Manisa Celal Bayar Üniversitesi Eğitim Fakültesi Dergisi
ISSN: 1309-8918
Yıl: 2025, Cilt: 13, Sayı: 1
https://dergipark.org.tr/en/pub/mcbuefd

Çok Boyutlu Madde Tepki Kuramı Test Eşitleme Yöntemlerinin Karşılaştırılması

Comparison of Multidimensional Item Response Theory Test Equating Methods

Burcu Demiröz*[1], Nuri Doğan[2]

Öz

Bu araştırmada, çok boyutlu testlerden elde edilen puanların eşitlenmesinde kullanılan çift-faktör çok boyutlu madde tepki kuramı gözlenen puan eşitleme, tam çok boyutlu madde tepki kuramı gözlenen puan ve çok boyutlu madde tepki kuramı gözlenen puan eşitleme tek boyutlu yaklaşım yöntemlerinden elde edilen sonuçların karşılaştırılması amaçlanmıştır. Karşılaştırma yaparken ortak madde deseni altında çeşitli faktörlere göre elde edilen eşitlenmiş puanlar, bu puanlara ait eşitlemenin standart hatası, yanlılık ve hata kareler ortalamasının karekökü değerleri incelenmiştir. Simülasyon verileri kullanılmıştır. Örneklem büyüklüğü, ortak madde oranı, boyutlar arasındaki ilişki düzeyi, çok boyutlu test eşitleme yöntemleri 3 ve kalibrasyon yöntemleri 2 farklı koşul içermektedir. Bu değişkenlerin farklı seviyelerinin kombinasyonu sonucunda 162 koşul oluşturulmuştur. Veri setlerinin üretilmesi ve eşitleme çalışmaları R programlama dili kullanılarak gerçekleştirilmiştir. Eş zamanlı ve ayrı kalibrasyon yöntemleri için örneklem büyüklüğü arttıkça hata değerlerinin azaldığı gözlenmiştir. Örneklem büyüklüğünün en az 3000 olması önerilmektedir. Eş zamanlı kalibrasyon yöntemi kullanıldığında ortak madde oranı %20; ayrı kalibrasyon yöntemi kullanıldığında ortak madde oranı %50 olduğunda en az hata değerleri gözlenmiştir. Ortak madde oranı eş zamanlı kalibrasyonda en çok %20; ayrı kalibrasyonda en az %50 olmalıdır. Eş zamanlı kalibrasyonda ayrı kalibrasyon yönteminden daha küçük hata değerleri gözlenmiştir.

Anahtar Kelimeler: test eşitleme, madde tepki kuramı, çift-faktör model, ölçek kalibrasyonu, hata

Abstract

In this study, it was aimed to compare the results obtained from different methods used in equating scores derived from multidimensional tests, including bifactor MIRT observed score equating, full MIRT observed score equating, and the unidimensional approximation of MIRT observed score equating. While making the comparison, equated scores obtained under the common-item design were examined along with the standard error of equating, bias, and the root mean square error. Simulation data were used. Sample size, common item rate, level of relationship between dimensions, multidimensional test equating methods include 3 and calibration methods include 2 different conditions. As a result of the combination of different levels of these variables, 162 conditions were created. Data generation and equating procedures were carried out using the R programming language. It was observed that the error values decreased as the sample size increased for the concurrent and separate calibration methods. A minimum sample size of 3000 is recommended. The lowest error values were observed when the common-item proportion was 20% for concurrent calibration and 50% for separate calibration. The common-item proportion should be at most 20% for concurrent calibration and at least 50% for separate calibration. Concurrent calibration yielded lower error values than separate calibration.

Keywords: test equating, item response theory, bi-factor model, scale linking, error

Başvuru tarihi: 20/11/2024 | **Kabul tarihi:** 28/05/2025 **Araştırma Makalesi / Research Article**

Önerilen atıf: Demiröz, B., & Doğan, N. (2025). Çok boyutlu madde tepki kuramı test eşitleme yöntemlerinin karşılaştırılması. *Manisa Celal Bayar Üniversitesi Eğitim Fakültesi Dergisi*, *13*(1), 156-176. https:/doi.org/10.52826/mcbuefd.1587864

***Sorumlu yazar:** bd14.olcme@gmail.com

Not: Bu makale birinci yazarın ikinci yazar danışmanlığında yürütmüş olduğu tez çalışmasından üretilmiştir.

*[1]Sorumlu yazar: MEB, Gaziantep, Türkiye
[2] Eğitimde Ölçme ve Değerlendirme, Hacettepe Üniversitesi, Ankara, Türkiye

Çok Boyutlu Madde Tepki Kuramı Test Eşitleme Yöntemlerinin Karşılaştırılması

Giriş

Pek çok test doğası gereği çok boyutludur. Bir testte birden fazla boyut ölçülüyorsa, karmaşık çok boyutlu yapının tanımlanması şartıyla, çok boyutlu MTK (ÇB-MTK) çerçevesi, yetenekleri ve maddeleri tek boyutlu bir MTK (TB-MTK) çerçevesinden daha doğru bir şekilde modelleyebilir. ÇB-MTK'nın geliştirilmesiyle birlikte çok boyutlu testler için kalibrasyon ve eşitleme problemine çözümler üretilmeye başlanmıştır. ÇB-MTK çerçevesi içerisinde çok boyutlu test eşitlemeye yönelik ilk çalışmalardan biri, rastgele grup deseni altında çok boyutlu iki parametreli lojistik modelle bağlantılı olarak gözlenen ve gerçek puan eşitleme prosedürlerini öneren Brossman (2010) tarafından yapılan çalışmadır. Brossman (2010) karmaşık bir ÇB-MTK modeli kullanmıştır. Brossman (2010) Tam ÇB-MTK gözlenen puan eşitleme (MOSE), ÇB-MTK gözlenen puan eşitleme tek boyutlu yaklaşım (AOSE), ÇB-MTK gerçek puan eşitleme tek boyutlu yaklaşım (ATSE) yöntemlerini geliştirmiştir. Diğer yandan basit yapılı ÇB-MTK gözlenen puan (SMO) (Lee ve Brossman, 2012), basit yapılı ÇB-MTK gerçek puan eşitleme (SMT) (Kim vd., 2019), çift-faktör ÇB-MTK gözlenen puan (Lee ve Lee, 2016), çift-faktör ÇB-MTK gerçek puan (Lee vd., 2015), madde takımı yanıt modeli ÇB-MTK gözlenen puan ve madde takımı yanıt modeli ÇB-MTK gerçek puan (Tao ve Cao, 2016) yöntemleri de ilerleyen süreçlerde geliştirilmiştir.

Tam ÇB-MTK gözlenen puan, MTK gözlenen puan eşitleme tek boyutlu yaklaşım eşitleme yöntemleri Brossman (2010) ve çift-faktör ÇB-MTK gözlenen puan eşitleme yöntemi Lee ve Lee (2016) tarafından geliştirilmiş olmasına rağmen bu yöntemlerin çeşitli koşullar altında nasıl performans gösterdiğini belirlemek için yeterince çalışma yapılmamıştır. Alan yazın incelendiğinde denk olmayan gruplarda ortak madde deseni altında gerçekleştirilen çok az çalışmaya ulaşılabilinmiştir. Yapılan çalışmaların genellikle rastgele grup deseni altında test eşitleme uygulamaları olduğu belirlenmiştir. Bunun yansıra kalibrasyon yöntemlerinin kullanılan test eşitleme yöntemlerine etkisinin de incelenmesi gerektiği belirlenmiştir. Bu nedenle, farklı test eşitleme desenleri altında gözlenen puan eşitleme yöntemlerinin performansının daha fazla araştırılması ve karşılaştırılması gereklidir. Bu durum göz önünde bulundurularak bu araştırmada, çok boyutlu testlerden elde edilen puanların eşitlenmesinde kullanılan çift-faktör ÇB-MTK gözlenen puan eşitleme (BF-MIRT), tam ÇB-MTK gözlenen puan (MOSE) ve ÇB-MTK gözlenen puan eşitleme tek boyutlu yaklaşım (AOSE) yöntemlerinden denk olmayan gruplarda ortak madde deseni altında elde edilen eşitlenmiş puanlar ve bu puanlara ait eşitlemenin standart hatası (SEE), yanlılık (BIAS) ve hata kareler ortalamasının karekökü (RMSE) değerlerinin çeşitli faktörlere göre karşılaştırılması amaçlanmıştır. Bu çerçevede aşağıdaki sorulara cevap aranmıştır.

Araştırma Sorusu 1: Denk olmayan gruplarda ortak madde deseni altında tam ÇB-MTK gözlenen puan, ÇB-MTK gözlenen puan eşitleme tek boyutlu yaklaşım ve çift-faktör ÇB-MTK gözlenen puan eşitleme yöntemleri kullanılarak elde edilen eşitlenmiş puanlar farklılaşmaktadır mıdır?

Araştırma Sorusu 2: Denk olmayan gruplarda ortak madde deseni altında tam ÇB-MTK gözlenen puan eşitleme, ÇB-MTK gözlenen puan eşitleme tek boyutlu yaklaşım ve çift-faktör ÇB-MTK gözlenen puan eşitleme yöntemleri kullanılarak elde edilen eşitlenmiş puanlar için yanlılık (BIAS), eşitlemenin standart hatası (SEE) ve hata kareler ortalamasının karekökü (RMSE) değerleri,

- Örneklem büyüklüğüne (500, 1500 ve 3000) göre nasıl değişmektedir?
- Ortak madde oranına (%20, %35 ve %50) göre nasıl değişmektedir?
- Çok boyutlu eş zamanlı kalibrasyon ve ayrı kalibrasyon (Stocking Lord) yöntemlerine göre nasıl değişmektedir?
- Boyutlar arasındaki ilişki düzeyine (r=0, r=0,30 ve r=0,60) göre nasıl değişmektedir?

Yöntem

Araştırma türü

Bu çalışmada, denk olmayan gruplarda ortak madde deseni altında üretilen simülasyon veri setlerinden yararlanılmıştır. Bu nedenle araştırma, Monte Carlo simülasyon araştırması niteliği taşımaktadır. Monte Carlo simülasyon araştırmaları, bilinen olasılık dağılımlarından sözde rastgele örnekleme yoluyla veri oluşturmayı içeren bilgisayar ortamında gerçekleştirilen deneysel desenlerdir.

Test eşitleme deseni

Araştırmada denk olmayan gruplarda ortak madde deseni kullanılmıştır. Bu desende ortak maddelerdeki performanslardan faydalanarak test formları için ortak bir ölçek oluşturulur.

Simülasyon koşulları

Çalışmada, örneklem büyüklüğü, ortak madde oranı, boyutlar arasındaki ilişki düzeyi, çok boyutlu test eşitleme yöntemleri 3 ve kalibrasyon yöntemleri 2 farklı koşul içermektedir. Bu değişkenlerin farklı seviyelerinin kombinasyonu sonucunda 162 (3x3x3x2x3) koşul oluşturulmuştur. Yapılan birçok araştırmada 100 replikasyon yapıldığı görülmüştür (Cao, 2008; Kim, 2018; Kumlu, 2019; Lee vd., 2012). Bu nedenle araştırmada her simülasyon koşulu 100 kere tekrarlanmıştır.

Boyut sayısı

Alan yazındaki çalışmalar incelendiğinde kullanılan veri setlerinin boyut sayılarının 2 ila 4 arasında değiştiği belirlenmiştir (Brossman ve Lee, 2013; Kim, 2018; Kumlu, 2019; Peterson, 2014;Uğurlu, 2020; Zhang, 2012). Yapılan incelemelerin sonuçları ve çift-faktör model yapısı göz önünde bulundurularak verilerin 3 boyutlu olmasına karar verilmiştir.

Örneklem büyüklüğü

Tek boyutlu ölçek kalibrasyonunda ve test eşitlemede örneklem büyüklüğü arttıkça sonuçların doğruluğunun ve kesinliğinin arttığı bilinmektedir (Çokluk vd., 2022; Gök ve Kelecioğlu, 2014; Lee vd., 2014; Kilmen ve Demirtaşlı, 2012; Kolen ve Brennan, 2014; Wang, 2006). Kolen ve Brennan'a (2004) göre MTK gerçek puan eşitleme yöntemlerinde denk olmayan gruplarda ortak madde deseni altında 3 PLM

kullanıldığında her form için örneklem büyüklüğünün 1500 olması gerekmektedir. Wang ve Liu (2018) 3000 ve 8000 örneklem büyüklükleri arasındaki değişimin daha az olduğunu belirlemiştir. Araştırma sonuçları göz önünde bulundurularak örneklem büyüklükleri küçük örneklemleri temsil etmesi amacıyla 500, orta örneklemleri temsil etmesi amacıyla 1500 ve büyük örneklemleri temsil etmesi amacıyla 3000 olarak belirlenmiştir.

Madde sayısı ve ortak madde oranı

Kolen ve Brennan (2004) tarafından önerilen bir kural, birden fazla içerik alanını ölçen testlerle eşitleme yapılırken testlerin en az 30-40 maddeden oluşması gerektiğidir. Öneriler göz önünde bulundurularak her boyutun 30, tüm testin ise 60 maddeden oluşmasına karar verilmiştir.

Angoff (1971) madde sayısı 40 ve altında olduğunda ortak madde oranının %20; madde sayısı 40'ın üstünde olduğunda ise ortak madde oranının %30 olması gerektiğini belirtmiştir. Çok boyutlu testler için ortak madde oranı %50 kullanılmaktadır (Peterson, 2014). Bunun sonucunda ortak madde oranı %20, %35 ve %50 olarak belirlenmiştir.

Boyutlar arasındaki ilişki düzeyi

Boyutlar arasındaki yüksek korelasyon boyutluluğun etkisini azaltmakta ve testler tek boyutluluğa yakınsamaktadır. Çok boyutlu test eşitleme yöntemleri karşılaştırılacağından verilerin tek boyutluluğa yakınsamasının önüne geçilmesi amacıyla boyutlar arasındaki ilişki düzeyi r =0, 0,3 ve 0,6 olarak belirlenmiştir.

Kalibrasyon yöntemleri

Denk olmayan gruplarda ortak madde deseni altında test karakteristik fonksiyonuna (TKF) dayalı yöntem ile yapılan ölçek kalibrasyonlarının iyi sonuçlar gösterdiği ortaya konulmuştur (Atar ve Yeşiltaş, 2017; Baker ve Al Karni, 1991; Yao ve Boughton, 2009). Bu nedenle araştırmada, eş zamanlı kalibrasyon yönteminin yanı sıra TKF'ye dayanan Stocking Lord ayrı kalibrasyon yönteminin kullanılması tercih edilmiştir.

Verilerin üretilmesi

Araştırmada uygulama ve anlaşılma kolaylığı nedeniyle ikili puanlanan madde cevap verisi üretilmiştir. Her bir koşul altında her grup için cevaplar telafi edici çok boyutlu üç parametreli lojistik (M-3PLM) MTK kullanılarak üretilmiştir. Veri setlerini oluşturmak için, bir genel faktör (g) ve iki spesifik faktörden (f1, f2) oluşan çift-faktör model kullanılmıştır. İlk otuz madde genel ve birinci özel faktöre; son otuz madde genel ve ikinci özel faktöre yüklenmiştir.

Madde ve yetenek parametrelerinin üretilmesi

Madde ayırıcılık parametresi, X ve Y testleri için minimum 0,3 maksimum 2 olacak şekilde sürekli tek biçimli (uniform); madde güçlük parametresi, her madde için bir tane olmak üzere X ve Y testleri için -3 ile +3 arasında ortalaması 0, standart sapması 1 olan normal; şans parametresi c, minimum 0,05 maksimum 0,20 olacak şekilde sürekli tek biçimli (uniform) dağılımdan üretilmiştir. Yetenek dağılımları her boyut için, Y formunu alan grubun yetenek dağılımı ortalaması 0 standart sapması 1 (N(0,1)); X formunu alan grubun yetenek dağılımı ortalaması 0,05 standart sapması 1 (N(0,05,1)) olan çok değişkenli normal dağılımdan üretilmiştir.

Verilerin analizi

Veri analizi R programlama dili (R Development Core Team, 2022) kullanılarak gerçekleştirilmiştir. Madde parametrelerini üretmek için 'stats' (R Development Core Team, 2022) ve yetenek parametreleri üretmek için 'mvtnorm' (Genz vd., 2021) paketi kullanılmıştır.

Birinci alt probleme yanıt aranırken sonuçların örneklem büyüklüğü, ortak madde oranı ve boyutlar arasındaki ilişki düzeyinden etkilenmemesi için örneklem büyüklüğü 3000, ortak madde oranı %20, boyutlar arasındaki ilişki düzeyi r = 0 olan veri seti kullanılmıştır. Eşitlenmiş puanların test eşitleme yöntemlerine göre farklılaşıp farklılaşmadığı tekrarlı ölçümler ANOVA (Varyans Analizi) testi ile incelenmiştir. Tekrarlı ölçümler ANOVA testi sonuçlarına göre eş zamanlı kalibrasyon yönteminden elde edilen eşitlenmiş puanlar arasında fark bulunmamıştır. Tekrarlı ölçümler ANOVA testi grupların ortalamalarını karşılaştırmıştır. Yöntemlerden elde edilen puanların ortalamaları benzer olduğu için yöntemler arasında fark bulunmamıştır. Gruplar oluşturulurken puanların normal dağıldığı varsayılmıştır. Bu nedenle puanlar alt, orta ve üst olmak üzere üç gruba ayrılmış ve grup bağımsız değişkeni oluşturularak uç değerlerde gözlenen farklılaşmaların istatistiksel olarak anlamlı olup olmadığı incelenmiştir. Normal dağılımda ortalama puandan -1, +1 standart sapma uzaklıkta bulunan puanlar orta grubu oluşturmuştur. Puanlar, ortalamanın -1 standart sapma altında kaldığında alt gruba, puanlar +1 standart sapma üstünde kaldığında ise üst gruba dahil edilmiştir. Gruplara ve yönteme göre eşitlenmiş puanların farklılaşıp farklılaşmadığı Tekrarlı ölçümler iki yönlü ANOVA testi ile incelenmiştir. Etki büyüklüğü, Cohen eta kare (η^2) değeri ile değerlendirilmiştir. Eşitlenmiş puanların hangi grup ve yöntem için farklılaştığı Tukey testi ile incelenmiştir (Ek-A).

Değerlendirme ölçütleri

Araştırmada eşitleme sonuçlarının doğruluğunu, her bir test eşitleme yönteminin performansını incelemek amacıyla eşitlenmiş puanlar için yanlılık (BIAS), eşitlemenin standart hatası (SEE) ve hata kareler ortalamasının karekökü (RMSE) değerleri hesaplanmıştır.

Bulgular

Bu bölümde veri analizleri sonucunda elde edilen bulgular ele alınan araştırma soruları dikkate alınarak

sunulmuştur.

Birinci araştırma sorusuna ilişkin bulgular

Tekrarlı ölçümler ANOVA testi sonuçları Tablo 1'de yer almaktadır.

Tablo 1. Tekrarlı ölçümler ANOVA sonuçları

		Sd	Kareler toplamı	Ortalama kareler	F
Eş zamanlı kalibrasyon	**Yöntem**	2	2,05	1,025	1,012
Ayrı kalibrasyon	**Yöntem**	2	179,1	89.57	12,58*

Not: * p<0,05

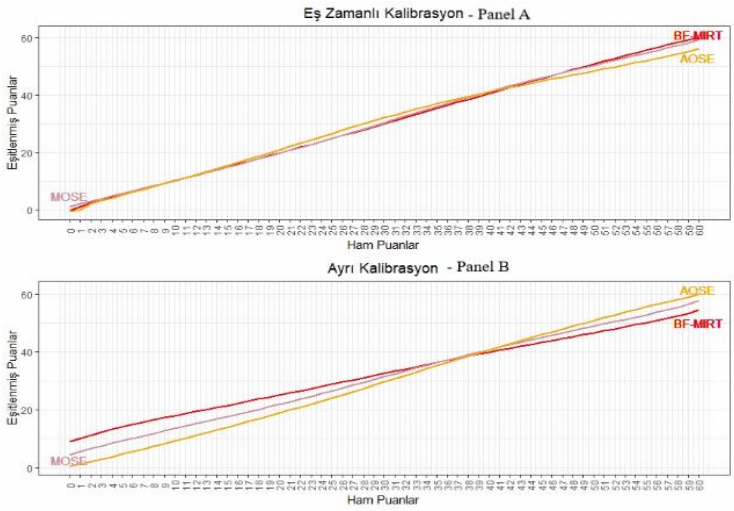
Tablo 1 incelendiğinde eş zamanlı kalibrasyon sonucunda elde edilen madde ve yetenek parametreleri kullanılarak gerçekleştirilen tam ÇB-MTK gözlenen puan (MOSE), ÇB-MTK gözlenen puan eşitleme tek boyutlu yaklaşım (AOSE) ve çift-faktör ÇB-MTK gözlenen puan eşitleme (BF-MIRT) uygulamalarından elde edilen eşitlenmiş puanlar arasında istatistiksel olarak anlamlı fark bulunmazken ayrı kalibrasyon (Stocking Lord) yöntemi kullanılarak gerçekleştirilen test eşitleme uygulamalarında ise istatistiksel olarak anlamlı fark bulunmuştur ($p < 0,05$). Tekrarlı ölçümler iki yönlü ANOVA testi sonuçları Tablo 2'de yer almaktadır.

Tablo 2. Tekrarlı ölçümler iki yönlü ANOVA sonuçları

		Sd	Kareler toplamı	Ortalama kareler	F
	Yöntem	2	2,05	1,025	2,843
Eş Zamanlı Kalibrasyon	**Grup**	2	37012	18506	59,19*
	Yöntem*Grup	4	79,78	19,944	55,237*
-	**Yöntem**	2	179,1	89,57	31,45*
Ayrı Kalibrasyon	**Grup**	2	29658	14829	60*
	Yöntem*Grup	4	524,1	131,02	46*

Not: * p<0,05

Tablo 2 incelendiğinde eş zamanlı ve ayrı kalibrasyon (Stocking Lord) yöntemleri kullanılarak gerçekleştirilen tam ÇB-MTK gözlenen puan (MOSE), ÇB-MTK gözlenen puan eşitleme tek boyutlu yaklaşım (AOSE) ve çift-faktör ÇB-MTK gözlenen puan eşitleme (BF-MIRT) uygulamalarında gruplardan elde edilen eşitlenmiş puanlar arasında istatistiksel olarak anlamlı fark bulunmaktadır ($p<0,05$). Puanların hangi grup ve yöntem için farklılaştığı ile ilgili Tukey testi sonuçları (Ek-A) incelendiğinde eş zamanlı kalibrasyon yöntemi için alt grupta istatistiksel olarak anlamlı fark yoktur ($p>0,05$). Fakat orta grupta AOSE yöntemi ile diğer yöntemler; üst grupta tüm yöntemler arasında istatistiksel olarak anlamlı fark vardır ($p<0,05$). Ayrı kalibrasyon (Stocking Lord) yönteminde orta grupta BF-MIRT ve MOSE yöntemleri arasında istatistiksel olarak anlamlı fark yoktur ($p>0,05$). Alt ve üst gruplarda ise tüm yöntemler arasında istatistiksel olarak anlamlı fark vardır ($p<0,05$). Puanlar Şekil 1'de çizgi grafiği ile incelenmiştir.



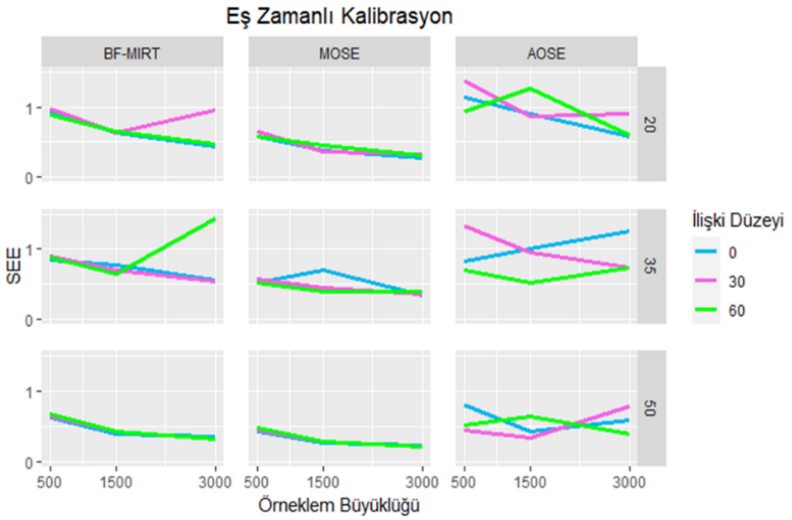
Şekil 1. Eşitlenmiş puanlar

Şekil 1 incelendiğinde genel olarak, ham puanlar ile eşitleme sonucu elde edilen puanlar arasındaki ilişkinin doğrusal olduğu görülmektedir. Eş zamanlı kalibrasyon yöntemi kullanılarak gerçekleştirilen test eşitleme uygulamaları sonucunda elde edilen eşitlenmiş puanların BF-MIRT ve MOSE eşitleme yöntemleri için benzer olduğu görülmüştür (Panel A). AOSE yönteminden elde edilen eşitlenmiş puanların ise x = 19 ile x = 37 ham puanları arasında diğer yöntemlerden elde edilen eşitlenmiş puanlardan farklılaştığı görülmektedir Benzer durum x = 46 ile x = 60 ham puanları arasında da görülmüştür. x = 46 ham puanından x = 60 ham puanına kadar farkların monoton arttığı belirlenmiştir. Ayrı kalibrasyon (Stocking Lord) yöntemi kullanılarak gerçekleştirilen test eşitleme uygulamaları sonucunda elde edilen eşitlenmiş puanların ise benzer olmadığı görülmüştür (Panel B).

İkinci araştırma sorusuna ilişkin bulgular[4]

İkinci alt problem doğrultusunda yöntemlerden elde edilen hata değerlerinin değişimi çizgi grafiği ile incelenmiştir. Eş zamanlı kalibrasyon sonucunda elde edilen SEE değerlerine ait grafikler Şekil 2'de gösterilmiştir.

[4] Anonim bir hakem tarafından bulguların, araştırmada incelenen koşullar arasındaki etkileşimler dikkate alınarak yeniden yorumlanması tavsiyesi gelmiş ancak çalışma tezden üretilirken sonuçların bir kısmı seçilerek alınması ve makaledeki biçimsel sınırlamalar nedeniyle analiz sonuçlarına ilişkin yorumlarda herhangi bir değişiklik yapılmamıştır.



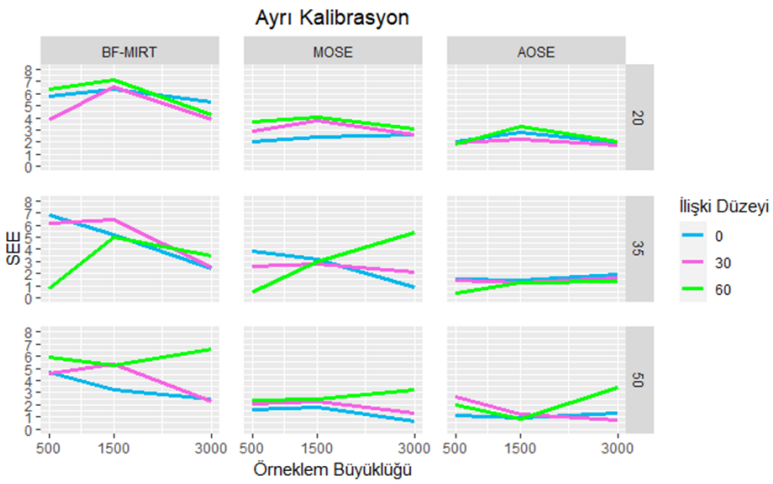
Şekil 2. Eş zamanlı kalibrasyon SEE değerleri

Şekil 2 incelendiğinde, eş zamanlı kalibrasyon sonucunda elde edilen madde ve yetenek parametreleri kullanılarak gerçekleştirilen test eşitleme uygulamalarının sonuçlarına göre, ortak madde oranı ve boyutlar arasındaki ilişki düzeyi sabit tutulduğunda, BF-MIRT ve MOSE yöntemleri için elde edilen SEE değerlerinin, örneklem büyüklüğü arttıkça azaldığı gözlenmiştir. AOSE yönteminde ise örneklem büyüklüğünün etkisi ortak madde oranına göre değişmektedir. Ortak madde oranı %20 olan veri setlerinde AOSE yönteminden elde edilen SEE değerlerinin örneklem büyüklüğü arttıkça azaldığı gözlenmiştir. Ortak madde oranı %35 ve %50 olan veri setlerinde örneklem büyüklüğü 500'den 1500'e çıkarıldığında SEE değerlerinin azaldığı; 1500'den 3000'e çıkarıldığında arttığı gözlenmiştir.

Örneklem büyüklüğü ve boyutlar arasındaki ilişki düzeyi sabit tutulduğunda, BF-MIRT ve AOSE yöntemleri için elde edilen SEE değerlerinin, ortak madde oranı arttıkça azaldığı gözlenmiştir. MOSE yönteminde ise elde edilen SEE değerlerinin, ortak madde oranı %20'den %35'e çıkarıldığında arttığı; %35'ten %50'ye çıkarıldığında azaldığı gözlenmiştir.

Örneklem büyüklüğü ve ortak madde oranı sabit tutulduğunda, BF-MIRT yöntemi için elde edilen SEE değerlerinin, boyutlar arasındaki ilişki düzeyi r=0'dan r=0,30'a çıkarıldığında arttığı; boyutlar arasındaki ilişki düzeyi r=0,30'dan r=0,60'a çıkarıldığında azaldığı gözlenmiştir. MOSE yöntemi için elde edilen SEE değerlerinin, boyutlar arasındaki ilişki düzeyi arttıkça arttığı gözlenmiştir. AOSE yönteminde ise elde edilen SEE değerlerinin, boyutlar arasındaki ilişki düzeyi arttıkça azaldığı gözlenmiştir.

Ayrı kalibrasyon sonucunda elde edilen SEE değerlerine ait grafikler Şekil 3'te gösterilmiştir.



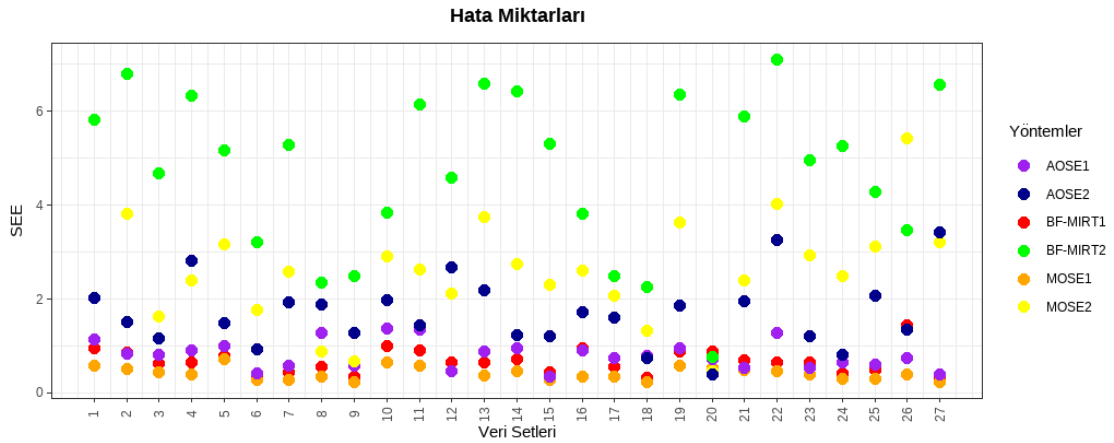
Şekil 3. Ayrı kalibrasyon SEE değerleri

Şekil 3 incelendiğinde, ayrı kalibrasyon sonucunda elde edilen madde ve yetenek parametreleri kullanılarak gerçekleştirilen test eşitleme uygulamalarının sonuçlarına göre, ortak madde oranı ve boyutlar arasındaki ilişki düzeyi sabit tutulduğunda, BF-MIRT ve MOSE yöntemleri için elde edilen SEE değerlerinin, örneklem büyüklüğü 500'den 1500'e çıkarıldığında arttığı; 1500'den 3000'e çıkarıldığında azaldığı gözlenmiştir. AOSE yönteminde ise elde edilen SEE değerlerinin, örneklem büyüklüğü 500'den 1500'e çıkarıldığında azaldığı; 1500'den 3000'e çıkarıldığında arttığı gözlenmiştir.

Örneklem büyüklüğü ve boyutlar arasındaki ilişki düzeyi sabit tutulduğunda, tüm yöntemler için elde edilen SEE değerlerinin, ortak madde oranı arttıkça azaldığı gözlenmiştir. Örneğin, örneklem büyüklüğü 3000 ve boyutlar arasındaki ilişki düzeyi r=0,30 olan veri setlerinde AOSE yöntemi kullanılarak elde edilen SEE değerleri, ortak madde oranı %20'den (SEE = 1,71) %35'e (SEE = 1,59) çıkarıldığında 0,12; ortak madde oranı %35'ten %50'ye (SEE = 0,73) çıkarıldığında 0,86 azalmıştır.

Örneklem büyüklüğü ve ortak madde oranı sabit tutulduğunda, BF-MIRT yöntemi için elde edilen SEE değerlerinin, boyutlar arasındaki ilişki düzeyi r=0'dan r=0,30'a çıkarıldığında azaldığı; boyutlar arasındaki ilişki düzeyi r=0,30'dan r=0,60'a çıkarıldığında arttığı gözlenmiştir. MOSE yöntemi için elde edilen SEE değerlerinin, boyutlar arasındaki ilişki düzeyi arttıkça arttığı gözlenmiştir. AOSE yönteminde ise elde edilen SEE değerlerinin, boyutlar arasındaki ilişki düzeyi arttıkça azaldığı gözlenmiştir.

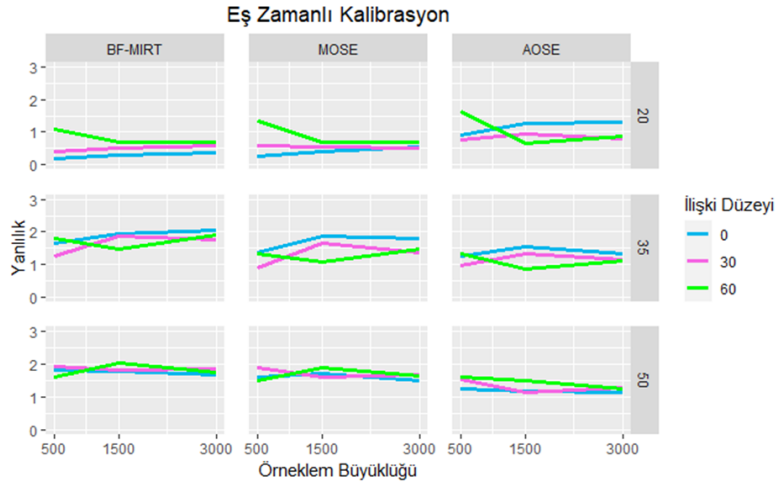
SEE değerlerinin eşitleme ve kalibrasyon yöntemlerine göre değişimi çizgi grafiği ile Şekil 4'te incelenmiştir. Bir ile belirtilen eşitleme yöntemleri (BF-MIRT1, MOSE1, AOSE1) eş zamanlı; iki ile belirtilen eşitleme yöntemleri (BF-MIRT2, MOSE2, AOSE2) ayrı (Stocking Lord) kalibrasyon sonucunda elde edilen SEE değerlerini her bir veri seti için göstermektedir.



Şekil 4. SEE ve yöntem etkileşimi

Şekil 4 incelendiğinde en yüksek SEE değerleri ayrı (Stocking Lord) kalibrasyon sonucunda elde edilen madde ve yetenek parametreleri kullanılarak gerçekleştirilen test eşitleme uygulamalarında gözlenmiştir. Eş zamanlı kalibrasyon ile gerçekleştirilen test eşitleme uygulamalarında en küçük SEE değerleri MOSE; en büyük SEE değerleri AOSE yönteminde gözlenmiştir. Ayrı kalibrasyon yönteminde ise en küçük SEE değerleri AOSE; en büyük SEE değerleri BFMIRT yönteminde gözlenmiştir.

Eş zamanlı kalibrasyon sonucunda elde edilen yanlılık değerlerine ait grafikler Şekil 5'te gösterilmiştir.



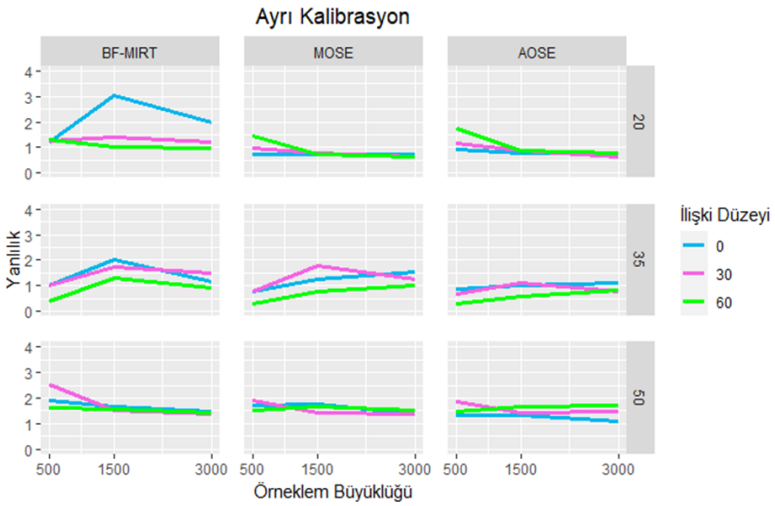
Şekil 5. Eş zamanlı kalibrasyon yanlılık değerleri

Şekil 5 incelendiğinde, eş zamanlı kalibrasyon sonucunda elde edilen madde ve yetenek parametreleri kullanılarak gerçekleştirilen test eşitleme uygulamalarının sonuçlarına göre, ortak madde oranı ve boyutlar arasındaki ilişki düzeyi sabit tutulduğunda BF-MIRT ve AOSE yöntemleri için elde edilen yanlılık değerlerine, örneklem büyüklüğünün etkisi ortak madde oranına göre değişmektedir. Ortak madde oranı %20 olan veri setlerinde BF-MIRT ve AOSE yöntemlerinden elde edilen yanlılık değerlerinin, örneklem büyüklüğü arttıkça arttığı gözlenmiştir. Ortak madde oranı %35 ve %50 olan veri setlerinde örneklem büyüklüğü arttıkça yanlılık değerlerinin azaldığı gözlenmiştir. MOSE yönteminde ise elde edilen yanlılık değerlerinin örneklem büyüklüğü 500'den 1500'e çıkarıldığında arttığı; 1500'den 3000'e çıkarıldığında azaldığı gözlenmiştir.

Örneklem büyüklüğü ve boyutlar arasındaki ilişki düzeyi sabit tutulduğunda, tüm yöntemler için elde edilen yanlılık değerlerinin, ortak madde oranı arttıkça arttığı gözlenmiştir. Örneğin, örneklem büyüklüğü 500 ve boyutlar arasındaki ilişki düzeyi r=0,30 olan veri setlerinde BF MIRT yöntemi kullanılarak elde edilen yanlılık değerleri, ortak madde oranı %20'den (Yanlılık = 0,41) %35'e (Yanlılık = 1,25) çıkarıldığında 0,83; ortak madde oranı %35'ten %50'ye (Yanlılık = 1,93) çıkarıldığında 0,68 artmıştır.

Örneklem büyüklüğü ve ortak madde oranı sabit tutulduğunda, BF-MIRT yöntemi için elde edilen yanlılık değerlerinin, boyutlar arasındaki ilişki düzeyi arttıkça arttığı gözlenmiştir. MOSE ve AOSE yöntemlerinde ise elde edilen yanlılık değerlerinin, boyutlar arasındaki ilişki düzeyi r=0'dan r=0,30'a çıkarıldığında azaldığı; r=0,30'dan r=0,60'a çıkarıldığında arttığı gözlenmiştir.

Ayrı kalibrasyon sonucunda elde edilen yanlılık değerlerine ait grafikler Şekil 6'da gösterilmiştir.



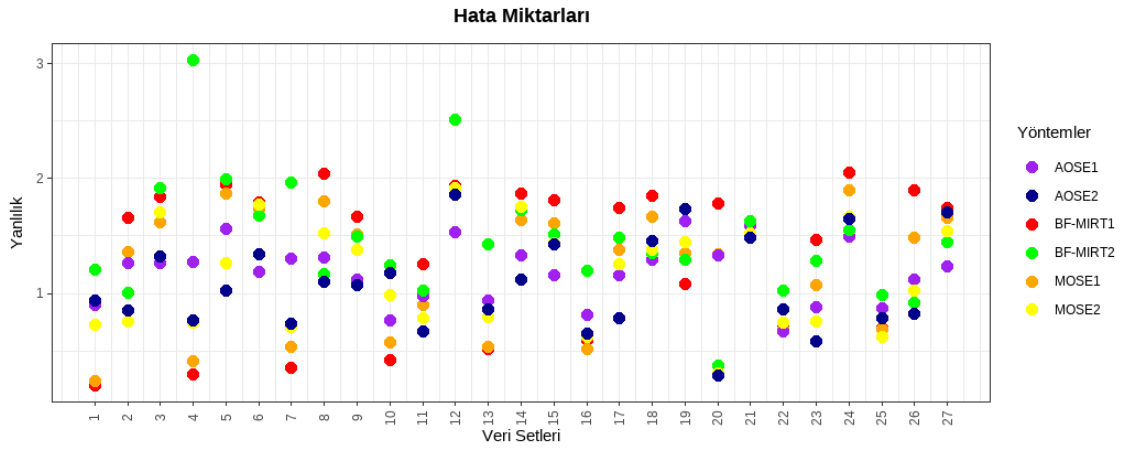
Şekil 6. Ayrı kalibrasyon yanlılık değerleri

Şekil 6 incelendiğinde, ayrı kalibrasyon sonucunda elde edilen madde ve yetenek parametreleri kullanılarak gerçekleştirilen test eşitleme uygulamalarının sonuçlarına göre, ortak madde oranı ve boyutlar arasındaki ilişki düzeyi sabit tutulduğunda, BF-MIRT ve MOSE yöntemleri için elde edilen yanlılık değerlerinin, örneklem büyüklüğü 500'den 1500'e çıkarıldığında arttığı; 1500'den 3000'e çıkarıldığında azaldığı gözlenmiştir. AOSE yönteminde ise örneklem büyüklüğünün etkisi ortak madde oranına göre değişmektedir. Ortak madde oranı %20 olan veri setlerinde AOSE yönteminden elde edilen yanlılık değerlerinin örneklem büyüklüğü arttıkça azaldığı gözlenmiştir. Ortak madde oranı %35 ve %50 olan veri setlerinde örneklem büyüklüğü arttıkça yanlılık değerlerinin arttığı gözlenmiştir.

Örneklem büyüklüğü ve boyutlar arasındaki ilişki düzeyi sabit tutulduğunda, BFMIRT yöntemi için elde edilen yanlılık değerlerinin, ortak madde oranı %20'den %35'e çıkarıldığında azaldığı; %35'ten %50'ye çıkarıldığında arttığı gözlenmiştir. MOSE ve AOSE yöntemlerinde ise elde edilen yanlılık değerlerinin, ortak madde oranı arttıkça arttığı gözlenmiştir.

Örneklem büyüklüğü ve ortak madde oranı sabit tutulduğunda, BF-MIRT ve MOSE yöntemleri için elde edilen yanlılık değerlerinin, boyutlar arasındaki ilişki düzeyi arttıkça azaldığı gözlenmiştir. AOSE yönteminde ise elde edilen yanlılık değerlerinin, boyutlar arasındaki ilişki düzeyi arttıkça arttığı gözlenmiştir.

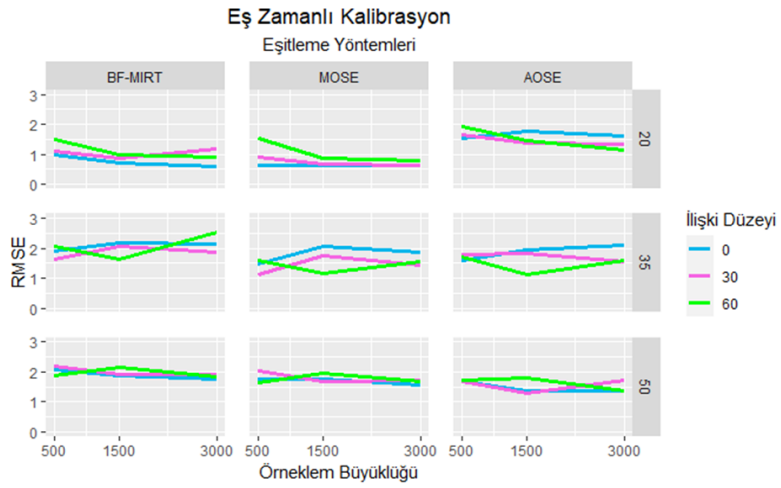
Yanlılık değerlerinin eşitleme ve kalibrasyon yöntemlerine göre değişimi çizgi grafiği ile Şekil 7'de incelenmiştir.



Şekil 7. Yanlılık ve yöntem etkileşimi

Şekil 7 incelendiğinde en yüksek yanlılık değerleri ayrı (Stocking Lord) kalibrasyon sonucunda elde edilen madde ve yetenek parametreleri kullanılarak gerçekleştirilen test eşitleme uygulamalarında gözlenmiştir. Eş zamanlı ve ayrı kalibrasyon ile gerçekleştirilen test eşitleme uygulamalarında en küçük yanlılık değerleri AOSE; en büyük yanlılık değerleri BF-MIRT yönteminde gözlenmiştir.

Eş zamanlı kalibrasyon sonucunda elde edilen RMSE değerlerine ait grafikler Şekil 8'de gösterilmiştir.



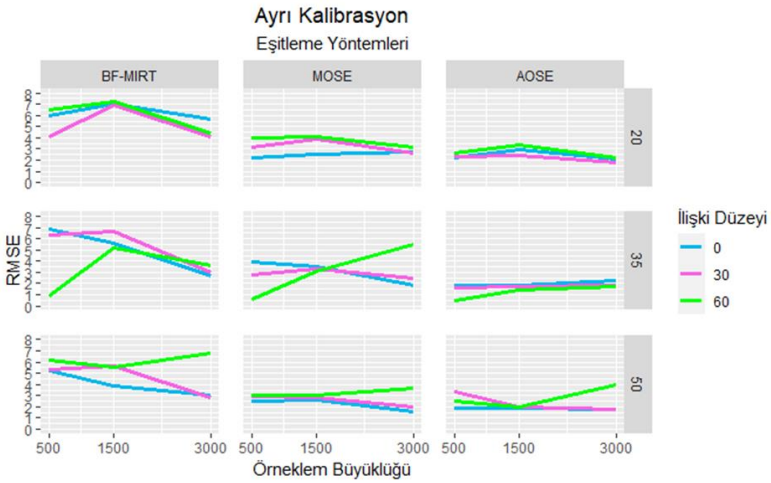
Şekil 8. Eş zamanlı kalibrasyon RMSE değerleri

Şekil 8 incelendiğinde, eş zamanlı kalibrasyon sonucunda elde edilen madde ve yetenek parametreleri kullanılarak gerçekleştirilen test eşitleme uygulamalarının sonuçlarına göre, ortak madde oranı ve boyutlar arasındaki ilişki düzeyi sabit tutulduğunda, tüm yöntemler için elde edilen RMSE değerlerinin, örneklem büyüklüğü arttıkça azaldığı gözlenmiştir. Örneğin, ortak madde oranı %20 ve boyutlar arasındaki ilişki düzeyi r=0,60 olan veri setlerinde AOSE yöntemi kullanılarak elde edilen RMSE değerleri, örneklem büyüklüğü 500'den (RMSE = 1,97) 1500'e (RMSE = 1,47) çıkarıldığında 0,46; örneklem büyüklüğü 1500'den 3000'e (RMSE = 1,15) çıkarıldığında 0,31 azalmıştır.

Örneklem büyüklüğü ve boyutlar arasındaki ilişki düzeyi sabit tutulduğunda, BFMIRT ve AOSE yöntemleri için elde edilen RMSE değerlerinin, ortak madde oranı %20'den %35'e çıkarıldığında arttığı; %35'ten %50'ye çıkarıldığında azaldığı gözlenmiştir. MOSE yönteminde ise elde edilen RMSE değerlerinin, ortak madde oranı arttıkça arttığı gözlenmiştir.

Örneklem büyüklüğü ve ortak madde oranı sabit tutulduğunda, BF-MIRT ve MOSE yöntemleri için elde edilen RMSE değerlerinin, boyutlar arasındaki ilişki düzeyi arttıkça arttığı gözlenmiştir. AOSE yönteminde ise elde edilen RMSE değerlerinin, boyutlar arasındaki ilişki düzeyi r=0'dan r=0,30'a çıkarıldığında azaldığı; r=0,30'dan r=0,60'a çıkarıldığında arttığı gözlenmiştir.

Ayrı kalibrasyon sonucunda elde edilen RMSE değerlerine ait grafikler Şekil 9'da gösterilmiştir.



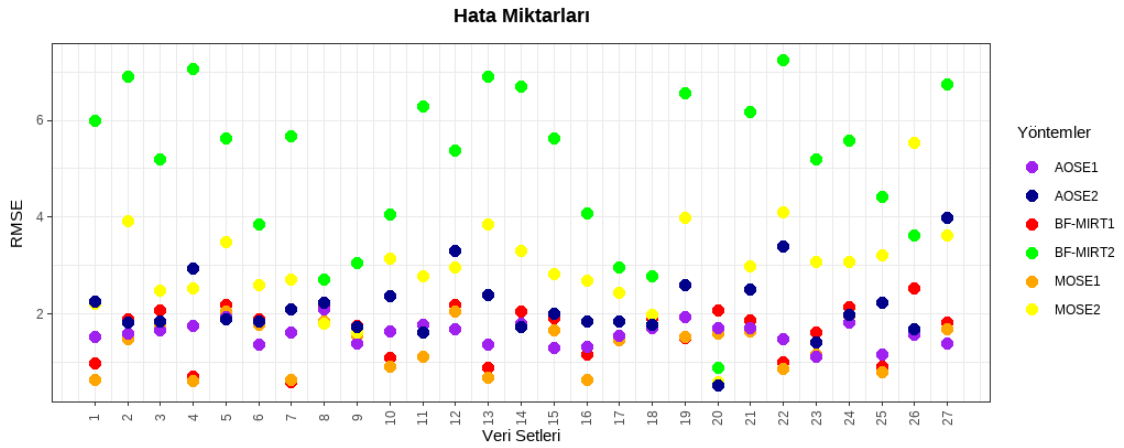
Şekil 9. Ayrı kalibrasyon RMSE değerleri

Şekil 9 incelendiğinde, ayrı kalibrasyon sonucunda elde edilen madde ve yetenek parametreleri kullanılarak gerçekleştirilen test eşitleme uygulamalarının sonuçlarına göre, ortak madde oranı ve boyutlar arasındaki ilişki düzeyi sabit tutulduğunda, tüm yöntemler için elde edilen RMSE değerlerinin, örneklem büyüklüğü 500'den 1500'e çıkarıldığında arttığı; 1500'den 3000'e çıkarıldığında azaldığı gözlenmiştir. Örneğin, ortak madde oranı %50 ve boyutlar arasındaki ilişki düzeyi r=0 olan veri setlerinde AOSE yöntemi kullanılarak elde edilen RMSE değerleri, örneklem büyüklüğü 500'den (RMSE = 1,83) 1500'e (RMSE = 1,84) çıkarıldığında 0,01 artarken; örneklem büyüklüğü 1500'den 3000'e (RMSE = 1,72) çıkarıldığında 0,12 azalmıştır.

Örneklem büyüklüğü ve boyutlar arasındaki ilişki düzeyi sabit tutulduğunda, BFMIRT ve MOSE yöntemleri için elde edilen RMSE değerlerinin, ortak madde oranı arttıkça azaldığı gözlenmiştir. AOSE yönteminde ise elde edilen RMSE değerlerinin, ortak madde oranı %20'den %35'e çıkarıldığında azaldığı; %35'ten %50'ye çıkarıldığında arttığı gözlenmiştir.

Örneklem büyüklüğü ve ortak madde oranı sabit tutulduğunda, BF-MIRT yöntemi için elde edilen RMSE değerlerinin, boyutlar arasındaki ilişki düzeyi r=0'dan r=0,30'a çıkarıldığında azaldığı; boyutlar arasındaki ilişki düzeyi r=0,30'dan r=0,60'a çıkarıldığında arttığı gözlenmiştir. MOSE yöntemi için elde edilen RMSE değerlerinin, boyutlar arasındaki ilişki düzeyi arttıkça arttığı gözlenmiştir. AOSE yönteminde ise elde edilen RMSE değerlerinin, boyutlar arasındaki ilişki düzeyi arttıkça azaldığı gözlenmiştir.

RMSE değerlerinin eşitleme ve kalibrasyon yöntemlerine göre değişimi çizgi grafiği ile Şekil 10'da incelenmiştir.



Şekil 10. RMSE ve yöntem etkileşimi

Şekil 10 incelendiğinde en yüksek RMSE değerleri ayrı (Stocking Lord) kalibrasyondan sonucunda elde edilen madde ve yetenek parametreleri kullanılarak gerçekleştirilen test eşitleme uygulamalarında gözlenmiştir. Eş zamanlı kalibrasyon ile gerçekleştirilen test eşitleme uygulamalarında en küçük RMSE değerleri MOSE; en büyük RMSE değerleri BF-MIRT yönteminde gözlenmiştir. Ayrı kalibrasyon yönteminde ise en küçük RMSE değerleri AOSE; en büyük RMSE değerleri BF-MIRT yönteminde gözlenmiştir.

Tartışma, Sonuç ve Öneriler

Bu araştırmada, çok boyutlu testlerden elde edilen puanların eşitlenmesinde kullanılan çift-faktör ÇB-MTK gözlenen puan eşitleme (BF-MIRT), tam ÇB-MTK gözlenen puan (MOSE) ve ÇB-MTK gözlenen puan eşitleme tek boyutlu yaklaşım (AOSE) yöntemlerinden denk olmayan gruplarda ortak madde deseni altında elde edilen eşitlenmiş puanlar ve bu puanlara ait eşitlemenin standart hatası (SEE), yanlılık (BIAS) ve hata kareler ortalamasının karekökü (RMSE) değerlerinin çeşitli faktörlere göre karşılaştırılması amaçlanmıştır.

Brossman (2010) MOSE, AOSE yöntemlerinden elde edilen eşitlenmiş puanların benzer olduğunu göstermiştir. Bu bulguyla tutarlı olarak alt grup incelendiğinde eş zamanlı kalibrasyon yöntemi kullanılarak gerçekleştirilen test eşitleme uygulamaları sonucunda elde edilen eşitlenmiş puanların benzer olduğu görülmüştür. Bu nedenle eş zamanlı kalibrasyon yöntemi kullanıldığında eşitleme uygulamalarını gerçekleştirmek için BF-MIRT, MOSE veya AOSE yöntemlerinden herhangi biri kullanılabilir. Üst grupta ise yöntemlerin farklılaştığı gözlenmiştir. Ayrı kalibrasyon (Stocking Lord) yöntemi kullanılarak gerçekleştirilen test eşitleme uygulamaları sonucunda elde edilen eşitlenmiş puanların ise benzer olmadığı görülmüştür. Choi (2019) ile benzer olarak ayrı kalibrasyon yöntemi kullanılarak gerçekleştirilen test eşitleme uygulamaları sonucunda elde edilen eşitlenmiş puanlar arasındaki fark eş zamanlı kalibrasyondan daha fazladır. Özellikle uç değerlere gidildikçe farkların arttığı görülmüştür. Bunun olası nedeni bu puanların frekanslarının düşük veya sıfır olmasıdır (Kim, 2018). Çünkü eşzamanlı kalibrasyon yöntemi, ortak maddeler için madde parametrelerini kestirirken iki grubun da yanıtlarını kullanır. Ayrı kalibrasyon yöntemi ise sadece bir grubun yanıtlarını kullanarak kestirim yapar (Kim ve Lee, 2018). Bu durumda yöntemler birbiri yerine kullanılamaz.

Lee (2013) çok boyutlu test eşitleme yöntemlerini (BF-MIRT, AOSE ve ATSE) incelediği çalışmasında örneklem büyüklüğü arttıkça ağırlıklandırılmış hata değerlerinin azaldığını gözlemlemiştir. Bu araştırmada da alan yazın ile benzer olarak eş zamanlı ve ayrı kalibrasyon yöntemi kullanılarak gerçekleştirilen test eşitleme uygulamaları sonucunda tüm yöntemler için örneklem büyüklüğü arttıkça hata değerlerinin azaldığı gözlenmiştir.

Kim ve Lee (2018), ortak madde oranı arttıkça hata değerlerinin azaldığını gözlemlemiştir. Meng (2007), Panidvadtana vd. (2021), eş zamanlı kalibrasyon yöntemi kullanılarak gerçekleştirilen test eşitleme uygulamaları sonucunda tüm yöntemler için %20 ortak madde oranının daha verimli olduğunu gözlemlemiştir. Araştırma sonuçlarına göre literatürle benzer olarak eş zamanlı kalibrasyon yönteminde ortak madde oranı %20 olduğunda en küçük hata değerleri gözlenmiştir. Ayrı kalibrasyon yönteminde ise ortak madde oranı arttıkça hata değerlerinin azaldığı ve %50 ortak madde oranının daha verimli olduğu gözlenmiştir. Ortak madde oranı arttıkça, ortak madde ile tüm test arasındaki ilişki daha güvenilir hale gelmektedir. Dolayısıyla daha az eşitleme hatası elde edilmesi beklenebilir (Hou, 2007).

Lee ve Lee (2016) boyutlar arasındaki ilişki düzeyi arttıkça (çok boyutluluk derecesi azaldıkça) BF-MIRT yönteminden elde edilen RMSD değerlerinin arttığını gözlemlemiştir. Bu bulgu ile benzer olarak BF-MIRT ve MOSE yöntemlerinden elde edilen hata değerleri boyutlar arasındaki ilişki düzeyi arttıkça yani çok boyutluluktan tek boyutluluğa gidildikçe artmaktadır. AOSE yönteminde ise boyutlar arasındaki ilişki düzeyi arttıkça yani çok boyutluluktan tek boyutluluğa gidildikçe hata değerlerinin azaldığı gözlenmiştir. AOSE yöntemi tek boyutlu gözlenen puan eşitleme yönteminin uzantısı olduğu için çok boyutluluk derecesi azaldıkça hata değerlerinin düşmesi beklenen bir sonuçtur.

Choi (2019), Kim (2017), Kim ve Lee (2018) eş zamanlı kalibrasyon yönteminin daha küçük hata değerleri ürettiğini belirtmiştirler. İki kalibrasyon yöntemi karşılaştırıldığında, bu bulgu ile tutarlı olarak eş zamanlı kalibrasyon yöntemi kullanılarak gerçekleştirilen test eşitleme uygulamalarında, ayrı kalibrasyon yöntemi kullanılarak gerçekleştirilen test eşitleme uygulamalarına göre daha küçük hata değerleri gözlenmiştir. Kalibrasyon yönteminden en çok etkilenen SEE; en az etkilenen yanlılık değerleridir. Zhang (2012) kalibrasyon yönteminin eşitleme sonuçlarına çok büyük etkisinin olduğunu belirtmiştir. Hata değerlerindeki artış incelendiğinde kalibrasyon yönteminden en çok etkilenen BF-MIRT; en az etkilenen AOSE yöntemidir.

Lee (2013) MOSE yönteminin diğer eşitleme yöntemlerinden daha iyi performans gösterdiğini belirlemiştir. Bununla benzer olarak eş zamanlı kalibrasyon yönteminde çalışmadaki tüm eşitleme yöntemleri arasında, MOSE yöntemi çoğu zaman hata değerleri karşılaştırıldığında en iyi performansı göstermiştir. Ayrı kalibrasyon yönteminde ise AOSE yöntemi en iyi performansı göstermiştir. Her iki kalibrasyon yöntemi içinde BF-MIRT yöntemi en kötü performansı göstermiştir. En büyük hata değerleri iki kalibrasyon yöntemi için de BF-MIRT yönteminde gözlenmiştir.

Elde edilen sonuçlar ve tartışmalara bağlı olarak test geliştiricilere ve uygulayıcılara aşağıdaki önerilerde bulunmak olası görünebilir:

(1) Eş zamanlı ve ayrı kalibrasyon yöntemi kullanılarak gerçekleştirilen test eşitleme uygulamaları için de örneklem büyüklüğü 3000 ve üzeri olmalıdır.

(2) Eş zamanlı kalibrasyon yöntemi kullanılarak gerçekleştirilen test eşitleme uygulamalarında ortak madde oranı %20 olduğunda en küçük yanlılık değerleri gözlenmiştir. Ortak madde oranını en çok %20 olmalıdır.

(3) Ayrı kalibrasyon yöntemi kullanılarak gerçekleştirilen test eşitleme uygulamalarında ortak madde oranı %50 olmalıdır.

(4) Eş zamanlı ve ayrı kalibrasyon yöntemi kullanılarak gerçekleştirilen test eşitleme uygulamalarında boyutlar arasındaki ilişki düzeyi BF-MIRT ve MOSE yöntemleri kullanıldığında r=0'a yakın olmalıdır. AOSE yöntemi kullanıldığında ise eş zamanlı kalibrasyonda r=0,30'a yakın; ayrı kalibrasyonda ise en az r=0,60' olmalıdır.

(5) İlişki düzeyi r=0'a yakın olduğunda eş zamanlı; ilişki düzeyi p=0,30 ve üzerinde olduğunda ayrı kalibrasyon yöntemi tercih edilebilir.

Çalışmada ele alınan koşullar ve çalışmanın sınırlılıkları göz önüne alındığında araştırmacılara yönelik olarak aşağıdaki önerilerde bulunmak olası görünebilir:

(1) Bu çalışmada BF-MIRT, MOSE veya AOSE yöntemleri kullanılmıştır. Alan yazın incelendiğinde tüm çok boyutlu test eşitleme yöntemlerini karşılaştıran herhangi bir çalışmaya rastlanmamıştır. Çok boyutlu test eşitleme yöntemlerinin tamamı farklı koşullar altında karşılaştırılabilir.

(2) BF-MIRT yöntemi uygulanırken Lord-Wingersky 2.0 algoritması kullanılmıştır. Çalışmada Lord-Wingersky 2.0 algoritmasının BF-MIRT yöntemi kullanılarak gerçekleştirilen test eşitleme uygulamalarında elde edilen hata değerlerinin yükselmesine neden olduğu düşünülmektedir. Lord-Wingersky 2.0 algoritması yeni geliştirilen bir yöntemdir. Algoritma veya algoritmanın farklı versiyonları kullanılarak BF-MIRT yöntemi uygulanabilir. Algoritmanın hata değerlerinin neden yüksek çıktığı incelenebilir

(3) Çalışmada veri setlerini oluşturmak için, bir genel faktör (g) ve iki spesifik faktörden (f1 ve f2) oluşan çift-faktör model kullanılmıştır. Bu konuyu yeninden araştırma sürecinde basit ve karmaşık yapı modeli ele alınabilir.

(4) Ayrı kalibrasyonun yönteminin eşitleme sonuçlarına etkisi incelenirken Stocking Lord yöntemi kullanılmıştır. Ortalama-ortalama, ortalama-sigma vb. yöntemler kullanılarak ayrı kalibrasyon yöntemleri karşılaştırılabilir.

(5) Çalışmada hata değerleri arasındaki farkların istatistiksel olarak anlamlılığı incelenmemiştir. Yapılacak çalışmalarda hata farklarının anlamlılığı incelenebilir.

Beyanlar

Etik beyan: Yukarıda bilgileri yer alan çalışmanın, etik kurul izni gerektirmeyen çalışmalar arasında yer aldığını beyan ederim/ederiz. Çalışmada bilimsel, etik ve alıntı kurallarına uyulduğu; toplanan veriler üzerinde herhangi bir tahrifatın yapılmadığı, karşılaşılacak tüm etik ihlallerde "Manisa Celal Bayar Üniversitesi Eğitim Fakültesi Dergisi ve Editörünün" hiçbir sorumluluğunun olmadığı, tüm sorumluluğun Sorumlu Yazara ait olduğu ve bu çalışmanın herhangi başka bir akademik yayın ortamına değerlendirme için gönderilmemiş olduğu sorumlu yazar tarafından taahhüt edilmiştir.

Çıkar çatışması: Yazarların çıkar çatışması yoktur.

Veri erişilebilirliği: Veriler yazarlardan talep üzerine sağlanabilir.

Kaynaklar

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

Atar, B., & Yeşiltaş, G. (2017). Çok boyutlu eşitleme yöntemlerinin eşdeğer olmayan gruplarda ortak madde deseni için performanslarının incelenmesi. *Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi, 8(*4), 421–434.

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.

Brossman, B. G. (2010). *Observed score and true score equating procedures for multidimensional item response theory* [Doctoral Dissertation]. University of Iowa.

Brossman, B. G., & Lee, W. (2013). Observed score and true score equating procedures for multidimensional item response theory. *Applied Psychological Measurement, 37*, 460-481.

Cao, L. (2008). *Mixed-format test equating: Effects of test dimensionality and common item sets* [Doctoral dissertation]. University of Maryland, College Park.

Choi, J. (2019). *Comparison of MIRT observed score equating methods under the commonitem nonequivalent groups design* [Doctoral Dissertation]. University of Iowa.

Çokluk, 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.

Genz, A., Bretz, F., Miwa, T., Mi, X., Leisch, F., Sheipl, F., & Hothorn, T. (2021). *mvtnorm: Multivariate Normal and t Distribution.* R package version 1.1-3, URL http://CRAN.Rproject.org/ package=mvtnorm.

Gök, B., & Kelecioğlu, H. (2014). Comparison of IRT equating methods using the commonitem nonequivalent groups design. *Mersin Üniversitesi Eğitim Fakültesi Dergisi, 10*(1), 120-136.

Hou, J. (2007). *Effectiveness of the hybrid Levine equipercentile and modified frequency estimation equating methods under the common-item nonequivalent groups design* [Doctoral Dissertation]. University of Iowa.

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.

Kim, K. (2017). *IRT linking methods for the bifactor model: a special case of the two-tier item factor analysis model* [Doctoral Dissertation]. The University of Iowa.

Kim, K. Y., & Lee, W. C. (2018). Linking methods for the full-information bifactor model under the common-item nonequivalent groups design. In M. J. Kolen & W. Lee (Eds.), *Mixed-format tests: Psychometric properties with a primary focus on equating* (Vol. 2.5, s. 243–261). Center for Advanced Studies in Measurement and Assessment.

Kim, S. Y. (2018). *Simple structure MIRT equating for multidimensional tests* [Doctoral Dissertation]. University of Iowa.

Kim, S. Y., Lee, W., & Kolen, M. J. (2019). Simple-structure multidimensional item response theory equating for multidimensional test. *Educational and Psychological Measurement, 80*(1), 91-125.

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking: Methods and practices*. Springer Science and Business Media.

Kumlu, G. (2019). *Test ve alt testlerde eşitlemenin farklı koşullar açısından incelenmesi* [Doktora Tezi]. Hacettepe Üniversitesi.

Lee, E. (2013). *Equating multidimensional test under a random groups design: A comparison of various equating procedures* [Doctoral Dissertation]. University of Iowa.

Lee, G., & Lee, W. (2016). Bi-factor MIRT observed-score equating for mixed-format tests. *Applied Measurement in Education, 29*, 224-241.

Lee, W., & Brossman, B. G. (2012). Observed score equating for mixed-format tests using a simple-structure multidimensional IRT framework. In M. J. Kolen & W. Lee (Eds.), *Mixed-format tests: Psychometric properties with a primary focus on equating* (Vol. 2.2, s. 115-142). Center for Advanced Studies in Measurement and Assessment.

Lee, W. C., He, Y., Hagge, S., Wang, W., & Kolen, M. J. (2012). Linking methods for the full-information bifactor model under the common-item nonequivalent groups design. In M. J. Kolen & W. Lee (Eds.), *Mixed-format tests: Psychometric properties with a primary focus on equatingn* (Vol. 2.2, s. 13–44). Center for Advanced Studies in Measurement and Assessment.

Meng, H. (2007). *A comparison study of IRT calibration methods for mixed-format tests in vertical scaling* [Unpublished Doctoral Dissertation]. University of Iowa.

Panidvadtana, P., Sujiva, S., & Srisuttiyakorn, S. (2021). A Comparison of the accuracy of multidimensional IRT equating methods for mixed-format tests. *Kasetsart Journal of Social Sciences, 42*, 215-220.

Peterson, J. L. (2014). *Multidimensional item response theory observed score equating methods for mixed-format tests* [Doctoral dissertation], University of Iowa.

R Development Core Team. (2022). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Tao, W., & Cao, Y. (2016). An extension of IRT-based equating to the dichotomous testlet response theory model. *Applied Measurement in Education, 29*, 108-121.

Uğurlu, S. (2020). *Comparison of equating methods for multidimensional test which contain items contain items with differential item functioning* [Doktora Tezi]. Hacettepe Üniversitesi.

Wang, S., & Liu, H. (2018). Minimum sample size needed for equipercentile equating under the random groups design. In M. J. Kolen & W. Lee (Eds.), *Mixed-format tests: Psychometric properties with a primary focus on equating* (Vol. 2.5, s. 107–126). Center for Advanced Studies in Measurement and Assessment.

Wang, T. (2006). Standard errors of equating for equipercentile equating with log-linear presmoothing using the delta method (CASMA Research Report No. 14). Center for Advanced Studies in Measurement and Assessment, The University of Iowa.

Yao, L., & Boughton, K. (2009). Multidimensional linking for tests with mixed item types. *Journal of Educational Measurement, 46*(2), 177–197.

Zhang, O. (2012). *Observed score and true score equating form multidimensional response theory under nonequivalent group anchor test design* [Doctoral Dissertation]. University of Florida.

Extended Abstract

Introduction

In order to ensure test security, when large-scale standard tests are applied, applications are carried out using different forms of the same test at different times. In these applications, alternative test forms are used to ensure test security. Equivalent test forms are used to make the scores comparable. Parallel test forms are developed to comply with the same content and statistical properties. However, it is almost impossible to create test forms that are completely equivalent in terms of statistical properties such as item difficulty. As a result of test equating, the effects caused by the differences in difficulty of the tests will be eliminated and the scores will be placed on the same scale. Although the full WM-RTK observed score, WM-RTK observed score equating one-dimensional approach equating methods were developed by Brossman (2010) and the Bi-factor WM-RTK observed score equating method by Lee and Lee (2016), not enough studies have been done to determine how these methods perform under various conditions. Considering this situation, the aim of this study is to compare the equated scores obtained under the common item design in unequal groups from the bi-factor AM-IMK observed score equating (BF-MIRT), full AM-IMK observed score (MOSE) and AM-IMK observed score equating one-dimensional approach (AOSE) methods used in equating the scores obtained from multidimensional tests, and the standard error of equating (SEE), bias (BIAS) and root mean square error (RMSE) values of these scores according to various factors.

Method

In the study, simulation data sets produced under the common item design in unequal groups were used. Therefore, the research has the nature of Monte Carlo simulation research. Simulation studies are experimental designs carried out in a computer environment that involve generating data through pseudo-random sampling from known probability distributions. In the study, the sample size, common item ratio, correlation level between dimensions, multidimensional test equating methods (3), and calibration methods (2) include different conditions. The combination of different levels of these variables resulted in a total of 162 (3x3x3x2x3) conditions. Each simulation condition was repeated 100 times. In the study, binary scored item response data was produced for ease of application and understanding. Responses for each group under each condition were produced using the compensatory multidimensional three-parameter logistic (M-3PLM) IRT. A bi-factor model consisting of one general factor (g) and two specific factors (f1 and f2) was used to create the data sets. The first thirty items loaded on the general and first specific factor. The last thirty items loaded on the general and second specific factor. Data analysis was performed using the R programming language (R Development Core Team, 2022). The package ‘stats’ (R Development Core Team, 2022) was used to generate item parameters and the package ‘mvtnorm’ (Genz et al., 2021) was used to generate ability parameters.

Findings

It was observed that the equated scores obtained as a result of the test equating applications performed using the concurrent calibration method were similar for the BF-MIRT and MOSE equating methods. It was observed that the equated scores obtained as a result of the test equating applications performed using the separate calibration (Stocking Lord) method were not similar. The highest SEE values were observed in the test equating applications performed using item and ability parameters obtained as a result of separate (Stocking Lord) calibration. In the test equating applications performed with concurrent calibration, the smallest SEE values were observed in the MOSE method, while the largest SEE values were observed in the AOSE method. In the separate calibration method, the smallest SEE values were observed in the AOSE method, whereas the largest SEE values were observed in the BF-MIRT method. The highest bias values were observed in the test equating applications performed using item and ability parameters obtained as a result of separate calibration. In the test equating applications performed with concurrent and separate calibration, the smallest bias values were observed in the AOSE method, while the largest bias values were observed in the BF-MIRT method.

Discussion, Conclusion and Suggestions

Lee (2013) observed in his study examining multidimensional test equating methods that the weighted error values decreased as the sample size increased. In this study, similar to the literature, it was observed that the error values decreased as the sample size increased for all methods as a result of the test equating applications performed using the concurrent and separate calibration method. In the separate calibration method, it was observed that as the common item ratio increases, the error values decrease and that the 50% common item ratio is more efficient. As the common item ratio increases, the relationship between the common item and the entire test becomes more reliable. Therefore, it can be expected to obtain less equating error (Hou, 2007). Lee and Lee (2016) observed that the RMSD values obtained from the BF-MIRT method increased as the level of relationship between dimensions increased. Similar to this finding, the error values obtained from the BF-MIRT and MOSE methods increased as the level of relationship between dimensions increased, that is, from multidimensionality to unidimensionality. Smaller error values were observed in the test equalization applications performed using the concurrent calibration method compared to the test equalization applications performed using the separate calibration method. Among both calibration methods, BF-MIRT method showed the worst performance. The largest error values were observed in BF-MIRT method for both calibration methods. For test equating applications using the concurrent and separate calibration method, the sample size should be 3000 or more. In test equalization applications performed using concurrent and separate calibration methods, the level of correlation between dimensions should be close to r = 0 when BF-MIRT and MOSE methods are used. When the AOSE method is used, it should be close to r = 0,30 in concurrent calibration and at least r = 0,60 in separate calibration. In this study, BF-MIRT, MOSE or AOSE methods were used. When the literature was reviewed, no study was found that compared all multidimensional test equating methods. All multidimensional test equating methods can be compared under different conditions. The statistical significance of the differences between error values was not examined in the study. The significance of error differences can be examined in future studies.

Ek-A: Post-Hoc Tukey Testi Sonuçları

Birinci alt probleme ait Tukey fark testi sonuçları

Tablo 3. Eş zamanlı kalibrasyona ait sonuçlar

	Fark	**SE**	**df**	**t**
BF-MIRT A – MOSE A	-0,45	0,26	116	-1,68
BF-MIRT A – AOSE A	0,32	0,26	116	1,19
MOSE A – AOSE A	0,77	0,26	116	2,88
BF-MIRT O – MOSE O	-0,16	0,13	116	-1,26
BF-MIRT O – AOSE O	-0,65*	0,13	116	-4,93
MOSE O – AOSE O	-0,48*	0,13	116	-3,66
BF-MIRT U – MOSE U	0,85*	0,26	116	3,19
BF-MIRT – AOSE U	3,56*	0,26	116	13,28
MOSE U – AOSE U	2,70*	0,26	116	10,09

Not: A: alt, O: orta, U: üst, * p<0,05.

Tablo 4. Ayrı kalibrasyona ait sonuçlar

	Fark	**SE**	**df**	**t**
BF-MIRT A - MOSE A	4,67*	0,75	116	6,20
BF-MIRT A – AOSE A	9,15*	0,75	116	12,13
MOSE A – AOSE A	4,47*	0,75	116	5,93
BF-MIRT O – MOSE O	1,14	0,37	116	3,06
BF-MIRT O – AOSE O	2,65*	0,37	116	7,11
MOSE O – AOSE O	1,50*	0,37	116	4,04
BF-MIRT U – MOSE U	-2,74*	0,75	116	-3,63
BF-MIRT – AOSE U	-5,26*	0,75	116	-6,97
MOSE U – AOSE U	-2,52*	0,75	116	-3,33

Not: A: alt, O: orta, U: üst, * p<0,05.