Gül, E., Doğan-Gül, Ç., Çokluk-Bökeoğlu, Ö. ve Özkan, M. (2017). Temel Eğitimden Ortaöğretime Geçiş Matematik Alt Testi Asıl Sınav ve Mazeret Sınavlarının Madde Tepki Kuramına Göre Eşitlenmesi**,** *Abant İzzet Baysal Üniversitesi Eğitim Fakültesi Dergisi, 17* (4), 1900-1915.
Geliş Tarihi: 07/06/2017 Kabul Tarihi: 09/10/2017

TEMEL EĞİTİMDEN ORTAÖĞRETİME GEÇİŞ MATEMATİK ALT TESTİ ASIL SINAV VE MAZERET SINAVLARININ MADDE TEPKİ KURAMINA GÖRE EŞİTLENMESİ*

Emrah GÜL**
Çilem DOĞAN GÜL***
Ömay ÇOKLUK BÖKEOĞLU****
Mustafa ÖZKAN*****

ÖZET

Bu araştırmanın amacını TEOG asıl sınav matematik alt testi ile TEOG mazeret sınavı matematik alt testlerinin Madde Tepki Kuramı'na (MTK) dayalı yöntemler kullanılarak eşitlenmesi ve hangi eşitleme yönteminin daha uygun olduğuna karar verilmesi oluşturmaktadır. Araştırmanın temel araştırma niteliği taşıdığı ifade edilebilir. Testlerin eşitlenmesi için, madde ve birey parametreleri flexMIRT programı kullanılarak kestirilmiştir. Eşitleme işlemi R "equateIRT" paketi kullanılarak yapılmıştır. Araştırmadan elde edilen sonuçlar, en düşük eşitleme hatasının karakteristik eğri yöntemlerinden elde edildiğini ortaya koymaktadır. Bu yöntemlerden de en düşük eşitleme hatasının Stocking-Lord yönteminden elde edildiği belirtilebilir. Ortalama-Ortalama yöntemi ise en yüksek eşitleme hatası üreten yöntemdir. Bu araştırmadan elde edilen bulgular Stocking-Lord yönteminin TEOG asıl sınav ve mazeret sınavı kapsamında yer alan matematik alt testlerinin eşitlenmesinde daha uygun olduğunu göstermiştir.

Anahtar Kelimeler: TEOG, Test Eşitleme, Madde Tepki Kuramı, Mazeret Sınavı

EQUATION OF MATHS SUBTESTS IN THE TRANSITION FROM PRIMARY TO SECONDARY EDUCATION (TEOG) AND MAKEUP EXAMINATION ACCORDING TO ITEM RESPONSE THEORY

ABSTRACT

The purpose of the study is to equate maths subtests of TEOG main exam and make-up examination by using Item Response Theory-based methods and to decide which equation method is appropriate. It could be suggested that it is a fundamental research. For test equalization, item-person parameters were estimated, using flexMIRT programme. R "equateIRT" pack was employed for the equation. The research results have shown that the minimum number of equation errors is obtained from characteristic curve methods. It could also be suggested that the Stocking-Lord method gives the lowest equation error number. The mean-mean method produces the highest equation error number. The findings of the study have shown that, when compared to the others, the Stocking-Lord method is more appropriate to equate the maths subtests of TEOG main exam and make-up examinations.

Key Words: TEOG, Test Equating, Item Response Theory, Make up Exam

---

* *Bu makale V. Eğitimde ve Psikolojide Ölçme ve Değerlendirme Kongresi'nde sözlü bildiri olarak sunulmuştur.*
** *Yrd. Doç. Dr., Hakkari Üniversitesi, Eğitim Fakültesi, Hakkari-Türkiye, e-posta: emrahgul@hakkari.edu.tr*
*** *Arş. Gör., Hakkari Üniversitesi, Eğitim Fakültesi, Hakkari-Türkiye, e-posta: cilemdogangül@hakkari.edu.tr*
**** *Doç. Dr., Ankara Üniversitesi, Eğitim Bilimleri Fak., Ankara-Türkiye, e-posta: cokluk@education.ankara.edu.tr*
***** *Uzm., MEB, Hayat Boyu Öğrenme Genel Müdürlüğü, Ankara-Türkiye e-posta: mustafaozkan@meb.gov.tr*

1.GİRİŞ

Yüksek risk içeren testler, bireyler için önemli sonuçları olan ve eğitimin her kademesinde öğrenciler, öğretmenler, okullar veya bölgeler hakkında alınan kararların belirleyicisi olarak kullanılan ölçme araçlarıdır. Bu testlere üniversiteye giriş, bir konuda yetkinlik kazanma veya akademik birtakım kararlar alma gibi amaçlarla ihtiyaç duyulur (Cizek, 2001; Resnick, 2004). Testlerin karar vermede bu denli belirleyici bir rol oynuyor olması, bireylerin belirli standartlara uygun olarak, eşit şartlarda test edilmesi sorununa ilginin, her geçen gün daha da artmasına neden olmaktadır. Bu durum ölçme ve değerlendirme alanında yapılan çalışmalarda da dikkat çekmektedir.

Farklı eğitim kurumlarına yerleşebilmek ya da kademeler arası geçişler için Türkiye'de dönem dönem değişen farklı sınav uygulamaları yapıldığı görülmektedir. Bireylerin bu anlamda ilk yol ayrımları, temel eğitimden ortaöğretime geçiş aşamasında olmaktadır. Türkiye'de temel eğitimden ortaöğretime geçiş kapsamında geçmişten bugüne LGS (Liselere Giriş Sınavı), OKS (Ortaöğretim Kurumları Sınavı) gibi farklı uygulamalar yapılmıştır. 2013-2014 eğitim-öğretim yılından itibaren ise Temel Eğitimden Ortaöğretime Geçiş (TEOG) sistemi kullanılmaya başlanmış ve bu kapsamda altı temel derse (Türkçe, Matematik, Fen ve Teknoloji, T.C. İnkılap Tarihi ve Atatürkçülük, İngilizce, Din Kültürü ve Ahlâk Bilgisi) ilişkin alt testler şeklinde uygulanan Merkezi Ortak Sınavlar (MOS) aracılığıyla, ortaokul 8. sınıf öğrencilerinin ortaöğretim kurumlarına yerleştirilmeleri sağlanmıştır. Ayrıca bu sistem ile ilk kez, herhangi bir probleminden dolayı (hastalık, mazeret vb.) sınava giremeyen öğrencilerin hak kaybına uğramalarını önlemek için, farklı sorulardan oluşan ancak aynı kapsama yönelik olduğu belirtilen bir "Mazeret Sınavı" uygulaması hayata geçirilmiştir. Bu uygulama ile mazereti olan öğrenciler asıl sınavdan 14 (on dört) gün sonra sınava girebilme hakkı elde etmişlerdir.

Temel Eğitimden Ortaöğretime Geçiş Sistemi kapsamında uygulanan testler, yukarıda da değinildiği gibi yüksek risk içeren testler olduğundan, doğal olarak öğrencilerde yüksek kaygı uyandırma potansiyeline sahiptir (Casbarro, 2004). Her yıl tekrarlanan ve aynı kapsamı ölçen bu testlerin psikometrik olarak eşitlenmesi; bu testlerden alınan puanların birbirleriyle karşılaştırılabilmesi, aynı ölçek üzerine yerleştirilerek bireyin başarı sırası ile ilgili daha fazla bilgi sahibi olunabilmesi vb. açılardan oldukça önemlidir. Test geliştiriciler, psikometrik açıdan ya da ölçülen özellik açısından benzer nitelikte testler oluşturmak istemelerine rağmen, her sınavda farklı soruların sorulması, testlerin psikometrik özelliklerinde de ister istemez farklılaşmalara neden olmaktadır (Tanguma, 2000).

Başta geniş ölçekli testler olmak üzere, tüm testlerin uygulanması sırasında yüksek düzeyde güvenlik ve gizlilik önlemlerinin alınması gerekir. Bu tür geniş ölçekli test uygulamalarında yasal, psikometrik ve pratik bir takım faktörlerin göz önünde bulundurulması gerekir. Bu anlamda üzerinde durulması gereken ilk nokta, bir testin aynı özelliği ölçen farklı formlarını oluşturmaktır. İki farklı oturumda aynı testin farklı ya da aynı bireylere uygulanması, testi sonra alanların, daha önce alanlara göre kesin bir şekilde daha avantajlı olmasını sağlayacak ve testin geçerliğini ve güvenirliğini tehdit edecektir. Ancak iki testin psikometrik açıdan birbirine eşdeğer olduğu kanıtlanırsa, aynı testin farklı formlarının, farklı bireylere uygulanması bile, bilimsel olarak savunulabilir bir durum haline gelir. Bütün bunların yanında test geliştiricilerin veya kurumların tamamen birbirine eşdeğer, fakat farklı sorulardan oluşan ve testi alan her bir birey için aynı ya da

benzer sonuçlar üreten testler yapılandırmaları oldukça zordur; hatta böyle bir beklenti gerçekçi de değildir. İşte tam da bu nokta "test eşitleme" sorununu gündeme getirmektedir. Test eşitleme kavramı, testin bir formunun, diğer formundaki karşılığı ve dönüşümü olarak tanımlanmaktadır (Angoff, 1981). Braun ve Holland (1982) ise test eşitlemeyi, güçlük düzeyi farklılık gösteren formlardan elde edilen puanların birbirlerinin yerine kullanılabilmesini sağlamak amacıyla yapılan sayısal dönüştürmeler / düzeltmeler olarak tanımlamaktadır.

Test eşitleme işlemleri, kuramsal olarak Klasik Test Kuramı (KTK) ve Madde Tepki Kuramı (MTK) temel alınarak yapılabilir. KTK'da kestirilen madde ve birey parametrelerinin testi alan gruba bağımlı olması durumu, aynı zamanda testlerin eşitlenmesi sürecinde de sorunlara neden olmaktadır. MTK ise görece karşılanması güç bazı varsayımlara sahip olmasına karşın, testlerin aynı ölçek üzerinde gösterilmesine olanak vermesi açısından KTK'ya göre daha avantajlıdır. Test eşitleme sürecine ilişkin olarak hem KTK, hem de MTK'ya dayalı farklı yöntemler geliştirilmiştir. KTK'ya dayalı yöntemleri; ortalama eşitleme, doğrusal eşitleme ve eşit yüzdelikli eşitleme olarak sıralamak mümkündür. Ancak bu çalışmada MTK temelli eşitleme yöntemleri kullanıldığından, aşağıda bu kuram temelli test eşitleme mantığı ve yöntemleri hakkında bilgi verilmeye çalışılmıştır.

Klasik Test Kuramı'na karşı ilk olarak 1930'lu yıllarda ortaya atılmış olan MTK ile ilgili çalışmalar 1950'lerde hız kazanmış, günümüzde de test geliştirme, soru bankası oluşturma, bireye uyarlanmış test, seçenek ağırlıklandırma, madde yanlılığı belirleme, test eşitleme gibi birçok konuda kullanılmakta ve bu anlamda pek çok soruna çözüm oluşturduğu ifade edilmektedir (Hambelton ve Swaminathan, 1985).

Madde Tepki Kuramı'nda test eşitleme işlemlerine yönelik olarak, öncelikle eşitleme aşamalarından söz etmekte yarar vardır. MTK'da eşitleme işlemleri; eşitleme deseninin seçilmesi, parametrelerin ortak bir ölçeğe yerleştirilmesi ve test puanlarının eşitlenmesi olarak tanımlanabilir (Loyd ve Hoover, 1980 ve Kolen ve Brennan, 2004). MTK'da testler, madde parametrelerinden elde edilen A ve B eğim katsayıları kullanılarak psikometrik olarak ortak bir ölçeğe yerleştirilir (Eşitlik 1) ve madde parametreleri de yine aynı eğim katsayıları kullanılarak dönüştürülebilir.

$$\theta^{*} = A\theta \mp B \qquad \text{(Eşitlik 1)}$$

Madde Tepki Kuramı'na dayalı eşitleme yöntemleri; "moment yöntemleri" ve "karakteristik eğri yöntemleri" olmak üzere iki temel grupta incelenebilir. Moment yöntemleri olarak adlandırılan birinci grupta; "Ortalama-Ortalama (Ort-Ort)" ve "Ortalama-Sigma (Ort-Sig)" olmak üzere iki yöntem yer almaktadır. Ort-Ort yönteminde güçlük ve ayırt edicilik parametre değerlerinin ortalaması ölçek puanlarının dönüşümünde kullanılmaktadır. A ve B katsayıları Eşitlik 2 ve Eşitlik 3'de görüldüğü gibi hesaplanmaktadır.

$$A = \frac{\mathrm{Ort}(a_i)}{\mathrm{Ort}\,(a_j)} \qquad \text{(Eşitlik 2)}$$

$$B = \mathrm{Ort}(b_j) - A\mathrm{Ort}(b_i) \qquad \text{(Eşitlik 3)}$$

Ortalama-Sigma yönteminde, A ve B katsayılarının hesaplanmasında güçlük parametrelerinin standart sapması ve ortalaması kullanılmaktadır (Marco, 1977). A ve B katsayılarının hesaplanması ise Eşitlik 4 ve Eşitlik 5' te görüldüğü gibidir.

$$A = \frac{SS(b_i)}{SS(b_j)} \qquad \text{(Eşitlik 4)}$$

$$B = Ort(b_j) - AOrt(b_i) \qquad \text{(Eşitlik 5)}$$

Karakteristik eğri yöntemleri olarak adlandırılan ikinci grupta ise; "Haebara (Ha)" ve "Stocking-Lord (StLr)" olarak adlandırılan iki yöntem yer almaktadır. Ha yönteminde, aynı yetenek düzeyindeki bireylerin madde karakteristik eğrileri arasındaki fark, her bir maddeye ait karakteristik eğriler arasındaki farkın kareleri toplamıdır (Haebara,1980; Kolen ve Brennan, 2004; Raju ve Arenson, 2002). Bu yöntemde A ve B eğim katsayılarının hesaplanması Eşitlik 6, 7 ve 8'de görüldüğü gibidir.

$$HaL(\theta_i) = \sum_{j=1}^{m}\left[p_{ij}(\theta_i, a_{1j}, b_{1j}, c_{1j}) - p_{ij}(\theta_i, a^*_{2j}, b^*_{2j}, c_{2j})\right]^2 \qquad \text{(Eşitlik 6)}$$

$$\alpha^*_{2j} = \frac{\alpha_{2j}}{A} \qquad \text{(Eşitlik 7)}$$

$$b^*_{2j} = Ab_{2j} + B \qquad \text{(Eşitlik 8)}$$

Stocking-Lord yönteminde, aynı yetenek düzeyindeki bireylerin madde karakteristik eğrileri arasındaki fark, her bir maddeye ait karakteristik eğriler arasındaki farkın toplamının karesi olarak tanımlanmaktadır (Stocking-Lord, 1983; Kolen ve Brennan, 2004; Raju ve Arenson, 2002). Bu yöntemde A ve B eğim katsayılarının hesaplanması Eşitlik 9, 10 ve 11'de görüldüğü gibidir.

$$StLrL(\theta_i) = \left[\sum_{j=1}^{m}\left[p_{ij}(\theta_i, a_{1j}, b_{1j}, c_{1j}) - p_{ij}(\theta_i, a^*_{2j}, b^*_{2j}, c_{2j})\right]\right]^2 \qquad \text{(Eşitlik 9)}$$

$$\alpha^*_{2j} = \frac{\alpha_{2j}}{A} \qquad \text{(Eşitlik 10)}$$

$$b^*_{2j} = Ab_{2j} + B \qquad \text{(Eşitlik11)}$$

Test eşitleme sorunu, aynı özelliğin farklı test formları kullanılarak ölçülmesi durumunda gündeme gelir. TEOG ve benzeri sınavlarda testlerin gizliliğini korumak için, farklı zamanlarda yapılan sınav uygulamalarında farklı sorular / formlar kullanılmaktadır. Farklı formların kullanılması, testlerin güçlük düzeylerinde de farklılaşmaya neden olmakta ve dolayısıyla bu durum bireylerin birbirleriyle karşılaştırılabilmesini olanaksız kılmaktadır. Bu tür sorunlar, test eşitleme çalışmalarının öneminin her geçen gün daha da iyi anlaşılmasına ve bu çalışmalara olan ihtiyacın artmasına yol açmaktadır. Bununla birlikte, Türkiye'de yapılan eşitleme çalışmalarının birçoğu, gerçek sınav verilerinin elde edilmesinin çoğu zaman mümkün olamamasından kaynaklı olarak ya yapay (simülatif) verilerle ya da sınavlarda çıkan soruların farklı gruplarda yeniden uygulanması yoluyla

toplanan verilerle gerçekleştirilmektedir. Dolayısıyla gerçek veriler ile yapılan araştırmalara ilişkin önemli bir ihtiyaç da ortaya çıkmakta ve bireyler hakkında önemli kararların verildiği bu sınavların eşitlenmesi, hangi eşitleme koşulunun daha uygun olduğunun belirlenmesi önem taşımaktadır.

Yukarıda da değinildiği üzere, TEOG sisteminde asıl sınav uygulamasının ardından, sınava giremeyen öğrenciler için bir de mazeret sınavı uygulaması yapılmaktadır. Mazeret sınavının da asıl sınav ile aynı kapsamı ölçtüğü, paralel olduğu varsayılmaktadır. Ancak bu bir varsayımdan ibarettir ve ne yazık ki bunu destekleyecek bilimsel kanıtlar mevcut değildir. TEOG asıl sınav ile TEOG mazeret sınavı sonuçlarının eşitlenmesi, sınavın geçerliği ve güvenirliği açısından önem taşımaktadır; çünkü testin, bireyler hangi formu alırsa alsın, herhangi bir grubun lehine ya da aleyhine sonuçlar doğurmaması gerekmektedir (Angoff, 1971). Bu kapsamda bu araştırmanın problemini TEOG asıl sınav (TEOG-A) matematik alt testi ile TEOG mazeret sınavı (TEOG-M) matematik alt testlerinin MTK'ya dayalı yöntemler kullanılarak eşitlenmesi ve hangi eşitleme yönteminin (ortalama-ortalama, ortalama-sigma, Haebara ve Stocking-Lord) daha uygun olduğuna karar verilmesi oluşturmaktadır.

2. YÖNTEM

2.1. Araştırma Modeli

Bu araştırmada, TEOG asıl sınav (TEOG-A) matematik alt testi ile TEOG mazeret sınavı (TEOG-M) matematik alt testi MTK'ya dayalı kestirim yöntemleri kullanılarak karşılaştırılmıştır. Burada amaçlanan en az hatalı sonuçlar üreten yöntem ve koşulların belirlenmesidir. Bu yönüyle araştırmanın temel araştırma niteliği taşıdığı ifade edilebilir.

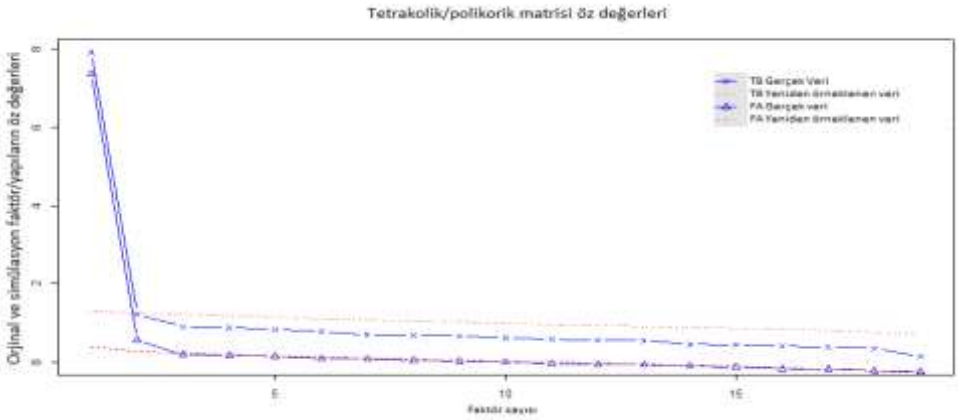
2.2. Çalışma Grubu

2013-2014 eğitim-öğretim yılı I. yarıyılında TEOG kapsamında uygulanan merkezi ortak sınavlardan matematik alt testi uygulanan N=1.275.541 öğrenci, mazeret sınavı kapsamında ise matematik alt testi uygulanan N=3392 öğrenci bulunmaktadır. Araştırmanın çalışma grubunun oluşturulmasında öncelikle matematik mazeret sınavına girip A kitapçığını alan öğrenci sayısı belirlenmiştir. Bu sayının n=1747 olduğu saptandıktan sonra, asıl sınav uygulamasından da eşit sayıda öğrencinin tesadüfî olarak seçimi sağlanmıştır. Böylelikle çalışma grubu A kitapçığını alan toplam n=3494 öğrenciden oluşmuştur.

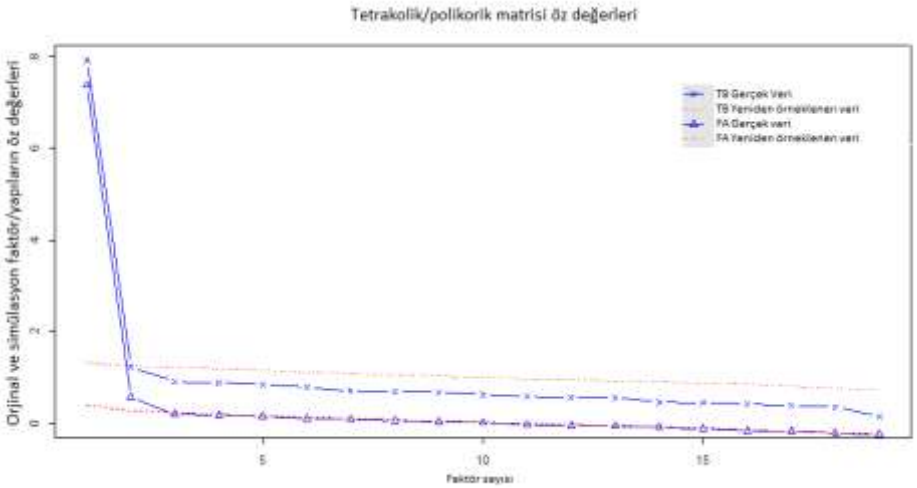
2.3. Veriler ve Analizi

Araştırmada MEB'den elde edilen TEOG-A-2014 ve TEOG-M-2014 matematik alt testlerine ilişkin hazır verileridir. Kestirimler MTK çerçevesinde yapıldığından, öncelikle kuramın varsayımları test edilmiştir. Kuramın en önemli varsayımı tek boyutluluktur. Tek boyutluluk test ile ölçülen özelliğin başat bir boyuta ait olması anlamına gelmektedir. Tek boyutluluk varsayımının test edilmesi için, R paketlerinden "polycor" kullanılmıştır. Bu paket ikili (1-0) puanlanan verilere, tekrakorik korelasyon matrisine dayalı olarak faktör analizi uygulanabilmesine olanak sağlamaktadır. Bu çalışma kapsamında tek boyutluluğun belirlenmesinde faktör sayısını belirlemenin Horn (1965) tarafından önerilen bir yöntemi olan "paralel analiz"den yararlanılmıştır. Paralel analiz, faktör sayısını belirlemek için Monte Carlo simülasyon yöntemi ile gerçek veri setine paralel

rastgele bir veri setinin üretilmesi ve iki veri setinin özdeğerlerinin karşılaştırılması mantığına dayanır. Böylece üretilen verilere ilişkin özdeğerin, gerçek verilere ilişkin özdeğerden yüksek olduğu noktadaki faktör sayısı, anlamlı faktör sayısı olarak kabul edilir. Grafik 1'de TEOG-A ve Grafik 2'de TEOG-M gerçek verileri ve bu verilere paralel olarak üretilmiş verilere ilişkin yamaç birikinti grafikleri sunulmaktadır.



Grafik 1. TEOG-A Gerçek Verileri ve Paralel Analizi ile Üretilmiş Verilere İlişkin Yamaç Birikinti Grafiği



Grafik 2. TEOG-M Gerçek Verileri ve Paralel Analizi ile Üretiliş Verilere İlişkin Yamaç Birikinti Grafiği

Grafik 1 ve Grafik 2 incelendiğinde, hem TEOG-A, hem de TEOG-M verilerinin tek boyutluluk varsayımını karşıladığı açıkça görülmektedir. Bu durum aynı zamanda test eşitlemenin temel koşulu olan "eşitlenecek iki testin aynı yapıyı, özelliği veya yeteneği ölçmesi" gerekliliğinin de karşılandığı anlamında değerlendirilebilir. Çünkü test

eşitlemede eşitlenecek iki testin sadece aynı özelliği ölçmesi yeterli değildir; aynı zamanda bu testlerin tek bir özelliği ölçmesi de gerekmektedir. Bunların yanında testlerin aynı yapıyı ölçüp ölçmediğinin diğer bir kanıtı olarak, Kan (2011)'de de belirtildiği gibi her iki testin aritmetik ortalamaları karşılaştırılmış ve aritmetik ortalamaları arasında manidar bir farklılık bulunamamıştır (t=0,320 ; p>0.05).

MTK'nın bir diğer varsayımı olan model-veri uyumunun 1, 2 ve 3 Parametreli Lojistik Modellere (PLM) göre incelenmesinde R paketlerinden "ltm" kullanılmış ve -2Loglikelihood değerleri hesaplanmıştır. Tablo 1'de model-veri uyumunun sınanamasına ilişkin -2Loglikelihood değerleri sunulmaktadır.

Tablo 1.
Model Veri Uyumunun Sınanması

		TEOG-A	TEOG-M
-2 Loglikelihood	**1 PLM**	-18751,81	-18732,58
	2 PLM	-18406,97	-18361,62
	3 PLM	-18120,76	-18029,61

Tablo 1 incelendiğinde, parametre sayısı arttıkça tüm modellerdeki loglikelihood değerlerinin de manidar bir şekilde düştüğü görülmektedir. Bu değerler her parametre sayısı (19 ayırt edicilik, 19 şans parametresi) arttığında ilgili serbestlik derecesinde (19) ki-kare değeri tablo değerini ($\chi = 36.19, \mathrm{p} = 0.01$) aşmaktadır. Bu durumda araştırmacıya en çok bilgi verecek olan loglikelihood değeri en düşük ve parametre sayısı en fazla olan modeli seçmelidir. Bu bulgu, her iki test için de 3 parametreli lojistik modelin daha uygun olduğu anlamında değerlendirilmiştir.

Verilerin yapısının test eşitleme koşullarına uygunluğu ve MTK analizleri için model veri uyumuna yönelik incelemenin ardından, öncelikle madde ve birey parametreleri FlexMIRT programı kullanılarak kestirilmiştir. Madde parametreleri için 3PLM ve birey parametreleri için Maksimum Likelihood (ML) yöntemi kullanılmıştır. Daha sonra R programlama dilinde oluşturulmuş "equateIRT" paketi kullanılarak tüm parametreler tanıtılmış ve her iki test formu MTK'ya dayalı ortalama-ortalama, ortalama-sigma, Haebara ve Stocking-Lord yötemleri ile eşitlenmiş, standart hataları ve RMSE (Root Mean Square Error) değerleri hesaplanmıştır (Battauz, 2013 ve Battauz, 2015). RMSE, maddelerin kestirilen parametreleri ile gerçek parametre değerleri arasındaki farkın kareleri toplamının tekrar sayısına oranının kareköküdür. Bu değer toplam hata miktarı olarak da adlandırılır. Hangi eşitleme yönteminin en iyi sonucu verdiğine karar vermenin ölçütü olarak RMSE ve standart hata değerleri incelenmiştir. Düşük RMSE ve standart hata değeri, iyi bir test eşitleme işleminin ya da yöntemlerden hangisinin daha uygun olduğunun bir göstergesidir; çünkü önemli olan işlemlerin minimum hata içerecek şekilde gerçekleştirilmesidir.

3. BULGULAR

Eşitleme işleminin yapılabilmesi, öncelikle eşitlenecek testlerin istatistiksel olarak da benzer olmasını gerektirmektedir. Bu nedenle testlerin ortalama, standart sapma ve varyans gibi betimsel istatistikler incelenmiş ve Tablo 3'te sunulmuştur.

Tablo 3.
TEOG-A ve TEOG-M Matematik Alt Testlerine İlişkin Betimsel İstatistikler

Test	Madde Sayısı	Ortalama	S	Varyans	Marjinal Güvenirlik	Ortalama Güçlük	Basıklık	Çarpıklık
TEOG-A	19	8.72	4.65	21.62	0.75	0.68	-0.59	0.62
TEOG-M	19	8.77	4.75	22.56	0.76	0.63	-0.61	0.63

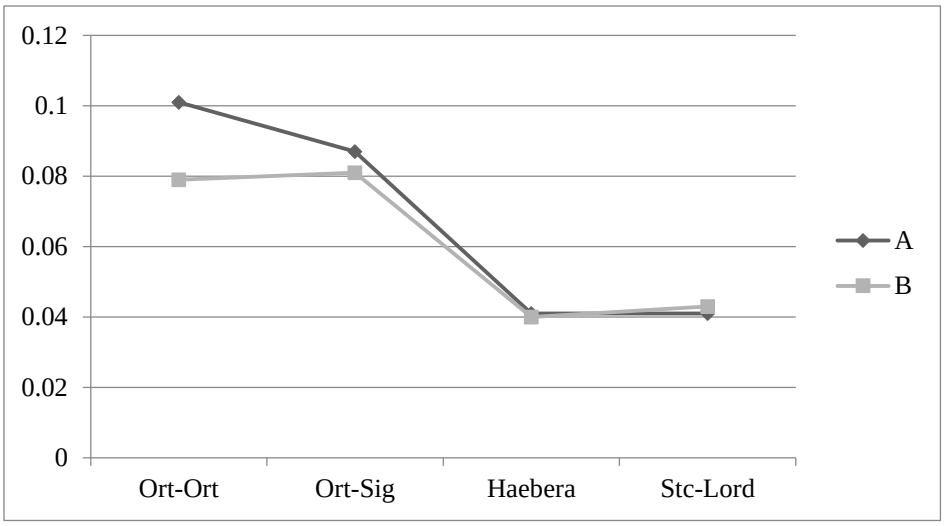
Tablo 3 incelendiğinde ortalama, standart sapma, güvenirlik ve diğer ölçülerin her iki test için de benzer olduğu ve testlerin eşitleme için uygun olduğu sonucuna ulaşılmıştır. Testten elde edilen tüm parametrelerin birbirine çok yakın olması testin aynı yapıyı ölçtüğü ve eşitleme varsayımlarını karşıladığı görülmektedir.

Madde Tepki Kuramı'na dayalı test eşitlemenin en önemli aşamalarından biri, tüm parametrelerin ortak bir ölçeğe yerleştirilmesidir (Loyd ve Hoover, 1980 ve Kolen ve Brennan, 2004). Testlerin ortak ölçeğe yerleştirilmesi madde parametrelerinden elde edilen A ve B eğim katsayılarının, Eşitlik 1 aracışığıyla psikometrik olarak dönüştürülmesi ile gerçekleştirilir. Tüm yöntemler için hesaplanan A ve B'ye ilişkin Eşitleme Katsayıları (EK) ve Eşitlemenin Standart Hataları (ESH) Tablo 4'te sunulmuştur.

Tablo 4.
Farklı Eşitleme Yöntemlerine İlişkin Eşitleme Katsayıları ve Eşitleme Hataları

Parametre	Moment Yöntemleri				Karakteristik Eğri Yöntemleri			
	Ort-ort EK	*ESH*	Ort-sigma EK	*ESH*	Haebara EK	*ESH*	Stocking-Lord EK	*ESH*
A	0.897	0.101	0.938	0.087	0.934	0.041	0.942	0.041
B	0.052	0.079	0.026	0.081	0.022	0.040	0.017	0.043

Tablo 4 incelendiğinde moment yöntemlerinin karakteristik eğri yöntemlerine göre daha büyük standart hatalar ürettiği görülmektedir. Tüm yöntemlerden elde edilen standart hatalar incelendiğinde Heabara ve Stocking-Lord yöntemlerinin daha düşük standart hatalarla A ve B katsayıları ürettiği görülmektedir. En fazla hata içeren yöntemin ise ortalama-ortalama yöntemi olduğu dikkat çekmektedir. Elde edilen standart hata miktarları, net olarak Grafik 3'te de görülmektedir. Bununla birlikte en yüksek eğim katsayılarının da Haebera ve Stocking-Lord yöntemlerinden elde edildiği belirlenmiştir. Bu bulgu, Stocking ve Lord (1983), Baker ve Al-Karni (1991) ve Ogasawara (2001) tarafından yapılan araştırmalarda karakteristik eğri yöntemlerinin moment yöntemlerine göre daha iyi sonuçlar ürettiği şeklindeki bulgularla da tutarlıdır.



Grafik 3. Farklı Eşitleme Yöntemleri ile Elde Edilen A ve B Değerlerine Ait Eşitleme Hataları

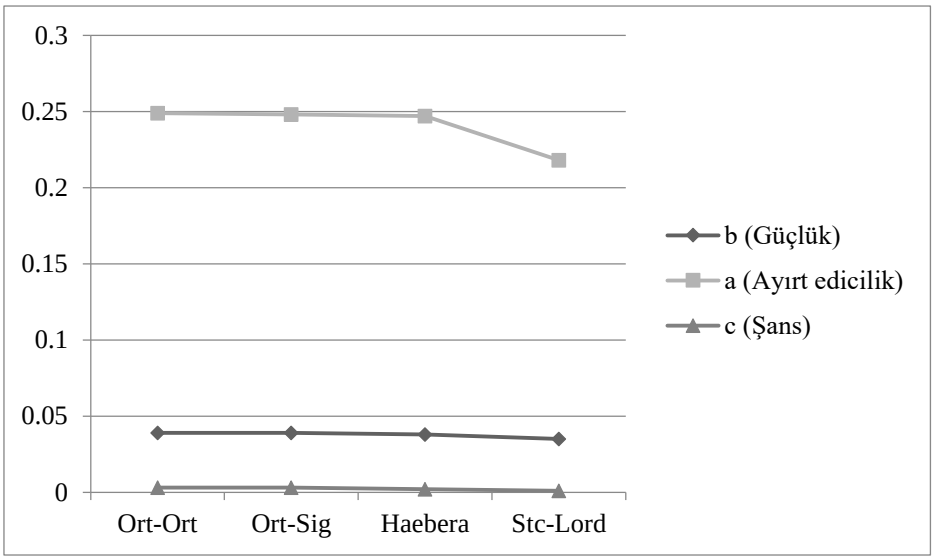
Elde edilen eğim katsayılarından (A ve B) yararlanarak tüm parametre değerleri ve yetenek dağılımları için ilgili test formalarının eşitliklerine bakılmıştır. Öncelikle Tablo 5'te sunulan, eğim katsayılarından elde edilen parametre değerlerine ait RMSE değerleri incelenmiştir.

Tablo 5.
Eşitlenen Madde Parametrelerine İlişkin RMSE Değerleri

Test	Madde Parametreleri	Moment Yöntemleri		Karakteristik Eğri Yöntemleri	
		Ortalama-ortalama	Ortalama-sigma	Haebara	Stocking-Lord
TEOG-A		RMSE	RMSE	RMSE	RMSE
TEOG-M	b	0.039	0.039	0.038	0.035
	a	0.249	0.248	0.247	0.218
	c	0.003	0.003	0.002	0.001

Tablo 5 incelendiğinde parametre değerlerinin dönüştürülmesinden elde edilen en düşük RMSE değerlerinin Haebera ve Stocking-Lord yöntemlerinden elde edildiği görülmektedir. TEOG-A ile TEOG-M'nin eşitlenmesi sonucunda tüm yöntemlerden elde edilen standart hatalar incelendiğinde, en düşük standart hatanın Stocking-Lord yöntemine, en yüksek hatanın ise ortalama-ortalama yöntemine ait olduğu belirlenmiştir. Bu bulgu Chu ve Kamata (2000) tarafından yapılan araştırmada elde edilen bulgularla örtüşmektedir.

Grafik 4'te madde parametrelerine ilişkin RMSE değerleri sunulmaktadır.



Grafik 4. Madde Parametrelerine İlişkin RMSE Değerleri

Grafik 4'te sunulan madde parametrelerinin eşitleme hataları incelendiğinde, tüm parametre değerlerinde en az hatanın yine Haebara ve Stocking-Lord yöntemlerinden elde edildiği görülmektedir. En düşük RMSE değeri Stocking-Lord yöntemine aittir. Bunun yanı sıra tüm eşitleme yöntemleri için madde ayırt edicilik parametresi için yüksek RMSE'ler elde edildiği görülmektedir.

Elde edilen eğim katsayılarından (A ve B) yararlanarak yetenek aralıkları için ilgili test formalarının eşitliği incelenmiştir. Öncelikle Tablo 6'da eğim katsayılarından elde edilen yetenek parametre değerlerine ilişkin RMSE değerleri sunulmuştur.

Tablo 6.

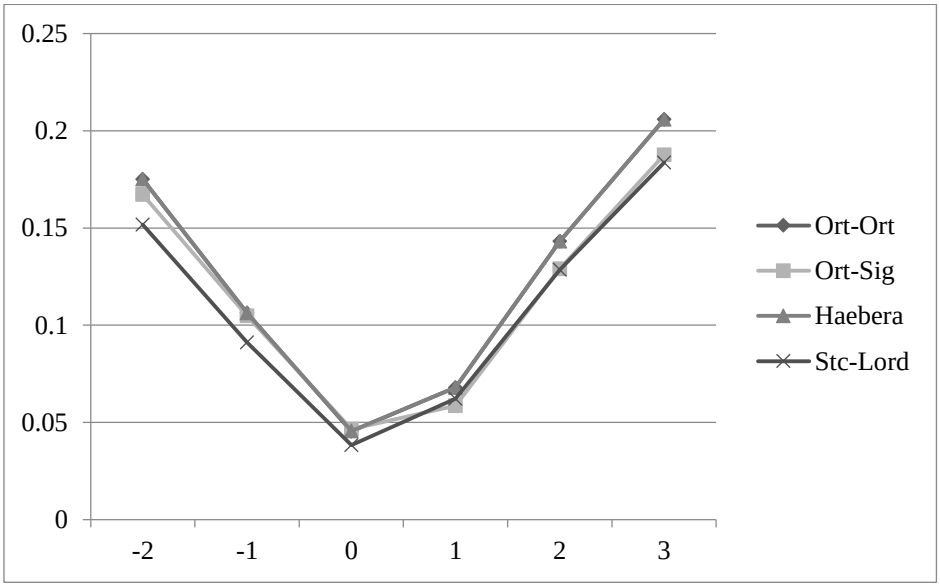
Eşitlenen Birey (Yetenek) Parametresine İlişkin RMSE Değerleri

Test	**Yetenek Aralığı**	**Moment Yöntemleri**		**Karakteristik Eğri Yöntemleri**	
		Ortalama-ortalama	**Ortalama-Sigma**	**Haebara**	**Stocking-Lord**
		RMSE	RMSE	RMSE	RMSE
TEOG-A TEOG-M	-2	0,1751	0,1673	0,1752	0,1517
	-1	0,1063	0,1049	0,1063	0,0912
	0	0,0454	0,0466	0,0455	0,0383
	1	0,0680	0,0586	0,0680	0,0622
	2	0,1432	0,1290	0,1432	0,1285
	3	0,2060	0,1876	0,2059	0,1837

Tablo 6'da sunulan TEOG-A ve TEOG-M testleri eşitlendiğinde yetenek parametresine ilişkin RMSE değerleri incelendiğinde, tüm yetenek aralıkları için en düşük RMSE

değerinin Stocking-Lord yönteminden elde edildiği görülmektedir. Özellikle "0" ile ifade edilen ortalama yetenek düzeyi için eşitleme hata miktarının oldukça düşük olduğu görülmektedir. Bu durum testlerin yalnızca ortalama düzeydeki bireyleri ayırmada etkili olduğu, buna karşılık uçlara doğru gidildikçe hata miktarının arttığı, bir başka deyişle testlerin eşitlikten uzaklaştığı ifade edilebilir.

Grafik 5'te yetenek parametrelerine ilişkin RMSE değerleri sunulmaktadır. Grafik 5 incelendiğinde, yetenek parametrelerine ait RMSE değerlerinden en düşüğünün Stocking-Lord yönteminde elde edildiği görülmektedir. Tüm yöntemlerden elde edilen eşitleme hataları incelendiğinde, en düşük hatanın "0" (ortalama yetenek) için üretildiği görülmektedir. Araştırmadan elde edilen bulguların alanyazınla da paralel olduğu ifade edilebilir (Cohen ve Kim, 1998; Hanson ve Beguin, 2002; Kim ve Kolen, 2006; Hung, Wu ve Chen, 1991; Way ve Tang, 1991; Karkee ve Wright, 2004; Kaskowitz ve De Ayala, 2001; Kim ve Lee, 2004; Kim ve Lee, 2006; Kim ve Kolen, 2004; Kim ve Song, 2004).



Grafik 5. Yetenek Parametresine İlişkin RMSE Değerleri

4.TARTIŞMA ve SONUÇ

Bu araştırmada TEOG asıl sınav ile TEOG mazeret sınavı matematik alt testlerinin eşitlenmesinde Madde Tepki Kuramına dayalı eşitleme yöntemlerinden (ortalama-ortalama, ortalama-sigma, Haebara ve Stocking-Lord) elde edilen eşitleme hataları karşılaştırılmış ve en düşük eşitleme hatasının hangi yöntem ile elde edildiğinin belirlenmesi amaçlanmıştır. Eşitleme işleminden önce, Madde Tepki Kuramı ve eşitlemenin varsayımlarından olan tek boyutluluk varsayımı test edilmiştir. Böylelikle veri yapısının tek boyutlu olduğu ve eşitleme için uygun olduğu belirlenmiştir. Bunun yanı sıra testlerin eşitlenebilmesi için gerekli olan ortalama, standart sapma vb. betimsel istatistikler incelenmiş ve bu varsayımların da karşılandığına karar verilmiştir.

Madde Tepki Kuramı'na dayalı olarak test eşitlemede en önemli aşamalarından biri olan ortak ölçeğe yerleştirme için madde parametrelerinden elde edilen A ve B eğim katsayıları kullanılarak, testler ortak bir ölçeğe yerleştirilmiş ve bunlara ilişkin RMSE değerleri hesaplanmıştır. A ve B eğim katsayılarını hesaplamada en düşük hatanın Karakteristik Eğri Yöntemleri grubunda yer alan Heabara ve Stocking-Lord yöntemlerinden, en yüksek hatanın ise ortalama-ortalama yönteminden elde edildiği sonucuna ulaşılmıştır. Madde ve yetenek parametrelerinin eşitlenmesinde de karakteristik eğri yöntemlerinin daha düşük hatalarla eşitleme yaptığı saptanmıştır.

Araştırmadan elde edilen sonuçlar, en düşük eşitleme hatasının karakteristik eğri yöntemlerinden elde edildiğini ortaya koymaktadır. Bu yöntemlerden de en düşük eşitleme hatasının Stocking-Lord yönteminden elde edildiği belirtilebilir. Ortalama-ortalama yöntemi ise en yüksek eşitleme hatası üreten yöntemdir. Bu sonuçlara göre Stocking-Lord yönteminin TEOG-A ve TEOG-M Matematik alt testlerinin eşitlenmesinde daha uygun yöntem olduğu sonucuna ulaşılmıştır.

Türkiye'de, öğrencileri bir üst öğretim kurumuna yerleştirmek amacıyla yapılan geniş ölçekli sınavlar, sınava giren öğrencilerin geleceklerini etkilemesi, ortaöğretim ve yükseköğretimin kalitesi açısından önem taşımaktadır. Bu sınavlara, bir üst öğrenim kurumuna yerleşme, bir konuda yetkinlik kazanma veya akademik birtakım kararlar alma gibi amaçlar için gerek duyulur (Cizek, 2001; Resnick, 2004). Her yıl testlerin farklı formlarının geliştirilerek kullanıldığı sınavlarda testlerin eşitliğinin sağlanması öğrenciler açısından çok önemlidir. Bu araştırmada da TEOG asıl sınav ile TEOG mazeret sınavlarının eşitlenmesi çalışmalarının yapılması gerekliliği vurgulanmıştır. Bu sınavlar için en uygun yöntemlerin karakteristik eğri yöntemlerinden Stocking-Lord yönteminin olduğu görülmektedir. Sınav uygulayıcılarının bu yöntemi ele alarak bu testlerin eşitliklerini sınaması ve sınavın bireyler arasında herhangi bir yanlılığa neden olmasını engellemeleri önerilebilir.

KAYNAKÇA

Angoff, W. H. (1971). Scales, norms and equivalent scores. In Thorndike, R. L. (Ed.), *Educational Measurement* (p. 509-600). Washington: American Council on Education.

Angoff, W. H. (1981). Summary and derivation of equating methods used at ETS. In P. W. Holland & D. B. Rubin (Eds.). Test Equating. New York: Academic Press.

Baker, F. B. & Al-Karni, A. (1991). A Comparison of two procedures for computing IRT Equating Coefficients. *Journal of Educational Measurement, 28 (2),* 147-162.

Barnard, J. J. (1996). In search for equity in educational measurement: Traditional versus modern equating methods. *Paper presented at ASEESA's national conference at the HRSC Conference Centre*, Pretoria, South Africa.

Battauz, M. (2013). IRT test equating in complex linkage plans. *Psychometrika*, *78*, 464-480.

Battauz, M. (2015). equateIRT: An R Package for IRT Test Equating. Accepted for publication in *Journal of Statistical Software.*

Braun, H. I., & Holland, P. W. (1982). Observed-score test equating: A mathematical analysis of some ETS equating procedures. In P. W. Holland & D. B. Rubin (Eds.), Test equating (pp. 9–49). New York: Academic.

Casbarro, J. (2004). Reducing anxiety in the era of highstakes testing. *Principals, 83*(5), 36-38.

Chu, K. L. & Kamata, A. (2000). Nonequivalent Group Equating via 1-P HGLLM. New Orleans, LA: *Paper presented at the annual meeting of the American Educational Research*

Cizek, G. J. (2001). Cheating to the test. *Education Matters Journal*, *1*(1), 40-47.

Cohen, A. S. & Kim, S. H. (1998). An ınvestigation of linking methods under the graded response model. *Applied Psychological Measurement, 22 (2),* 116-130.

Dorans, N. J. (2000). Research notes: distinctions among classes of linkages. The College Board, Office of Research and Development.

Haebara, T. (1980). Equating logistic ability scales by a weighted least squares method. Japanese *Psychological Research, 22 (3),* 144-149.

Hambleton, R. K. & Swaminathan, H. (1985). *Item Response Theory: Principles and Applications.* Boston: Academic Puslishers Group.

Hambleton, R. K., Swaminathan, H. & Rogers, H. (1991). *Fundamentals of Item Response Theory*. Newbury Park CA: Sage.

Hanson, B. A. & Beguin, A. A. (2002). Obtaining a common scale for ıtem response theory ıtem parameters using separate versus concurrent estimation in the common-ıtem equating design. *Applied Psychological Measurement, 26 (1),* 3–24.

Hung, P., Wu, Y., & Chen, Y. (1991). IRT Item Parameter Linking: Relevant Issues for the Purpose of Item Banking. *Paper presented at the International Academic Symposium on Psychological Measurement, Taiwan.*

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

Kaskowitz, G. S. & De Ayala, R. J. (2001). The effect of error in ıtem parameter estimates on the test response function method of linking. *Applied Psychological Measurement, 25 (1),* 39-52.

Kim, S. & Kolen, M. J. (2006). Robustness of Format Effects of IRT Linking Methods for Mixed Format Tests. *Applied Measurement in Education, 19* (4), 357-381

Kim, S. & Kolen, M. J. (2004*). Optimally defining criterion functions for the characteristic curve procedures in the ırt scale linking*. Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA.

Kim, S. & Lee, W. (2004). IRT Scale linking methods for mixed-format tests (ACT research report 2004). Iowa City, IA: ACT, Inc.

Kim, S. & Lee, W. C. (2006). An Extension of Four IRT Linking Methods for Mixed-Format Tests. *Journal of Educational Measurement, 43 (1),* 53–76.

Kim, S. & Song, M.-Y. (2004). *Least squares estimation of IRT scale linking coefficients under the graded response model.* Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA.

Kolen, M. J. & Brennan, R. L. (2004). Test equating: Methods and practices. New York: Springer.

Loyd, B. H. & Hoover, H. D. (1980). Vertical equating using the rasch model. *Journal of Educational Measurement, 17 (3),* 179-193.

Marco, G. L. (1977). Item characteristic curve solutions to three ıntractable testing problems. *Journal of Educational Measurement, 14* (2), 139- 160.

Mâsse, L. C., Allen, D., Wılson, M., ve Wılliams, G. (2006). Introducing equating methodologies to compare test scores from two different self-regulation scales". *Health Education Research 21*, 110-120.

Ogasawara, H. (2001). Standard errors of ıtem response theory equating/linking by response function methods. *Applied Psychological Measurement, 25* (1), 53–67.

Raju, N. S. & Arenson, E. A. (April, 2002). *Developing a common metric in ıtem response theory: an area-minimization approach*. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.

Resnick, M. (2004). The educated student: Defining and advancing student achievement. Alexandria VA: *National School Boards Association*

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

Tanguma, J. (2000). "Equating test scores using the linear method: A primer." *Paper presented at the annual meeting of the Southwest Educational Research Association.* Dallas, TX.

Way, W. D. & Tang, K. L. (1991, April 4-6). *A comparison of four logistic model equating methods.* Paper presented at the Annual Meeting of the American Educational Research Association, Chicago, IL.

EXTENDED ABSTRACT

1. Introduction

The purpose of the study was to equate the main maths test of "Transition from Primary Education to Secondary Education System Examination (TEOG)" and the maths test in the make-up exam of TEOG by using Item Response Theory-based methods, and to decide which equation method (mean-mean, mean-sigma, Haebara and Stocking-Lord) was more appropriate to use.

2. Method

In the study, Item Response Theory-based predictive methods were compared using the data from the main maths test of Transition from Primary Education to Secondary Education System Examination (TEOG-A) and the maths test used in the make-up exam of TEOG (TEOG-M). Here, the purpose was to define methods and conditions that would give results with least errors. In this sense, it could be suggested that this was a baseline study. During the fall semester of the 2013-2014 academic year, there were 1.275.541 students who were given the maths test of TEOG, one of the central system common exams, and the number of those who took the maths test of the make up examination of TEOG was 3392. The number of students who were given Booklet A in the maths test of the make up examination of TEOG was the first thing to determine to form the study group of the research. When the number was found as 1747, 1747 students who were given Booklet A in the maths test of TEOG-A were randomly chosen. As a result, the study group consisted of a total of 3494 students. The ready dataset was employed for the study. The research data were from the maths tests of TEOG-A-2014 and TEOG-M-2014. The assumptions of the theory were primarily tested using "polycor" and "ltm", two R packages, since the estimations were based on the Item Response Theory (IRT) framework. FlexMIRT program was used to estimate item and person parameters for the equation of the tests. 3PLM was employed for the item parameters, and Maximum Likelihood (ML) method was employed for the person parameters. The equation process was completed using R "equateIRT" package.

3. Findings, Discussion and Results

For proper equation, tests under question must be similar regarding statistics. For this reason, descriptive statistics such as mean, standard deviation and variance were examined for the tests in the study and the tests were found to have similar means, standard deviations, reliabilities and other measures, and thus eligible for the equation.

One of the most significant phases of test equation based on Item Response Theory is to place all parameters in a common scale (Loyd and Hoover, 1980; Kolen and Brennan, 2004). This placement process is performed when A and B curve coefficients obtained from item parameters are psychometrically transformed. When A and B curve equation coefficients calculated for all the methods were examined, it was seen that moment methods gave higher standard errors when compared to characteristic curve methods. When standard errors obtained from all the methods were examined, it was seen that Heabara and Stocking-Lord methods generated A and B coefficients with less standard errors. The method with the most errors was found to be mean-mean. However, it was concluded that the highest curve coefficients were obtained from Haebera and Stocking-

Lord methods. This finding was consistent with the ones from the studies by Stocking and Lord (1983), Baker and Al-Karni (1991) and Ogasawara (2001) that concluded characteristic curve methods give better results than moment methods. These results supported the finding of this research.

The equation of the test forms was tested for all the parameter values and ability distributions with the help of the obtained curve coefficients. First of all, it was concluded that the lowest RMSE values as a result of parameter value transformation were provided by Haebera and Stocking-Lord methods when RMSE values of the parameters obtained from the curve coefficients were examined. As a result of the equation of the maths tests of TEOG-A and TEOG-M, it was concluded that the lowest standard error was provided by Stocking-Lord method, and the highest by the mean-mean method, when all the standard errors obtained from all the methods were compared. This finding supported the ones from the study by Chu and Kamata (2000).

The equation of the related test forms was tested for ability intervals, employing the obtained curve coefficients (A and B). When RMSE values obtained from the ability parameter values in the curve coefficients were examined, it was seen that the lowest RMSE value of the ability parameters was provided by Stocking-Lord method. When equation errors from all the methods were examined, it was seen that the lowest error was generated for "0" (mean ability). It was also shown that the equation error particularly for "0" (mean ability level) was rather low. This case could be interpreted in the following way: the tests were efficient to differentiate those at mean levels, but they became distanced from equation when the number of errors increased down-line. The research findings were found to be parallel to the ones in the literature (Cohen and Kim, 1998; Hanson and Beguin, 2002; Kim and Kolen, 2006; Hung, Wu and Chen, 1991; Way and Tang, 1991; Karkee and Wright, 2004; Kaskowitz and De Ayala, 2001; Kim and Lee, 2004; Kim and Lee, 2006; Kim and Kolen, 2004; Kim and Song, 2004). The research results have shown that the lowest equation error could be obtained from characteristic curve methods. Among those, Stocking-Lord method has been found to give the lowest equation error whereas mean-mean method has been concluded to give the highest. According to these results, it was concluded that Stocking-Lord method was more appropriate to use for the equation of the maths tests of TEOG-A and TEOG-M.

In our country, the large-scale examinations to place students in higher level schools are crucial in terms of their influence on the future of students and the quality of secondary and higher level education. Such examinations are needed for certain aims such as university entrance, competence gain or particular academic decision-making processes (Cizek, 2001; Resnick, 2004). Providing equation in those examinations in which different test forms are developed every year plays an important role for students. Hence, every year, it is essential to provide equation in examinations and to choose the most appropriate methods. Also, in this study, it is highlighted that equation studies for TEOG-A and TEOG-M are needed. It is seen that the most eligible method for the examinations is Stocking-Lord, one of the characteristic curve methods. Examiners need to test equation of tests based on this method and prevent examinations from giving biased results for individuals.