TURKISH ACADEMIC RESEARCH REVIEW
Türk Akademik Araştırmalar Dergisi
e-ISSN: 2602-2923

Çok Kategorili Puanlanan Maddelerden Oluşan Testlerde Klasik Test Kuramı ve Madde Tepki Kuramı'na Dayalı Test Eşitleme Yöntemlerinin Karşılaştırılması*

Comparison of Test Equating Methods Based on Classical Test Theory and Item Response Theory in Polytomously Scored Tests

Öz

Aynı örtük özelliği ölçen testin benzer zorluklara sahip iki formunun puanlarının birbirine dönüştürülmesini içeren istatistiksel süreç test eşitleme olarak tanımlanır. Bu çalışma* çok kategorili puanlanan maddelerden oluşan test formlarının eşitlenmesi sürecinde Klasik Test Kuramı ve Madde Tepki Kuramı'na dayalı test eşitleme yöntemlerinden elde edilen hataların karşılaştırılmasını amaçlayan bir olasılıksal simülasyon çalışmasıdır. Yürütülen çalışmada örneklem büyüklüğü 50, 100, 300, 1000 ve 3000 kişi ve test uzunluğu 10, 20, 30, 40 ve 50 madde olmak üzere 2 değişimlenen koşul için 25 alt koşul belirlenmiştir. Bu alt koşullara uygun olarak 0 (yanlış), 1(kısmen doğru) ve 2 (tam doğru) şeklinde puanlanan genelleştirilmiş kısmi puan modeline uyumlu simülatif veriler kullanılmıştır. Çalışmada veri setlerini üretmek için WinGen3 programından faydalanılmıştır. Üretilen veri setlerinin eşitlenme süreci R programı ile yürütülmüştür. Bu çalışmada doğrusal, ortalama ve eşit yüzdelikli eşitleme yöntemleri olmak üzere üç Klasik Test Kuramı yöntemi (KTK) ve ortalama-ortalama ve ortalama-standart sapma olmak üzere iki Madde Tepki Kuramı (MTK) yöntemi kullanılmıştır. Üretilen test formları kullanılarak üç KTK yönteminden ve iki MTK yönteminden elde edilen eşitlemenin standart hataları hesaplanmış ve yöntemlerden elde edilen hata miktarları karşılaştırılmıştır. Araştırma sonucunda, farklı örneklem büyüklüğü ve farklı test uzunluğuna göre üretilen test formlarının eşitlenmesi sürecinde en az hata ile eşitleme yapan yöntemin eşit yüzdelikli eşitleme yöntemi ve en fazla hata ile eşitleme yapan yöntemin ortalama-ortalama yöntemi olduğu saptanmıştır. Çalışmanın sonucunun literatürle uyumlu olduğu görülmüştür. Test eşitleme yöntemleri örneklem uzunluğu bağlamında incelendiğinde örneklem büyüklüğü arttıkça test eşitleme hatasının azaldığı belirlenmiştir. Yöntemler, test uzunluğu bağlamında değerlendirildiğinde ise madde sayısı arttıkça hesaplanan hata miktarının da arttığı sonucuna ulaşılmıştır.

Anahtar kelimeler: Eğitim Bilimleri, Test eşitleme, Madde Tepki Kuramı, Klasik Test Kuramı, Eşitleme hatası, Çok kategorili puanlanan testler.

Abstract

The statistical process of converting the scores of two forms of a test measuring the same latent trait with similar difficulties is defined as test equating. This study is a stochastic simulation study aiming to compare the errors obtained from test equating methods based on Classical Test Theory and Item Response Theory in the process of equating test forms consisting of polytomously scored items. In the study, 25 sub-conditions were determined for 2 varying conditions with sample sizes of 50, 100, 300, 1000, and 3000 people and test lengths of 10, 20, 30, 40 and 50 items. In accordance with these sub-conditions, simulative data compatible with the Generalized Partial Credit Model, which is scored as 0 (incorrect), 1 (partially correct), and 2 (fully correct), were used. WinGen3 program was used to generate the data sets in the study. The equating process of the generated data sets was carried out with the R program. In this study, three Classical Test Theory (CTT) methods, namely linear, mean, and equipercentile equating methods, and two Item Response Theory (IRT) methods, namely mean-mean and mean-sigma, were used. Using the test forms produced, the standard errors of equating obtained from the three CTT methods and the two IRT methods were calculated, and the error amounts obtained from the methods were compared. As a result of the study, it was found that the method that equalized with the least error in the process of equating the test forms produced according to different sample sizes and different test lengths was the equipercentile equating method and the method that equalized with the most error was the mean-mean method. The results of the study were found to be consistent with the literature. When the test equating methods were analyzed in terms of sample length, it was determined that the test equating error decreased as the sample size increased. When the methods were evaluated in terms of test length, it was concluded that the amount of error calculated increased as the number of items increased.

Keywords: Educational Sciences, Test equating, Item Response Theory, Classical Test Theory, Equating error, Polytomously scored tests.

Merve ÇÖRTÜK
Doktorant, Akdeniz Üniversitesi Eğitim Fakültesi, Eğitimde Ölçme ve Değerlendirme Anabilim Dalı / PhD student, Akdeniz University, Faculty of Education Department of Measurement and Evaluation in Education
mcortuk95@gmail.com,
https://orcid.org/0000-0002-7687-2206
https://ror.org/01m59r132

Alper SİNAN
Doç. Dr., Akdeniz Üniversitesi Eğitim Fakültesi, Eğitimde Ölçme ve Değerlendirme Anabilim Dalı / Associate Professor, Akdeniz University, Faculty of Education Department of Measurement and Evaluation in Education
asinan@akdeniz.edu.tr,
https://orcid.org/0000-0001-6632-5500
https://ror.org/01m59r132

*Bu çalışma Merve Çörtük'ün 2022 yılı "Çok Kategorili Puanlanan Maddelerden Oluşan Testlerde Klasik Test Kuramı ve Madde Tepki Kuramı'na Dayalı Test Eşitleme Yöntemlerinin Karşılaştırılması" isimli yüksek lisans tezinden üretilmiştir.

Makale Türü-Article Type: Araştırma Makalesi/Research Article
Geliş Tarihi/Recieved: 04.10.2023
Kabul Tarihi/Accepted: 30.11.2023
Yayım Tarihi/Date Published: 31.12.2023

***Atıf/Cite as*:** Çörtük, M.-Sinan, A. (2023). Çok kategorili puanlanan maddelerden oluşan testlerde klasik test kuramı ve madde tepki kuramı'na dayalı test eşitleme yöntemlerinin karşılaştırılması. *Turkish Academic Research Review, 8 (4),* 1429-1439.

Değerlendirme/Peer-Review: Ön İnceleme: İç Hakem (Editörler). İçerik İnceleme: İki Dış Hakem/Çift taraflı körleme. Single anonnymized-One internal (Editorial Board). Double anonymized-Two extarnal.

Benzerlik Taraması/Plagiarism Checks: *Yapıldı-Turnitin/Yes-Turnitin.*

Yayıncı/Published: Published by Mehmet ŞAHİN Since 2016- Akdeniz University, Faculty of Theology, Antalya, 07058 Turkey.

***Etik Beyan/Ethical Statement*:** Bu çalışmanın hazırlanma sürecinde bilimsel ve etik ilkelere uyulduğu ve yararlanılan tüm çalışmaların kaynakçada belirtildiği beyan olunur. / It is declared that scientific and ethical principles have been followed while carrying out and writing this study and that all the sources used have been properly cited. Merve Çörtük - Alper Sinan

Çıkar Çatışması/Conflicts of Interest: Çıkar çatışması beyan edilmemiştir. / The author(s) has no conflict of interest to declare.

Finansman/Grant Support: Bu araştırmayı desteklemek için dış fon kullanılmamıştır. / The author(s) acknowledge that they received no external funding in support of this research.

Etik Bildirim/Complaints: turkisharr@gmail.com

1.Giriş

Eğitim ve psikolojide ölçülmek istenen özellik her zaman doğrudan ölçülebilir değildir. Bu nedenle çeşitli ölçme araçları kullanılır. Özellikle eğitimde ölçme araçları önemli yere sahiptir. Okullarda ilgi, yetenek gibi duyuşsal özellikleri ölçen araçların yanısıra maksimum performansı ölçen araçlara da sıklıkla yer verilmektedir. Öğretmenler ve eğitim yöneticileri sıklıkla birkaç farklı test türüyle elde edilen sonuçlara göre hareket etmektedir. Bu testlerin yanı sıra belirli bir eğitim veya öğretim programının etkilerini ölçmek için de başarı testleri kullanılmaktadır (Anastasi, 1976). Sadece okullar değil ülke genelinde de her yıl onlarca test uygulanmaktadır. Lise ve üniversite gibi eğitim kademelerine geçiş için yapılan testler, meslek hayatına geçişte kurumların yaptığı testler bunların sadece bir kısmını oluşturmaktadır. Yapılan bu testlerin sonuçlarıyla bireyin hayatını etkileyen önemli kararlar alınmaktadır. Dolayısıyla da sonuçların geçerli ve güvenilir olması istenmektedir. Sonuçların güvenirliğini etkileyen en önemli faktörlerden birisi bu testlerde yer alan maddelerin yalnızca bir kez kullanılıyor oluşu ve yılda bir ya da daha fazla yapılması halinde farklı test formlarının yer almasıdır. Geniş ölçekli testleri bireylerin yıl içerisinde birden fazla alma hakkının olması aynı test formunun kullanılmasını imkânsız hale getirmektedir. Bu duruma bağlı olarak da farklı test formları kullanılmaktadır. Aynı testin farklı formlarını alan bireylerden birinin diğerine karşı avantaj ya da dezavantaj sahibi olmaması için bu formların benzer olması gerekmektedir. Formların benzerliğinin kontrol edilmesi test ilişkilendirme işlemi ile mümkün olmaktadır (Kolen & Brennan, 2014).

Mislevy (1992) ve Linn (1993), ilişkilendirme terimini, bir testin sonuçlarını diğer testlerle karşılaştırılabilir hale getirmek amacıyla kullanılan bir genel terim olarak kullanmışlardır; bu terimi beş ana başlık altında toplamışlardır: eşitleme, kalibrasyon, istatistiksel moderasyon, kestirim ve sosyal moderasyon.

Eşitleme, bir testten elde edilen puanların diğer bir testten elde edilen puanlar yerine kullanılabilmesini sağlayan bir ilişkilendirme yöntemidir. Diğer ilişkilendirme yöntemlerine göre daha zor istatistiksel sürece sahip olmasına rağmen en güçlü ilişkilendirme yöntemi olması nedeniyle avantaj sağlamaktadır. Varsayımların sağlandığı durumda, eşitlenmiş iki test puanları birbirinin yerine kullanılabilir. Kalibrasyon, bir testin daha kısa ve tipik olarak daha az güvenilir bir versiyonunun, daha uzun ve güvenilir bir versiyonu ile karşılaştırılmasını içerir; bu, farklı test formlarını karşılaştırılabilir kılar. İstatistiksel moderasyon, farklı kaynaklardan gelen veya farklı özellikleri ölçen testlerin karşılaştırılmasını sağlayan bir yöntemdir ve genellikle bu tür karşılaştırmalar için ankor testleri kullanılır. Kestirim, bir test veya formdan elde edilen puanlarla başka bir formun veya testin puanlarını tahmin etmeyi içerir ve genellikle grup ve zaman bağlamında kullanılır. Sosyal moderasyon, farklı kişiler tarafından ve standart bir puanlama düzenine göre elde edilen puanların karşılaştırılmasını ifade eder, bu özellikle sosyal bilimler veya toplumsal faktörlerin etkisi altında olan alanlarda kullanışlıdır.

Test eşitleme yöntemleri genel olarak geleneksel eşitleme yöntemleri ve Madde Tepki Kuramı'na dayalı eşitleme yöntemleri olmak üzere iki başlık altında yer almaktadır. Geleneksel eşitleme yöntemleri Klasik Test Kuramı'na dayalı eşitleme yöntemleri olarak da bilinmektedir. Bu çalışmada geleneksel test eşitleme yöntemlerinden ortalama, eşit yüzdelikli ve doğrusal eşitleme; Madde Tepki Kuramı'na dayalı test eşitleme yöntemlerinden ortalama-ortalama ve ortalama-standart sapma yöntemleri kullanılmıştır.

Eşit Yüzdelikli Eşitleme. Aynı özelliği ölçen testin bir formunda yer alan bir puan ve diğer formunda yer alan bir puanın yüzdelik sıralaması aynıysa testi alan grup içerisinde eşit olduğu kabul edilir. Buna göre, puanları

ayarlamak için testin bir formunda yer alan puanları testin diğer formunda yer alan puanlara eşitlemek için bir formun puanları aynı yüzdelik dilimde yer aldığı diğer form puanlarına dönüştürülür. Bu işlem eşit yüzdelikli eşitleme işlemi olarak adlandırılır (Livingston, 2014).

Doğrusal eşitleme. Bu yöntemde, ortalamalar ve standart sapma dışında iki test formuna ait puanların eşit olduğu varsayılır (Muraki, Hombo, & Lee, 2000). İki form arasındaki farkları sabit olarak düşünmek yerine iki test formu arasındaki zorluk farklılıklarının puan ölçeği boyunca değişmesine izin verir. Doğrusal eşitlemede, ortalamalarına eşit uzaklıkta olan puanlar standart sapma birimlerinde eşit olarak ayarlanır. Bu nedenle, doğrusal eşitleme, iki formun araçlarının yanı sıra ölçek birimlerinin de farklı olmasına izin vermektedir.

Ortalama eşitleme. Bu yöntem, doğrusal eşitleme ve eşit yüzdelikli eşitlemeye göre daha az katıdır. Testin iki formundan alınan puan ortalamalarının eşit olması gerektiği varsayımına dayanır (Sansivieri, Wiberg, & Matteucci, 2017).

Ortalama-standart sapma eşitleme. Madde güçlük parametresinin standart sapması ve ortalaması kullanılarak elde edilen katsayılar aracılığıyla yürütülen bir test eşitleme yöntemidir (Marco, 1977).

Ortalama-ortalama eşitleme. Madde güçlük parametresiyle birlikte ayırt edicilik parametresinin de dikkate alındığı ve bu parametrelerin ölçek puan dönüşümünde kullanıldığı eşitleme yöntemidir (Loyd & Hover, 1980).

Bootstrap yöntemi, standart hataları tahmin etmek için kullanılan bir yöntemdir ve bu yöntem Efron (1982) tarafından geliştirilmiştir. Bootstrap yöntemi ile standart hata hesaplama işlemi beş adımda gerçekleştirilir, bu adımları Kolen ve Brennan (2014) aşağıdaki gibi açıklar:

1. İlk olarak, N büyüklüğünde bir örneklemle başlanır. Bu örnek veri, orijinal veriden çekilir.
2. Ardından, bu N boyutundaki örnek veriler kullanılarak çeşitli örnek veriler oluşturulur. Her bir örneklem, rastgele verileri içerir.
3. Her bir örneklem için örneklem ortalaması hesaplanır. Bu, her örneklem için ayrı ayrı yapılır.
4. Adım 2 ve 3, R kez tekrarlanır. Yani, R farklı örneklem oluşturulur ve her bir örneklemin ortalaması hesaplanır.
5. Son olarak, bu R örneklem ortalamalarının standart sapması hesaplanır. Bu standart sapma, istatistiğin tahmini bootstrap standart hatasıdır ve istatistiğin güvenilirliğini değerlendirmek için kullanılır.

Bu araştırmanın ana amacı, değişen test uzunlukları ve örneklem büyüklükleri altında çeşitli test eşitleme yöntemlerini (ortalama, eşit yüzdelikli, doğrusal, ortalama-ortalama ve ortalama-standart sapma) kullanarak test formlarını eşitleme hatalarını karşılaştırmaktır. Bu kapsamda, farklı test uzunluklarına sahip testler (10, 20, 30, 40 ve 50 madde) ve değişen örneklem büyüklüklerine sahip gruplar (50, 100, 300, 1000 ve 3000) üzerinde çalışılarak, hangi test eşitleme yönteminin en düşük hata oranına sahip olduğu sorusu yanıtlanmaya çalışılmıştır.

2.Yöntem

2.1.Araştırmanın Modeli

Bu çalışma, temel bir araştırma türüdür ve temel araştırma, bir teori veya hipotezin altında yatan süreçleri açıklamaya yönelik bir çalışma olarak tanımlanır (Fraenkel, Wallen, & Hyun, 2012). Bu çalışma, farklı test uzunlukları ve örneklem büyüklükleri üzerinde odaklanarak test eşitleme yöntemlerini karşılaştırmayı amaçlamaktadır. Bu çalışmanın temel amacı, hangi yöntemin en düşük hata oranını verdiğini belirlemek ve teorik çalışmalara katkıda bulunmaktır. Bu nedenle, bu çalışma temel bir araştırma türünü yansıtmaktadır.

Eğitim alanında önemli katkılar sağlayan simülasyon çalışmaları genellikle iki ana kategori altında incelenir: belirleyici ve olasılıksal simülasyon çalışmaları. Belirleyici simülasyon çalışmalarında, tüm koşullar sabit tutulurken, olasılıksal simülasyon çalışmalarında en az bir koşul değişken olarak ele alınır (Cohen, Manion, & Morrison, 2007). Bu bağlamda, bu çalışma ayrıca bir olasılıksal simülasyon çalışmasıdır.

2.2.Verilerin Üretilmesi

Bu çalışmada, verilerin oluşturulmasında Madde Tepki Kuramı (MTK) temelinde çok kategorili puanlama sistemine sahip testler için geliştirilen genelleştirilmiş kısmi kredi modeli kullanılmıştır. Bu çalışmanın kapsamında kullanılan veriler, erişimi ücretsiz olan WinGen3 programı kullanılarak üretilmiştir (K. T. Han, 2007). Çalışma boyunca veri üretme sürecinde sabit tutulan koşullar Tablo 1'de sunulmuştur.

Tablo 1. Sabit Tutulan Koşullar

	En küçük değer	En büyük değer
a parametresi	0.2	2.8
b parametresi	-3	3
c parametresi	-	-
Yetenek	-3	3
Tekrar Sayısı	25	
Model	Genelleştirilmiş Kısmi Kredi Modeli	
Kategori Sayısı	3	

Değişimlenen tutulan koşullar Tablo 2'de gösterilmektedir.

Bu çalışmada, kişilerin yetenek düzeyleri, a ve b parametrelerinin dağılımları ve tekrar sayısı sabit tutulmuştur. Ayrıca, tüm veri setleri genelleştirilmiş kısmi kredi modeline (GPCM) uygun olarak üretilmiştir. Bu seçim, GPCM'nin başarı testlerinde kullanılmasının ve a parametresinin de dahil edilmesinin avantajları nedeniyle yapılmıştır (de Ayala, 2009). Veriler, çok kategorili puanlama sistemine sahip olan GPCM'ye uygun olarak 0 (yanlış), 1 (kısmen doğru) ve 2 (tam doğru) puanlama şeklinde üretilmiştir.

Verilerin genel popülasyonu temsil etmesi amacıyla, b parametresi için geniş bir aralık seçilmiş ve yetenek parametresi de b parametresine uygun bir şekilde üretilmiştir. Benzer şekilde, a parametresi için de geniş bir aralık tercih edilmiştir.

Bu araştırmada simülasyon çalışmalarından elde edilen sonuçların daha genellemeye uygun olabilmesi için yinelemeler yapılmıştır. Madde Tepki Kuramı temelli çalışmalarda, 25 yinelemenin yeterli olduğu literatürde belirtilmektedir (Harwell, Stone, Hsu, & Kirisci, 1996). Bu nedenle, bu çalışmada her alt koşul için 25 yineleme kullanılmıştır.

Tablo 2. Değişimlenen Koşullar

	Değerler				
Örneklem Büyüklüğü	50	100	300	1000	3000
Test Uzunluğu	10	20	30	40	50

Tablo 2 incelendiğinde iki değişimlenen koşul olduğu görülmektedir: örneklem büyüklüğü ve test uzunluğu. Bu çalışmada Klasik Test Kuramı ve Madde Tepki Kuramı'na dayalı eşitleme yöntemleri karşılaştırılması amaçlandığı için örneklem büyüklükleri geniş bir aralıkta seçilmiştir. Türkiye'de uygulanan geniş ölçekli testler göz önünde bulundurularak test uzunluğu olarak 10, 20, 30, 40 ve 50 maddeden oluşan testler ile çalışma yürütülmüştür.

Değişimlenen koşullar göz önünde bulundurulduğunda; örneklem büyüklüğü koşulu (5 koşul) ve test uzunluğu koşulu (5 koşul) olmak üzere 25 koşul olmaktadır. Bu durumda üretilen her koşul için bir adet X formu ve bir adet Y formu olmak üzere toplam 50 test formu oluşturulmuştur. Çalışma verileri 25 tekrarla üretildiğinde 1250 adet cevap örüntüsü elde edilmiştir.

2.3.Verilerin Analizi

Bu araştırmada kullanılan veriler WinGen3 programı aracılığıyla üretilmiştir (K. T. Han & Hambleton, 2007). Verilerin üretilmesinde MTK temelinde çok kategorili puanlanan testler için oluşturulan genelleştirilmiş kısmi kredi modeli (GPCM) kullanılmıştır. Çalışma kapsamında üretilen verilerin KTK'ye dayalı eşitleme yöntemleri ile eşitlenmesinde R 4.1.2 programı (R core team, 2021) Equate paketi (Albano, 2016) kullanılmıştır. MTK'ye dayalı test eşitleme yöntemleri için ise IRTEQ programı kullanılmıştır. Eşitleme süreci boyunca verilerin düzenlenmesi ve bootstrap hata miktarlarının hesaplanması Microsoft Excel aracılığıyla yapılmıştır.

3.Bulgular

Tablo 3. Madde Sayıları ve Örneklem Büyüklüklerine Göre Hata Miktarları

Madde Sayısı	Örneklem Büyüklüğü	ESH				
		O	D	E.Y.	O-O	O-S
	50	1.51	1.56	1.35	1.94	1.99

	100	1.73	1.76	1.50	2.84	1.95
10	300	1.62	1.66	1.47	1.99	1.67
	1000	1.47	1.49	1.34	2.71	1.76
	3000	1.42	1.41	1.27	1.71	1.30
	50	4.18	4.30	3.58	3.09	2.58
	100	2.75	2.84	2.51	3.44	2.61
20	300	2.59	2.64	2.22	3.50	2.10
	1000	2.00	2.00	1.70	3.09	1.63
	3000	2.52	2.48	2.24	2.65	2.35
	50	4.38	4.44	3.72	4.39	2.98
	100	3.45	3.38	2.87	4.76	2.40
30	300	2.73	2.70	2.41	3.78	2.64
	1000	2.65	2.63	2.39	3.67	2.16
	3000	2.12	2.13	1.91	3.55	2.12
	50	5.03	5.22	4.43	4.99	3.24
	100	3.90	3.89	3.42	4.95	2.71
40	300	3.18	3.25	2.86	4.79	2.98
	1000	2.76	2.76	2.49	5.06	3.29
	3000	2.68	2.68	2.36	5.05	2.41
	50	6.80	7.14	6.01	5.12	3.68
	100	4.42	4.50	4.04	5.03	4.27
50	300	3.94	3.96	3.44	6.46	3.44
	1000	3.30	3.29	2.91	4.96	3.64
	3000	2.98	3.00	2.69	4.71	2.23

10 maddeden oluşan testlerin eşitlenmesi sırasında elde edilen hatalar incelendiğinde, beş farklı örneklem büyüklüğü için en küçük hatanın eşit yüzdelikli eşitleme yöntemi tarafından üretildiği görülmüştür. Diğer taraftan, en yüksek hataların ortalama-ortalama ve ortalama-standart sapma yöntemleri kullanılarak elde edildiği belirlenmiştir. Klasik Test Kuramı'na dayalı ortalama, doğrusal ve eşit yüzdelikli eşitleme yöntemlerinde

örneklem büyüklüğü sırasıyla 300, 1000 ve 3000 olduğunda hata miktarlarının azaldığı gözlemlenirken, Madde Tepki Kuramı'na dayalı ortalama-ortalama ve ortalama-standart sapma yöntemlerinden elde edilen hata miktarlarının önce artma sonra azalma eğiliminde oldukları gözlemlenmiştir.

20 maddeden oluşan testlerin eşitlenmesinden elde edilen hatalar incelendiğinde, en düşük hata eşit yüzdelikli ve ortalama-standart sapma yöntemlerinden elde edilirken en yüksek hatayı veren yöntemlerin doğrusal eşitleme ve ortalama-ortalama eşitleme yöntemleri olduğu belirlenmiştir. Genel olarak en düşük hatanın 20 madde için de eşit yüzdelikli eşitleme yönteminden elde edildiği söylenebilir.

30 maddeden oluşan testlerin eşitlenmesi sırasında elde edilen hatalar incelendiğinde, en düşük hata Madde Tepki Kuramı (MTK)'ye dayalı eşitleme yöntemlerinden ortalama-standart sapma ve Klasik Test Kuramı (KTK)'ye dayalı eşitleme yöntemlerinden eşit yüzdelikli eşitleme yöntemi kullanılarak elde edilmiştir. Diğer yandan, en yüksek hatalar MTK'ye dayalı eşitleme yöntemlerinden ortalama-ortalama ve KTK'ye dayalı eşitleme yöntemlerinden doğrusal eşitleme yöntemi kullanılarak elde edilmiştir.

40 ve 50 maddeden oluşan testlerin eşitlenmesi sırasında elde edilen hatalar incelendiğinde, diğer test uzunluklarına ait bulgulara benzer şekilde en düşük hatanın eşit yüzdelikli ve ortalama-standart sapma yöntemleri kullanılarak elde edildiği görülmüştür. Ayrıca, en yüksek hataların doğrusal eşitleme ve ortalama-ortalama eşitleme yöntemleri kullanılarak elde edildiği belirlenmiştir. Bu sonuçlar, farklı test uzunlukları için de benzer eşitleme eğilimlerinin olduğunu göstermektedir.

4.Sonuç ve Tartışma

Bu araştırmada farklı örneklem büyüklüğü ve test uzunluğu koşullarında, Klasik Test Kuramı'na dayalı ortalama, doğrusal, eşit yüzdelikli eşitleme yöntemleri ve Madde Tepki Kuramı'na dayalı ortalama-ortalama ve ortalama-standart sapma yöntemlerinin karşılaştırılması amaçlanmıştır.

Literatürde test eşitleme süreci için yöntemlerin karşılaştırıldığı benzer çalışmalar bulunmaktadır. Kolen (1981), Klasik Test Kuramı'na dayalı eşitleme yöntemi olan eşit yüzdelikli eşitlemeyi ve Madde Tepki Kuramı'na dayalı 1PL ve 3PL yöntemlerini karşılaştırmış ve eşit yüzdelikli eşitleme ile 3PL eşitleme yöntemlerinin 1PL'den daha iyi sonuçlar verdiğini bulmuştur. Kolen ve Whitney (1982), yaptıkları çalışmada eşit yüzdelikli eşitlemenin daha iyi sonuçlar verdiğini bulmuşlardır. Şahhüseyinoğlu (2006), İngilizce puanlarını doğrusal, eşit yüzdelikli ve Rasch eşitleme yöntemlerine göre eşitlemiş ve eşit yüzdelikli eşitlemenin Rasch yöntemi kadar güvenilir sonuçlar verdiğine ulaşmıştır. Kilmen ve Demirtaşlı (2012), farklı örneklem büyüklüğüne sahip gruplarda MTK'ye dayalı test eşitleme yöntemlerini karşılaştırmış ve test karakteristik eğrisi yöntemlerinin daha az hata ile eşitleme yaptığını ve örneklem büyüklüğünün artmasının hesaplanan hata miktarını azalttığını belirtmişlerdir. Öztürk ve Anıl (2012) 2008 yılının ilkbahar ve sonbahar dönemlerinde yapılan ALES'in sayısal testinden elde edilen puanlarla doğrusal ve eşit yüzdelikli eşitleme yöntemlerini kullanarak test eşitleme işlemini yapmışlar ve eşit yüzdelikli eşitleme yönteminin daha düşük hatalı eşitleme yaptığı dolayısıyla bu yöntemin kullanılmasının daha doğru sonuçlar vereceği yargısına ulaşmışlardır. Gök ve Kelecioğlu (2014), test uzunluğu ve örneklem büyüklüğü koşullarına bağlı olarak test eşitleme çalışması yapmışlar ve ortalama-ortalama yönteminin daha az hata ile eşitleme yaptığını ve en fazla hata veren yöntemin ortalama-standart sapma yöntemi olduğunu gözlemlemişlerdir. Ayrıca, madde sayısının ve örneklem büyüklüğünün arttığı durumlarda daha kararlı eşitlemeler yapıldığını ve eşitlemenin standart hatasının daha düşük olduğunu belirtmişlerdir. Uysal ve Kilmen (2016), 3PL ve GPCM'ye

göre üretilen veriler ile test eşitleme çalışması yapmışlar ve test karakteristik eğrisi yönteminin daha az hata ile eşitleme yaptığını belirtmişlerdir. Pektaş ve Kılınç (2016) PISA 2012 Türkiye örnekleminden elde edilen verilerle ortalama eşitleme, doğrusal eşitleme ve eşit yüzdelikli eşitleme yöntemlerini kullanılarak test eşitleme çalışması yürütmüş ve eşitlemeden elde edilen hataların karşılaştırılması sonucunda puan dağılımlarının en doğru şekilde eşitlenmesini sağlayan eşitleme yönteminin eşit yüzdelikli eşitleme yöntemi olduğu sonucuna ulaşılmışlardır. Nisa ve Retnawati (2018), MTK'ye dayalı yöntemleri karşılaştırmış ve en az hata ile eşitleme yapan yöntemin ortalama-ortalama yöntemi olduğunu bulmuşlardır.

İlgili literatür incelediğinde, Klasik Test Kuramı'na (KTK) ve Madde Tepki Kuramı'na (MTK) dayalı eşitleme yöntemlerinin karşılaştırıldığı ve genellikle daha az hata ile eşitleme yapan yöntemin KTK yöntemlerinden eşit yüzdelikli eşitleme ve MTK yöntemlerinden de ortalama-ortalama ve test karakteristik eğrisi yöntemleri olduğunu gözlemlenmiştir. Bu çalışmada da literatürle uyumlu olarak eşit yüzdelikli eşitlemenin daha az hata ile eşitleme yaptığı sonucuna ulaşılmıştır. Ancak ortalama-ortalama yöntemi ile yapılan eşitlemelerde diğer yöntemlere göre hesaplanan hata miktarının daha fazla gözlenmiştir ve çalışma bu yönüyle literatürden farklılaşmaktadır.

Örneklem büyüklüğünün değişken olarak alındığı çalışmalarda, örneklem büyüklüğünün arttıkça hesaplanan hata miktarının azaldığı sonucuna ulaşılmaktadır. Bu sonuç test eşitleme konusunda yapılan diğer çalışmalarla benzerlik göstermektedir. Örneklem büyüklüğü dışında, test uzunluğunun değişken olarak ele alındığı çalışmalarda, madde sayısının arttıkça hata miktarının azaldığı gözlemlenmektedir. Literatürde gözlemlenen bulguların aksine bu çalışmada farklı bir sonuç elde edilmiş ve test uzunluğunun arttıkça yöntemlerden elde edilen hatanın arttığı görülmüştür. Bu sonuçlar, eşitleme sürecinde örneklem büyüklüğü ve test uzunluğunun etkilerinin karmaşık olabileceğini ve çalışma koşullarına bağlı olarak değişebileceğini göstermektedir.

5.Öneriler

Bu araştırmadan elde edilen bulgulara dayanarak test eşitleme alanında yapılabilecek araştırmalar için aşağıdaki öneriler sunulabilir:

1. Gerçek Verilerle Yapılan Araştırmalar: Bu çalışma bilgisayar ortamında üretilmiş simülatif verilerle yürütülmüştür. Benzer bir çalışma gerçek bir test uygulamasından elde edilen verilerle yapılabilir. Bu, laboratuvar koşullarının gerçek dünya uygulamalarına nasıl uyarlandığının daha iyi anlaşılmasına yardımcı olabilir.

2. Farklı Model ve Kategori Sayıları: Bu çalışmada veriler MTK'ye dayalı genelleştirilmiş kısmi puan modeli kullanılarak incelenmiştir. Farklı modellerin ve farklı kategori sayılarının eşitleme süreçlerine etkisini inceleyen çalışmalar yapılabilir.

3. Parametre Aralıkları: Bu çalışmada madde ve yetenek parametreleri için geniş bir aralık kullanılmıştır. Farklı parametre aralıklarının test eşitleme sonuçlarına etkisini inceleyen çalışmalar yapılabilir.

4. Hata Kestirim Yöntemleri: Bu çalışmada hata kestirim yöntemlerinden bootstrap kullanılmıştır. Farklı hata kestirim yöntemlerinin kullanıldığı çalışmalar yapılabilir ve bu yöntemlerin farklı eşitleme süreçleri için etkileri karşılaştırılabilir.

Kaynakça | References

Albano, A. D. (2016). {equate}: An {R} package for observed-score linking and equating. Journal of Statistical Software, 74(8), 1-36. doi:10.18637/jss.v074.i08

Anastasi, A. (1976). Psychological testing (4. Baskı). London: Collier Macmillan Publishers

Cohen, L., Manion, L., & Morrison, K. (2007). Internet-based research and computer usage. İçinde Research Methods in Education (6. Baskı, s. 226-252). New York: Routledge.

de Ayala, R. J. (2009). The theory and practıce of item response theory. New York: The Guilford Press.

Efron, B. (1982). The jackknife, the bootstrap and other resampling plans. Retrieved from Philadelphia:

Fraenkel, J. R., Wallen, N. E., & Hyun, H. H. (2012). The nature of research. İçinde How to design and evaluate research in education (8. Baskı). New York,NY: McGraw-Hill Education.

Gök, B., & Kelecioğlu, H. (2014). Denk olmayan gruplarda ortak madde deseni kullanılarak madde tepki kuramına dayalı eşitleme yöntemlerinin karşılaştırılması. Mersin Üniversitesi Eğitim Fakültesi Dergisi, 10(1), 120-136.

Han, K. T., & Hambleton, R. K. (2007). User's manual: WinGen (642). Retrieved from Amherst, MA:University of Massachusetts:

Harwell, M., Stone, C. A., Hsu, T., & Kirisci, L. (1996). Monte Carlo studies in Item Response Theory. Applied Psychological Measurement, 20(2), 101-125.

Kilmen, S., & Demirtaşlı, N. (2012). Comparison of test equating methods based on item response theory according to the sample size and ability distribution Procedia - Social and Behavioral Sciences, 46, 130-134. doi: 10.1016/j.sbspro.2012.05.081

Kolen, M. J. (1981). Comparison of traditional and item response theory methods for equating tests. Journal of Educational Measurement Spring, 18(1), 1-11.

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

Kolen, M. J., & Whitney, D. R. (1982). Comparison of four procedures for equating the tests of general educational development. Journal of Educational Measurement, 19(4), 279-293.

Linn, L. L. (1993). Linking results of distinct assessments. Applied Measurement in Education, 6(1), 83-102. doi:10.1207/s15324818ame0601_5

Livingston, S. A. (2014). Equipercentile equating. İçinde Equating Test Scores (Without IRT) (2. Baskı, s. 17-23): Educational Testing Service.

Loyd, B. H., and Hoover, H. D. (1980). Vertical Equating Using the Rasch Model. Journal of Educational Measurement, 17 (3), 179-193.

Marco, G. L. (1977). Item Characteristic Curve Solutions to Three Intractable Testing Problems. Journal of Educational Measurement, 14 (2), 139- 160.

Mislevy, R. J. (1992). Linking educational assessments: concepts, issues, methods and prospects. Retrieved from Princeton,NJ:

Muraki, E., Hombo, C. M., & Lee, Y. (2000). Equating and linking of performance assessments. Psychological Measurement, 24(4), 325-337.

Nisa, C., & Retnawati, H. (2018). Comparing the methods of vertical equating for the math learning achievement tests for junior high school students. REiD (Research and Evaluation in Education), 4(2), 164-174.

Öztürk, N., & Anıl, D. (2012). Akademik personel ve lisansüstü eğitimi giriş sınavı puanlarının eşitlenmesi üzerine bir çalışma. Eğitim ve Bilim, 37(165), 180-193.

Pektaş, S., & Kılınç, M. (2016). PISA 2012 matematik testlerinden iki kitapçığın gözlenen puan eşitleme yöntemleri ile eşitlenmesi Mehmet Akif Ersoy Üniversitesi Eğitim Fakültesi Dergisi (40), 432-444. doi:10.21764/efd.49376

R core team. (2021). R: a language and environment for statistical computing. Retrieved from https://www.R-project.org/

Sansivieri, V., Wiberg, M., & Matteucci, M. (2017). A review of test equating methods with a special focus on IRT-based approaches. Statistica, 77(4), 329-352.

Şahhüseyinoğlu, D. (2006). İngilizce yeterlik sınavı puanlarının üç farklı eşitleme yöntemine göre karşılaştırılması. Hacettepe Üniversitesi Eğitim Fakültesi Dergisi (31), 115-125.

Uysal, İ., & Kilmen, S. (2016). Comparison of item response theory test equating methods for mixed format tests. International Online Journal of Educational Sciences, 8(2), 1-11.

Beyan ve Açıklamalar (Disclosure Statements)

Araştırmacıların katkı oranı beyanı / Contribution rate statement of researchers:

1. Yazar/First author % 51

2. Yazar/Second author % 49

Yazarlar tarafından herhangi bir çıkar çatışması beyan edilmemiştir. (No potential conflict of interest was reported by the authors).

Structured Abstract

The trait to be measured in education and psychology is often not observable. Therefore, various measurement tools are used to measure these characteristics. In education, many different characteristics, from affective characteristics to measuring maximum performance, are evaluated with measurement tools. These tests are important for teachers, educational administrators, and other decision-makers. In addition, many tests are administered throughout the country and the results of these tests affect important decisions in the lives of individuals. Therefore, test results must be reliable and valid. One of the reasons for using different test forms is the need to be able to take the same test more than once, which makes it impossible to use the same test form continuously. Therefore, different test forms should be similar, and this similarity can be controlled by test association. Test association can be categorized under five headings: equating, calibration, statistical moderation, estimation, and social moderation. Equating is an association method that allows scores obtained from one test to be used instead of scores obtained from another test. It involves a more demanding statistical process than other methods, but it is one of the most powerful association methods. When the assumptions are met, two equalized test scores can be used interchangeably. Test equating methods can be categorized under two headings: methods based on Classical Test Theory and methods based on Item Response Theory. In this study, mean, linear, and equipercentile equating methods from Classical Test Theory and mean-mean and mean-standard deviation methods from Item Response Theory were used in the equating process. Mean equating is based on the assumption that the mean scores of two test forms should be equal. Equipercentile equating assumes that the percentile rankings of the scores of the two forms of the test measuring the same trait are the same. The scores are transformed so that the scores of one form are in the other form's percentile. Linear equating assumes that the scores of the two test forms are equal except for their means and standard deviations. This method offers a flexible approach where scores can vary across scale units. The mean-mean method performs a test equating using the item difficulty parameter and the discrimination parameter. These parameters are used to convert scores into scale scores. The mean-standard deviation method equalizes scores using the standard deviation and means of the item difficulty parameter. The bootstrap method was used in the study to calculate the standard error of equating. This study, it was aimed to compare the errors of equalizing test forms using various test equating methods (mean, equal equipercentile, linear, mean-mean, and mean-standard deviation) under varying test lengths and sample sizes. Accordingly, tests with different test lengths (10, 20, 30, 40, and 50 items) and groups with varying sample sizes (50, 100, 300, 1000, and 3000) were studied to answer the question of which test equating method has the lowest error rate.

In this study, the WinGen3 program was used in the production of the data because of its free access. The data were generated in the form of 0 (incorrect), 1 (partially correct), and 2 (fully correct) scores in accordance with the generalized partial credit model, which is one of the Item Response Theory models with a polytomously scored system. The ability levels of the individuals, the distributions of the parameters a and b, and the number of repetitions were kept constant. For the data to be representative of the general data, a wide range was chosen for the b parameter and the ability parameter was generated in the same range as the b parameter. Similarly, a wide range was also preferred for a parameter. Since this study aims to compare test equating methods based on Classical Test Theory and Item Response Theory, the sample sizes were chosen from a wide range. In addition, different test lengths were examined by considering the large-scale tests used in Turkey. These tests include 10, 20, 30, 40 and 50 items. The study data were generated with 25 repetitions and 1250 response patterns were obtained in this case.

When the results obtained from the research were examined, it was seen that the lowest error was obtained from the equipercentile equating method in Classical Test Theory methods and from the mean-standard deviation method in Item Response Theory methods. When compared with the related literature, it is seen that there are similar results in the studies. The highest error was generally calculated in the linear equating method in Classical Test Theory methods and in the mean-mean method in Item Response Theory methods. When analyzed in the context of the changed conditions, it was concluded that the error decreased as the sample size increased. While this result is consistent with the literature, the increase in error as the number of items in the test increases is a result that this study differs from the literature.

Based on the findings obtained from this study, it is recommended to conduct studies with real data, to prefer different numbers of models and/or categories, and to choose narrower parameter ranges for future research in the field of test equating.