Hacettepe Üniversitesi Eğitim Fakültesi Dergisi (H. U. Journal of Education) 34(4): 1061-1074 [2019]
doi: 10.16986/HUJE.2019049186 [Araştırma Makalesi / Research Article]

Test Eşitlemede Çok Boyutluluğun Eş Zamanlı ve Ayrı Kalibrasyona Etkisi*

The Effect of Multidimensionality on Concurrent and Separate Calibration in Test Equating

Neşe ÖZTÜRK GÜBEŞ**

• ***Geliş Tarihi:*** *04.01.2018* • ***Kabul Tarihi:*** *04.01.2019* • ***Yayın Tarihi:*** *31.10.2019*

Kaynakça Bilgisi: Öztürk Gübeş, N. (2019). Test eşitlemede çok boyutluluğun eş zamanlı ve ayrı kalibrasyona etkisi. *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi, 34*(4), 1061-1074. doi: 10.16986/HUJE.2019049186

Citation Information: Öztürk Gübeş, N. (2019). The effect of multidimensionality on concurrent and separate calibration in test equating. *Hacettepe University Journal of Education, 34*(4), 1061-1074. doi: 10.16986/HUJE.2019049186

ÖZ: Bu araştırmanın amacı, çok boyutluluğun eş zamanlı ve ayrı kalibrasyon yapılarak elde edilen eşitlenmiş puanlara etkisini incelemektir. Araştırma simülasyon verileri kullanılarak yürütülmüştür. Araştırma kapsamında [5 (boyutluluk düzeyi: 0.90, 0.75, 0.50, 0.25 ve 0.00) x 2 (kalibrasyon yöntemi: eş zamanlı ve ayrı kalibrasyon) x 2 (ölçek dönüştürme yöntemi: Stocking-Lord ve Haebara) x 2 (test eşitleme yöntemi: Madde Tepki Kuramı gerçek puan eşitleme ve gözlenen puan eşitleme)] olmak üzere toplam 40 koşul incelenmiştir. Çok boyutluluk testlerin Θ_1 ve Θ_2 olmak üzere farklı iki yeteneği ölçtüğü varsayılarak oluşturulmuştur. İki yetenek arasındaki korelasyonun değeri düştükçe çok boyutluluğun derecesi artmaktadır. İki yetenek arasındaki korelasyonun 0.90 olduğu koşul çok boyutluluğun derecesinin en düşük, iki yetenek arasındaki korelasyonun 0.00 olduğu koşul çok boyutluluğun derecesinin en yüksek olduğu koşulu temsil etmektedir. Eşdeğer olmayan gruplar ortak test deseni altında testler birbirine eşitlenmiştir. Elde edilen eşitlenmiş puanlar yanlılık, standart sapma ve RMSE ölçütleri kullanılarak değerlendirilmiştir. Araştırma bulguları, tüm koşullarda eş zamanlı kalibrasyon yapılarak elde edilen eşitleme sonuçlarının ayrı kalibrasyon ile elde edilenlere göre genel olarak daha yanlı ve daha fazla eşitleme hatasına sahip olduğunu göstermiştir. Standart sapma ölçütüne göre ise çok boyutluluğun derecesinin düşük olduğu koşullarda eş zamanlı kalibrasyon ve ayrı kalibrasyon yapılıp Stocking-Lord ve Haebara ölçek dönüştürme yöntemleri ile elde edilen eşitleme sonuçları benzer performans göstermiştir fakat çok boyutluluğun derecesinin ciddi olduğu koşullarda en az tesadüfi hataya sahip eşitleme sonuçlarının eş zamanlı kalibrasyon yapılarak elde edildiği görülmüştür.
Anahtar Sözcükler: Test eşitleme, çok boyutluluk, eş zamanlı kalibrasyon, ayrı kalibrasyon, ölçek dönüştürme yöntemleri

ABSTRACT: The aim of this study is to investigate the effects of multidimensionality on equating results which obtained from separate and concurrent calibration methods. The study was conducted with using simulated data. In the scope of research, totally 40 simulation conditions [5 (degree of multidimensionality: 0.90, 0.75, 0.50, 0.25, and 0.00) x 2 calibration methods (separate and concurrent) x 2 (scale transformation methods: Stocking-Lord and Haebara) x 2 (test equating methods: IRT true score equating and observed score equating)] were examined. Multidimensionality was constructed as assuming the two test forms measuring Θ_1 and Θ_2 abilities. While the simulation condition which has correlation between abilities 0.90 represents weak multidimensional case, the correlation between abilities 0.00 represents the severe multidimensional case. Tests were equated under common-item non-equivalent groups design. Equating results were evaluated by using bias, standard deviation and RMSE evaluation criteria. The results showed that, generally under all conditions equating results provided from concurrent calibration more biased and had higher RMSE values than equating results provided by separate calibration. Based on standard deviation criteria, when the degree of multidimensionality was low, equating results which got from

* Bu araştırma, 1-3 Eylül 2016 tarihleri arasında Antalya'da düzenlenmiş olan V. Eğitimde ve Psikolojide Ölçme ve Değerlendirme Kongresi'nde sözlü bildiri olarak sunulmuştur.

** Dr. Öğr. Üyesi, Mehmet Akif Ersoy Üniversitesi, Eğitim Fakültesi, Eğitim Bilimleri Bölümü, Eğitimde Ölçme ve Değerlendirme Ana Bilim Dalı, Burdur-TÜRKİYE. e-posta: nozturk@mehmetakif.edu.tr (ORCID: 0000-0003-0179-1986)

concurrent calibration and separate calibration with Stocking-Lord or Haebara scale transformation methods showed similar performance but when the degree of multidimensionality was severe equating results which had lowest random error were provided by concurrent calibration.
Keywords: Test equating, concurrent calibration, separate calibration, scale transformation methods, multidimensionality.

1. GİRİŞ

Bir testin aynı yapıyı ve kapsamı ölçen birden fazla formunun kullanılması testlerden elde edilen puanların karşılaştırılabilmesi için eşitlenmesi gerektiği gerçeğini oluşturur. Test eşitleme, bir testin farklı formlarından elde edilen puanların birbiri yerine kullanılması sağlayan sürece verilen addır. Klasik test kuramı ve madde tepki kuramına (MTK) dayalı birçok test eşitleme yöntemi geliştirilmiştir. MTK'ya dayalı test eşitleme yöntemleri uygulamada birçok test programı tarafından sıklıkla kullanılmaktadır (Kolen ve Brennan, 1995).

MTK'ya dayalı test eşitleme madde kalibrasyonu, ölçek dönüştürme ve eşitleme olmak üzere üç aşamada tamamlanır. İlk basamağı gerçekleştirmek bir diğer ifadeyle madde parametrelerini kestirebilmek için öncelikle uygun MTK modeli ya da modellerine karar vermek gerekir. Tek bir boyutu ölçmek için geliştirilen testler için tek boyutlu MTK modelleri, birden fazla boyutu ölçmek için hazırlanan testler için çok boyutlu MTK modelleri geliştirilmiştir. Tek boyutlu MTK modellerinde bireylerin cevaplarını sadece tek bir örtük değişkenin etkilediği varsayılır. İki kategorili bir diğer ifadeyle 1-0 şeklinde puanlanan maddelere ilişkin madde parametreleri genellikle bir-parametreli lojistik model (1PL), iki-parametreli lojistik (2PL) model ve üç-parametreli lojistik (3PL) modellerinden biri seçilerek kestirilir. Tek boyutlu ikili puanlanan maddeler için geliştirilen MTK modelleri içerisinde geniş ölçekli sınavlarda en çok kullanılanı 3PL modeldir (Huggins, 2012). 3PL model altında Θ yeteneğine sahip bir i bireyinin j maddesine doğru cevap verme olasılığı matematiksel olarak (1) numaralı eşitlikte olduğu gibi ifade edilebilir:

$$P_{ij} = P(\Theta_i; a_j, b_j, c_j) = c_j + (1 - c_j)\frac{\exp[Da_i(\Theta_i - b_j)]}{1+\exp[Da_j(\Theta_i - b_j)]} \qquad (1)$$

Eşitlik 1'de yer alan a_j j maddesinin ayırt edicilik parametresi; b_j güçlük parametresi ve c_j şans parametresidir. Eşitlikteki D ise ölçekleme sabitidir ve değeri 1.7'dir.

Tek boyutlu MTK modelleri, bir bireyin test maddeleri ile etkileşiminin tek bir örtük değişkenle modellenebileceği varsayımına dayanmaktadır (Reckase, 2009). Fakat bireylerin test maddelerine cevap verirken kullandıkları bilişsel süreçler bu kadar basit değildir. Birçok araştırmacı bireylerin testteki performansını çok boyutlu yeteneklerin ya da özelliklerin etkilediğini savunmaktadır (Ackerman, 1996; Reckase, 2009). Örneğin, sözel ifadeler içeren bir matematik problemine öğrencinin cevap verebilmesi için sadece hesaplamaları bilmesi yetmez aynı zamanda sözel ifadeyi anlaması gerekir. Dolayısıyla bu matematik probleminin hem sayısal yeteneği hem de sözel yeteneği ölçtüğü varsayılır (Min, 2007).

Maddelerinin birden fazla yeteneği ölçmeye duyarlı olan testler MTK uygulamalarında birtakım sorunlara neden olurlar. Açıkçası, eğer testteki maddelere verilen cevapları birden fazla örtük özellik etkiliyor ise tek boyutluluk varsayımı ihlal ediliyor demektir ve böyle bir durumda tek boyutlu MTK modellerinin kullanılması tavsiye edilmez (Zhang, 2009). Tek boyutlu MTK modellerinde yer alan yetenek parametrelerinin yenilerinin eklenmesi ile tek boyutlu MTK modellerinin bir uzantısı olarak çok boyutlu MTK modelleri geliştirilmiştir (Ackerman, 1994). Çok boyutlu MTK modellerini Ackerman (1989), telafi edici (compensatory) ve telafi edici olmayan (non-compensatory) modeller olmak üzere iki kategoride sınıflandırmıştır. Telafi edici model, bir yetenek düzeyindeki yetkinliğin diğer boyut veya boyutlardaki eksik yetkinliği tamamlaması durumunda kullanılmaktadır. Telafi edici model boyutların etkileşmesine, bir boyuttaki yüksek yeteneğin ikinci boyuttaki daha düşük yeteneği tamamlamasına izin

vermektedir. Telafi edici olmayan modellerde ise yüksek puan alabilmek için bir bireyin her iki boyutta da yetkin olması gerekmektedir (Yao ve Schwarz, 2006).

Çok boyutlu üç parametreli lojistik model (M-3PL), tek boyutlu 3PL modelin genelleştirilmiş halidir. Çok boyutlu modellerde tıpkı tek boyutlu modellerde olduğu gibi bir maddeye verilen doğru cevap $X_{ij}=1$ ve yanlış cevap $X_{ij}=0$ olarak kodlanır. Madde karakteristik eğrisi yerini doğru cevap verme yüzeyine bırakır. M-3PL modeli (2) numaralı eşitlikte olduğu gibi ifade edilebilir (Reckase, 2009):

$$P(X_{ij}=1|\ \Theta_j, a_i, d_i, c_i) = c_i + (1-c_i)\frac{\exp[a_i'\Theta_j + d_i]}{1+\exp[a_i'\Theta_j + d_i]} \qquad (2)$$

Eşitlik 2'de yer alan Θ_j, M boyutluluktaki örtük yeteneklerin bir vektörü olmak üzere; a_j madde eğimlerinin bir vektörü, d_i madde güçlüğü ile ilişkili skaler bir parametredir.

Uygun MTK modelleri seçildikten sonra test eşitleme sürecinin ilk basamağı olan ayrı (separate) ya da eş zamanlı (concurrent) kalibrasyon yapılarak madde ve yetenek parametreleri kestirilebilir. Ayrı kalibrasyonda her iki test formuna ait madde ve yetenek parametreleri bilgisayar programı her bir form için ayrı çalıştırılarak kestirilir. Eş zamanlı kalibrasyonda ise her iki test formuna ait madde ve yetenek parametreleri eş zamanlı olarak tek bir seferde kestirilir. Her iki test formuna ait bütün parametrelerin eş zamanlı olarak kestirilmesi aynı ölçekte elde edilmelerini garantilerken ayrı kalibrasyon yapılarak kestirilen madde parametreleri aynı ölçekte olmayacaktır. Ayrı kalibrasyon yapıldığında, örtük değişkene ait ölçeğin ortalama ve standart sapması örtük değişkenin dağılımına sabitlenmektedir. Eğer madde parametrelerini kestirmede kullanılan örneklemler farklı evrenlerden geliyor ise kestirilen parametreler farklı ölçeklerde elde edilir (Hanson ve Beguin, 2002).

Test eşitlemede eşdeğer olmayan gruplar ortak test deseni kullanıldığında, farklı formlara ait kestirilen madde ve yetenek parametreleri farklı ölçeklerde elde edilecektir. Testleri eşitleyebilmek için farklı formlara ait madde ve yetenek parametrelerinin aynı ölçekte olması gerekir (Kolen ve Brennan, 2004). Ölçek dönüştürme işleminin temel amacı yeni formun madde ve yetenek parametrelerini eski formun madde ve yetenek parametrelerinin ölçeğine dönüştürecek iki bağlanma (linking) katsayısını bir diğer deyişle eğim (A) ve kesişim (B) sabitlerini bulmaktır. Eski formu "E", yeni formu "Y" harfleri temsil etmek üzere yeni formun yetenek parametreleri (3) numaralı verilen eşitlik ile madde parametreleri (ayırt edicilik ve güçlük) (4) ve (5) numaralı eşitlik ile eski formun ölçeğine dönüştürülebilir. Şans parametresi olasılık ölçeğinde olduğu için herhangi bir dönüştürme işleminin uygulanmasına gerek yoktur (Kolen ve Brennan, 2004):

$$\theta_E = A\theta_Y + B \qquad (3)$$
$$a_{jE} = a_{jY}/A \qquad (4)$$
$$b_{jE} = Ab_{jY} + B \qquad (5)$$

Ayrı kalibrasyon sonucu elde edilen madde ve yetenek parametreleri ancak ölçek dönüştürme işlemi uygulanarak aynı ölçekte ifade edilebilir. Ortalama/sigma (Marco, 1977), ortalama/ortalama (Loyd ve Hoover, 1980), Haebara (Haebara, 1980) ve Stocking-Lord (Stocking ve Lord, 1983) yöntemi yaygın olarak kullanılan dört ölçek dönüştürme yöntemidir (akt. Kolen ve Brennan, 2004). Ortalama/ortalama ve ortalama/sigma yöntemlerine moment yöntemleri de denilmektedir. Ortalama/sigma yönteminde, 4 ve 5 numaralı eşitliklerdeki A (eğim) ve B (kesişim) sabitlerini elde etmek için ortak maddelerden elde edilen b-parametre kestirimlerinin ortalama ve standart sapmaları alınır. Ortalama/ortalama yönteminde ise ortak maddelerden elde edilen a-parametre kestirimlerinin ortalaması alınarak A-sabiti elde edilir. Eşitliklerdeki B-sabitini elde etmek için ise ortak maddelerden elde edilen b-parametre kestirimlerinin ortalaması alınır (Kolen ve Brennan, 2004).

Eğim A ve kesişim B sabitlerinin kestirimleri, ortak maddelerin madde parametrelerinden ziyade madde karakteristik eğrileri arasındaki farka dayalı olarak tanımlanan ölçüt fonksiyonu minimize edilerek de elde edilebilir. Bundan dolayı Stocking ve Lord (1983) bu tür ölçek dönüştürme yöntemlerine karakteristik eğri yöntemleri adını vermiştir. Karakteristik eğri yöntemlerinde, bir test karakteristik fonksiyonu oluşturmak üzere toplanan yeni forma ait madde tepki eğrileri, eski formun ölçeğine dönüştürülür. Haebara ve Stocking-Lord literatürde yaygın olarak kullanılan iki karakteristik eğri ölçek dönüştürme yöntemidir (Kim, 2004). Haebara yönteminde, karakteristik eğriler arasındaki farkın karelerinin toplamı minimize edilerek ölçüt fonksiyon elde edilir. Madde karakteristik eğrileri arasındaki fark, belirli bir yetenek düzeyindeki bireylerin her bir madde için elde edilen madde karakteristik eğrileri arasındaki farkın karesi alınıp toplanarak bulunur. Stocking-Lord yönteminde ise Haebara yönteminin tam tersi olarak maddeler üzerinden toplamın farkının karesi alınmaktadır. Stocking-Lord yönteminde, karesi alınmadan önce her bir parametre seti kestirimi üzerinden toplam alınır. Bu yöntemde süreç, test karakteristik eğrileri üzerinden işlemektedir (Kolen ve Brennan, 2004).

Madde ve yetenek parametreleri kestirilip aynı ölçekte elde edildikten sonra MTK gerçek puan eşitleme ya da MTK gözlenen puan eşitleme yapılarak testler birbirine eşitlenebilir. MTK gerçek puan eşitleme yönteminde, bir test formuna ait θ yetenek düzeyi ile ilişkili gerçek puanın diğer formun θ yetenek düzeyi ile ilişkili gerçek puanına eşdeğer olduğu kabul edilir. MTK'ya dayalı yürütülen bir diğer eşitleme yöntemi MTK gözlenen puan eşitlemedir. MTK gözlenen puan eşitleme yönteminde, MTK modelleri kullanılarak her bir formun gözlenen puan dağılımları kestirildikten sonra eşit yüzdelikli eşitleme yöntemi ile puanlar birbirine eşitlenir (Kolen ve Brennan, 2004).

Literatürde eş zamanlı ve ayrı kalibrasyonun yatay ve dikey eşitlemede karşılaştırıldığı çok sayıda çalışma bulunmaktadır (Albayrak-Sarı ve Kelecioğlu, 2017; Altun ve Kelecioğlu, 2016; Hanson ve Beguin, 1999; Hanson ve Beguin, 2002; Kang ve Petersen, 2009; Kim ve Cohen, 1998; Petersen, Cook ve Stocking, 1983; Wingersky, Cook ve Eignor, 1987). Bu araştırmaların ortak özelliği eş zamanlı ve ayrı kalibrasyon yöntemlerini tek boyutlu veri üzerinde test etmiş olmalarıdır. Fakat gerçek test uygulamalarında veri seti her zaman MTK'nın tek boyutluluk varsayımını sağlayamayabilir. Çok boyutluluğun oluştuğu durumlarda eş zamanlı ve ayrı kalibrasyon sonucu elde edilen test eşitleme sonuçları bu durumdan nasıl etkilenmektedir? Bu sorunun ikili puanlanan testlerde araştırıldığı çalışma literatürde az sayıdadır (Beguin ve Hanson, 2001; Beguin, Hanson ve Glass, 2000).

Beguin, Hanson ve Glass (2000) araştırmalarında çok boyutluluğun ayrı ve eş zamanlı kalibrasyon üzerine etkisini incelemişlerdir. Simülasyon verisi kullanarak yaptıkları araştırmalarında, çok boyutluluğun derecesini boyutlar arasındaki kovaryansı 0.50, 0.70 ve 0.90 olarak alıp üç düzeyde, grupların yetenek dağılımları arasındaki farkı eşdeğer ve eşdeğer olmayan gruplar olmak üzere iki düzeyde alarak incelemişlerdir. Çok boyutluluğu ikili puanlanan maddeler için telafi edici çok boyutlu 3PL model kullanarak oluşturmuşlar ve her bir koşul için parametre kestirimini BILOG-MG programında ayrı ve eş zamanlı kalibrasyon yaparak kestirmişlerdir. Ayrı kalibrasyon yaptıkları koşullarda ölçek dönüştürmeyi Stocking-Lord yöntemini kullanarak gerçekleştirmişlerdir. Araştırmalarının sonucunda, grupların eşdeğer olduğu koşullarda eş zamanlı kalibrasyon yapılarak elde edilen sonuçların Stocking-Lord ölçek dönüştürme yöntemi ile elde edilen sonuçlarla aynı ya da daha düşük değerlendirme ölçütü değerlerine sahip olduğunu bulmuşlardır. Grupların eşdeğer olmadığı koşullarda boyutlar arasındaki kovaryans arttıkça hatanın da arttığını ve çok boyutluluktan eş zamanlı kalibrasyon yapılarak elde edilen sonuçları daha fazla etkilendiği sonucuna ulaşmışlardır. Eşdeğer olmayan gruplar koşullarında boyutlar arasındaki kovaryansın 0.70 ve 0.90 olduğu durumlarda ayrı kalibrasyon yapılarak parametre kestirilmesi daha iyi performans göstermiştir.

Beguin ve Hanson (2001) çalışmalarında çok boyutluluğun MTK test eşitlemede eş zamanlı ve ayrı kalibrasyona etkisini incelemişlerdir. İki boyutlu telafi edici olmayan model altında eşdeğer ve eşdeğer olmayan gruplar ortak test deseni için veri türetmişlerdir. İki örtük değişken arasındaki korelasyonu 0.00, 0.30 ve 0.50 alarak eşdeğer ve eşdeğer olmayan gruplar için veri türetmişlerdir. BILOG-MG ve EPDIRM (Hanson, 2000) bilgisayar programlarını kullanarak ayrı ve eş zamanlı kalibrasyon yaparak parametreleri kestirmişlerdir. Ayrı kalibrasyon yaptıkları koşullarda Stocking-Lord ölçek dönüştürme yöntemini kullanarak formlara ait parametreleri aynı ölçeğe dönüştürmüşlerdir. Araştırmalarında üç faktörü incelemişlerdir: (1) eş zamanlı ve ayrı kalibrasyonu, (2) eş zamanlı kalibrasyonda BILOG-MG ve EPDIRM programlarının performansını, (3) tek boyutlu parametre kestirimi ve çok boyutlu telafi edici olmayan parametre kestirimini. Araştırmalarının sonucunda tek boyutlu model altında, eş zamanlı kalibrasyonun ayrı kalibrasyondan daha düşük ya da ayrı kalibrasyona eşit toplam hata verdiği görülmüştür. Gruplar eşdeğer olduğu durumlarda, ölçek dönüştürme yapılmayan ayrı kalibrasyon sonuçlarının ölçek dönüştürme yapılan ayrı kalibrasyon sonuçlarına benzer ya da onlardan daha iyi performans gösterdiğini bulmuşlardır. Çoğu koşulda, çok boyutlu model ile yapılan kestirimler tek boyutlu modele göre daha düşük toplam hata vermiştir. Çok boyutlu model parametre kestiriminde kullanıldığında, boyutlar arasındaki korelasyon arttıkça toplam hatada artış olduğunu gözlemlemişlerdir.

MTK'nın tek boyutluluk varsayımı gerçek test uygulamalarında her zaman karşılanmayabilir. He ne kadar test verisinde oluşan çok boyutluluğun, tek boyutlu MTK modellerine dayalı uygulanan test eşitlemeye olumsuz etkisinin olduğu araştırmacılar tarafından gösterilmiş olsa da çok boyutlu MTK modellerini kullanmak çok fazla tercih edilmemektedir. Bunun sebepleri arasında kompleks yapıda olmaları ve doğru parametre kestirimlerinin daha zor sağlanması gösterilebilir (Zhang, 2009). Stout (1987) testlerin katı tek boyutluluğundan ziyade temelde tek boyutlu olup olmadığının test edilmesini önermiştir. Bir testin baskın bir boyut ile daha zayıf boyutları ölçmesi *temel olarak tek boyutlu (essentially unidimensional)* olduğunu göstermektedir. Zhang (2009) boyutlar arasındaki korelasyonun yüksek olmasının test verisinin temel olarak tek boyutlu olduğunun bir göstergesi olabileceğini belirtmiştir. Reckase, Ackerman ve Carlson (1988) bir testin bir ölçüde baskın bir boyutun yanında farklı boyutları ölçmesi durumunda tek boyutlu modellerin kullanımının uygun olduğunu analitik ve deneysel olarak göstermiştir. Zhang (2009) eğitimde uygulanan testlerde genellikle ölçülen yapı ile ilişkili ikincil boyutların ana yapıyla ilişkili olduğunu dolayısıyla temel olarak verinin tek boyutluluk varsayımını sağladığını ve tek boyutlu modellerin testle ölçülen ana yapıyı yansıttığını belirtmiştir.

Daha önce de belirtildiği gibi çok boyutluluğun oluştuğu durumlarda eş zamanlı ve ayrı kalibrasyon sonucu elde edilen test eşitleme sonuçlarının bu durumdan nasıl etkilendiğinin incelendiği literatürde çok az sayıda çalışma bulunmaktadır (Beguin ve Hanson, 2001; Beguin, Hanson ve Glass, 2000). Bu araştırmanın amacı, ikili puanlanan testlerde çok boyutluluğun eş zamanlı ve ayrı kalibrasyon yapılarak elde edilmiş eşitleme sonuçlarına etkisini incelemektir. Beguin, Hanson ve Glass (2000) ile Beguin ve Hanson'ın (2001) yaptıkları araştırmalardan bu araştırmanın farkı, çok boyutluluğun etkisini karakteristik eğri ölçek dönüştürme yöntemlerinin her ikisini (Stocking-Lord ve Haebara) ve MTK gerçek puan eşitleme ile gözlenen puan eşitleme yöntemlerini kullanarak elde edilen eşitlenmiş puanlar üzerindeki etkisini incelemesidir. Araştırmanın literatürde bahsedilen bu boşluğu doldurması ve test eşitleme uygulamalarına katkı getirmesi açısından önemli olduğu düşünülmektedir.

2. YÖNTEM

2.1. Araştırma Verileri

Araştırma simülasyon verileri kullanılarak yürütülmüştür. Bu araştırmada madde cevapları Hanson ve Beguin'nin (2002) çalışmasında yer alan madde parametreleri kullanılarak üretilmiştir. Test eşitleme çalışmasında eşdeğer olmayan gruplar ortak test deseni kullanılmıştır. Form X ve Form Y olmak üzere iki test formu oluşturulmuştur. Her bir test formu, 20'si her iki test formunda ortak olmak üzere, toplam 60 çoktan seçmeli maddeden oluşmaktadır.

Araştırma kapsamında [5 (boyutluluk düzeyi) x 2 (kalibrasyon yöntemi) x 2 (ölçek dönüştürme yöntemi) x 2 (test eşitleme yöntemi)] olmak üzere toplam 40 koşul incelenmiştir. Araştırmada çok boyutluluk maddelerin yarısının Θ_1 ve diğer yarısının Θ_2 yeteneğini ölçtüğü varsayılarak oluşturulmuştur. Çok boyutluluğun derecesi madde cevaplarını üretmede kullanılan SimuMIRT (Yao, 2003) programının girdi dosyasında yer alan varyans-kovaryans matrisi aracılığı ile düzenlenmiştir. İki yetenek arasında 0.90, 0.75, 0.50, 0.25 ve 0.00 olmak üzere beş farklı korelasyon değeri alınarak çok boyutlu veri üretilmiştir. Madde parametreleri çok boyutlu 3PL (M-3PL) modeli kullanılarak üretilmiştir. Form X'i (yeni form) Grup 2 ve Form Y'yi (eski form) Grup 1'in aldığı varsayılmıştır.

Simülasyon verilerinin güvenirliğini kontrol etmek için "psych" (Revelle, 2018) R paket programında faktör sayısı iki ile sınırlandırılarak tetrakorik korelasyon matrisine dayalı olarak "principal axis faktör analizi" yapılmıştır. Tablo 1'de Form X ve Form Y için faktörler arası korelasyonların 50 tekrar için hesaplanan ortalamaları görülmektedir.

Tablo 1: Simülasyon verilerinin kontrolüne ait bulgular

Koşullar	Form X	Form Y
Koşul1 ($r_{\Theta 1\Theta 2}$=0.90)	0.80	0.80
Koşul2 ($r_{\Theta 1\Theta 2}$=0.75)	0.70	0.70
Koşul3 ($r_{\Theta 1\Theta 2}$=0.50)	0.49	0.49
Koşul4 ($r_{\Theta 1\Theta 2}$=0.25)	0.25	0.25
Koşul5 ($r_{\Theta 1\Theta 2}$=0.00)	0.00	0.00

Tablo 1'de yer alan veri üretiminde kullanılan ve üretilen veriler üzerinden faktör analizi yapılarak elde edilen faktörler arası korelasyonların Koşul 1, Koşul 2 ve Koşul 3'te her iki form için de öngörülen korelasyon değerine yakın, Koşul 4 ve Koşul 5'te ise birebir aynı olduğu görülmektedir.

Araştırma, eşdeğer olmayan gruplar ortak test deseni kullanılarak yürütüldüğü için çok boyutlu koşul altında, Grup 1'in (Y formunu alan grup) her iki yetenekteki ortalaması sıfır ve standart sapması 1'dir. Grup 2'nin (X formunu alan grup) ise her iki yetenekteki ortalaması 1 birim Grup 1'den daha yüksek olup, standart sapması Grup 1 ile aynı ve 1'dir. Çok boyutlu koşul altında iki değişkenli normal dağılım temel alınarak veri üretilmiştir. Örneklem büyüklüğü 3000 olarak belirlenmiştir ve araştırmada kullanılan tekrar sayısı 50'dir. Veri üretiminde kullanılan koşullar Tablo 2'de özetlenmiştir.

Tablo 2: Veri üretmede kullanılan koşullar

Koşullar	Grupların Yetenek Dağılımı
Koşul1 ($r_{\Theta 1\Theta 2}$=0.90)	
Koşul2 ($r_{\Theta 1\Theta 2}$=0.75)	$(\Theta_1, \Theta_2)_Y \sim BN(\mu_1=0, \mu_2=0, \sigma_1=1, \sigma_2=1)$
Koşul3 ($r_{\Theta 1\Theta 2}$=0.50)	$(\Theta_1, \Theta_2)_X \sim BN(\mu_1=1, \mu_2=1, \sigma_1=1, \sigma_2=1)$
Koşul4 ($r_{\Theta 1\Theta 2}$=0.25)	
Koşul5 ($r_{\Theta 1\Theta 2}$=0.00)	

2.2. Verilerin Analizi

Veri analizi dört aşamada tamamlanmıştır. Veri analizinin ilk aşamasında Bilog-MG (Zimowski, Muraki, Mislevy ve Bock, 1996) bilgisayar programı kullanılarak her iki test formuna ait madde parametreleri eş zamanlı ve ayrı kalibrasyon yapılarak kestirilmiştir. Eş zamanlı ve ayrı kalibrasyon yapımında kullanılan BILOG-MG kodları Ek-1'de sunulmuştur. Eş zamanlı kalibrasyon yapılarak elde edilen Form X ve Form Y'ye ait madde ve yetenek parametreleri aynı ölçekte elde edildiği için ölçek dönüştürme yapmadan MTK gerçek puan eşitleme ve MTK gözlenen puan eşitleme yöntemleri ile testler birbirine eşitlenmiştir. Ayrı kalibrasyon yapılarak elde edilen Form X ve Form Y' ye ait madde ve yetenek parametrelerini aynı ölçekte elde etmek amacıyla veri analizinin ikinci aşamasında Stocking-Lord ve Haebara ölçek dönüştürme yöntemleri kullanılmıştır. Ölçek dönüştürme STUIRT (Kim ve Kolen, 2004) bilgisayar programı aracılığı ile yapılmıştır. Veri analizinin üçüncü aşamasında POLYEQUATE (Kolen, 2004) bilgisayar programı kullanılarak testler birbirine eşitlenmiştir. Veri analizinin son aşamasında ise RMSE (Root Mean Squared Error; hata kareleri ortalamasının karekökü), yanlılık ve standart sapma indeksleri kullanılarak elde edilen eşitlenmiş puanlar değerlendirilmiştir. Değerlendirme indeksleri her bir puan düzeyinde hesaplanmış ve her bir koşul için tüm puanlar için hesaplanan indekslerin ortalaması alınmıştır. RMSE, yanlılık ve standart sapma (SS) indeksleri her bir puan düzeyinde eşitlik (6), (7) ve (8) görüldüğü gibi matematiksel olarak ifade edilebilir:

$$RMSE\ (x) = \left\{\frac{1}{M}\sum_{i=1}^{M}[\hat{e}_i(x) - e(x)]^2\right\}^{1/2} \quad (6)$$

$$Yanlılık\ (x) = \frac{1}{M}\sum_{i=1}^{M}[\hat{e}_i(x) - e(x)] \quad (7)$$

$$SS(x) = \left\{\frac{1}{M}\sum_{i=1}^{M}\left[\hat{e}_i(x) - \bar{\hat{e}}_i(x)\right]^2\right\}^{1/2}, \quad \bar{\hat{e}}_i(x) = \frac{1}{M}\sum_{i=1}^{M}\hat{e}_i\ (x) \quad (8)$$

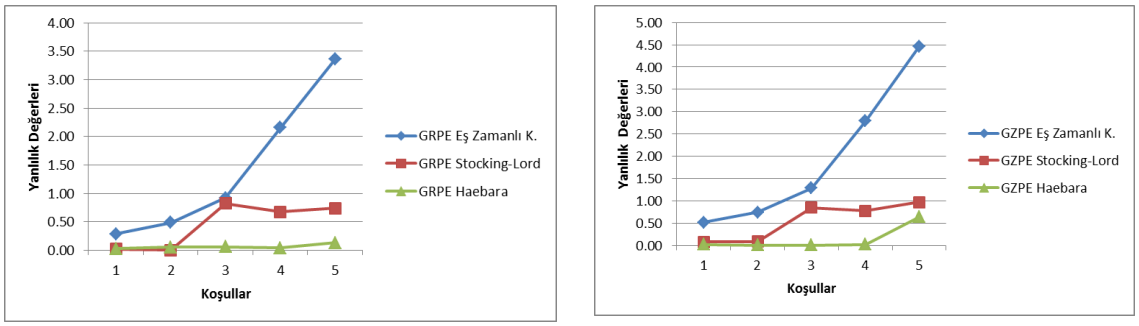
Yukarıda verilen eşitliklerde *M* tekrar sayısını göstermek üzere; $\hat{e}_i(x)$ *i.* tekrarda elde edilen *x* ham puanını Y ölçeğinin ham puanına çeviren eşitleme fonksiyonunu ve $e(x)$ ölçüt eşitleme fonksiyonu göstermektedir. Bu araştırmada ölçüt eşitleme fonksiyonu her bir puan için 50 tekrardan elde edilen eşitleme fonksiyonlarının ortalaması alınarak hesaplanmıştır. Sinharay ve Holland (2007) yanlılık ölçütünün eşitleme sonuçlarına karışan sistematik hatanın, standart sapma ölçütünün ise eşitleme sonuçlarının sahip olduğu tesadüfi hatanın bir ölçüsü olarak kullanılabileceğini belirtmiş ve bu üç değerlendirme ölçütü arasındaki ilişkiyi matematiksel olarak şu şekilde ifade etmiştir:

$$[\text{RMSE}(x)]^2=[\text{SS}(x)]^2+[\text{Yanlılık}(x)]^2$$

3. BULGULAR

Eş zamanlı kalibrasyon yapılarak ve ayrı kalibrasyon yapılıp Stocking-Lord ile Haebara ölçek dönüştürme yöntemleri kullanılarak elde edilen MTK gerçek-puan eşitleme (GRPE) ve gözlenen-puan eşitleme (GZPE) sonuçlarına ilişkin yanlılık, standart sapma ve RMSE değerlendirme ölçütlerine göre çizdirilen grafikler sırayla Şekil 1, Şekil 2 ve Şekil 3'te verilmiştir. Grafiklerde yer alan yanlılık değerleri mutlak değeri alınarak $|\text{yanlılık}|$ mutlak büyüklüğüne dayalı olarak şekillerde sunulmuştur (Sinharay ve Holland, 2007).

Eş zamanlı kalibrasyon yapılarak elde edilen GRPE ve GZPE sonuçları ile ayrı kalibrasyon yapılıp Stocking-Lord ve Haebara ölçek dönüştürme uygulanarak elde edilen eşitleme sonuçlarına ilişkin yanlılık değerlerine ait çizilen grafikler Şekil 1'de görülmektedir.

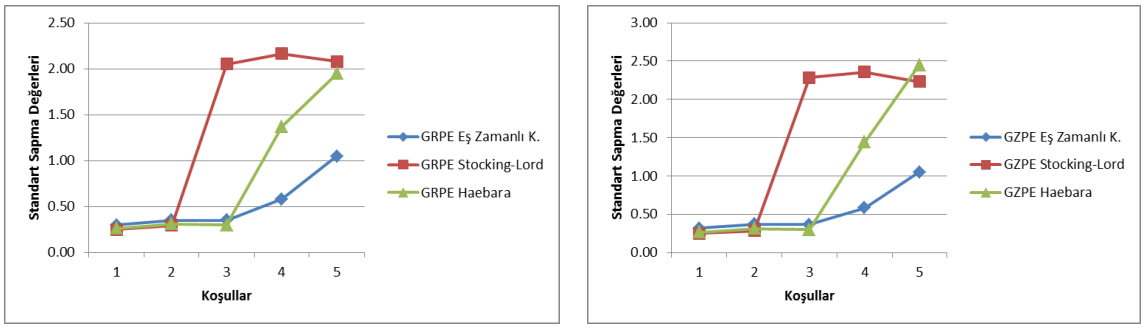


Şekil 1. Yanlılık değerlendirme ölçütüne ilişkin çizilen grafikler

Not. *GRPE: Gerçek-puan eşitleme; GZPE: Gözlenen-puan eşitleme; Koşul1:* $r_{\Theta 1\Theta 2}$*=0.90; Koşul2:*$r_{\Theta 1\Theta 2}$*=0.75; Koşul3:* $r_{\Theta 1\Theta 2}$*=0.50; Koşul4:* $r_{\Theta 1\Theta 2}$*=0.25; Koşul5: 0.00*

Şekil1'de görülen grafikler incelendiğinde eş zamanlı kalibrasyon yapılarak elde edilen GRPE ve GZPE sonuçlarında en yansız kestirimler boyutlar arasındaki korelasyonun en yüksek olduğu ($r_{\Theta 1\Theta 2}$=0.90) koşullarda elde edilirken en yanlı kestirimler boyutlar arasındaki korelasyonun sıfır olduğu koşullarda elde edilmiştir. Boyutlar arasındaki korelasyon değeri azaldıkça eş zamanlı kalibrasyon yapılarak elde GRPE ve GZPE sonuçlarının yanlılık değerleri artmış dolayısıyla daha yanlı eşitleme sonuçları elde edilmiştir. Ayrı kalibrasyon yapılıp Stocking-Lord ölçek dönüştürme yöntemi uygulanılarak elde edilen GRPE ve GZPE, eşitleme sonuçları iki boyut arasındaki korelasyonun 0.90 ve 0.75 olduğu koşullarda en düşük yanlılık değerlerine sahip olduğu boyutlar arasındaki korelasyonun 0.50'ye düşmesiyle yanlılık değerlerinde hızlı bir artış olduğu söylenebilir. Şekil 1'de görüldüğü üzere Haebara ölçek dönüştürme yöntemi uygulanarak elde edilen GRPE ve GZPE sonuçlarına ait yanlılık değerleri boyutlar arasındaki korelasyonun 0.90, 0.75, 0.50 ve 0.25 olduğu koşullarda sıfıra çok yakın değerler almış ve boyutlar arasındaki korelasyonun 0.00 olduğu koşulda artış göstermiştir. Şekil 1'de verilen grafiklere dayalı olarak tüm koşullarda en yanlı GRPE ve GZPE sonuçlarının eş zamanlı kalibrasyon yapıldığında elde edildiği, Haebara ve Stocking-Lord ölçek dönüştürme yöntemlerinin çok boyutluluğun derecesinin düşük olduğu koşullarda (Koşul 1: $r_{\Theta 1\Theta 2}$=0.90 ve Koşul 2: $r_{\Theta 1\Theta 2}$=0.75) benzer performans gösterdiği fakat çok boyutluluğunun derecesinin yüksek olduğu koşullarda (Koşul 3: $r_{\Theta 1\Theta 2}$=0.50, Koşul 4: $r_{\Theta 1\Theta 2}$=0.25 ve Koşul 5: $r_{\Theta 1\Theta 2}$=0.00) en yansız sonuçların Haebara yöntemiyle elde edildiği söylenebilir.

Eş zamanlı kalibrasyon yapılarak elde edilen GRPE ve GZPE sonuçları ile ayrı kalibrasyon yapılıp Stocking-Lord ve Haebara ölçek dönüştürme yöntemleri uygulanarak elde edilen eşitleme sonuçlarına ilişkin standart sapma değerlerine ait çizilen grafikler Şekil 2'de görülmektedir.

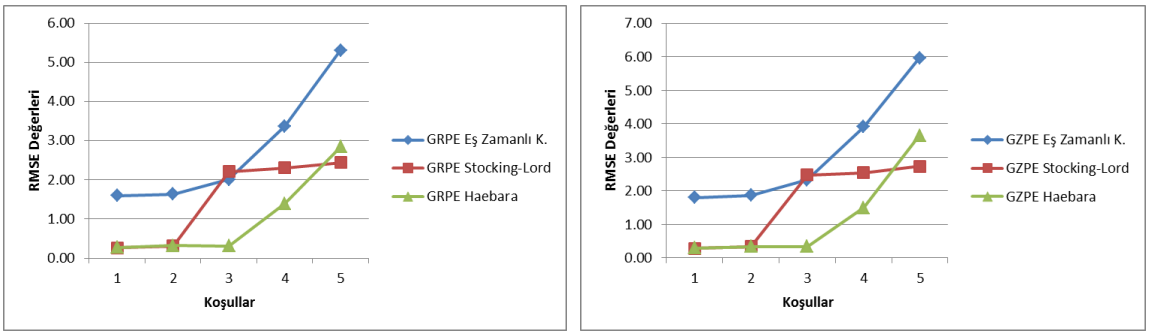


Şekil 2. Standart sapma değerlendirme ölçütüne ilişkin çizilen grafikler

Not. GRPE: Gerçek-puan eşitleme; GZPE: Gözlenen-puan eşitleme; Koşul1: $r_{\Theta 1\Theta 2}$*=0.90; Koşul2:*$r_{\Theta 1\Theta 2}$*=0.75; Koşul3:* $r_{\Theta 1\Theta 2}$*=0.50; Koşul4:* $r_{\Theta 1\Theta 2}$*=0.25; Koşul5:* $r_{\Theta 1\Theta 2}$*=0.00*

Şekil 2'de görülen grafikler incelendiğinde, GRPE ve GZPE sonuçlarının boyutlar arasındaki korelasyonun yüksek olduğu (Koşul 1: $r_{\Theta 1\Theta 2}$=0.90 ve Koşul 2: $r_{\Theta 1\Theta 2}$=0.75) bir diğer ifadeyle çok boyutluluğun derecesinin düşük olduğu koşullarda eş zamanlı kalibrasyon, Stocking-Lord ve Haebara ölçek dönüştürme yöntemleri ile elde edilen GRPE ve GZPE sonuçlarının düşük standart sapma değerlerine sahip olarak benzer performans gösterdiği söylenebilir. Daha önce de belirtildiği gibi standart sapma değerlendirme ölçütü eşitleme sonuçlarına karışan tesadüfi hatanın bir göstergesidir. Bu bulguya dayalı olarak çok boyutluluğun derecesinin düşük olduğu koşullarda üç yöntem ile elde edilen GRPE ve GZPE sonuçlarının tesadüfi hata ölçütüne göre benzer performans gösterdiği söylenebilir. Şekil 2'de verilen grafiklerde görüleceği üzere iki boyut arasındaki korelasyon değerinin 0.75'ten 0.50'ye düşmesi Stocking Lord ölçek dönüştürme yöntemi kullanılarak elde edilen GRPE ve GZPE sonuçlarının standart sapma değerlerinde hızlı bir artışa sebep olmuştur. Eş zamanlı kalibrasyon ve Haebara ölçek dönüştürme yöntemi ile elde edilen eşitleme sonuçları boyutlar arasındaki korelasyonun 0.50 olduğu koşullarda benzer ve SL yöntemine göre daha düşük standart sapma değerlerine sahip olarak daha iyi performans göstermiştir. Ancak boyutlar arasındaki korelasyonun 0.25'e ve 0.00'a düşmesi eş zamanlı kalibrasyon ve Haebara ölçek dönüştürme yapılarak elde edilen GRPE ve GZPE sonuçlarının standart sapma değerlerinde hızlı bir yükselişe sebep olmuştur ve bu artışın Haebara yöntemi ile elde edilen eşitleme sonuçlarında daha fazla olduğu görülmektedir. Çok boyutluluğun derecesinin çok ciddi olduğu koşullarda (Koşul 4: $r_{\Theta 1\Theta 2}$=0.25 ve Koşul 5: $r_{\Theta 1\Theta 2}$= 0.00) en düşük standart sapma değerine sahip GRPE ve GZPE sonuçları eş zamanlı kalibrasyon yapılarak elde edilmiştir.

Eş zamanlı kalibrasyon yapılarak elde edilen GRPE ve GZPE sonuçları ile ayrı kalibrasyon yapılıp Stocking-Lord ve Haebara ölçek dönüştürme yapılarak elde edilen eşitleme sonuçlarına ilişkin RMSE değerlerine ait çizilen grafikler Şekil 3'te görülmektedir.



Şekil 3. RMSE değerlendirme ölçütüne ilişkin çizilen grafikler

Not. GRPE: Gerçek-puan eşitleme; GZPE: Gözlenen-puan eşitleme; Koşul 1: $r_{\Theta 1\Theta 2}$=0.90; Koşul 2: $r_{\Theta 1\Theta 2}$=0.75; Koşul 3: $r_{\Theta 1\Theta 2}$=0.50; Koşul 4: $r_{\Theta 1\Theta 2}$=0.25; Koşul 5: $r_{\Theta 1\Theta 2}$=0.00

Şekil 3'te yer alan RMSE değerlendirme ölçütüne göre çizilen grafikler incelendiğinde, çok boyutluluğun derecesinin düşük olduğu koşullarda (Koşul 1: $r_{\Theta 1\Theta 2}$=0.90; Koşul 2: $r_{\Theta 1\Theta 2}$=0.75) Stocking Lord ve Haebara ölçek dönüştürme uygulanarak elde edilen GRPE ve GZPE sonuçları benzer performans göstermiş ve eş zamanlı kalibrasyon yapılarak elde edilen eşitleme sonuçlarından daha düşük RMSE değerlerine sahip olmuştur. Boyutlar arasındaki korelasyonun 0.75'ten 0.50'ye düşmesi Stocking-Lord yöntemi ile elde edilen GRPE ve GZPE sonuçlarının RMSE değerlerinde hızlı bir artışa neden olmuş, Koşul 3'te ($r_{\Theta 1\Theta 2}$=0.50) en düşük RMSE değerleri Haebara yöntemiyle elde edilmiştir. Çok boyutluluğun derecesinin çok ciddi olduğu koşullarda (Koşul4: $r_{\Theta 1\Theta 2}$=0.25; Koşul5: $r_{\Theta 1\Theta 2}$=0.00) en yüksek RMSE değerleri eş zamanlı kalibrasyon sonucu elde edilen GRPE ve GZPE sonuçlarına ait olmuştur. Koşul 4 ve Koşul 5'te Haebara ve Stocking-Lord yöntemleri eş zamanlı kalibrasyondan daha iyi performans göstermekle birlikte Koşul 4'te Haebara yöntemiyle elde edilen GRPE ve GZPE sonuçları SL

yöntemiyle elde edilenlerden daha düşük RMSE değerlerine sahip olurken Koşul 5'te, çok boyutluluğun derecesinin en yüksek olduğu koşulda, en iyi performans gösteren yöntem Stocking-Lord olmuştur.

4. TARTIŞMA ve SONUÇ

Bu araştırmada çok boyutluluğun eş zamanlı ve ayrı kalibrasyon yapılıp karakteristik eğri ölçek dönüştürme yöntemleri (Stocking-Lord ve Haebara) uygulanarak elde edilen MTK gerçek-puan eşitleme ve MTK gözlenen-puan eşitleme sonuçlarına etkisi incelenmiştir.

Araştırma bulguları yanlılık ve RMSE değerlendirme indeksi ölçütlerine göre çok boyutluluğun derecesinin düşük olduğu koşullarda ($r_{\Theta 1\Theta 2}$=0.90 ve $r_{\Theta 1\Theta 2}$=0.75) en yanlı ve en yüksek eşitleme hatasına sahip gerçek puan ve gözlenen puan eşitleme sonuçlarının eş zamanlı kalibrasyon ile elde edildiğini göstermiştir. Ancak iki boyut arasındaki korelasyonun 0.50'ye düştüğü koşullarda Stocking-Lord yöntemiyle elde edilen eşitleme sonuçlarının sahip olduğu yanlılık değerleri ve eşitleme hataları artarak eş zamanlı kalibrasyon ile elde edilen hata değerlerine yaklaşmıştır. Benzer şekilde Beguin, Hanson ve Glass (2000) çok boyutluluğun ayrı ve eş zamanlı kalibrasyon üzerine etkisini inceledikleri araştırmalarında eşdeğer olmayan gruplar deseninde boyutlar arasındaki korelasyonun 0.90 ve 0.70 olduğu koşullarda eş zamanlı kalibrasyon yapılarak elde edilen eşitlenmiş puanların yanlılık ve eşitleme hatalarının ayrı kalibrasyon yapılıp Stocking-Lord ölçek dönüştürme yöntemi uygulanarak elde edilenlere göre daha yüksek olduğu fakat iki yetenek arasındaki korelasyonun 0.50 olduğu koşullarda eş zamanlı kalibrasyon yapılarak elde edilen eşitleme sonuçlarının daha düşük yanlılık ve eşitleme hatasına sahip olduğu bulgusunu elde etmişlerdir.

Livingston, Dorans ve Wright (1989) yanlılık değerlendirme ölçütünün eşitlenmiş puanların sistematik olarak çok yüksek ya da çok düşük olma eğiliminin bir ölçüsünü verdiğini negatif yanlılık değerlerinin pozitif yanlılık değerlerini elediği için eşitlenmiş puanlar çok yüksek ya da çok düşük olmadığı sürece yanlılık istatistiğinin iyi bir değerlendirme ölçütü olmadığını belirtmiştir. Ancak yanlılık istatistiğinin, büyük RMSE istatistiği değerlerinin nedenini belirlemek için her zaman değerli bir istatistik olduğunu vurgulamışlardır. Araştırmamızda veri setinde oluşan çok boyutluluğun derecesinin en yüksek olduğu Koşul 5'te ($r_{\Theta 1\Theta 2}$=0.00) özellikle eş zamanlı kalibrasyon yapılarak elde edilen yanlılık değerlerinin gerçek puan eşitlemede 3.5 değerine gözlenen puan eşitleme 4.5 değerine ulaştığı görülmektedir. Bu bulgu çok boyutluluğun çok ciddi olduğu koşullarda Livingston ve diğerlerinin (1989) de belirttiği gibi eşitlenmiş puanların sistematik olarak çok yüksek ya da çok düşük kestirildiğinin bir göstergesidir. Ayrıca Şekil1'deki yanlılık değerlerine dayalı olarak çizdirilen grafikler ile Şekil 3'te yer alan RMSE değerlerine ait çizdirilen grafikler benzer örüntü göstermiş yüksek yanlılık değerlerine sahip koşullarda yüksek RMSE değerleri elde edilmiştir.

Araştırmanın bir diğer bulgusu çok boyutluluğun derecesinin çok ciddi olduğu ($r_{\Theta 1\Theta 2}$=0.50, $r_{\Theta 1\Theta 2}$=0.25 ve $r_{\Theta 1\Theta 2}$=0.00) koşullarda yine en yanlı ve en yüksek eşitleme hatasına sahip kestirimlerin eş zamanlı kalibrasyon yapılarak elde edilmesidir. Beguin ve Hanson (2001) ise tam tersi olarak araştırmalarında eş değer olmayan gruplar deseninde iki yetenek arasındaki korelasyonun 0.00, 0.30 ve 0.50 olduğu koşullarda eş zamanlı kalibrasyon ile elde edilen eşitleme sonuçlarının daha düşük yanlılık ve eşitleme hatasına sahip olduğu sonucuna ulaşmışlardır. Araştırmamız bulguları ile Beguin ve Hanson'ın (2001) bulguları arasında fark olmasının en önemli sebeplerinden biri olarak veri üretmede esas alınan çok boyutlu MTK modeli gösterilebilir. Beguin ve Hanson (2001) araştırmalarında çok boyutlu veriyi telafi edici olmayan modelleri dikkate alarak üretirken bu araştırmada çok boyutlu veriler telafi edici modeller dikkate alınarak üretilmiştir.

Araştırmada kullanılan bir diğer değerlendirme ölçütü standart sapmadır. Standart sapma ölçütü, eşitleme sonuçlarına karışan tesadüfî hatanın bir göstergesidir. Kolen ve Brennan'ın

(2004) da belirttiği üzere eşitleme sonuçlarına tesadüfî hatanın karışması örneklemden ve evrenden elde edilen eşitleme ilişkisinin farklılaştığı anlamına gelmektedir. Araştırma bulguları, standart sapma değerlendirme ölçütüne göre boyutlar arasındaki korelasyonun 0.90 ve 0.75 olduğu koşullarda eş zamanlı kalibrasyon ve ayrı kalibrasyon yapılarak elde edilen gerçek puan ve gözlenen puan eşitleme sonuçlarının benzer performansa sahip olduğunu göstermiştir. Fakat iki boyut arasındaki korelasyonun 0.50 olduğu koşullarda Stocking-Lord yöntemi ile elde edilen eşitlenmiş puanlar Haebara ve eş zamanlı kalibrasyon ile elde edilen eşitlenmiş puanlardan daha yüksek standart sapma değerlerine sahip olmuştur. Beguin ve diğerleri (2000) araştırmalarında iki yetenek arasındaki korelasyonun 0.90 ve 0.70 olduğu koşullarda ayrı kalibrasyon yapılarak elde edilen eşitlenmiş puanların eş zamanlı kalibrasyon ile karşılaştırıldığında daha az tesadüfi hataya sahip olduğunu fakat korelasyon değerinin 0.50 olduğu koşulda tıpkı bizim araştırmamızda olduğu gibi eş zamanlı kalibrasyonun Stocking-Lord ölçek dönüştürme yönteminden daha iyi performans gösterdiği sonucuna ulaşmışlardır.

Bu araştırmanın en çarpıcı bulgusu, Haebara yönteminin genel olarak çok boyutluluğun derecesinin yüksek olduğu koşullarda ($r_{\Theta 1\Theta 2}$=0.50 ve $r_{\Theta 1\Theta 2}$=0.25) Stocking-Lord yönteminden daha iyi performans göstermesidir. Literatürde eş zamanlı kalibrasyon ile ayrı kalibrasyonun karşılaştırıldığı araştırmalarda genellikle test karakteristik eğrisi ölçek dönüştürme yöntemlerinden sadece Stocking-Lord yöntemi tercih edilerek araştırmalar yürütülmekte, eş zamanlı kalibrasyon ile karşılaştırılınca MTK'nın tek boyutluluk varsayımının ihlaline daha dayanıklı olduğu rapor edilerek önerilmektedir (Kim, 2007). Bu araştırmanın yanlılık ve eşitleme hatası bulgularına dayalı olarak çok boyutluluğun derecesinin düşük olduğu koşullarda ayrı kalibrasyon yapılıp Haebara ya da Stocking-Lord yöntemlerinden biri seçilip ölçek dönüştürmenin yapılması, çok boyutluluğun derecesinin yüksek olduğu durumlarda ise Haebara yönteminin kullanılarak ölçek dönüştürmenin yapılması önerilebilir. Gelecekte yapılacak olan araştırmalarda Haebara ile Stocking-Lord yöntemlerinin performansının farklı koşullar altında karşılaştırılması yararlı olacaktır.

Çok boyutluluğun genel olarak eş zamanlı ve ayrı kalibrasyon yapılarak elde edilen eşitlenmiş puanları etkilediği (Hanson & Beguin, 2001; Hanson ve diğerleri, 2000) sonucunu bizim araştırmamız da desteklemektedir. Kolen ve Brennan (2004) MTK'nın varsayımlarının ihlaline ayrı kalibrasyon yapılıp karakteristik eğri ölçek dönüştürme yöntemleri uygulanarak elde edilen kestirimlerin eş zamanlı kalibrasyonla karşılaştırıldığında daha dayanıklı olduğu ve daha doğru sonuçlar vereceğini belirtmişlerdir. Ayrıca, uygulamada ayrı kalibrasyonun daha güvenilir olduğunu vurgulamışlardır. Araştırmamızın bulguları yanlılık ve RMSE değerlendirme ölçütleri temel alındığında Kolen ve Brennan'ın (2004) önerisini destekler niteliktedir.

Bu araştırma simülasyon verileri, iki kategoride puanlanan madde cevapları ve çok boyutluluğun testlerin sadece iki yeteneği ölçmesi ile sınırlıdır. İleride yapılacak olan araştırmalarda çok boyutluluk testlerin ikiden fazla yeteneğini ölçtüğü varsayılıp oluşturularak eş zamanlı ve ayrı kalibrasyon yöntemleri ile test eşitlemeye çok boyutluluğun etkisi incelenebilir. Karma testlerin eşitlenmesinde eş zamanlı ve ayrı kalibrasyon yöntemleri karşılaştırılabilir. Simülasyon verileri her ne kadar araştırmacılara aynı anda bir çok faktörü çalışma imkanı vermesi açısından değerli olsa (Harris ve Crouse, 1993) da gerçek test durumlarını yansıtmayabilir. Dolayısıyla bu araştırma kapsamında ele alınan ve önerilen tüm koşulların gerçek veri seti kullanılarak sınanmasında fayda vardır.

5. KAYNAKLAR

Ackerman, T. A. (1989). Unidimensional IRT calibration of compensatory and noncompensatory items. *Applied Psychological Measurement, 13,* 113-127. [Available online at: https://conservancy.umn.edu/bitstream/handle/11299/107494/v13n2p113.pdf?sequence=1&isAllowed=y], Retrieved on December 12, 2017.

Ackerman, T. A. (1994). Creating a test information profile for a two-dimensional latent space. *Applied Psychological Measurement, 18(3),* 257-275. [Available online at: https://conservancy.umn.edu/bitstream/handle/11299/117004/v18n3p257.pdf;sequence=1], Retrieved on December 10, 2015.

Ackerman, T. A. (1996). Graphical representation of multidimensional item response theory analyses. *Applied Psychological Measurement, 20*(4), 311-329. [Available online at: https://conservancy.umn.edu/bitstream/handle/11299/119465/v20n4p311.pdf?sequence=1&isAllowed=y], Retrieved on December 12, 2015.

Albayrak-Sarı, A. & Kelecioğlu, H. (2017). A comparison of IRT vertical scaling methods determining the increase in science achievement. *Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergis*i, *8*(1), 98-111. [Çevrim-içi: http://dergipark.gov.tr/epod/issue/28110/286221], Erişim tarihi: 11 Ekim 2018.

Altun, A. & Kelecioğlu, H. (2016). Dikey ölçeklemede madde tepki kuramına dayalı kalibrasyon ve yetenek kestirim yöntemlerinin karşılaştırılması. *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi, 31*(3), 447-460. [Çevrim-içi: http://www.efdergi.hacettepe.edu.tr/yonetim/icerik/makaleler/2171-published.pdf], Erişim tarihi: 3 Ocak 2018.

Beguin, A. A., & Hanson, B. A. (2001, April). *Effect of noncompensatory multidimensionality on separate and concurrent estimation in IRT observed score equating*. Paper presented at the The Annual Meeting of the National Council on Measurement in Education, Seattle, WA.

Béguin, A. A., Hanson, B. A., & Glas, C.A.W. (2000, April). *Effect of multidimensionality on separate and concurrent estimation in IRT equating.* Paper presented at the Annual Meeting of the National Council on Measurement in Education, New Orleans, LA. [Available online at: http://www.bh.com/papers/paper0002.html], Retrieved on December 28, 2015.

Hanson, B. A., & Beguin, A.A. (2002). Obtaining a common scale for item response theory item parameters using separate versus concurrent estimation in the common-item equating design. *Applied Psychological Measurement, 26*(1), 3-24. [Available online at: https://journals.sagepub.com/doi/10.1177/0146621602026001001 , Retrieved on December 28, 2015.

Huggins, A.C. (2012). *The effect of differential item functioning on population invariance of item response theory true score equating.* Unpublished doctoral dissertation, University of Miami.

Kang, T., & Petersen, N. (2009). *Linking item parameters to a base scale*. ACT Research Report Series 2009-2, Iowa City, IA: ACT, Inc.

Kim, S. (2004). *Unidimensional IRT scale linking procedures for mixed-format tests and their robustness to multidimensionality*. Unpublished doctoral dissertation, University of Iowa, Iowa City.

Kim, S., & Cohen, A. S. (1998). A comparison of link-ing and concurrent calibration under item response theory. *Applied Psychological Measurement, 22*(2),131-143. [Available online at: https://journals.sagepub.com/doi/10.1177/01466216980222003], Retrieved on December 28, 2015.

Kim, S., & Kolen, M. J. (2004). STUIRT: A computer program for scale transformation under unidimensional item response theory models [Computer Software]. Iowa City: IA: The Center for Advanced Studies in Measurement and Assessment, The University of Iowa.

Kolen, M.J. (2004). POLYEQUATE [computer program].Iowa City,IA: The Center for Advanced Studies in Measurement and Assessment (CASMA), The University of Iowa.

Kolen, M. J., & Brennan, R. L. (1995).Test equating methods and practices. New York: Springer-Verlag.

Kolen, M.J., & Brennan, R.L. (2004). Test equating: Methods and practices (2nd ed.).New York, NY: Springe-Verlag.

Min, K-S. (2007). Evaluation of linking methods for multidimensional IRT calibrations. *Asia Pacific Education Review, 8*(1), 41-45. [Available online at: https://link.springer.com/article/10.1007/BF03025832] Retrieved on December 12, 2017.

Petersen, N. S., Cook, L. L., & Stocking, M. L.(1983). IRT versus conventional equating methods: A comparative study of scale stability. *Journal of Educational Statistics, 8*(2), 137-156. [Available online at: https://www.jstor.org/stable/1164922?seq=1/analyze] Retrieved on March 12, 2014.

Reckase, M. D. (2009). *Multidimensional item response theory*. New York: Springer.

Reckase, M. D., Ackerman, T. A., & Carlson, J. E. (1988). Building a unidimensional test using multidimensional items. *Journal of Educational Measurement,* 25(3), 193-203. [Available online at: https://www.jstor.org/stable/1434499?seq=1#page_scan_tab_contents] Retrieved on October 18, 2018.

Revelle, W. (2018). *Procedures for psychological, psychometric, and personality*. R package version 1.8.4.

Sinharay, S., & Holland, P. W. (2007). Is it necessary to make anchor tests mini-versions of the tests being equated or can some restrictions be relaxed? *Journal of Educational Measurement*, *44*(3), 249-275. [Available online at: https://www.jstor.org/stable/20461859?seq=1#page_scan_tab_contents] Retrieved on October 3, 2015.

Stocking, M. L., & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement, 7*(2), 201-210.

Stout, W. (1987). A nonparametric approach for assessing latent trait unidimensionality. *Psychometrika*, *52*(4), 589-617.

Wingersky, M. S., Cook, L. L., & Eignor, D. R.(1987). *Specifying the characteristics of linking items used for item response theory item calibration.* ETS Research Report 87-24. Princeton, NJ: Educational Testing Service.

Yao, L. (2003). SimuMIRT [Computer Software]. Monterey, CA: Defense Manpower Data Center.

Yao, L., & Schwarz, R.D. (2006). A multidimensional partial credit model with associated item and test statistics: An application to mixed-format tests. *Applied Psychological Measurement, 30*, 469-492.

Zhang, B. (2009). Application of unidimensional item response models to tests with item sensitive to secondary dimensions. *The Journal of Experimental Education*, *77*(2), 147-166.

Zimowski, M. F., Muraki, E., Mislevy, R. J., & Bock, R. D. (1996). BILOG-MG: Multiple group IRT analysis and test maintenance for binary items [Computer program]. Chicago: Scientific Software International.

Extended Abstract

The use of different test forms of the same test which measures the same construct and content leads test equating reality to compare test scores from different forms. Test equating is the name of process that is used to adjust scores on test forms and end of this process that scores on different forms can be used interchangeably (Kolen & Brennan, 2004). Based on classical test theory and item response theory (IRT), many equating methods have been developed. Equating test forms via item response theory equating methods include generally three steps: item calibration or estimating item parameters, scale transformation and test equating. Item parameters can be calibrated through separate or concurrent calibration.

While the item parameters for each test forms are estimated using separate computer runs in separate calibration, in concurrent calibration item parameters for both forms are estimated simultaneously in one computer run. Estimating parameters for all items simultaneously assures that all parameter estimates are on the same scale. In the common-item nonequivalent groups design, different tests with same common items are administered to samples from different populations. If item parameters for each form estimated by separate calibration, item parameter estimates from two groups will not be on the same scale. As Hanson and Beguin (2002) indicated that constraining the scale of the latent variable by fixing the mean and standard deviation of the latent variable distribution in the common-item nonequivalent groups design yields item parameters in different scales. As Kolen and Brennan (2004) emphasized parameter estimates must be on the same scale to conduct test equating. There are four scale transformation methods which commonly used to put item parameter estimates of one form on the scale of the item parameter estimates for the other form: the mean/sigma (Marco, 1977), mean/mean (Loyd &Hoover, 1980), Haebara (Haebara, 1980) and Stocking-Lord (Stocking & Lord, 1983) methods (as cited in Kolen and Brennan, 2004).

A number of studies have been carried out to compare the performance of concurrent and separate calibration (Albayrak-Sarı & Kelecioğlu, 2017; Altun & Kelecioğlu, 2016; Hanson & Beguin, 1999; Hanson & Beguin, 2002; Kang & Petersen, 2009; Kim & Cohen, 1998; Petersen, Cook & Stocking, 1983; Wingersky, Cook & Eignor, 1987). The common properties of these researches are separate and concurrent calibration methods were tested by using unidimensional data. In reality, the unidimensionality assumption of IRT may not be maintained and multidimensionality may affect separate and concurrent calibration results in test equating. Little research (Beguin & Hanson, 2001; Beguin, Hanson & Glass, 2000) has been done comparing the concurrent and separate estimation methods in tests equating with dichotomously scored tests. The aim of this study is investigating the effects of multidimensionality on equating results which obtained from separate and concurrent calibration methods.

The study was conducted with using simulated data. Data was simulated based on item parameters from Hanson and Beguin (2002) study. Test equating was conducted under common item nonequivalent groups design. Item responses were simulated for Form X and Form Y. There were 20 common items for each form and totally each form was comprised of 60 multiple choice items. In the scope of research, totally 40 simulation conditions [5 (degree of multidimensionality: 0.90, 0.75, 0.50, 0.25, and 0.00) x 2 calibration methods (separate and concurrent) x 2 (scale transformation methods: Stocking-Lord and Haebara) x 2 (test equating methods)] were examined. For each condition 50 samples of both forms were generated with 3000 persons per form. Multidimensionality was constructed as assuming the two test forms were measuring Θ_1 and Θ_2 abilities. While the simulation condition which has correlation between abilities 0.90 represents the weak multidimensional case, the correlation between abilities 0.00 represents the severe multidimensional case. BILOG-MG program was used to estimate item parameters via separate and concurrent calibration methods. In the separate estimation conditions, the parameters of Form X and Form Y were brought on to a common scale by Stocking-Lord and Haebara scale transformation methods via STUIRT computer program. In the last step of data analyses, test scores from Form X were equated to Form Y test scores with IRT true-score equating and observed score equating methods by using POLYEQUTE computer program. And finally equating results were evaluated by using Bias, standard deviation and RMSE evaluation criteria.

The results showed that, generally under all conditions equating results provided from concurrent calibration more biased and had higher RMSE values than equating results provided by separate calibration. Based on standard deviation criteria, when the degree of multidimensionality was low, equating results which got from concurrent calibration and separate calibration with Stocking-Lord or Haebara scale transformation methods showed similar performance but when the degree of multidimensionality was severe ($r_{\Theta 1 \Theta 2}$=0.25 ve 0.00) equating results which had lowest random error were provided by concurrent calibration.

EK 1: BILOG-MG Eş Zamanlı ve Ayrı Kalibrasyon Kodları

Eş Zamanlı Kalibrasyon

```
>COMMENTS
concurrent calibration for form Y and form X
>GLOBAL DFName='simyx1.prn',NPArm=3,LOGistic,SAVE;
>SAVE PARM='simyx1.PAR';
>LENGTH NITEMS=(100);
>INPUT NTOtal = 100, NIDchar = 6, NGRoup = 2, NFOrm = 2, KFName = 'KEY.txt';
>ITEMS INUM = (1(1)100);
>TEST TNAme = 'EN', INUmber = (1(1)100);
>FORM1 LENgth = 60, INUmbers = (1(1)60);
>FORM2 LENgth = 60, INUmbers = (41(1)100);
>GROUP1 GNAme = 'Y', LENgth = 60, INUmbers = (1(1)60);
>GROUP2 GNAme = 'X', LENgth = 60, INUmbers = (41(1)100);
(6A1,T1,I1,T1,I1,T8,60A1)
>CALIB CYCles = 40, NEWton = 20, TPRior,NORMAL, REFERENCE = 1;
```

Ayrı Kalibrasyon

```
>COMMENTS
separate calibration for form X
>GLOBAL DFName='sim1.dat',NPArm=3,LOGistic,SAVE;
>SAVE PARM='sim1.PAR';
>LENGTH NITEMS=(60);
>INPUT NTOtal = 60, NIDchar = 4;
>ITEMS INUM = (1(1)60);
>TEST TNAme = 'EN', INUmber = (1(1)60);
(4A1,60A1)
>CALIB CYCles = 40, NEWton = 20, NQPT=30, TPRior,NORMAL;
```