görev, puanlayıcı ve bunların etkileşimleri gibi farklı kaynaklara atfetmektedir. G Kuramı bu yönüyle klasik yaklaşımlara önemli bir üstünlük sağlar: hangi hata kaynağının ölçme sonuçlarını ne ölçüde etkilediğini nicel olarak gösterebilir. "Karar çalışmaları" aracılığıyla ise farklı ölçme desenlerinin —örneğin puanlayıcı sayısının artırılması— güvenirlik üzerindeki etkisi simüle edilebilir (Brennan, 2001; Shavelson ve Webb, 1991).

Bununla birlikte G Kuramı'nın da önemli sınırlılıkları vardır. Model, her bir puanlayıcının katılık ya da cömertlik düzeyini bireysel bir parametre olarak doğrudan tahmin etmemekte; bunun yerine puanlayıcı kaynaklı varyansı bir bütün olarak raporlamaktadır. Dolayısıyla "Puanlayıcı 3

diğerlerinden önemli ölçüde daha katı mıdır?" gibi tanılayıcı sorulara net yanıt vermek güçtür. Ayrıca G Kuramı, ham puanların aralıklı ölçekte olduğunu varsayarak ordinal ölçekteki rubrik verilerini doğrudan işlemekte; bu durum ölçüm hatalarını artırabilmektedir (Bond ve Fox, 2015).

ÇYRÖM, her iki yaklaşımın bu sınırlılıklarını aşmak üzere tasarlanmıştır: bireysel puanlayıcı parametrelerini doğrudan tahmin eden, ordinal verileri eşit aralıklı logit ölçeğine dönüştüren ve birden fazla hata kaynağını eş zamanlı olarak modelleyebilen bu yaklaşım, performans değerlendirme araştırmalarında giderek daha merkezi bir konuma gelmektedir.

Rasch Modelinden Çok Yüzeyli Rasch Modeline

Rasch Modelinin Temel Mantığı

Georg Rasch (1960) tarafından geliştirilen tek parametreli madde tepki kuramı modeli, bireyin bir maddeye doğru yanıt verme olasılığının iki büyüklüğün farkıyla belirlendiğini öne sürmektedir: bireyin yetenek düzeyi (β) ve maddenin güçlük düzeyi (δ). Temel Rasch modeli aşağıdaki log-olasılık oranı biçiminde ifade edilmektedir:

$$\ln[\, P(X_{ni} = 1) / P(X_{ni} = 0) \,] = \beta_n - \delta_i \quad (1)$$

Bu eşitlikte X_{ni}, n bireyinin i maddesine verdiği yanıtı; β_n bireyin yeterlik düzeyini ve δ_i maddenin güçlük düzeyini göstermektedir. Bireyin yeterliği arttıkça ve madde güçlüğü azaldıkça doğru yanıt olasılığı yükselmektedir. Modelin temel üstünlüğü, uygun model-veri uyumu koşullarında madde ve birey parametrelerinin ortak bir ölçek üzerinde karşılaştırılabilmesini sağlayan değişmezlik ilkesidir. Rasch ölçümü, sıralı ham puanları eşit aralıklı logit ölçeğine dönüştürerek ölçme sonuçlarının daha anlamlı biçimde yorumlanmasına olanak tanır (Bond ve Fox, 2015; Wright ve Masters, 1982).

Dereceli Puanlama Modeli

Performans değerlendirmelerinde puanlar genellikle ikili (0–1) değil, sıralı kategorilerden (1–2–3–4 gibi) oluşmaktadır. Wright ve Masters (1982), bu yapıyı ele almak üzere Rasch modelini Dereceli Puanlama Modeli'ne (Rating Scale Model, RSM) genişletmiştir. Modelde her kategoriden bir üst kategoriye geçişin zorluğunu temsil eden eşik parametreleri (τ_k) yer almaktadır:

$$\ln[\, P(X_{ni} = k) / P(X_{ni} = k{-}1) \,] = \beta_n - \delta_i - \tau_k \quad (2)$$

Burada $P(X_{ni} = k)$, bireyin k kategorisinde puan alma olasılığını; τ_k ise k−1. kategoriden k. kategoriye geçişin zorluğunu ifade etmektedir. Bu genişleme, rubrik puanlama anahtarlarında yaygın olan çok kategorili verilerin Rasch çerçevesinde analiz edilmesine zemin hazırlamıştır.

Çok Yüzeyli Rasch Modeline Geçiş

Linacre (1989), Dereceli Puanlama Modeli'ni insan yargısına dayalı performans değerlendirme bağlamına uyarlarken modele puanlayıcı yüzeyini eklemiş ve böylece ÇYRÖM'ün temelini atmıştır. Temel üç yüzeyli model aşağıdaki biçimde gösterilebilir:

$$\ln[\, P(X_{nij} = k) / P(X_{nij} = k{-}1) \,] = \beta_n - \delta_i - \alpha_j - \tau_k \quad (3)$$

Bu eşitlikte β_n bireyin yeterlik düzeyini, δ_i görevin güçlüğünü, α_j puanlayıcının katılık-cömertlik eğilimini ($\alpha_j > 0$ katı; $\alpha_j < 0$ cömert) ve τ_k puanlama kategorisi eşiğini ifade etmektedir. Öğrenci, görev, puanlayıcı ve rubrik ölçütü olmak üzere dört yüzey içeren bir model ise aşağıdaki biçimde genişletilebilir:

$$\ln[\, P(X_{nijc} = k) / P(X_{nijc} = k-1) \,] = \beta_n - \delta_i - \alpha_j - \varphi_c - \tau_k \quad (4)$$

Bu yapıda φ_c, c. rubrik ölçütünün yapısal güçlüğünü göstermektedir. Araştırma bağlamına göre zaman, değerlendirme koşulu ya da öğrenci profili gibi ek yüzeyler de modele dâhil edilebilir (Linacre, 1994).

Modelin en ayırt edici özelliği, tüm bu parametreleri aynı ortak logit ölçeğinde eş zamanlı olarak tahmin etmesidir. Böylece katı puanlayan bir puanlayıcı tarafından değerlendirilen bir öğrencinin puanı, farklı bir puanlayıcı tarafından değerlendirilen öğrencilerle adil biçimde karşılaştırılabilecek şekilde düzeltilebilmektedir. Bu özellik, ÇYRÖM'ü "adil puanlama" (fair scoring) yaklaşımının temel psikometrik aracı kılmaktadır (Engelhard ve Wind, 2018).

Temel Kavramlar ve Analiz Çıktıları

Yüzey Kavramı

ÇYRÖM'de "yüzey" (facet), ölçme sürecini sistematik olarak etkileyen her bir değişken kümesini ifade etmektedir. Her yüzey, aynı logit ölçeğinde konumlandırılan ve kendi bağlamında yorumlanan parametreler bütününü oluşturur. Bir yazma becerisi araştırmasında tipik yüzeyler şunlardır: (a) öğrenci —gerçek yazma yeterliğini temsil eden birey yüzeyi—; (b) puanlayıcı —değerlendirme yapan kişilerin katılık-cömertlik eğilimlerini içeren yüzey—; (c) görev —öğrencilerin performans sergilediği yazma etkinliklerinin güçlük düzeyleri—; (d) ölçüt —rubrikte yer alan içerik, örgütleme, dil kullanımı gibi boyutların işleyiş güçlükleri.

Tablo 1.

ÇYRÖM'de Temel Yüzeyler, Parametreleri ve Yorumlanması

Yüzey	Parametre	Yüksek Logit → …	Düşük Logit → …
Öğrenci	β (Beta)	Yüksek yeterlik / başarı performansı	Düşük yeterlik / zayıf performans
Puanlayıcı	α (Alfa)	Katı puanlama → öğrenci puanları baskılanır	Cömert puanlama → öğrenci puanları şişirilir
Görev / Madde	δ (Delta)	Zor görev → başarı olasılığı düşer	Kolay görev → başarı olasılığı artar
Ölçüt / Boyut	φ (Fi)	Yüksek puan almanın güç olduğu ölçüt	Yüksek puan almanın görece kolay olduğu ölçüt
Kategori Eşiği	τ (Tau)	Bir üst kategoriye geçiş daha zor	Bir üst kategoriye geçiş daha kolay

Not. Parametre sembolleri Linacre (1989, 1994) ve Bond ile Fox (2015) esas alınarak düzenlenmiştir.

Logit Ölçeği

ÇYRÖM'ün temel çıktısı, ham puanlardan farklı olarak eşit aralıklı olan logit birimiyle ifade edilmektedir. Bir logit, olasılık oranının doğal logaritmasıdır ve matematiksel olarak logit = ln[P / (1 − P)] biçiminde tanımlanmaktadır. Ortalama sıfır etrafında simetrik dağılan bu ölçek, farklı yüzeylerin parametrelerinin doğrudan karşılaştırılmasına olanak tanımaktadır.

Uygulamada tipik yorum: öğrencinin yeterlik düzeyi +2,00 logit iken bir görevin güçlüğü de +2,00 logit ise, söz konusu öğrencinin o görevi başarıyla tamamlama olasılığı yaklaşık %50'dir. Aynı öğrencinin −1,00 logit güçlüğündeki bir görevi başarıyla tamamlama olasılığı ise e^(2−(−1))

/ (1 + e^(2−(−1))) ≈ %95'e yükselir. Bu yorumlama çerçevesi, öğretmen ve araştırmacıların model çıktılarını pedagojik anlamda işlevselleştirmesini kolaylaştırmaktadır (Bond ve Fox, 2015; Eckes, 2011).

Uyum İstatistikleri: Infit, Outfit ve ZSTD

Modelin veriyle ne ölçüde uyuştuğunu değerlendirmek için iki temel istatistik kullanılmaktadır. Infit (bilgi ağırlıklı uyum istatistiği), bireyin ya da puanlayıcının yetenek/katılık düzeyine yakın noktalardaki beklenmedik yanıt örüntülerine duyarlı olup sistematik tutarsızlıkları daha iyi yakalar. Outfit (sıradan uyum istatistiği) ise beklenti dışında kalan aşırı uç noktalardaki yanıtlara daha duyarlıdır (Bond ve Fox, 2015; Linacre, 1994).

Her iki istatistik için kare ortalama (mean square, MNSQ) değerlerinin 0,50–1,50 aralığında olması genel kabul görmüş bir ölçüttür; bu aralık örneklem büyüklüğüne ve araştırma amacına göre 0,70–1,30'a daraltılabilir. 1,00 değeri mükemmel model uyumunu ifade eder: MNSQ > 1,50 beklenenden fazla tutarsızlığa; MNSQ < 0,50 ise mekanik biçimde tahmin edilebilir, bilgi üretmeyen puanlama örüntüsüne işaret eder.

MNSQ değerleriyle birlikte raporlanan standartlaştırılmış artık (ZSTD) istatistiği, gözlenen uyumsuzluğun istatistiksel anlamlılığını test eder. Genel kural olarak |ZSTD| > 2,0 kritik sınır kabul edilmekte; örneklem büyüklüğü arttıkça ZSTD'nin aşırı duyarlı hale geldiği göz önünde bulundurularak MNSQ ve ZSTD'nin birlikte değerlendirilmesi önerilmektedir (Bond ve Fox, 2015; Linacre, 1994). Özellikle yüksek outfit değeri olan bir puanlayıcı için MNSQ ve ZSTD ikilisinin analizi, rastlantısal mı yoksa sistematik bir hata mı söz konusu olduğunu ayırt etmeye yardımcı olmaktadır.

Ayırma İndeksi ve Güvenirlik Katsayısı

Her yüzey için ayrı ayrı hesaplanan ayırma indeksi (separation index, G), ilgili yüzeydeki elemanların birbirinden ne ölçüde ayırt edilebildiğini göstermektedir. Güvenirlik katsayısı ise bu ayrımın ne denli tekrarlanabilir olduğunu ortaya koymaktadır; iki istatistik arasındaki ilişki $G^2 / (1 + G^2)$ dönüşümüyle sağlanmaktadır.

Öğrenci yüzeyinde yüksek ayırma ve güvenirlik istenen bir durumdur: öğrencilerin performans düzeylerinin güvenilir biçimde birbirinden ayrıldığını ve ölçme aracının yeterli ayırt edici güce sahip olduğunu gösterir. Puanlayıcı yüzeyinde yüksek ayırma ise farklı bir anlam taşır: puanlayıcıların katılık-cömertlik bakımından önemli ölçüde birbirinden farklılaştığını, yani puanlama standartlarının homojen olmadığını göstermekte ve bu nedenle puanlayıcı eğitimi ile kalibrasyon süreçlerine duyulan gereksinimi ortaya koymaktadır (Eckes, 2011; Myford ve Wolfe, 2003). Görev yüzeyinde yüksek ayırma ise görevlerin güçlük bakımından gerçekten farklılaştığını; dolayısıyla görev puanlarının doğrudan karşılaştırılmasında dikkatli olunması gerektiğini işaret etmektedir.

Kategori İşleyişi

Rubrik gibi dereceli puanlama araçlarında kategorilerin beklenen sırada işleyip işlemediği, ÇYRÖM çerçevesinde Andrich eşik parametreleri (threshold parameters) aracılığıyla incelenmektedir. Teorik olarak her eşik değerinin bir öncekinden büyük olması —yani $\tau_1 < \tau_2 < \tau_3$ sırasının korunması— beklenmektedir; düzensiz eşik sırası, bazı kategorilerin puanlayıcılar tarafından yeterince ayırt edilemediğine işaret eder.

Örneğin 1–4 arasında puanlanan bir rubrikte $\tau_2 > \tau_1$ koşulunun sağlanmaması, 2. ve 3. kategorilerin anlamsız hâle geldiğini ve iki bitişik kategorinin birleştirilmesinin ölçme kalitesini artırabileceğini göstermektedir. FACETS çıktılarında kategori gözlem frekansları (her kategorinin ne sıklıkla kullanıldığı), kategoriye verilen ortalama logit ölçüleri ve eşik sıralamaları bu inceleme için temel veri kaynaklarıdır (Linacre, 1994; Wright ve Masters, 1982). Genel uygulama kılavuzu olarak her kategorinin en az 10 gözlem içermesi ve gözlem frekanslarının tek modlu bir dağılım sergilemesi önerilmektedir.

ÇYRÖM'de Puanlayıcı Etkilerinin İncelenmesi

ÇYRÖM'ün performans değerlendirme araştırmalarına en özgün katkısı, puanlayıcı kaynaklı sistematik hata türlerini nicel parametreler düzeyinde ayrıştırabilmesidir. Myford ve Wolfe (2003, 2004), bu konudaki kapsamlı çalışmalarıyla modelin hangi puanlayıcı etkilerini tanılayabileceğini sistematik biçimde ortaya koymuştur. Aşağıda başlıca puanlayıcı etkileri açıklanmaktadır.

Katılık ve Cömertlik Efektleri

Katılık (severity) efekti, bir puanlayıcının gerçek performans düzeyine kıyasla sistematik olarak düşük puan verme eğilimini ifade eder. Bunun tersi olan cömertlik (leniency) efektinde ise puanlayıcı sistematik olarak yüksek puan vermektedir. ÇYRÖM'de her puanlayıcı ortak logit ölçeğinde bir noktaya yerleştirilmekte; pozitif değerler katılığı, negatif değerler cömertliği göstermektedir. Örneğin FACETS çıktısında Puanlayıcı A'nın +0,85 logit, Puanlayıcı B'nin −0,62 logit değerine sahip olduğu görülürse; A sistematik olarak daha katı, B ise daha cömert bir puanlama eğilimi sergilemektedir.

Bu parametrenin pratik önemi büyüktür: farklı öğrencilerin farklı puanlayıcılar tarafından değerlendirildiği durumlarda gözlenen ham puan farklılıkları, gerçek yeterlik farklılıklarını değil puanlayıcı eğilimlerini yansıtıyor olabilir. ÇYRÖM, "adil puan" (fair score) hesaplama yöntemiyle her öğrencinin puanını aldığı puanlayıcının katılık düzeyine göre düzelterek daha adil karşılaştırmalar yapılmasına olanak sağlar (Eckes, 2008; Engelhard ve Wind, 2018).

Halo Etkisi

Thorndike (1920) tarafından tanımlanan halo etkisi, puanlayıcının bir ölçüte ya da genel izlenime ilişkin yargısının diğer bağımsız ölçütlerin puanlanmasını da etkilemesi durumudur. Yazma değerlendirmesinde örneğin yazı estetiğinden ya da dil akıcılığından olumlu etkilenen bir puanlayıcı, içerik ve örgütleme ölçütlerini de olduğundan yüksek puanlayabilmektedir.

ÇYRÖM çerçevesinde halo etkisi, puanlayıcının farklı ölçütlerde verdiği puanların beklenen örüntüden sapması incelenerek değerlendirilebilmektedir. Bir puanlayıcının tüm ölçütlerde benzer puan dağılımı göstermesi —yani ölçütler arasında yeterli varyasyon yaratmaması— halo etkisinin varlığına işaret edebilir. Bununla birlikte Murphy vd.'nin (1993) uyardığı üzere, yazma becerisinin bazı boyutları gerçekten de birbiriyle ilişkili olduğundan her yüksek ölçütler arası korelasyonu halo hatası olarak yorumlamak yanıltıcı olacaktır.

Horn Etkisi

Horn etkisi, halo etkisinin olumsuz yönlü biçimidir: puanlayıcının bir boyuttaki olumsuz yargısının diğer bağımsız ölçütleri de olumsuz etkilemesi durumudur. Yazma değerlendirmesinde örneğin yazım-noktalama hatalarının çokluğundan olumsuz etkilenen bir puanlayıcı, içerik zenginliği ve özgünlük ölçütlerini de olduğundan düşük puanlayabilmektedir. Horn etkisi,

potansiyel olarak güçlü öğrencilerin yeteneklerinin gözden kaçırılmasına yol açtığından ölçme adaleti açısından ciddi sonuçlar doğurabilmektedir (Saal vd., 1980).

Merkeze Yönelme

Merkeze yönelme (central tendency), puanlayıcının uç kategorileri kullanmaktan kaçınarak performans düzeylerinden bağımsız biçimde orta kategorilere yönelmesidir. Bu durum, yüksek performanslı öğrencilerin hak ettiğinden düşük, düşük performanslı öğrencilerin ise gereğinden yüksek puanlanmasına; dolayısıyla öğrenciler arası gerçek yeterlik farklılıklarının gizlenmesine yol açmaktadır. ÇYRÖM çıktılarındaki kategori kullanım frekansları incelenerek belirli kategorileri nadiren ya da hiç kullanmayan puanlayıcılar saptanabilmektedir.

Farklılaşan Puanlayıcı İşleyişi

Farklılaşan puanlayıcı işleyişi (Differential Rater Functioning, DRF), bir puanlayıcının belirli öğrenci gruplarına, belirli görevlere ya da belirli rubrik ölçütlerine karşı sistematik olarak farklı bir katılık düzeyi sergilemesidir. Bu kavram, madde tepki kuramındaki "Farklılaşan Madde İşleyişi" (Differential Item Functioning, DIF) olgusunun puanlayıcı bağlamına uyarlamasıdır. Örneğin bir puanlayıcı kız öğrencilerin yazılı ürünlerini erkek öğrencilere kıyasla sistematik olarak daha yüksek ya da düşük puanlıyorsa DRF söz konusudur. DRF, eğitimde ölçme adaleti açısından ciddi sonuçlar doğurabilmekte; belirli demografik özelliklere sahip öğrencilerin sistematik biçimde dezavantajlı konuma düşmesine yol açabilmektedir (Myford ve Wolfe, 2004).

Tablo 2.

Performans Değerlendirmede Başlıca Puanlayıcı Etkileri: ÇYRÖM Tanılaması

Puanlayıcı Etkisi	Tanımı	Ölçme Adaletine Etkisi	ÇYRÖM ile Tanılama Yolu
Katılık / Cömertlik	Sistematik düşük ya da yüksek puan eğilimi	Ham puanlar gerçek yeterliği yansıtmaz	Logit ölçeğindeki α parametresi; adil puan hesabı
Halo Etkisi	Genel olumlu izlenimi diğer ölçütlere yansıtma	Ölçütler arası yapay yüksek korelasyon	Ölçüt bazlı puan örüntüleri; infit/outfit istatistikleri
Horn Etkisi	Genel olumsuz izlenimi diğer ölçütlere yansıtma	Güçlü boyutlar gözden kaçırılır	Ölçüt bazlı puan asimetrisi; çok boyutlu uyum analizi
Merkeze Yönelme	Uç kategorilerden kaçınarak ortalamaya yönelme	Öğrenciler arası gerçek farklar gizlenir	Kategori kullanım frekansları ve Andrich eşik sıralamaları
DRF	Belirli gruplara / görevlere / ölçütlere farklı katılık	Grup bazlı ölçme adaleti ihlali	Yüzey etkileşim analizi; ki-kare uyum testi
Rastgelelik / Tutarsızlık	Sistematik olmayan, değişken puanlama davranışı	Ölçme hatası artar, güvenirlik düşer	Yüksek infit ve outfit MNSQ değerleri; yüksek ZSTD

Not. DRF: Farklılaşan Puanlayıcı İşleyişi. Sınıflandırma Myford ve Wolfe (2003, 2004), Engelhard (1994) ve Saal vd. (1980) esas alınarak hazırlanmıştır.

FACETS Yazılımı ve Analiz Süreci

ÇYRÖM analizleri için en yaygın kullanılan yazılım, Linacre (1989) tarafından geliştirilen ve etkin biçimde güncellenen FACETS'tir. Yazılım, araştırmacının tanımladığı yüzey yapısına göre ham verileri ortak logit ölçeğindeki parametrelere dönüştürmekte ve her yüzey için kapsamlı

çıktı tabloları üretmektedir (Linacre, 2024). Alternatif yazılımlar arasında R ortamında çalışan TAM paketi (Robitzsch vd., 2022) ile WINMIRA sayılabilir; ancak FACETS, kullanım kolaylığı ve çıktı zenginliği bakımından alanyazında en geniş kabul gören araç olmaya devam etmektedir.

Veri Yapısı ve Analiz Deseni

FACETS analizi için veri genellikle her satırın bir gözlemi temsil ettiği düz metin ya da CSV formatında hazırlanır. Tipik bir yazma değerlendirmesi için veri kaydı aşağıdaki yapıyı izler:

Tablo 3.

FACETS Veri Girişi: Örnek Yapı (Öğrenci × Puanlayıcı × Ölçüt)

Öğrenci	Puanlayıcı	Görev	Ölçüt	Puan
1	A	1	İçerik	3
1	A	1	Örgütleme	2
1	A	1	Dil Kullanımı	3
1	B	1	İçerik	4
…	…	…	…	…

Not. Bu yapıda her satır bağımsız bir gözlemi temsil eder. Tam çapraz desende her öğrenci her puanlayıcı tarafından, her görev için tüm ölçütlerde puanlanır.

Veri deseni iki ana türde olabilmektedir. Tam çapraz (fully crossed) desende her öğrenci her puanlayıcı tarafından değerlendirilmekte; bu desen en yüksek bilgi düzeyini ve en güvenilir parametre tahminlerini sağlamaktadır. Kısmen bağlı (partially connected) desende ise bazı öğrenciler yalnızca belirli puanlayıcılarla eşleştirilmekte; parametrelerin doğru tahmin edilebilmesi için desende yeterli bağlantı olması zorunludur. Bağlantısız desenler parametre tahminini matematiksel olarak imkânsız kıldığından araştırma planlaması aşamasında desen bütünlüğü dikkate alınmalıdır.

Temel Çıktı Tabloları

FACETS analizi sonucunda her yüzey için ayrı bir çıktı tablosu üretilmektedir. Bu tablolarda her eleman için logit ölçüsü, standart hata, MNSQ infit, MNSQ outfit, ZSTD infit, ZSTD outfit, gözlem sayısı ve ortalama puan yer almaktadır.

Değişken haritası (Wright haritası ya da "bubble chart"), tüm yüzeylerin parametrelerini aynı logit ölçeğinde görsel olarak sunan ve analiz raporlamasının en kritik unsurlarından biridir. Bu harita üzerinde yüksek yeterlikli öğrenciler, zor görevler, katı puanlayıcılar ve yüksek puan gerektiren ölçütler ölçeğin üst bölgelerinde; düşük yeterlikli öğrenciler, kolay görevler, cömert puanlayıcılar ve kolay ölçütler alt bölgelerinde konumlanmaktadır. Harita, öğrenci yeterliğinin görev güçlüğüyle ne ölçüde eşleştiğini —yani testin hedef kitlesine uygunluğunu— tek bir bakışta değerlendirmeye olanak tanımaktadır.

FACETS Analiz Süreci

Tablo 4.

FACETS ile ÇYRÖM Analiz Süreci: Aşamalar ve İçerikler

Aşama	Adı	İçeriği / Yapılan İşlem	FACETS Çıktısı
1	Araştırma Deseni Belirleme	Yüzeyler tanımlanır; tam çapraz / kısmen bağlı desen seçilir	Spesifikasyon dosyası (.fac)
2	Veri Hazırlama	Her gözlem için öğrenci × puanlayıcı × görev × ölçüt × puan kaydı düzenlenir	Veri dosyası (.dat)
3	İlk Analiz	Model parametreleri JMLE algoritmasıyla kestirilir; yakınsama kontrol edilir	Parametre kestirimleri; standart hatalar
4	Model-Veri Uyumu	Infit ve Outfit MNSQ / ZSTD incelenir; uyumsuz elemanlar saptanır	Uyum tabloları; baloncuk grafiği
5	Kategori İşleyişi	Andrich eşik sıralaması ve kategori frekansları değerlendirilir	Kategori olasılık eğrileri
6	Puanlayıcı Analizi	Katılık-cömertlik sıralaması, DRF ve merkeze yönelme incelenir	Puanlayıcı tablosu; değişken haritası
7	Adil Puan Hesabı	Puanlayıcı etkisinden arındırılmış öğrenci ölçüleri hesaplanır	Adil ortalama puanlar tablosu
8	Raporlama	Her yüzey için parametre kestirimleri, uyum, ayırma ve güvenirlik raporlanır	Tablo ve şekiller; yorumlar

Not. JMLE: Joint Maximum Likelihood Estimation (Birleşik En Çok Olabilirlik Kestirimi). Süreç Linacre (1989, 2024) ve Eckes (2011) esas alınarak düzenlenmiştir.

Raporlamada Dikkat Edilmesi Gereken Noktalar

ÇYRÖM sonuçlarının raporlanmasında yalnızca istatistiksel değerlerin sunulması yeterli değildir; bu değerlerin ölçme bağlamında ne anlama geldiğinin de açıklanması gerekmektedir. Öncelikle modele dâhil edilen yüzeyler ve analiz deseni (tam çapraz mı, kısmen bağlı mı) açık biçimde belirtilmelidir. Her yüzey için logit ölçüleri, standart hatalar, infit ve outfit MNSQ/ZSTD değerleri ile ayırma indeksleri ve güvenirlik katsayıları raporlanmalıdır.

Puanlayıcı yüzeyi ayrıntılı biçimde yorumlanmalı; en katı ve en cömert puanlayıcılar, model uyumundan sapan puanlayıcılar ve bunların öğrenci puanlarına olası etkileri tartışılmalıdır. Öğrenci yüzeyinde ham puanlar ile düzeltilmiş (adil) puanlar arasındaki uyum ya da farklılıklar örneklenerek sunulmalıdır. Kategori işleyişine ilişkin bulgular, rubriğin revize edilip edilmeyeceği konusunda somut önerilere dönüştürülmeli; durum yalnızca teknik bir sorun olarak değil, öğretim için çıkarımlar üretecek biçimde yorumlanmalıdır.

ÇYRÖM ve Genellenebilirlik Kuramı: Karşılaştırmalı Bir Değerlendirme

Performans değerlendirme araştırmalarında hata kaynaklarını incelemede en yaygın kullanılan iki yaklaşım ÇYRÖM ve G Kuramı'dır. Her iki yaklaşım da birden fazla hata kaynağını aynı analizde ele alabilmektedir; ancak aralarında köklü kavramsal ve metodolojik farklılıklar bulunmaktadır.

G Kuramı'nın temel gücü, her yüzeyin ve yüzeyler arası etkileşimlerin toplam varyansa katkısını yüzde olarak göstermesindedir. Bu özellik "hangi hata kaynağı en fazla ölçme hatasına katkıda bulunuyor?" sorusuna yanıt vermede son derece işlevseldir. Karar çalışmaları aracılığıyla

ise farklı ölçme desenlerinin güvenirlik katsayısına etkisi simüle edilerek pratik ölçme deseni önerileri sunulabilir (Brennan, 2001; Shavelson ve Webb, 1991).

ÇYRÖM'ün G Kuramı'na göre üstün olduğu nokta, her bireysel puanlayıcıyı ortak bir ölçüm noktasında konumlandırmasıdır. "Puanlayıcı 3 diğerlerinden sistematik olarak 0,80 logit daha katıdır ve bu fark istatistiksel olarak anlamlıdır" biçiminde tanılayıcı bir ifade G Kuramı çerçevesinde dile getirilememektedir. Üstelik ÇYRÖM, ordinal ham puanları aralıklı logit ölçeğine dönüştürerek ölçüm hatalarını azaltmakta; G Kuramı ise genellikle ham puanların aralıklı olduğunu varsaymaktadır.

Öte yandan G Kuramı'nın bazı önemli üstünlükleri de bulunmaktadır. G Kuramı, karma ve dengesiz desenlere ÇYRÖM'e kıyasla daha esnek biçimde uygulanabilmektedir. Bunun yanı sıra karar çalışmaları, araştırma öncesi örneklem büyüklüğü planlaması için son derece kullanışlı bir araçtır. İki yaklaşımın rekabet değil tamamlama ilişkisi içinde olduğu ve ideal durumda birlikte kullanılabileceği görüşü alanyazında giderek yaygınlaşmaktadır (Brennan, 2001; Eckes, 2011).

Tablo 5.

Çok Yüzeyli Rasch Modeli ve Genellenebilirlik Kuramı: Karşılaştırmalı Özellikler

Özellik	Çok Yüzeyli Rasch Modeli	Genellenebilirlik Kuramı
Bireysel puanlayıcı parametresi	Her puanlayıcı için ayrı logit ölçüsü (α)	Puanlayıcı varyans bileşeni (toplu)
Ölçek türü	Eşit aralıklı logit ölçeği	Ham puan (aralıklı ölçek varsayımı)
Adil puan hesabı	Evet (puanlayıcı ve görev etkisinden arındırılmış)	Hayır
Karar çalışması	Kısmi (bootstrap simülasyon)	Evet (standart G ve D çalışmaları)
Eksik veri	Bağlantı koşuluyla esnek	Kısmi; denge ve bağlantı gerektirir
Kategori işleyişi	Andrich eşik parametreleri ile incelenir	Doğrudan incelenmez
Ölçüm değişmezliği	Temel ilke; örneklemden bağımsız tahmin	Test edilmez
Yazılım	FACETS, TAM (R), jMetrik	GENOVA, urGENOVA, R (gtheory)

Not. Karşılaştırma Brennan (2001), Eckes (2011) ve Bond ile Fox (2015) kaynaklarına dayanmaktadır.

ÇYRÖM'ün Güçlü Yanları ve Sınırlılıkları

Güçlü Yanlar

ÇYRÖM'ün performans değerlendirme araştırmalarına en belirgin katkısı ölçme adaletine yaptığı katkıdır. Model, öğrencilerin yeterlik tahminlerini puanlayıcı katılığı ve görev zorluğu gibi sistematik hatalardan arındırmakta; böylece ham puan sıralaması değişebilmektedir. Bu durum özellikle farklı öğrencilerin farklı puanlayıcılar ya da farklı görevlerle değerlendirildiği yüksek riskli (high-stakes) değerlendirmelerde —sınıf geçme, burs belirleme, öğretmen yeterliliği— kritik önem taşımaktadır.

İkinci güçlü yan tanılayıcı işlevidir: model, puanlama sisteminin hangi bileşenlerinin sorunlu işlediğini somut biçimde ortaya koymakta; bu tanılama doğrudan puanlayıcı eğitimi programlarına, rubrik revizyonuna ve ölçme deseni iyileştirmelerine rehberlik edebilmektedir.

Üçüncü üstünlük, ortak logit ölçeğinin sağladığı ölçüm bütünlüğüdür. Farklı yüzeylerin parametrelerinin aynı ölçekte ifade edilmesi "öğrencinin yeterliği görev güçlüğüyle eşleşiyor mu?" ya da "puanlayıcı katılığının büyüklüğü gerçek öğrenci yeterlik farklılıklarını gölgeleyecek düzeye mi ulaşmış?" gibi sorgulara sayısal yanıtlar üretmeyi olanaklı kılmaktadır.

Dördüncü güçlü yan, öğretimsel uygulanabilirliğidir: örneğin öğrencilerin en çok zorlandığı ölçütün "örgütleme" olduğunu gösteren bir ÇYRÖM bulgusu, yazma öğretiminde paragraf yapısı ve metin planlama etkinliklerine daha fazla yer verilmesi gerektiğine dair somut bir öğretim kararını gerekçelendirebilmektedir.

Sınırlılıklar

ÇYRÖM'ün temel sınırlılığı, sağlıklı parametre tahmini için yüzeyler arası yeterli bağlantı (connectivity) gerektirmesidir. Tamamen kopuk desenlerde —örneğin hiçbir öğrencinin birden fazla puanlayıcı tarafından değerlendirilmediği durumlarda— puanlayıcı parametrelerini öğrenci parametrelerinden matematiksel olarak ayrıştırmak imkânsız hale gelmektedir.

İkinci sınırlılık, model çıktılarının yorumlanmasının uzmanlık gerektirmesidir: logit ölçeği, infit ve outfit MNSQ/ZSTD değerleri ve Andrich eşik parametrelerinin doğru yorumlanması için Rasch ölçme kuramına ilişkin sağlam bir teorik altyapı zorunludur.

Üçüncü sınırlılık, geçerlik bütünlüğüne ilişkindir: ÇYRÖM, puanlama sürecinin iç yapısına dair güçlü kanıtlar üretir; ancak performans görevinin öğretim hedefleriyle uyumu, rubrik içerik geçerliği ve sonuç geçerliği gibi geçerliğin diğer boyutları ayrıca incelenmelidir. Model tek başına ölçme aracının geçerli olduğunu kanıtlamaz (AERA, APA ve NCME, 2014).

Dördüncü sınırlılık, tek boyutluluk varsayımıdır. ÇYRÖM, ölçülen özelliğin baskın bir latent boyut boyunca tanımlanabildiğini varsaymaktadır. Yazma becerisi gibi çok bileşenli performans alanlarında bu varsayım; kuramsal yapı, ölçütler arası ilişkiler, model-veri uyumu ve uygun artık temelli tanılamalar birlikte değerlendirilerek incelenmelidir. ÇYRÖM tek başına performansın tüm çok boyutlu yapısını açıklamak yerine, tanımlanan ölçme modelinin veriye ne ölçüde uyduğuna ilişkin kanıt üretir.

Türkiye'de ÇYRÖM Araştırmaları ve Yönelimler

Türkiye'de ölçme-değerlendirme alanında ÇYRÖM kullanımı 2010'ların ortasından itibaren belirgin biçimde artmıştır. Ulusal tez veritabanları ve hakemli dergi taramaları incelendiğinde bu çalışmaların dört ana başlık altında toplandığı görülmektedir: (a) yazılı anlatım ve yazma becerisi değerlendirmeleri; (b) yabancı dil sözlü ve yazılı performans ölçmeleri; (c) öğretmen adaylarının mesleki performanslarının puanlayıcı güvenirliği açısından incelenmesi; (d) okul öncesi, beden eğitimi, müzik ve görsel sanatlar gibi performans ağırlıklı derslerdeki rubrik temelli değerlendirmelerin psikometrik sınanması.

Bu çalışmaların ortak bulgusu, puanlayıcılar arasında anlamlı katılık-cömertlik farklılıklarının varlığı ve bu farklılıkların ham puan sıralamalarını çoğu zaman değiştiren büyüklüklere ulaşmasıdır. Söz konusu bulgu, Türk eğitim sisteminde performans değerlendirme bağlamında "puanlayıcı kalibrasyonu" ve "puanlayıcı eğitimi" kavramlarının araştırma ve uygulamada daha merkezi bir konuma gelmesini gerektirmektedir.

Türk eğitim sisteminin belirli özellikleri ÇYRÖM araştırmaları için özel bir bağlam oluşturmaktadır. Merkezi sınav kültürünün baskınlığı nedeniyle performans değerlendirme pratiği görece geç gelişmiş; ancak Türkiye Yüzyılı Maarif Modeli (Millî Eğitim Bakanlığı

[MEB], 2024) ile beceri odaklı ve süreç temelli değerlendirme yaklaşımlarının öğretim programlarına sistematik biçimde girmesi, rubrik temelli değerlendirmelerde psikometrik kalite güvencesini güncel bir araştırma önceliğine dönüştürmüştür.

TARTIŞMA, SONUÇ VE ÖNERİLER

Tartışma

Bu derleme makale, Çok Yüzeyli Rasch Ölçme Modeli'ni kuramsal temelleri, matematiksel altyapısı, temel analiz parametreleri, puanlayıcı etkileri ve sınırlılıkları bakımından kapsamlı biçimde incelemiştir. Elde edilen sentez, performans değerlendirme araştırmaları açısından birkaç temel sonucu öne çıkarmaktadır.

İlk olarak, performans değerlendirmelerde ham puanların öğrenci yeterliğini doğrudan temsil ettiği varsayımı epistemolojik olarak savunulamaz görünmektedir. Linacre (1989) ve Engelhard (1994) tarafından ortaya konulduğu üzere, puanlayıcı katılık farklılıkları zaman zaman gerçek öğrenci yeterlik farklılıklarına yaklaşan büyüklüklere ulaşmaktadır. Bu saptama, puanlama güvenirliği sorununu salt teknik bir ölçme meselesinden çıkararak eğitimde fırsat eşitliği ve adalet boyutuna taşımaktadır. ÇYRÖM, bu sorunu ampirik olarak görünür kılma ve düzeltme kapasitesine sahip nadir yaklaşımlardan biridir.

İkinci olarak, ÇYRÖM'ün tanılayıcı işlevi betimleyici çıktıların çok ötesine geçmektedir. Hangi puanlayıcının eğitime ihtiyaç duyduğu, rubriğin hangi ölçütünün yeniden yapılandırılması gerektiği, hangi görevin yapısal güçlük bakımından beklenmeyen biçimde çalıştığı ve hangi öğrencinin adil olmayan bir değerlendirme koşulunda puanlandığı gibi eylem odaklı sorulara yanıt üretme kapasitesi, modeli yalnızca bir psikometri aracı olmaktan çıkarmakta; ölçme kalitesini sistematik olarak geliştirmeye yönelik bir araştırma-uygulama köprüsüne dönüştürmektedir.

Üçüncü olarak, ÇYRÖM ve G Kuramı'nın tamamlayıcı biçimde kullanılması, ölçme sürecine ilişkin daha bütüncül ve daha güçlü bir kanıt tabanı oluşturabilmektedir. G Kuramı'nın varyans bileşeni perspektifi —hangi yüzey toplam hataya en çok katkıda bulunuyor?— ile ÇYRÖM'ün bireysel parametre tanılaması —tam olarak hangi puanlayıcı ne kadar katı?— bir arada sunulduğunda araştırmacılar ve uygulayıcılar için daha güçlü ve işlevsel bir bilgi zemini ortaya çıkmaktadır.

Dördüncü olarak, ÇYRÖM'ün önemi yalnızca araştırma düzeyiyle sınırlı değildir. Yazma öğretimi, dil sınavları, öğretmen yeterliliği değerlendirmeleri ve Türkiye'nin yeni öğretim programlarında yer alan süreç temelli değerlendirme anlayışları bağlamında ÇYRÖM'ün uygulamaya aktarılması; öğrencilerin daha adil değerlendirilmesine, öğretmenlerin daha nesnel dönüt almasına ve eğitim yöneticilerinin kanıta dayalı kararlar almasına zemin hazırlayabilmektedir.

Son olarak, alanyazında giderek yaygınlaşan yapay zekâ tabanlı otomatik puanlama sistemleri karşısında ÇYRÖM'ün yeni bir rol üstlenebileceği görülmektedir. İnsan puanlayıcı ile yapay zekâ tabanlı puanlayıcı arasındaki katılık-cömertlik farklılıklarını, DRF örüntülerini ve kategori işleyişini karşılaştıran çalışmalar; hem geleneksel puanlama güvenirliğine hem de yapay zekâ değerlendirme sistemlerinin doğrulanmasına yeni bir metodolojik pencere açmaktadır.

Sonuç

Sonuç olarak, insan yargısına dayalı her türlü değerlendirme bağlamında —ilkokul yazma değerlendirmesinden yabancı dil sertifika sınavlarına, öğretmen adayı değerlendirmesinden proje puanlamalarına kadar— ÇYRÖM, ölçme sonuçlarının geçerliği, güvenirliği ve adaleti açısından vazgeçilmez bir yöntemsel çerçeve sunmaktadır. Eğitim araştırmacılarının ve ölçme uzmanlarının bu modeli yalnızca teknik bir analiz prosedürü olarak değil, eğitimde adil ve bilimsel değerlendirme kültürünün inşasına katkıda bulunan kapsamlı bir çerçeve olarak değerlendirmeleri önerilmektedir.

Öneriler

Gelecek araştırmalar için önerilen öncelikli yönelimler şunlardır: (a) ilkokul yazma değerlendirmelerinde ÇYRÖM ile rubrik işleyişinin ve puanlayıcı güvenirliğinin sınanması; (b) ortaöğretim yabancı dil sözlü beceri sınavlarında DRF analizleri; (c) öğretmen puanlayıcı eğitim programlarının etkililiğinin ÇYRÖM parametreleri üzerinden boylamsal izlenmesi; (d) G Kuramı ve ÇYRÖM'ün aynı veri setine birlikte uygulandığı karşılaştırmalı metodoloji çalışmaları; (e) yapay zekâ tabanlı otomatik puanlama sistemlerinin (örneğin büyük dil modeli destekli değerlendirme) insan puanlayıcıyla karşılaştırıldığı ÇYRÖM çalışmaları.

Uygulama düzeyinde, performans değerlendirmelerinin planlanması aşamasında puanlayıcılar arası bağlantıyı güvence altına alan desenler kurulmalı; puanlayıcı eğitimi yalnızca başlangıçta değil, ara kalibrasyon oturumlarıyla süreç boyunca sürdürülmelidir. Rubrik kategorilerinin işleyişi pilot uygulamalarla sınanmalı, ham puanların yanı sıra puanlayıcı ve görev etkilerinden arındırılmış ölçüler raporlanmalı ve yüksek riskli kararlarda ÇYRÖM bulguları geçerliğin diğer kanıtlarıyla birlikte değerlendirilmelidir.

KAYNAKÇA

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

Baird, J.-A., Andrich, D., Hopfenbeck, T. N., & Stobart, G. (2017). Assessment and learning: Fields apart? *Assessment in Education: Principles, Policy & Practice, 24*(3), 317–350. https://doi.org/10.1080/0969594X.2017.1319337

Bond, T. G., & Fox, C. M. (2015). *Applying the Rasch model: Fundamental measurement in the human sciences* (3rd ed.). Routledge.

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

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*(1), 37–46. https://doi.org/10.1177/001316446002000104

Cohen, J. (1968). Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin, 70*(4), 213–220. https://doi.org/10.1037/h0026256

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles*. Wiley.

Crocker, L., & Algina, J. (2008). *Introduction to classical and modern test theory* (2nd ed.). Cengage Learning.

Eckes, T. (2005). Examining rater effects in TestDaF writing and speaking performance assessments: A many-facet Rasch analysis. *Language Assessment Quarterly, 2*(3), 197–221. https://doi.org/10.1207/s15434311laq0203_2

Eckes, T. (2008). Rater types in writing performance assessments: A classification approach to rater variability. *Language Testing, 25*(2), 155–185. https://doi.org/10.1177/0265532207086780

Eckes, T. (2011). *Introduction to many-facet Rasch measurement: Analyzing and evaluating rater-mediated assessments*. Peter Lang.

Engelhard, G., Jr. (1994). Examining rater errors in the assessment of written composition with a many-faceted Rasch model. *Journal of Educational Measurement, 31*(2), 93–112. https://doi.org/10.1111/j.1745-3984.1994.tb00436.x

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

Hoyt, W. T. (2000). Rater bias in psychological research: When is it a problem and what can we do about it? *Psychological Methods, 5*(1), 64–86. https://doi.org/10.1037/1082-989X.5.1.64

Linacre, J. M. (1989). *Many-facet Rasch measurement*. MESA Press.

Linacre, J. M. (1994). *Many-facet Rasch measurement* (2nd ed.). MESA Press.

Linacre, J. M. (2024). *Facets computer program for many-facet Rasch measurement* (Version 3.87.0) [Computer software]. Winsteps.com. https://www.winsteps.com

Millî Eğitim Bakanlığı. (2024). *Türkiye Yüzyılı Maarif Modeli İlkokul Türkçe Dersi Öğretim Programı: 1, 2, 3 ve 4. sınıflar*. Millî Eğitim Bakanlığı.

Murphy, K. R., Jako, R. A., & Anhalt, R. L. (1993). Nature and consequences of halo error: A critical analysis. *Journal of Applied Psychology, 78*(2), 218–225. https://doi.org/10.1037/0021-9010.78.2.218

Myford, C. M., & Wolfe, E. W. (2003). Detecting and measuring rater effects using many-facet Rasch measurement: Part I. *Journal of Applied Measurement, 4*(4), 386–422.

Myford, C. M., & Wolfe, E. W. (2004). Detecting and measuring rater effects using many-facet Rasch measurement: Part II. *Journal of Applied Measurement, 5*(2), 189–227.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Danish Institute for Educational Research.

Robitzsch, A., Kiefer, T., & Wu, M. (2022). *TAM: Test analysis modules* (R package version 4.1-4). https://CRAN.R-project.org/package=TAM

Saal, F. E., Downey, R. G., & Lahey, M. A. (1980). Rating the ratings: Assessing the psychometric quality of rating data. *Psychological Bulletin, 88*(2), 413–428. https://doi.org/10.1037/0033-2909.88.2.413

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

Stemler, S. E. (2004). A comparison of consensus, consistency, and measurement approaches to estimating interrater reliability. *Practical Assessment, Research & Evaluation, 9*(4), 1–19. https://doi.org/10.7275/96jp-xz07

Thorndike, E. L. (1920). A constant error in psychological ratings. *Journal of Applied Psychology, 4*(1), 25–29. https://doi.org/10.1037/h0071663

Weigle, S. C. (2002). *Assessing writing*. Cambridge University Press.

Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis: Rasch measurement*. MESA Press.