Hacettepe Üniversitesi Eğitim Fakültesi Dergisi (H. U. Journal of Education) 29(3), 23-34 [Temmuz 2014]

Dikey Ölçeklemede Klasik Test ve Madde Tepki Kuramına Dayalı Yöntemlerin Karşılaştırılması *

Comparison of the Methods of Classical Test Theory and Item Response Theory on Vertical Scaling

Emre ÇETİN **, Selahattin GELBAL***

ÖZ: Dikey ölçekleme, öğrencilerin, ardışık sınıf ya da yaş seviyelerinde matematik ya da okuma becerileri gibi alanlarda, ne kadar gelişim ortaya koyduğunu belirlemeye yarayan bir test bağlama türüdür. Bu araştırmanın amacı, dikey ölçekleme işlemi sonucunda, sınıf seviyesi arttıkça, bu seviyelerde meydana gelen gelişimin örüntüsünü çıkarmaktır. Araştırmanın verilerini, 2005 yılında Türkiye genelinde yapılan İlköğretim Öğrencilerinin Başarılarının Belirlenmesi Sınavı'na (ÖBBS) ait 6., 7. ve 8. sınıf öğrencileri oluşturmaktadır. Dikey ölçekleme, Klasik Test Kuramı (KTK) ve Madde Yanıt Kuramı (MTK) temelinde uygulanmıştır. KTK'ya dayalı olarak Thurstone (1938), Madde yanıt Kuramına dayalı olarak yapılan yetenek kestirimlerinde de Expected A Posteriori (EAP) puanlama yolu kullanılmıştır. Dikey Ölçekleme sonuçlarının değerlendirme ölçütü olarak; ortalama, standart sapma ve etki büyüklüğü değerleri kullanılmıştır. Araştırma sonunda, Thurstone ölçeklemede Matematik ve Türkçe testlerinde sınıf seviyesi ile birlikte standart sapmalar artış göstermektedir. Araştırmaya dayalı bulgular incelendiğinde, gerek KTK gerekse MTK uygulamalarında, ortalamaların seyrinden farklı olarak standart sapmaların arttığı söylenebilir..

Anahtar sözcükler: dikey ölçekleme, test eşitleme, bağlama, örneklem büyüklüğü, akademik gelişim

ABSTRACT: Vertical Scaling is a kind of linking which is used to determine how much the students of adjacent grades or ages have improved in the subject areas such as Maths and Language. The purpose of this research is, as a result of vertical scaling, to establish the pattern of students' improvement in certain levels as class grade increases. The data in this research were obtained from the Achievement Exam (OBBS) results for Turkish primary school students of the 6th, 7th and 8th grades in 2005. Vertical Scaling was conducted based on Classic Test Theory (CTT) and Item Response Theory (IRT). Thurstone (1938) scaling method was used based on CTT, and Expected A Posteriori (EAP) scaling method was used in IRT estimation. As an evaluation criteria of vertical scaling, mean, standard deviation and effect size figures in academic growth were used. At the end of the research, in Thurstone Scaling, standard deviations in Maths and Turkish Tests rise as class levels increase. In the conducts of CTT and IRT, it can be said that standard deviations increase free from the increase or decrease of means.

Keywords: vertical scaling, test equating, linking, sample size, academic growth

1. GİRİŞ

Özellikle uluslararası sınavların (TIMMS, PISA, PIRLS vb.) kullanım sıklığının da artmasıyla, öğrencilerin akademik başarılarının yıldan yıla incelemek ve öğrenci başarısındaki gelişimin ne kadar olduğunu ortaya koyabilmek amacıyla yapılan çalışmaların sayısında ciddi bir artış meydana gelmiştir. Bu çalışmalardaki artışın bir diğer nedeni, 'gelişmiş' ülkeler öncülüğünde öğrenci başarısının izlenmesi ve gereken önlemlerin alınmasına yönelik kabul edilmiş yasalardır (NCLB; Public Law 107-110). Literatürde, öğrencilerin bir kademeden daha üst bir kademeye geçerken gösterdikleri gelişimi belirlemek amacıyla üzerinde hemfikir olunan bir yöntem bulunmamaktadır (Tong, 2005). Bu durumda eşitleme ya da ölçekleme olarak adlandırılan bağlama türlerine ihtiyaç duyulmaktadır. Dikey ölçekleme çalışmaları sonucunda, öğrencilerin ardışık sınıf seviyelerindeki gelişiminin izlenmesine ek olarak, okullar arası başarının izlenmesi ve bunun sonucunda okullar arasındaki başarı farklılıklarının giderilmesine yönelik önlemler de alınabilmektedir. Benzer şekilde, cinsiyet ve etnik alt gruplar arasındaki başarı farklarının miktarı dikey ölçekleme yoluyla belirlenebilir ve giderilebilir.

---

* Birinci yazarın doktora tezinden türetilmiştir.

** Yrd.Doç.Dr.,Doğu Akdeniz Üniversitesi Eğitim Bilimleri Bölümü, e-posta: kiriemre@gmail.com

***Prof.Dr. Hacettepe Üniversitesi Eğitimde Ölçme ve Değerlendirme Anabilim Dalı, e-posta: gelbal@hacettepe.edu.tr

Dikey ölçekleme çalışması yütülmesinde karşılaşacak problemler, ölçekleme ile ilgili psikometrik kararlar ve ölçekleme sonuçlarına dayalı olarak verilecek pratik uygulamalara dönük kararlar olmak üzere iki türe ayrılabilir. Bu kararları, pratik ve teorik kararlar olarak adlandırmak mümkündür. Teorik (psikometrik) kararlar, ölçekleme yönteminin (KTK, MTK), ölçekleme deseninin, ölçeklemede değerlendirme ölçütlerinin, uygun örneklem büyüklüğünün seçilmesi gibi kararlardan; pratik kararlar ise ölçekleme sonunda, başarısız okullara ya da öğrencilere yönelik ne gibi uygulamaların yapılacağı, başarısızlığa ya da başarıya yönelik sorumluluğun, il, eyalet, okul ya da merkezi bir kuruma verilmesi, başarısız okullar için ne kadar bütçe ayrılması veya ne tür önlemler alınması gerektiği gibi aşamalardan oluşabilmektedir. Dikey ölçeklemede kullanılan üç farklı yöntem bulunmaktadır. *Hieronymus ölçekleme, Thurstone ölçekleme* ve *Madde Tepki Kuramına* dayalı ölçekleme (Kolen ve Brennan, 2004). Hieronymus ölçekleme yaygın olarak kullanılan bir yöntem olmadığından ve bu araştırmada kullanılmadığından bu yöntem açıklanmamıştır.

1.1. Thurstone Dikey Ölçekleme

Thurstone 1925 ve 1938 olmak üzere iki ayrı ölçekleme yöntemi geliştirmiştir. İlk geliştirilen Thurstone ölçeklemede (1925), ilk olarak maddelerin güçlük indeksleri elde edilir. Daha sonra elde edilen güçlük indeksleri normalleştirilmiş *z* puanlarına dönüştürülür. Thurstone daha sonra (1938) madde güçlükleri yerine ham puanlara dayalı yeni bir ölçekleme yöntemi önermiştir. Bu yöntemde bütün sınıf seviyelerinde dağılımın normal olduğu sayıltısı vardır. Bu yöntemin normallik sayıltısından öte gizli sayıltısı da maddelerin ayırıcılık güçlerinin yüksek olmasıdır (Gulliksen, 1950). Thurstone yöntemi ile ölçeklemede ilk olarak her bir test puanına ait yüzdelik sıralar bulunur; daha sonra ham puanlar normalleştirilmiş *z* puanlarına dönüştürülür. Farklı seviyelerden elde edilecek *z* puanları grupların yetenek düzeyleri farklı olduğundan eşit olmayacaktır. Bu puanlar araştırma deseninin türüne göre (ortak madde, ya da ortak grup) ortak bir ölçeğe dönüştürülür. Thurstone ölçekleme temel olarak şu aşamalarla gerçekleştirilir:

1) Her bir seviye için ham puanlar elde edilir.

2) Her bir seviye için ham puanlar yüzdelik puanlara dönüştürülür.

3) Yüzdelik puanlar normalleştirilmiş *z* puanlarına dönüştürülür

4) Bağlanacak ardışık seviyelere ait *z* puanları dağılımının saçılım grafiği çıkarılır.

5) Eşitlik 1 ve 2 kullanılarak ardışık seviyeler ortak ölçeğe dönüştürülür.

Thurstone ölçekleme ardışık seviyelere ait gruplar için aşağıdaki şekilde elde edilir (Gulliksen, 1950).

$$\sigma_2(SC) = \frac{\sigma[z_1^*(y)]}{\sigma[z_2^*(y)]}\sigma_1(SC) \quad (1)$$

$$\mu_2(SC) = \sigma_1(SC)\left[\mu[z_1^*(y)] - \frac{\sigma[z_1^*(y)]}{\sigma[z_2^*(y)]}\mu[z_2^*(y)]\right] + \mu_1(SC) \quad (2)$$

$\mu_1(SC)$= Ardışık alt seviyedeki gruba ait ortalama.

$\mu_2(SC)$= Ardışık üst seviyedeki gruba ait ortalama.

$\sigma_1(SC)$= Ardışık alt seviyedeki gruba ait standart sapma

$\sigma_2(SC)$= Ardışık üst seviyedeki gruba ait standart sapma.

$\sigma[z_1^*(y)]$= Ortak maddelerin alt seviyedeki gruba ait standart sapması.

$\sigma[z_2^*(y)]$= Ortak maddelerin üst seviyedeki gruba ait standart sapması.

$\mu[z_2^*(y)]$= Ortak maddelerin üst seviyedeki gruba ait ortalaması.

$\mu[z_1^*(y)]$= Ortak maddelerin alt seviyedeki gruba ait ortalaması

KTK, madde ve test istatistiklerinin örnekleme bağlı olması, hata varyansının ve standart hatanın tüm bireyler için eşit olması, bireylerin yetenek düzeyinin maddelere bağlı olması, üst ve alt gruptaki yetenek kestirimleri için uygun olmaması, testin güvenirliğinin örnekleme bağlı olması gibi nedenlerden dolayı sınırlılıklar içermektedir. KTK'nın bazı zayıf noktaları, MTK'nın ortaya çıkmasına zemin hazırlamıştır. Bu kuram, bireyin yetenek düzeyinin, belirli bir madde grubundan bağımsız olarak kestirilebileceği sayıltısı üzerine kurulmuştur (Hambleton, 1985).

1.2. Madde Tepki Kuramı İle Dikey Ölçekleme

Madde tepki kuramı diğer ölçekleme yöntemlerine göre daha güçlü sayıltılara sahiptir. Tek boyutlu modellerin en önemli sayıltıları *tek boyutluluk* ve *yerel bağımsızlıktır*. Dikey ölçeklemede, Madde Tepki Kuramı (MTK) uygulandığında, tek boyutluluk, parametre sayısı (1, 2 ya da 3 parametreli modeller), kalibrasyon yöntemi (ayrı, ortak), puanlama yolları ve yetenek kestirimi gibi çok sayıda karar vermek gerekmektedir. MTK'ya dayalı ölçeklemenin karmaşık ve tutarsız sonuçlar vermesi, bu kararların birbirleriyle ilişkileri ile ilgilidir (Tong, 2005; Kolen ve Brennan, 2004).

Madde parametreleri kalibrasyon yoluyla ortak bir ölçeğe (0,1) dönüştürüldükten sonra, kestirimlerinin yapılması gerekmektedir. Yetenek kestirimi için**,** Quadrature Dağılım (QD), Maksimum Olabilirlik Kestirimi (Maximum Likelihood Estimation), Expexted A Posteriori (EAP) en yaygın kullanılan yöntemlerdendir.

1.3. Dikey Ölçeklemede Kullanılan Desenler

Dikey ölçeklemede, ölçekleme testi deseni (*scaling test)* ve ortak madde deseni olmak üzere iki temel yöntem bulunmaktadır. Ölçekleme testi deseninde ölçeklenecek sınıflar ya da seviyelere kendi sınıf seviyelerine ve programlarına uygun bir test uygulanır. Daha sonra ölçekleme testi adı verilen tüm sınıf seviyelerinin programlarına uygun hazırlanmış bir test tüm gruplara uygulanır. Ölçekleme testi deseninde her sınıf kendi seviyesine uygun test alır. Uygulanan bu testler ölçekleme çalışmasını gerçekleştirmek amacıyla birbirlerine bağlanarak ortak bir ölçeğe dönüştürülmüş olur. Ortak madde deseninde ise, her grup kendi seviyesine uygun bir test alır. Fakat ardışık sınıfların aldığı testlerin içinde ortak maddeler bulunur. Bu ortak maddeler iç ortak madde denir (Kolen ve Brennan, 2004). Bu araştırmada 6., 7. ve 8. sınıflar üzerinde ölçekleme çalışması yapılmıştır. Her sınıf seviyesinde aynı olan ortak maddeler kullanılmıştır.

1.4. Değerlendirme Ölçütleri

Dikey ölçekleme temelinde yürütülen araştırmalarda, bağlama sonuçlarını değerlendirmek amacıyla, sınıf seviyeleri arasındaki ortalama farkı, standart sapma farkı ve etki büyüklüğü kullanılmaktadır (Kolen ve Brennan, 2004).

1.4.1. Etki büyüklüğü

Sınıf seviyeleri arasındaki ortalama farkları, akademik gelişim için bilgi verse de yeterli değildir. Yen (1986) akademik gelişim için standart sapmaları da dikkate alan, etki büyüklüğü indeksini önermiştir.

$$\text{Etki Büyüklüğü} = \frac{\bar{X}_{\text{üst}} - \bar{X}_{\text{alt}}}{\sqrt{\frac{S^2_{\text{üst}} - S^2_{\text{alt}}}{2}}} \qquad (3)$$

$\bar{X}_{\text{üst}}$:Üst sınıfa ait ortalama

$\bar{X}_{alt}$:Alt sınıfa ait ortalama

$S^2_{üst}$:Üst sınıfa ait standart sapma

S^2_{alt}:Alt sınıfa ait standart sapma

Etki büyüklüğü ardışık iki sınıf başarısı arasındaki farklılığı standardize etmek amacıyla kullanılır. Bu çalışmada, 6 ile 7 ve 7 ile 8. Sınıflar arasında olmak üzere iki ayrı etki büyüklüğü hesaplanmıştır.

Ölçekleme çalışmalarının, çoğunlukla sınıf seviyesi ile birlikte ortalama ve standart sapma değerlerindeki değişim üzerinde ve bu değişimi etkileyen, test uzunluğu, güvenirlik gibi faktörler üzerinde yoğunlaştığı görülmektedir.

1.5. Araştırmanın Amacı ve Önemi

Bireylerin, belirli bir özellikte bir yıldan diğerine ne kadar gelişim gösterdiklerinin belirlenmesi, karmaşık bir işlemdir. Bağlama sonuçlarını etkileyen değişkenlerin tek tek ya da birbirleri ile etkisine ek olarak, testlerin psikometrik özellikleri, grupların yetenek düzeyi, ortak maddelerin içerikleri de ölçekleme işlemi üzerinde etkili olabilmektedir.

Dikey ölçekleme çalışması yürütülmesi sırasında, verilecek kararlar iki ana grupta toplanabilir: a) Ölçekleme işlemi ile ilgili kararlar, b) Ölçekleme sonuçlarına dayalı olarak verilen kararlar. Ölçekleme çalışması yürütülmesinde, kullanılacak ölçekleme yöntemi (MTK, KTK) en önemli kararlardan birisini oluşturmaktadır. Dikey ölçekleme çalışmalarının en önemli avantajlarından birisi, ölçekleme sonunda elde edilen puanların testten bağımsız olmasıdır. Dikey ölçekleme, avantajlarının yanında dezavantajlar da doğurabilmektedir. Ölçekleme sonunda okullara, verilen ceza uygulamaları, öğretimin test tekniğine yönelik olmasına ve daha alt düzey davranışların kazanılmasına ağırlık verilmesine neden olabilmekte. Ölçekleme sonunda elde edilen başarı puanlarına dayalı olarak verilecek kararların, kaynağı ve yönü ülkelerin ekonomik politikalarına ve yapılanma türlerine de bağlı olabilmektedir. Örneğin, Amerika Birleşik Devletlerinde, okullar öngördükleri başarı düzeyine erişemediklerinde, okulun bir şirket tarafından işletilmesine karar verecek kadar değişik çözüm yolları aranmakta iken Türkiye gibi okulların ve öğrencilerin merkezi kararlarla yönetildiği ülkelerde daha farklı önlemler almak gerekebilmektedir.

Dikey ölçekleme sonunda farklı okullar ya da iller birbirleriyle karşılaştırılabilir ve ortalamanın ya da standartların altın da kalan kurum ya da iller için önlemlerin alınması sağlanmış olur. Ölçekleme çalışması sonunda, kullanılacak ölçütlerin standart olması, ve kurumdan kuruma ya da ilden ile değişmemesi gerekmektedir. Yaygın olarak kullanılan ölçütler, sınıf seviyesi ile birlikte ortalamada ya da değişkenlikteki farklılaşmalardır. Bu çalışmanın amacı, gerek klasik gerekse madde tepki kuramına dayalı olarak dikey ölçekleme çalışması yürütmek ve ardışık sınıf seviyeleri arasındaki ortalama, standart sapma, etki büyüklüğü değerlerini karşılaştırmaktır.

1.6. Alt Problemler

1.6.1. KTK ile ölçeklemede ortalama, standart sapma ve etki büyüklüğü değerleri sınıf seviyesi ile birlikte değişim göstermekte midir?

a. *Thurstone ölçeklemede Matematik dersinde ortalama, standart sapma ve etki büyüklüğü değerleri sınıf seviyesi ile birlikte değişme göstermekte midir?*
b. *Thurstone ölçeklemede Türkçe dersinde ortalama, standart sapma ve etki büyüklüğü değerleri sınıf seviyesi ile birlikte değişme göstermekte midir?*

1.6.2. MTK ile ölçeklemede ortalama, standart sapma ve etki büyüklüğü değerleri sınıf seviyesi ile birlikte değişim göstermekte midir?

a. *MTK ile ölçeklemede Matematik dersinde ortalama, standart sapma ve etki büyüklüğü değerleri sınıf seviyesi ile birlikte değişmekte midir?*
b. *MTK ile ölçeklemede Türkçe dersinde ortalama, standart sapma ve etki büyüklüğü değerleri sınıf seviyesi ile birlikte değişmekte midir?*

2.YÖNTEM

2.1. Araştırma Verileri

Araştırmanın verilerini, 2005 yılında Türkiye genelinde yapılan İlköğretim Öğrencilerinin Başarılarının Belirlenmesi Sınavı (ÖBBS)'na ait 6., 7. ve 8. sınıf öğrencilerinin Türkçe ve Matematik testlerine ait puanları oluşturmaktadır. Bu testlere ait soru sayıları, öğrenci sayıları ve ortak madde sayıları Tablo 2.1'de verilmiştir.

Tablo 1: Matematik ve Türkçe Dersi Test ve Örneklem Büyüklükleri

	Matematik			Türkçe		
	6.sınıf	7.sınıf	8. sınıf	6.sınıf	7.sınıf	8. sınıf
Soru sayısı	20	20	25	20	20	25
Ortak Madde Sayısı	4	4	4	4	4	4
N	13401	5368	11200	11168	9892	13329

Tablo 1 incelendiğinde, Matematik ve Türkçe testinde 6 ve 7. sınıflarda 20, 8. sınıfta ise 25 soru bulunmaktadır. Her testte ortak madde sayısı 4'tür. Bu sonuçlar, ölçekleme çalışması yapmak için gereken asgari sayı ölçütünü karşılamaktadır.

2.2. Evren ve Örneklem

Bu araştırmanın evrenini, Türkiye'de öğrenim görmekte olan 6, 7 ve 8. sınıf öğrencileri oluşturmaktadır. Araştırma gerçek veri üzerinden yürütülmüş, evreni temsil eden örneklem Milli Eğitim Bakanlığı, EARGED birimi tarafından tabakalı örnekleme yoluyla seçilmiştir (MEB, 2008).

2.3. Araştırma Deseni

Bu araştırma, yetenek düzeyi eşit olmayan gruplara uygulanan ortak madde deseni üzerinden yürütülmüştür. Ortak madde sayısının tüm testin %20'si kadar olması yeterlidir (Kolen ve Brennan, 2004). Bu araştırmanın ortak maddeleri 6., 7. ve 8. sınıflarda aynıdır. Bu yönüyle araştırma karma bir modele dayalıdır. Bu yöntem, seçilen temel sınıfın, diğer sınıflara doğrudan ölçeklenebilmesini sağlar; bu nedenle zincirleme ölçekleme işlemi uygulanmamıştır.

2.4. Kullanılan Yöntemler

Bu araştırmada, KTK'ya dayalı Thurstone Ölçekleme ve MTK'ya dayalı dikey ölçekleme tekniği kullanılmıştır. Thurstone ölçeklemede, ilk olarak tüm örneklem için 6, 7 ve 8. sınıf seviyelerine ait ölçekleme işlemi ham puanlar ve frekanslara dayalı olarak normalleştirilmiş *z* puanlarına dönüştürülmüştür. Aynı işlemler, her bir testteki ortak maddeler için de tekrarlanmıştır. Daha sonra, 6. sınıf temel sınıf olarak ortalaması 0 ve standart sapması 1 olacak şekilde belirlenmiştir. Ortak maddelerin ortalama ve standart sapma değerleri kullanılarak bölüm 1'deki eşitlik 1ve 2 yardımıyla ardışık seviyelere ait ortalama ve standart sapmalar elde edilmiştir. Thurstone ölçeklemenin hangi puan ranjında yürütüleceğine karar vermek keyfi olsa da, seçilecek ölçek puanları araştırma sonuçlarını etkilemektedir (Williams ve diğerleri, 1998). Bu araştırmada, en düşük ve en yüksek 3 puan çıkarıldığında, 6 ve 7. sınıflarda 14, 8. sınıfta ise 19 puan üzerinden ölçekleme yürütülmüştür.

2.5. Bilgisayar Programları

Farklı sınıf seviyelerinde, gerek MTK için gerekse Thurstone ölçekleme için temel sayıltı olan tek boyutluluk testi için DFA analizi yürüten LISREL 8 (Jöreskog ve diğerleri, 1999) Programı kullanılmıştır. Dikey ölçeklemede, *z* puanları ve yüzdelikler Excel programında analiz edilmiştir. MTK'ya dayalı ölçeklemede farklı grupların karşılaştırmasını ortak kalibrasyon yöntemi ile sağlayan BILOG-MG 3 (Zimowski ve diğerleri, 1996) kullanılmıştır.

2.6. Verilerin Analizi

Analizler yapılmadan, önce eksik ve kayıp veriler veri setinden çıkarılmıştır. Daha sonra elde edilen matrislerden ortak maddeler için ve tüm test için, toplam puanlar elde edilmiştir. Analiz sonuçları Matematik ve Türkçe testlerine göre ayrı ayrı verilmiştir.

Tablo 2: Matematik ve Türkçe Testlerine Ait Betimsel İstatistikler

	Matematik			**Türkçe**		
	6.sınıf	7.sınıf	8. sınıf	6.sınıf	7.sınıf	8. sınıf
Ortalama	7.49	7.28	11.18	10.03	10.31	14.79
Ortalama Güçlük	0,37	0,36	0,44	0,501	0,51	0,591
Standart sapma	3.80	4.01	5.65	4.77	4.71	5.99
Medyan	7	6	10	9	10	15
Standart Hata	1.98	1.98	2.20	1.98	1.98	2.11

Tablo 2 incelendiğinde, Matematik testinin ortalama güçlüğünün Türkçe testinden daha küçük olduğu, her üç sınıf seviyesinde de görülmektedir. Matematik dersi 6. ve 7. sınıfta benzer güçlüklerde, 8. sınıfta ise gruba daha kolay gelmiştir. Benzer şekilde, Türkçe testinde de ortalama güçlük değerleri 6 ve 7. sınıflarda aynı iken, 8. sınıfta daha yüksek bulunmuştur. Standart sapmalarda, ortalamalar ile benzer bir seyir izlemekte, Matematik testinde sınıf seviyesi arttıkça standart sapmalar artmış, Türkçe testinde de 6, ve 7. sınıfta sabit kalırken 8. sınıfta artmıştır.

Thurstone ölçeklemenin temel sayıltıları, maddelerin, ayrıcılık güçlerinin yüksek olması, ve testlerin yeterli kabul edilebilecek güvenirlik değerlerinde olmasıdır. Bu sayıltıları test etmek amacıyla, KTK madde analizleri yapılmış, madde güçlük indeksi değerleri ve nokta çift serili ayrıcılık gücü indeksleri (rjx) elde edilmiştir. Testin güvenirliği KR-20 güvenirlik formülü ile hesaplanmış ve elde edilen güvenirlik katsayıları, MEB tarafından elde edilen güvenirlik katsayıları ile tutarlı bulunmuştur (MEB, 2008). Klasik test kuramına dayalı olarak elde edilen madde analizlerine bakıldığında, ölçekleme yapabilemek için gereken sayıltıların karşılandığı görülmüştür.

2.6.1. Normallik

Thurstone ölçeklemenin yürütülmesi için en önemli sayıltı normalleştirilmiş puanlarda meydana gelebilecek uç değerlerin olmamasıdır. Gulliksen (1950), ardışık seviyelere ait saçılım grafiklerine bakmanın ve eğer doğrusallığı bozan bir durum varsa, eğer mümkünse madde çıkartmayı önermiş aksi takdirde ölçeklemenin yürütülemeyeceğini belirtmiştir. Matematik ve Türkçe testlerinde 6-7, ve 7-8 sınıflara ait *z* puanlarının doğrusal olduğu, bu durumunda Thurstone ölçeklemede ardışık sınıflar arasındaki normalleştirilmiş *z* puanlarının doğrusallık sayıltısının karşıladığı görülmektedir. Doğrusallığı bozan, yüksek *z* puanları bulunmakla birlikte, bu puanların etkisi, en yüksek ve en düşük puanlar dağılımdan çıkarıldığından giderilmiştir.

2.6.2. Tek Boyutluluk

Tek boyutluluk sayıltısını test etmek amacıyla Matematik ve Türkçe testlerine LISREL programı yardımıyla doğrulayıcı faktör analizi (DFA) kullanılmıştır. Bu çalışmada model

uyumunu araştırmak için, Ki Kare uyum testi, yaklaşık Hataların Ortalama Karekökü (Root Mean Square Error of Approximation, RMSEA), İyilik Uyum İndeksi (Goodness of Fit Index,GFI), Düzeltilmiş İyilik Uyum İndeksi (Adjusted Goodness of Fit Index, AGFI), Karşılaştırmalı Uyum Indeksi (Comparative Fit Index, CFI), Normlaştırılmış Uyum Indeksi (Normed Fit Index, NFI), Göreli Uyum İndeksi (Relative Fit Index, RFI), Fazlalık Uyum İndeksi (Incremental Fit Index, IFI) kullanılmıştır.

RMSEA değeri matematik dersinde 6, 7 ve 8. sınıflarda 0,002, Türkçe dersinde 6 ve 8 sınıflarda 0.01, 7. sınıfta da 0.02 olarak bulunmuştur. Her iki test için elde edilen tüm RMSEA değerleri 0,01 ile 0,02 arasında bulunmuştur. Bu durum da uyumun mükemmele yakın olduğunu göstermektedir. NFI, CFI, IFI ve GFI değerleri 1'e yaklaştıkça uyum mükemmelleşir, bu indeksler için tablo değeri incelendiğinde, bu indekslerin 0,93 ile 0,98 arasında değiştiği ve bu değerler de verinin modele yüksek derecede uyduğunu göstermiştir.

2.6.3. MTK Model Veri Uyumu

Model veri uyumunun testi her bir maddeye ait gözlenen ve beklenen doğru cevaplanma olasılıklarının karşılaştırılması yoluyla yapılır. Bu karşılaştırmayı görmenin en açık yolu, her bir madde için elde edilen gözlenen ve beklenen değerleri arasındaki farkı gösteren değerlere bakmaktır (Hambelton ve diğ, 1991). Bu araştırmada matematik ve Türkçe derslerine ait, 1, 2 ve 3 parametreli modellerden elde edilen uyum iyiliği grafikleri her bir madde için çıkarılmış ve bu grafikler sonucunda, daha iyi uyum gösteren 2 parametreli model seçilmiştir.

3. BULGULAR

Araştırma bulguları, sınıf seviyelerine, kullanılan dikey ölçekleme yöntemlerine ve üç farklı bağlama tekniği değerlendirme yöntemine göre verilmiştir. Araştırmanın alt problemlerine ilişkin bulgular verilmeden önce araştırmada sınıflara göre eşitlenen testlere ilişkin betimsel istatistikler Tablo 3.1'de verilmiştir.

Tablo 3: Sınıflara göre Matematik ve Türkçe Testinden Alınan Puanların Betimsel İstatistikleri

	Matematik			Türkçe		
	6.sınıf	7.sınıf	8. sınıf	6.sınıf	7.sınıf	8. sınıf
Madde sayısı	20	20	25	20	20	25
Ortak Madde Sayısı	4	4	4	4	4	4
N	13401	5368	11200	11168	9892	13329
Ortalama Güçlük	0,37	0,36	0,44	0,50	0,51	0,59
Sx	3.80	4.01	5.65	4.77	4.71	5.99
KR-20	.72	.75	.84	0.82	0.82	0.87
Ortalama Ayırıcılık (rjx)	0.51	0.54	.58	0.61	0.60	.64

Tablo 3'te araştırma verilerindeki uç değerler ve kayıp veriler çıkarıldıktan sonra, öğrenci sayıları matematik testinde, 6. sınıfta 13401, 7. sınıfta 5368 ve 8. sınıfta da 11200'dür. 7. sınıftaki öğrenci sayısı diğer seviyelere göre az olmasına rağmen, literatürde önerilen örneklem sayılarının (250, 500, 1000.vb.) çok üstünde olduğundan yeterli sayılmıştır.

3.1. Alt Problem 1.6.1.a'ya İlişkin Bulgular ve Yorumlar

a. Thurstone ölçeklemede Matematik dersinde ortalama, standart sapma, ve etki büyüklüğü değerleri sınıf seviyesi ile birlikte değişme göstermekte midir?

Bu alt probleme cevap bulabilmek için, Matematik testin Thurstone ölçeklemenin ilk aşamasında bölüm 1'de belirtildiği gibi ham puanlar üzerinden her bir sınıf seviyesine ait normalleştirilmiş z puanları tüm test için ve ortak maddeler için çıkarılmıştır.

Elde edilen, z puanları ile ilişkilendirmek amacıyla her iki ardışık seviyede de ortak olan maddelere ait, normalleştirilmiş z puanları da elde edilmiştir. Elde edilen z puanlarına dayalı ortalama 6. sınıf ortalaması 0 standart sapması 1 olacak şekilde temel sınıf olarak seçilmiş ve diğer sınıfların ortalaması yeniden ölçeklenmiş olarak elde edilmiştir. Ölçeklemenin son aşaması olarak, normalleştirilmiş z puanları ile yeniden ölçekleme sonucu elde edilen ortalama ve standart sapmaya dayalı olarak ölçeklenmiş puanlara ulaşılmıştır.

Tablo 4: Matematik Dersine Ait Ölçeklenmiş Puanlar (Sonuç)

	6.sınıf	7. sınıf	8.sınıf
Ortalama	0,472	0,555	0,249
Sx	0,995	0,977	1,126
Etki Büyüklüğü	0,08	-0,29	

Tablo 4 incelendiğinde, ölçeklenmiş puanlara ait ortalamanın 6 ve 7. sınıflarda birbirine yakın değerler alırken, 8. sınıfa gelindiğinde bu ortalama düşmüştür. Fakat standart sapmalar incelendiğinde, 6. sınıftan 7.sınıfa düşmüş, 8. sınıfa gelindiğinde yükselmiştir; ölçekleme sonucunda beklenti sınıf seviyesi arttıkça ölçeklenmiş puanların da tutarlı olarak artmasıdır; fakat ortalamadan çok standart sapmalar sınıf seviyesi ile tutarlı artış göstermektedir.

Etki büyüklükleri incelendiğinde 6 ve 7. sınıf arasındaki standartlaştırılmış etki büyüklüğü 0,08, 7 ve 8. Sınıflar arasındaki ise -0,29 bulunmuştur. Buradan hareketle gerek ortalama gerekse etki büyüklüğü matematik dersinde her üç seviyede tutarlı bir örüntü oluşturmamaktadır.

3.2 Alt Problem 1.6.1.b'ye İlişkin Bulgular ve Yorumlar

a. Thurstone ölçeklemede Türkçe dersinde ortalama, standart sapma, ve etki büyüklüğü değerleri sınıf seviyesi ile birlikte değişme göstermekte midir?

Bu alt probleme cevap bulabilmek için, Türkçe dersinde Thurstone ölçeklemenin ilk aşamasında bölüm 1'de belirtildiği gibi ham puanlar üzerinden her bir sınıf seviyesine ait normalleştirilmiş z puanları tüm test için ve ortak maddeler için çıkarılmıştır.

Elde edilen, z puanları normalleştirilmiş z puanlarına dönüştürülür. Elde edilen z puanlarına dayalı ortalama 6. sınıf ortalaması 0 standart sapması 1 olacak şekilde temel sınıf olarak seçilmiş ve diğer sınıfların ortalaması yeniden ölçeklenmiş olarak elde edilmiştir. Normalleştirilmiş z puanları ile yeniden ölçekleme sonucu elde edilen ortalama ve standart sapmaya dayalı olarak ölçeklenmiş puanlara ulaşılmıştır.

Tablo 5: Türkçe Dersine Ait Ölçeklenmiş Puanlar (Sonuç)

	6.sınıf	7. sınıf	8.sınıf
Ortalama	-0,119	-0,207	-0,487
Sx	0,838	0,915	1,231
Etki Büyüklüğü	-0,09	-0,26	

Tablo 5'te, Türkçe dersinde tüm örnekleme ait ölçeklenmiş puanların ortalaması seviyeler arttıkça düşmektedir; fakat standart sapmalar sınıf seviyesi ile birlikte artmaktadır bu bulgu matematik dersindeki sonuçlarla tutarlılık göstermektedir. Ortak maddelere ait standart sapmaların sınıf seviyesi arttıkça gözlendiği düşünüldüğünde, ölçekleme sonucundaki seviyeler

arasındaki değişkenlik ortak maddelere bağlı olabilir. Özetle, ölçeklenmiş puanlarda sınıf seviyesi arttıkça standart sapmalar da artmaktadır.

Etki büyüklükleri incelendiğinde, 6 ile 7. sınıf arasında -0,09 7 ile 8. sınıf arasında ise -0,26 bulunmuştur. Bu bulgu Matematik testine ait sonuçlar ile tutarlılık göstermektedir. Türkçe testinde de en düşük etki büyüklükleri 7 ile 8. sınıf arasında gözlenmektedir.

3.3. Alt Problem 1.6.2.a'ya İlişkin Bulgular ve Yorumlar

a. MTK ile ölçeklemede Matematik dersinde ortalama, standart sapma, ve etki büyüklüğü değerleri sınıf seviyesi ile birlikte değişmekte midir?

Kestirimler yapılırken, tıpkı Thurstone ölçeklemedeki gibi, temel grup 6. sınıf olarak seçilmiş ve ilk ardışık olarak 6 ile 7. sınıf daha sonra da 6 ile 8. sınıf ölçeklenerek, ortak dönüşüm tamamlanmıştır. Ortak seviye belirlendikten sonra, toplam puanlar üzerinden yetenek kestirimi yapılmasına olanak sağlayan EAP yöntemiyle ardışık seviyeler ölçeklenmiştir. Bu ölçekleme sonucunda BILOG-MG 3 (Zimowski ve diğerleri, 1996) yeniden ölçeklenmiş ve sonsal dağılım (posterior distribution) olmak üzere iki farklı ortalama ve standart sapma değeri vermektedir.

Tablo 6: Matematik Dersine MTK'na Dayalı Ortalama, Standart Sapma ve Etki Büyüklükleri

	6.sınıf		7.sınıf		8.sınıf	
	Ortalama	Sx	Ortalama	Sx	Ortalama	Sx
Yeniden Ölçekleme	0	1	-0,101	1,062	0,291	1,308
Etki büyüklüğü		-0,09		0,35		

Tablo, 6 incelendiğinde, 6. sınıftan 7. sınıfa ortalamalar yeniden ölçekleme sonucunda ve örtük dağılım sonucunda düşerken, 8. sınıfa gelindiğinde en yüksek seviyeye ulaşmaktadır. Standart sapmalar ise, sınıf seviyesi arttıkça düzenli artış göstermektedir. Bu sonuçlar Thurstone ölçekleme ile karşılaştırıldığında, ortalamalar tutarsızlık gösterirken, standart sapmalar her iki ölçekleme yönteminde de sınıf seviyesine bağlı olarak görece artmıştır.

Ortalamalar ve Standart sapmalardan elde edilen etki büyüklükleri ise: 6 ve 7. sınıflar arasında yeniden ölçekleme için -0,09, 7. İle 8. sınıflar arasındaki ise 0,36 bulunmuş, örtük dağılıma ait etki büyüklüğü 6. 7. sınıflar için yine -0,09 iken 7 ve 8. sınıflar için 0,35 bulunmuştur. Thurstone ölçeklemeden elde edilen etki büyüklüğü ile MTK'ya dayalı etki büyüklüğü karşılaştırıldığında, 6 ile 7. sınıflar arasındaki yakın iken, 7 ve 8. sınıflar arasında tutarsız gözükmektedir.

3.4. Alt Problem 1.6.2.b'ye İlişkin Bulgular ve Yorumlar

a. MTK ile ölçeklemede Türkçe dersinde ortalama, standart sapma, ve etki büyüklüğü değerleri sınıf seviyesi ile birlikte değişmekte midir?

Tablo 7: Türkçe Dersine MTK'ya Dayalı Ortalama ve Standart sapmalar

	Ortalama			Sx		
	6.sınıf	7.sınıf	8.sınıf	6.sınıf	7.sınıf	8.sınıf
Yeniden Ölçekleme	0	-	-0,001	1	-	1,069

Tablo 7 incelendiğinde, Türkçe dersinde MTK'na dayalı olarak gerçekleştirilen yeniden ölçekleme sonucunda, 6. sınıftan 8. sınıfa ortalama ve standart sapma değerlerinde dikkate değer bir değişim gözlenmemektedir. Türkçe testinde, 6-7. sınıflar arasında model veri uyumu sağlanmadığından MTK'ya dayalı ölçekleme çalışması gerçekleştirilememiştir.

4. TARTIŞMA VE SONUÇ

Araştırma sonunda, Alt Problem 1'den elde edilen bulgular çerçevesinde, Thurstone ölçeklemede, matematik testinde ortalamalar sınıf seviyesi ile birlikte düzenli artış göstermezken, standart sapmalar tutarlı artma eğilimindedir. Türkçe testinde ise ölçeklenmiş ortalamalar, sınıf seviyesi arttıkça düşerken, standart sapmalar yine artış göstermiştir. Bu durumda, ortalama örüntüsünden bağımsız olarak Thurstone ölçekleme her iki derste de sınıf seviyesi ile birlikte standart sapmalarda artış göstermektedir. Bu bulgu, Becker ve Forsyth (1992) tarafından elde edilen bulgularla tutarlılık göstermektedir.

Alt problem 2'ye ait bulgular incelendiğinde, MTK'ya dayalı ölçekleme sonunda, matematik dersinde, sınıf seviyesi ile ortalamalar tutarlı bir artış ya da azalma göstermez iken, standart sapmalar seviyelere bağlı olarak artış göstermiştir. Fakat Türkçe testine ait, MTK ölçeklemesinde gerek ortalama gerekse standart sapmalarda düzenli bir örüntü meydana gelmemiştir.

4.1. Öneriler

Dikey ölçekleme doğası ve kullanılan yöntemlerin karmaşıklığı gibi nedenlerden dolayı, üzerinde hem fikir olunan bir yöntem değildir**.** Araştırma sonunda görülmüştür ki dikey ölçekleme işlemi yürütülmesinde, gerek MTK, gerekse KTK uygulamalarında ölçekleme sonucunun değerlendirilmesi için standart sapmalar ortalamalara göre daha uygun ve tutarlı sonuçlar vermektedir.

Ölçekleme sonucunda kullanılan yöntemin doğruluğu kadar, ölçekleme sonucuna bağlı olarak verilecek kararlarda eğitim doğurguları açısından önemlidir. Eğer kullanılan yöntemler, yeteri kadar hatadan arınık ve "doğru" değilse verilecek kararlar, bu sonuçlar çerçevesinde olmalıdır.

1. Bu araştırma, sadece Türkçe ve matematik testleri üzerinden yürütülmüştür, diğer testlerde de bulguların nasıl olacağı incelenebilir.
2. Bu araştırmada gerçek veri seti kullanılmıştır, aynı analizler yapay veri üzerinde de yürütülüp sonuçları karşılaştırılabilir.
3. Ölçekleme sürekliliği açısından, benzer araştırmalar değişik örneklem ve sınıf seviyelerinde tekrar edilebilir.
4. Farklı madde formatları üzerinde de ölçekleme çalışması gerçekleştirilebilir.
5. Ölçekleme çalışması, cinsiyet ya da daha farklı alt gruplar içinde tekrarlanabilir ve bu alt gruplar arsındaki başarı farklılıkları karşılaştırılabilir.

5. KAYNAKLAR

Becker, D. F., & Forsyth, R. A. (1992). An empirical investigation of Thurstone and IRT methods of scaling achievement tests. *Journal of Educational Measurement, 29*, 341–354.

Gulliksen, H. (1950). *Theory of mental tests*. New York: Wiley.

Hambleton, R. K. & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Boston: Kluwer Nijhoff Publishing.

Jöreskog, K., Sörbom, D., Du Toit, S.H.C. & Du Toit, M. (1999)..*LISREL 8: New Statistical Features*. Chicago, Illinois: Scientific Software International, Inc.

Kolen, M. J & Brennan, R. L. (2004). *Test equating, scaling, and linking: methods and practices* (2nd edn) (New York, Springer Verlag).

No Child Left Behind Act of 2001. (2002). Pub. L. No. 107-110, 115 Stat. 1425.

Thurstone, L. L. (1925). A method of scaling psychological and educational tests. *Journal of Educational Psychology,16*(7), 433-451.

Thurstone, L. L. (1938). Primary Mental Abilities. Psychometric Monographs, No 1. *35*, 93-107.

Tong, T (2005). *Comparison of Methodologies And Results in Vertical Scaling for Educational Achievements Tests.*Unpublished Ph.D. Thesis, University of Iowa, Iowa.

Williams, V.S.L., Pommerich, M., & Thissen, D. (1998). A comparison of developmental scales based on Thurstone methods and item response theory. *Journal of Educational Measurement*, *35*, 93-107.

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.(1996). BILOG-MG:Multiple-group IRT analysis and test maintenance for binary items [Computer software]. Chicago: Scientific Software International.

Extended Abstract

Vertical Scaling is a kind of linking which is used to determine how much the students of adjacent grades or ages have improved in the subject areas such as Maths and Language. The purpose of this research is, as a result of vertical scaling, to establish the pattern of students' improvement in certain levels as class grade increases in both Item Response Theory (IRT) and Classical Test Theory (CTT), which is also called Thurstone scaling method.

Traditional group comparison methods are insufficient to compare student achievement between adjacent grades so it is unlikely to take a picture of growth in those grades. The reason is that the group achievement levels and tests are different in adjacent grades so there is nothing left common that can be used to make comparisons.

Vertical scaling designs claim that we can make possible comparisons between adjacent grades by using single group design or anchor item design. In literature, there are three vertical scaling methods used to see a pattern of growth: 1) Thurstone method which is based on CTT, 2) IRT models, and 3) Hireynomus scaling method. Thurstone suggested two different scaling methods: the first (1925) was conducted by item difficulty and the second (1938) was based on total number correct, which also assumes that scores are normally distributed in both grades. Thurstone methods can be described in three basic steps: a) take the raw score for each grade; b) convert the raw scores into normalized z scores and, c) use the anchor items chaining adjacent grades. Thurstone scaling method also has a hidden assumption, i.e. high discrimination index of the items. To conduct vertical scaling based on IRT, many decisions have to be made: a) IRT model (1, 2 and 3 parameter), b) estimation methods (concurrent, separate) and scoring types (quadrature distribution, maximum likelihood estimation, expected aposteriori). Each decision interacts with each other and it makes models complicated. When we use separate calibration in ability estimation, scale transformation has to be made to obtain standard scores, which means 0 and standard deviation 1. Nonlinear transformations techniques should not be used because it might change the mean and standard deviation of scores. In this research, concurrent calibration was used to obtain standard scales so any of the transformation method can be used.

The data in this research were obtained from the Achievement Exam (OBBS) results for Turkish primary school students of the 6^{th}, 7^{th} and 8^{th} grades in 2005. The research was conducted among 13401 students from the 6^{th} grade, 5368 from the 7^{th} and 11200 from the 8^{th}. Vertical Scaling was conducted based on CTT and IRT.

Thurstone (1938) scaling method was used based on CTT, and Expected A Posteriori (EAP) scaling method was used in IRT estimation. In CTT and IRT unidimensionality is a strong assumption. As a previous data analysis, to test unidimensionality, Confirmatory Factor Analysis was used in Lisrel 8 software, and Root Mean Square Error of Approximation, Goodness of Fit Index, Adjusted Goodness of Fit Index, Normed Fit Index, showed that assumption held. IRT ability estimation and model data fit was conducted by using Bilog MG. To test model data fit, added FIT in >SCORE option. When IRT was chosen, as a result of model-data fit, the two-parameter model was found to fit the data better than other models. As a part of pre- analysis, CTT item analysis was conducted for all grades, and reliability and discriminations of the items was found satisfactory to perform scale transformation. Another assumption

of Thurstone scaling method is, linearity of adjacent z scores between two grades, was also tested and this assumption held too.

As the evaluation criteria of vertical scaling, standard deviations of related statistics obtained from different samplings were used in sampling size; and mean, standard deviation and effect size figures in academic growth were used.

At the end of the research, free from its mean pattern in Thurstone Scaling based on CTT, standard deviations in Maths and Turkish Tests rise as class levels increase. At the end of the scaling based on IRT, whereas class levels and means in maths do not increase or decrease consistently, standard deviations increase related to certain levels. However, in IRT scaling of the Turkish Test, a consistent pattern is obtained neither in means nor in standard deviations. According to the results and conclusions, when criteria to observe growth are studied, it is found that as class levels increase so do standard deviations consistently. In the conducts of CTT and IRT, it can be said that standard deviations increase free from the increase or decrease of means.

Kaynakça Bilgisi

Çetin, E. ve Gelbal, S. (2014). Dikey ölçeklemede klasik test ve madde tepki kuramına dayalı yöntemlerin karşılaştırılması. *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi [Hacettepe University Journal of Education]*, *29*(3), 23-34.

Citation Information

Çetin, E. & Gelbal, S. (2014). Comparison of the methods of classical test theory and item response theory on vertical scaling. [in Turkish]. *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi [Hacettepe University Journal of Education]*, *29*(3), 23-34.