Hacettepe Üniversitesi Eğitim Fakültesi Dergisi (H. U. Journal of Education) 31(3): 447-460 [2016]

DOI: 10.16986/HUJE.2016015183
Makalenin Geliş Tarihi: 03.01.2014
Makalenin Kabul Tarihi: 22.12.2015
Online Yayın Tarihi: 14.01.2016

Dikey Ölçeklemede Madde Tepki Kuramına Dayalı Kalibrasyon ve Yetenek Kestirim Yöntemlerinin Karşılaştırılması*

A Comparison of Calibration Methods and Proficiency Estimators Based on Item Response Theory in Vertical Scaling

Ayşegül ALTUN**, Hülya KELECİOĞLU***

ÖZ: Bu araştırma kapsamında; ortak madde deseninde madde tepki kuramına dayalı ayrı ve eş zamanlı kalibrasyon ve Expected A Posteriori, Maximum A Posteriori ve Maksimum Olabilirlik yetenek kestirim yöntemleri kombinasyonuyla elde edilen dikey ölçeklerin karşılaştırılması yapılmıştır. Bu karşılaştırmayı yapabilmek için öncelikle 2008-2010 yılları arasında yapılan SBS sınavında sorulan sorulardan 6., 7. ve 8. sınıf matematik testleri oluşturulmuştur. Daha sonra 6. sınıftaki 503, 7. sınıftaki 502 ve 8. sınıftaki 500 öğrencinin zorluk düzeyleri farklı olan matematik testlerine verdikleri cevaplardan elde edilen puanlar kullanılarak dikey ölçekler geliştirilmiştir. Dikey ölçekleme süreciyle aynı ölçeğe yerleştirilen bu puanlardan elde edilen ölçek puanları kullanılarak ortalamalar, ortalamalar arasındaki fark, etki büyüklükleri ve yatay uzaklıklar hesaplanmış ve böylece ölçekleme sürecinde yapılan seçimlerin dikey ölçeklemeyi nasıl etkilediği belirlenmeye çalışılmıştır.

Anahtar sözcükler: Dikey ölçekleme, madde tepki kuramı (MTK), matematik başarısı

ABSTRACT: In this study the comparison of the vertical scales, which are obtained through the combination of separate and concurrent calibration based on item response theory and Expected A Posteriori,, Maximum A Posteriori and Maximum Likelihood proficiency estimation methods, take place. For this comparison firstly, math tests for the 6th, 7th and 8th grades were composed from the questions asked in SBS (high school entrance exam) between 2008-2010 years. Then, the vertical scales were developed by using the scores obtained from the answers of 503 6th grade, 502 7th grade and 500 8th grade students to the math tests in different difficulty levels. By using the scale scores which were obtained from these scores placed in the same scale with the vertical scaling process, means and the difference between the means, effect sizes and horizontal distances were calculated so it was tried to be determined how the choices in scaling process affected the vertical scaling

Keywords: Vertical scaling, Item response theory (IRT), Mathematic achievements

1. GİRİŞ

Başarı testlerinin kullanım amaçlarından biri öğrencinin anlık başarı düzeyine ilişkin bilgi vermektir. Okullarda öğrencinin anlık başarı düzeyine ilişkin bilginin yanı sıra bir alanda sınıf ve öğrenci düzeyinde gelişiminin bilinmesine de gereksinim duyulur. Yıllara göre gelişim düzeyinin bilinmesi ile başarının sürekliliğine ilişkin bilgi elde edilebilir. Bu bilgiler ayrıca, öğrenci ve sınıf düzeyinde yapılacak iyileştirme çalışmalarının temelini oluşturur. Ancak öğrencilerin bir derste her sınıf düzeyinde aldığı puanlar aynı ölçekte yer almadığı için doğrudan karşılaştırılamaz. Karşılaştırmayı sağlayabilmek için aynı konu alanındaki ancak farklı güçlük düzeylerinde olan testlerin puanlarının aynı ölçeğe yerleştirilmesi gerekir. Böylece farklı sınıf

*Bu makale birinci yazarın doktora tezinden üretilmiştir.

** Yrd.Doç.Dr.,Ondokuz Mayıs Üniversitesi Eğitim Fakültesi, Samsun-TÜRKİYE, aysegul.altun@omu.edu.tr
*** Prof.Dr., Hacettepe Üniversitesi, Eğitim Fakültesi, Ankara-TÜRKİYE, hulyaebb@hacettepe.edu.tr

düzeylerindeki gelişim izlenebilir. Aynı konu alanında, ancak farklı sınıf düzeylerindeki testlerden alınan puanları aynı ölçeğe yerleştirme sürecine dikey ölçekleme, bu ölçeklere de dikey ölçek adı verilir (McBridge ve Wise, 2001).

Dikey ölçekleme sürecinde farklı veri toplama desenleri ve ölçekleme yöntemleri bulunmaktadır. Desenler ve yöntemler ölçekleme sonucunu da farklılaştırabilmektedir (Tong ve Kolen, 2007).

1.1. Veri Toplama Desenleri

Dikey ölçek geliştirmede ortak madde deseni, ölçekli test deseni ve eşit grup deseni en yaygın olarak kullanılan veri toplama desenleridir. Bu araştırmada ortak madde deseni kullanıldığından, aşağıda kısaca bu yöntem hakkında bilgi verilmiştir.

Ortak madde deseni: Ortak madde deseninde, her sınıf düzeyi için o sınıf düzeyine uygun test geliştirilir ve her test ait olduğu sınıf düzeyine uygulanır. Farklı sınıf düzeylerindeki testleri birbirine bağlamak için ard arda gelen sınıf düzeylerindeki testlerde ortak maddeler yer alır. Ortak maddeler ana testleri örnekleyecek biçimde oluşturulmalıdır. Ortak maddelerin farklı sınıf düzeylerindeki testleri nasıl örnekleyeceğinin belirlenmesi önemlidir. Ortak maddeler her sınıf düzeyinden seçildiğinde, alt sınıftaki öğrenciler öğrenmedikleri konulara ait sorularla karşılaşmakta; sadece alt sınıf düzeyinde olduğunda ise ortak madde seti ana testi örneklememektedir. Bunlara ek olarak ortak maddelerin seçiminde, ortak maddelerin testin genelinin bir minyatürü mü olacağı yoksa test edilen alanın iki sınıf düzeyinde ortak olan yapısını en iyi yansıtacak şekilde mi seçilmesi gerektiğinin kararının verilmesi gerekmektedir.

Ortak madde deseninde bir sınıf temel düzey olarak tanımlanır ve öğrencilerin ortak maddelerdeki performansları ardışık sınıf düzeyindeki puanların birbiriyle ilişkilendirilmesinde kullanılır. Daha sonra zincirleme bir süreçle diğer sınıf düzeyleri temel düzeyle bağlanır. Örneğin 5. sınıftan 7. sınıfa kadar öğrenci gelişimini ölçen testlerden 5. sınıf temel düzey olarak seçilebilir. Bu durumda 7. sınıf puanlarını temel düzeyle bağlamak için 7. sınıf puanları 6. ve 7. sınıfta ortak olan maddeler yardımıyla 6. sınıfla aynı ölçeğe yerleştirilir. Daha sonra 6. sınıf puanları 5. sınıf puan ölçeğine 5. ve 6. sınıflarda ortak olan maddeler yardımıyla yerleştirilir. Bu zincirleme kullanılarak 7. sınıf puanları temel düzey olan 5. sınıf ölçeğine yerleştirilmiş olur (Tong ve Kolen, 2007).

1.2. Ölçekleme Yöntemleri

En çok kullanılan üç ölçekleme yöntemi Hieronymus ölçekleme, Thurstone ölçekleme ve madde tepki kuramına (MTK) dayalı ölçeklemedir. Bu genel istatistiksel yaklaşımlarda kullanılan süreçler, kullanılan veri toplama desenine bağlıdır (Tong ve Kolen, 2010). Bu araştırmada MTK'ya dayalı ölçekleme yöntemi kullanılmıştır.

MTK ile ölçekleme: MTK test performansı ile yetenek arasında matematiksel bir fonksiyon tanımlar ve bu fonksiyonla yeteneği tahmin eder. MTK ile ölçekleme yöntemi tüm veri toplama deseninde uygulanabilmektedir. MTK'da ölçekleme farklı sınıf düzeylerinin parametrelerinin ayrı ayrı ya da birlikte kalibre edilebilmektedir. Bunun yanı sıra parametreleri kestirmek için kullanılan yöntemler de ölçekleme sonuçlarını etkilemektedir. Bu araştırmada MTK ile ölçekleme yöntemi kullanılmıştır.

1.2.1. Kalibrasyon yöntemleri

Dönüşüm yoluyla parametre tahminlerini aynı ölçeğe yerleştirme süreci kalibrasyon olarak adlandırılmaktadır (Kolen,2004). Farklı gruplardan elde edilen verileri ortak bir ölçeğe yerleştirmek için çeşitli kalibrasyon yöntemleri kullanılmaktadır. Bunların içinde en yaygın olarak kullanılanları ayrı kalibrasyon ve eş zamanlı kalibrasyon yöntemleridir (Meng, 2007).

Ayrı kalibrasyon: Ayrı kalibrasyonda her sınıf düzeyi için madde ve yetenek kestirimleri ayrı ayrı hesaplanır. Sınıflardan biri başlangıç ölçeği olarak alınır ve daha sonra bütün sınıf düzeyleri için yapılan parametre tahminleri başlangıç ölçeğine dönüştürülür. Ölçek dönüşümü, ortak maddeler için iki grupta elde edilen parametre tahminleri aracılığı ile gerçekleştirilir.

Ayrı kalibrasyon sonuçlarını ortak bir ölçeğe yerleştirmek için moment yöntemleri ve karakteristik eğri yöntemlerinden biri kullanılır. Moment yöntemleri ortalama/ortalama yöntemi (MM) (Lyod ve Hover, 1980), ortalama/sigma yöntemi (MS) (Macro, 1977) dir. Karakteristik eğri yöntemleri Stocking-Lord (SL) (Stocking ve Lord, 1983) ve Haebara (Meng, 2007) yöntemleridir. Bu çalışmada ayrı kalibrasyon yöntemlerinden Stocking ve Lord (SL) kullanılmıştır. Bu yöntem, aşağıdaki eşitlik yoluyla puanları aynı ölçeğe yerleştirir.

$$SLDiff(\theta_i) = \left[\sum_{j:V} p_{ij}(\theta_{Ji}; \hat{a}_{Jj}, \hat{b}_{Jj}, \hat{c}_{Jj}) - \sum_{j:V} p_{ij}(\theta_{Ji}; \frac{\hat{a}_{Jj}}{A}, A\hat{b}_{Jj} + B, \hat{c}_{Jj}) \right]^2$$

Aşağıdaki ölçütü en küçük yapan A ve B kombinasyonu bulunarak kestirim sürdürülür (Kolen ve Brennan, 2004).

$$SLcrit = \sum_i SLDiff(\theta_i)$$

Eşitlikteki A ve B katsayılarının çözümü yoğun bir iteratif süreci gerektirir.

Eş zamanlı kalibrasyon: Eş zamanlı kalibrasyonda bütün sınıf düzeylerindeki verinin aynı anda kalibre edilmesiyle dikey ölçek oluşturulmaktadır (Meng, 2007). Bu yöntemde tüm parametreler aynı anda birlikte ölçeklendiğinden parametreler aynı ölçek üzerinde yer almaktadır. Bu nedenle parametre dönüşümlerine ihtiyaç duyulmamaktadır. (Kim, Lee, Kim ve Kelly, 2009).

1.2.3. Yetenek kestirimi

MTK'ya dayalı modellerde yetenek ve madde parametrelerinin kestirimi için faklı yöntemler kullanılır. Bu araştırmada yetenek kestiriminde yaygın olarak kullanılan maksimum olabilirlik (Maksimum Likelihood) (MO), maximum a posteriori (MAP) ve expected a posteriori (EAP) kullanılmış veher üç yöntemle kestirilen dikey eşitleme sonuçları karşılaştırılmıştır.

1.3. Değerlendirme ölçütleri

Ölçekleme sonuçlarını karşılaştıracak mutlak bir ölçüt olmadığından dikey ölçeklerin özellikleri kendi içlerinde karşılaştırılmaktadır. Dikey ölçeklemede sonuçları bir yıldan diğer yıla ne kadar büyüme olduğuna, aynı sınıf düzeyinde sınıflar içindeki dağılımın zamanla nasıl değiştiğine ve sınıf düzeyleri arasındaki dağılıma bakarak değerlendirilir. Değerlendirmede ortalama, standart sapma, etki büyüklüğü ve yatay uzaklık gibi istatistikler kullanılmaktadır (Tong ve Kolen, 2007).

Bir sınıf düzeyinden diğer sınıf düzeyine olan büyüme: Burada ard arda gelen sınıf düzeylerinin ortalamaları arasındaki farklar karşılaştırılır. Ölçekleri karşılaştırmak için kesin bir ölçüt yoktur ancak ortalama tahminlerinin konu alanı, kalibrasyon yöntemleri ve yetenek kestirimlerine bakılmaksızın artması beklenmektedir.

Sınıf düzeylerinin değişkenliği: Sınıf düzeyleri arttıkça her sınıf düzeyinin kendi içerisindeki değişkenliği ele alınır. Bunun için her sınıf düzeyinde dikey ölçeklerden elde edilen standart sapmalardır karşılaştırılır. Bu çalışmada her sınıf düzeyi için farklı olarak geliştirilen dikey ölçeklerden hesaplanan standart sapmaların, kalibrasyon metotları ve yetenek kestirimlerine bağlı olup olmadığına bakılmıştır. *Sınıf dağılımları arasındaki ayrım:* Ard arda

gelen sınıf düzeyleri arasındaki ölçek puan dağılımlarının üst üste gelme derecesidir. Bunu hesaplayan indekslerden biri etki büyüklüğü, diğeri ise yatay uzaklıktır (Kim, 2007).

Etki Büyüklüğü: Yen (1986) sınıflar arasındaki olası değişkenlik farklılıklarını göz önüne alarak sınıf dağılımları arasındaki ayrışmayı değerlendirmek için etki büyüklüğü indekslerini kullanıştır. Etki büyüklüğü iki farklı düzeyin ortalamaları ve varyansları kullanılarak kestirilir.

$$\text{Etki büyüklüğü} = \frac{\overline{x}_{üst} - \overline{x}_{alt}}{\sqrt{S^2_{üst} + S^2_{alt}}}$$

Etki büyüklüğü ne kadar büyük ise başarıdaki artış o kadar fazladır ve sınıf düzeyleri arasında daha fazla ayrım vardır.

Yatay Uzaklık: Yatay uzaklık, iki puan dağılımının aynı yüzdelik dilimlere karşılık gelen yüzdelerin farkı olarak tanımlanmaktadır (Holland, 2002). Fark ne kadar büyük ise bir düzeyden diğer düzeye o kadar büyüme olmuş demektir. Yüzdelik dilimlerin karşılaştırması ayrıca, büyümenin hangi düzeydeki öğrencilerde daha fazla, hangi düzeydeki öğrencilerde daha az olduğunu da belirlemeyi sağlamaktadır.

Yüzdelik dilimler arasındaki yatay uzaklık şöyle hesaplanır:

$$D(p) = Y(p) - X(p)$$

D(p): İki grubun p. yüzdelik dilimleri arasındaki yatay uzaklık
Y(p): Birinci grupta p. yüzdelik dilimde yer alanların yüzdesi
X(p): İkinci grupta p. yüzdelik dilimde yer alanların yüzdesi

Sınıf düzeyleri arasındaki değişkenliği, sınıf dağılımları arasındaki ayrımı ölçekler açısından değerlendirirken herhangi bir değer veya standart yoktur. Ancak sınıf düzeylerine göre standart sapmalar arasındaki farkın, etki büyükleri arasındaki fark ve yatay uzaklıklar veya yüzdelik dilimler arasındaki farkın sınıf düzeylerindeki değerleri arasındaki fark on kat veya daha fazla olması, on kat ve ya daha az olması ölçeğin iyi işlemediğini gösterir (Kim,2007).

1.4. Araştırmanın Amacı ve Önemi

Herhangi bir konu alanında hem sınıf düzeyinde hem de bireysel olarak öğrencilerin gelişimlerinin belirlenmesi öğretimin etkililiğini sağlamada ve öğrencilere etkili rehberlik hizmeti sunmada büyük önem taşır. Bu çalışmada odak noktası büyüme modeli bağlamında elde edilen ölçekleme özellikleridir. Büyüme modellerinin altında yatan varsayım ise matematik, okuma vb. gibi herhangi bir konu alanında küçük sınıflardan elde edilen test puanları ile daha sonraki sınıf düzeylerinden elde edilen test puanlarının kesin ve belli bir ölçekle karşılaştırılabileceğidir. Birçok büyüme modeli test puanlarının dikey olarak ölçeklenebileceğini varsayar (Briggs ve Weeks, 2009). Öğrencilerin başarılarındaki artış testlerden elde edilen ölçek puanlarının farklı sınıf düzeyleri arasında karşılaştırılmasıyla belirlenebilmektedir. Sınıf ve öğrenci düzeyindeki gelişim dikey ölçekleme ile belirlenebilir. Bu karşılaştırmanın yapılabilmesi zorluk düzeyleri farklı olan testlere farklı sınıf seviyesindeki öğrencilerin verdikleri cevaplardan elde edilen puanların aynı ölçeğe yerleştirilmesini gerektirmektedir. Bu süreç dikey ölçeklemedir ve karmaşık bir süreçtir. Ölçekleme süreci boyunca kullanılacak ölçekleme deseni, ölçekleme yöntemi gibi birçok kararların verilmesi gerekmektedir. Farklı kararlar farklı ölçeklerin ortaya çıkmasına neden olmaktadır. Alan yazında hangi yöntemin ya da yöntemlerin öğrencilerin başarılarındaki artışı en iyi ve doğru ortaya koyduğu konusunda bir uzlaşma yoktur. Buna rağmen dikey ölçekleme birçok test geliştiricisi tarafından kullanılmaktadır. Ancak her test geliştirici geliştirdiği ölçek için dikey ölçek geliştirme süreçlerini kendisi belirlemektedir (Tong ve Kolen, 2007). Yapılan bu çalışmanın amacı farklı yaklaşımlar kullanılarak elde edilen dikey ölçekleme sonuçlarının bir fonksiyonu olarak elde edilen büyüme örüntülerinin deneysel olarak

karşılaştırılmasıdır. Bu nedenle ayrı ve eş zamanlı kalibrasyon yöntemleriyle, ve MTK yetenek tahminlerinin (MO, EAP ve MAP) kombinasyonlarıyla dikey ölçekler geliştirilmesi ve bu ölçekleme sonucunda elde edilen sonuçların büyüme modeli çerçevesinde karşılaştırılmasıdır.

Bu çalışmada farklı eğitim seviyelerindeki hedefleri kapsamak ve böylece oluşturulan testlerden elde edilen puanları aynı ölçeğe yerleştirip karşılaştırma yapabilmek için testler geliştirilmiştir. Oluşturulan bu testler yardımıyla elde edilen veriler kullanılarak parametre tahminleri aynı ölçeğe yerleştirildikten sonra analizler yapılmıştır. Bu çalışmada, 6., 7. ve 8. Sınıf öğrencilerinin matematik dersindeki gelişimlerini görmek amacıyla yapılan dikey ölçeklemenin hangi koşullarda nasıl sonuçlar verdiğini ortaya koymaktadır. Öğrencilerin bir alanda yıllara göre gelişimlerini izlemek için puanların aynı ölçekle ifade edilmesi gerekir. Farklı zamanlarda ve farklı sınavlardan alınan puanların doğrudan karşılaştırılması anlamlı olmadığından, öğrencinin gelişimini göstermez. Başarının izlenmesi, öğrenciyi daha iyi tanımayı ve etkili bir dönüt vermeyi sağladığı gibi öğretim programlarının ve öğrenme ortamlarının değerlendirilmesinde önemli bilgiler verir. Bu nedenle bu araştırma dikey ölçekleme konusunda da bir örnek sunması bakımından önem taşımaktadır.

2. YÖNTEM

2.1. Çalışma Grubu

Bu araştırmanın çalışma grubunu 2011-2012 öğretim yılında 6., 7. ve 8. Sınıflarda öğrenim gören öğrenciler oluşturmaktadır. Çalışmaya Ankara'da bulunan beş farklı ilköğretim okulundan toplam 1504 öğrenci katılmıştır. Testler 6. sınıfta 229 erkek, 245 kız öğrenci ve 24 de cinsiyetini belirtmeyen toplamda 503 (%33) öğrenciye, 7. sınıfta 207 erkek, 242 kız ve 52'de cinsiyetini belirtmeyen toplamda 501 (%33) öğrenciye, 8. sınıfta ise 221 erkek, 175 kız ve cinsiyetini belirtmeyen 103 toplamda 500 (%33) öğrenciye uygulanmıştır.

2.2. Araştırma Deseni

Bu araştırmada eşitleme desenlerinden, ortak madde deseni kullanılmıştır. Her testte ait olduğu sınıf düzeyine uygun maddeler ve farklı sınıf düzeyindeki testlerde de bulunan ortak maddeler bulunmaktadır. Örneğin 7. sınıf testinde 7. sınıf düzeyindeki maddelerin yanı sıra 6. sınıf testinde de yer alan ortak maddeler bulunmaktadır. Ortak maddelerde gösterilen performans bir sınıf düzeyinden diğer sınıf düzeyine ne kadar gelişme olduğunu belirlemede kullanılmaktadır. Kolen (2004) ortak madde sayısının testteki madde sayısının %20'si kadar olmasının yeterli olduğunu belirtirken Lorie ve Yao (2005) yaptıkları çalışmada ortak madde sayısındaki artışın testteki ölçmenin standart hatasını azalttığını bulmuşlardır. Kim, Lee, Kim ve Kelley (2009) tarafından Rasch modelde dikey ölçekleme yapılırken ortak madde sayısının tüm madde sayısının %25'inden fazla olarak alınması önerilmektedirler. Bu çalışmada ortak madde sayısı her üç testte de testteki toplam madde sayısının %25'inden fazladır.

Aşağıdaki tabloda ortak ve ortak olmayan maddelerin testlere dağılımlarını göstermektedir.

Tablo 1: Ortak ve ortak olmayan maddelerin testlere dağılımı

		Ortak maddelerin düzeylere göre dağılımı		
Testler	Ortak olmayan maddeler	6.sınıf düzeyi	7. sınıf düzeyi	8.sınıf düzeyi
6. sınıf	6,8,9,10,11,12,13,14, 15,16,17,18,19,20,21	1*,2*,3**	4*,5*,7**	-
7. sınıf	9,10,11,12,13,14,15,16 ,17,18,19,20,21,22,23	1*,2*,4**	5**,6***,7*,8*	3***
8. sınıf	7,8,9,10,11,12,13,14, 15,16,17,18,19,20,21	2*,3*,	1*,5**,6*	4***

*: Her üç testte de ortak olan maddeler
**: 6 ve 7. Sınıf testlerinde ortak olan maddeler
***: 7 ve 8. Sınıf testlerinde ortak olan maddeler

Tabloda da gösterildiği gibi ortak maddelerin 4'ü her üç sınıf düzeyinde de ortaktır. Bu dört ortak maddeye ek olarak iki madde 6 ve 7. sınıf testlerinde, iki madde de 7. ve 8. sınıf testlerinde ortaktır. Arda arda gelen sınıf düzeylerinde aşamalı olarak ortak olan maddelerden, 1 tane 6. sınıf sorusu ve 1 tane 7. sınıf sorusu 6. ve 7. sınıfta ortak; 1 tane 7. sınıf sorusu ve 1 tane 8. sınıf sorusu 7. ve 8. sınıfta ortak olacak şekilde seçilmiştir. Farklı sınıf düzeyleri için hazırlanan testler bu ortak maddeler kullanılarak birbiriyle bağlanmakta, böylece farklı testlerden elde edilen puanlar aynı ölçeğe yerleştirilmektedir. Altıncı sınıf temel sınıf olarak alındığında 6. ve 7. sınıftaki ortak maddeler bu iki sınıf düzeyindeki bağlantıyı sağlamakta, 7. ve 8. sınıftaki ortak maddeler de 7. ve 8. sınıf düzeyleri arasındaki bağlantıyı sağlamaktadır. Böylece 6. ve 8. sınıf arasındaki bağlantı 7. sınıf üzerinden yapılmaktadır. Ortak madde deseninin eleştirilerinden biri olan küçük sınıflardaki öğrencilerin üst sınıflara ait soru cevaplamak zorunda olmasının dezavantajı veya büyük sınıfların alt sınıflara ait soruları cevaplayıp avantajlı olmaları durumuyla baş edebilmek için bu şekilde karma bir desen oluşturulmuştur.

2.3. Veri Toplama Aracının Hazırlanması ve Verilerin Toplanması

Araştırmada kullanılan testler, 2007-2010 yılları arasında 6., 7. ve 8. sınıflara uygulanan SBS sınavlarındaki matematik testlerinden ayırt edicilikleri yüksek olan maddelerden seçilerek oluşturulmuştur. Madde güçlükleri açısından ise testlerde zor ve kolay maddeler bulunmaktadır, ancak testin geneli ortalama zorluktaki maddelerden oluşmaktadır. Test maddeleri seçilirken 6., 7. ve 8. sınıf matematik programlarındaki öğrenme alanları dikkate alınmıştır. Bu öğrenme alanları sayılar, cebir, olasılık ve istatistik, geometri ve ölçmedir (MEB, 2009). 6. ve 8. sınıf testinde 21 madde, 7. sınıf testinde ise 23 madde yer almaktadır. Testler ortak madde desenine göre hazırlandığı için 6. ve 8. sınıf testlerindeki 6 madde, 7. sınıf testindeki 8 madde ortak maddelerdir. Her üç testte de ortak maddeler dışında kalan maddeler ait olunan sınıf düzeyine uygun ve sadece bu sınıf düzeyindeki öğrencilerin cevapladıkları maddelerdir. Yani 6. sınıf testinde 6. sınıf SBS sorularından oluşan ve sadece bu sınıf düzeyindeki öğrencilerin cevapladığı toplam 15 madde ve 6 da ortak madde yer almaktadır. 7. sınıf testinde 7. sınıf SBS sorularından oluşturulan ve sadece bu sınıf düzeyindeki öğrencilerin cevapladığı 15 madde ve 8 ortak madde yer almaktadır. Benzer şekilde 8. sınıf testinde 8. sınıf SBS sorularından oluşturulan ve sadece bu sınıf düzeyindeki öğrencilerin cevaplayacakları 15 madde ve 6 de ortak madde bulunmaktadır. Her üç sınıf düzeyinde de ortak olmayan maddeler 5 sorusu Sayılar, 3 sorusu Cebir, 3 sorusu Olasılık ve İstatistik, 2 sorusu Geometri ve diğer 2 soruda Ölçme öğrenme alanından olacak şekilde seçilmiştir.

Ortak olmayan maddelere benzer şekilde ortak sorular seçilirken de öğrenme alanları ve alt öğrenme alanları dikkate alınmıştır. Ortak soruların 4'ü 6, 7 ve 8. sınıf testlerinin üçünde de bulunmaktadır. Bu dört ortak soruya ek olarak 1 adet 6. sınıf ve 1 adet 7. sınıf sorusu 6. ve 7. sınıf testlerinde ortak; 1 adet 7. sınıf ve 1 adet 8. sınıf sorusu 7. ve 8. sınıf testlerinde ortaktır. Ortak sorular Sayılar, Cebir, Ölçme ve Olasılık ve İstatistik öğrenme alanlarının her birinden 2 tane olacak şekilde seçilmiştir.

2.4. Verilerin Analizi

6., 7. ve 8. sınıf testlerine ait KR-20 güvenirlik katsayısı incelendiğinde, 6. sınıfa ait testin güvenirliği (.73), 7. sınıfa ait testin güvenirliği (.79) ve 8. sınıfa ait testin güvenirliği de (.82) olarak bulunmuştur. Bunlara ek olarak her üç testin ayırıcılık gücü indekslerinin ortalamaları 6, 7 ve 8. sınıf testleri için sırasıyla, (.50), (.47) ve (.56) olduğu için, her üç testin de ayırıcı sonuçlar verdiği ve testlerin ayırıcılık gücünün yüksek olduğu söylenebilir.

Her üç sınıf düzeyi için madde tepki kuramına dayalı yapılan analizler, verilerin 2PL modele uygun olduğunu göstermektedir.

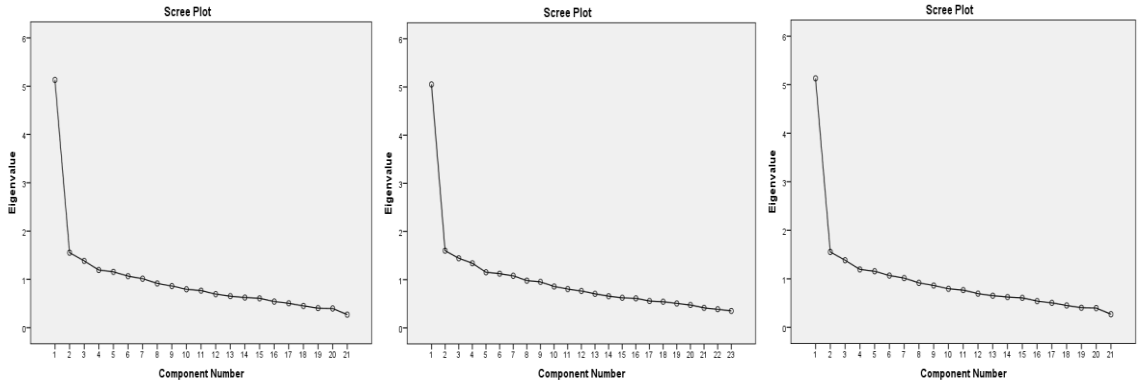
Her üç testteki maddeler BILOG-MG (Zimowski, Muraki, Mislevy, ve Bock, 1996) programı kullanılarak ölçeklenmiştir. Eş zamanlı kalibrasyon için bütün sınıf düzeylerindeki

veriler bir araya getirilmiştir. Bütün tahminler ortak bir ölçek üzerine bir defa BILOG-MG programının çalıştırılmasıyla yerleştirilmektedir. Ayrı kalibrasyonda ise her sınıf düzeyindeki veriler BILOG-MG kullanılarak ayrı olarak kalibre edilmektedir. Daha sonra her sınıf düzeyinin parametre tahminlerini temel olarak kabul edilen ölçeğe yerleştirmek için ST (Hanson ve Zeng. Rev. Cui,2004) programı kullanılarak SL dönüşümünde kullanılan eğim ve kesişim değerleri hesaplanmıştır.

3. BULGULAR

Araştırma sonucunda üç farklı yetenek kestirim yöntemi ile iki farklı kalibrasyon yönteminin kombinasyonuyla elde edilen dikey ölçeklemelerin sonuçlarının öğrencilerin büyüme örüntüsünü nasıl belirledikleri karşılaştırılmıştır.

MTK'nın tek boyutluluk sayıltısı test edilirken her üç test için de ayrı ayrı yapılan açımlayıcı faktör analizinden elde edilen yamaç eğim grafikleri Şekil 1'de aşağıda verilmiştir.



Şekil 1: 6, 7 ve 8. Sınıf verileri için yamaç eğim grafikleri

6,7 ve 8.sınıf verilerinin tek boyutluluk sayıltısını sağladığı Şekil 1'de görülmektedir.

Elde edilen farklı dikey ölçekleme sonuçlarını karşılaştırmak ve değerlendirebilmek için arda arda gelen sınıf düzeyleri arasındaki ortalama θ değerleri, ortalamalar arasındaki fark, standart sapma, etki büyüklüğü ve ortalama yatay uzaklık değerleri ayrı ve eş zamanlı kalibrasyon yöntemleriyle ve MO, EAP ve MAP ile kestirilmiş ve elde edilen sonuçlar Tabloda gösterilmiştir.

Tablo 2: Ardarda gelen sınıf düzeylerinde ortalamalar, ortalamalar arasındaki fark, standart sapma, etki büyüklüğü ve ortalama yatay uzaklıklar

	Sınıf	Ayrı Kalibrasyon			Eş Zamanlı Kalibrasyon		
	Düzeyleri	EAP	MAP	MO	EAP	MAP	MO
Ortalama	6	0.007	-0.020	0.007	0.000	0.000	0.000
	7	-0.204	-0.161	-0.155	-1.859	-2.124	-1.492
	8	0.054	-0.038	0.082	0.338	0.509	0.244
Ortalamalar arası fark	6-7	-0.211	-0.141	-0.148	-1.859	-2.124	-1.492
	7-8	0.258	0.199	0.237	2.197	2.633	1.736
Standart sapma	6	0.893	0.844	1.191	1.000	1.000	1.000
	7	0.251	0.408	0.493	0.700	0.459	0.482
	8	0.702	0.652	0.930	1.115	1.326	1.117
Etki Büyüklükleri	6-7	-0.322	-0.213	-0.178	-2.151	-2.723	-1.900
	7-8	1.345	0.366	0.318	2.360	1.336	2.016
Ortalama Yatay	6-7	-0.23	-0.13	-0.20	-1.82	-2.08	-1.46
Uzaklıklar	7-8	0.29	0.21	0.22	2.25	2.68	1.91

Ortalama θ değerleri ayrı kalibrasyonda 6. ve 8. sınıflarda MAP kestirimi MO ve EAP'a göre daha düşük sonuçlar verirken EAP ile MO karşılaştırıldığında EAP ile hesaplanan değerlerin daha düşük olduğu görülmektedir. 7. sınıfta ise MO en büyük değeri alırken EAP en düşük ortalama θ değerine sahiptir. Eş zamanlı kalibrasyonda 6. sınıf referans sınıf olarak alındığı için her üç kalibrasyon yönteminde de ortalama θ değerleri sıfırdır. 8. sınıf düzeyinde MAP ile kestirilen 0.509 değeri en yüksek ortalama θ değeri olarak bulunurken 7. sınıf düzeyinde MAP ile kestirilen -2.124 en düşük değer olarak bulunmuştur. Farklı sınıf düzeyleri için yapılan dikey ölçeklemeler sonucunda elde edilen ortalama θ değerleri kalibrasyon yöntemleri dikkate alınarak incelendiğinde; sınıf düzeyleri için farklı yetenek kestiricileri ve kalibrasyon yöntemleriyle elde edilen dikey ölçekleme sonuçlarına göre genel olarak en düşük ortalamaya sahip sınıfın 7. sınıf olduğu görülmektedir. Tablo incelendiğinde her üç yetenek kestirim yönteminde de ortalama θ değerlerinin 6. sınıf düzeyinden 7. sınıf düzeyine azalırken 8. sınıf düzeyinde ise tekrar artmakta olduğu görülmektedir. Ortalama θ değerleri 6. sınıf düzeyinde MAP ile ayrı kalibrasyon yöntemi kombinasyonundan elde edilen değer dışında diğer yetenek kestirim ve kalibrasyon yöntemlerinde pozitif değerler almaktadır. 7. sınıf düzeyi için hesaplanan ortalama θ değerlerinin hepsi her iki kalibrasyon yöntemi ve her üç yetenek kestirimi sonucunda negatif olarak bulunmuştur. 8. sınıfta ise MAP kestirimiyle ayrı kalibre edilerek elde edilen ortalama -0.038 değerleri dışındaki diğer θ değerleri pozitif değerler almaktadır.

Sınıf düzeyleri arasındaki büyüme örüntüsünü belirlemek için ardışık sınıf düzeyleri arasındaki ortalama farkları incelendiğinde 6. ve 7. sınıflar arasındaki ortalama θ farkları (-2.124 ile -0.141) arasında değerler almakta iken 7. ve 8. sınıflar arasında hesaplanan ortalama θ farkları ise (0.199 ile 2.633) arasındadır. En yüksek ortalama farkı MAP ve eş zamanlı kalibrasyon sonucunda 7. ve 8. sınıflar arasında 2.633 olarak hesaplanırken en düşük ortalama θ farkı MAP ve ayrı kalibrasyon ile 6. ve 7. sınıflar arasında -0.141 olarak hesaplanmıştır. Ortalama θ farkları eş zamanlı kalibrasyonda daha büyük sonuçlar vermektedir. Tablo 2 incelendiğinde ayrı kalibrasyon yöntemi ve MO, EAP ve MAP ile kestirilen 6. ve 7. sınıflar arasındaki ortalama θ farkları negatif iken 7. ve 8. sınıf düzeylerinde ise pozitif olduğu görülmektedir.

Sınıf düzeyleri arasındaki benzerlik veya farklılıkları gösteren standart sapma değerleri incelendiğinde 6. sınıftan 7. sınıfa azalmakta, 8. sınıfta ise tekrar artmaktadır En yüksek standart sapma değeri MAP ve eş zamanlı kalibrasyon kombinasyonu sonucu 8. sınıf düzeyinde 1.326 olarak hesaplanırken en düşük standart sapma değeri ise EAP ve ayrı kalibrasyon sonucunda 7. sınıf düzeyinde 0.251 olarak hesaplanmıştır. Kalibrasyon yöntemlerine göre karşılaştırıldığında ayrı kalibrasyonda EAP ve MAP eş zamanlı kalibrasyona kıyasla daha düşük standart sapma değerleri üretirken MO (8. sınıf düzeyi hariç) daha yüksek standart sapma değeri üretmektedir. . Ayrı kalibrasyonda 6. ve 8. sınıf düzeyinde de MAP kestirimi en düşük değerleri verirken 7. sınıf düzeyinde en düşük değeri EAP vermektedir. En yüksek değerlerin her üç sınıf düzeyinde de MO kestirimi ile hesaplanan değerler olduğu görülmektedir. Eş zamanlı kalibrasyon yöntemi ve MO, EAP ve MAP yetenek kestirim yöntemleri kombinasyonunda standart sapmaların nasıl değiştiği incelendiği zaman 6. sınıf ortalaması 0 standart sapması 1 olan referans grup olduğu için 6. sınıfın standart sapma değerleri her üç yetenek kestirim yöntemi içinde 1 olarak alınmıştır. Eş zamanlı kalibrasyonda 7. sınıf düzeyinde MAP ve MO yöntemleriyle hesaplanan standart sapma değerleri birbirine yakın sonuçlar verirken, EAP ile daha büyük standart sapma değeri elde edilmektedir. 8. sınıfta en düşük standart sapma değerini EAP alırken en büyük değeri MAP almaktadır.

Yetenek kestirimlerinin sınıf düzeyleri arasındaki ayrımı ya da benzerliği etki büyüklüğü ve yatay uzaklık birimleri ile analiz edilmiştir. Etki büyüklükleri incelendiğinde 6. ve 7. sınıflar -2.151 ile -0.178 arasında değerler almakta iken 7 ve 8. sınıflar ise 0.318 ile 2.360 arasında değerler almaktadır. Ard arda gelen sınıf düzeyleri arasındaki etki büyüklüklerinin her iki kalibrasyon yönteminde de 6 ve 7. sınıflar arasında negatif değerler alırken 7 ve 8. sınıflar arasında ise pozitif değerler aldığı görülmektedir. Eş zamanlı kalibrasyon yönteminde 6 ve 7.

sınıflar arasında MAP kestirimiyle hesaplanan etki büyüklükleri EAP ve MO kestirimlerinden elde edilen değerlere kıyasla daha büyük değerler vermektedir. 6 ve 7. sınıflar arasında en büyük etki büyüklüğü MAP ile hesaplanan değer olurken 7 ve 8. sınıflar arsında da en büyük değerin EAP ile hesaplanan değer olduğu tablodan görülmektedir.

Ortalama yatay uzaklıklar ise 6 ve 7. sınıflar arasında (-2.08 ile -0.13) arasında değerler alırken 7 ve 8. sınıflar arasında ise (0.22 ile 2.68) arasında değerler almaktadır. Eş zamanlı kalibrasyondaki değerler ayrı kalibrasyona göre daha büyüktür. Ayrı kalibrasyonda en düşük ortalama yatay uzaklık değerini 6. ve 7. sınıflar arasında MAP verirken en yüksek değeri ise EAP vermektedir. 7. ve 8. sınıflar arasında da benzer bir durum söz konusudur. Eş zamanlı kalibrasyonda ortalama yatay uzaklıklar MAP ile en büyük değerler alırken en küçük değerleri MO vermektedir.

4. TARTIŞMA ve SONUÇ

Farklı MTK modeli, bağlama yöntemi ve yetenek kestirim yaklaşımıyla dikey ölçek esnemekte yada daha da daralmaktadır. Buna ek olarak aynı soruya aynı test içinde aynı öğrenci grubunun verdiği cevaplarla büyüme örüntüsünün kesin olarak belirlenmesi seçilen modelden etkilenir. iki kalibrasyon yönteminde de 6. sınıftan 7. sınıfa geçildiğinde ortalamaların düştüğü, 8. sınıfta ise tekrar arttığı görülmektedir. Benzer sonuçlar Kolen ve Tong (2008), Boughton, Lorié ve Yao (2005) ve Tong ve Kolen (2007) tarafından yapılan çalışmalarda da görülmüştür. Kolen ve Tong'un (2008) çalışmalarında İngilizce testinde genel olarak sınıf düzeyi arttıkça diğer sınıflarda ortalamalarda artmıştır. Bu artma 6. ve 7. sınıflar arasında gözlenmemiş ve 6. sınıfın ortalaması 7. sınıfın ortalamasından daha büyük olarak bulunmuştur. Boughton, Lorié ve Yao'nun (2005) çalışmalarında ise 7. sınıfın 8 ve 9. sınıflara göre daha başarılı olduğu bulunmuştur. Ayrıca Tong ve Kolen'in (2007) çalışmasında MTK ile ölçeklemede öğrencilerin başarılarındaki dağılım sınıf düzeyleri boyunca kararsızlık göstermekte veya azalmakta olduğu bulunmuştur

Ortalamalar arasındaki farklar incelendiği zaman ayrı kalibrasyonda $EAP \rangle MO \rangle MAP$ şeklinde iken eş zamanlı kalibrasyonda ise $MAP \rangle EAP \rangle MO$ şeklindedir ve eş zamanlı kalibrasyondaki ortalama farkları ayrı kalibrasyona göre daha büyüktür. Her iki kalibrasyon yönteminde de MO, EAP ve MAP ile kestirilen 6. ve 7. sınıflar arasındaki ortalama θ farkları negatif iken 7. ve 8. sınıf düzeylerinde ise pozitif olduğu görülmektedir. Bu durumda 6 ve 7. sınıflar arasında başarıda azalma olduğu yani 6. sınıf öğrencilerinin 7. sınıf öğrencilerine göre daha başarılı olduğu şeklinde yorumlanabilir. 8. sınıf ile 7. sınıf arasında ortalamalar arasındaki farkın pozitif olması ise 7. sınıftan 8. sınıfa geçince öğrencilerin başarılarının arttığını belirtmektedir. Bu sonuçtan farklı olarak Kolen ve Tong (2010) sınıf düzeyleri arasındaki ortalama farkların düşük sınıflarda fazla, sınıf düzeyi arttıkça ise azalmakta olduğunu bulmuşlar ve bu durumu düşük sınıf düzeylerinde daha fazla büyüme olduğu şeklinde yorumlamışlardır.
Yaptığı çalışmada Kim (2007) küçük sınıflarda eş zamanlı kalibrasyonun ayrı kalibrasyona göre daha düşük ortalama farkı verdiğini bulurken büyük sınıflarda eş zamanlı kalibrasyonun hesaplanamadığını belirtmiştir. Bu bulgulardan farklı olarak matematik testinde 5. sınıftan 10. sınıfa kadar yapılan dikey eşitlemede ayrı kalibrasyonun daha düşük sonuçlar verdiği bulunmuştur (Karkee et al., 2003). Kim, Lee, Kim ve Kelly (2009) ortalamalar arasındaki farkın yetenek kestiriminden etkilenmezken kalibrasyon yönteminden etkilendiğini bulmuşlardır. Sınıf düzeyleri arttıkça ortalamalar artmaktadır ve yetenek kestirim yöntemlerinde benzer sonuçlar vermektedir (Kolen ve Tong, 2010). Bu bulgulardan farklı olarak Chin, Kim ve Nering (2006) tarafından yapılan çalışmada eş zamanlı kalibrasyonda ortalamalar arasındaki farklar için artma ya da azalma örüntüsü gözlenmemiştir. Ortalamalar arasındaki farklar her iki kalibrasyon yönteminde ve her üç yetenek kestirim yönteminde aynı örüntüyü vermektedir.

Standart sapma değerleri incelendiği zaman ise ayrı kalibrasyonda 6.ve 8. sınıflarda: MO⟩ EAP⟩ MAP şeklinde iken 7. sınıfta ise MO⟩ MAP⟩ EAP şeklinde, ve eş zamanlı kalibrasyonda ise 7. sınıfta: EAP⟩ MO⟩ MAP şeklinde iken 8. sınıfta ise MAP ⟩ MO⟩ EAP şeklinde olduğu görülmektedir. Buna ek olarak 6. sınıf MO ve 7. sınıf ve MO kombinasyonu dışında eş zamanlı kalibrasyon sonucu elde edilen değerler ayrı kalibrasyondaki değerlerden daha büyüktür. Standart sapmalar 6. sınıftan 7. sınıfa azalmakta, 8. Sınıfta ise tekrar artmaktadır. Bir başka ifadeyle her iki kalibrasyon yöntemi ve her üç yetenek kestiriminde en homojen sınıf 7. sınıftır. Bu sonuca benzer şekilde Kolen ve Tong (2007) yaptıkları çalışmalarda ayrı kalibrasyonda standart sapmaların artıp azalmakta olduğunu bulmuşlardır. Ancak bu durum Chin, Kim,ve Nering'in (2006) standart sapmaların sınıf düzeyi arttıkça azalma eğiliminde olduğu bulgusuyla çelişmektedir. Bu bulgulardan farklı olarak bir başka çalışmada standart sapmaların azalıp daha sonra düz bir çizgi izlemesi puan dağılımlarının sınıf düzeyleri arasındaki ayrımının açıkça sabit kaldığının ve bu durumun başarılı ve başarısız öğrencilerin sınıf düzeyleri arttıkça birbirleriyle neredeyse aynı olduğunun göstergesi olarak kabul edilmiştir (Tong ve Kolen, 2008). Standart sapma kriterinin kalibrasyon yönteminden değil de yetenek kestirim yöntemlerinden daha çok etkilendiği belirtilmektedir (Kim, Lee, Kim ve Kelly, 2009; Kolen ve Tong, 2010; Kim, 2007). Yaptığı çalışmada Kim (2007) matematik testinde MO'nun EAP'a göre daha büyük standart sapma değerleri ürettiğini bulmuştur. Bu çalışmada sınıf düzeyleri arasındaki standart sapma değerleri arasındaki farkların 10 kat veya daha fazla olmaması ölçeklerin iyi işlediğini göstermektedir. Standart sapma değerleri ayrı ve eş zamanlı kalibrasyonda tüm yetenek kestirim yöntemleri için aynı örüntüyü göstermektedir (6'sınıftan 7'ye geçişte azalmakta, 8. Sınıfa geçişte ise artmaktadır). Her iki kalibrasyon yönteminde sınıf düzeyleri arasında standart sapma değerleri arasındaki artma veya azalma miktarları dikkate alındığında MO kestiriminin diğer iki kestirim yöntemine göre daha yakın sonuçlar ürettiği görülmektedir.

Yetenek kestirimlerinin sınıf düzeyleri arasındaki ayrımı ya da benzerliği etki büyüklüğü ve yatay uzaklık birimleri ile analiz edilmiştir. Etki büyüklüğü standardize edilmiş ortalama farklarıdır. Etki büyüklüğünün küçük olması ard arda gelen sınıf düzeyleri arasında yetenek düzeyleri arasındaki farkın az olduğunun göstergesidir. Ard arda gelen sınıf düzeyleri arasındaki etki büyüklüklerinin her iki kalibrasyon yönteminde de 6 ve 7. sınıflar arasında negatif değerler alırken 7 ve 8. sınıflar arasında ise pozitif değerler aldığı görülmektedir. Bu durum 6 ve 7. sınıflar arasında yetenek düzeyleri arasındaki fark 6. sınıf lehine iken 7 ve 8. sınıf düzeylerinde ise farkın 8. sınıf lehine arttığını göstermektedir. Etki büyüklüğünün küçük olması ard arda gelen sınıf düzeyleri arasında yetenek düzeyleri arasındaki farkın az olduğunun göstergesidir. Bu nedenle yapılan çalışmada artan bir büyüme örüntüsü olduğu söylenebilir. Ayrı kalibrasyonda etki büyüklükleri EAP⟩ MAP⟩ MO şeklinde değerler alırken eş zamanlı kalibrasyonda 6. ve 7. sınıflar arasında MAP⟩ EAP⟩ MO şeklinde, 7. ve 8. sınıflar arasında ise EAP⟩ MO⟩ MAP şeklinde değerler almaktadır. Alan yazında etki büyüklükleriyle ilgili farklı bulgulara rastlanmıştır. EAP ile kestirilen etki büyüklüklerinin diğer yetenek kestirimlerine göre daha büyük olduğu bulunmuştur (Kim, 2007). Bir başka çalışmada etki büyüklüklerinin ayrı ve eş zamanlı kalibrasyon yöntemlerinde benzer sonuçlar verdiği görülmüştür. Ancak yetenek kestirim yöntemleri dikkate alındığında EAP ve MAP ile hesaplanan etki büyüklüklerinin MO ve test karakteristik eğrisi yöntemleriyle hesaplanan değerlere göre daha büyük olduğu belirtilmiştir. Buna ek olarak etki büyüklüğünün küçük sınıflarda yüksek, büyük sınıflarda düşük olduğu ve bunun azalan bir büyüme örüntüsünün göstergesi olduğunu belirtilmiştir (Tong ve Kolen, 2010). Benzer şekilde bir diğer çalışmada etki büyüklüklerinin sınıf düzeyleri arttıkça azaldığı görülmüştür (Tong & Kolen, 2008). Ayrı kalibrasyonda her üç yetenek kestirim yönteminde de benzer örüntü görülmektedir. Kolen ve Tong (2007) çalışmalarında etki büyüklüğünün ayrı kalibrasyonda genellikle EAP'ın etki büyüklüğünün MO ve MAP'dan daha küçük olduğunu bulmuşlardır. Bu çalışmada da EAP ile hesaplanan etki büyüklükleri MAP ve MO ile

hesaplanan etki büyüklüklerinden daha büyüktür. Bu sonuç Beard (2008)'in simülasyon verileri kullanarak yaptığı çalışma sonucu elde ettiği bulguyla örtüşmektedir. Bu çalışmada sınıf düzeyleri arasındaki etki büyüklükleri arasındaki farkların 10 kat veya daha fazla olmaması ölçeklerin iyi işlediğini göstermektedir. Etki büyüklükleri ayrı ve eş zamanlı kalibrasyonda ve MO, EAP ve MAP için 6'sınıftan 7'ye geçişte azalan, 8. sınıfa geçişte ise artacak şekilde aynı örüntüyü göstermekte. Her iki kalibrasyon yönteminde etki büyüklükleri eş zamanlı kalibrasyonda daha büyük ayrı kalibrasyonda ise küçük değerler almaktadır.

Ortalama yatay uzaklıklar incelendiği zaman ayrı kalibrasyonda $EAP \rangle MO \rangle MAP$ şeklinde değerler alırken eş zamanlı kalibrasyonda $MAP \rangle EAP \rangle MO$ şeklinde değerler almaktadır. Sınıf düzeyleri arttıkça ortalama yatay uzaklıklar artmaktadır. Kim (2007) çalışmasında ise sınıf düzeyi arttıkça yatay uzaklıkların azalmakta olduğunu bulunmuştur.

4.1. Öneriler

Öğrencilerin anasınıfından 12. Sınıfa kadar olan süreçte başarılarındaki değişimlerinin izlenmesi isteniyorsa bunun için dikey ölçekleme yönteminin kullanılması kaçınılmazdır. Ancak dikey ölçekleme karmaşık bir süreçtir ve bu süreç boyunca yapılan seçimler ölçeklemenin sonucunu yorumlamayı bir diğer ifadeyle öğrencilerin başarılarındaki değişimi yorumlamayı da etkilemektedir. Farklı kalibrasyon yöntemleri ve yetenek kestirim yöntemleriyle yapılan dikey ölçekleme sonuçları aynı örüntüyü verse de büyüme miktarındaki değişimi farklı olarak tahmin etmektedir. Dolayısıyla dikey ölçekleme sonucunda öğrencilerin başarıları hakkında önemli kararlar verilecekse farklı ölçme sonuçlarından elde edilen bulguları da dikkate almak gerekmektedir. Bu çalışma gerçek veri seti ile kalibrasyon yöntemlerinin karşılaştırılmasını içermektedir. Bu nedenle bu ölçeklerin tam anlamıyla karşılaştırılması mümkün değildir. Gerçek puan değerlerinin bilinebileceği bir simülasyon çalışmasıyla bu karşılaştırma daha doğru bir şekilde yapılabilir. Bu çalışmada sadece çoktan seçmeli maddeler kullanılmıştır. Farklı soru formatlarının farklı kalibrasyon yöntemleri veya farklı yetenek kestirim yöntemleriyle elde edilen sonuçları nasıl etkilediği bir başka araştırmanın konusu olabilir.

5. KAYNAKLAR

Beard, J. J.(2008). An Investigation of vertical scaling with item response theory using a multistage testing framework. Yayınlanmamış Doktora Tez. University of Iowa, Iowa.

Boughton, K.A., Lorie, W. & Yao, L. (2005). *A Multidimensional Multigrup IRT Models for Vertical Scales with Complex Test Structure: An Emprical Evaluation of Student Growth using Real Data*. National Council on Measurement in Education: 2005. Monreal/ Quebec/ Canada.

Briggs, D.C. & Weeks, J.P. (2009). The İmpact of Vertical Scaling Decisions on Growth Interpretations. *Educational Measurments 28(4), 3-14*

Burg, S. (2008). An investigation of dimensionality across grade levels and effects on vertical linking for elementary gerade mathematics achievement tests. NCME: NYC, 2008.

Chin, T.Y. , Kim,W. & Nering, M. L. (2006). *Five Statistical Factors That Influence IRT Vertical Scaling*. Paper presented at the annual meeting of National Council on Measurement in Education (NCME) at San Francisco, April 2006.

Hanson, B.& Zeng, L.(Rev. Cui,Z 2004). *ST: A Computer Program for IRT Scale Transformation*.

Holland, P. W. (2002). Two measures of changes in the gaps between the CDFs of test score distributions. *Journal of Educational and Behavioral Statistics*, 27, 3- 17.

Karkee,T.; Lewis, D.M.; Hoskens, M.; Yao, L.& Haug, C. (2003). *Seperate vs Concurrente Calibration Methods in Vertical Scaling*. Paper presented at the annual meeting of National Council on Measurement in Education .Chicago, IL, April 22-24, 2003.

Kim, J. (2007). A comparison of calibration methods and proficiency estimators for creating IRT vertical scales. (Yayınlanmamış Doktora Tezi), University of Iowa, Iowa.

Kim,J., Lee, W.C., Kim,D. & Kelley, K. (2009). *Investigation of Vertical Scaling Using the Rasch Model*. National Council on Measurement in Education: April 2009.

Kolen, J. M. (2004). Linking Assessments: Concept and History. *Applied Psychological Measurement,* 28(4), 219-226.

Kolen M.. J. & Tong, Y. (2010). Psychometric Properties of IRT Proficiency Estimates. *Educational Measurement Issues and Practice,* 29(3), 8-14.

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

Macro, G. L. (1977). Item Characteristic Curve Solutions to Three Intractable Testing Problems. *Journal of Educational Measurement*, 14 (2), 139- 160

MEB. (2009). İlköğretim Matematik Dersi 6-8. Sınıflar Öğretim Programı. Ankara, 2009.

MEB (2008a). 6. Sınıf seviye belirleme sınavı soru kitapçığı.

MEB (2008b). 7. Sınıf seviye belirleme sınavı soru kitapçığı.

MEB (2008c). 8. Sınıf seviye belirleme sınavı soru kitapçığı.

MEB (2009a). 6. Sınıf seviye belirleme sınavı soru kitapçığı.

MEB (2009b). 7. Sınıf seviye belirleme sınavı soru kitapçığı.

MEB (2009c). 8. Sınıf seviye belirleme sınavı soru kitapçığı.

MEB (2010a). 6. Sınıf seviye belirleme sınavı soru kitapçığı.

MEB (2010b). 7. Sınıf seviye belirleme sınavı soru kitapçığı.

MEB (2010c). 8. Sınıf seviye belirleme sınavı soru kitapçığı.

Meng, H (2007). A comparison study of IRT calibration methods for mixed-format tests in vertical scaling.Unpublished Ph.D. Thesis, University of Iowa, Iowa.

McBridge, J.& Wise, L. (2001) Developing the Vertical Scale fort he Florida Comprehensive Assesment Test (FCAT). A Harcourt Educational Measurment, San Antonio, Texas.

Stocking, M. L. & Lord, F. M. (1983). Developing a Common Metric İn Item Response Theory. *Applies Psychological Measurment,* 7(2), 201-2010.

Tong, Y. & Kolen, M. (2010) Scaling: An ITEMS Module. *Educational Measurement: Issues and Practice, 29(4), 39-48.*

Tong, Y. & Kolen (2007). Comparison of Methodologies and Results in Vertical Scaling for Educational Achievement Tests. *Applied Measurement in Education, 20(2), 227-253.*

Tong, Y. & Kolen, M. (2008). *Maintenance of Vertical Scales.* National Council on Measurement in Education: March 2008. New York City.

Yen, W. M. (1986). The choice of scale for educational measurement: An IRT perspective. Journal of Educational Measurement, 23, 299–325.

Zimowski, M. F., Muraki, E., Mislevy, R. J.,& Bock, R. D. (2003). BILOG-MG 3 [computer program]. Chicago: Scientific Software Corporation.

Extended Abstract

In this study the comparison of the vertical scales, which are obtained through the combination of separate and concurrent calibration based on item response theory and EAP, MAP and MO proficiency estimation methods, take place. For this comparison firstly, math tests for the 6th, 7th and 8th grades were composed from the questions asked in SBS (high school entrance exam) between 2008-2010 years. 6, 7 and 8 grade tests' items selected according to learning goals which are 6^{th}, 7^{th} and 8^{th} grade mathematics curriculum. There are 21 items in the 6th and 8th grade level test and 23 items in the 7th grade level tests. In every

tests there are items appropriate to the grade level and the common items taking place in different grade level tests. In 7th grade level test for example, besides the7th grade level items there are also common items from the 6th grade tests. 6 items in the 6th and 8th grade level tests and 8 items in the7th grade level test are common. 4 of the common items in tests are common in all three level tests. In addition to these 4 common items there are 2 common items in the 6th and 7th grade level tests and 2 other items common in the 7th and 8th grade tests. From the items which are gradually common in consecutive grade levels were chosen like this: one of the 6th grade and one of the 7th grade question are common in the 6th and the 7th grade; one of the 7th grade and the 8th grade question are common in the 7th and the 8th grade. Tests for different grade levels are linked to each other by using these common items, then the scores obtained from the different tests are placed in the same scale. When the 6th grade is taken as the base grade, the common items in 6th and 7th grade level tests provides the connection between these two grades level and it is also the same for the common items in the 7th and 8th grades for the connection between the 7th and the 8th levels. So, the connection between the 6th and the 8th grades are provided through the 7th grade. Chain scaling was applied in this study. As a critic for the common item design, in order to cope with the disadvantage of lower grades being obliged to answer the questions belong to the upper grades or upper grades' having the advantage in answering the questions of lower grades, a mixed design was used. In the study the number of the common items' is more than the 25% of all items' number in all three tests.

Then, the vertical scales were developed by using the scores obtained from the answers of 503 6th grade, 502 7th grade and 500 8th grade students to the math tests in different difficulty levels. By using the scale scores which were obtained from these scores placed in the same scale with the vertical scaling process, means and the difference between the means, effect sizes and horizontal distances were calculated so it was tried to be determined how the choices in scaling process affected the vertical scaling.

When calibration methods are compared it is noticed that there is a decrease in the achievements from the 6th to the 7th grade however there is an increase from the 7th to the 8th grade in both calibration methods. Examining the standard deviations while the 7th grade is more homogenous in both calibration methods, concurrent calibration gives larger standard deviation values than separate calibration for all three grades. When the effect sizes are examined it is noticed that it takes negative values between the 6th and 7th grades however it takes positive values between the 7th and 8th grades. On the other hand effect sizes in the separate calibration are lower than those in the concurrent calibration. Mean horizontal distances give the same results with the effect size. When the horizontal distances are examined, as the achievement level between the 6th and 7th grades increases the difference between the proficiency levels also increases in favor of the 6th grades in separate calibration. Besides, in the 7th and 8th grades the difference also increases with the increase in the achievement level. In concurrent calibration the horizontal distances between the 6th and 7th grades taking negative and decreasing values (according to absolute value) which means that the difference between the 6th and the 7th grades increases slowly in favor of the 7th grades besides taking positively decreasing values between the 7th and the 8th grades which means that the difference between the proficiency levels decreases between the 7th and the 8th grades.

When proficiency estimation methods are examined, in terms of means while MAP usually creates lower values comparing with EAP and MO it is noticed that values calculated with EAP are lower. According to the examination of the standard deviation values while MO gives larger values than EAP and MAP such a pattern is not there for the proficiency estimations in the concurrent calibration. Although in the separate calibration the effect sizes are in the form of EAP MAP MO, in concurrent calibration EAP takes larger values comparing to MO. While mean horizontal distance values are in the form of EAP MO MAP in the separate calibration, they take values in the form of MAP EAP MO in the concurrent calibration. In horizontal distances there is a pattern in the form of EAP MO MAP in separate calibration while the pattern is in the form of MAP EAP MO in concurrent calibration.

Kaynakça Bilgisi

Altun, A., & Kelecioğlu, H. (2016). Dikey ölçeklemede madde tepki kuramına dayalı kalibrasyon ve yetenek kestirim yöntemlerinin karşılaştırılması. *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi [Hacettepe University Journal of Education*], 31(3), 447-460.

Citation Information

Altun, A., & Kelecioğlu, H. (2016). A comparison of calibration methods and proficiency estimators based on item response theory in vertical scaling [in Turkish]. *Hacettepe University Journal of Education [Hacettepe Üniversitesi Eğitim Fakültesi Dergisi*], 31(3), 447-460.