ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
2018; 9(1);49-63

Kategori Sayısının Psikometrik Özellikler Üzerine Etkisinin Mokken Homojenlik Modeli'ne Göre İncelenmesi*

Investigation of the Effects of the Number of Categories on Psychometric Properties According to Mokken Homogeneity Model*

Asiye ŞENGÜL AVŞAR**

Öz

Araştırmanın amacı çok kategorili puanlanan maddelerden oluşan testlerde kategori sayısının psikometrik özellikler üzerindeki etkisinin parametrik olmayan madde tepki kuramı (POMTK) modeli ile belirlenmesidir. Belirlenen amaç doğrultusunda iki farklı büyüklükte (100 ve 500), çeşitli dağılım özelliklerine sahip (normal dağılan, sağa çarpık dağılan ve sola çarpık dağılan) örneklemler için iki farklı test uzunluğunda (10 madde ve 30 madde), üç farklı sayıda kategoriye (üç, beş ve yedi) sahip maddeler simülatif olarak üretilmiştir. Kategori sayısının psikometrik özellikler üzerindeki etkisi POMTK modellerinden Mokken Homojenlik Modeli (MHM) ile araştırılmıştır. Yapılan araştırma temel araştırma olarak tasarlanmıştır. Verilerin üretilmesinde ve verilerin analizinde R Studio 3.4.0 yazılımı kullanılmıştır. R Studio yazılımında MHM'ye göre analizler Mokken paketi ile yapılmıştır. MHM'ye göre yapılan ölçekleme sonucunda kategori sayısının değişmesiyle birlikte maddelerin MHM'ye uyumunda belli bir örüntü gözlenmemiştir. Genel olarak hem kısa testlerde, hem de uzun testlerde kategori sayısının güvenirlik değerlerinin kestiriminde etkili olmadıkları gözlenmiştir. Araştırmada belirlenen test koşullarında testler MHM'ye düşük düzeyde uyumlu çıkmıştır.

Anahtar Kelimeler: çok kategorili puanlanan maddeler, kategori sayısı, parametrik olmayan madde tepki kuramı, mokken homojenlik modeli

Abstract

The aim of the research was to examine the effects of the number of categories for polytomous items on psychometric properties in a nonparametric item response theory (NIRT) model. For the purpose of the study, data sets with two different sample sizes (100 and 500) that come from different sample distribution shapes (normal distribution, positively skewed distribution, and negatively skewed distribution), two different test lengths (10 items and 30 items), and three different number of categories (three, five, and seven) were generated. The effects of the number of categories on psychometric properties of polytomous items were analyzed by Mokken Homogeneity Model (MHM) under NIRT model. The research was designed as a basic research. In the generation and analysis of data sets, R Studio 3.4.0 software was used. For analysis conducted with MHM, Mokken package was used in R Studio. According to scaling with MHM, specific pattern of item fit to MHM with changing the number of categories was not observed. In general, it was found that the number of categories has no effect on reliability estimate. It was determined that tests have weak fit to MHM under test conditions in the research.

Keywords: polytomous items, number of category, nonparametric item response theory, mokken homogeneity model

* Bu araştırma 1-3 Eylül 2016 tarihinde Akdeniz Üniversitesi'nde düzenlenen V. Eğitimde ve Psikolojide Ölçme ve Değerlendirme Kongresi'nde sözlü olarak sunulan bildiriden türetilmiştir.

** Dr. Öğretim Üyesi, Recep Tayyip Erdoğan Üniversitesi, Eğitm Fakültesi, Rize-Türkiye, asiye.sengul@erdogan.edu.tr, ORCID ID: orcid.org/0000-0001-5522-2514

DOI: 10.21031/epod.357160
Geliş Tarihi: 23.11.2017
Kabul Tarihi: 29.01.2018

GİRİŞ

Bireylerin duyuşsal özelliklerinin ölçülmesi, eğitimde ve psikolojide önemli yer tutmaktadır. Bu özellikler ölçülürken kullanılan ölçme araçlarının ikili puanlanan maddelerden daha çok tepkilerin dereceli ya da çoklu olarak sunulduğu yanıt kategorilerine sahip maddelerden oluştukları görülmektedir. Çok kategorili puanlanan maddeler, ikili puanlanan maddelerle karşılaştırıldığında tepki kategorilerinin sayısı daha fazla olduğundan ilgilenilen örtük özelliği daha geniş ranjlarda ölçebilmekte ve buna bağlı olarak örtük özellikle ilgili daha fazla bilgiye ulaşılmasını sağlamaktadır (Ostini ve Nering, 2006).

Ölçme araçlarının psikometrik özelliklerinin belirlenmesi, bu araçlara göre verilen kararların doğruluğu ve uygunluğu açısından çok önemlidir. Psikometrik özellikler, ölçmeleri etkileyen problemleri araştıran ve bu problemleri mümkün olduğunca azaltmaya çalışan bir disiplin olarak tanımlanan test kuramları ile belirlenir (Crocker ve Algina, 1986). Eğitimde ve psikolojide sıklıkla Klasik Test Kuramı'na (KTK) ve Madde Tepki Kuramı'na (MTK) göre ölçekleme yapılarak ölçme araçlarının psikometrik özelliklerinin belirlendiği görülmektedir.

Genellikle istatistiksel bir model olarak tanımlanan MTK, bireyin madde ve test performansını test eden ve bireyin performansının altında yatan yeteneğini, maddeler ve test aracılığıyla kestiren bir kuram olarak tanımlanabilir (Hambleton ve Jones, 1993). MTK literatürde son zamanlarda yapılan çalışmaların katkısıyla genel olarak parametrik madde tepki kuramı (PMTK) ve parametrik olmayan MTK (POMTK) (Sijtsma ve Molenaar, 2002) modelleri olacak şekilde genel iki başlık altında sınıflandırılmıştır.

PMTK modellerine göre yeni sayılabilen POMTK modelleri, PMTK modelleri ile karşılaştırıldığında daha az varsayım gerektirmektedir (Štochl, 2007). POMTK modelleri kısa testlerde, küçük örneklemlerde uygulamalarda kolaylık sağlayan modellerdir (Junker ve Sijtsma, 2001; Meijer, 2004; Molenaar, 2001). Literatür incelendiğinde POMTK modellerinin Mokken model ve parametrik olmayan regresyon modelleri olacak şekilde sınıflandırılabileceği görülmektedir (Şengül Avşar ve Tavşancıl, 2017). Mokken model; Monoton Homojenlik Modeli (MHM) ve İkili Monotonluk Modeli (İMM) olacak şekilde kendi için alt modellere ayrılmaktadır (Sijtsma ve Molenaar, 2002).

MHM sıralama düzeyindeki ölçmelerde, bireylerin sıralama amacıyla değerlendirilmesinde kullanılan (Štochl, 2007) ve bir testi alan bireylerin puanlarını kullanarak onları örtük özellikleri boyunca sıralayan (Tendeiro ve Meijer, 2013) bir model olarak tanımlanabilir. MHM hem ikili (1-0) puanlanan maddelerden oluşan ölçme araçlarının hem de çok kategorili puanlanan maddelerden oluşan ölçme araçlarının POMTK'ya göre ölçeklenmesini sağlayan bir modeldir. MHM, çok kategorili puanlanan maddeler için PMTK modellerinden dereceli tepki modelinin (DTM-graded response model) parametrik olmayan karşılığı olarak tanımlanmaktadır (Hemker, Sijtsma, Molenaar ve Junker, 1996; Sijtsma ve Molenaar, 2002).

Ölçme araçlarının psikometrik özelliklerinin MHM'ye göre belirlenmesinde hem ikili puanlanan maddeler hem de çok kategorili puanlanan maddeler için MHM'ye göre parametre kestirimleri, "ölçeklenebilirlik katsayısı (scalability coefficient-H)" ile yapılmaktadır (van Onna, 2004). İkili puanlanan maddeler için Loevinger (1947,1948) tarafından geliştirilen H katsayısı; Mokken (1971) tarafından MHM'de bir set içinde yer alan madde çiftleri (i, j madde çiftleri) için (H_{ij}), tek bir maddenin setteki diğer maddelerle (H_i) ve madde setlerinin tamamıyla (H) olan ilişkisini tanımlanmak için yeniden düzenlenmiştir (Mokken, 1997).

Mokken modelleriyle yapılan ölçeklemede çok önemli bir yere sahip olan H katsayısı iki veya üç parametreli lojistik modellerde yer alan "a" katsayısının (madde ayırt edicilik indeksi) parametrik olmayan karşılığı olarak yorumlanabileceği gibi ölçme araçlarının MHM'ye göre ölçeklenip ölçeklenmediğinin belirlenmesinde kullanılan ölçeklenebilirlik indeksi anlamlarını taşımaktadır (Meijer, 2004; Mokken, 1997; van Onna, 2004).

Mokken modellerinde güvenirlik hesaplamaları için Cronbach α, Guttman lambda 2 (λ) ve Rho katsayılarının raporlaştırıldığı görülmektedir (Şengül Avşar ve Tavşancıl, 2017). Mokken (1971)

tarafından önerilen ve literatürde aynı zamanda Molenaar Sijtsma (MS) istatistiği olarak bilinen Rho katsayısı Mokken modellerinden İMM'ye uygun bir güvenirlik katsayısıdır (Štochl, 2007; van der Ark, 2015). MHM'ye göre ölçeklenen ölçme araçlarının güvenirlik kestirimleri için ayrıca van der Ark, van der Palm ve Sijstma'nın (2011) geliştirdikleri örtük sınıf güvenirlik katsayısı (LCRC-latent class reliability coefficient) kullanılabilir.

POMTK modellerinden; tek boyutluluk, yerel bağımsızlık ve monotonluk varsayımlarını gerektiren MHM'nin (Sijtsma ve Molenaar, 2002) örtük değişkenle, homojen ve monoton madde karakteristik eğrilere (MKE) sahip maddeler arasındaki ilişkiyi tanımlayan bir model olduğu ifade edilmektedir (Meijer ve Baneke, 2004). Burada MHM'nin tek boyutlu PMTK modelleriyle benzer varsayımları gerektiği açık bir şekilde görülmektedir. Ancak PMTK ve POMTK modelleri karşılaştırıldığında iki model arasındaki temel farkın ikili puanlanan maddeler için MKE'lere, çok kategorili puanlanan maddeler için madde adım fonksiyonlarına bağlı olduğu bilinmektedir (Şengül Avşar ve Tavşancıl, 2017). MKE'ler PMTK modellerinde monoton ve lojistik olarak kestirilirken bu eğriler POMTK modellerinde monoton olmalarına rağmen lojistik olarak kestirilmezler (Sijtsma ve Molenaar, 2002).

Literatür taramasında yurt içinde ve yurt dışında ölçeklerin psikometrik özelliklerinin POMTK modellerine göre incelendiği çeşitli çalışmaların yapıldığı görülmüştür (Galindo Garre ve diğerleri, 2014; Sachs, Law ve Chan, 2003; Koğar, 2015; Şengül Avşar ve Tavşancıl, 2017; Young, Blodgett ve Reardon, 2003). Özellikle çok kategorili puanlanan maddeleri konu alan simülatif araştırmalarda örneklem büyüklüğü, madde sayısı, örneklemin dağılım şekli gibi çeşitli faktörlerin POMTK modellerinden MHM'ye göre elde edilen geçerlik ve güvenirlik katsayılarına etkisi araştırılırken gerçek veri setleriyle yapılan araştırmalarda ölçeklerin psikometrik özelliklerinin belirlenmesi amaçlanmıştır.

Duyuşsal özelliklerin ölçülmesinde Likert tipi ölçekler sıklıkla kullanılmaktadır. Bu ölçekler sıralama düzeyinde, çok kategorili puanlanan maddelerden oluşmaktadır. Literatürde Likert tipi, çok kategorili puanlanan maddelerden oluşan ölçme araçlarının psikometrik özelliklerini etkileyen önemli bir faktörün kategori sayısı olduğu ifade edilmektedir (Fabiola, Iwin, Jennifer ve Zaira, 2012; Leung, 2011; Lozano, García-Cueto, ve Muñiz, 2008; Maydeu-Olivares, Kramp, García-Forero, Gallardo-Pujol ve Coffman, 2009; Preston ve Colman, 2000; Weng, 2004).

Weng (2004) araştırmasında kategori sayısının psikometrik özellikler üzerindeki etkisinin literatürde farklı araştırmacılar tarafından pek çok kez araştırıldığını, bazı araştırmacıların özellikle iç tutarlılık katsayısı olarak yorumlanan Cronbach α üzerindeki etkilerinin önemli olduğunu belirtmiştir. Bununla birlikte Maydeu-Olivares ve diğerleri (2009) araştırmalarında ideal kategori sayısının ne olması gerektiği konusunda literatürde fikir birliği olmadığını ancak kategori sayısının ölçeklerin psikometrik özellikleri üzeride etkileri olduğunu belirtmiştir.

Yapılan literatür taramasında KTK kapsamında kategori sayının psikometrik özellikler üzerindeki etkisinin araştırıldığı çeşitli çalışmaların yapıldığı görülmektedir (Erkuş, Sanlı, Bağlı ve Güven, 2000; Leung, 2011; Lozano ve diğerleri, 2008; Maydeu-Olivares ve diğerleri, 2009; Preston ve Colman, 2000; Weng, 2004; Uyumaz ve Çokluk, 2016). Bu araştırmalarda genel olarak kategori sayısının güvenirlik kestirimleri üzerinde etkili olduğu, kategori sayısı arttıkça güvenirlik değerlerinin arttığı sonucuna ulaşılmıştır.

Kategori sayısının psikometrik özellikler üzerindeki etkisini PMTK modelleriyle araştıran çeşitli çalışmalar da bulunmaktadır (Fabiola ve diğerleri, 2012; İlhan ve Güler, 2017; Lee ve Paek, 2014; Maydeu-Olivares ve diğerleri, 2009). Fabiola ve diğerleri (2012) ve Maydeu-Olivares ve diğerleri (2009) kategori sayısının psikometrik özellikler üzerinde etkili olduğunu ifade ederken Lee ve Paek (2014) kategori sayısının psikometrik özellikler üzerinde etkili olmadığını ifade etmiştir. Ayrıca İlhan ve Güler (2017) yaptıkları araştırmada kategori sayısının model veri uyumu üzerinde önemli bir etkisinin olmadığını belirtmişlerdir.

Yapılan literatür taramasında küçük örneklemlere uygulanan ölçme araçlarının psikometrik özelliklerinin belirlenmesinde PMTK modellerinin oldukça sınırlayıcı olduğu, bu gibi durumlarda POMTK modellerinden MHM'nin PMTK modellerine göre kullanışlı bir model olduğu ve

uygulamalarda kolaylık sağlayabileceği ifade edilmiştir. Çok kategorili puanlanan maddelerden oluşan ölçme araçlarının psikometrik özelliklerinin incelenmesinde kategori sayısının önemli bir faktör olduğu belirtilmiştir. Yapılan bu araştırmada farklı kategori sayılarına sahip ölçme araçlarının psikometrik özelliklerinin simülatif veriler aracılığıyla uygulamalarda kolaylık sağlayan POMTK modellerinden MHM ile belirlenmesi gerekli görülmüştür.

Araştırmanın Amacı

Bu araştırmanın genel amacı, simülatif olarak üretilen "üç", "beş" ve "yedi" kategoriden oluşan 10 maddelik kısa ve 30 maddelik uzun testlerin, çeşitli dağılım özelliklerine (normal dağılan, sağa çarpık dağılan ve sola çarpık dağılan) sahip 100 ve 500 kişilik örneklemlerden oluşan test koşullarında psikometrik özelliklerinin MHM'ye göre belirlenmesidir. Belirlenen genel amaç doğrultusunda aşağıdaki sorulara yanıt aranmıştır.

Çeşitli özelliklere sahip örneklemelere uygulanan kategori sayısı ""üç", "beş" ve "yedi" olan 10 maddelik kısa ve 30 maddelik uzun testlerin MHM'ye göre elde edilen:

1. Maddelere ait model veri uyumu düzeyleri nasıldır?
2. Maddelere ait model veri uyumu değerleri için kestirilen standart hata değerleri nelerdir?
3. Testlere ait model veri uyumu değerleri nelerdir?
4. Testlere ait model veri uyumu değerleri için kestirilen standart hata değerleri nelerdir?
5. Testler için kestirilen güvenirlik değerleri (α, λ ve LCRC) nelerdir?

YÖNTEM

Bu araştırma; kategori sayısının çok kategorili puanlanan maddelerden oluşan testlerin psikometrik özellikleri üzerine etkilerinin simülatif olarak üretilen veriler aracılığıyla belirlenmesinin amaçlandığı temel bir araştırmadır. Temel araştırmalarda mevcut teorilere dayanılarak yeni bilgi ya da yeni teoriler üretilir (Freankal, Wallen ve Hyun, 2012).

Araştırmanın test koşulları Tablo 1'de verilmiştir. Araştırmada yer alan koşullar ve verilerin üretilmesi için gerekli parametrelerin seçimi, literatür taraması sonuçlarına göre daha önce yapılan benzer çalışmalar incelenerek belirlenmiştir.

Tablo 1. Test Koşulları

		Madde Sayısı					
		10 Madde			30 Madde		
Örneklemin Dağılım Şekli	Örneklem Büyüklüğü	3 kategori	5 kategori	7 kategori	3 kategori	5 kategori	7 kategori
Normal Dağılım	100	X	X	X	X	X	X
	500	X	X	X	X	X	X
Sağa Çarpık Dağılım	100	X	X	X	X	X	X
	500	X	X	X	X	X	X
Sola Çarpık Dağılım	100	X	X	X	X	X	X
	500	X	X	X	X	X	X

Tablo 1'de araştırmada; üç farklı örneklem dağılım şekli, iki farklı örneklem büyüklüğü, iki farklı test uzunluğu ve üç farklı kategori sayısı olmak üzere 36 farklı (3*2*2*3) test koşulunun incelendiği görülmektedir. Tüm bu koşullar için 100 replikasyon yapılarak toplamda 3600 veri seti oluşturulmuştur. Tablo 1'de yer alan tüm test koşullarına literatüre dayalı yapılan incelemeler sonucunda karar verilmiştir.

Örneklem büyüklüğü için POMTK'ya göre yapılacak çözümlemelerde, Molenaar (2001) 300-400 kişilik örnekleme, Ramsay (1991) ise en az 100 kişilik örnekleme ihtiyaç duyulduğunu belirtmiştir. Ayrıca veriler POMTK modellerinden MHM'nin, PMTK'da parametrik karşılığı olan DTM'ye göre üretilmiştir. DTM'ye uyum gösteren veriler MHM'ye de uyum göstermektedir (Emons, 2008). DTM için en az 500 kişiden elde edilecek kestirimlerin doğru olduğu belirtilmektedir (Jiang, Wang ve Weiss, 2016). Literatürdeki bu bilgilere dayalı olarak araştırmada 100 kişiden oluşan küçük örneklemler ve 500 kişiden oluşan büyük örneklemler tercih edilmiştir. Ayrıca araştırmada yetenek dağılımlarının normal, sağa çarpık ve sola çarpık dağıldığı durumlar incelenmiştir. Yetenek dağılımları ortalamaları sıfır, standart sapmaları bir olan normal dağılımdan üretilmiş olup Tablo 2'de yetenek dağılımlarına ilişkin betimsel istatistikler verilmiştir.

Tablo 2. Yetenek Dağılımlarına İlişkin Betimsel İstatistikler

Örneklemin Dağılım Şekli	Örneklem Büyüklüğü	Çarpıklık Katsayısı	Basıklık Katsayı	Standart Hata
Normal Dağılım	100	-0.01	-0.18	0.10
	500	-0.01	-0.06	0.04
Sağa Çarpık Dağılım	100	0.54	0.06	0.10
	500	0.59	0.24	0.04
Sola Çarpık Dağılım	100	-0.55	0.01	0.10
	500	-0.57	0.18	0.04

Çarpıklık katsayısının mutlak değerinin; birden büyük olduğu durumda örneklemlerin dağılım şekillerinin yüksek düzeyde çarpık, 0.50 ve bir arasında orta düzeyde çarpık, 0.50'den küçük olduğu durumlarda ise yaklaşık olarak simetrik olduğu ifade edilmektedir (Bulmer, 1979, s. 63). Tablo 2 incelendiğinde çarpık dağılan örneklemlerin orta düzeyde çarpık olduğu görülmektedir. Eğitimde ve psikolojide uygulamalarda normal dağılımlı veri setlerine ulaşmak hedeflenmektedir. MTK uygulamalarında da mümkün olduğunca veri setinin normal dağılımlı olması beklenir. Bu bağlamda çarpık dağılımların normal dağılımdan aşırı düzeyde sapmış olmaması gerekir. Bunun için gerçek uygulamaları düşünerek araştırmada çarpık dağılımlar orta düzeyde çarpık olacak şekilde seçilmiştir.

Madde parametreleri literatüre bağlı kalınarak normal ve tek biçimli olacak şekilde belirlenmiştir. Buna göre b parametresi $N\sim(0, 1)$ normal, a parametresi ise tek biçimli $a \in U\ [1,2]$ olacak şekilde seçilmiştir (Bahry, 2012; Cohen, Kim ve Baker, 1993; DeMars, 2002; Syu, 2013).

Literatürde test uzunluğunu ya da madde sayısını konu alan araştırmalar incelendiğinde 10'dan 80'e kadar çeşitli sayılarda maddelerden oluşan testlerin kullanıldığı görülmektedir (Lee, 2007; Lee, Wollack ve Douglas, 2009; Liang, Wells ve Hambleton, 2014; Patsula ve Gessaroli, 1995; Stone, 1992; Stone ve Zhang, 2003; Sueiro ve Abad, 2011). Bu araştırmada POMTK modellerinin kısa testlerde uyum göstermesi avantajı düşünülerek az sayıda maddelerden oluşan testler tercih edilmiştir. Buna göre araştırma kapsamında 10 maddelik testler kısa, 30 maddelik testler uzun testler olarak belirlenmiştir.

Araştırmada kategori sayılarının belirlenmesinde KTK ve MTK kapsamında kategori sayısının psikometrik özellikler üzerindeki etkisini araştıran çalışmalardan yararlanılmıştır (Fabiola ve diğerleri, 2012; Leung, 2011; Uyumaz ve Çokluk, 2016; Weng, 2004).

Verilerin Analizi

Araştırmada belirlenen test koşullarına uygun simülatif verilerin üretilmesi ve verilerin analizi R Studio 3.4.0 yazılımı ile gerçekleştirilmiştir. MHM'ye göre yapılan çözümlemeler ve güvenirlik katsayıları hesaplamaları için van der Ark (2007) tarafından geliştirilen Mokken paketi kullanılmıştır.

Araştırmada maddelerin ve testlerin MHM'ye uyumları H katsayıları ile belirlenmiştir. H katsayılarının değerlendirilmesinde Mokken (1971) tarafından tanımlanmış değerlendirme ölçütleri kullanılmaktadır. Bu ölçütler; $0.30 \leq H < 0.40$ için düşük, $0.40 \leq H < 0.50$ için orta ve $H \geq 0.50$ için yüksek olacak şekilde belirlenmiştir (Meijer ve Baneke, 2004; Mokken, 1971, 1997; Sijtsma, Debets ve Molenaar, 1990; van Onna, 2004). Araştırmada ayrıca maddelere ve testlere ait H değerleri için kestirilen standart hata değerleri incelenmiştir.

BULGULAR

Araştırma kapsamında elde edilen bulgular araştırma sorularının sırasına bağlı olarak sunulmuştur.

Maddelere Ait Model Veri Uyumu Düzeyleri

MHM'ye göre ölçeklenen maddelerin model veri uyumu değerleri olan H katsayılarının Mokken'a (1971) göre belirlenen değerlendirme ölçütleriyle incelenmesi sonucunda elde edilen değerlerin uyum düzeyleri 10 madde ve 30 madde için sırasıyla Tablo 3'te ve Tablo 4'te verilmiştir.

Tablo 3. 10 Madde İçin Model Veri Uyumu Düzeyleri

		Normal Dağılan		Sağa Çarpık Dağılan		Sola Çarpık Dağılan	
	Uyum Düzeyi	100	500	100	500	100	500
3 kategori	Düşük	8	8	7	7	7	8
	Orta	-	1	1	-	3	2
	Yüksek	-	-	-	-	-	-
5 kategori	Düşük	7	8	7	2	7	8
	Orta	2	1	1	6	3	2
	Yüksek	-	-	-	2	-	-
7 kategori	Düşük	6	7	4	6	8	8
	Orta	2	1	2	1	2	2
	Yüksek	-	-	-	-	-	-

Tablo 3 incelendiğinde çeşitli dağılım özelliklerine sahip iki farklı büyüklükteki örneklemlerden elde edilen maddelerin MHM'ye göre ölçeklenmesi sonucu elde edilen H değerlerinin uyum düzeylerinin verildiği görülmektedir. Tablo 3'te sola çarpık dağılan örneklemler hariç diğer örneklemlerde MHM'ye uyum göstermeyen maddelerin olduğu anlaşılmaktadır.

Normal dağılan ve sağa çarpık dağılan hem küçük hem de büyük örneklemlerde maddelerin MHM'ye en iyi uyum gösterdiği kategori sayısının beş olduğu söylenebilir. Sola çarpık dağılan küçük örneklemlerde maddelerin MHM'ye en iyi uyum gösterdiği kategori sayılarının üç ve beş olduğu söylenebilir. Sola çarpık dağılan büyük örneklemlerde kategori sayısının maddelerin MHM'ye uyum düzeyini etkilemedikleri görülmektedir. Araştırma kapsamında 10 maddelik testler için genel olarak maddelerin MHM'ye en iyi uyum gösterdiği kategori sayısının beş olduğu sonucuna ulaşılabilir.

Tablo 4. 30 Madde İçin Model Veri Uyumu Düzeyleri

		Normal Dağılan		Sağa Çarpık Dağılan		Sola Çarpık Dağılan	
	Uyum Düzeyi	100	500	100	500	100	500
3 kategori	Düşük	20	20	20	20	22	23
	Orta	7	6	4	3	7	7
	Yüksek	-	1	1	1	-	-
5 kategori	Düşük	23	23	21	19	22	23
	Orta	6	6	5	6	7	7
	Yüksek	-	-	-	-	-	-
7 kategori	Düşük	22	22	19	20	22	22
	Orta	7	7	7	6	6	7
	Yüksek	-	-	-	-	-	-

Tablo 4 incelendiğinde çeşitli dağılım özelliklerine sahip iki farklı büyüklükteki örneklemlerden elde edilen maddelerin MHM'ye göre ölçeklenmesi sonucu elde edilen H değerlerinin uyum düzeylerinin verildiği görülmektedir. Tablo 4'te sola çarpık dağılan büyük örneklemlerde üç ve beş kategoriden oluşan test koşulları hariç diğer tüm test koşullarında MHM'ye uyum göstermeyen maddelerin olduğu anlaşılmaktadır.

Normal dağılan hem küçük hem de büyük örneklemlerde maddelerin MHM'ye en iyi uyum gösterdiği kategori sayıları beş ve yedidir. Sağa çarpık dağılan örneklemlerde genel olarak kategori sayısı arttıkça maddelerin MHM'ye uyum düzeyleri artmıştır. Bu koşulda MHM'ye en iyi uyum kategori sayısının yedi olduğu görülse de beş ve yedi kategorili puanlanan maddeler arasında MHM'ye uyum düzeyi açısından farkın tek bir madde için olduğu belirtilmelidir. Sola çarpık dağılan hem küçük hem de büyük örneklemlerde maddelerin MHM'ye en iyi uyum gösterdiği kategori sayılarının üç ve beş kategori olduğu görülse de yedi kategorili puanlanan maddelerle bu maddeler arasında MHM'ye uyum düzeyi açısından farkın tek bir madde için olduğu belirtilmelidir.

Araştırma kapsamında 30 maddelik testler için tüm test koşulları birlikte değerlendirildiğinde, genel olarak MHM'ye en iyi uyumların kategori sayısının beş olduğu durumlarda oluştuğu sonucuna ulaşılabilir. Bununla birlikte örneklemlerin dağılım şekilleri kendi içlerinde incelendiğinde sola çarpık dağılan örneklemler hariç diğer durumlarda genel olarak kategori sayısı arttıkça maddelerin MHM'ye uyum düzeylerinin arttığı da söylenebilir.

Araştırmada belirlenen test koşullarında (farklı dağılım özelliklerine sahip farklı büyüklükteki örneklemlerden 10 maddelik kısa ve 30 maddelik uzun testler için) genel olarak kategori sayısının maddelerin MHM'ye uyum düzeyleri üzerinde çok etkili bir faktör olmadığı sonucuna ulaşılmıştır. Her ne kadar araştırmada belirlenen test koşullarının çoğunda kategori sayısı arttıkça maddelerin MHM'ye uyumu artmış gibi görünse de bu artışın tek madde ile sınırlı olduğu belirtilmelidir. Genel olarak tüm test koşulları birlikte değerlendirildiğinde MHM'ye en iyi uyuma beş kategorili puanlanan maddelerden ulaşıldığı ifade edilebilir.

Maddelere Ait Model Veri Uyumu Değerleri İçin Kestirilen Standart Hata Değerleri

Tablo 5'te farklı dağılım özelliklerine sahip farklı büyüklükteki örneklemlerden, 10 maddelik kısa ve 30 maddelik uzun testlerden elde edilen tüm maddelere ait model veri uyumu değerleri için kestirilen standart hata değerlerinin en küçük (SH_{iek}) ve en büyük (SH_{ieb}) değerleri verilmiştir.

Tablo 5. Maddelere Ait Model Veri Uyumu Değerleri İçin Kestirilen En Küçük ve En Büyük Standart Hata Değerleri

		Normal Dağılan				Sağa Çarpık Dağılan				Sola Çarpık Dağılan			
		100 kişi		500 kişi		100 kişi		500 kişi		100 kişi		500 kişi	
		SH_{iek}	SH_{ieb}	SH_{iek}	SH_{ieb}	SH_{iek}	SH_{ieb}	SH_{iek}	SH_{ieb}	SH_{iek}	SH_{ieb}	SH_{iek}	SH_{ieb}
10 madde	3	0.06	0.10	0.03	0.04	0.06	0.09	0.03	0.04	0.06	0.10	0.03	0.05
	5	0.06	0.10	0.03	0.05	0.06	0.09	0.03	0.05	0.06	0.10	0.03	0.05
	7	0.07	0.08	0.03	0.05	0.06	0.07	0.02	0.04	0.06	0.10	0.03	0.05
30 madde	3	0.05	0.08	0.02	0.05	0.05	0.10	0.02	0.05	0.05	0.10	0.02	0.05
	5	0.05	0.09	0.02	0.05	0.05	0.10	0.02	0.05	0.05	0.09	0.02	0.04
	7	0.05	0.09	0.02	0.05	0.05	0.11	0.02	0.05	0.05	0.09	0.02	0.04

Tablo 5'te MHM'ye göre 10 madde için elde edilen maddelere ait SH değerleri; normal dağılan küçük örneklemlerde en küçük 0.06, en büyük 0.10 değerlerini alırken normal dağılan büyük örneklemlerde en küçük 0.03, en büyük 0.05 değerlerini almıştır. Sağa çarpık dağılan küçük örneklemlerde SH değerleri; en küçük 0.06, en büyük 0.09 değerlerini alırken sağa çarpık dağılan büyük örneklemlerde en küçük 0.02, en büyük 0.05 değerlerini almıştır. Sola çarpık dağılan küçük örneklemlerde SH değerleri; en küçük 0.06, en büyük 0.10 değerlerini alırken sola çarpık dağılan

büyük örneklemlerde en küçük 0.03, en büyük 0.05 değerlerini almıştır. Genel olarak örneklem büyüklüğünün artmasıyla birlikte standart hata değerlerinde azalma olduğu görülmektedir.

Tablo 5'te MHM'ye göre 30 madde için elde edilen maddelere ait SH değerleri; normal dağılan küçük örneklemlerde en küçük 0.05, en büyük 0.09 değerlerini alırken normal dağılan büyük örneklemlerde en küçük 0.02, en büyük 0.05 değerlerini almıştır. Sağa çarpık dağılan küçük örneklemlerde SH değerleri; en küçük 0.05, en büyük 0.11 değerlerini alırken sağa çarpık dağılan büyük örneklemlerde en küçük 0.02, en büyük 0.05 değerlerini almıştır. Sola çarpık dağılan küçük örneklemlerde SH değerleri; en küçük 0.05, en büyük 0.10 değerlerini alırken büyük örneklemlerde en küçük 0.02, en büyük 0.05 değerlerini almıştır. Genel olarak örneklem büyüklüğünün artmasıyla birlikte standart hata değerlerinde azalma olduğu görülmektedir.

Kategori sayısı dikkate alındığında SH değerlerinde hem 10 maddelik hem de 30 maddelik testler için belli bir örüntü gözlenmemiştir. Diğer bir deyişle kategori sayısının değişmesinin SH değerleri üzerinde önemli bir etkisinin olmadığı sonucuna ulaşılmıştır. SH değerleri üzerinde en etkili faktörün örneklem büyüklüğü olduğu görülmektedir. Örneklem büyüklüğü arttıkça SH değerleri azalmaktadır.

Testlere Ait Model Veri Uyumu Değerleri ve Bu Değerler İçin Kestirilen Standart Hata Değerleri

Tablo 6'da testlere ait model veri uyumu değerleri ve bu değerler için kestirilen standart hata değerleri verilmiştir.

Tablo 6. Testlere Ait Model Veri Uyumu Değerleri ve Bu Değerler İçin Kestirilen Standart Hata Değerleri

		Normal Dağılan				Sağa Çarpık Dağılan				Sola Çarpık Dağılan			
		100 kişi		500 kişi		100 kişi		500 kişi		100 kişi		500 kişi	
		H	SH	H	SH	H	SH	H	SH	H	SH	H	SH
10 madde	3	0.33*	0.05	0.33*	0.02	0.32*	0.04	0.31*	0.02	0.35*	0.05	0.35*	0.02
	5	0.34*	0.05	0.34*	0.02	0.33*	0.04	0.43**	0.02	0.36*	0.05	0.36*	0.02
	7	0.33*	0.05	0.34*	0.02	0.33*	0.04	0.32*	0.02	0.36*	0.05	0.35*	0.02
30 madde	3	0.34*	0.03	0.35*	0.02	0.35*	0.03	0.34*	0.02	0.36*	0.04	0.36*	0.02
	5	0.36*	0.03	0.35*	0.02	0.35*	0.03	0.35*	0.02	0.36*	0.04	0.36*	0.02
	7	0.37*	0.04	0.36*	0.02	0.36*	0.03	0.35*	0.02	0.36*	0.03	0.36*	0.02

* düşük uyum, ** orta uyum

Tablo 6'da verilen testlere ait model veri uyumu değerleri Mokken'a (1971) göre belirlenen değerlendirme ölçütleriyle incelendiğinde, kısa testler için MHM'ye en iyi uyuma beş kategorili puanlanan maddeler için sağa çarpık dağılan 500 kişilik örneklemden ulaşılmıştır. Bu durumda testlerin MHM'ye uyumu orta düzeydedir. Diğer tüm durumlarda testlerin MHM'ye düşük düzeyde uyumlu oldukları görülmektedir. Kategori sayısı dikkate alındığında kısa testlerde genel olarak en iyi uyuma beş kategoriden ulaşılmıştır. Ancak burada H katsayısının sayısal değeri üzerinden yorum yapılmaktadır. Genel olarak testlerin MHM'ye uyum düzeylerinin Mokken'a (1971) göre belirlenen değerlendirme ölçütlerine göre düşük düzeyde oldukları unutulmamalıdır.

Benzer inceleme 30 maddelik uzun testler için yapıldığında tüm test koşullarının MHM'ye uyum düzeyinin düşük düzeyde olduğu görülmektedir. Kategori sayısı dikkate alındığında uzun testlerde genel olarak en iyi uyuma yedi kategoriden ulaşılmıştır. Ancak burada H katsayısının sayısal değeri üzerinden yorum yapılmaktadır. Genel olarak testlerin MHM'ye uyum düzeylerinin Mokken'a (1971) göre belirlenen değerlendirme ölçütlerine göre düşük düzeyde oldukları unutulmamalıdır.

Tablo 6'da verilen SH değerleri incelendiğinde hem kısa hem de uzun testlerde örneklem büyüklüğünün artmasıyla birlikte bu değerlerin azaldığı görülmektedir. Genel olarak kategori sayısı SH değerleri üzerinde etkili değildir.

Testler İçin Kestirilen Güvenirlik Değerleri (α, λ ve LCRC)

Tablo 7'de testler için kestirilen α, λ ve LCRC güvenirlik değerleri verilmiştir.

Tablo 7. Testler İçin Kestirilen Güvenirlik Değerleri

Dağılım Şekli	Örneklem Büyüklüğü	Güvenirlik Katsayısı	10 madde			30 madde		
			3	5	7	3	5	7
Normal Dağılan	100	α	0.74	0.74	0.73	0.91	0.91	0.92
		λ	0.75	0.76	0.75	0.92	0.92	0.93
		LCRC	0.80	0.79	0.78	0.94	0.93	0.93
	500	α	0.75	0.75	0.75	0.91	0.91	0.92
		λ	0.75	0.75	0.76	0.91	0.92	0.92
		LCRC	0.79	0.78	0.78	0.92	0.92	0.93
Sağa Çarpık Dağılan	100	α	0.73	0.73	0.76	0.91	0.91	0.92
		λ	0.74	0.75	0.77	0.91	0.92	0.92
		LCRC	0.80	0.79	0.79	0.93	0.93	0.93
	500	α	0.73	0.79	0.74	0.91	0.91	0.92
		λ	0.74	0.80	0.75	0.91	0.92	0.92
		LCRC	0.77	0.82	0.77	0.92	0.92	0.92
Sola Çarpık Dağılan	100	α	0.76	0.76	0.77	0.91	0.92	0.92
		λ	0.77	0.77	0.78	0.91	0.92	0.92
		LCRC	0.82	0.81	0.80	0.93	0.93	0.93
	500	α	0.76	0.76	0.77	0.91	0.92	0.92
		λ	0.76	0.77	0.78	0.91	0.92	0.92
		LCRC	0.79	0.79	0.79	0.92	0.93	0.93

Tablo 7'de verilen testlere ait güvenirlik kestirimleri incelendiğinde örneklemlerin tüm dağılım şekillerinde 30 maddeden oluşan uzun testlerden elde edilen güvenirlik katsayılarının, 10 maddeden oluşan kısa testlerden elde edilen güvenirlik katsayılarına göre daha yüksek olduğu görülmektedir. Kategori sayısının değişmesiyle birlikte güvenirlik kestirimlerinin birbirlerinden çok farklı değerler almadıkları görülmektedir. Diğer bir deyişle kategori sayısının MHM'ye göre yapılan güvenirlik kestirimlerinde etkili bir faktör olduğu söylenemez. Belirlenen test koşullarında güvenilirlik kestirimi üzerinde en etkili faktörün madde sayısı olduğu görülmektedir. Madde sayısının artmasıyla birlikte tüm test koşullarında güvenirlik katsayıları artmıştır. Ayrıca Tablo 7'de görüldüğü gibi genel olarak α güvenirlik katsayısı, λ ve LCRC güvenirlik katsayılarının altında değerler vermiştir.

SONUÇLAR ve TARTIŞMA

Bu araştırmada simülatif veriler kullanılarak kategori sayısının maddelerin psikometrik özellikleri üzerindeki etkisinin POMTK modellerinden MHM ile kestirimlerinin incelenmesi amaçlanmıştır. Araştırmanın amacı doğrultusunda iki farklı büyüklükteki çeşitli dağılım özelliklerine sahip (normal dağılan, sağa çarpık dağılan ve sola çarpık dağılan) örneklemler için iki farklı test uzunluğunda (10 madde (kısa) ve 30 madde (uzun)), üç farklı sayıda kategoriye (üç, beş ve yedi) sahip maddeler 100 replikasyon yapılarak simülatif olarak üretilmiştir.

Araştırmada oluşturulan test koşullarında örneklemlerin dağılım özelliği ve büyüklüğüne göre MHM'ye uyum göstermeyen maddelerin olduğu belirlenmiştir. Kısa testlerde en fazla uyumsuzluk yedi kategoride puanlanan maddeler için sağa çarpık dağılan küçük örneklemlerde gözlenmiştir. Burada dört maddenin MHM'ye uyum göstermediği belirlenmiştir. Uzun testlerde en fazla uyumsuzluk üç kategoride puanlanan maddeler için sağa çarpık dağılan büyük örneklemlerde gözlenmiştir. Burada altı maddenin MHM'ye uyum göstermediği belirlenmiştir. Bunların dışındaki diğer koşullarda maddelerin tamamına yakının MHM'ye uyum gösterdiği belirlenmiştir. Elde edilen bu sonuç araştırma kapsamında oluşturulan koşullarda küçük örneklemlere uygulanan testlerin MHM'ye uyumlu olduğunu göstermektedir. Araştırmanın bu bulgusu, literatürde belirtilen küçük

örneklemlere uygulanan kısa testlerin MHM'ye uyumlu olduğu bulgusuyla (Junker ve Sijtsma, 2001; Meijer, 2004; Molenaar, 2001) paralellik göstermektedir. Ayrıca araştırma kapsamında oluşturulan test koşullarında testlerden elde edilen H katsayıları incelendiğinde Mokken (1971) tarafından belirlenen değerlendirme ölçütlerine göre testlerin genelde MHM'ye düşük düzeyde uyumlu oldukları belirlenmiştir.

Araştırmada belirlenen test koşullarında maddelerin kategori sayısının değişmesiyle birlikte MHM'ye uyum düzeyinde belli bir örüntüye göre değişim gözlenmemiştir. Ancak genel olarak uyum düzeyi değişmemekle birlikte kategori sayısının artmasıyla H katsayılarının daha yüksek değerler aldığı gözlenmiştir. MHM'ye uyum göstergesi olarak tanımlanan H katsayılarının değerleri arttıkça daha ayırt edici maddelere ulaşılmaktadır (Sijtsma ve Molenaar, 2002). Bu bağlamda MHM'ye uyum düzeyinin artması, maddelerin ve testlerin psikometrik açıdan daha geçerli ölçme yapabildiğini göstermektedir. Yapılan bu araştırmada kategori sayısının artması testlerin MHM'ye uyumunu değiştirmemiştir. Diğer bir deyişle araştırma koşullarında kategori sayısının MHM'ye göre yapılan ölçeklemede geçerlik üzerinde etkisi olmamıştır. Araştırmanın bu bulgusu kategori sayısının artmasına bağlı olarak ölçme geçerliliğinin arttığını ifade eden KTK kapsamında yürütülen bazı araştırmalardan (Lozano, García-Cueto ve Muñiz, 2008; Preston ve Colman, 2000) farklı çıkmıştır. Ayrıca KTK kapsamında yürütülen ve kategori sayısının geçerliği etkilemediğini ifade eden araştırmalara (Erkuş, Sanlı, Bağlı ve Güven, 2000; Maydeu-Olivares ve diğerleri, 2009; Uyumaz ve Çokluk, 2016) paralel çıkmıştır. Yine araştırmanın bu bulgusu PMTK kapsamında yürütülen araştırmalardan İlhan ve Güler (2017) ve Lee ve Paek (2014) tarafından yürütülen araştırmanın bulgularına paralel çıkarken Fabiola ve diğerlerinin (2012) ve Maydeu-Olivares ve diğerlerinin (2009) bulgularından farklılaşmıştır.

Araştırmada maddeler ve testler için kestirilen SH değerleri incelendiğinde, kategori sayısının bu değerler üzerinde etkili olmadığı belirlenmiştir. SH değerleri için yapılan kestirimler örneklem büyüklüğünün artmasıyla birlikte azalmıştır. Diğer bir deyişle örneklem büyüklüğü arttıkça H kestirimleri için yapılan hata değerleri azalmıştır. Araştırmanın bu bulgusu literatüre paraleldir (Smits, Timmerman ve Meijer, 2012; Şengül Avşar ve Tavşancıl, 2017; Koğar, 2015).

Araştırmada madde sayısına bağlı olarak tüm güvenirlik değerlerinin arttığı görülmüştür. Madde sayısının artmasıyla birlikte güvenirliğin de arttığı (Crocker ve Algina,1986) bilinmektedir. Ayrıca araştırmanın pek çok koşulunda kategori sayısının artmasıyla birlikte güvenirlik değerlerinde küçük artışlar gözlenmiştir. Araştırmanın bu bulgusu KTK kapsamında yürütülen araştırma bulgularına paraleldir (Erkuş, Sanlı, Bağlı ve Güven, 2000; Leung, 2011; Lozano, García-Cueto ve Muñiz, 2008; Maydeu-Olivares ve diğerleri, 2009; Preston ve Colman, 2000, Uyumaz ve Çokluk, 2016; Weng, 2004). PTMK kapsamında yürütülen araştırmalarda da (Maydeu-Olivares ve diğerleri, 2009; Lee ve Paek, 2014; Pozehl, 1990; Wang, 2004; Zenisky, Hambleton ve Sireci, 2002; Zhang, 2010) benzer sonuçlara ulaşılmıştır.

Çok kategorili puanlanan maddelerden oluşan testlerin, MHM'ye göre kestirilen bir başka güvenirlik değeri olan LCRC'nin de yüksek kestirildiği görülmüştür. Bu bulgu Rivas, Bersabé ve Berrocal'ın (2005) yaptıkları araştırmada belirtilen MHM'ye göre yapılan ölçekleme ile yüksek güvenirlikte testlere ulaşılacağı bulgusunu desteklemektedir. Araştırmada en yüksek güvenirlik kestirimlerine küçük farklarla da olsa hem kısa hem de uzun testlerde genellikle beş ve yedi kategorili puanlanan maddelerden ulaşılmıştır. Araştırmanın bu sonucu, araştırmalarında aynı kategori sayılarını seçen İlhan ve Güler'in (2017) araştırma bulgusuna paraleldir.

Araştırmadan elde edilen sonuçlar özetlendiğinde kategori sayısının MHM'ye göre yapılan kestirimler üzerinde çok büyük etkilerinin olmadığı sonucuna ulaşılmıştır. Genellikle kategori sayısının artmasıyla birlikte geçerlik katsayısı gibi yorumlanabilen H katsayılarının MHM'ye uyum düzeylerini değiştirmediği ancak bazı durumlarda uyum düzeyi aynı kalmakla birlikte daha yüksek değerler aldığı görülmüştür. Kategori sayısının güvenirlik kestirimleri için seçilen α, λ ve LCRC katsayıları üzerinde etkili bir faktör olmadığı belirlenmiştir. Madde sayısı arttıkça α, λ ve LCRC katsayıları yüksek değerler almıştır. Ayrıca kategori sayısı, maddeler ve testler için kestirilen SH değerleri üzerinde etkili bulunmamıştır.

Araştırmadan elde edilen sonuçlardan POMTK modellerinden MHM'nin kısa testlerde ve küçük örneklemlerde uyumlu olduğu görülmüştür. Bu nedenle küçük örneklemlere ulaşabilen araştırmacılar için verilerini MHM'ye göre ölçeklemeleri önerilebilir. Araştırma sonuçlarından özellikle çok kategorili puanlanan maddelerden oluşan ve MHM'ye göre ölçekleme yaparak ölçek geliştirmek isteyen araştırmacıların beş kategorili puanlanan maddeleri kullanmaları önerilebilir. Ancak bu önerilerde araştırmanın sınırlılıkları göz önünde bulundurulmalıdır.

Araştırma kapsamında getirilen önerilere ek olarak simülatif bir çalışma olarak yürütülen bu araştırmada belirlenen test koşulları dışında, farklı madde ve kategori sayılarına sahip testler, değişen örneklem büyüklükleri ve değişen çarpıklık durumları gibi farklı test koşulları oluşturularak yeni çalışmalar yapılabilir. Gerçek veri setleriyle de benzer bir çalışma yapılarak bu araştırmanın sonuçlarıyla karşılaştırılabilir.

KAYNAKÇA

Bahry, L. M. (2012). *Polytomous item response theory parameter recovery: An investigation of nonnormal distributions and small sample size* (Master's Thesis). Available from ProQuest Dissertations and Theses database. (UMI No. MR90146)

Bulmer, M. G. (1979). *Principles of statistics.* New York: Dover Publications.

Cohen, A. S., Kim, S. H., & Baker, F. B. (1993). Detection of differential item functioning in the graded response model. *Applied Psychological Measurement*, *17*(4), 335-350. doi:10.1177/01466216930170040

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. Orlando: Harcourt Brace Jovanovich Inc.

DeMars, C. E. (2002, April). *Recovery of graded response and partial credit parameters in Multilog and Parscale.* Paper presented at the annual meeting of American Educational Research Association, Chicago.

Emons, W. H. M. (2008). Nonparametric person-fit analysis of polytomous item scores. *Applied Psychological Measurement, 32*(3), 224- 247. doi:10.1177/0146621607302479

Fraenkel, J. R., Wallen, N. E., & Hyun, H. H. (2012). *How to design and evaluate research in education* (8th Edition). New York: McGraw-Hill.

Erkuş, A., Sanlı, N., Bağlı, M. ve Güven, K. (2000). Öğretmenliğe ilişkin tutum ölçeği geliştirilmesi. *Eğitim ve Bilim, 25*(116), 27-33. http://egitimvebilim.ted.org.tr/index.php/EB/article/view/5276/1439 adresinden erişildi.

Fabiola, G., Iwin, L., Jennifer, L., & Zaira, V. (2012). The effect of the number of answer choices on the psychometric properties of stress measurement in an instrument applied to children. *Evaluar*, *12*, 43-59. Retrieved from https://revistas.unc.edu.ar/index.php/revaluar/article/viewFile/4694/4488

Galindo-Garre, F., Hendriks, S. A., Volicer, L., Smalbrugge, M., Hertogh, C. M., & van der Steen, J. T. (2014). The bedford alzheimer nursing-severity scale to assess dementia severity in advanced dementia: A nonparametric item response analysis and a study of its psychometric characteristics. *American Journal of Alzheimer's Disease and Other Dementias, 29*(1), 84-90. doi:10.1177/1533317513506777

Hambleton, R. K., & Jones, R. W. (1993). Comparison of classical test theory and item response theory and their applications to test development. *Educational Measurement, 12*, 38-47. doi:10.1111/j.1745-3992.1993.tb00543.x

Hemker, B. T., Sijtsma, K., Molenaar, I. W., & Junker, B. W. (1996). Polytomous IRT models and monotone likelihood ratio of the total score. *Psychometrika, 61*(4), 679-693.

İlhan, M., & Güler, N. (2017). The number of response categories and the reverse directional item problem in likert-type scales: A study with the rasch model. *Journal of Measurement and Evaluation in Education and Psychology, 8*(3), 321-343. doi:10.21031/epod.321057

Jiang, S., Wang, C., & Weiss, D. J. (2016). Sample size requirements for estimation of item parameters in the multidimensional graded response model. *Frontiers in psychology*, *7*, 109, 1-10. doi:10.3389/fpsyg.2016.00109

Junker, B., & Sijtsma, K. (2001). Nonparametric item response theory in action: An overview of the special issue. *Applied Psychological Measurement*, *25*(3), 211- 220. doi:10.1177/01466210122032028

Koğar H. (2015). Madde tepki kuramına ait parametrelerin ve model uyumlarının karşılaştırılması: Bir monte carlo çalışması. *Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi*, *6*, 142-157. doi:10.21031/epod.02072

Lee, Y. S. (2007). A comparison of methods for nonparametric estimation of item characteristic curves for binary items. *Applied Psychological Measurement, 31*(2), 121–134. doi:10.1177/0146621606290248

Lee, J., & Paek, I. (2014). In search of the optimal number of response categories in a rating scale. *Journal of Psychoeducational Assessment, 32*(7), 663-673. doi:10.1177/0734282914522200

Lee, Y. S., Wollack, J. A., & Douglas, J. (2009). On the use of nonparametric item characteristic curve estimation techniques for checking parametric model fit. *Educational and Psychological Measurement, 69*(2), 181–197. doi:10.1177/0013164408322026

Leung, S. O. (2011). A comparison of psychometric properties and normality in 4-, 5-, 6-, and 11-point likert scales. *Journal of Social Service Research, 37*(4), 412-421. doi:10.1080/01488376.2011.580697

Liang, T., Wells, C. S., & Hambleton, R. K. (2014). An assessment of nonparametric approach for evaluating the fit of item response models. *Journal of Educational Measurement, 51*(1), 1–17. doi:10.1111/jedm.12031

Lozano, L. M., García-Cueto, E., & Muñiz, J. (2008). Effect of the number of response categories on the reliability and validity of rating scales. *Methodology*: *European Journal of Research Methods for the Behavioral and Social Sciences, 4*(2), 73-79. doi:10.1027/1614-2241.4.2.73

Maydeu-Olivares, A., Kramp, U., García-Forero, C., Gallardo-Pujol, D., & Coffman, D. (2009). The effect of varying the number of response alternatives in rating scales: Experimental evidence from intra-individual effects. *Behavior Research Methods, 41*(2), 295-308. doi:10.3758/BRM.41.2.295

Meijer, R. R. (2004, March). *Investigating the quality of items in cat using nonparametric IRT* (Report No. 04-05). Law School Admission Council Computerized Testing Report. A Publication of the Law School Admission Council.

Meijer, R. R., & Baneke, J. J. (2004). Analyzing psychopathology items: A case for nonparametric item response theory modeling. *Psychological Methods, 9*(3), 354-368. doi:10.1037/1082-989X.9.3.354

Mokken, R. J. (1971). *A theory and procedure of scale analysis: With applications in political research*. The Hague: Mouton.

Mokken, R. J. (1997). Nonparametric models for dichotomous responses. In W. J. van der Linden, & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 351-368). New York: Springer-Verlag.

Molenaar, I. W. (2001). Thirty years of nonparametric item response theory. *Applied Psychological Measurement, 25*(3), 295-299. doi:10.1177/01466210122032091

Ostini, R., & Nering, M. L. (2006*). Polytomous item response theory models*. Thousand Oaks, CA: Sage

Patsula, N. L., & Gessaroli, E. M. (April, 1995). *A comparison of item parameter estimates and iccs produced with testgraf and bilog under different test lengths and sample sizes.* Paper presented at the Annual Meeting of the National Council on Measurement in Education, San Francisco, CA.

Pozehl, J. B. (1990). *Application of item response theory to criterion-referenced measurement: An investigation of the effects of model choice, sample size, and test length on reliability and estimation accuracy* (Doctoral Dissertation). Available from ProQuest Dissertations and Theses database. (UMI No. 9030146)

Preston, C. C., & Colman, A. M. (2000). Optimal number of response categories in rating scales: Reliability, validity, discriminating power, and respondent preferences. *Acta Psychologica, 104*(1), 1-15. doi:10.1016/S0001-6918(99)00050-5

Ramsay, J. O. (1991). Kernel smoothing approaches to nonparametric item characteristic curve estimation. *Psychometrika, 56*(4), 611-630.

Rivas, T., Bersabé, R., & Berrocal, C. (2005). Application of double monotonicity model to polytomous items: Scalability of the beck depression items on subjects with eating disorders. *European Journal of Psychological Assessment, 21*(1), 1-10. doi:10.1027//1015-5759.21.1.1

Sachs, J., Law, Y. K., & Chan, C. K. K. (2003). A nonparametric item analysis of a selected item subset of the learning process. *British Journal of Educational Psychology, 73*(3), 395–423. doi:10.1348/000709903322275902

Sijtsma, K., & Molenaar, W. I. (2002). *Introduction to nonparametric item response theory*. USA: Sage Publications.

Sijtsma, K., Debets, P., & Molenaar, W. I. (1990). *Mokken scale analysis for polychotomous items: Theory, a computer program and an empirical application.* Netherlands: Quality and Quantity, Kluwer academic publishers.

Smits, I. A. M., Timmerman, M. E., & Meijer, R. R. (2012). Exploratory mokken scale analysis as a dimensionality assessment tool: Why scalability does not imply unidimensionality. *Applied Psychological Measurement, 36*(6), 516-539. doi:10.1177/0146621612451050

Stone, C. A. (1992). Recovery of marginal maximum likelihood estimates in the two parameter logistic response model: An evaluation of Multilog. *Applied Psychological Measurement, 16*(l), 1-16. doi:10.1177/014662169201600101

Stone, C. A., & Zhang, B. (2003). Assessing goodness of fit of item response theory models: A comparison of traditional and alternative procedures. *Journal of Educational Measurement, 40*(4), 331-352. doi:10.1111/j.1745-3984.2003.tb01150.x

Štochl, J. (2007). Nonparametric extension of item response theory models and its usefulness for assessment of dimensionality of motor tests. *Acta Universitatis Carolinae*, *42*(1), 75-94.

Sueiro, M. J., & Abad, F. J. (2011). Assessing goodness of fit in item response theory with nonparametric models: A comparison of posterior probabilities and kernel smoothing approach. *Educational and Psychological Measurement, 71*(5), 834-848. doi:10.1177/0013164410393238

Syu, J. J. (2013). *Applying person fit-in faking detection-the simulation and practice of non parametric item response theory*. (Doctoral Dissertation, National Chengchi University). Retrieved from http://nccur.lib.nccu.edu.tw/bitstream/140.119/58646/1/251501.pdf

Şengül Avşar, A., & Tavşancıl, E. (2017). Examination of polytomous items' psychometric properties according to nonparametric item response theory models in different test conditions. *Educational Sciences: Theory & Practice*, *17*(2). doi:10.12738/estp.2017.2.0246

Tendeiro, J. N., & Meijer, R. R. (2013). The probability of exceedance as a nonparametric person fit statistic for tests of moderate length. *Applied Psychological Measurement, 37*(8), 653–665. doi:10.1177/0146621613499066

Uyumaz, G., & Çokluk, Ö. (2016). An investigation of item order and rating differences in likert-type scales in terms of psychometric properties and attitudes of respondents. *Journal of Theoretical Educational Science, 9*(3), 400-425. doi:10.5578/keg.10011

van der Ark, L. A. (2007). Mokken scale analysis in R. *Journal of Statistical Software, 20*(11), 1-19.

van der Ark, L. A. (2015). Package 'mokken'. Retrieved from http://cran.rproject.org/web/packages/mokken/mokken.pdf

van der Ark, L. A., van der Palm, D. W., & Sijtsma, K. (2011). A latent class approach to estimating test-score reliability. *Applied Psychological Measurement*, *35*(5), 380-392. doi:10.1177/0146621610392911

van Onna, M. J. H. (2004). Estimates of the sampling distribution of scalability coefficient h. *Applied Psychological Measurement, 28*(6), 427-449. doi:10.1177/0146621604268735

Wang, W. C. (2004). Direct estimation of correlation as a measure of association strength using multidimensional item response models. *Educational and Psychological Measurement*, *64*(6), 937-955. doi:10.1177/0013164404268671

Weng, L. J. (2004). Impact of the number of response categories and anchor labels on coefficient alpha and test-retest reliability. *Educational and Psychological Measurement*, *64*(6), 956-972. doi:10.1177/0013164404268674

Young, M. A., Blodgett, C., & Reardon, A. (2003). Measuring seasonality: Psychometric properties of the seasonal pattern assessment questionnaire and the inventory for seasonal variation. *Psychiatry Research, 117*(1), 75-83. doi: 10.1016/S0165-1781(02)00299-8

Zhang, O. (2010). *Polytomous IRT or testlet model: An evaluation of scoring models in small testlet size situations* (Master's Thesis, University of Florida). Retrieved from http://ufdc.ufl.edu/UFE0042638/00001

Zenisky, A. L., Hambleton, R. K., & Sireci, S. G. (2002). Identification and evaluation of local item dependencies in the medical college admissions test. *Journal of Educational Measurement*, *39*(4), 291 -309. doi:10.1111/j.1745-3984.2002.tb01144.x

EXTENDED ABSTRACT

Introduction

Measuring the affective properties of individuals takes an important place in education and psychology. It is seen that the measurement tools used to measure these properties are mostly composed of items that have response categories in which graded or polytomous response are presented, rather than items that are scored binary.

The identification of the psychometric properties of measurement tools is very important in terms of the accuracy and suitability of the decisions made according to these tools. Psychometric properties are determined by various test theories such as Classical Test Theory (CTT) and Item Response

Theory (IRT). These theories are often used in educational and psychological research. With the contributions of recent studies conducted in the literature, IRT has been classified as Parametric Item Response Theory (PIRT) and Non-parametric Item Response Theory (NIRT) (Sijtsma and Molenaar, 2002).

NIRT models are models that provide convenience in applications for short tests and small samples. Upon analysing the literature, it is seen that NIRT models can be classified as Mokken model and nonparametric regression models (Şengül Avşar and Tavşancıl, 2017). Mokken model is divided into sub-models as Monotone Homogeneity Model (MHM) and Double Monotonicity Model (DMM) (Sijtsma and Molenaar, 2002).

MHM, which is one of the NIRT models, is a statistical measurement model that requires fewer assumptions compared to PIRT models. MHM requires unidimensionality, local independence, and monotonicity assumptions (Sijtsma and Molenaar, 2002). It is seen clearly that MHM requires similar assumptions as unidimensional PIRT models. However, when the PIRT and NIRT models are compared, it is known that the essential difference between the two models depends on the item characteristic curves (ICC) (Şengül Avşar and Tavşancıl, 2017). While ICCs are monotonically and logistically estimated in PIRT models, these curves cannot be estimated in NIRT models logistically, even though they are monotone (Sijtsma and Molenaar, 2002).

Parameter estimations according to MHM are made with "scalability coefficient (H)". The H coefficient, which is developed by Loevinger (1947,1948) for the binary scored items has been regulated by Mokken (1971) to identify the relationship of a single item with other items in the set (H_i) and the entire set of items (H) for the item pairs (i, j item pairs) (H_{ij}) that are in a set in MHM (Mokken, 1997). Aside from its being interpreted as the nonparametric equivalent of the coefficient "a" (item discrimination index) in the logistic models with two or three parameters, the H coefficient also means scalability index which is employed in determining whether the measurement tools are scaled according to MHM (Meijer, 2004; Mokken, 1997; van Onna, 2004). Reliability estimations in MHM are made with Cronbach α, Guttman lambda 2 (λ), and latent class reliability coefficient (LCRC).

It is observed in the literature review that several studies in which the psychometric properties of the scales are examined according to the NIRT models are carried out in Turkey and abroad (Galindo Garre et al., 2014; Sachs, Law, and Chan, 2003; Koğar, 2015; Şengül, Avşar, and Tavşancıl, 2017; Young, Blodgett and Reardon, 2003). Especially in the studies on the polytomous items, researches are carried out on both simulated and real data sets according to various factors such as sample size, the number of items and distribution of sample.

Another factor affecting the psychometric properties of measurement tools, which are composed of polytomous items and used in the measurement of affective properties, is the number of categories (Leung, 2011; Lozano, García-Cueto, and Muñiz, 2008; Preston and Colman, 2000; Weng, 2004). In this context, it is seen that various studies investigating the effect of category number on psychometric properties within the scope of CTT and PIRT have been made.

It is seen in the literature that MHM is a useful model in determining the psychometric properties of measurement tools and the number of categories is an important factor in analysing the psychometric properties of measurement tools consisting of polytomous items. It has been found necessary to determine the psychometric properties of the measurement tools with different category numbers via the simulated data sets with the MHM from the NIRT models.

Method

This research is a basic research which aims to determine the effects of the number of categories on the psychometric properties of tests consisting of polytomous items via simulated data.

The study consists of 36 different test conditions; two different sample size (100 and 500), three different sample distribution shapes (normal distribution, positively skewed distribution and

negatively skewed distribution), two different test lengths (10 items and 30 items), three number of categories (three, five and seven). The conditions and the selection of the parameters required for the generation of data in the study are determined by examining similar studies that are conducted in the literature. The data sets are produced in the conditions indicated in the research by replicating 100 times.

R Studio 3.4.0 software is employed to generate the simulated data. R Studio 3.4.0 software was used for making analyses of the data according to the MHM and for the LCRC reliability coefficient calculations, and the Mokken package which was developed by van der Ark (2007) was used in the R Studio software.

Results and Discussion

When the results gathered from the study are summarized, it was observed that there was not any specific pattern of item fit to MHM with changing the number of categories of polytomous items. It was found that tests have weak fit to MHM under test conditions in the research. In general, it was seen that the number of categories has no effect on reliability estimates for both short and long tests. Reliability coefficients α, λ, and LCRC are estimated as having higher values for long tests than short tests. Nevertheless, it should be pointed out that the reliability estimations are generally high, and the differences emerged depending on the number of categories are not immense. Furthermore, the number of categories was not found to be influential in standard error values estimated for the tests. In addition to the fact that findings obtained from the study are obtained from estimations made with MHM, which is a NIRT model, the findings are also parallel to some findings in the literature obtained from the analyses conducted with PIRT and CTT.