Araştırma Makalesi
Research Article

https://doi.org/10.21764/maeuefd.1366213

Madde Tepki Kuramına Dayalı Test Eşitlemede Ölçek Dönüştürme Yöntemlerinin, Ortak Madde Oranının ve Madde Ayırt Ediciliğinin Eşitleme Hatasına Etkisi*

The Effect of Calibration Methods, Common Item Ratio and Item Discrimination On Equating Error in Test Equating Based on Item Response Theory

Yıldız YILDIRIM [1] **Tuba GÜNDÜZ** [2] **Fazilet Gül iNCE ARACI** [3]

[1] Aydın Adnan Menderes Üniversitesi, Eğitim Fakültesi, Aydın, Türkiye
[2] Muğla Sıtkı Koçman Üniversitesi, Eğitim Fakültesi, Muğla, Türkiye
[3] Karamanoğlu Mehmetbey Üniversitesi, Eğitim Fakültesi, Karaman, Türkiye

Özet: Bu araştırmanın amacı Madde Tepki Kuramına madde tepki kuramına dayalı test eşitlemede ortak madde oranının, madde ayırt ediciliğinin ve ölçek dönüştürme yönteminin eşitlemenin standart hatasına etkisini incelemektir. Bu temel araştırma kapsamında, iki madde ayırt edicilik düzeyi (orta ($a_{log\text{-}ort}$ = 0,00) ve yüksek ($a_{log\text{-}ort}$ = 0,50)) × üç ortak madde oranı (%10, %20 ve %30) × dört ölçek dönüştürme yöntemi (Stocking & Lord, Haebara, Ortalama-Standart sapma ve Ortalama-Ortalama) olmak üzere toplam 24 simülasyon koşulu bulunmaktadır ve her koşul için 100 tekrar yapılmıştır. Araştırma sonucunda eşitlemenin standart hatasının en düşük olduğu ölçek dönüştürme yönteminin Stocking & Lord'un yöntemi, en yüksek olduğu yöntemin ise ortalama-ortalama yöntemi olduğu belirlenmiştir. Ayrıca hem ortak madde oranı hem de madde ayırt ediciliği arttıkça eşitlemenin standart hatasının azaldığı sonucuna varılmıştır. Ek olarak eşitlemenin standart hatasının en düşük olduğu koşul, ölçek dönüştürme yönteminin Stocking & Lord, ayırt edicilik düzeyinin yüksek ve ortak madde oranının %30 olduğu koşuldur. Eşitlemenin standart hatasının en yüksek olduğu koşul ise ölçek dönüştürme yönteminin ortalama-ortalama, ayırt edicilik düzeyinin orta ve ortak madde oranının %10 olduğu koşuldur. Son olarak eşitlemenin standart hatasının ortak madde oranına göre madde ayırt ediciliğinden daha çok etkilendiği sonucuna varılmıştır. Bu sonuçlara dayalı olarak araştırmacılara ve test geliştiricilere öneriler sunulmuştur.

Anahtar Kelimeler: *Ölçek dönüştürme yöntemleri, ortak madde oranı, madde ayırt ediciliği, MTK'ya dayalı test eşitleme*

Abstract: The purpose of this research is to examine the effects of common item ratio, item discrimination and calibration method on equating error in test equating based on Item Response Theory. Within the scope of this basic research, there are a total of 24 simulation conditions [two item discrimination levels (medium ($a_{log\text{-}mean}$ = 0.00), high ($a_{log\text{-}mean}$ = 0.50), high (a_{mean} = 1.00)) × three common item ratios (10%, 20% and 30%) × four calibration methods (Stocking & Lord, Haebara, Mean-Sigma and Mean-Mean)] and 100 replications were made for each condition. As a result of the research, it was determined that the calibrating method with the lowest equating error was Stocking & Lord's method and the method with the highest equating error was the mean-mean method. Also, it was concluded that as both the common item ratio and item discrimination increased, the equating error decreased. Additionally, the lowest equating error was found when the calibration method was Stocking & Lord, the discrimination level was high and the common item ratio was 30%. The highest equating error was observed when the calibration method was mean-mean, the discrimination level was medium, and the common item ratio was 10%. Finally, it was concluded that the equating error was more affected by item discrimination than the common item ratio. Based on these results, recommendations are made to researchers and test developers.

Keywords: *Calibration methods, common items ratio, item discrimination, IRT test equating*

* Bu çalışma, 7. Uluslararası Eğitimde ve Psikolojide Ölçme ve Değerlendirme Kongresi'nde sözlü olarak sunulan "Madde Tepki Kuramına Dayalı Test Eşitlemede Ortak Madde Oranının ve Madde Ayırt Ediciliğinin Eşitleme Hatasına Etkisi" başlıklı bildirinin genişletilmiş hâlidir.

Sorumlu Yazar/Corresponding author: yildizyldrm@gmail.com
Geliş/Received: *25/09/2023*
Revizyon/Revised: *16/01/2026*
Kabul/Accepted: *09/02/2026*

1. Giriş

Günümüzde test puanları, bireysel ve kurumsal düzeylerde önemli kararlar alırken bilgi sağlamak için sıklıkla kullanılmaktadır (Kolen & Brennan, 2014). Türkiye'de de geniş ölçekli sınavlarda, kurumlara giriş ve sertifika sınavlarında test puanlarının kullanımı söz konusudur. Bu sınavlardan farklı zamanlarda elde edilen test puanlarının bazı durumlarda karşılaştırılabilir olması gerekmektedir. Örneğin Türkiye'de lisans mezunu öğrenciler lisansüstü eğitimlerine başvurmak için Yabancı Dil Bilgisi Seviye Tespit Sınavı'na (YDS) girmektedir. Bu sınav yılda iki kez uygulanmakta olup her sınavda farklı maddelerden oluşan farklı formlar kullanılmaktadır. Ayrıca YDS'nin geçerlilik süresi kurumdan kuruma değişmekle birlikte genel olarak beş yıldır (ÖSYM, 2023). Aynı lisansüstü programa başvuran ve farklı yıllarda YDS'ye giren öğrencilerin birbirleriyle karşılaştırılması, maddelerin farklı olması nedeniyle madde güçlükleri de farklı olabileceği için uygun olmayabilir. Çünkü aynı amacı ölçmek için hazırlanan ve farklı maddeler kullanılan bu tür sınavların güçlük düzeyi farklı olabileceğinden aynı yetenek düzeyindeki bireyler farklı puanlar alabilir. Bu gibi durumlarda bir katılımcının başka bir katılımcıdan daha zor ya da daha kolay bir test formu almasıyla test adaletinin sağlanamaması durumu söz konusu olabilir. Bu sorunla başa çıkabilmek için ise aynı testin pek çok formu üzerinde karşılaştırılabilir puanlar elde etmek için test formlarının eşitlenmesine ihtiyaç duyulur. Böylelikle test puanlarının karşılaştırılabilirliği için test eşitleme yapılarak adil bir değerlendirmeye olanak sağlanabilir. Test eşitleme, Crocker ve Algina (1986) tarafından iki testten eşdeğer puanlar oluşturma süreci olarak tanımlanmaktadır. Test eşitlemede farklı veri toplama desenleri kullanılmaktadır ve bunlardan biri ortak maddeli eşdeğer olmayan gruplar desenidir.

Ortak maddeli eşdeğer olmayan gruplar deseni bazı maddelerin ortak olduğu iki farklı test formunun iki farklı gruba uygulandığı test eşitleme desenidir. Ortak maddeli eşdeğer olmayan grup deseninde, ortak maddeler içsel ya da dışsal olmak üzere iki türlü olabilir (Kolen & Brennan, 2014). Ortak maddeler içsel olduğunda bu maddelerden elde edilen puan test puanına dâhil edilirken, dışsal olduğunda dâhil edilmemektedir. Literatürde ortak madde sayısı tartışmalı bir konu olarak ele alınmış ve araştırmalar yapılmıştır. Bu araştırmaların birçoğunun ortak yanı, ortak madde sayısı arttıkça eşitleme hatasının azaldığı sonucuna ulaşmalarıdır (Bastari, 2000; Cohen & Kim, 1998; Hanson & Beguin, 2002; Kaskowitz, 1998; Kim & Cohen, 2002; Meng, 2012; Tate, 2000; Uysal, 2014; Walker & Kim, 2009; Yang & Houang, 1996). Ancak bu araştırmaların yanında ortak madde sayısının artmasının eşitleme hatası üzerinde önemli bir farklılık yaratmadığı ya da azalmasıyla da yeterli test eşitleme sonuçları sağlayabildiği sonucuna varan araştırmalar bulunmaktadır (He, 2011; Hills, Subhiyah & Hirsch, 1988). Örneğin, Çokluk-Bökeoğlu, Uçar ve Balta (2022)'nın çalışmalarında bazı simülasyon koşullarında ortak madde oranının artmasıyla hatanın azaldığı, bazı simülasyon koşullarında ise ortak madde oranının azalmasıyla hatanın azaldığı gözlemlenmiştir. Bu araştırmacılar çalışmalarında örneklem büyüklüğü, test uzunluğu, ortak madde oranı, Madde Tepki Kuramı (MTK) parametre kestirim modelleri ve yetenek dağılımları gibi değişkenlerin farklı simülasyon koşullarını ele almış olup dört farklı ölçek dönüştürme yöntemini de test eşitlemede kullanmıştır. İnal ve Anıl (2018), belirledikleri simülasyon koşulları altında gerçekleştirilen test eşitlemelerde değişen madde fonksiyonu gösteren ortak madde oranı arttıkça grup değişmezliğinin bir ölçüsü olan REMSD (Root Expected Mean Square Difference)'ın da manidar şekilde arttığını tespit etmişlerdir. Özetle farklı koşullar altında ortak madde oranındaki değişiklikler eşitleme hatasını farklı yönde etkileyebilmektedir. Bu nedenle eşitleme hatasının düşük olması için en uygun ortak madde oranının ne olduğunu belirlemek amacıyla yapılan araştırmalar devam etmektedir. Dolayısıyla farklı koşullar altında eşitleme hatasının en aza indirileceği ortak madde oranının belirlenmesinin araştırmaya değer bir konu olduğu düşünülmektedir. Bununla birlikte maddelerin ayırt edicilik düzeylerinin farklı olduğu koşullar altında ortak madde oranının eşitleme hatasını nasıl etkilediğine yönelik çalışmalara rastlanılmamıştır. Bu

araştırmanın amaçlarından biri de "Ortak Maddeli Eşdeğer Olmayan Gruplar Deseni" kullanarak farklı ayırt edicilik düzeyi koşulları altında değişen ortak madde oranının eşitleme hatasına etkisinin incelenmesidir.

Farklı test eşitleme desenleri olduğu gibi farklı test eşitleme yöntemleri de bulunmaktadır. Bu yöntemler Klasik Test Kuramına (KTK) ve MTK'ya dayalı olmak üzere iki ana başlıkta yer almaktadır. MTK'ya dayalı test eşitleme yöntemlerinin KTK'nın sahip olduğu bazı sınırlılıkları (parametrelerin test ve gruba bağımlı olması, ölçmenin standart hatasının tek değer olması vs.) nedeniyle birtakım üstünlükleri söz konusudur. Özellikle MTK'ya dayalı test eşitleme yöntemlerinin KTK'ya dayalı test eşitleme yöntemlerine göre ölçme kesinliğinin daha yüksek ve tutarlı olduğu bilinmektedir (Alordiah & Oji, 2024; Hambleton & Jones, 2003, Han, Kolen & Pohlmann, 1997; Kolen & Brennan, 2014). Diğer yandan MTK'ya dayalı test eşitleme yöntemleri bilgisayar ortamında bireye uyarlanmış testler, çok aşamalı testler gibi MTK'ya dayalı uygulamalarda da kullanılabilmesi avantajına sahiptir (Seo, 2017). Literatürde KTK'ya ve MTK'ya dayalı test eşitlemelerin eşitleme hatasına etkisini inceleyen araştırmalarda da bu üstün yanlarını kanıtlayan bulgulara rastlanmıştır (Caldwell, 1984; Cho, 2007; Chen, 2001; Hills vd., 1988; Kelecioğlu, 1994; Mutluer & Çakan, 2023; Peterson, Cook & Stocking, 1983; Suanthong, 1998; Şahhüseyinoğlu, 2005). Bu üstün yanları ve literatürde sıklıkla kullanılması nedeniyle bu araştırmada MTK'ya dayalı test eşitleme yöntemi ele alınmıştır. MTK'ya dayalı test eşitleme yönteminin aşamaları bulunmaktadır: (i) eşitleme deseninin (veri toplama deseninin) belirlenmesi, (ii) kestirilen madde ve yetenek parametrelerinin ortak bir ölçeğe dönüştürülmesi ve (iii) test puanlarının eşitlenmesidir (Cook & Eignor, 1991).

MTK'ya dayalı test eşitleme yapmadan önce test formlarına ait maddelerin parametrelerinin kestirilmesi gerekir ve bu işleme madde kalibrasyonu denir. Literatürde iki tür kalibrasyon yöntemi vardır: ayrı kalibrasyon ve eş zamanlı kalibrasyon. Test formlarına ait madde ve yetenek parametresi kestirimleri her bir test formu üzerinde ayrı ayrı yapıldığında kestirilen bu parametreler, farklı ölçek üzerinde yer alırlar (Kolen & Brennan, 2014). Bunun üstesinden gelmek için ölçek dönüştürme yöntemlerinden faydalanılmaktadır. MTK'ya dayalı test eşitleme yapılmadan önce formlardan elde edilen bu parametrelerin aynı MTK ölçeği üzerine getirilmesi gereklidir. Literatürde bulunan ölçek dönüştürme yöntemleri moment yöntemleri ve karakteristik eğri yöntemleri olarak ikiye ayrılmaktadır. Moment yöntemleri; Ortalama-Standart sapma (OS - Marco, 1977) ve Ortalama-Ortalama (OO - Loyd & Hoover, 1980) yöntemleridir. Karakteristik eğri yöntemleri ise Haebara (HA - Haebara, 1980), Stocking–Lord (SL - Stocking & Lord, 1983) yöntemleridir. OS yönteminde, iki testteki ortak maddelerin kestirilen *b* parametrelerinin standart sapması ve ortalaması kullanılmaktayken OO yönteminde, ortak maddelere ait *b* parametrelerinin yanında *a* parametrelerinin ortalaması da kullanılmaktadır. HA yöntemi ile madde karakteristik eğrileri arasındaki farkı belirlemek için, her bir madde karakteristik eğrisinin farkının kareleri toplamı alınırken SL yöntemi ile, her bir madde karakteristik eğrisindeki farkın toplamlarının karesi alınarak madde parametreleri tek bir ölçeğe yerleştirilir (Haebara, 1980; Kolen & Brennan, 2014; Marco, 1977; Loyd & Hoover, 1980; Stocking & Lord, 1983). Bu araştırmada eşitlenecek olan test formalarına ait madde parametreleri ayrı kalibrasyon yapılarak kestirildiğinden ölçek dönüştürme yöntemlerinin de ele alınması ihtiyaç haline gelmiş olup ölçek dönüştürme yöntemlerinin tamamı (OO, OS, HA, SL) araştırma kapsamında ele alınmıştır. Ölçek dönüştürme yöntemlerini karşılaştıran pek çok araştırma alanyazında mevcuttur (Baker & Al-Karni, 1991; Çokluk-Bökeoğlu vd., 2022; Dilek, Atalay-Kabasakal & Gören, 2025; Gök & Kelecioğlu, 2014; Gündüz, 2015; Hanson & Béguin, 2002; Li, Jiang & von Davier, 2012; Öztürk-Gübeş, 2014; Uysal & Kilmen, 2016; Wang & Kolen, 2016; Yurtçu & Güzeller, 2022; Zor, 2023).

Bu araştırmada, alanyazındaki diğer araştırmalardan farklı olarak madde ayırt ediciliğinin test eşitleme hatasına etkisi incelenmiştir. Literatürde madde ayırt ediciliğinin eşitleme hatasına etkisini inceleyen bir araştırmaya rastlanmadığı için bu araştırmanın önemli olduğu düşünülmektedir. Bununla birlikte farklı ayırt edicilik düzeyi koşulları altında ölçek

dönüştürme yöntemlerinin ve ortak madde oranının eşitlemenin standart hatasını nasıl etkilediği de incelenmiştir. MTK'ya dayalı gözlenen puan eşitlemede ayırt edicilik düzeyinin daha önce alanyazında ele alınmamış olmasıyla birlikte bu koşulların da farklı ayırt edicilik düzeylerinde hata bağlamında nasıl tepki göstereceğinin bilinmediği söylenebilir. Ayrıca test geliştirme sürecinde ortaya çıkabilecek farklı ayırt edicilik düzeylerine sahip maddelerden oluşan testlere yönelik performansı yüksek olan ortak madde oranının ve ölçek dönüştürme yönteminin belirlenmesinin pratik anlamda alanyazına katkı sunacağı düşünülmektedir. Diğer bir deyişle bu araştırmanın sonuçları, test geliştiriciler ortak maddeli eşdeğer olmayan gruplar deseni kullandığında farklı ayırt edicilik düzeylerine sahip maddelere karşı ne tür önlemler alabileceğini kestirmesine kılavuzluk edebilir. Bu araştırmanın test geliştiricilerinin ve uygulayıcılarının yanı sıra test eşitleme alanında çalışan araştırmacılara da önemli bilgiler sunacağı öngörülmektedir. Diğer yandan MTK'nın eğitimde ve psikoloji alanında yaygın olarak uygulanmasıyla birlikte test eşitlemede ölçek dönüştürme yöntemleri de alan uzmanları ve uygulayıcılarının giderek daha fazla ilgisini çekmektedir. Bir test eşitleme çalışmasında ortak madde oranından ve madde ayırt edicilik düzeylerinden bağımsız bir şekilde ölçek dönüştürme yöntemlerinin kullanımı, MTK ile test eşitlemenin avantajlarının hayata geçirilmesini potansiyel olarak tehdit edebileceğinden, test eşitleme yapılmadan önce ortak madde oranı ve madde ayırt edicilik düzeyleri ile ölçek dönüştürme yönteminin birlikte değerlendirilmesinin önemli olduğu düşünülmektedir. Araştırmacılar bu çalışmadan yola çıkarak yürüttükleri simülasyona ya da gerçek veriye dayalı araştırmalarda ayırt edicilik düzeyini koşul olarak ele alarak teoriyi güçlendirmeye katkı sağlayabilirler. Bu doğrultuda araştırmanın amacı madde tepki kuramına dayalı test eşitlemede ortak madde oranının, madde ayırt ediciliğinin ve ölçek dönüştürme yöntemlerinin test eşitleme hatasına etkisini incelemektir.

2. Yöntem

2.1. Araştırma Modeli

Bu araştırma ile MTK'ya dayalı gözlenen puan eşitlemede farklı ortak madde oranı, farklı madde ayırt edicilik düzeyleri ve farklı ölçek dönüştürme yöntemlerinin, eşitleme hatalarına etkisinin incelenmesi amaçlanmıştır. Çalışma belirli koşullar altında karşılaştırma amacı taşıdığından simülasyon verileri ile yapılan bir temel araştırma niteliği taşımaktadır (Kothari, 2004).

2.2. Verilerin Üretilmesi

Araştırmada kullanılan veriler R ortamında üretilmiştir. Araştırmada iki madde ayırt edicilik düzeyi (orta ($a_{log\text{-}ort}$ =0,00) ve yüksek ($a_{log\text{-}ort}$ =0,50)) × üç ortak madde oranı (%10, %20 ve %30) × dört ölçek dönüştürme yöntemi (Stocking & Lord, Haebara, Ortalama-Standart sapma ve Ortalama-Ortalama) olmak üzere toplam 24 koşul bulunmaktadır. Alanyazındaki diğer çalışmalar da incelenerek replikasyon (tekrar) sayısı 100 olarak belirlenmiştir (Kim & Cho, 2020; Uçar & Sünbül, 2024).

Birey parametreleri dağılımı eşdeğer olmayan gruplarda normal dağılım baz alınarak üretilmiştir. Araştırmada yer alan X formu (yeni form) ile Y formunun (eski form) ortalamaları sırayla 0 ve 0,50; standart sapmaları ise 1 olarak ele alınmıştır. Araştırma kapsamında X formu (yeni form) Y formuna eşitlenmiştir. Veri setleri 1500 bireyden oluşurken üretilen test 40 maddeden (Kim & Cho, 2020; Uysal & Kilmen, 2016) oluşmaktadır. Bununla birlikte ortak maddeler içseldir. Bu doğrultuda %10 ortak madde oranı için 36 madde + 4 ortak madde, %20 ortak madde oranı için 32 madde + 8 ortak madde ve %30 ortak madde oranı için 28 madde + 12 ortak madde olarak madde dağılımları belirlenmiştir. Koşul olarak ele alınan ortak madde oranlarının belirlenmesinde alanyazında sıklıkla kullanılan ve önerilen oranlar belirlenmiştir (Çokluk-Bökeoğlu vd., 2022; Uysal, Şahin-Kürşad & Kılıç, 2022; Wolkowitz & Wright, 2019). Araştırma

kapsamında kullanılan verilerin üretilmesinde, testlerin yapısının genellikle çoktan seçmeli (1-0) olması nedeniyle şans faktörünü de göz önünde bulundurmak için üç parametreli lojistik model (3PLM) kullanılmıştır. Araştırmanın amacı doğrultusunda *a* parametrelerinin üretilmesinde iki farklı durum ele alınmış olup üretilen parametrelerin logaritmik ortalamaları 0 ve 0,5; yani aritmetik ortalamaları yaklaşık 1,13 ve 1,87'dir. Bu parametrelerin üretilmesinde Gök ve Kelecioğlu'nun (2014) da çalışmasında ele aldığı gibi standart sapması 0,5 olan log-normal dağılımlardan yararlanılmıştır. Teorik olarak -∞ ile +∞ arasında değer alabilen *a* parametresi uygulamalarda sıklıkla 0 ile 2 arasında değerler almakta ve 0'a yaklaştığı durumda bireyleri ayırt etme gücü düşmekte bu parametrenin sayısal değeri yükseldikçe ayırt edicilik de yükselmektedir (Hambleton, Swaminathan & Rogers, 1991). Araştırma kapsamında üretilen *a* parametrelerinin aritmetik ortalamaları dikkate alındığında Baker (2001)'in ölçüt değerlerine paralel olacak şekilde *a* parametresinin düzeylerine ilişkin koşullar "orta" ve "yüksek" ayırt edicilik düzeyi olarak tanımlanmıştır. Verilerin üretilmesinde b parametreleri, literatürde sıklıkla ele alındığı gibi, tüm koşullarda ortalaması 0, standart sapması 1 olmak üzere normal dağılıma uyacak şekilde üretilmiştir (Andersson & Wiberg, 2017; Bulut, 2013; Leôncio, Wiberg & Battauz, 2023). Şans parametresi olarak da tanımlanan c parametreleri ise 0,20-0,30 aralığında olmak üzere tek biçimli dağılımdan üretilmiştir. Ayrıca araştırmada diğer bir koşul olan ölçek dönüştürme yöntemlerinden OO, OS, HA ve SL de ele alınmıştır. Bu yöntemlerin ele alınmasında alanyazında sıklıkla kullanılmaları etkili olmuştur (Çokluk-Bökeoğlu vd., 2022; Uyar, Aksekioğlu & Öztürk-Gübeş, 2018). Son olarak her bir koşul için bir X formu bir de Y formu oluşturulmuştur. Bu doğrultuda toplam 6 koşul×4 ölçek dönüştürme yöntemi×100 tekrar sayısı olmak üzere toplam 2400 test eşitleme yapılmıştır. Tüm koşulların düzeyleri ve üretilen verilerin özellikleri Tablo 1'de verilmiştir.

Tablo 1

Simülasyon koşullarının düzeyleri ve üretilen verilerin özellikleri

		Kategoriler
Koşullar	*a* parametresi	Orta, $a \sim L$ (μ=0,00, σ=0,50) Yüksek, $a \sim L$ (μ=0,50, σ=0,50)
	Ortak madde oranı	%10, 36 madde + 4 ortak madde %20, 32 madde + 8 ortak madde %30, 28 madde + 12 ortak madde
		Özellikler
Üretilen Veri	MTK Modeli	3 parametreli lojistik model
	b parametresi	$b \sim N$ (μ=0,00, σ=1,00)
	c parametresi	$c \sim U$ (0,20-0,30)
	Birey parametreleri	$\theta_X \sim N$ (μ=0,00, σ=1,00) $\theta_Y \sim N$ (μ=0,50, σ=1,00) n = 1500
Ölçek Dönüştürme Yöntemleri	Ortalama-Ortalama Ortalama-Standart sapma Haebera Stocking & Lord	

2.3. Verilerin Analizi

Verilerin analizinde R (v.4.4.1) ortamında equateIRT paketi (Battauz, 2015) kullanılmıştır. Ölçek dönüştürme işlemi OO, OS, HA ve SL yöntemleri ile yapıldıktan sonra her bir tekrar sayısı ve her bir koşul için MTK'ya dayalı olarak ortak maddeli eşdeğer olmayan gruplar desenine göre gözlenen puan test eşitleme yapılmıştır. Gözlenen puan eşitlemede eşitlenen test formlarının aynı yapıyı ölçmesi ve formlar arasında ölçülen özelliğin değişmemesi temel varsayımlardır. Gözlenen puan eşitleme, gerçek puan eşitlemeye göre daha yaygın kullanılmaktadır (Kolen & Brennan, 2014).

Her bir koşul ve her bir tekrar için test eşitleme yapıldıktan sonra eşitlemenin standart hatası değerleri incelenmiştir. Eşitlemenin standart hatası tekrar sayısına bölünerek ortalama hata değerleri elde edilmiştir ve bu değerler karşılaştırılarak yorumlanmıştır. Hesaplamada kullanılan test eşitlemenin ortalama standart hatasına (ESH) ait formül Eşitlik 1'de verilmiştir (Ogasawara, 2001; Ogasawara, 2003; Zhang, 2021).

$$\underline{X}_{\text{ESH}}(\hat{\eta}) = \frac{\sqrt{\frac{\partial_\eta}{\partial_\alpha} acov(\hat{\alpha}) \frac{\partial_\eta}{\partial_\alpha}}}{N} \qquad (1)$$

Eşitlik 1'de N tekrar sayısını belirtmektedir. ESH ise eşitlemenin standart hatasını temsil etmektedir ve $\frac{\partial_\eta}{\partial_\alpha}$ ve $acov(\hat{\alpha})$ ile hesaplanmaktadır. Eşitlikte $\frac{\partial_\eta}{\partial_\alpha}$, gözlenen puanlar için madde parametreleri ile ölçek dönüştürme katsayılarının kısmi türevleri olan bir vektörü temsil ederken, $acov(\hat{\alpha})$ ise gözlenen puanlar için $\hat{\alpha}$'nın asimptotik varyans kovaryans matrisini göstermektedir.

3. Bulgular ve Yorum

Araştırmanın amacı doğrultusunda gözlenen puan eşitlemede OO, OS, HA ve SL ölçek dönüştürme yöntemleri kullanıldığında orta ve yüksek ayırt edicilik düzeylerinin ve %10, %20 ve %30 ortak madde oranlarının eşitleme hatasını nasıl etkilediği incelenmiştir. Bu doğrultuda elde edilen ortalama eşitleme hataları Tablo 2'de verilmiştir.

Tablo 2

Her bir koşula göre eşitlemenin standart hata değerlerinin ortalamaları

	Ortak madde oranı	Orta ayırt edicilik	Yüksek ayırt edicilik
SL	%10	0,344	0,279
	%20	0,313	0,222
	%30	0,291	0,214
HA	%10	0,354	0,277
	%20	0,325	0,240
	%30	0,320	0,227
OS	%10	0,311	0,305
	%20	0,286	0,281
	%30	0,246	0,243
OO	%10	0,377	0,319
	%20	0,354	0,308
	%30	0,324	0,255

Bu araştırmada MTK'ya dayalı test eşitlemede farklı ölçek dönüştürme yöntemleri, farklı ayırt edicilik düzeyleri ve ortak madde oranının eşitleme hatasını nasıl etkilediği incelenmiştir. Tablo 2, değişen ortak madde oranı bağlamında incelendiğinde tüm ölçek dönüştürme ve ayırt edicilik düzeyleri koşullarında en yüksek eşitleme hatası ortak madde oranı %10 olduğunda, en düşük eşitleme hatası ise ortak madde oranı %30 olduğunda görülmüştür. İlk olarak orta ayırt edicilik düzeyi koşulu incelendiğinde SL ve HA ölçek dönüştürme yöntemleri için ortak madde oranının %10'dan %20'ye yükselmesiyle hatanın yaklaşık 0,029 ila 0,031 aralığında azaldığı, OS ve OO yöntemleri için ise yaklaşık 0,025 ila 0,033 aralığında azaldığı görülmüştür. Ortak madde oranı %20'den %30'a yükseldiğinde ise SL ölçek dönüştürme yöntemi için hatanın yaklaşık 0,022 azaldığı, HA ölçek dönüştürme yöntemi için 0,005 düzeyinde azaldığı görülmüştür. OS ve OO yöntemlerinde ortak madde oranı %20'den %30'a yükseldiğinde ise eşitleme hatasının yaklaşık 0,030-0,040 aralığında azaldığı görülmüştür. Bu doğrultuda orta ayırt edicilik düzeyi koşulu altında ortak madde oranının düzeyinin

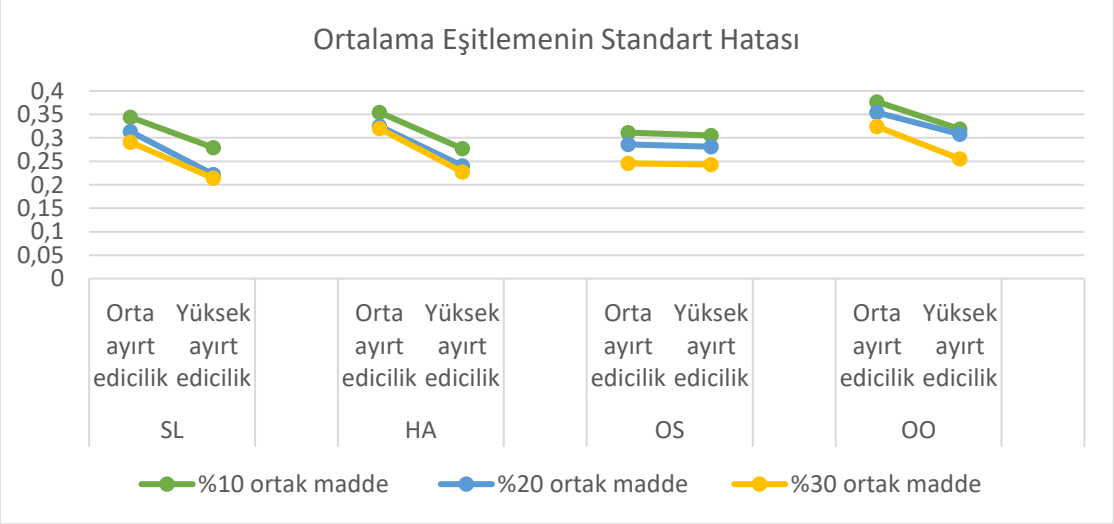
değişmesinden en az etkilenen ölçek dönüştürme yönteminin HA olduğu görülmüştür. Ayrıca orta ayırt edicilik koşulu altında tüm ortak madde oranları için en düşük eşitleme hatalarının da OS yönteminde olduğu belirlenmiştir.

Yüksek ayırt edicilik düzeyinde genel olarak SL ölçek dönüştürme yönteminin %20 ve %30 ortak madde oranı koşullarında en düşük eşitleme hatalarını verdiği söylenebilir. Ortak madde oranı %10'dan %20'ye yükseldiğinde hatanın yaklaşık 0,057 azaldığı, %20'den %30'a yükseldiğinde ise hatanın yaklaşık 0,008 azaldığı görülmüştür. Bu yöntemden sonra %20 ve %30 ortak madde oranı koşullarında en düşük eşitleme hatalarına sahip olan yöntem HA ölçek dönüştürme yöntemidir. Ölçek dönüştürme yöntemlerinden HA'da ortak madde oranın %10'dan %20'ye yükseldiğinde eşitleme hatasının yaklaşık 0,037, %20'den %30'a yükseldiğinde ise yaklaşık 0,013 düştüğü gözlemlenmiştir. Yüksek düzey ayırt edicilik koşulu altında OO ve OS yöntemleri ele alındığında ise bu yöntemlerin ortak madde oranı değişimine benzer tepkiler verdikleri görülmüştür. Bu yöntemlerin her ikisinde de ortak madde oranı %10'dan %20'ye yükseldiğinde eşitleme hatasının yaklaşık 0,011 ila 0,024 aralığında azaldığı, %20'den %30'a yükseldiğinde ise 0,038 ila 0,053 aralığında azaldığı gözlemlenmiştir.

Genel itibariyle tüm ölçek dönüştürme yöntemleri ve her iki ayırt edicilik koşulları altında ortak madde oranı arttıkça eşitlemenin standart hatasının azaldığı görülmektedir. Bu bulgu tüm ayırt edicilik düzeylerinde maddelerin bulunduğu testlerde ortak madde sayısı arttığında eşitleme hatasının azaldığını kanıtlamaktadır. Bununla birlikte sonuçlar genel olarak değerlendirildiğinde, aradaki farkın pratik olarak manidar olmayabileceği unutulmamalıdır. Farklı ortak madde oranlarında ve ölçek dönüştürme yöntemleri koşulları altında ayırt edicilik düzeylerine göre ortalama eşitleme standart hatalarına ilişkin grafik Şekil 1'de verilmiştir.

Şekil 1

Farklı ölçek dönüştürme yöntemleri ve ortak madde oranları için ayırt edicilik düzeylerine göre eşitleme hata değerleri



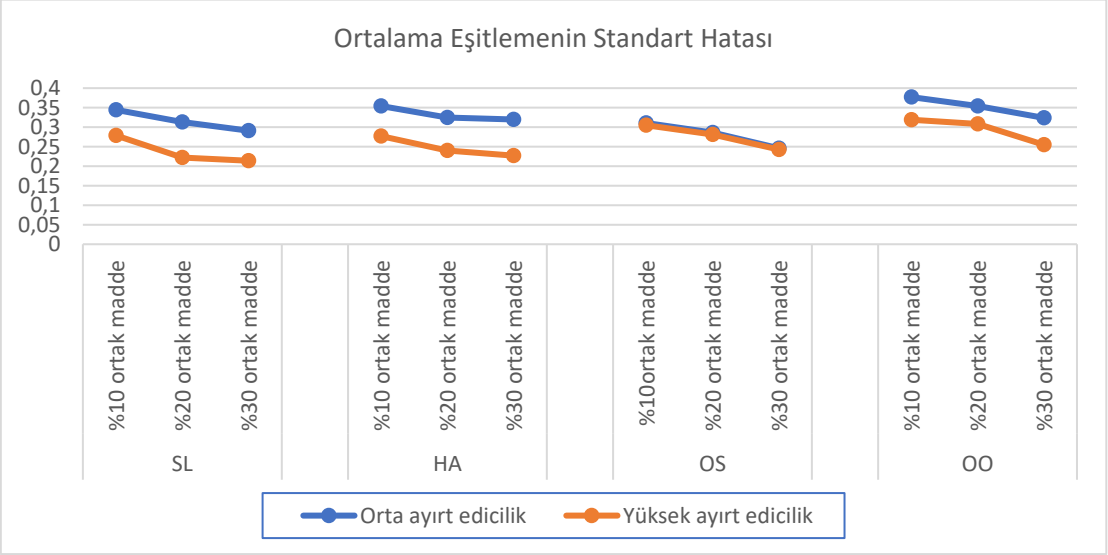
Şekil 1 incelendiğinde de her iki ayırt edicilik düzeyinde ve tüm ölçek dönüştürme yöntemlerinde ortak madde oranının artmasının eşitleme hatasını azalttığı görülmektedir. Bu azalma oranı her iki ayırt edicilik düzeyinde küçük düzeyde gibi görünse de %10 ortak madde oranı artışıyla yaklaşık %1,538 (0,005/0,325) ile %23,223 (0,049/0,211) aralığında değişmektedir. Yüksek ayırt edicilik düzeyinde SL ve HA yöntemlerinde ortak madde oranının %20 ve %30 olmasının arasında eşitlemenin standart hatası açısından genel olarak büyük bir değişiklik görünmemektedir. Ayrıca OS yöntemi için orta ayırt edicilik düzeyinde ortak madde oranının %20'den %30'a yükselmesiyle, SL yöntemi için yüksek ayırt edicilik düzeyinde ortak madde oranının %10'dan %20'ye yükselmesiyle ve OO yöntemi için ortak madde oranının %20'den

%30'a yükselmesiyle standart hatanın diğer koşullara göre daha fazla düşüş gösterdiği görülmektedir. Son olarak tüm ölçek dönüştürme yöntemlerinde ve her iki ayırt edicilik düzeyinde en düşük eşitleme hatalarının ortak madde oranının %30 olduğu ve 12 ortak maddenin olduğu koşulda olduğu söylenebilir.

Araştırma kapsamında farklı ölçek dönüştürme yöntemleri koşulları altında madde tepki kuramına dayalı gözlenen puan test eşitlemede %10, %20 ve %30 ortak madde oranları için madde ayırt ediciliğinin eşitleme hatasını nasıl etkilediği de incelenmiştir. Tablo 2 incelendiğinde ortak madde oranı %10 olduğunda SL ölçek dönüştürme yöntemi için ayırt edicilik düzeyi orta düzeyden yüksek düzeye ulaştığında eşitleme hatasının yaklaşık 0,065 azaldığı görülmektedir. SL ölçek dönüştürme yöntemi için ortak madde oranı %20 olduğunda ayırt edicilik düzeyi orta düzeyden yüksek düzeye ulaştığında eşitleme hatasının yaklaşık 0,091 azaldığı görülmektedir. Ortak madde oranı %30 olduğunda ise ayırt edicilik düzeyi orta düzeyden yüksek düzeye ulaştığında eşitleme hatasının değerinin yaklaşık 0,077 azaldığı görülmektedir. HA ölçek dönüştürme yöntemi için ise ortak madde oranı %10 olduğunda ayırt edicilik düzeyi orta düzeyden yüksek düzeye ulaştığında hatanın yaklaşık 0,077 azaldığı görülmüştür. Ortak madde oranının %20 ve %30 olduğu koşullar için de orta ayırt edicilik düzeyinden yüksek ayırt edicilik düzeyine ulaştığında sırasıyla yaklaşık 0,085 ve 0,093 azaldığı görülmüştür. OS yöntemi incelendiğinde ise tüm ortak madde oranı düzeylerinde ayırt ediciliğin yükselmesinin eşitleme hatası üzerinde büyük bir etkisi olmadığı gözlemlenmiştir. OO ölçek dönüştürme yönteminde ise %10 ortak madde oranı koşulunda ayırt edicilik düzeyi orta düzeyden yüksek düzeye ulaştığında eşitleme hatasında yaklaşık 0,058 azalma görülmüştür. Ortak madde oranı %20 koşulunda ise ayırt edicilik düzeyi orta düzeyden yüksek düzeye ulaştığında eşitleme hatasında yaklaşık 0,046 azalma gözlemlenmiştir. Son olarak %30 ortak madde oranı koşulunda da ayırt edicilik düzeyi orta düzeyden yüksek düzeye ulaştığında yaklaşık 0,069'luk bir azalma görülmüştür. Bu bulgular tüm ortak madde oranlarında test çiftlerinde madde ayırt edicilik düzeyi arttığında eşitleme hatasının azaldığını kanıtlamaktadır. Her iki ayırt edicilik düzeyi için ortak madde oranlarına göre ortalama eşitleme standart hatalarına ilişkin grafik Şekil 2'de verilmiştir.

Şekil 2

Farklı ölçek dönüştürme yöntemleri ve ayırt edicilik düzeyleri için ortak madde oranlarına göre eşitleme hata değerleri



Şekil 2 incelendiğinde de tüm ölçek dönüştürme yöntemleri ve ortak madde oranlarında test çiftlerindeki madde ayırt edicilik düzeylerinin artmasının eşitleme hatasını azalttığı görülmektedir. Bu azalma oranı tüm ortak madde oranlarında

küçük düzeyde gibi görünse de yaklaşık %1,22 (0,003/0,246) ile %29,07 (0,091/0,313) aralığında değişmektedir. SL yöntemi için en büyük değişim ise %20 ortak madde oranında ayırt edicilik düzeyinin orta düzeyden yüksek düzeye ulaştığı durumdadır. HA ölçek dönüştürme yöntemi incelendiğinde, SL yöntemine ilişkin sonuçlara benzer sonuçlar üretildiği görülmektedir. OS yönteminde ise tüm ortak madde oranı koşulları için ayırt edicilik düzeyinin değişmesinin eşitleme hatasını neredeyse hiç etkilemediği görülmektedir. Ayrıca OO yönteminin, tüm ortak madde oranı koşullarında her iki ayırt edicilik düzeyinde de en yüksek eşitleme hatasına sahip olduğu gözlemlenmiştir.

Bu bulgular doğrultusunda, tüm ölçek dönüştürme yöntemlerinde ve tüm ortak madde oranlarında test çiftlerinde maddelerin ayırt ediciliğinin daha yüksek olmasının genel itibariyle eşitleme hatasını büyük oranda azalttığı söylenebilir. Bu nedenle farklı ortak madde oranları kullanıldığında ölçek dönüştürme yöntemi fark etmeksizin madde ayırt ediciliklerinin olabildiğince yüksek olması önemlidir. Son olarak tüm koşullar altında en düşük eşitleme hatalarının yüksek ayırt edicilikte olduğu yani *a* parametresinin logaritmik ortalamasının 0,5 olduğu koşulda (aritmetik ortalamasının yaklaşık 1,87 olduğu koşul) olduğu söylenebilir.

Sonuçlar birlikte değerlendirildiğinde eşitleme hatasının en düşük olduğu koşulun SL ölçek dönüştürme yönteminin kullanıldığı ve en yüksek ortak madde oranı ve ayırt ediciliğin bulunduğu (%30 ortak madde oranı ve $a_{log\text{-}ort}$=0,50) koşul olduğu görülmüştür. Eşitleme hatasının en yüksek olduğu koşul ise OO ölçek dönüştürme yönteminin kullanıldığı, en düşük ortak madde oranı ve orta ayırt edicilik düzeyinde (%10 ortak madde oranı ve $a_{log\text{-}ort}$=0,00) bulunmuştur. Hem ortak madde oranının artması hem de ayırt ediciliğin artması eşitleme hatasını azaltmaktadır. Ancak eşitleme hatasının ayırt edicilik düzeyindeki değişimden ortak madde oranına nazaran daha çok etkilendiği görülmektedir. Bunun yanı sıra ölçek dönüştürme yöntemlerinden de en düşük eşitleme hatalarının yüksek ayırt edicilik düzeyinde genel itibariyle SL ölçek dönüştürme yönteminde olduğu ve bunu HA ölçek dönüştürme yönteminin takip ettiği görülmüştür. Orta ayırt edicilik düzeyinde ise tüm ortak madde oranı koşullarında en düşük eşitleme hatası OS ölçek dönüştürme yönteminde bulunmuştur. Bu ayırt edicilik düzeyinde eşitleme hatasının en düşük olduğu ikinci ölçek dönüştürme yöntemi ise SL'dir. Araştırma kapsamında ele alınan koşullar altında eşitleme hatasının en yüksek olduğu ölçek dönüştürme yönteminin ise OO ölçek dönüştürme yöntemi olduğu tespit edilmiştir.

4. Sonuç, Tartışma ve Öneriler

Bu araştırmada MTK'ya dayalı test eşitlemede ölçek dönüştürme yönteminin, madde ayırt edicilik düzeyinin ve ortak madde oranının eşitleme hatasını nasıl etkilediği incelenmiştir. Araştırmanın bulguları doğrultusunda ortak madde oranı arttıkça eşitleme hatasının azaldığı sonucuna varılmıştır. Bu araştırmanın sonuçlarına benzer olarak ortak madde oranının artmasıyla eşitleme hatasının azaldığı sonucunu elde eden araştırmalar literatürde mevcuttur (Kaskowitz, 1998; Bastari, 2000; Kim & Cohen, 2002; Hanson & Beguin, 2002). Kumlu (2019), çalışmasında elde ettiği bulgulara göre ortak madde oranının değişmesinin eşitleme sonuçlarına ilişkin hata değerlerine fazla yansımadığını belirtmiştir. Araştırmacının bu sonuçlara ulaşmasında, ele aldığı diğer koşulların eşitleme hataları üzerinde daha baskın düzeyde etkisinin olmasından kaynaklanabileceği düşünülmektedir. Bu araştırmada da çalışma kapsamında ele alınan madde ayırt ediciliğinin eşitleme hatası üzerinde daha baskın bir etkisi olduğu görülmüştür. Öyle ki ortak madde oranının eşitleme hatası üzerindeki etkisinin madde ayırt edicilik düzeyine göre daha az olduğu tespit edilmiştir. Hills vd. (1988) ise bu araştırmanın aksine ortak madde oranının belli bir noktada artmasının hatayı azalttığını bulmuştur. Ancak araştırmacılar ortak madde oranı %33 olduğunda mutlak ortalama farkın %50 olduğu duruma göre daha düşük olduğunu tespit etmişlerdir. Diğer yandan çoğunlukla test çiftlerindeki madde ayırt edicilik düzeylerinin artmasının eşitleme

hatasında daha büyük değişimlere yol açtığı sonucuna varılmıştır. Örneğin, SL yöntemi için ortak madde oranının %10 olduğu durumda orta ayırt edicilik düzeyinin orta ayırt edicilik düzeyiyle eşitleme hatası farkı, madde ayırt ediciliğinin orta olduğu durumda ortak madde oranının %20 olduğu koşulun %10 olduğu koşula eşitleme hatası farkından büyüktür. Ortak madde oranında olduğu gibi madde ayırt ediciliğinde de düzey arttıkça eşitleme hatasının azaldığı görülmüştür.

Araştırma kapsamında ele alınan ölçek dönüştürme yöntemine göre sonuçlar incelendiğinde SLyöntemi (Stocking & Lord, 1983) yüksek ayırt edicilik altındaki her bir koşulda en iyi performansı göstermiştir. Buna paralel olarak Kolen ve Brennan (1995), pek çok çalışmada SL yönteminin iyi performans gösterdiğini belirtmiştir. Buna benzer sonuçlara ulaşan çalışmalara literatürde sıklıkla rastlanmaktadır (French, 1996; Hanson & Béguin, 1999; Hanson & Béguin, 2002; Karkee & Wright, 2004; Kilmen, 2010; Kim & Kolen, 2006; Meng, 2012; Speron, 2009; Tian, 2011; Uysal, 2014; Yurtçu & Güzeller, 2022). Her ne kadar yüksek ayırt edicilik düzeyinde SL yöntemi en iyi performansı göstermiş olsa da OS yöntemi ayırt ediciliğin orta olduğu tüm koşullarda en düşük eşitleme hatasına sahip yöntem olmuştur. Bunun nedeninin OS yönteminin yalnızca madde güçlük parametresine dayalı hesaplanması olduğu düşünülmektedir. Dolayısıyla madde ayırt ediciliği yüksek olmadığı durumlarda OS yönteminin kullanılması önerilebilir. Ayrıca OO yönteminin tüm koşullarda en yüksek eşitleme hatasına sahip olduğu sonucuna varılmıştır. Bu araştırmanın aksine Gök ve Kelecioğlu (2014) yaptıkları araştırma kapsamında genel itibariyle en iyi eşitleme sonuçlarının OO yöntemi kullanılarak elde edilebildiğini belirtmiştir.

Bu çalışmanın sonuçlarına dayalı olarak, ortak madde oranın artırılması ve formlarda yer alan maddelerin ayırt edicilik düzeylerinin yüksek olması önerilmektedir. Ancak bu çalışma kapsamında daha değerli bir öneri olarak, uygulayıcıların ortak madde oranlarını artırmayı ele almaktansa, maddelerin ayırt edicilik düzeylerinin yüksek olmasını göz önünde bulundurmaları gerektiği söylenebilir. Böylelikle yüksek düzeyde madde ayırt ediciliğine sahip maddeler kullanarak ortak madde oranı artırılmadan da hata oranları nispeten düşük test eşitleme gerçekleştirilebilir. MTK'ya dayalı eşitlemede yalnızca kuramın varsayımlarının kontrolü ile sınırlı kalmayıp maddelerin ayırt edicilik kestirimlerinin de incelenmesi ortak madde oranının belirlenmesinde önemli hale gelmektedir. Diğer yandan Hills vd. (1988) de minimum sayıda ortak madde kullanımının test güvenliği açısından avantaj sağladığını vurgulamışlardır. Herhangi bir ölçek dönüştürme yöntemini uygulamadan önce madde ayırt edicilikleri ile ortak madde oranını birlikte değerlendirmenin önemli olduğu sonucuna ulaşılmıştır. Diğer bir deyişle madde ayırt ediciliği ve ortak madde oranları birlikte değerlendirildikten sonra bir ölçek dönüştürme yöntemi seçilmesi gerektiği araştırma sonuçlarına dayalı olarak söylenebilmektedir. Bu nedenle test eşitleme yapacak araştırmacılara ve test merkezlerine amaçları ve koşullara uygun olan ölçek dönüştürme yönteminin seçilmesi önerilebilir. Diğer araştırmacılara yönelik ise, bu çalışmadan farklı olarak, yalnızca ortak maddelerin ayırt ediciliği farklılaştığında, maddelerin diğer parametreleri farklılaştığında veya farklı eşitleme yöntemleri kullanıldığında eşitleme hatasını araştırmak için başka çalışmalar yapmaları önerilebilir.

Çıkar Çatışması Bildirimi

Yazarların bildireceği herhangi bir çıkar çatışması bulunmamaktadır.

Üretken Yapay Zekâ Kullanımına Dair Beyan

Bu araştırmanın hazırlanmasında üretken yapay zekâ araçları kullanılmamıştır.

Etik Beyan

Bu araştırmada gerçek bireylere ait veriler kullanılmayıp veriler simülasyon yöntemiyle üretildiğinden çalışma, etik kurul onayı gerektirmeyen çalışmalar kapsamında yer almaktadır.

Yazar Katkıları

Bu araştırmanın tasarlanmasına tüm yazarlar katkı sağlamıştır. Yazarların katkı sağladığı bölümler aşağıda verilmiştir.

Yıldız Yıldırım: Makalenin ilk taslağının yazılması, kuramsal çerçevenin oluşturulması, veri simülasyonu, veri analizi, görselleştirme.

Tuba Gündüz: Makalenin ilk taslağının yazılması, kuramsal çerçevenin oluşturulması, veri simülasyonu, veri analizi, görselleştirme.

Fazilet Gül İnce Aracı: Makalenin ilk taslağının yazılması, kuramsal çerçevenin oluşturulması, görselleştirme.

Kaynaklar

Alordiah, C., & Oji, J. (2024). Test equating in educational assessment: A comprehensive framework for promoting fairness, validity, and cross-cultural equity. *Asian Journal of Assessment in Teaching and Learning*, *14*(1), 70-84. https://doi.org/10.37134/ajatel.vol14.1.7.2024

Andersson, B., & Wiberg, M. (2017). *Item response theory observed-score Kernel equating. Psychometrika, 82(1), 48–66.* https://doi.org/10.1007/s11336-016-9528-7

Baker, F. (2001). *The Basics of Item Response Theory*. ERIC Clearinghouse.

Baker, F. B. & Al-Karni, A. (1991). A comparison of two procedures for computing IRT equating coefficients. *Journal of Educational Measurement, 28*(2), 147–162. https://doi.org/10.1111/j.1745-3984.1991.tb00350.x

Bastari, B. (2000). *Linking multiple choice and constructed response items to a common proficiency scale* (Order No. 44070296). [Unpublished doctoral dissertation, University of Massachusetts Amherst]. https://doi.org/10.7275/16132240

Battauz, M. (2015). equateIRT: An R package for IRT test equating. *Journal of Statistical Software*, *68*(7), 1–22. https://doi.org/10.18637/jss.v068.i07

Bulut, O. (2013). *Between-person and within-person subscore reliability: Comparison of unidimensional and multidimensional IRT models* (Order No. 3589000). Available from ProQuest Dissertations & Theses Global. (1429501632). https://www.proquest.com/dissertations-theses/between-person-within-subscore-reliability/docview/1429501632/se-2

Caldwell, L. J. (1984). A *comparison of equating error in linear and Rasch model test equating methods* (Order No. 8427294). Available from ProQuest Dissertations & Theses Global. (303292556). https://www.proquest.com/dissertations-theses/comparison-equating-error-linear-rasch-model-test/docview/303292556/se-2

Chen, H. (2001). *Calibration of the ITBS Survey Test Battery to the complete test battery: A comparison of five linking methods* (Order No. 3009576). Available from ProQuest Dissertations & Theses Global. (304701160). https://www.proquest.com/dissertations-theses/calibration-itbs-survey-test-battery-complete/docview/304701160/se-2

Cho, Y. (2007). *Comparison of bootstrap standard errors of equating using IRT and equipercentile methods with polytomously -scored items under the common -item nonequivalent -groups design* (Order No. 3301690). Available from ProQuest Dissertations & Theses Global. (304858423). https://www.proquest.com/dissertations-theses/comparison-bootstrap-standard-errors-equating/docview/304858423/se-2

Cohen, A. S., & Kim, S. H. (1998). An investigation of linking methods under the graded response model. *Applied Psychological Measurement, 22(2), 116–130.* https://doi.org/10.1177/01466216980222002

Cook, L. L., & Eignor, D. R. (1991). IRT equating methods. *Educational measurement: Issues and practice*, *10*(3), 37-45. https://doi.org/10.1111/j.1745-3992.1991.tb00207.x

Çokluk-Bökeoglu, Ö., Uçar, A., & Balta, E. (2022). Madde tepki kuramına dayalı gerçek puan eşitlemede ölçek dönüştürme yöntemlerinin incelenmesi. *Ankara Üniversitesi Eğitim Bilimleri Fakültesi Dergisi, 55*(1), 1-36. https://doi.org/10.30964/auebfd.1001128

Dilek, H., Atalay-Kabasakal, K., & Gören, S. (2025). Examination of Scale Transformation and Test Equating Methods in Testlet Based Tests, *Kastamonu Education Journal, 33*(3), 658-671. https://doi.org/10.24106/kefdergi.1750267

French, D. C. (1996). *The utility of Stocking & Lord's equating procedure for equating norm-referenced and criterion-referenced tests with both dichotomous and polytomous components* (Order No. 9719355). Available from ProQuest Dissertations & Theses Global. (304284607). https://www.proquest.com/dissertations-theses/utility-stocking-amp-lords-equating-procedure/docview/304284607/se-2

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. https://dergipark.org.tr/tr/pub/mersinefd/issue/17393/181786

Gündüz, T. (2015). *Test eşitlemede Madde Tepki Kuramına dayalı yetenek parametresine yönelik ölçek dönüştürme yöntemlerinin karşılaştırılması.* Yayımlanmamış Yüksek lisans tezi, Gazi Üniversitesi, Eğitim Bilimleri Enstitüsü, Ankara.

Haebara, T. (1980). Equating logistic ability scales by a weighted least squares method. *Japanese Psychological Research,* 22, 144-149. https://www.jstage.jst.go.jp/article/psycholres1954/22/3/22_3_144/_pdf

Hambleton, R. K., & Jones, R. W. (1993). Comparison of classical test theory and item response theory and their applications to test development. *Educational Measurement: Issues and Practice*, *12*(3), 38-47. https://doi.org/10.1111/j.1745-3992.1993.tb00543.x

Hambleton, R.K., Swaminathan, H., & Rogers, H.J. (1991). *Fundamentals of item response theory.* Sage.

Han, T., Kolen, M., & Pohlmann, J. (1997). A comparison among IRT true-and observed-score equatings and traditional equipercentile equating. *Applied Measurement in Education*, *10*(2), 105-121. https://doi.org/10.1207/s15324818ame1002_1

Hanson, B. A. & Beguin, A. A. (1999). Separate versus concurrent Estimation of IRT item parameters in the common item equating design. *ACT Research Report Series*, 99: 8. https://www.act.org/content/dam/act/unsecured/documents/ACT_RR99-08.pdf

Hanson, B. A., & Béguin, A. A. (2002). Obtaining a common scale for item response theory item parameters using separate versus concurrent estimation in the common-item equating design. *Applied Psychological Measurement, 26*(1), 3–24. https://doi.org/10.1177/0146621602026001001

He, Y. (2011). *Evaluating equating properties for mixed-format tests* (Order No. 3461151). Available from ProQuest Dissertations & Theses Global. (879634637). https://www.proquest.com/dissertations-theses/evaluating-equating-properties-mixed-format-tests/docview/879634637/se-2

Hills, J. R., Subhiyah, R. G., & Hirsch, T. M. (1988). Equating minimum- competency tests: Comparisons of methods. *Journal of Educational Measurement, 25*(3), 221-231. https://doi.org/10.1111/j.1745-3984.1988.tb00304.x

İnal, H., & Anıl, D. (2018). Investigation of group invariance in test equating under different simulation conditions. *Eurasian Journal of Educational Research, 18*(78), 67-86. https://dergipark.org.tr/en/download/article-file/626510

Karkee, T. B. & Wright, K. R. (2004, April). Evaluation of linking methods for placing three parameter logistic item parameter estimates onto a one-parameter scale. Paper presented at the *Annual Meeting of the American Educational Research Association*, San Diego, California.

Kaskowitz, G. S. (1998). *The effect of error in item parameter estimates on linking and equating with the IRT test characteristic curve method* (Order No. 9836419). Available from ProQuest Dissertations & Theses Global. (304426493). https://www.proquest.com/dissertations-theses/effect-error-item-parameter-estimates-on-linking/docview/304426493/se-2

Kelecioğlu, H. (1994). *Öğrenci seçme sınavı puanlarının eşitlenmesi üzerine bir çalışma.* Yayımlanmamış doktora tezi, Hacettepe Üniversitesi, Sosyal Bilimler Enstitüsü, Ankara.

Kilmen, S. (2010). *Madde Tepki Kuramı'na dayalı test eşitleme yöntemlerinden kestirilen eşitleme hatalarının örneklem büyüklüğü ve yetenek dağılımına göre karşılaştırılması*. Doktora Tezi, Ankara Üniversitesi Eğitim Bilimleri Enstitüsü, Ankara.

Kim, K., Y. & Cho, U. H. (2020). Approximating bifactor IRT true-score equating with a projective item response model. *Applied Psychological Measurement, 44*(3), 215-218. https://doi.org/10.1177/0146621619885903

Kim, S., & Cohen, A. S. (2002). A comparison of linking and concurrent calibration under the graded response model. *Applied Psychological Measurement*, *26*(1), 25-41. https://doi.org/10.1177/0146621602026001002

Kim, S., & Kolen, M. J. (2006). Robustness to format effects of IRT linking methods for mixed-format tests. *Applied Measurement in Education*, *19*(4), 357-381. https://doi.org/10.1207/s15324818ame1904_7

Kumlu, G. (2019). *Test ve alt testlerde eşitlemenin farklı koşullar açısından incelenmesi.* Doktora Tezi, Hacettepe Üniversitesi Eğitim Bilimleri Enstitüsü, Ankara. http://www.openaccess.hacettepe.edu.tr:8080/xmlui/handle/11655/8877

Kolen, M. J., & Brennan, R. L. (1995). *Test equating methods and practices.* Springer.

Kolen, M. J. & Brennan, R. L. (2014). *Test equating, scaling, and linking* (3rd Ed.). Springer.

Kothari, C. R. (2004). *Research methodology: Methods and techniques* (2nd Ed.). New Age International.

Lee, W.C., & Ban, J.C. (2010). A comparison of IRT linking procedures. *Applied Measurement in Education, 23*(1), 23-48. https://doi.org/10.1080/08957340903423537

Leôncio, W., Wiberg, M., & Battauz, M. (2023). Evaluating equating transformations in IRT observed-score and kernel equating methods. *Applied Psychological Measurement, 47*(2), 123-140. https://doi.org/10.1177/01466216221124087

Li, D., Jiang, Y., & von Davier, A. A. (2012). The accuracy and consistency of a series of IRT true score equatings. *Journal of Educational Measurement, 49*(2), 167-189. https://doi.org/10.1111/j.1745-3984.2012.00167.x

Loyd, B. H. & Hoover, H. D. (1980). Vertical equating using the Rasch model. *Journal of Educational Measurement,* 17, 179-193. https://doi.org/10.1111/j.1745-3984.1980.tb00825.x

Marco, G. L. (1977). Item characteristic curve solutions to three intractable testing problems. *Journal of Educational Measurement,* 14, 139-160. https://doi.org/10.1111/j.1745-3984.1977.tb00033.x

Meng, Y. (2012). *Comparison of Kernel equating and Item Response Theory equating methods.* Available from Education Research Index. (1651827164; ED546709). https://www.proquest.com/dissertations-theses/comparison-kernel-equating-item-response-theory/docview/1651827164/se-2

Mutluer, C., & Çakan, M. (2023). Comparison of test equating methods based on classical test theory and item response theory. *Journal of Uludag University Faculty of Education, 36*(3), 866-906. https://doi.org/10.19171/uefad.1325587

Ogasawara, H. (2001). Item response theory true score equatings and their standard errors. *Journal of Educational and Behavioral Statistics*, *26*(1), 31-50. https://doi.org/10.3102/10769986026001031

Ogasawara, H. (2003). Asymptotic standard errors of IRT observed-score equating methods. *Psychometrika, 68(2), 193–211.* https://doi.org/10.1007/bf02294797

ÖSYM. (2023). *Yabancı dil bilgisi seviye tespit sınavı (YDS/1) kılavuzu*. https://dokuman.osym.gov.tr/pdfdokuman/2023/YDS-1/bkilavuz08032023.pdf

Öztürk-Gübeş, N. (2014). *Testlerin boyutluluğunun, ortak madde formatının, yetenek dağılımının ve ölçek dönüştürme yöntemlerinin karma testlerin eşitlenmesine etkisi.* Yayımlanmamış Doktora Tezi, Hacettepe Üniversitesi, Eğitim Bilimleri Ensitütüsü, Ankara. https://openaccess.hacettepe.edu.tr/xmlui/handle/11655/1761

Peterson, N. S., Cook, L. L., & Stocking, M. L. (1983). IRT versus conventional equating methods: a comparative study of scale stability. *Journal of Educational Statistics, 8*(2), 137-156. https://doi.org/10.3102/10769986008002137

Seo, D. G. (2017). Overview and current management of computerized adaptive testing in licensing/certification examinations. *Journal of Educational Evaluation for Health Professions*, *14,* 17. https://doi.org/10.3352/jeehp.2017.14.17

Speron, E. (2009). *A comparison of metric linking procedures in Item Response Theory* (Order No. 3370885). Available from ProQuest Dissertations & Theses Global. (304900819). https://www.proquest.com/dissertations-theses/comparison-metric-linking-procedures-item/docview/304900819/se-2

Stocking, M. L. & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement*, *7*(2), 201–210. https://doi.org/10.1177/014662168300700208

Suanthong, S. (1998). *An investigation of factors affecting test equating in latent trait theory* (Order No. 9841455). Available from ProQuest Dissertations & Theses Global. (304466901). https://www.proquest.com/dissertations-theses/investigation-factors-affecting-test-equating/docview/304466901/se-2

Şahhüseyinoğlu, D. (2005). *İngilizce yeterlik sınavı puanlarının üç farklı eşitleme yöntemine göre karşılaştırılması.* Yayımlanmamış doktora tezi, Hacettepe Üniversitesi Sosyal Bilimler Enstitüsü, Ankara.

Tate, R. (2000). Performance of a proposed method for the linking of mixed format tests with constructed response and multiple-choice items. *Journal of Educational Measurement, 37*(4), 329-346. https://doi.org/10.1111/j.1745-3984.2000.tb01090.x

Tian, F. (2011). *A comparison of equating/linking using the Stocking-Lord method and concurrent calibration with mixed-format tests in the non-equivalent groups common-item design under IRT* (Order No. 3475252). Available from ProQuest Dissertations & Theses Global. (900304732). https://www.proquest.com/dissertations-theses/comparison-equating-linking-using-stocking-lord/docview/900304732/se-2

Uçar, A., & Sünbül, Ö. (2024). Comparing equating errors on various factors for subtests which have added value. *Journal of Advanced Education Studies, 6*(1), 92-111. https://doi.org/10.48166/ejaes.1438652

Uyar, Ş., Aksekioğlu, B., & Öztürk-Gübeş, N. (2018). PISA 2012 matematik okuryazarlığı testinde farklı ölçek dönüştürme yöntemlerinin karşılaştırılması. *Mehmet Akif Ersoy Üniversitesi Eğitim Fakültesi Dergisi*, 46, 121-148. https://doi.org/10.21764/maeuefd.330613

Uysal, İ. (2014). *Madde tepki kuramına dayalı test eşitleme yöntemlerinin karma modeller üzerinde karşılaştırılması*. Yüksek Lisans Tezi, Abant İzzet Baysal Üniversitesi Eğitim Bilimleri Enstitüsü, Bolu.

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. https://iojes.net/?mod=makale_tr_ozet&makale_id=40844

Uysal, İ., Şahin-Kürşad, M., & Kılıç, A. F. (2022). Effect of item parameter drift in mixed format common items on test equating. *Participatory Educational Research, 9*(5), 143-160. https://doi.org/10.17275/per.22.108.9.5

Walker, M. E., & Kim, S. (2009, April). Linking mixed-format tests using multiple choice anchors. Paper presented at the *Annual Meeting of the American Educational Research Association (AERA) and the National Council on Measurement in Education (NCME)*, San Diego.

Wang, S., & Kolen, M. J. (2016). Evaluation of scale transformation methods with stabilized conditional standard errors of measurement for mixed-format tests. In M. J. Kolen & W. Lee (Eds.) Mixed-format tests: Psychometric properties with a primary focus on equating (Volume 4) (CASMA Monograph Number 2.4, pp. 205–222). Iowa City: CASMA, The University of Iowa. https://education.uiowa.edu/sites/education.uiowa.edu/files/2021-11/casma-monograph-2.4.pdf#page=217

Wolkowitz, A. A., & Wright, K. D. (2019). Effectiveness of equating at the passing score for exams with small sample sizes. *Journal of Educational Measurement, 56*(2), 361–390. https://doi.org/10.1111/jedm.12212

Yang, W., & Houang, R. T. (1996, April). The effect of anchor length and equating method on the accuracy of test equating: Comparisons of linear and IRT-based equating using anchor-item design. Paper presented at the *AERA Annual Conference*, New York.

Yurtçu, M., & Güzeller, C. O. (2022). Comparison of item response theory scaling methods with ROC analysis. *Journal of Measurement and Evaluation in Education and Psychology, 13*(1), 15-22. https://doi.org/10.21031/epod.892079

Zhang, Z. (2021). Asymptotic standard errors of generalized partial credit model true score equating using characteristic curve methods. *Applied Psychological Measurement, 45*(5), 331-345. https://doi.org/10.1177/01466216211013101

Zor, Y. M. (2023). Investigation of multidimensional scale transformation methods applied to multidimensional tests according to various conditions. *Adıyaman University Journal of Educational Sciences, 13*(1), 41-53. https://doi.org/10.17984/adyuebd.1239198

Extended Abstract

Introduction

Test scores are frequently used to provide information when making important decisions at the individual and institutional levels (Kolen & Brennan, 2014). Test scores are also used in large-scale exams, school-level exams, institution entrance and certificate exams. Test scores obtained from these exams at different times should be comparable in some cases. Because two individuals with the same ability level may take different exams and receive different scores. In such cases, there is a need to equate test forms to eliminate problems that may arise when one participant takes a more difficult or easier test form than another participant and to obtain comparable scores on many forms of the same test.

In a test equating study, the use of calibration methods independent of common item ratio and item discrimination levels and the realization of the advantages of test equating with IRT are potential threats. For this reason, it is important to evaluate the common item ratio and item discrimination levels together with the calibration method before equating the test. The aim of the research is to examine the effects of common item ratio, item discrimination and calibration method on equating error in test equating based on item response theory.

Method

Research Model

Since this research aims to compare the equating error under different conditions, it is a basic research conducted with simulation data (Kothari, 2004).

Data Simulation

We generated the simulated data used in the research in the R. There are six conditions in the study: two item discrimination levels (medium (a_{mean}=0.00) and high (a_{mean}=0.50) × three common item ratios (10%, 20% and 30%). 100 replications were performed for each condition. Data sets consist of 1500 individuals and the test produced consists of 40 items and, common items are internal. Accordingly, the item distribution was determined as 36 items + 4 common items for a 10% common item ratio, 32 items + 8 common items for a 20% common item ratio, and 28 items + 12 common items for a 30% common item ratio. In generating the response patterns, a three-parameter logistic model (3PLM) was used to take into account the chance factor since the structure of the tests is multiple choice (1-0). In

addition, another condition in the current research, calibrating methods Mean-mean (MM), Mean-Sigma (MS), Haebara (HA) and Stocking & Lord (SL), was also discussed. Finally, a Form X and a Form Y were generated for each condition. In this study, 2400 test equations were conducted, including 6 conditions × 4 calibrating methods × 100 replications.

Data Analysis

The equateIRT (Battauz, 2015) package was used to analyze the data in R (v.4.4.1). After we performed the calibrating process with MM, MS, HA and SL methods, the test equation was conducted according to the common item non-equivalent groups design based on IRT observed score equating for each replication and each condition.

After equating for each condition and each replication, we examined the standard error values of the equating. Mean error values were obtained by dividing the standard error of the equation by the number of replications, and these values were compared and interpreted.

Results and Conclusion

In the research, it was observed that HA was the calibration method least affected by the change in the level of common item ratio under the condition of medium discrimination level. In addition, it was determined that the MS method had the lowest equating errors for all common item ratios under the medium discrimination condition. It can be said that the high discrimination level gives the lowest equating errors in all common item ratio conditions of the SL method. In general, it is seen that the standard error of equating decreases as the ratio (number) of common items increases under all calibration methods and both of discrimination conditions. This finding showed that the equating error decreases when the number of common items increases in tests containing items at the both of discrimination levels.

When the common item ratio increases in the SL method, the equating errors decrease at each discrimination level. But if the discrimination level is high, there does not appear to be an important differentiation in terms of the standard error of the equating. When the HA method is examined, it is seen that results similar to the results of the SL method are produced. In the MS method, it is seen that changing the discrimination level for all common item ratio conditions does not affect the equating error. It was also observed that the MM method had the highest equating error under all calibrating methods and common item ratio conditions.

In line with these findings, higher discrimination of items in test pairs in all calibrating methods and all common item ratios reduces the equating error. Therefore, when different common item ratios are used, it is important that item discrimination is high, regardless of the calibrating method. Finally, it can be said that under all conditions, the lowest equating errors are higher level of discrimination.