RANDOM GRUP DESENİ ALTINDA TAM MIRT EŞİTLEMEDE ÖRNEKLEM BÜYÜKLÜĞÜNÜN ETKİSİ

THE EFFECT OF SAMPLE SIZE ON FULL MIRT EQUATING UNDER RANDOM GROUP DESİGN

Burcu DEMİRÖZ[1] Nuri DOĞAN[2]

Başvuru Tarihi: 15.09.2023 Yayına Kabul Tarihi:11.05.2024 DOI: 10.21764/maeuefd.1361350
(*Araştırma Makalesi)*

Özet: Bu çalışmanın amacı random grup deseni altında Tam ÇB-MTK gözlenen puan eşitleme yönteminin doğruluğu üzerinde farklı örneklem büyüklüğü düzeylerinin etkisini boyutlarda yer alan madde sayısı koşulu altında belirlemektir. Çalışmada random grup deseni altında üretilen simülasyon veri setlerinden yararlanılmıştır. Geniş bir örneklem büyüklüğü aralığını incelemek için, örneklem büyüklüğü (N) 500'den 8.000'e kadar 500'er arttırılmıştır. 16 örneklem büyüklüğü düzeyi ve boyutlarda yer alan 3 farklı madde sayısı koşulu için eşitlemenin standart hatası (SEE), yanlılık (BIAS) ve hata kareler ortalamasının karekökü (RMSE) değerleri incelenmiştir. Araştırmanın sonuçlarına göre SEE ve RMSE değerleri uç değerlere doğru artmaktadır. BIAS değerleri, ortalama ham puanın altındaki puanlar için negatif, üstündeki puanlar için ise pozitiftir. Örneklem büyüklüğü arttıkça SEE, BIAS ve RMSE değerlerinin azaldığı gözlenmiştir. Örneklem büyüklüğü 4000 ve üzerinde olduğunda hata değerlerinde önemli değişim gözlenmemektedir. Random grup deseni altında Tam ÇB-MTK gözlenen puan eşitleme yöntemi için boyutlarda yer alan madde sayısı koşulu altında 4000 örneklem büyüklüğünün yeterli olduğu sonucuna ulaşılmıştır.

Anahtar Sözcükler: *Test eşitleme, hata, örneklem büyüklüğü, çok boyutlu madde tepki kuramı, tam ÇB-MTK gözlenen puan eşitleme*

Abstract: The purpose of this study is to determine the effect of different sample size levels on the Full MIRT observed score equating method under the random group design contingent upon the number of items present in the dimensions. The study utilized simulated data sets generated under the random group design condition. To examine a wide range of sample size, the sample size (N) was incremented by 500 from 500 to 8,000. The standard error of equating (SEE), bias (BIAS), and root mean square error (RMSE) values of equating were examined for 16 sample size levels and 3 different numbers of items in dimensions. According to the results of the research, SEE and RMSE values increase towards outlier values. BIAS values are negative for scores below the mean raw score and positive for scores above the mean raw score. It was observed that as the sample size increases, SEE, BIAS and RMSE values decrease. When the sample sizes are 4000 or above, there is no significant change in error values. The study concluded that sample size of 4000 is sufficient under the item number condition in the dimensions for the Full MIRT observed score equating method under the random group design.

Keywords: *Test equating, error, sample size, mirt, full MIRT observed score equating*

Giriş

Günümüzde, standartlaştırılmış testlerin test güvenliğini sağlamak ve test maddelerinin gizliliğini korumak amacıyla yıl içerisinde birden fazla kez uygulandığı bilinmektedir. Örneğin ALES, nisan, temmuz ve kasım aylarında olmak üzere yılda üç defa uygulanmaktadır. Yabancı Dil Bilgisi Seviye Tespit Sınavı (YDS) ilkbahar ve sonbahar dönemi olmak üzere yılda iki defa yapılır. Bu uygulamaların adil olabilmesi için, sınava giren tüm kişilerin testin farklı formlarından karşılaştırılabilir puanlar alması gereklidir. Çok sayıda test formu kullanıldığında test formları farklı psikometrik özellikler gösterebilmektedir. Örneğin testlerin güçlükleri farklılaşabilmektedir. Test puanlarının karşılaştırılabilir olması için test formlarının güçlük farklılıkları düzenlenmelidir. Test eşitleme, aynı özelliklere sahip formlar arasındaki güçlük farklılıklarını düzenlemek için kullanılan istatistiksel bir yöntemdir. Kolen ve Brennan'a (2014) göre test eşitleme, test formlarından elde puanların birbirinin yerine kullanılabilmesi amacıyla kullanılan istatistiksel bir süreçtir.

Test birden fazla gizil değişkeni, yeteneği veya özelliği ölçüyor, birden fazla içerik alanından veya birden fazla madde türünden oluşuyor ise testler arasında eşitleme çok boyutlu test eşitleme yöntemleri kullanılarak gerçekleştirilir. Çok boyutlu testlerin yaygınlaşması ve çok boyutlu madde tepki kuramındaki ilerlemeler sonucunda çok boyutlu test eşitleme yöntemleri de geliştirilmeye başlanmıştır. Çok boyutlu test eşitleme ile ilgili ilk çalışmalar Brossman (2010) tarafından gerçekleştirilmiştir. Bross (2010) Tam Çok Boyutlu Madde Tepki Kuramı (ÇB-MTK) Gözlenen Puan Eşitleme Yöntemi, Çok Boyutlu Madde Tepki Kuramı Gözlenen Puan Eşitleme Tek Boyutlu Yaklaşım, Çok Boyutlu Madde Tepki Kuramı Gerçek Puan Eşitleme Tek Boyutlu Yaklaşım yöntemlerini geliştirmiştir. Daha sonra basit yapılı ÇB-MTK gözlenen puan (Lee ve Brossman, 2012), basit yapılı ÇB-MTK gerçek puan eşitleme (Kim, Lee ve Kolen, 2019), bi-faktör ÇB-MTK gözlenen puan (Lee ve Lee, 2016), bi-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 önerilmiştir.

Çok boyutlu test eşitleme yöntemleri üzerinde çalışmalar devam etmektedir. Çok boyutlu test eşitleme yöntemlerinin çeşitli koşullar altında nasıl performans gösterdiğini belirlemek için incelemeler yapılmaktadır (Choi,2019; Kim,2022; Peterson,2014; Zhang, 2012). Örneklem büyüklüğü, madde sayısı, ortak madde oranı, boyut sayısı, boyutlar arası ilişki düzeyi ve kalibrasyon yöntemleri bu araştırmalarda incelenen koşullardan bazılarıdır (Bolt, 1999; Kumlu, 2019; Panidvadtana ve vd., 2019; Pekmezci, 2018; Zor, 2023). Test eşitlemede doğru sonuçlar elde etmek için yeterince büyük bir örneklem büyüklüğüne sahip olmak gerekmektedir (Atar ve Yeşiltaş, 2017;

Kim, Lee ve Kolen, 2019, Lee, 2013; Li ve Lissitz, 2000; Skaggs ve Lissitz, 1986; Tate, 2003). Örneklem büyüklüğünü artırmak, eşitleme sonuçlarının doğruluğunu iyileştirmek için en sık kullanılan yöntemlerden biridir. Örneklem büyüklüğündeki artışlar eşitlemenin doğruluğunu arttırmaya yani eşitlemenin standart hatasını azaltmaya yardımcı olmaktadır (Çokluk ve vd., 2022; Gök ve Kelecioğlu, 2014; Lee ve vd., 2014; Kilmen ve Demirtaşlı, 2012; Kolen ve Brennan, 2014; Tsai, 1997; Wang, 2006). Küçük örneklemlerle eşitleme yapmanın kritik yönü, eşitleme hatalarının oldukça büyük olabilmesidir. Eşitlemenin standart hatasını kestirmeye yönelik formüller bu küçük örneklemlerde geçerli olmayabilir. Bu eşitlemenin kesinliğine ilişkin tahminlerin yanlış olduğu anlamına gelir (Parshall, Houghton ve Kromrey,1995). Örneklem büyüklüğü ile ilgili bir diğer durum ise madde tepki kuramına dayalı yöntemler kullanıldığında madde ve yetenek parametre kalibrasyonları ile eşitlemenin parametre tahminlerinden etkilenmesidir. Parametre tahminlerinin kararlılığını ve doğruluğunu etkileyen iki faktör vardır: örneklem büyüklüğü ve madde sayısı (Hambleton ve Cook, 1983). Madde ve yetenek parametreleri örneklemden kestirilir. Bu nedenle örneklem büyüklüğü parametre tahminlerini etkilemektedir. Örneklem büyüklüğü, madde ve yetenek parametre kalibrasyonları ile eşitlemeyi dolaylı olarak etkilemiş olur. Örneklem büyüklüğünün parametre tahminlerine etkilerini gösteren önemli çalışmalar bulunmaktadır (Baldwin, 2006; Barnes ve Wise, 1991; Harwell ve Janosky, 1991; Linacre, 1994; Lord ve Wingersky, 1984; Parshall ve vd., 1997; Ree ve Jensen, 1983). Örneğin Ree ve Jensen a, b ve c parametrelerinin tahmin edilmesi için geniş bir yetenek aralığında büyük örneklemlere ihtiyaç duyulduğunu belirtmiştir. Linacre (1994), örneklem büyüklüğü arttıkça, madde ve yetenek parametre kalibrasyonlarında daha küçük hata değerleri gözlenmiştir. Örneklem büyük ise, örneklemdeki kalibrasyon ve eşitleme ilişkileri popülasyondaki kalibrasyon ve eşitleme ilişkilerini doğru bir şekilde temsil edebilir. Bununla birlikte, büyük örneklemlere ulaşmak zor ve maliyetlidir. Bu nedenle, kabul edilebilir kalibrasyon ve eşitleme sonuçları elde etmek için uygun örneklem büyüklüğünü belirlemek gereklidir. Alanyazın incelendiğinde tek boyutlu test eşitleme yöntemleri için gerekli olan örneklem büyüklüğünün incelendiği görülmüştür (Gök ve Kelecioğlu, 2014; Livingston ve Kim,2010; Kilmen,2010; Wang ve Liu, 2018). Alanyazında Tam ÇB-MTK gözlenen puan eşitleme yöntemi kullanılarak test eşitleme yapabilmek ve doğru sonuçlar elde edebilmek için gerekli örneklem büyüklüğünün henüz incelenmediği gözlenmiştir. Bu nedenle eşitleme yöntemin performansının farklı örneklem büyüklüğü düzeylerinde incelenmesi gerekmektedir. Bu çalışmanın amacı random grup deseni altında Tam ÇB-MTK gözlenen puan eşitleme yönteminin doğruluğu üzerinde farklı örneklem büyüklüğü düzeylerinin etkisini boyutlarda yer alan madde sayısı koşulu altında belirlemektir. Bu amaç doğrultusunda aşağıdaki problem cümlesine yanıt aranmıştır.

Problem Cümlesi

Çok boyutlu Madde Tepki Kuramına dayalı Tam ÇB-MTK gözlenen puan eşitleme yönteminden elde edilen eşitlemenin standart hatası (SEE), yanlılık (BIAS) ve hata kareler ortalamasının karekökü (RMSE) değerleri, boyutlarda yer alan madde sayısı (5-15, 10-10, 15-5) koşulu altında farklı örneklem büyüklüğü düzeylerine bağlı olarak nasıl değişmektedir?

Yöntem

Araştırmanın Modeli

Bu çalışmada, random grup deseni altında üretilen simülasyon veri setlerinden yararlanılmıştır. Bu nedenle çalışma, simülasyon araştırması niteliği taşımaktadır. Simülasyon araştırması, birçok senaryoyu kapsayabilen daha genel analitik sonuçların aksine, belirli senaryolarda istatistiksel yöntemlerin performansı hakkında ampirik sonuçlar elde etmek için kullanılır (Morris, White ve Crowther, 2019).

Araştırma Deseni

Çalışmada random grup deseni kullanılmıştır. Random grup deseninde; ortak bir evrenden gelen bireyler test formlarına random olarak atanır. İki gruba farklı test formları uygulanır (Cook ve Eignor, 1991). Böylece ortak bir evrenden gelen bireyler, benzer X ve Y formlarına rastgele atanmış olur (Kolen ve Brennan, 2014). Random grup deseninin pratik bir özelliği, her bireyin sadece bir test formunu almasıdır. Bu da sınav süresini kısaltır.

Verilerin Üretilmesi

Madde veri setleri üretilirken R 4.2.2 (R Development Core Team, 2022) yazılımı “stats” paketi kullanılarak araştırmacı tarafından üretilmiştir. Çalışmada iki boyutlu basit yapılı çok boyutlu madde tepki kuramı (Basit yapı ÇB-MTK) için madde parametreleri üretilmiştir. Basit yapı, her bir maddenin yalnızca bir faktöre yüklendiği ve diğer faktörlerde çapraz yüklenmelerinin olmadığı yapılardır (McDonald, 2000; Sass & Schmitt, 2010). Diğer bir ifade ile basit yapı, her bir faktörün çok yüksek bir şekilde birkaç maddeye yüklendiği durumlardır. Bunun sonucunda temel gizil yetenek ve madde arasında net bir ilişki gözlenir. Faktörler diğer maddeler ile çok düşük yüklenir. Yetenek ve diğer maddeler arasında bir ilişki yoktur (Finch, 2006; Swygert, McLeod ve Thissen, 2001).

Swaminathan ve Gifford (1983), örneklem büyüklüğü 1000 ve üzerinde olsa dahi çoktan seçmeli testlerinin uzunluğunun 15 maddenin altına düştüğünde madde ayırıcılık parametresinin kötü kestirimler verdiğini rapor etmiştir. Hambleton ve Cook (1983), testlerden istikrarlı sonuçlar elde

etmek isteniyorsa en az 200 sınav katılımcısı ve 20 madde kullanılması gerektiğini belirtmiştir. Bu nedenle bu araştırmada toplam test uzunluğu yirmi madde şeklinde belirlenmiştir. Örneklem büyüklüğü ve boyutlarda yer alan madde sayısı manipüle edilen faktörler olarak ele alınmıştır. Birinci koşulda ilk beş madde birinci, son on beş madde ikinci faktöre yüklenmiştir. İkinci koşulda ilk on madde birinci, son on madde ikinci faktöre yüklenmiştir. Üçüncü koşulda ise ilk on beş madde birinci, son beş madde ikinci faktöre yüklenmiştir.

Eşitleme için iki test formu üretilmiştir. Y eski (referans alınan) test formu, X ise eşitleme yapılacak yeni test formudur. Veriler, iki kategorili ve üç parametreli lojistik madde tepki kuramı modeli (3PL) kullanılarak üretilmiştir. İki form için madde ayırıcılık (a) ve şans (c) parametreleri eşit, madde güçlük parametresi (b) ise eşit olmayacak şekilde üretilmiştir. a parametreleri ortalaması 0.5 standart sapması 0.1 olan çok değişkenli lognormal dağılımdan üretilmiştir. Basit yapının sağlanabilmesi amacıyla her madde için ait olduğu boyuta ait bir a parametresi üretilirken diğer boyut için a parametresi sıfır alınmıştır. c parametreleri 0.05 ile 0.25 arasında değişen tek-biçimli (uniform) dağılımdan üretilmiştir. c parametresinin alabileceği maksimum değer, sınavlarda tercih edilen 4 seçenekli maddelerin şansla doğru yanıtlama olasılığının %25 olması göz önüne alınarak belirlenmiştir. b parametreleri X ve Y formları için -4 ve +4 arasında değişen ortalaması 0 ve standart sapması 1 olan normal dağılımdan üretilmiştir. Yetenek parametreleri (θ) iki boyut için -4 ve +4 arasında değişen çok değişkenli normal dağılımdan aralarında ilişki olmayacak şekilde üretilmiştir.

Bu çalışmanın odak noktası örneklem büyüklüğünün Tam ÇB-MTK gözlenen puan eşitleme yönteminin doğruluğu üzerinde etkisidir. Geniş bir örneklem büyüklüğü aralığını incelemek için, örneklem büyüklüğü (N) 500'den 8000'e kadar 500'er artırılmıştır. Buna göre çalışmada 16 örneklem büyüklüğü düzeyi incelenmiştir. Tablo 1'de çalışmada incelenen boyutlarda yer alan madde sayısı koşulları ve örneklem büyüklüğü düzeyleri belirtilmiştir. Tablo 2'de ise formlara ait betimsel istatistikler yer almaktadır. Tüm örneklem büyüklüğü düzeyleri için betimsel istatistikler aynıdır.

Tablo 1

İncelenen Koşullar ve Düzeyler

Boyutlarda Yer Alan Madde Sayısı	Örneklem Büyüklükleri
Koşul 1: 5 - 15	Düzey 1: 500
Koşul 2: 10 - 10	Düzey 2: 1000
Koşul 3: 15 - 5	Düzey 3: 1500
	Düzey 4: 2000
	Düzey 5: 2500
	Düzey 6: 3000
	Düzey 7: 3500
	Düzey 8: 4000
	Düzey 9: 4500
	Düzey 10: 5000
	Düzey 11: 5500
	Düzey 12: 6000
	Düzey 13: 6500
	Düzey 14: 7000
	Düzey 15: 7500
	Düzey 16: 8000

Tablo 2

Formların Betimsel İstatistikleri

		Ortalama	Standart Sapma	Çarpıklık	Basıklık	Ortalama Farkı
5- 15	X	13.153	2.230	0.181	3.183	0.667
	Y	13.820	3.133	-0.370	2.355	
10- 10	X	11.576	2.843	-0.173	2.718	0.526
	Y	12.102	2.229	-0.230	2.947	
15- 5	X	11.103	2.564	-0.269	2.582	0.187
	Y	10.915	2.208	0.050	2.787	

Simülasyon prosedürü. Basit yapı ÇB-MTK modeli, bir testin iki farklı yeteneği değerlendirmek için tasarlandığını varsayarak, madde ve birey parametrelerini üretmek için bir üretici model olarak kullanılmıştır. Her iki formun da aynı sınav grubundan rastgele seçilen eşdeğer sınav gruplarına verildiği varsayılmıştır. Test eşitleme işlemi ise R 4.2.2 (R Development Core Team, 2022) yazılımında "equate" paketi (Albano, 2016) kullanılarak gerçekleştirilmiştir. Spesifik simülasyon süreci aşağıdaki gibidir;

- Çok boyutlu test eşitlemenin ön koşulu madde ve yetenek parametrelerinin aynı ölçekte olmasıdır. Ölçek kalibrasyon yöntemleri, madde ve kişi parametre tahminlerini aynı ölçeğe yerleştirmek için geliştirilmiştir. Ölçek kalibrasyonları, rotasyonel belirsizlik, korelasyon belirsizliği, orijin ve ölçü birimindeki belirsizliği gidererek yetenek tahminlerini ve madde parametre tahminlerini aynı ölçeklere yerleştirir. Veriler random grup deseni altında basit yapı ÇB-MTK modelinde üretilmiştir. Random grup deseninde aynı evrenden rastgele seçilen eşdeğer gruplar oluşturulduğu için m boyutlu test uzayında orijinlerinin ve ölçü birimlerinin aynı olduğu varsayılmıştır. Böylece orijin ve ölçü birimlerindeki belirsizlik giderilmiştir. Korelasyon belirsizliği sorunu genellikle, gizil yeteneklerin çok değişkenli bir normal dağılım izlediğini ve karşılıklı olarak birbirine dik olduğunu belirterek çözülür (Brossman, 2010; Lee ve Lee, 2016). Bunun için yetenek parametreleri çok değişkenli normal dağılımdan üretilmiştir. Yetenek parametrelerinin diklik koşulu varyans-kovaryans matrisi aracılığı ile ayarlanmıştır. Her iki form da farklı koordinat sistemlerinde olsa da sonuç olarak ortaya çıkan marjinal gözlenen puan dağılımı, madde tepki kuramı modellerinin değişmezlik özelliğinden dolayı koordinat sistemi seçiminden etkilenmemiştir. Böylece tüm belirsizlikler giderildiği için aynı ölçekte parametre tahminleri sağlanmıştır. Bu nedenle herhangi bir kalibrasyon işlemi gerçekleştirilmemiştir.

- Madde ve yetenek parametreleri kullanarak maddelerin doğru cevap olasılıkları tahmin edilmiştir.

- Tam ÇB-MTK gözlenen puan eşitleme prosedürünü gerçekleştirmek için, ÇB-MTK çerçevesinde Lord-Wingersky (Lord ve Wingersky, 1984) algoritmasının değiştirilmiş bir versiyonu kullanılarak θ değerlerinin her bir kombinasyonu için koşullu gözlenen puan dağılımları ($f(x|u)$) belirlenmiştir.

- Çok değişkenli yetenek yoğunluğu $\psi(\theta)$) koşullu gözlenen puan dağılımları ($f(x|u)$) ile çarpılmıştır. Çok değişkenli yetenek yoğunluğunu modellemek için ilişkisiz eksenlere sahip çok

değişkenli standart normal dağılım (θ~MVN(0, I)) kullanılmıştır. Ardından tüm yetenek seviyelerinde toplanarak marjinal gözlenen puan dağılımı elde edilmiştir.

- İki form için de marjinal gözlenen puan dağılımı elde edildikten sonra Tam ÇB-MTK gözlenen puan eşitleme prosedürünün son adımı olan geleneksel eşit yüzdelikli eşitleme yapılmıştır. Eşit yüzdelikli eşitlemede küçük örneklemlerde diğer düzleştirme yöntemlerine göre daha az hata verdiği için (Karagül, 2020; Pak ve Lee, 2014; Puhan, 2011) loglinear pre-smoothing yöntemi kullanılmıştır.

Efron ve Tibshirani (1993) standart hatanın hesaplanabilmesi için 200 tekrarın yeterliği olduğunu ancak bootstrap güven aralığı için 1000 ile 2000 arasında yeniden örnekleme (bootstrap) yapılmasının gerektiğini belirtmişlerdir. Bu nedenle bu çalışmada her örneklem büyüklüğü düzeyi için nispeten kararlı eşitleme sonuçları elde etmek amacıyla 1000 bootstrap yapılmıştır. Böylece her örneklem düzeyi ve boyutlarda yer alan madde sayısı koşulu için 1000 eşitleme işlemi gerçekleştirilmiştir. Farklı örneklem büyüklüğü düzeyleri ve boyutlarda yer alan madde sayısı koşullarından elde edilen eşitleme sonuçlarını değerlendirmek için eşitlemenin standart hatası (SEE), yanlılık (BIAS) ve hata kareler ortalamasının karekökü (RMSE) olmak üzere tüm puan ölçeğindeki hata miktarını gösteren genel istatistikler hesaplanmıştır. Bu üç istatistik her e_x x puanında kriter eşitlenmiş puan ve $\widehat{e_{xr}}$ r. tekrardan elde edilen x puanındaki eşitlenmiş puan olmak üzere aşağıdaki denklemlerle hesaplanmıştır.

$$SEE(x) = \sqrt{\frac{1}{R}\sum_{r=1}^{R}\left[\widehat{e_{xr}} - \left(\frac{1}{R}\sum_{r=1}^{R}\widehat{e_{xr}}\right)\right]^2}$$

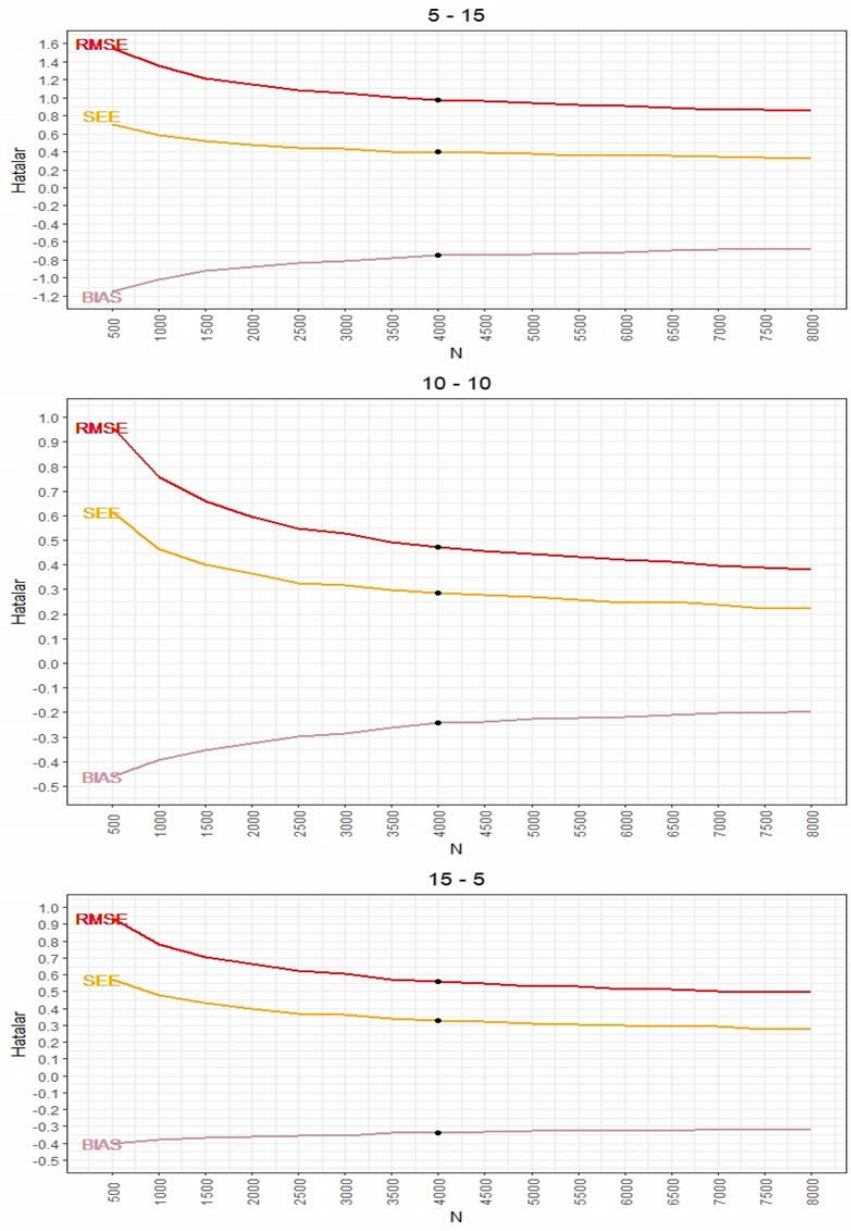
$$BIAS\ (x) = \left(\frac{1}{R}\sum_{r=1}^{R}\widehat{e_{xr}}\right) - e_x$$

$$RMSE(x) = \sqrt{(SEE(x))^2 + (BIAS(x))^2}$$

Her örneklem büyüklüğü düzeyi için tekrarlar sonucunda elde edilen hata değerlerinin ortalaması alınarak örneklem büyüklüğüne ait ortalama bir hata değeri bulunmuştur. Ortalama hata değerlerindeki değişimleri değerlendirmek için ΔHata (ΔBIAS, ΔSEE ve ΔRMSE) değerleri hesaplanmıştır. Örneklem büyüklüğü arttıkça hata değerlerinin önemli ölçüde değişip değişmediğini belirlemek için önerilen kesme kriterleri, ΔSEE ≤ 0,02, ΔBIAS≤ 0,02 ve ΔRMSE ≤ 0,02 olarak alınmıştır (Chen, 2007; Cheung ve Rensvold, 2002).

Bulgular

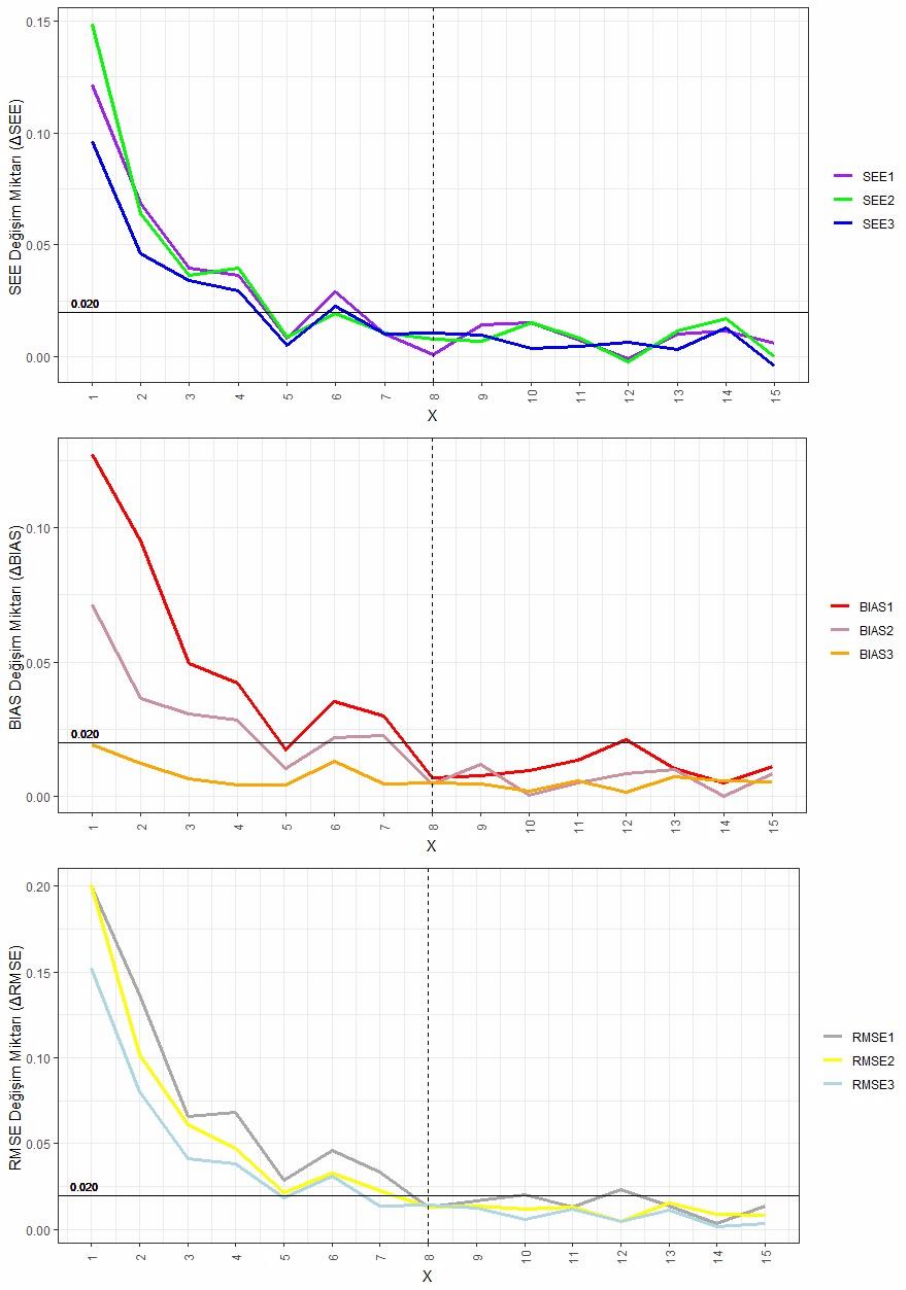
Analizler sonucunda elde edilen SEE, BIAS ve RMSE değerlerinin 16 farklı örneklem büyüklüğü düzeyi ve boyutlarda yer alan 3 farklı madde sayısı koşulu altında aldığı değerler Ek 1'de yer almaktadır. Farklar alındıktan sonra oluşan gruplar X ile temsil edilmektedir. Örneğin N = 500 ile N = 1000 örneklemleri arasındaki fark X = 1 ile temsil edilmiştir. Hata değerlerinin eğilimi Şekil 1'de gösterilmiştir.



Şekil 1. Hata değerleri

Şekil 1 incelendiğinde örneklem büyüklüğü arttıkça SEE, BIAS ve RMSE değerlerinin boyutlarda yer alan madde sayısının üç düzeyi için de azaldığı gözlenmiştir. Hata değerleri incelendiğinde ikinci ve üçüncü koşullarda SEE, BIAS ve RMSE değerlerinin birbirine yakın olduğu gözlenmiştir. En yüksek hata değerleri birinci koşulda gözlenmiştir. Örneğin birinci koşulda N = 500 için SEE değeri 0.70 iken üçüncü koşulda 0.57'dir. En küçük hata değerleri ikinci koşulda gözlenmiştir.

Hata değerlerindeki değişim miktarları (ΔHata) Şekil 2'de gösterilmiştir. Birinci koşula ait ΔHata değerleri SEE1, BIAS1 ve RMSE1; ikinci koşula ait ΔHata değerleri SEE2, BIAS2 ve RMSE2, üçüncü koşula ait ΔHata değerleri ise SEE3, BIAS3 ve RMSE3 şeklinde sunulmuştur.



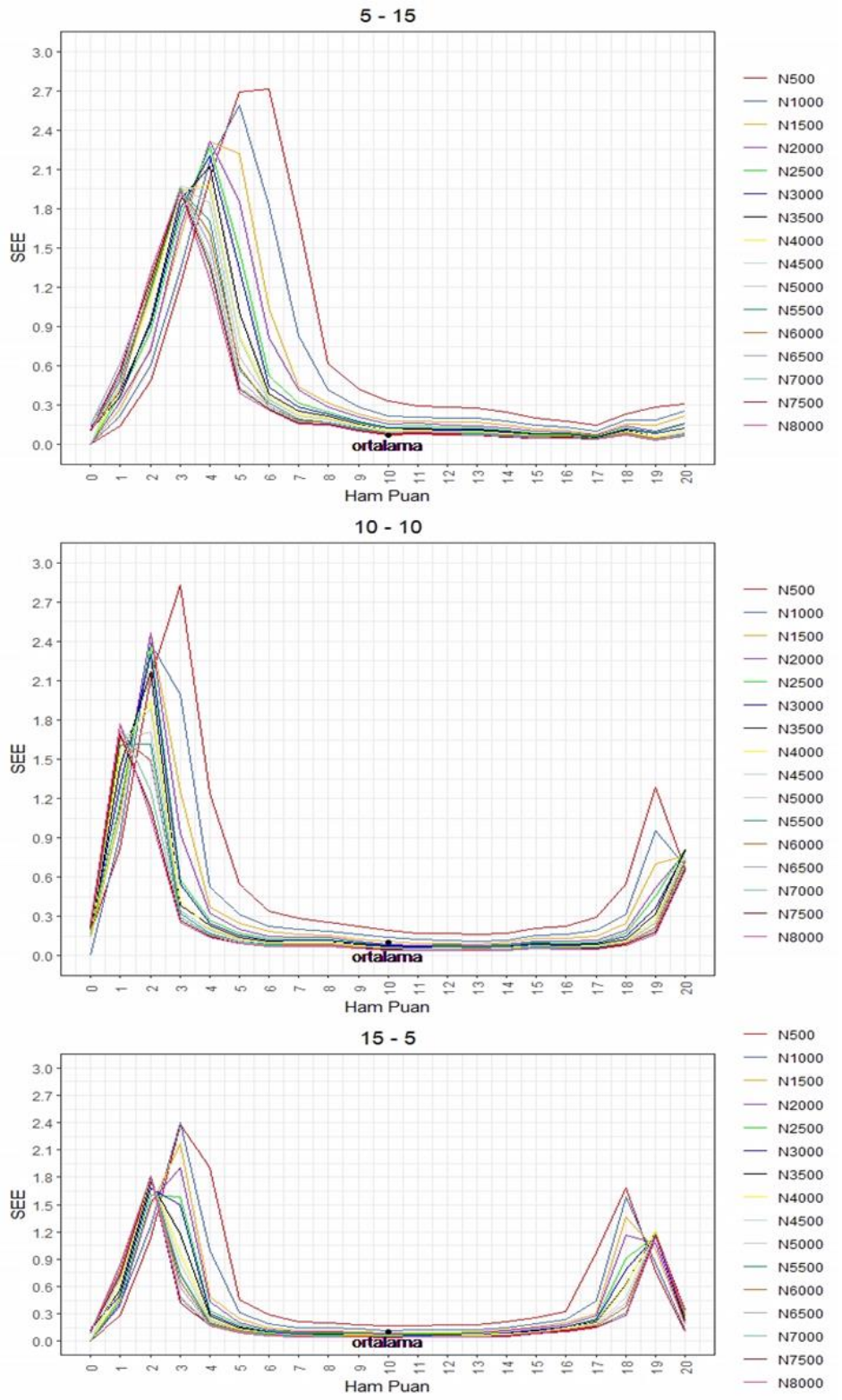
Şekil 2. ΔHata değerleri

Şekil 2 incelendiğinde ΔHata değerlerinin boyutlarda yer alan madde sayısının üç koşulu için de örneklem büyüklüğü arttıkça azaldığı gözlenmiştir. Örneklem büyüklüğü 4000 ve üzerinde olduğunda genel olarak ΔHata değerlerinin 0.020'nin altına düştüğü belirlenmiştir. Hata değerlerinde en büyük azalma, örneklem büyüklüğü 500'den 1000'e çıkarıldığında meydana gelmiştir (Ek 1). Bir başka ifade ile en büyük ΔHata değeri X = 1 için gözlenmiştir. Örneğin ikinci koşulda SEE değeri 0.15, BIAS değeri 0.07 ve RMSE değeri 0.2 azalmıştır.

Tüm koşullarda örneklem büyüklüğü 500'den 4000'e çıkarken SEE, BIAS ve RMSE değerindeki azalma daha çok belirgindir. SEE değerleri arasındaki fark 0.009 ila 0.14 arasında değişmektedir. Örneklem büyüklüğü 4000'den 8000'e çıkarken SEE değerindeki azalma daha az belirgindir. SEE değerleri arasındaki fark 0.0040 ila 0.017 arasında değişmektedir.

BIAS değerindeki azalma koşulların hepsi için örneklem büyüklüğü 500'den 4000'e çıkarken daha çok belirgindir. BIAS değerleri arasındaki fark 0.0041 ila 0.1272 arasında değişmektedir. Örneklem büyüklüğü 4000'den 8000'e çıkarken BIAS değerindeki azalma daha az belirgindir. BIAS değerleri arasındaki fark 0.0001 ila 0.0211 arasında değişmektedir.

Üç koşul için de örneklem büyüklüğü 500'den 4000'e çıkarken RMSE değerindeki azalma daha çok belirgindir. RMSE değerleri arasındaki fark 0.0132 ila 0.2 arasında değişmektedir. Örneklem büyüklüğü 4000'den 8000'e çıkarken RMSE değerindeki azalma daha az belirgindir. RMSE değerleri arasındaki fark 0.0019 ila 0.023 arasında değişmektedir. SEE değerlerinin değişimi Şekil 3'te gösterilmiştir.



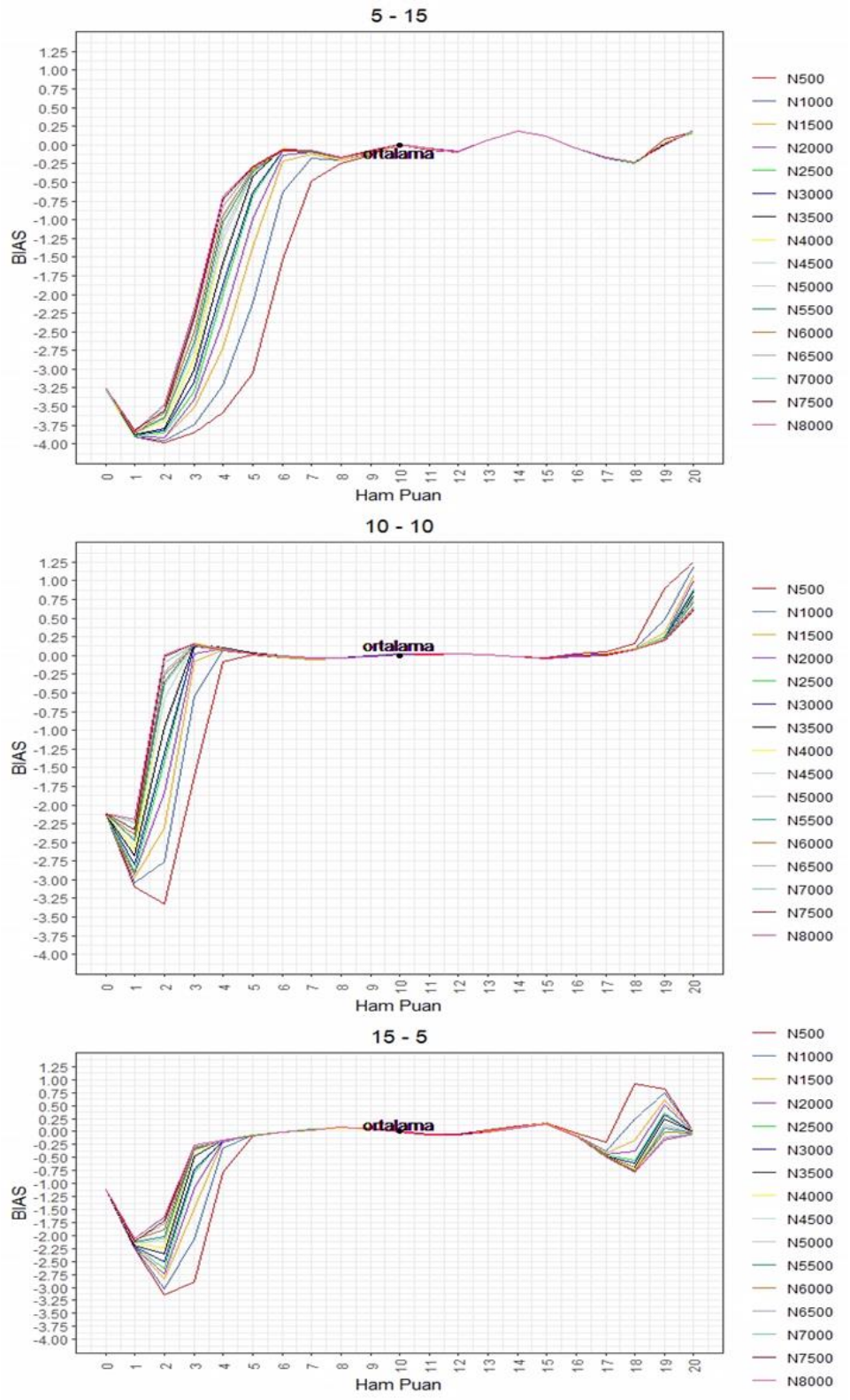
Şekil 3. SEE değişimi

Şekil 3 incelendiğinde genel olarak boyutlarda yer alan madde sayısının üç koşulu için de SEE değerleri ortalama ham puan (x = 10) yakınında en küçük değerdedir. Üç koşul için de SEE değerleri, puanların ortalamadan sapmasının bir fonksiyonu olarak monoton bir şekilde (ancak doğrusal değil) uç değerlere doğru artmaktadır. Örneğin, ikinci koşulda N = 500 için ortalama ham puanda SEE

değeri 0.19 iken x = 2 ham puanında 2.12 olarak gözlenmiştir. İkinci ve üçüncü koşullarda en yüksek SEE değerleri x = 3 ham puanında N = 500 için gözlenmiştir. Fakat birinci koşul için örneklem büyüklüğü 3000 ve üzerinde olduğunda ortalamanın üstündeki ham puanlar için SEE değerlerinin azaldığı gözlenmiştir. Örneğin N = 3000 için ortalama ham puanda SEE değeri 0.121 iken x = 19 ham puanında 0,087 olarak gözlenmiştir.

SEE, üç koşul için de her ham puan düzeyinde örneklem büyüklüğü arttıkça azalmaktadır. En yüksek SEE değerleri N= 500 için gözlenmiştir. En küçük SEE değerleri N = 8000 için gözlenmiştir. Fakat üç örneklem için de x = 1 ham puanında örneklem büyüklüğü arttıkça SEE değeri artmıştır. Örneğin, üçüncü koşulda x = 1 ham puanında N = 500 için SEE değeri 0.28 iken N = 8000 için SEE değeri 0.83 olarak gözlenmiştir.

Tüm koşullar için örneklem büyüklüğündeki farklılıklarla ilişkili SEE değerlerindeki farklılıkların, ortalama ham puandan daha uzak mesafelerdeki puanlar için daha belirgin hale gelmektedir. Tüm örneklem büyüklükleri arasındaki en büyük farklar ham puan ölçeğinin alt ve üst ucundaki değerler için gözlenmiştir. Örneğin, ikinci koşulda ortalama ham puanda N = 500 için SEE değeri (0.19) ile N = 8000 için SEE değeri (0.04) arasındaki fark (0.14) bir puandan azdır. Ancak ortalama ham puandan eksi bir standart sapma uzaklaşıldığında (x = 4) fark (1.09) bir puandan daha fazladır. Benzer durum ortalamadan yaklaşık artı bir buçuk standart sapma uzaklaşıldığında da gözlenmiştir. x = 19 ham puanında N = 500 için SEE değeri (1.28) ile N = 8000 için SEE değeri (0.16) arasındaki fark (1.12) bir puandan daha fazladır. Fakat üç koşul için de x = 0 ham puanında bu durum farklılık göstermektedir. Örneğin, üçüncü koşulda x = 0 ham puanında N = 3000 için SEE değeri (0.13) ile N=8000 için SEE değeri (0.07) arasındaki fark (0.05) bir puandan azdır. BIAS değerlerinin değişimi Şekil 4'te gösterilmiştir.



Şekil 4. BIAS değişimi

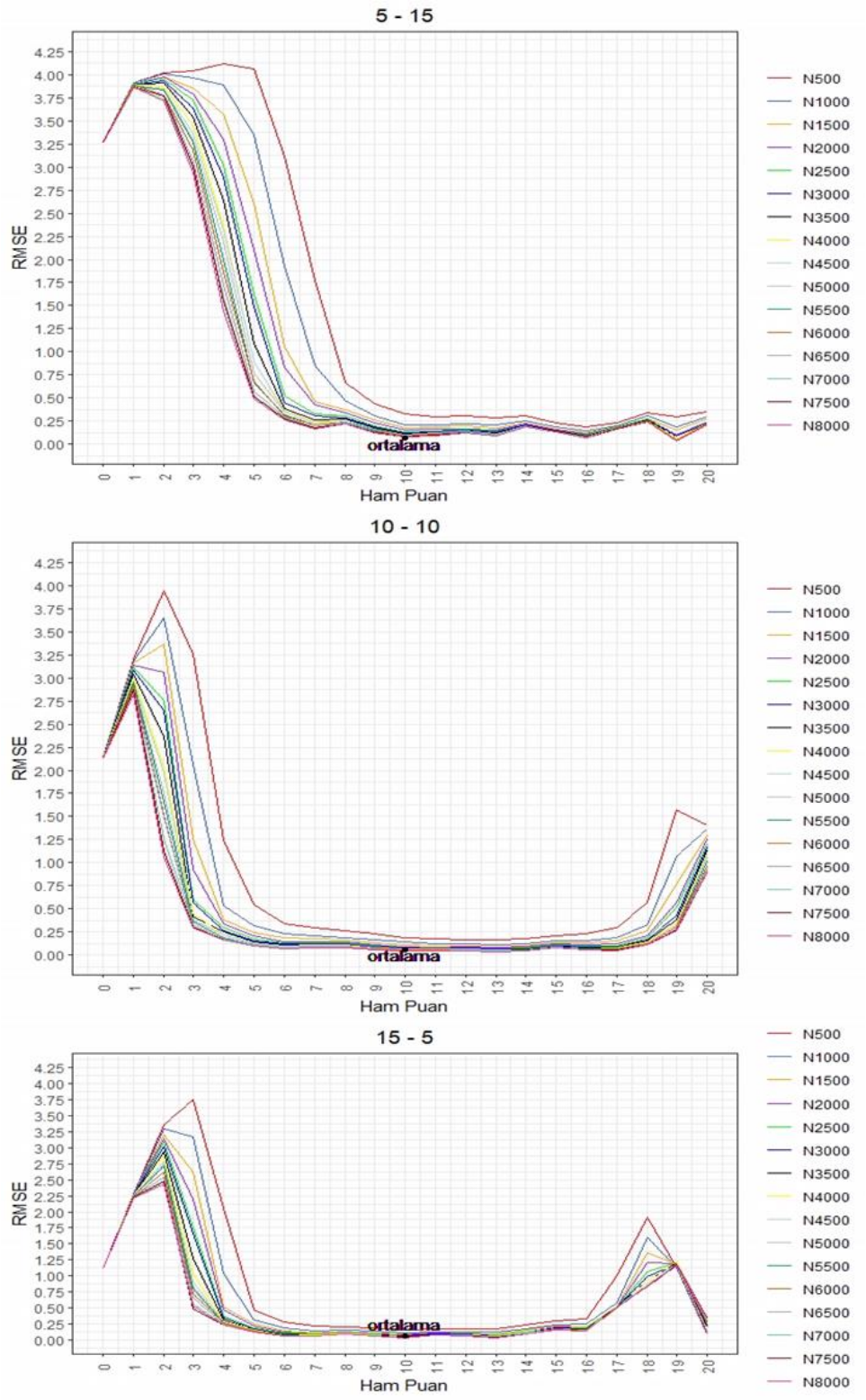
Şekil 4 incelendiğinde boyutlarda yer alan madde sayısının üç koşulu için de BIAS değerleri ortalama ham puan (x = 10) yakınında en küçük değerdedir. Üç koşul için de BIAS değerleri, uç değerlere doğru artar. İkinci ve üçüncü koşullarda 4 ila 19 aralığında yer alan orta ham puanlar için tüm

örneklem büyüklüklerinde küçük BIAS (≤1) değerleri gözlenmiştir. Fakat birinci koşul için örneklem büyüklüğü 6000 ve üzerinde olduğunda benzer durum gözlenmiştir. Örneklem büyüklüğü 5500 ve altında olduğunda 5 ila 19 aralığında yer alan ham puanlar için küçük BIAS (≤1) değerleri gözlenmiştir. İkinci ve üçüncü koşullarda örneklem büyüklüğü 5000 ve üzerinde olduğunda ise 3 ila 17 aralığında yer alan ham puanlar için çok küçük BIAS (≤ 0.5) değerleri gözlenmiştir.

BIAS, üç koşul için de her ham puan düzeyinde BIAS, örneklem büyüklüğü arttıkça azalmaktadır. En yüksek BIAS değerleri N= 500 için gözlenmiştir. En küçük BIAS değerleri N = 8000 için gözlenmiştir. Örneklem büyüklüğü arttıkça BIAS değeri azalmasına rağmen örneklem büyüklüğü 8000 için dahi BIAS mevcuttur.

Tüm koşullar için örneklem büyüklüğündeki farklılıklarla ilişkili BIAS değerlerindeki farklılıkların, ortalama ham puandan daha uzak mesafelerdeki puanlar için daha belirgin hale gelmektedir. 4 ila 18 aralığında yer alan orta ham puanlar için BIAS değerindeki değişim oldukça küçüktür (<1). Örneğin, birinci koşulda x = 8 ham puanında N = 500 için BIAS değeri (-0.24) ile N = 8000 için BIAS değeri (-0.15) arasındaki fark (0.08) bir puandan azdır. Ancak x = 4 ham puanda N = 500 için BIAS değeri (-3.57) ile N = 8000 için BIAS değeri (-0.68) arasındaki fark (2.88) bir puandan daha fazladır.

BIAS değerleri, ortalama ham puanın altındaki puanlar için negatif, üstündeki puanlar için ise pozitiftir. Bu durum üç farklı koşulda da geçerlidir. Örneğin, üçüncü koşulda x = 2 ham puanında N = 500 için BIAS değeri -3.15 iken x = 18 ham puanında 0.92 olarak gözlenmiştir. RMSE değerlerinin değişimi Şekil 5'te gösterilmiştir.



Şekil 5. RMSE değişimi

Şekil 5 incelendiğinde boyutlarda yer alan madde sayısının üç koşulu için de RMSE değerleri ortalama ham puanın (x = 10) yakınında en küçük değerdedir. Üç koşul için de RMSE değerleri, puanların ortalamadan sapmasının bir fonksiyonu olarak monoton bir şekilde (ancak doğrusal değil)

uç değerlere doğru artar. Örneğin, birinci koşulda N = 4000 için ortalama ham puanda RMSE değeri 0.10 iken x = 1 ham puanda 3.88 olarak gözlenmiştir. RMSE, üç koşul için de her ham puan düzeyinde örneklem büyüklüğü arttıkça azalmaktadır. En yüksek RMSE değerleri N= 500 için gözlenmiştir. En küçük RMSE değerleri N = 8000 için gözlenmiştir.

Tüm koşullar için örneklem büyüklüğündeki farklılıklarla ilişkili RMSE değerlerindeki farklılıkların, ortalama ham puandan daha uzak mesafelerdeki puanlar için daha belirgin hale gelmektedir. İkinci ve üçüncü koşullarda tüm örneklem büyüklükleri arasındaki en büyük farklar ham puan ölçeğinin alt ve üst ucundaki değerler için gözlenmiştir. Örneğin, ikinci koşulda ortalama ham puanda N = 500 için RMSE değeri (0.19) ile N = 8000 için RMSE değeri (0.04) arasındaki fark (0.15) bir puandan azdır. Ancak ortalama ham puandan eksi bir standart sapma uzaklaşıldığında (x = 4) fark (1.07) bir puandan daha fazladır. Benzer durum ortalamadan yaklaşık artı bir buçuk standart sapma uzaklaşıldığında da gözlenmiştir. x = 19 ham puanda N = 500 için RMSE değeri (1.56) ile N = 8000 için RMSE değeri (0.26) arasındaki fark (1.30) bir puandan daha fazladır. Fakat birinci koşulda ortalamanın üstündeki ham puanlar için RMSE değerlerindeki farklılıkların azaldığı gözlenmiştir. Örneğin ortalama ham puandan yaklaşık artı bir standart sapma uzaklaşıldığında (x = 16) N = 500 için RMSE değeri (0.18) ile N = 8000 için RMSE değeri (0.07) arasındaki fark (0.11) bir puandan azdır. Benzer durum üç koşul için de x = 0 ham puanında gözlenmiştir. Örneğin, birinci koşulda x = 0 ham puanında N = 500 için RMSE değeri (3.264) ile N=8000 için RMSE değeri (3.261) arasındaki fark (0.003) bir puandan azdır.

Tartışma ve Sonuç

Bu çalışmanın amacı random grup deseni altında Tam ÇB-MTK gözlenen puan eşitleme yönteminin doğruluğu üzerinde farklı örneklem büyüklüğü düzeylerinin etkisini boyutlarda yer alan madde sayısı koşulu altında belirlemektir. Bu amaç doğrultusunda simülasyon çalışması gerçekleştirilmiştir. Örneklem büyüklüğü ve boyutlarda yer alan madde sayısı manipüle edilen faktörler olarak ele alınmıştır. Birinci koşulda ilk beş madde birinci, son on beş madde ikinci faktöre yüklenmiştir. İkinci koşulda ilk on madde birinci, son on madde ikinci faktöre yüklenmiştir. Üçüncü koşulda ise ilk on beş madde birinci, son beş madde ikinci faktöre yüklenmiştir. Örneklem büyüklüğü (N) 500'den 8.000'e kadar 500'er artırılmıştır. Bu çalışmada 16 farklı örneklem büyüklüğü düzeyi ve boyutlarda yer alan 3 farklı madde sayısı olmak üzere toplam 48 koşul incelenmiştir. Farklı örneklem büyüklüğü düzeylerinden elde edilen eşitleme sonuçlarını değerlendirmek için SEE, BIAS ve RMSE değerlendirme kriterlerinden yararlanılmıştır.

Eşitleme hatasının büyük çoğunluğu örnekleme hatasıdır. Literatürdeki çalışmalar tek boyutlu test eşitlemede, örneklem büyüklüğü arttıkça örnekleme hatasının en aza indirilebileceğini göstermiştir (Asiret ve Sünbül, 2016; Cui ve Kolen, 2009; Liu ve Kolen, 2011). Goodman (2020)'e göre yeni veya eski form örneklem büyüklüğü arttığında SEE azalmıştır. Araştırmanın sonuçlarına göre tek boyutlu testler ile benzer olarak çok boyutlu test eşitlemede örneklem büyüklüğü arttıkça SEE, BIAS ve RMSE değerleri genel olarak azalmaktadır. Livingston (1993) ve Skaggs (2005) incelenen her örneklem büyüklüğünde, Kök Ortalama Kare Sapma (Root-Mean-Square Deviation / RMSD) değerlerinin azaldığını ancak her seferinde daha küçük miktarlarda değişim olduğunu göstermişlerdir. Bu bulguyla tutarlı olarak SEE, BIAS ve RMSE değerleri her örneklem büyüklüğünde daha küçük miktarda azalmıştır. Wang ve Liu (2018) karma testler için random grup deseni altında eşit yüzdelikli eşitleme için eşitleme doğruluğunun 500 ile 3000 örneklem arasında daha fazla değiştiği belirlenmiştir. 3000 ve 8000 örneklem büyüklükleri arasındaki değişimin daha az olduğu görülmüştür. Bu çalışmada Wang ve Liu (2018)'in çalışmalarıyla benzer olarak örneklem büyüklüğü 4000 ve üzerinde olduğunda hata değerlerinde önemli değişim gözlenmemiştir (ΔHata $\leq$ 0.02). Random grup deseni altında Tam ÇB-MTK gözlenen puan eşitleme yöntemi için boyutlarda yer alan madde sayısı (5-15, 10-10, 15-5) koşulu altında 4000 örneklem büyüklüğünün yeterli olduğu sonucuna ulaşılmıştır.

Bireyler hakkında toplam puana göre kararların verildiği test uygulamaları için Tam ÇB-MTK gözlenen puan eşitleme yöntemi kullanılırken, kesme puanı ve yakınında hata değerleri incelenmelidir. Tsai (1997) random grup deseni altında doğrusal ve eşit yüzdelikli eşitleme yöntemleri için standart hata değerlerinin ham puan değeri ortalama puana yaklaştıkça azaldığını gözlemiştir. Bu bulguyla benzer olarak testin ortalama ham puanı ve çevresinde tüm örneklem büyüklüğü düzeyleri için hata değerleri minimum değerleri almaktadır. Kesme puanı bu aralıkta ise Tam ÇB-MTK gözlenen puan eşitleme yöntemi tüm örneklem büyüklüğü düzeyleri için kullanılabilirdir. Kesme puanları ortalamanın yakınında olmadığında hata değerleri artmaktadır. Özellikle küçük örneklemlerde ciddi artış gözlenmektedir. Bu nedenle kesme puanı uç değerlere yaklaştığında örneklem büyüklüğü mümkün olduğunca büyük olmalıdır. Benzer şekilde, kesme puanına göre karar vermek yerine testten alınan tüm puanları kullanan uygulamalar için, puan ölçeğinin tümü boyunca standart hatalara dikkat edilmelidir.

Eşitlenen test formları farklı olduğunda BIAS değerleri SEE değerlerinden daha büyüktür. Bu çalışmada boyutlarda yer alan madde sayısı 5-15 ve 15-5 olan koşullarda BIAS değerlerinin SEE değerlerinden daha büyük olduğu gözlenmiştir. BIAS değerleri, SEE değerlerinden daha büyük olduğu için Tam ÇB-MTK gözlenen puan eşitleme yönteminin doğruluğunun test formlarının farklılıklarına bağlı olduğu belirlenmiştir. Boyutlarda yer alan madde sayısı 10-10 olan ikinci koşulda

ise BIAS değerlerinin SEE değerlerinden daha küçük olduğu gözlenmiştir. BIAS değerleri, SEE değerlerinden daha küçük olduğu için Tam ÇB-MTK gözlenen puan eşitleme yönteminin doğruluğunun test formlarının farklılıklarına bağlı olmadığı belirlenmiştir. BIAS miktarını en aza indirebilen bir eşitleme prosedürü, SEE miktarını azaltabilen bir prosedüre tercih edilir (Kim ve vd., 2019). Bu nedenle testler eşitlenirken ikinci koşulda olduğu gibi boyutlarda yer alan madde sayılarının eşit olması tercih edilmelidir. Aynı zamanda boyutlarda yer alan madde sayısı 5-15 ve 15-5 olan koşullarda BIAS değerleri, SEE değerlerinden daha büyük olduğu için, RMSE değerlerine daha fazla katkıda bulunmaktadır. Bunun sonucunda RMSE değerlerine ait grafikler BIAS değerlerine ait grafiklere daha çok benzemektedir. Boyutlarda yer alan madde sayısı 10-10 olduğunda ise SEE değerleri, BIAS değerlerinden daha büyük olduğu için, RMSE değerlerine daha fazla katkıda bulunmaktadır. RMSE değerlerine ait grafikler SEE değerlerine ait grafiklere daha çok benzemektedir.

Çalışmanın birkaç sınırlılığı bulunmaktadır. İlk olarak, bu bir simülasyon çalışması olduğu için sonuçlar yorumlanırken dikkatli olunmalıdır. Tüm simüle edilmiş veriler belirli koşullara dayanmaktadır. 1-0 puanlanan veriler ile sınırlıdır. Boyut sayısı 2 ile sınırlandırılmıştır. Boyutlarda yer alan madde sayısı 3 koşul ile sınırlandırılmıştır. Basit yapı modeli kullanılmıştır. Random grup deseni altında iki grup denk olacak şekilde yetenek parametreleri sınırlandırılmıştır.

Öneriler

Uygulayıcılara Öneriler

- Tam ÇB-MTK gözlenen puan eşitleme yöntemi için boyutlarda yer alan madde sayısı (5-15, 10-10, 15-5) koşulu altında 4000 örneklem büyüklüğünün yeterli olduğu sonucuna ulaşılmıştır. Bu nedenle benzer koşullarda eşitleme uygulaması gerçekleştirileceği zaman 4000 örneklem büyüklüğü kullanılması önerilmektedir.
- Boyutlarda yer alan madde sayısı koşulu incelendiğinde en küçük hata değerleri boyutlarda yer alan madde sayılarının eşit olduğu koşulda gözlenmiştir. Bu nedenle Tam ÇB-MTK gözlenen puan eşitleme yöntemi kullanılırken boyutlarda yer alan madde sayılarının eşit olması önerilmektedir.

Araştırmacılara Öneriler

- Kesin tavsiyelerde bulunmak için daha fazla araştırmaya ihtiyaç vardır, ancak uygulayıcılar bu çalışmanın sonuçlarını eşdeğer testlerin uygulanması durumunda karar verme sürecine rehberlik etmek için kullanabilirler. Bulguların diğer veri kümeleriyle doğrulanması da faydalı olacaktır.

- Çalışmada simülasyon verileri kullanılmıştır. Gerçek veriler kullanılarak çalışma tekrarlanabilir.
- Random grup deseni yerine tek grup ya da denk olmayan gruplarda ortak madde deseni kullanılarak çalışma tekrarlanabilir.
- Kullanılan desenin değişmesi ile hem ikili hem de çoklu madde yanıtlarına izin veren kalibrasyon prosedürleri kullanılabilir. Boyut ve madde sayısı arttırılabilir. Boyutlar arası ilişki düzeyinin farklılaşması incelenebilir. Boyutlarda yer alan madde sayısı oranı farklılaştırılabilir.

Kaynakça

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

Asiret, S., & Sünbül, S. Ö. (2016). Investigating test equating methods in small samples through various factors. *Educational Sciences: Theory and Practice, 16(2)*, 647-668.

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.

Baldwin, P. (2006, April). A modified IRT model intended to improve parameter estimates under small sample conditions. Presented at the National Council on Measurement in Education, San Francisco, USA.

Barnes, L. L. B., & Wise, S. L. (1991). The utility of a modified one-parameter IRT model with small samples. *Applied Measurement in Education, 4*(2), 143–157.

Bolt, D. M. (1999). Evaluating the effects of multidimensionality on IRT true-score equating. *Applied Measurement in Education, 12*(4), 383-407.

Brossman, B. G. (2010). *Observed score and true score equating procedures for multidimensional item response theory* [Doktora Tezi, Iowa Üniversite].

Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling. A Multidisciplinary Journal, 14*, 464–504. https://doi.org/10.1080/10705510701301834

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling, 9*, 233–255.

Choi, J. (2019). *Comparison of mirt observed score equating methods under the common-item nonequivalent groups design* [Doktora Tezi, Iowa Üniversite].

Cui, Z., & Kolen, M. J. (2009). Evaluation of two new smoothing methods in equating: The cubic b-spline presmoothing method and the direct presmoothing method. *Journal of Educational Measurement, 46*(2), 135–158.

Çokluk, Ö., Uçar, A., & Balta, E. (2022). Madde tepki kuramına dayalı gerçek puan eşitlemede ölçek dönüştürme yöntemlerinin incelenmesi. *Ankara Universitesi Egitim Bilimleri Fakultesi Dergisi*. https://doi.org/10.30964/auebfd.1001128

Finch, H. (2006). Comparison of the performance of varimax and promax rotations: Factor structure recovery for dichotomous items. *Journal of Educational Measurement, 43*, 39-52.

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

Hambleton, R. K., & Cook, L. L. (1983). Robustness of item response models and effects of test length and sample size on the precision of ability estimates. In D. J. Weiss (Ed.), *New horizons in testing* (pp. 31–49). New York: Academic.

Harwell, M. R., & Janosky, J. E. (1991). An empirical study of the effects of small datasets and varying prior variances on item parameter estimation in BILOG. *Applied Psychological Measurement, 15*(3), 279–291.

Karagül, A. E. (2020). *Küçük örneklemlerde çok kategorili puanlanan maddelerden oluşan testlerde klasik test eşitleme yöntemlerinin karşılaştırılması* [Yüksek Lisans Tezi, Ankara Üniversitesi]. Yöktez.

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]. Yöktez.

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

Kim, K. Y. (2022). Item response theory true score equating for the bifactor model under the common-item nonequivalent groups design. *Psychological Measurement 46*(6). 479-493.

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. (2014). *Test equating, scaling, and linking: Methods and practices* (3). New York, NY: Springer.

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

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

Lee, E., Lee, W. C., & Brennan, R. L. (2014). *Equating multidimensional tests under a random groups design: A comparison of various equating procedures.* (CASMA Research Report No. 40). Center for Advanced Studies in Measurement and Assessment, The University of Iowa

Lee, G., Lee, W., Kolen, M. J., Park, I. Y., Kim, D. I., & Yang, J. S. (2015). Bi-factor mirt true-score equating for testlet-based tests. *Journal of Educational Evaluation, 28*, 681-700.

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. M. J. Kolen ve W. Lee (Ed.), *Mixed-format tests: Psychometric properties with a primary focus on equating* (vol 2.2, s. 115-142) içinde. Center for Advanced Studies in Measurement and Assessment.

Li, Y. H., & Lissitz, R. W. (2000). An evaluation of the accuracy of multidimensional IRT linking. *Applied Psychological Measurement, 24*(2), 115-138.

Linacre, J. M. (1994). Sample size and item calibration [or person measure] stability. *Rasch Measurement Transactions, 7*(4), 328.

Liu, C., & Kolen, M. J. (2011). Automated selection of smoothing parameters in equipercentile equating. In M. J. Kolen & W. Lee (Eds.), *Mixed-format tests: Psychometric properties with a primary focus on equating (Volume 1)*. (CASMA Monograph Number 2.1) (pp. 237–261). Iowa City, IA: CASMA, The University of Iowa.

Livingston, S. A. (1993). Small-sample equating with log-linear smoothing. *Journal of Educational Measurement, 30*(1), 23–39.

Livingston, S. A., & Kim, S. (2010). Ransom group equating with samples of 50 to 400 test takers. *Journal of Educational Measurement, 47*(2), 175-185.

Lord, F. M., & Wingersky, M. S. (1984). Comparison of IRT true-score and equipercentile observed-score “equatings.” *Applied Psychological Measurement, 8*, 452-461.

McDonald, R. P. (2000). A basis for multidimensional item response theory. *Applied Psychological Measurement, 24*, 99-114.

Morris, T. P., White, I. R., & Crowther, M. J. (2019). Using simulation studies to evaluate statistical methods. *Statistics in Medicine, 1-29*. doi:10.1002/sim.8086

Pak, S., & Lee, W. C. (2014). An investigation of performance of equating for mixed-format tests using only multiple-choice common items. In M. J. Kolen & W. Lee (Eds.), *Mixed-format tests: Psychometric properties with a primary focus on equating (Volume 3).* (CASMA Monograph Number 2.3) (pp. 7–23). Iowa City, IA: CASMA, The 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.

Parshall, C. G., Du Bose Houghton, P., & Kromrey, J.D. (1995). Equating error and statistical bias in small sample linear equating. *Journal of Educational Measurement, 32*, 37–54.

Parshall, C. G., Kromrey, J. D., Chason, W., & Yi, Q. (1997, June). Evaluation of parameter estimation under IRT models and small samples. Paper presented at the Psychometric Society, Gatlinburg, USA.

Pekmezci, F. B. (2018). *İki faktör modelde (bifactor) diklik varsayımının farklı koşullar altında sınanması* (Doktora Tezi), Ankara Üniversitesi, Ankara.

Peterson, J. L. (2014). *Multidımensional item response theory observed score equating methods for mixed-format tests* [Doktora Tezi, Iowa Üniversitesi].

Puhan, G. (2011). Futility of Log-Linear Smoothing when Equating with Unrepresentative Small Samples. *Journal of Educational Measurement, 48*(3), 274-292.

R Core Team (2022). *R: A language and environment for statistical computing* [Computer software]. R Foundation for Statistical Computing, Vienna, Austria. Retrieved from http://www.R-project.org/

Ree, M. J., & Jensen, H. E. (1983). Effects of sampe size on linear equating of item characteristic curve parameters. In Weiss, D. J. (Ed.), *New Horizons in Testing* (pp. 135–146). Elsevier.

Sass, D. A., & Schmitt, T. A. (2010). A comparative investigation of rotation criteria within exploratory factor analysis. *Multivariate Behavioral Research, 45*, 73-103.

Skaggs, G. (2005). Accuracy of random groups equating with very small samples. *Journal of Educational Measurement, 42(4)*, 309-330.

Skaggs, G., & Lissitz, R. W. (1986). IRT test equating: Relevant issues and a review of recent research. *Review of Educational Research, 56* (4), 495-529.

Swaminathan, J., & Gifford, J. A. (1983). Estimation of parameters in the three-parameter latent trait model. In D. J. Weiss (Ed.), *New horizons in testing* (pp. 13–30). New York: Academic

Swygert, K. A., McLeod, L. D., & Thissen, D. (2001). Factor analysis for items or testlets scored in more than two categories. In D. Thissen & H. Wainer (Eds.), *Test scoring* (pp. 217–250). Mahwah, NJ: Erlbaum

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.

Tate, R. (2003). A comparison of selected empirical methods for assessing the structure of responses to test items. *Applied Psychological Measurement, 27*, 159–203.

Tsai, T, H. (1997, March). Estimating minimum sample sizes in random groups equating. Presented at the National Council of Measurement in Education Meeting, Chicago, USA.

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

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

Zhang, O. (2012). *Observed score and tru score equating form multidimensional response theory under nonequivalent group anchor test design* [Doktora Tezi, Florida Üniversitesi].

Zor, Y. M. (2023). Investigation of multidimensional scale transformation methods applied to multidimensional test according to various conditions. *Adiyaman University Journal of Educational Sciences, 13*(1),41-53.

Extended Abstract

Introduction

In recent times, it is known that standardized tests are administered multiple times within a year to ensure the security and confidentiality of the tests. For these test administrations, fairness requires that all individuals taking the test should receive comparable scores from different forms of the test. When multiple test forms are used, the test forms may exhibit different psychometric properties. According to Kolen and Brennan (2014), the definition of equating is a statistical process used to adjust the scores obtained from test forms so that these scores can be interchangeably used. With the widespread use of multidimensional tests and the development of multidimensional item response theory, multidimensional test equating methods have also started to be developed. The initial studies about multidimensional test equating were conducted by Brossman (2010). Brossman (2010) developed the methods of Full Multidimensional IRT observed score equating (MOSE), observed score unidimensional approximation of MIRT equating (AOSE), and true score unidimensional approximation of MIRT equating (ATSE). Subsequently, simple structured MIRT observed score (SMO) (Lee and Brossman, 2012), simple structured MIRT true score equating (SMT) (Kim, Lee, and Kolen, 2019), bi-factor MIRT observed score (Lee and Lee, 2016), bi-factor MIRT true score (Lee et al., 2015), testlet response model MIRT observed-score equating, and testlet response model MIRT true score (Tao and Cao, 2016) methods have also been developed. In the conducted studies, it has been observed that the necessary sample size for performing test equating using the Full Multidimensional IRT observed score equating method and obtaining accurate results has not yet been determined. Hence, it is essential to examine the performance of equating methods across different sample sizes. The purpose of this research is to determine the effect of different sample size levels on the accuracy of the Full Multidimensional IRT observed score equating method under a Random Group Design. With this purpose, answer is sought for the following research question.

Research Question

How does the Standard Error of Equating (SEE), Bias and Root Mean Square Error (RMSE) values of equating obtained from the Full Multidimensional IRT observed score equating method vary with different sample sizes under the conditions of varying item counts in different dimensions (5-15, 10-10, 15-5)?

Method

The software R 4.2.2 (R Development Core Team, 2022) was used for generating item datasets. Item parameters were generated for a two-dimensional simple structure (SS-MIRT) in the study. Data were generated using dichotomous and the three parameter logistic model. Sample size and the number of items in dimensions were treated as manipulated factors. Focus of this study is the effect of sample size on the accuracy of the Full Multidimensional IRT observed score equating. In order to examine a wide sample size range, sample size was increased with an increment of 500 from 500 to 8000. For each sample size levels, 1000 repetitions were performed to obtain relatively stable matching results. To evaluate the equating results obtained under different sample sizes levels and item count conditions, general statistics indicating the error amount on the entire score scale, Standard Error of Equating (SEE), BIAS, and Root Mean Squared Error (RMSE), were calculated. The average error value for each sample sizes levels was obtained by taking the average of the calculated error values for each sample sizes levels. To evaluate the averaged error value, the amounts of change in error values ΔError (ΔBIAS, ΔSEE ve ΔRMSE) were computed.

Findings

It was observed that as the sample size increases, SEE, BIAS, and RMSE values decrease across all three samples. The ΔError values decrease as the sample size increases within the three conditions. It has been determined that when the sample size is 4000 or more, ΔError values generally fall below 0.020. The greatest decrease in ΔError values occurred when the sample size was increased from 500 to 1000. SEE values are lowest near the mean raw score (x = 10) within the three conditions. SEE values across the three conditions increase monotonically (but not linearly) towards extreme values as a function of deviation of scores from the mean. BIAS values are lowest near the mean raw score (x = 10) within the three conditions. Across the three conditions, BIAS values increase as scores move towards extreme values. In the second and third conditions, small BIAS values (≤1) were observed for middle raw scores ranging from 4 to 19 for all sample sizes. RMSE values are lowest near the mean raw score (x = 10) within the three conditions. The three conditions, RMSE values increase monotonically (but not linearly) towards extreme values as a function of deviation of scores from the mean. The highest error values were observed in the first condition.

Discussion, Conclusion and Recommendation

SEE, BIAS, and RMSE values generally decrease as sample size increases. In this study, SEE, BIAS, and RMSE values decreased in smaller increments with increasing sample size. No significant change in error values was observed when the sample size reached 4,000 or more ($\Delta Error \leq 0.02$). Therefore, a sample size of 4,000 is sufficient to obtain equating results with acceptable accuracy for the Tam MIRT observed score equating method under the Random group design. For test applications where decisions about individuals are made based on the total score, the cut score and the error values near it should be examined when using the Tam MIRT observed score equating method. Error values are minimized for all sample size levels around the mean raw score of the test. If the cut score falls within this range, the Tam MIRT observed score equating method can be used for all sample sizes. When cut scores are not near the mean, error values increase, especially for small sample sizes.

NOT: Bu çalışma Hacettepe Üniversitesi Lisansüstü Eğitim-Öğretim ve Sınav Yönetmeliğinin "Tez savunma sınavı jürisinin oluşturulması" bölümünün 5. maddesi kapsamında tez ile ilişkili araştırma makalesi kapsamında geliştirilmiştir.

ETİK BEYAN: "*Random Grup Deseni Altında Tam Mırt Eşitlemede Örneklem Büyüklüğünün Etkisi*" başlıklı çalışmanın yazım sürecinde bilimsel, etik ve alıntı kurallarına uyulmuş; toplanan veriler üzerinde herhangi bir tahrifat yapılmamıştır. Simülasyon araştırması türünde yürütülen bu araştırma, insan ve hayvanların (materyal/veriler dâhil) deneysel ya da diğer bilimsel amaçlarla kullanılması ve insanlar üzerinde yapılan klinik araştırmalardan olmadığından **etik kurul izni gerektirmemektedir.** Karşılaşılacak tüm etik ihlallerde "Mehmet Akif Ersoy Üniversitesi Eğitim Fakültesi Dergisi Yayın Kurulunun" hiçbir sorumluluğunun olmadığı, tüm sorumluluğun Sorumlu Yazara ait olduğu ve bu çalışmanın herhangi başka bir akademik yayın ortamına değerlendirme için gönderilmemiş olduğunu taahhüt ederim.

Ek-1 1000 Tekrar Sonucunda Elde Edilen Ortalama Hata Değerleri

	5 - 15			10 - 10			15 - 5		
N	SEE	BIAS	RMSE	SEE	BIAS	RMSE	SEE	BIAS	RMSE
500	0,7065	-1,1452	1,5509	0,6139	-0,4626	0,9596	0,5743	-0,3998	0,9357
1000	0,5852	-1,0180	1,3506	0,4654	-0,3914	0,7592	0,4781	-0,3808	0,7838
1500	0,5165	-0,9228	1,2142	0,4014	-0,3550	0,6573	0,4319	-0,3686	0,7033
2000	0,4768	-0,8733	1,1482	0,3649	-0,3243	0,5960	0,3978	-0,3621	0,6622
2500	0,4404	-0,8310	1,0800	0,3252	-0,2961	0,5486	0,3684	-0,3579	0,6238
3000	0,4319	-0,8139	1,0512	0,3164	-0,2856	0,5269	0,3632	-0,3538	0,6054
3500	0,4030	-0,7785	1,0053	0,2969	-0,2637	0,4941	0,3407	-0,3407	0,5745
4000	0,3927	-0,7487	0,9719	0,2863	-0,2410	0,4711	0,3304	-0,3361	0,5610
4500	0,3918	-0,7417	0,9587	0,2786	-0,2366	0,4577	0,3200	-0,3312	0,5469
5000	0,3776	-0,7340	0,9421	0,2716	-0,2247	0,4439	0,3101	-0,3266	0,5344
5500	0,3622	-0,7243	0,9218	0,2564	-0,2244	0,4319	0,3063	-0,3284	0,5287
6000	0,3547	-0,7109	0,9084	0,2481	-0,2196	0,4190	0,3016	-0,3227	0,5168
6500	0,3557	-0,6898	0,8851	0,2504	-0,2113	0,4141	0,2951	-0,3243	0,5122
7000	0,3455	-0,6794	0,8716	0,2391	-0,2012	0,3985	0,2918	-0,3170	0,5008
7500	0,3341	-0,6843	0,8679	0,2221	-0,2012	0,3898	0,2788	-0,3229	0,4989
8000	0,3282	-0,6733	0,8539	0,2219	-0,1929	0,3816	0,2828	-0,3177	0,4956