Eğitim ve Bilim

Orijinal Makale

2025 Ek Sayı 1 93-110

Karşılaştırma Yargılarıyla Puanlamada Uyarlamalı Eşleme ve Standart Hata Sonlandırma Kuralı: Yazma Becerisinin Ölçülmesine Yönelik bir Uygulama *

Sungur Gürel [1], Murat Doğan Şahin [2], İbrahim Uysal [3], Ali İhsan İbileme [4], Tuba Gündüz [5]

Öz

Mevcut çalışmanın amacı örneklem büyüklüğü ve standart hata sonlandırma kuralı koşulları altında karşılaştırma yargılarıyla puanlamanın güvenirliğini incelemektir. Bu amaç doğrultusunda 250, 500 ve 1000 örneklem büyüklükleri ile 0.40, 0.35 ve 0.30'luk standart hata sonlandırma kuralları çaprazlanarak 9 koşullu, 82 tekrarlı bir Monte Carlo simülasyonu gerçekleştirilmiştir. Ayrıca, 50 öğrencilik bir örneklemle 0.40 standart hata durdurma kuralı ve 40 en fazla karşılaştırma sayısı kullanılarak gerçek bir uygulama yapılmıştır. Simülasyon çalışmasında puanlama güvenirliği gerçek güvenirlik, sıralama güvenirliği ve ölçek ayrıştırma güvenirliği ile belirlenmiştir. Gerçek veri ile yapılan uygulamada ise bütüncül ve analitik puanlama ile karşılaştırma yargılarıyla puanlama arasındaki korelasyon incelenmiş ve karşılaştırma yargılarıyla puanlama için ölçek ayrıştırma güvenirliği hesaplanmıştır. Simülasyon sonuçları tüm koşullarda yüksek düzeyde puanlama güvenirliği göstermiştir. Dahası puanlama güvenirliği örneklem büyüklüğünden bağımsız bulunmuştur. Daha katı standart hata sonlandırma kurallarının, daha yüksek güvenirlik düzeyleri sağladığı, ancak bunun için performansların daha yüksek sayıda ikili karşılaştırmaya tabi tutulması gerektiği sonucuna ulaşılmıştır. Gerçek uygulama sonuçları, 0.89'luk yüksek ölçek ayrıştırma güvenirliği göstermiş ve rubrik kullanılarak verilen puanlarla 0.70'in üzerinde bir korelasyon ortaya koymuştur. Genel olarak araştırma sonuçları, karşılaştırma yargılarıyla puanlamanın hem sınıf içi hem de geniş ölçekli uygulamalarda kullanılabilir olduğunu göstermektedir. Ayrıca, karşılaştırma yargılarıyla puanlamanın daha kolay uygulanması, test sürecinde bir farklılık gerektirmemesi ve puanları sürekli bir ölçek üzerine yerleştirmesi nedeniyle avantaj sağladığı düşünülmektedir.

Anahtar Kelimeler

Karşılaştırma yargıları
Bütünsel değerlendirme
Ölçek ayrıştırma güvenirliği
İkili karşılaştırma

Makale Hakkında

Gönderim Tarihi: 06.10.2024
Kabul Tarihi: 07.01.2025
Elektronik Yayın Tarihi: 03.03.2025

DOI: 10.15390/EB.2025.14123

* Bu çalışmanın bir bölümü 4-6 Ekim 2024 tarihleri arasında düzenlenen Uluslararası Ölçme, Seçme ve Yerleştirme Sempozyumu'nda sözlü bildiri olarak sunulmuştur.

[1] Siirt Üniversitesi, Eğitim Fakültesi, Eğitim Bilimleri Bölümü, Türkiye, s.gurel@siirt.edu.tr
[2] Anadolu Üniversitesi, Eğitim Fakültesi, Eğitim Bilimleri Bölümü, Türkiye, muratdogansahin@gmail.com
[3] Bolu Abant İzzet Baysal Üniversitesi, Eğitim Fakültesi, Eğitim Bilimleri Bölümü, Türkiye, ibrahimuysal@ibu.edu.tr
[4] Eskişehir Teknik Üniversitesi, Sürekli Eğitim Uygulama ve Araştırma Merkezi, Türkiye, aiibileme@gmail.com
[5] Muğla Sıtkı Koçman Üniversitesi, Eğitim Fakültesi, Eğitim Bilimleri Bölümü, Türkiye, tubagunduz@mu.edu.tr

Giriş

Bireylerin bir performans sonucu ortaya koydukları ürünün değerlendirilmesinde, puanlama güvenirliğini sağlamak amacıyla rubriklerden yararlanılır. Tek puanlayıcıdan kaynaklanabilecek yanlılıkların önüne geçmek amacıyla en sık kullanılan yöntem, performansın üç ile beş kategoriden oluşan rubriklerle en az iki puanlayıcı tarafından bağımsız olarak puanlanmasıdır ve iki puanlama arasında yüksek bir uyum olması beklenir. En fazla beş kategorili rubriklerden yararlanılmasındaki temel amaç, puanlayıcılar arasındaki uyumu mümkün olduğunca yüksek tutmaktır. Öyle ki kategori sayısı arttığında uyumsuzluk düzeyi de artabilir (Goossens ve De Maeyer, 2018). Bu durum, öğrenciden istenilen performansın sınırlı sayıda kategorisi olan rubriğe uygun biçimde yapılandırılmasını, bir başka ifadeyle geçerlikten ziyade güvenirliğin ön plana çıkarılmasını berberinde getirir (van Daal, Lesterhuis, Coertjens, Donche ve De Maeyer, 2019).

Özellikle üst düzey becerilerin ölçülmesinde kullanılan rubriklerin birden fazla bileşeni olması, bir başka ifadeyle analitik rubrik kullanılması, her bir bileşenin ayrıca puanlanmasını gerektirir. Bu durum, değerlendirmeye ayrılan süreyi artırmaktadır. Bir diğer önemli husus ise, yaratıcılık gibi becerilere odaklanan ürünlerin de değerlendirilmesinde, değerlendirmeye esas olan ürünün rubriğin parçalarının toplamından farklı bir hüviyet göstermesidir; bir başka ifadeyle ölçülen özellik, rubrikte yer bulan parçaların toplamından farklıdır (Jones ve Davies, 2023). Bu problemin önüne geçmek amacıyla analitik yerine bütünsel bir puanlama yapılması durumunda ise, bireylerin performanslarındaki farklılıkları ortaya koyabilecek hassasiyette ölçüm yapılması mümkün olmamaktadır. Bu ikilem, ölçülen performans açısından bireyler arasındaki farklılıkların ortaya konulması gereken durumlarda önemli bir sınırlılığa neden olmaktadır.

İki puanlayıcıya dayalı geleneksel sistemlerdeki bir diğer sınırlılık ise, uzman puanlayıcılara olan ihtiyaçtır. Buradaki uzmanlıktan kasıt, sadece alandaki yetkinlik ve deneyim değil, puanlama sürecindeki standardizasyonu sağlamadır. Uzman puanlayıcı sayısının sınırlı olduğu durumlarda, nispeten daha az uzmanlık gerektiren bir yaklaşıma dayalı karar verilebilen bir puanlama yaklaşımına ihtiyaç duyulduğu söylenebilir.

Her ne kadar bir rubrik yardımıyla performansın mutlak bir yetkinlik düzeyi referans alınarak ölçülmesi mümkün görünse de puanlayıcıların daha önce yaptıkları puanlamalardan etkilendikleri ve bu durumun puanlama performanslarına yansıdığı bilinmektedir (Bloxham, 2009; Crisp, 2013). Bu durum, deneysel psikolog Laming'in (2003), "Mutlak bir yargı yoktur. Tüm yargılar, bir şeyle diğerinin karşılaştırmalarıdır." savını desteklemektedir. Laming burada, her ne kadar ölçülecek özelliğin doğası ve ölçme işleminin amacı gereği mutlak bir değerlendirme yapılmak istense de bunun, puanlama davranışının doğası gereği mümkün olmadığını belirtmektedir. Performansın bütünsel olarak değerlendirilmesinde rubriğe dayalı geleneksel yöntemlerin kullanımına ilişkin sözü geçen tüm bu sınırlılıkların, karşılaştırma yargılarıyla puanlama (KYP) ile en aza indirilmesi mümkün görünmektedir.

KYP, temelde Thurstone'un karşılaştırma yargıları ilkesine dayanır; puanlayıcının önüne iki farklı bireyin performansı çıkar, puanlayıcıdan beklenen ise yalnızca hangi performansın bütünsel olarak diğerinden daha iyi olduğuna karar vermesidir. Bu biçimde her bir performans çok sayıda ikili karşılaştırmaya tabi tutulur. Yetenek kestirimi ise, Rasch modele çok benzeyen Bradley-Terry-Luce (BTL) modeli (Bradley ve Terry 1952; Luce 1959) ile gerçekleştirilir. KYP'nin temel dayanağı, bireylerin göreli yargıda bulunmalarının, mutlak yargıya varmalarından daha kolay olduğu yönündedir (Bramley, 2005). Bunların yanında iki puanlayıcı iki farklı performansın puanlamasında uzlaşmayabilirler; ancak hangisinin "daha iyi" olduğu konusunda uzlaşmaları muhtemeldir (Bramley, 2007; Steedle ve Ferrara, 2016). Öyle ki ikili karşılaştırma yöntemi kullanılarak yapılan değerlendirmede değerlendiricilere yüzeysel bir eğitimin verilmesi yeterli görülmektedir (Heldsinger ve Humphry, 2013).

KYP, temel olarak getirdiği avantajlarla öncelikle K-12 düzeyinde, öğretmenlerin sınıf içinde yaptığı performans değerlendirme uygulamalarında kullanımıyla kendine yer bulmuştur (Steedle ve Ferrara, 2016). Özellikle öğretmenlerin yaptığı puanlamaların güvenirliğine yönelik araştırmaların artması ve bu çalışmalarda çoğunlukla öğretmen puanlamalarının değişken olduğunun görülmesi (Humphry ve Heldsinger, 2019), KYP'nin rubriğe dayalı puanlamalara göre çok daha kolay ve dolayısıyla yüksek güvenirlikli sonuçlar veren uygulamalarının sayısında artışı da beraberinde getirmiştir. Bu konuda özellikle nomoremarking.com, temel eğitim düzeyinde yazma becerisinin boylamsal olarak ölçülmesi amacıyla Birleşik Krallık kapsamında beş yıldır yaklaşık 2000 okulda öğretmenlerin karşılaştırmalı yargılarına dayalı olarak yaptığı çalışmalarla öne çıkmaktadır (Christodoulou, 2024). Bir başka ifadeyle son yıllarda gerek sınıf içinde biçimlendirici ve değer biçmeye dönük uygulamalarda, gerekse geniş ölçekli uygulamalarda KYP kendine daha fazla uygulama alanı bulmaktadır.

Türkiye'deki ölçme ve değerlendirme uygulamalarında, yerleştirmeye dönük ulusal sınavların çoktan seçmeli maddelerle yapılıyor olmasının her kademede çoktan seçmeli madde kullanımını artırdığı söylenebilir. Ancak T.C. Millî Eğitim Bakanlığı (MEB), Türkiye Yüzyılı Maarif Modeli çerçevesinde oluşturduğu yeni öğretim programlarıyla üretimsel dil becerilerini de ön plana çıkararak öğretmenleri çoktan seçmeli yerine açık uçlu madde kullanımına yönlendiren bir yönetmelik yayımlamıştır (MEB Ölçme ve Değerlendirme Yönetmeliği, 2023). Bu durumun sonucunda K-12 düzeyinde açık uçlu maddelerin kullanımında bir artış beklenmektedir. Bu yönetmelikle birlikte sınıf içi bu kullanımın yanında orta öğretim düzeyindeki yazılı sınavların bir kısmı da ülke çapında ortak yazılı sınavlar olarak yapılandırılmıştır. Bu sürecin devamında, MEB tarafından gerçekleştirilen, yüksek önem arz eden yerleştirmeye dönük ulusal sınavlarda da açık uçlu maddelerin kullanılması beklenebilir. Yüzbinlerle ifade edilecek geniş kitlelere yapılacak sınavlarda uzman puanlayıcı sayısının yeterli düzeye ulaşması mümkün olmayacağından, puanlama güvenirliğine yönelik kayda değer endişeler ortaya çıkabilecektir. KYP'nin, gerek sınıf içi uygulamalarda gerekse geniş ölçekli yüksek önem arz eden sınavlarda uzman olmayan puanlayıcılarla da yüksek güvenirlik değerlerini mümkün kılması, Türkiye'de yapılacak bu uygulamalarda yaşanabilecek olası sorunları bertaraf etmesi açısından önemli bir fırsat sunmaktadır.

Karşılaştırma Yargılarıyla Puanlamanın Gelişimi

KYP uygulamasında temel bileşenler performansların eşlenmesi, yetenek kestirimi, güvenirlik ve sonlandırma kuralı olmak üzere dört kısımda incelenebilir. Her ne kadar ilk adım performansların eşlenmesi olsa da modelin gelişiminde kronolojik olarak yetenek kestirimi üzerindeki çalışmalar önde olduğundan, bu bileşenlerin tanıtımına da yetenek kestirimiyle başlanmıştır.

Yetenek Kestirimi

Thurstone'un (1927) karşılaştırmalı yargı yasası, gizil bir ölçek üzerinde nesneler arasındaki mesafeleri tahmin etmeye yönelik ilk yöntemi sunmuştur. Sonrasında, Bradley ve Terry (1952) ile Luce (1959), karşılaştırmalı yargı verilerinin analizinde lojistik fonksiyonların nasıl uygulanabileceğini göstermiş, Andrich (1978) ise Thurstone'un modelinin Rasch lojistik modeliyle örtüştüğünü ortaya koymuştur:

$$P_k(a_j) = p(X_{jk} = 1|a_j, a_k)\frac{\exp{(a_j - a_k)}}{1 + \exp{(a_j - a_k)}}$$

$P_k(a_j)$ veya $p(X_{jk} = 1|a_j, a_k)$ ifadesi, j performansının k performansına tercih edilme olasılığını belirtir. Bu durumda, $X_{jk} = 1$ ifadesi, j performansının k performansına tercih edildiği anlamına gelir. Burada a_j ve a_k, sırasıyla j ve k performanslarına ait yetenek puanlarını logit biriminde temsil eder.

KYP uygulamaları mutlak bir değerlendirmenin insanın doğası gereği mümkün olmayacağı temelinde, göreli bir değerlendirmeyi esas almaktadır. Bu durum, araştırmacıların aklına mutlak performans beklentilerinin olduğu durumlarda bu yöntemin nasıl kullanılabileceği konusunu getirebilir. Burada yapılan puanlamaları mutlak bir ölçek üzerine yerleştirmek amacıyla çapa performanslardan yararlanılması söz konusudur (Heldsinger ve Humphry, 2013; Using anchors to link judging sessions, 2016). Buna göre bir uzman grup tarafından mutlak olarak puanlanmış performanslar

da KYP sürecine dâhil edilirler. Bu yolla puanlamaların hem mutlak bir ölçek üzerine yerleştirilmesi hem de farklı zamanlarda aynı amaçla yapılan uygulamalarda bireylerin aynı ölçek üzerinde eşitleme yapılarak değerlendirilebilmesi mümkün olabilmektedir (Benton, 2021).

Performansların Eşleştirilmesi

KYP üzerine yapılan çalışmaların son yıllarda arttığı görülmektedir (Benton, 2021; Bramley ve Vitello, 2019; Crompvoets, Béguin ve Sijtsma, 2020; Crompvoets, Béguin ve Sijtsma, 2022; Holmes, Meadows, Stockford ve He, 2018; Humphry ve Heldsinger; 2019; Lesterhuis, Bouwer, Van Daal, Donche ve De Maeyer, 2022; van Daal vd., 2019; Verhavert, Bouwer, Donche ve De Maeyer, 2019; Verhavert, Furlong ve Bouwer, 2022). Bunların arasında özellikle Pollit'in (2012) yaptığı çalışma son derece önemlidir. Bu çalışma performans çiftlerinin rassal olarak oluşması yerine performansların İsviçre Sistemi Satranç Eşlemesi adı verilen bir yöntemle eşlenmesi fikrini ortaya atmış ve bu eşleme işlemiyle yönteminin adı uyarlamalı karşılaştırma yargılarıyla puanlama (UKYP) adını almıştır. Ancak bu eşleme yöntemi, KYP uygulamalarındaki güvenirlik ölçüsü olan ölçek ayrıştırma güvenirliğinin gerçekçi olmayan şekilde yüksek kestirilmesine yol açtığı iddiasıyla eleştirilmektedir (Bramley, 2005; Bramley ve Vitello, 2019). KYP'de ölçme kesinliği için yüksek ölçek ayrıştırma güvenirliğine düşük sayıda karşılaştırma sayısıyla ulaşılması gerektiği düşünüldüğünde, eşlemede uyarlamayı sağlamanın bir gereklilik olduğu söylenebilir. Bu durum, Pollit'in önerisi dışında yeni uyarlamalı yöntemlerin ortaya çıkmasını gerektirmiştir.

UKYP'deki süreç, tipik bir bireyselleştirilmiş bilgisayarlı test (BBT) uygulamasına benzetilebilir (Crompvoets vd., 2020). Bu durum, özellikle UKYP'de karşılaştırılacak performans çiftlerini eşlemenin uyarlamalı olmasından kaynaklanmaktadır. BBT'deki madde seçim algoritmasına benzer biçimde, UKYP'de performansların eşlenmesi Fisher bilgi fonksiyonu kullanılarak yapılır (Pollit, 2012). Bununla birlikte Crompvoets ve diğerlerinin (2020) önerdiği metotta, her geçici yetenek kestirimi sonrası havuzdaki performanslardan standart hatası en yüksek olan performans belirlenir ve karşısına gelecek eş, bir olasılık yoğunluk fonksiyonuna göre seçilir. $\theta_i \sim N[\theta_i, \mathrm{SE}(\theta_i)]$ olmak üzere (SE kestirilen standart hatayı, θ_i ise kestirilen başarı seviyesini temsil eder) mümkün olan her bir j nesnesinin yoğunluk değerinin eşlenmesi mümkün her bir j nesnesinin toplam yoğunluk değerine bölünmesiyle olasılıklar elde edilir. Olasılık yoğunluk fonksiyonuna göre eşleştirme yapıldığında karşıya gelecek performans, ilk performansa başarı seviyesi bakımında benzer ancak standart hatası yüksek bir performans olacaktır. Çalışma kapsamında, bu uyarlamalı eşleme metodundan yararlanılmıştır.

Güvenirlik

KYP'de güvenirlik, ölçek ayrıştırma güvenirliği (İng. scale seperation reliability-SSR) ile hesaplanır. SSR değerlendiricilerin performansların düzeylerine ilişkin hem fikir olma düzeylerine yönelik bilgi verir (Verhavert vd., 2019). Güvenirlik düzeyinin göstergesi olan SSR, Andrich ve Douglas (1977) tarafından şu şekilde formüle edilmiştir (Aktaran Gustafsson, 1977);

$$\mathrm{SSR} = \frac{\sigma_a^2 - MSE}{\sigma_a^2}$$

yukarıdaki formülde σ_a^2 yetenek kestirimlerinin varyansını, MSE ise standart hataların karelerinin aritmetik ortalamasını temsil etmek üzere:

$$MSE = \frac{\sum_j^n se_{aj}^2}{n}$$

(se standart hatayı temsil etmek üzere) biçiminde hesaplanır.

Sonlandırma Kuralı

KYP'de bir diğer önemli bileşen ise sonlandırmadır. Yapılan çalışmalar incelendiğinde, çoğunlukla sabit bir karşılaştırma sayısı ya da ortalama karşılaştırma sayısına dayalı bir sonlandırma kuralının kullanıldığı görülmektedir (Crompvoets vd., 2020; Lesterhuis vd., 2022; Pollit, 2012; Sims, Cox, Eckstein, Hartshorn, Wilcox ve Hart, 2020; Thwaites, Kollias ve Paquot, 2024). Bununla birlikte

farklı yetenek düzeylerindeki performansların eşit sayıda karşılaştırmaya tabi tutulmaları kestirilen yeteneklerin hata düzeylerinin farklılaşmasını beraberinde getirmektedir; bu durumda yetenek dağılımının iki ucuna gidildikçe kestirimlerin standart hatası artmaktadır (Crompvoets vd., 2020; Uysal, Gürel, Şahin, İbileme ve Yıldırım Görgülü, 2024). Bu duruma bir çözüm bulmak amacıyla Verhavert ve diğerleri (2022), sonlandırma kuralı olarak SSR'yi kullanmışlardır; buna göre önceden belirlenen bir güvenirlik değerine ulaşıldığında karşılaştırma işlemi sonlandırılmaktadır. Ancak, yüksek bir SSR'nin sağlanması da tüm kestirimler için belirli bir standart hata değerini garanti etmez. Bu nedenle, tüm performanslar için kestirilen yetenek düzeylerinin standart hatalarının önceden belirlenen bir düzeyde olmasını sağlamanın, bir başka ifadeyle BBT uygulamalarına benzer olarak, sonlandırma kuralının yetenek kestiriminin standart hatasına dayalı olarak yapılmasının daha isabetli olacağı değerlendirilmektedir.

Mevcut Çalışma

Alanyazındaki çalışmaların bir kısmı performansların eşlenmesinde yarı rassal eşlemeyi kullanırken, son yıllarda uyarlamalı eşlemenin kullanıldığı çalışmaların ön plana çıktığı görülmektedir (Crompvoets vd., 2020; Crompvoets vd., 2022; Holmes vd., 2018; Lesterhuis vd., 2022; Pollit, 2012; Sims vd., 2020; Thwaites vd., 2024; van Daal vd, 2019; Verhavert vd., 2019; Verhavert vd., 2022). Bu çalışmalarda performansların ait olduğu alan esas olmak üzere farklı karşılaştırma sayılarıyla 0.70 ile 0.95 arasında SSR değerleri elde edilmiştir. Bununla beraber sonlandırma kuralı genel olarak kâğıt başına düşen ortalama karşılaştırma sayısına dayanmaktadır; yukarıda belirtildiği üzere SSR'ye dayalı sonlandırmayı esas alan yeni çalışmalar da yapılmıştır. Ancak bu sonlandırma yaklaşımları ile bazı performanslar için düşük standart hatalar elde edilirken bazıları için yüksek standart hatalı kestirimler elde edilebilmektedir. Her kestirimin standart hatası için belirli bir sınır değerin belirlenebilmesi, aynı BBT uygulamalarında olduğu gibi standart hataya dayalı bir sonlandırma kuralının kullanılmasıyla mümkün olacaktır. Ancak alanyazında, UKYP çalışmalarında standart hataya dayalı sonlandırma kuralının kullanıldığı bir çalışma yer almamaktadır. Bu çalışma ile uyarlamalı eşleme ve standart hata sonlandırma kuralına dayalı olarak oluşturulan bir algoritmanın performansı ilk kez ortaya konmuştur. Bu amaç doğrultusunda araştırmanın ilk kısmında farklı örneklem büyüklükleri için üretilmiş veri setleri kullanılarak, farklı standart hata değerleri sonlandırma kuralı olarak kullanıldığında ortalama karşılaştırma sayısı, gerçek güvenirlik, sıralama güvenirliği ve SSR değerlerinin değişimi incelenmiştir. UKYP çalışmaları göz önüne alındığında orta (250) ve geniş (500 ve 1000) olarak değerlendirilebilecek üç farklı örneklem büyüklüğünde 0.30, 0.35 ve 0.40 standart hataya dayalı sonlandırma kuralının performanslarını karşılaştıran bu Monte Carlo simülasyon çalışmasının sonuçlarının gerçek veriye dayalı uygulamalarda araştırmacılara yol göstereceği değerlendirilmektedir. Standart hata değerleri olarak 0.30, 0.35 ve 0.40'ın tercih edilmesi, bu değerlerin BBT uygulamalarındaki standart hata sonlandırma kuralı çerçevesinde en çok tercih edilen değerler olmasından ileri gelmektedir. Elbette, standart hata değeri düştükçe güvenirliğin yükseleceği öngörülebilir; ancak bunun bir sonucu olarak karşılaştırma sayısında önemli bir artış meydana geleceği göz ardı edilmemelidir. Bu nedenle bu araştırmanın temel amacı, en uygun standart hata değerini bularak gerçek uygulamada ortaya çıkacak karşılaştırma sayısının makul olduğu bir değere ulaşmak, bir başka ifadeyle ölçme kesinliğine odaklanmaktadır. Standart hata ve SSR için çok iyi değerler elde edilirken, ortalama karşılaştırma sayısında dramatik artışların ortaya çıkmasının, gerçek uygulamalarda kayda değer bir kullanışlılık problemi yaratacağı unutulmamalıdır. Bu nedenle araştırmanın ikinci kısmı olan gerçek uygulamada 0.40 standart hata değeri sonlandırma kuralı için tercih edilmiştir. Bununla birlikte yine BBT uygulamalarında olduğu gibi en çok karşılaştırma sayısı belirlenmiştir. Buna göre 40 karşılaştırma yapılması durumunda 0.40 standart hatanın yakalanamaması durumunda dahi ilgili kâğıt için karşılaştırma süreci sonlandırılmıştır. En fazla 40 karşılaştırma değeri, simülasyon sonuçlarına göre belirlenmiştir.

Yöntem

Araştırma simülasyon çalışması ve gerçek uygulama olmak üzere iki kısımdan oluşmaktadır. Aşağıda simülasyon çalışması ve gerçek uygulama sırasıyla açıklanmaktadır.

Simülasyon Çalışması

Bir Monte Carlo simülasyon çalışması olarak gerçekleştirilen ilk kısımda tamamen çaprazlanmış desen kullanılmıştır. Üç örneklem büyüklüğü (250, 500 ve 1000), sonlandırmada kullanılan üç standart hata değeri (0.40, 0.35 ve 0.30) olmak üzere toplam 2 faktör ve 9 koşul üzerinde çalışılmıştır. Mevcut araştırma geniş ölçekli uygulamalarda da karşılaştırmalı yargıyla puanlamayı incelediğinden örneklem büyüklüğüne 500 ve 1000 koşulu tanımlanmıştır. Bunun yanında 250 koşulu ise orta genişlikte örneklem büyüklüğünü temsil etmesi için ele alınmıştır. Steedle ve Ferrara (2016) araştırmasında 200 kişilik bir örneklem kullanırken Pollitt (2012) 1000 kişilik örneklemden yararlanmıştır. Çalışma bu örneklem büyüklükleri arasındaki değişimi incelemek üzere tasarlanmıştır.

Katılımcıların yetenekleri ortalaması 0, standart sapması 1 olan normal dağılım kullanılarak üretilmiştir. Üretilen yeteneklerin standart sapma değerleri değiştikçe standart hata kestirimleri değişkenlik göstermektedir. Mantıklı standart hatalar üretebilmek için Crompvoets, Béguin ve Sijtsma (2021) tarafından önerildiği üzere standart sapma 1 olarak alınmıştır. Uyarlamalı eşleme algoritması ve yetenek kestirimlerinin kodlanmasında Crompvoets ve diğerlerinin (2020) yaklaşımı referans alınmıştır. Sonlandırma kuralının kodlanması ise özgündür. Veriler Linux işletim sistemi üzerinde R yazılımı (R Core Team, 2023) kullanılarak üretilmiş ve analiz edilmiştir. Tübitak Türk Ulusal Bilim Altyapısı (TRUBA) üzerinde analizler gerçekleştirilmiş olup 56 çekirdekli bilgisayarlardan yararlanılmıştır. Bilgisayarda yer alan çekirdeklere görevler paralelleştirme işlemiyle dağıtılmıştır. Bu aşamada *doParallel* paketi (Daniel, Microsoft Corporation, Weston ve Tenenbaum, 2022) kullanılmıştır. TRUBA sistemine görevler gönderilirken *slurm*'dan yararlanılmıştır. Veri matrisinin büyüklüğünden dolayı işlemlerin oldukça uzun sürmesi nedeniyle TRUBA sistemine başvurulmuş olup TRUBA sistemi çalıştırılırken süre kısıtlamasına gidilmektedir. Bu nedenle araştırmada her bir koşul için 82 tekrar tamamlanabilmiştir.

Simülasyon çalışması sonucunda elde edilen bulgular değerlendirilirken her bir kâğıdın kaç kez karşılaştırmaya tabi tutulduğunu gösteren ortalama karşılaştırma sayısı, yukarıda açıklaması yapılan SSR, sıralama güvenirliği ve gerçek güvenirlik değerleri kullanılmıştır. Sıralama güvenirliği üretilen yetenekler ve kestirilen yetenekler arasındaki Spearman sıra farkları korelasyonu aracılığıyla hesaplanırken gerçek güvenirlik üretilen ve kestirilen yetenekler arasındaki Pearson momentler çarpımı korelasyon katsayısının karesi aracılığıyla hesaplanmaktadır.

Gerçek Uygulama

Gerçek uygulama tarama deseni altında gerçekleştirilmiştir. Bu kapsamda öğrencilerin yetenekleri karşılaştırmalı yargıyla puanlama aracılığıyla kestirilmiştir. Gerçek uygulamada ÖSYM Araştırma ve Geliştirme Daire Başkanlığı tarafından desteklenen "Açık Uçlu Maddelerin Uyarlamalı Karşılaştırmalı Yargıyla Puanlanması için bir Sistem ve Yazılım Geliştirme" adlı proje kapsamında araştırmacılar tarafından geliştirilen bir yazılım kullanılmıştır. Araştırma kapsamında ELLIPSE Corpus (Crossley vd., 2023) çalışmasında yer alan veri seti kullanılmıştır. ELLIPSE Corpus veri seti, bir otomatik puanlama çalışması kapsamında toplanarak CC BY-NC-SA 4.0 DEED uluslararası lisanslı açık kaynakla araştırmacıların bilimsel amaçlarla kullanımına açılmış bir veri setidir. Söz konusu veri setinde ABD'deki ikinci dilleri İngilizce olan 8. sınıf öğrencilerinin "teknolojinin insan hayatına olan etkilerine yönelik" yazdıkları kompozisyonlar ve bu kompozisyonların uzman puanlayıcılarca değerlendirmesiyle elde edilen puanlamalar yer almaktadır. Kompozisyonlar daha öncesinde iki puanlayıcı tarafından hem genel İngilizce dil yeterliği açısından bütüncül olarak hem de bağdaşıklık,

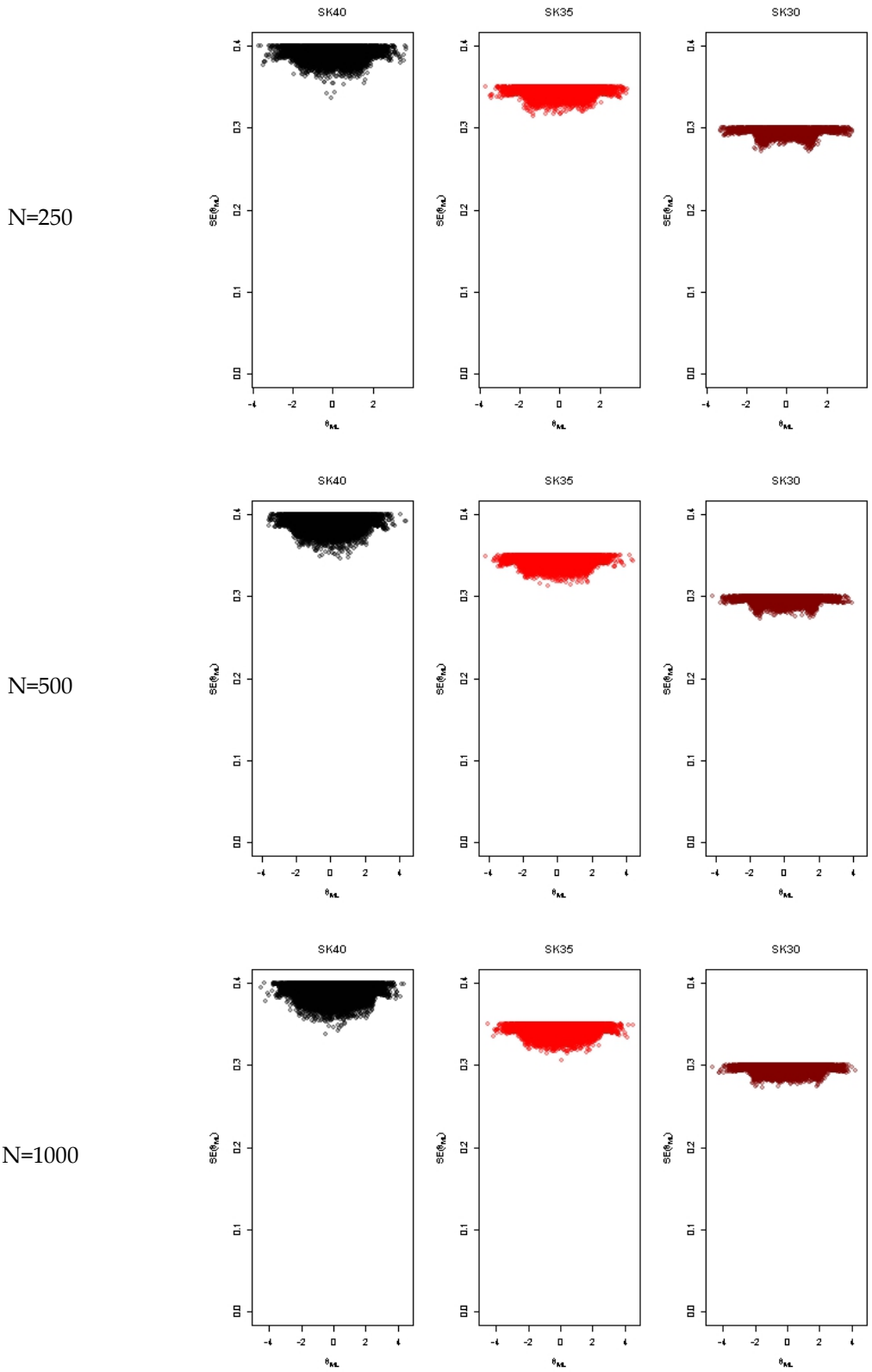
sözdizimi, kelime dağarcığı, ifade, dilbilgisi ve kurallar alt boyutlarıyla analitik olarak puanlanmıştır. Hem bütüncül olarak elde edilen puan hem de 6 alt boyutun toplamıyla analitik olarak elde edilen puanlar bu çalışmaya aktarılmıştır. İlgili başlıkta verilen 250 yazma performansından 50'si, standartlaştırılmış analitik toplam puanı mümkün olduğu kadar standart normal dağılımı takip edecek şekilde seçilmiştir. Standartlaştırılmış analitik toplam puanları -1.88 ile 2.63 aralığında olup ortalaması 0.10, standart sapması ise 0.96'dır. Bu seçimin gerekçesi, simülasyon çalışmasında öğrenci başarısının standart normal dağılım ile üretilmiş olmasıdır. 50 yanıtın İngilizce dil yeterliği bütüncül puanları ise 2 ile 5 aralığındadır. Bir öğrenci 2, yedi öğrenci 2.5, on üç öğrenci 3, on sekiz öğrenci 3.5, sekiz öğrenci 4, iki öğrenci 4.5 ve bir öğrenci 5 puan almıştır. 50 performansta yer alan toplam kelime sayısı 152 ile 1532 arasında olup ortancası 449 kelime, ortalaması 535 kelimedir. Kâğıtların dil kullanım becerileri açısından bütünsel olarak değerlendirilmesi amaçlandığından, karşılaştırmalı yargılarla puanlamaya uygundur. İlgili veri setine ilişkin tüm ayrıntılara https://github.com/scrosseye/ELLIPSE-Corpus adresinden ulaşılabilir.

Karşılaştırmalı yargıyla puanlama işlemi, bu araştırmanın yazarlarından dördü tarafından yapılmıştır. Puanlama yapan yazarlar C1 seviyesinde İngilizce yeterliğine sahip olup genel anlamda daha iyi yazma performansını seçme dışında bir puanlama yönergesi almamışlardır. Buradaki temel amaç, KYP'nin temel avantajlarından biri olarak ön plana çıkan alan uzmanı olmayan puanlayıcılarla dahi yüksek güvenirlik elde edilmesi beklentisini test etmeye yöneliktir. Gerçek uygulama sonucunda UKYP'nin performansını değerlendirmek amacıyla SSR ile birlikte UKYP ile elde edilen yetenek kestirimleriyle uzmanlarca puanlanan İngilizce dil yeterliği bütüncül puanları ve standartlaştırılmış analitik toplam puanları arasındaki korelasyon raporlanmıştır.

Bulgular

Simülasyon Çalışmasına Yönelik Bulgular

Karşılaştırma yargılarıyla puanlamada uyarlamalı eşleme tercih edildiğinde farklı örneklem büyüklüklerinde ve standart hata sonlandırma kuralları uygulandığında yetenek kestirimlerinin tutarlılığının belirlenmesi amaçlanan bu çalışmada öncelikli olarak yetenek kestirimleri ve standart hatalar arasındaki ilişki sunulmuştur. Yetenek kestirimlerinin tutarlılığını belirlemek için farklı simülasyon koşullarında elde edilen gerçek güvenirlik, sıralama güvenirliği ve SSR raporlanmıştır. Son olarak uygulamanın kullanışlılığını değerlendirmek amacıyla örneklem büyüklüğünde ve her bir sonlandırma kuralını sağlamak üzere ortalama kaç karşılaştırma yapıldığı raporlanmıştır.



Şekil 1. Farklı örneklem büyüklüklerinde yetenek kestirimleri ile standart hata arasındaki ilişki.

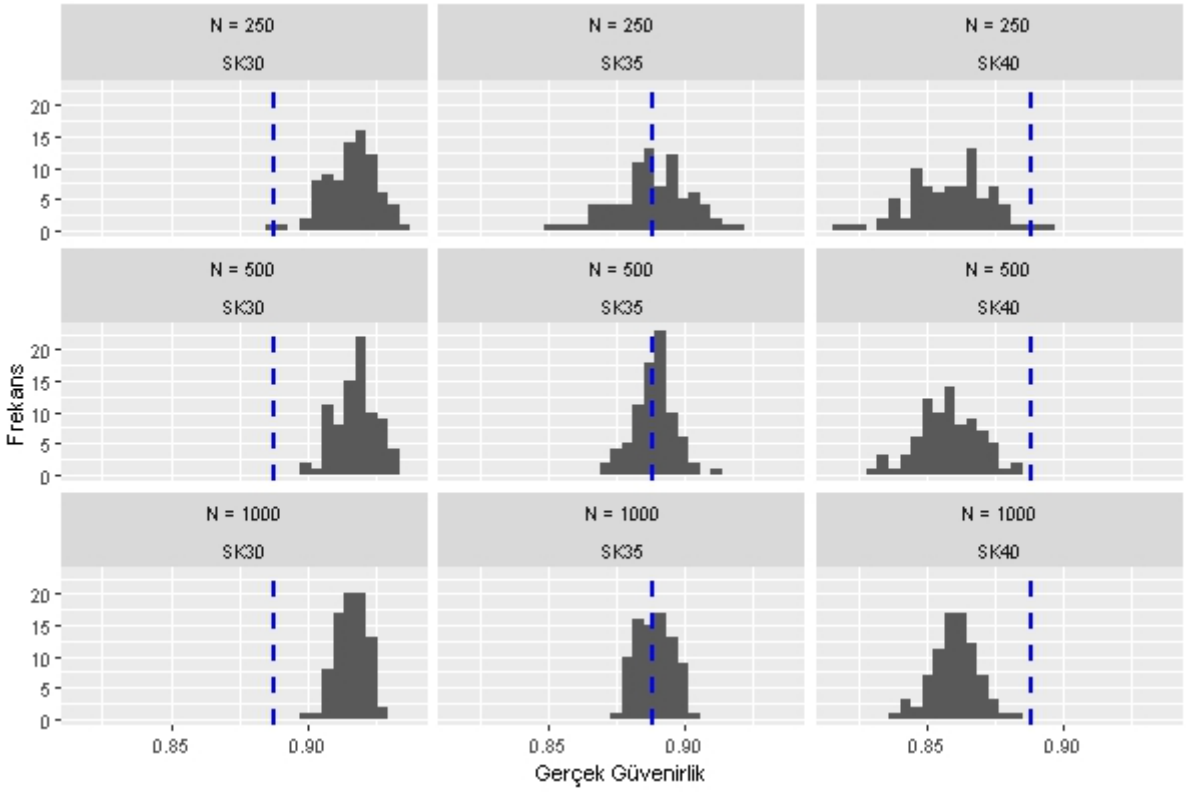
Şekil 1'de farklı örneklem büyüklüklerinde farklı standart hata sonlandırma kuralları ile elde edilen yetenek kestirimleri ile kestirimlerin standart hataları arasındaki ilişki sunulmuştur. Şekil 1'de SK40, 0.40'lık standart hata sonlandırma kuralı ile elde edilen sonuçları, SK35 0.35'lik standart hata sonlandırma kuralı ile elde edilen sonuçları ve SK30 ise 0.30'luk standart hata sonlandırma kuralı ile

elde edilen sonuçları ifade eder. Analiz edilen koşulların tamamında sonlandırma kuralına temel olan standart hata değeri sağlanmıştır. Yetenek dağılımının merkezinde sonlandırma kuralından çok daha iyi standart hata kestirimleri yapılabilmiştir. Ayrıca 0.40'lık sonlandırma kuralında daha geniş bir aralıkta standart hata kestirimleri yapılmışken 0.30'luk sonlandırma kuralında daha dar bir aralıkta standart hata kestirimleri yapılmıştır.

Tablo 1. Farklı simülasyon koşullarında gerçek güvenirlik, sıra korelasyonu ve SSR ortalamaları

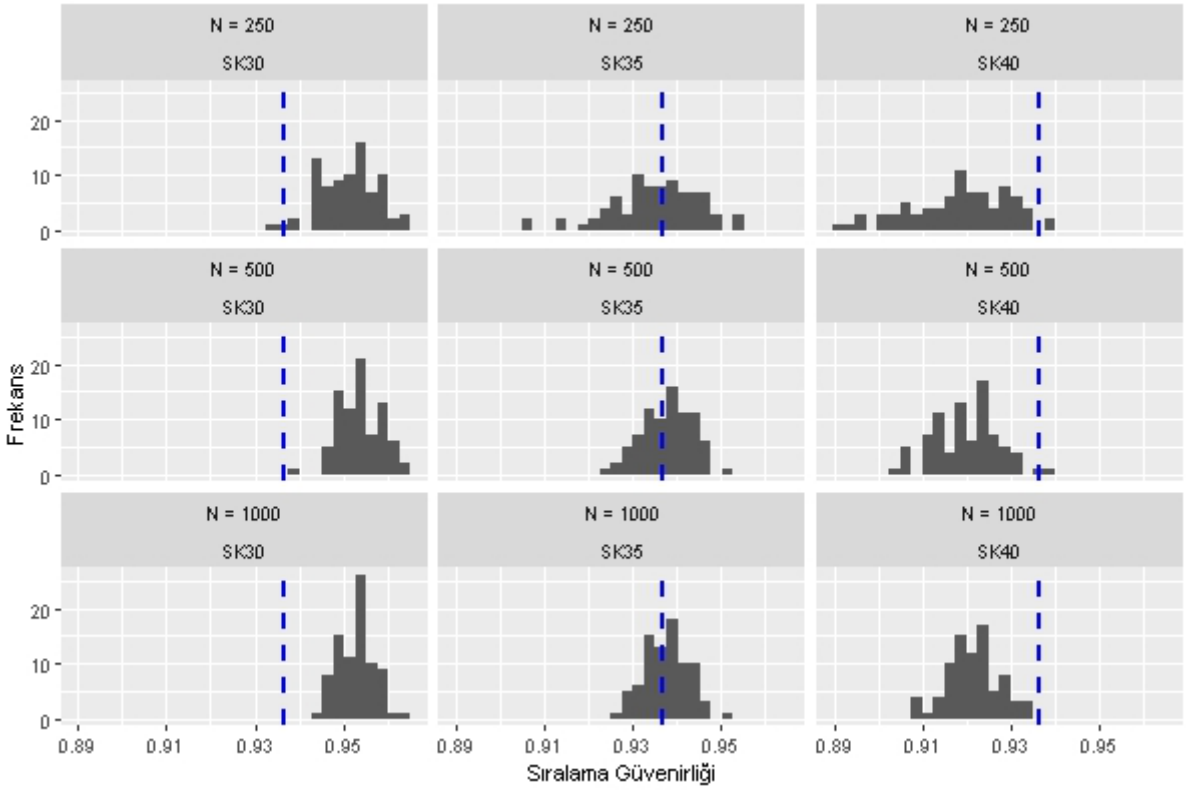
	Gerçek Güvenirlik			Sıra Korelasyonu			SSR		
N	**SK40**	**SK35**	**SK30**	**SK40**	**SK35**	**SK30**	**SK40**	**SK35**	**SK30**
250	0.859	0.888	0.916	0.918	0.936	0.951	0.841	0.881	0.913
500	0.858	0.889	0.917	0.920	0.938	0.954	0.843	0.883	0.916
1000	0.860	0.889	0.916	0.922	0.938	0.953	0.844	0.883	0.916

Tablo 1'de gerçek güvenirlik, sıra korelasyonu ve SSR ortalamaları sunulmuştur. Tablo 1 incelendiğinde kestirilen istatistiklerinin ortalamalarının örneklem büyüklüğüne göre pek farklılık göstermediği belirlenmiştir. Elde edilen tutarlılık istatistiklerinin tamamı iyi derecede tutarlılığa (Pollit, 2012) işaret etmektedir. Gerçek güvenirlik, sıra korelasyonu ve SSR istatistiklerinin nispeten esnek sayılabilecek SK40 sonlandırma kuralından nispeten katı sayılabilecek SK30 sonlandırma kuralına doğru ilerlenmesiyle iyileştiği sonucuna ulaşılabilir. Her bir simülasyon koşulunda 82 tekrarda elde edilen tutarlılık istatistiklerinin dağılımları Şekil 2, Şekil 3 ve Şekil 4'te sunulmuştur.



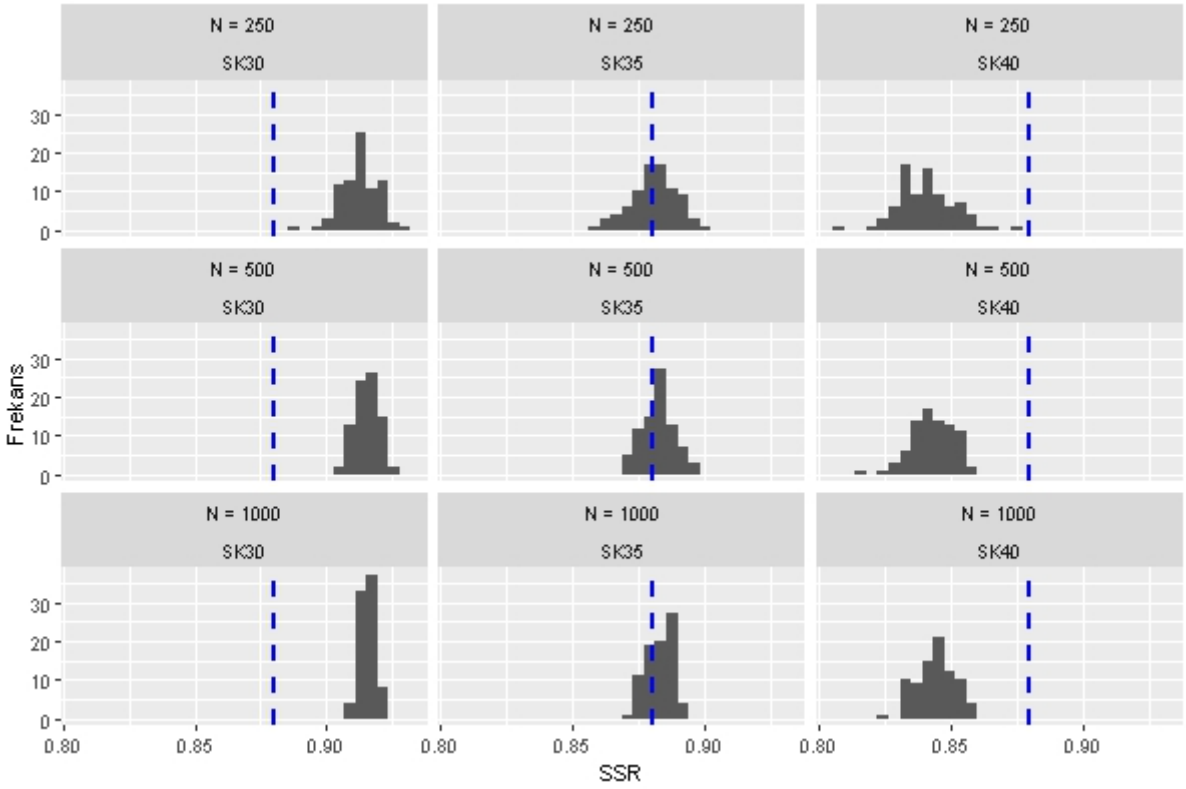
Şekil 2. Farklı simülasyon koşullarında elde edilen gerçek güvenirlik istatistiklerinin dağılımı

Şekil 2'de sunulan farklı simülasyon koşullarında elde edilen gerçek güvenirlik istatistikleri incelendiğinde her bir koşulda dağılımın şeklinin farklılaştığı tespit edilmiştir. Her ne kadar her koşulda kabul edilebilir ya da iyi seviyede gerçek güvenirlik istatistikleri elde edilmiş olsa da örneklem büyüklüğünün artışıyla kestirilen gerçek güvenirlik istatistiğinin daha dar bir alana yayıldığı tespit edilmiştir. Ayrıca daha katı standart hata sonlandırma kuralı ile daha yüksek seviyelerde gerçek güvenirlik elde edilmiştir.



Şekil 3. Farklı simülasyon koşullarında elde edilen sıralama güvenirliği istatistiklerinin dağılımı

Şekil 3'te sunulan farklı simülasyon koşullarında elde edilen sıralama güvenirliği istatistikleri incelendiğinde her bir koşulda dağılımın şeklinin farklılaştığı tespit edilmiştir. Her ne kadar her koşulda kabul edilebilir ya da iyi seviyede sıralama güvenirliği istatistikleri elde edilmiş olsa da örneklem büyüklüğünün artışıyla kestirilen sıralama güvenirliği istatistiğinin daha dar bir alana yayıldığı tespit edilmiştir. Ayrıca daha katı standart hata sonlandırma kuralı ile daha yüksek seviyelerde sıralama güvenirliği elde edilmiştir. Gerçek güvenirlik ve sıralama güvenirliği karşılaştırıldığında ise sıralama güvenirliği kısmen daha yüksek bulunmuştur.



Şekil 4. Farklı simülasyon koşullarında elde edilen SSR istatistiklerinin dağılımı

Şekil 4'te sunulan farklı simülasyon koşullarında elde edilen SSR istatistikleri incelendiğinde her bir koşulda dağılımın şeklinin farklılaştığı tespit edilmiştir. Her ne kadar her koşulda kabul edilebilir ya da iyi seviyede SSR istatistikleri elde edilmiş olsa da örneklem büyüklüğünün artışıyla kestirilen SSR istatistiğinin daha dar bir alana yayıldığı tespit edilmiştir. Ayrıca daha katı standart hata sonlandırma kuralı ile daha yüksek seviyelerde SSR elde edilmiştir. Sonlandırma kuralı olarak farklı seviyelerde standart hata değerleri kullanıldığında performanslar için gereken ortalama karşılaştırma sayısına ilişkin bulgular Tablo 2'de sunulmaktadır.

Tablo 2. İlgili sonlandırma kuralını sağlamak için gerekli ortalama karşılaştırma

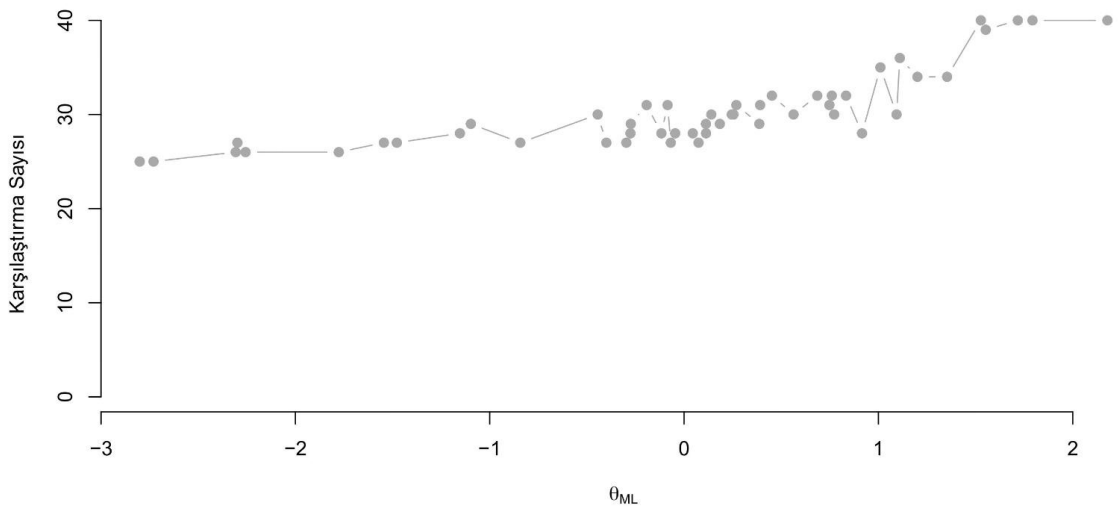
	Standart Hata Sonlandırma Kuralı		
N	**SK40**	**SK35**	**SK30**
250	30.56	39.16	52.55
500	30.46	38.84	51.68
1000	30.42	38.74	51.38

Tablo 2'de farklı simülasyon koşullarında ilgili standart hata sonlandırma kuralının sağlanması için her bir performansın ortalama kaç karşılaştırmaya dahil olması gerektiği belirtilmiştir. Tablo 2'de sunulan bilgilere göre ortalama karşılaştırmaya dâhil olma sayısı örneklem büyüklüğünden etkilenmemektedir. Ancak daha katı bir standart hata sonlandırma kuralını sağlamak için çok daha fazla karşılaştırma yapılması gerekmektedir.

Gerçek Uygulama Bulguları

Araştırmanın ikinci kısmında, 50 yazma performansına dair UKYP yetenek kestirimlerini elde edebilmek için toplam 754 ikili karşılaştırma gerçekleştirilmiştir. Ortalama olarak her bir karşılaştırmada 2 dakika 41 saniyede karar verilmiş, toplamda 33 saat 48 dakika 28 saniye puanlama yapılmıştır.

UKYP yetenek kestirimleri -2.80 ile 2.18 aralığında olup ortalaması 0, standart sapması ise 1.19'dur. 50 yazma performansının 46'sı 0.40'lık standart hata kuralını sağlamış, sadece 4 performans 40 en fazla karşılaştırma sayısına dayalı olarak daha fazla ikili karşılaştırmaya dâhil edilmemiştir. Söz konusu 4 performansa dair elde edilen standart hata kestirimleri sırasıyla 0.403, 0.409, 0.417 ve 0.472'dir. Bu performansların tamamı yüksek başarı seviyesindeki performanslar olup yetenek kestirimleri 1.53 ile 2.18 aralığındadır. Karşılaştırma sayısı ve yetenek kestirimleri arasındaki ilişki Şekil 5'te sunulmuştur.



Şekil 5. UKYP yetenek kestirimleri ile karşılaştırma sayısı arasındaki ilişki

Şekil 5'te sunulan bulgulara göre daha yüksek başarı seviyesindeki performansların daha fazla sayıda ikili karşılaştırmaya dâhil olduğunu söylemek mümkündür. Ayrıca her bir performansın ortalama 30.16 karşılaştırmaya dâhil olması daha önce Tablo 2'de sunulan örneklem büyüklüğünden bağımsız olarak 0.40'lık standart hata sonlandırma kuralının sağlanması için her bir performansın ortalama kaç karşılaştırmaya dâhil olması gerektiği bulgusu ile örtüşmektedir.

UKYP yetenek kestirimlerine dair elde edilen SSR istatistiği 0.89'dur. Bu bulgu puanlama güvenirliğinin oldukça yüksek olduğunu göstermektedir. Bu sonuç alan uzmanı olmayan puanlayıcıların bile yüksek güvenirlikli puanlama yapabileceğinin göstergesi olarak değerlendirilebilir. Ayrıca UKYP yetenek kestirimleri ile standartlaştırılmış analitik toplam puanlar arasında 0.71'lik, İngilizce dil yeterliği bütüncül puanları arasında 0.73'lük korelasyonlar hesaplanmıştır. UKYP yetenek kestirimlerinin diğer iki puanla ilişkisi Ek olarak sunulan Şekil 6'da görselleştirilmiştir. Bu bağlamda UKYP ile elde edilen yetenek kestirimlerinin analitik ve bütüncül rubrik kullanılarak gerçekleştirilen puanlama ile uyumlu olduğu sonucuna ulaşılabilir.

Tartışma, Sonuç ve Öneriler

Bu araştırma ile farklı örneklem büyüklüklerinde uyarlamalı eşleme ve farklı seviyelerde standart hata sonlandırma kuralı ile gerçekleştirilen karşılaştırma yargılarıyla puanlamanın güvenirliğini karşılaştırmak amaçlanmıştır. Monte Carlo simülasyonu sonuçları göstermektedir ki 0.30, 0.35 ve 0.40 standart hata sonlandırma kuralları uygulandığında yeteneğin orta noktalarında daha düşük standart hatalara rastlanmakta; dahası 0.40 standart hata sonlandırma kuralında 0.35 ve 0.30 standart hata sonlandırma kurallarına göre daha geniş bir aralıkta standart hata kestirimi yapılmaktadır. Örneklem büyüklüğü; gerçek güvenirlik, sıralama güvenirliği ve SSR açısından elde edilen sonuçlar üzerinde oldukça küçük farklılıklar oluşturmaktadır. Ancak mevcut araştırmanın aksine Crompvoets ve diğerleri (2020) örneklem büyüklüğünün sıralama güvenirliği ve gerçek güvenirlik değeri üzerinde etkili olduğunu belirtmiştir. Bu noktada Crompvoets ve diğerlerinin (2020) araştırmasında örneklem büyüklüğünün 20, 25, 30 ve 100 şeklinde belirlendiğini dikkate almak önemlidir. Mevcut çalışma ise en az 250 kişilik örneklemler içermektedir. Ancak yapılan tekrarlar dikkate alındığında örneklem büyüklüğü arttıkça gerçek güvenirlik, sıra güvenirliği ve SSR değerlerinin daha dar bir aralıkta bulunduğu fark edilmektedir.

Genel olarak sıralama güvenirliği, gerçek güvenirlik ve SSR değerlerinden daha yüksek bulunmuştur. Ortalama karşılaştırma sayısı açısından örneklem büyüklüğü bir farklılık ortaya çıkarmamış olup daha katı standart hata sonlandırma kuralının sağlanabilmesi için daha fazla sayıda karşılaştırma yapılması gerekmiştir. Araştırmada standart hata sonlandırma kuralına göre gerçek güvenirlik, sıra güvenirliği, SSR ve ortalama karşılaştırma sayısı farklılaşmaktadır. Sonuçlar genel olarak ele alındığında sıra korelasyonları 0.40, 0.35 ve 0.30 standart hata sonlandırma kuralında sırasıyla yaklaşık 0.92, 0.94 ve 0.95 değerlerini almaktadır. Elde edilen bu sonuç göstermektedir ki sıralamanın ön plana çıktığı durumlarda simülasyonda ele alınan bütün standart hata durdurma kurallarının kullanılması uygun olup alınacak kararların önemine göre kabul edilebilir standart hata büyüklüğüne göre standart hata sonlandırma kuralı belirlenebilir. Ancak 0.40, 0.35 ve 0.30 standart hata sonlandırma kuralı uygulandığında sırasıyla yaklaşık 30, 39 ve 52 ortalama karşılaştırma yapılması gerektiği hatırda tutulmalıdır. Gerçek güvenirlik değerleri ele alındığında 0.40, 0.35 ve 0.30 standart hata sonlandırma kuralı için sırasıyla yaklaşık 0.86, 0.89 ve 0.92 değerlerine ulaşılmıştır. Karşılaştırmalı yargıyla puanlama araştırmalarında ön plana çıkan SSR güvenirliği ise 0.40, 0.35 ve 0.30 standart hata sonlandırma kuralı için sırasıyla 0.84, 0.88 ve 0.92 bulunmuştur. Crompvoets ve diğerleri (2020), 20-22 karşılaştırma ile 0.80 gerçek güvenirlik ve SSR değerine ulaşılabileceğini belirtmektedir. Elde edilen sonuçlar bu kanıtla tutarlıdır. Verhavert vd., (2019) 0.90 civarında bir güvenirliğe ulaşmak için yaklaşık 26-37 karşılaştırma yapılması gerektiğini belirtmiştir. Mevcut araştırmada da ortalama yaklaşık 39 karşılaştırmada 0.90'a yakın güvenirlik değerlerine ulaşılmıştır. Pollitt (2012) ise araştırmasında 0.90'ın üzerinde bir güvenirliğe 9 tur karşılaştırma sonrasında erişmiştir. Ancak Pollitt'in (2012) araştırmasında yetenek kestirimlerine dair standart sapma değerinin 3.85 olmasının bu sonucu ortaya çıkardığı düşünülmektedir.

Araştırma tüm kâğıtlar için en fazla 0.40'lık standart hata koşulunda gerçekleştirilmiştir. Söz konusu standart hata değeriyle her bir öğrencinin yetenek kestirimi için hata miktarının düşük olması hedeflenmiştir. Böylece öğrenciler hakkında bireysel olarak alınan kararlar daha nitelikli olabilecektir. Nitekim alanyazın incelendiğinde yetenek kestirimlerinin standart hata ortalamalarının 1.5'e kadar ulaştığı (Verhavert vd., 2022) görülmektedir. Mevcut araştırma sonucunda KYP'de uyarlamalı seçim algoritmasının yeterli performansı gösterdiği sonucuna ulaşılmıştır. Alınacak kararların önemine göre mevcut araştırmadaki standart hata sonlandırma kurallarından birisinin tercih edilebileceği belirlenmiştir. Dahası araştırma sonuçları göstermiştir ki örneklem büyüklüğünün artması araştırma sonuçlarını farklılaştırmamaktadır. Dolayısıyla gerçek uygulamalarda yeterli puanlayıcıya ulaşıldığında büyük örneklemlerde de uyarlamalı seçim algoritması aracılığıyla KYP işlemi yapılabilir.

Araştırmanın ikinci kısmında ise simülasyon sonuçlarının elde edilmesinin ve örneklem büyüklüğünün sonuçları farklılaştırmadığının belirlenmesinin ardından 50 öğrencilik bir örneklemle gerçek uygulama gerçekleştirilmiştir. Simülasyon çalışmasına benzer koşullar altında yürütülen gerçek uygulamada 0.40 standart hata sonlandırma kuralı ve en fazla 40 karşılaştırma yapılması dikkate alınmıştır. Gerçek uygulama sonucunda simülasyon sonucuna benzer bulgulara erişilmiştir. Nitekim gerçek uygulamada kâğıtlar için ortalama 30 karşılaştırma yapılmış olup sonuç simülasyon çalışmasındaki 0.40 standart hata sonlandırma kuralı için belirlenen ortalama karşılaştırma sayısına oldukça yakındır. SSR değeri gerçek uygulamada 0.89 iken 0.40'lık standart hata sonlandırma kuralı ile gerçekleştirilen simülasyon çalışmasında 0.84'tür. Gerçek uygulamada kestirilen yeteneklerin standart sapma değerlerinin 1.19 olarak bulunması SSR değerinin gerçek uygulamada daha yüksek bulunmasıyla sonuçlanmış olabilir.

Gerçek uygulama sonuçlarında dikkat çeken bir unsur üst yetenek düzeylerinde karşılaştırma sayısının artması ve standart hataların daha yüksek olmasıdır. Bu durum üst yetenek düzeylerindeki bireylerin cevaplarından hangisinin daha iyi olduğuna karar verilmesinde daha fazla zorlanıldığına işaret etmekte olabilir. Bunun yanında gerçek uygulamada kullanılan veri setindeki bütüncül ve standartlaştırılmış analitik puanlar ile UKYP aracılığıyla kestirilen yetenekler arasındaki korelasyonlar 0.70'in üzerinde bulunmuştur. Literatürde rubrik ile KYP arasındaki korelasyonların 0.38 ile 0.92 aralığında değişebildiği belirtilmektedir (Steedle ve Ferrara, 2016). Gerçek uygulamadaki puanlama işlemini İngiliz dili alanında uzman olmayan ve C1 düzeyinde İngilizce becerisine sahip dört araştırmacının gerçekleştirmiş olması ve 0.70'in üzerinde bir korelasyona ulaşılması UKYP kullanımı konusunda umut vericidir. Bartholomew, Nadelson, Goodridge ve Reeve (2018), mevcut çalışmaya benzer şekilde araştırmalarında öğretmenlik yapmayan bireylerle KYP işlemi gerçekleştirmiştir. Gerçek uygulama sonucunda elde edilen standart hatalar ve güvenirlik değerleri K12 düzeyinde sınıf içi ölçme ve değerlendirmede KYP kullanılabileceğine işaret etmektedir. Bunun yanında KYP'yi kullanmak öğrencilerin alışageldiği test uygulamalarında bir farklılık oluşturmamaktadır (Pollitt, 2012).

UKYP, puanlama işleminin çok daha kolay bir kararla gerçekleşmesini sağlamaktadır. Bunun yanında elde edilen kestirimlerin madde tepki kuramı modellerine çok benzer bir biçimde elde edilmesi önemli bir avantajdır. Ayrıca rubriklerden farklı olarak, yetenek kestirimleri sürekli bir ölçek üzerinde elde edilmektedir; bu da ölçülen özellik bakımından bireyler arasındaki farklılıkların daha hassasiyetle ortaya konulmasını mümkün kılmaktadır. Araştırma kapsamındaki gerçek uygulamada yaklaşık 0.90'lık bir güvenirlik değeri, ortalama 30 karşılaştırma ile elde edilebilmektedir. Gerçek uygulamalarda karşılaştırma sayısını azaltmak için, nispeten daha düşük güvenirlik düzeyleri hedeflenebilir. Hele ki sınıf içi uygulamalarda 0.70 düzeyinde bir güvenirlik hedeflenmesi durumunda bu hedefe daha az sayıda karşılaştırma ile ulaşılması mümkün görünmektedir.

754 karşılaştırma için elde edilen ortalama karşılaştırma süresi göz önüne alındığında, bir performansın nihai olarak puanlanmasında yaklaşık 40 dakikalık bir süre gerektiği görülmektedir. Bu sürenin İngiliz Dili uzmanı olmayan puanlayıcılarla ve ortalama kelime sayısının yüksek olduğu bir durumda ortaya çıktığı göz ardı edilmemelidir. Bunun yanında, MEB'in yakın zamanda aldığı kararlar göz önüne alındığında kısa-orta vadede açık uçlu soruların yüksek önem arz eden sınavlarda kullanılması olası görünmektedir. Bu sınavlarda beş-altı kategoriden oluşan bütüncül rubriklerle puanlama yapılması durumunda, temelinde sıralama amaçlı olan bu sınavlarda ayırt edicilik problemi ortaya çıkabilir; bu nedenle de analitik bir puanlamayla bu çalışma kapsamında olduğu gibi bir kompozisyonun birkaç boyutta ayrı ayrı puanlanması gerekecektir. Rubriğe dayalı yöntemlerde her kâğıdın en az iki puanlayıcı tarafından puanlanması gerektiği ve uyuşmazlık durumunda deneyimli puanlayıcıların da sürece dâhil olduğu göz önüne alındığında bu sürenin oldukça makul olduğu söylenebilir.

MEB ve ÖSYM, bir süredir dört dil becerisi sınavları üzerinde çalışmaktadırlar. Bu sınavlarda sadece yazma becerileri değil, konuşma becerileri de ölçülmek istenmektedir. KYP, sadece yazma becerisinin değil, bir diğer üretimsel beceri olan konuşma becerisinin de puanlanmasında kullanılabilir. Verilen kararın sadece hangi performansın daha iyi olduğuna odaklanılmasıyla yüksek güvenirlik değerlerinin elde edilebilmesi, KYP'nin bu tür sınavların bütünsel değerlendirmeye uygun farklı boyutlarında da kullanılabilmesini mümkün kılmaktadır. Ancak bu çalışmanın, Türkiye'de K-12 düzeyinde KYP yönteminin kullanımına yönelik ilk gerçek uygulama araştırması olduğu unutulmamalıdır. Bu nedenle, KYP'nin K-12 düzeyinde kullanılabilirliğine yönelik daha fazla araştırma yapılması gerektiği belirtilebilir.

Sonuçlara dayalı olarak araştırma kapsamında ele alınan tüm koşullarda yüksek seviyede puanlama güvenirliği tespit edilmiştir. Genel bir kabul olarak UKYP'den yararlanıldığında değerlendiricilerden kararlarını verirken yalnızca geçerliği göz önünde bulundurmaları istenebilecektir. Nitekim sonuçların son derece yüksek güvenilirlikte olduğu (Jones ve Davies, 2023; Pollitt, 2012) belirlenmiştir.

Mevcut araştırma sonuçlarına dayalı olarak ileride yapılacak çalışmalarda araştırmacılara uyarlamalı seçim algoritmasına alternatif bir seçim algoritması geliştirmeleri önerilebilir. Bunun yanı sıra standartlaştırma işleminin de yapılmasına aracılık edebilecek referans set temelli UKYP çalışması gerçekleştirilebilir. Referans set temelli yaklaşımda ise uyarlamalı seçim algoritması, bayese dayalı uyarlamalı seçim algoritması ya da araştırmacılar tarafından ortaya konulacak yeni seçim algoritmalarından yararlanılabilir. Araştırma, simülasyon faktörleri açısından standart hata sonlandırma kuralı ve örneklem büyüklüğü ile sınırlıdır. Sonraki araştırmalarda simülasyon faktörlerinin sayısı artırılabilir. Bu yönde yetenek dağılımı normal dağılımla üretilirken farklı standart sapma değerleri kullanılabilir. Öğrencilerin yetenek düzeylerindeki farklılaşmanın daha fazla olması durumunda standart hata sonlandırma kuralı ile uyarlamalı seçim algoritması kullanmanın etkileri incelenebilir. Araştırma gerçekleştirilirken puanlayıcı sayısı üzerine bir incelemede bulunulmamıştır. Yapılacak araştırmalarda puanlayıcı sayısı farklılaştırılabilir, puanlayıcıların arasındaki uyum düzeyleri üzerinde değişiklikler yapılarak UKYP güvenirliği üzerindeki etkileri incelenebilir.

Teşekkür

Bu çalışmanın ortaya çıkarılmasında yararlanılan "Açık Uçlu Maddelerin Uyarlamalı Karşılaştırmalı Yargıyla Puanlanması için bir Sistem ve Yazılım Geliştirme" adlı projeyi destekleyen ÖSYM Araştırma ve Geliştirme Daire Başkanlığı'na teşekkürlerimizi sunarız.

Kaynakça

Andrich, D. (1978). Relationships between the Thurstone and Rasch approaches to item scaling. *Applied Psychological Measurement*, *2*(3), 451-462. doi:10.1177/014662167800200319

Bartholomew, S. R., Nadelson, L. S., Goodridge, W. H. ve Reeve, E. M. (2018). Adaptive comparative judgment as a tool for assessing open-ended design problems and model eliciting activities. *Educational Assessment, 23*(2), 85-101. doi:10.1080/10627197.2018.1444986

Benton, T. (2021). Comparative judgement for linking two existing scales. *Frontiers in Education*, 6, 775203. doi:10.3389/feduc.2021.775203

Bloxham, S. (2009). Marking and moderation in the UK: False assumptions and wasted resources. *Assessment & Evaluation in Higher Education*, *34*(2), 209-220. doi:10.1080/02602930801955978

Bradley, R. A. ve Terry, M. E. (1952). Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika, 39*(3-4), 324-345. doi:10.1093/biomet/39.3-4.324

Bramley, T. (2005). A rank-ordering method for equating tests by expert judgment. *Journal of Applied Measurement, 6*(2), 202-223.

Bramley, T. (2007). Paired comparison methods. P. Newton, J. Baird, H. Goldstein, H. Patrick ve P. Tymms (Ed.), *Techniques for monitoring the comparability of examination standards* içinde (s. 246-294). Qualifications and Curriculum Authority.

Bramley, T. ve Vitello, S. (2019). The effect of adaptivity on the reliability coefficient in adaptive comparative judgement. *Assessment in Education: Principles, Policy & Practice, 26*(1), 43-58. doi:10.1080/0969594X.2017.1418734

Christodoulou, D. (2024). *Using comparative judgement to improve writing [webinar]. The Education Hub*. https://theeducationhub.org.nz/using-comparative-judgement-to-improve-writing/#:~:text=Comparative%20judgement%20is%20a%20process,double%20marking%2C%20but%20much%20quicker adresinden erişildi.

Crisp, V. (2013). Criteria, comparison and past experiences: How do teachers make judgements when marking coursework? *Assessment in Education: Principles, Policy & Practice, 20*(1), 127-144. doi:10.1080/0969594X.2012.741059

Crompvoets, E. A. V., Béguin, A. A. ve Sijtsma, K. (2020). Adaptive pairwise comparison for educational measurement. *Journal of Educational and Behavioral Statistics, 45*(3), 316-338. doi:10.3102/1076998619890589

Crompvoets, E. A. V., Beguin, A. A. ve Sijtsma, K. (2021). *Pairwise comparison using a Bayesian selection algorithm: Efficient holistic measurement*. PsyArXiv. doi:10.31234/osf.io/32nhp

Crompvoets, E. A. V., Béguin, A. A., ve Sijtsma, K. (2022). On the bias and stability of the results of comparative judgment. *Frontiers in Education*, 6, 788202. doi:10.3389/feduc.2021.788202

Crossley, S. A., Tian, Y., Baffour, P., Franklin, A., Kim, Y., Morris, W. … Boser, U. (2023). Measuring second language proficiency using the English Language Learner Insight, Proficiency and Skills Evaluation (ELLIPSE) corpus. *International Journal of Learner Corpus Research, 9*(2), 248-269. doi:10.1075/ijlcr.22026.cro

Daniel, F., Microsoft Corporation, Weston, S. ve Tenenbaum, D. (2022). doParallel: Foreach parallel adaptor for the 'parallel' package (Version 1.0.17) [Bilgisayar yazılımı]. https://CRAN.R-project.org/package=doParallel adresinden erişildi.

Goossens, M. ve De Maeyer, S. (2018). How to obtain efficient high reliabilities in assessing texts: Rubrics vs comparative judgement. Technology enhanced assessment. TEA 2017. Communications in Computer and Information Science, Springer, Cham. doi:10.1007/978-3-319-97807-9_2

Gustafsson, J.-E. (1977). *The Rasch model for dichotomous items: Theory, applications and a computer program*. Göteborg: Göteborg University.

Heldsinger, S. ve Humphry, S. (2013). Using calibrated exemplars in the teacher-assessment of writing: an empirical study. *Educational Research, 55*(3), 219-235. doi:10.1080/00131881.2013.825159

Holmes, S. D., Meadows, M., Stockford, I. ve He, Q. (2018). Investigating the comparability of examination difficulty using comparative judgement and Rasch modelling. *International Journal of Testing, 18*(4), 366-391. doi:10.1080/15305058.2018.1486316

Humphry, S. M. ve Heldsinger, S. (2019). A two-stage method for classroom assessments of essay writing. *Journal of Educational Measurement, 56*(3), 505-520. doi:10.1111/jedm.12223

Jones, I. ve Davies, B. (2023). Comparative judgement in education research. *International Journal of Research & Method in Education, 47*(2), 170-181. doi:10.1080/1743727X.2023.2242273

Laming, D. (2003). *Human judgment: The eye of the beholder*. Thomson Learning.

Lesterhuis, M., Bouwer, R., Van Daal, T., Donche, V. ve De Maeyer, S. (2022). Validity of comparative judgment scores: How assessors evaluate aspects of text quality when comparing argumentative texts. *Frontiers in Education, 7*, 823895. doi:10.3389/feduc.2022.823895

Luce, R. D. (1959). *Individual choice behaviours: A theoretical analysis*. New York: John Wiley & Sons.

MEB Ölçme ve Değerlendirme Yönetmeliği. (2023). *Resmi Gazete* (Sayı: 32304). https://www.resmigazete.gov.tr/eskiler/2023/09/20230909-2.htm adresinden erişildi.

Pollitt, A. (2012). The method of adaptive comparative judgement. *Assessment in Education: Principles, Policy & Practice, 19*(3), 281-300. doi:10.1080/0969594X.2012.665354

R Core Team. (2023). *R: A language and environment for statistical computing* (Version 4.3.0) [Bilgisayar yazılımı]. Retrieved from https://www.r-project.org/

Sims, M. E., Cox, T. L., Eckstein, G. T., Hartshorn, K. J., Wilcox, M. P. ve Hart, J. M. (2020). Rubric rating with MFRM versus randomly distributed comparative judgment: A comparison of two approaches to second-language writing assessment. *Educational Measurement: Issues and Practice, 39*(4), 30-40. doi:10.1111/emip.12329

Steedle, J. T. ve Ferrara, S. (2016). Evaluating comparative judgment as an approach to essay. *Applied Measurement in Education, 29*(3), 211-223. doi:10.1080/08957347.2016.1171769

Thurstone, L. L. (1927). A law of comparative judgment. *Psychological Review, 34*(4), 273-286. doi:10.1037/h0070288

Thwaites, P., Kollias, C. ve Paquot, M. (2024). Is CJ a valid, reliable form of L2 writing assessment when texts are long, homogeneous in proficiency, and feature heterogeneous prompts?. *Assessing Writing*, 60. https://doi.org/10.1016/j.asw.2024.100843

Using anchors to link judging sessions. (2016). https://nomoremarkingltd.freshdesk.com/support/solutions/articles/16000029952-using-anchors-to-link-judging-sessions adresinden erişildi.

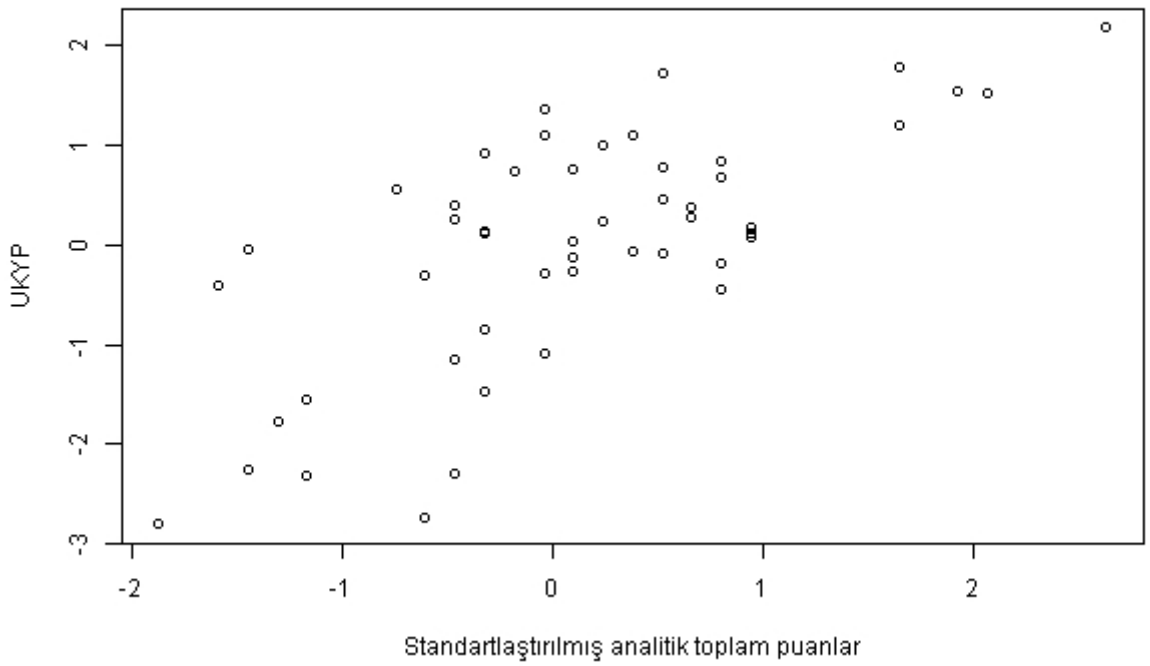
Uysal, İ., Gürel, S., Şahin, M. D., İbileme, A. İ. ve Yıldırım Görgülü, Y. (2024). *Açık uçlu maddelerin karşılaştırmalı yargıyla puanlanmasında sabit sayıda uyarlamalı ve rassal eşlemeye dayalı bir simülasyon çalışması.* 9. Eğitimde ve Psikolojide Ölçme ve Değerlendirme Kongresi, Anadolu Üniversitesi, Eskişehir.

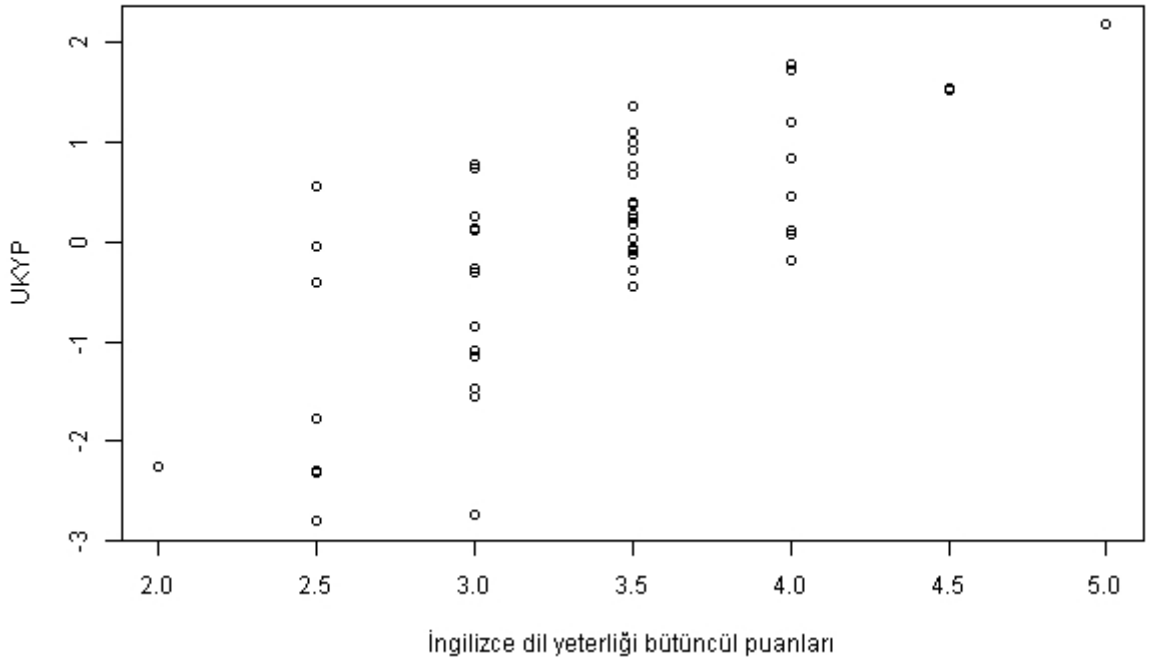
van Daal, T., Lesterhuis, M., Coertjens, L., Donche, V. ve De Maeyer, S. (2019). Validity of comparative judgement to assess academic writing: Examining implications of its holistic character and building on a shared consensus. *Assessment in Education: Principles, Policy & Practice, 26*(1), 59-74. doi:10.1080/0969594X.2016.1253542

Verhavert, S., Bouwer, R., Donche, V. ve De Maeyer, S. (2019). A meta-analysis on the reliability of comparative judgement. *Assessment in Education: Principles, Policy & Practice, 26*(5), 541-562. doi:10.1080/0969594X.2019.1602027

Verhavert, S., Furlong, A. ve Bouwer, R. (2022). The accuracy and efficiency of a reference based adaptive selection algorithm for comparative judgment. *Frontiers in Education*, 6, 785919. doi:10.3389/feduc.2021.785919

Ek 1





Şekil 6. UKYP yetenek kestirimleri ile uzmanlarca puanlanan İngilizce dil yeterliği bütüncül puanları ve standartlaştırılmış analitik toplam puanları arasındaki ilişki