ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
2018; 9(1);17-32

Madde Tepki Modellemesinde Genellenebilirlik İle İki Yüzeyli Desenlerin İncelenmesi*

Investigation of Two Facets Design With Generalizability In Item Response Modeling

Gülden KAYA UYANIK** Selahattin GELBAL***

Öz

Bu çalışmada, Madde Tepki Modellemesinde Genellenebilirlik (MTMG) yaklaşımı iki yüzeyli bx(m:t) deseni ile incelenmiş ve Genellenebilirlik Kuramından (GK) elde edilen sonuçlar ile karşılaştırılmıştır. Çalışmada simülasyon verisi kullanılmıştır. Genellenebilirlik Kuramı doğrusal veri seti bx (m:t) dengelenmiş rastgele deseni için üretilmiştir. Üretilen veriler madde takımı etkisi, madde takımı uzunluğu ve madde takımı sayısı açısından farklılık göstermektedir. Veriler toplamda iki evrenden ve her evren dört farklı koşuldan oluşmaktadır. Araştırmanın sonucu tüm evrenlere ait koşulların varyans kestirimlerinin MTMG yaklaşımı ve GK ile elde edilen sonuçlar arasında bir fark olmadığını göstermektedir. Elde edilen bu sonuç MTMG yaklaşımını ortaya atan ve tek yüzeyli desen üzerinde inceleyen Briggs ve Wilson'ın yapmış oldukları çalışma ile desteklenmektedir. MTMG yaklaşımı ve GK ile kestirilen değerler arasında fark yoktur; ancak MTMG yaklaşımında hata varyansı etkileşim varyansından ayrı olarak gözlenebilir. Çalışmada ayrıca madde takımları güvenirliği farklı koşullar altında incelenmiştir. Birey-madde takımı etkileşiminin küçük olduğu durumlarda etkileşimin büyük olduğu durumlara göre daha yüksek güvenirlik elde edilmiştir. Bunun yanında madde takımı etkisi arttıkça güvenirliğin düştüğü gözlenmiştir. Ayrıca tüm evrenlere ait koşullar incelendiğinde madde takımları için madde sayısı artıkça güvenirliğin arttığı gözlenmiştir.

Anahtar Kelimeler: Madde tepki modellemesinde genellenebilirlik, genellenebilirlik kuramı, madde takımı, madde sayısı, güvenirlik

Abstract

An approach called generalizability in item response modeling (GIRM) is investigated with two facets sx(i:t) design and results are compared with results of generalizability theory in this study. In this study simulated data is used. In Generalizability Theory linear model random facets balanced bx(m:h) design are used for generating data. Generated data are differed by factors. These factors are testlet effect, testlet length and number of testlet. All generated data consist of two different universes and all universes have four different conditions. According to the results of this study the estimates of variance components obtained using GIRM approach are generally quite similar to those obtained using GT approach. Briggs and Wilson's study is supported this result. There is no difference between results of GIRM and GT but error variance could be separated from residual variance with GIRM. This study also examines the reliability of testlets under different conditions. Testlets are more reliable when person-item variance is smaller. Furthermore, when testlet effect is increased,reliability is decreased. When conditions of all universes are investigated it is concluded that it is effective to have more items to increase reliability.

Keywords: Generalizability in item response modeling, generalizability theory, testlet, number of item, reliability

* Bu çalışma Gülden KAYA UYANIK'a ait doktora tezinden üretilmiştir.

** Doktor Öğretim Üyesi, Sakarya Üniversitesi, Eğitim Fakültesi, Sakarya-Türkiye, guldenk@sakarya.edu.tr. ORCID ID: orcid.org/0000-0002-8100-6994
*** Prof. Dr. , Hacettepe Üniversitesi, Eğitim Fakültesi, Ankara-Türkiye, sgelbal@gmail.com. ORCID ID: orcid.org/0000-0001-5181-7262

Geliş Tarihi:07.11.2017
DOI: 10.21031/epod.349718 *Kabul Tarihi:01.03.2018*

GİRİŞ

Eğitim çalışmalarında Klasik Test Kuramı (KTK), Genellenebilirlik Kuramı (GK) ve Madde Tepki Kuramı (MTK) olmak üzere üç temel kuram yer almaktadır. Bazı araştırmacılar test kuramlarını puanların analiz edilmesi ve yorumlanması bakımından klasik ve modern olmak üzere ikiye ayırırlar. Klasik Test Kuramı ve Genellenebilirlik Kuramı klasik; Madde Tepki Kuramı ise modern kuram olarak ele alınmakta olup, bu kuramlarda farklı matematiksel modeller kullanılmaktadır. Modern test kuramının popülerliği gün geçtikçe artsa da KTK hâlâ en pratik kuram olarak görülmektedir. Günümüzde hala birçok test KTK'ya göre geliştirilmektedir. Genellenebilirlik Kuramının hata kaynaklarını çözümlemede ANOVA'yı kullanıyor olması GK'nın KTK'nın uzantısı olduğunu göstermektedir (Feldt ve Quails,1989; Shavelson ve Webb, 1991). Bu nedenle GK'da sıklıkla MTK'dan farklı olarak klasik kuramın içinde ele alınmaktadır.

Modern ve klasik kuramların tamamen farklı olmadığı, bir arada ya da birbirinin tamamlayıcısı olarak kullanılabileceğine yönelik iddialar da söz konusudur. Buna dayalı olarak araştırmacılar KTK ve GK'yı, MTK ile birbirine bağlayan çalışmalar üzerine yoğunlaşmışlardır. Örneğin Kolen ve Harris (1987) hem MTK hem de KTK'ya dayalı olarak çok değişkenli test modelleri ortaya atmışlardır. Benzer şekilde güvenirlik konusu içinde MTK'yı klasik kuramlarla birleştiren çalışmalar da söz konusudur. Samejima (1977,1994). güvenirlik ve ölçmenin standart hatası tahmini için KTK ile MTK'yı birbirine test bilgi fonksiyonu üzerinden bağlamıştır ve 1994 yılında yaptığı çalışma ile test bilgi fonksiyonu için tahmini güvenirlik önermiştir. Lord (1983), paralel formların yeteneğe dayalı güvenirlik katsayılarının kestirimi için eşitlikler ileri sürmüştür. Raju ve Oshime (2005), kısa ve uzun testler için yeteneğe dayalı güvenirlik kestirimini yapan iki eşitlik ortaya koymuştur ve bu eşitliklerden birinin Spearman Brown eşitliği ile aynı olduğunu ispatlamıştır. Dimitrov (2003), ikili puanlanan maddeler için gerçek puan kestirimlerini, MTK ve KTK'yı birleştirerek elde edilen eşitlikler üzerine çalışmıştır.

Madde Tepki Kuramında bireylerin belli bir alanda doğrudan gözlenemeyen yetenekleri ile bu alanı yoklayan sorulardan oluşan test maddelerine verdikleri yanıtlar arasındaki matematiksel ilişki yer alırken, Genellenebilirlik Kuramında ölçme sonuçlarının güvenirliği belirlenir, güvenilir gözlemler tasarlanır, araştırılır ve kavramsallaştırılır. Madde Tepki Kuramı (MTK) ve Genellenebilirlik Kuramı (GK) en azından yüzeysel açıdan birbirinden farklı olarak görülür. Örneğin; Brennan (2001), Genellenebilirlik Kuramının örnekleme modeli, Madde Tepki Kuramının ise bir ölçekleme modeli olduğunu belirtmiştir. Ancak her iki yaklaşım da desene ve ölçme araçlarının analizine ilişkin önemli bilgiler sağladığı için MTK'yı ve GK'yı hem büyük ölçekli sınavlarda hem de daha küçük ölçek çalışmalarında beraber kullanmak yararlı olabilir (Bock, Brennan ve Muraki, 2002). GK ve MTK'nın beraber kullanılmasının önemi anlaşılmış olmasına rağmen birleşimi oluşturmak zaman almıştır. Konuyla ilgili ilk adım Linacre (1989, 1999) tarafından atılmıştır. Linacre puanlayıcılar tarafından ikili puanlanan madde puanlarını incelemiştir. Elde ettiği model Rasch modelin (Rasch, 1960) basit bir genellemesi olarak sunulmuştur.

Alanyazında GK ile MTK'nın birlikte kullanıldığı çalışmalarda genel olarak her maddenin birden çok puanlayıcı tarafından puanlandığı desenler üzerinde çalışılmıştır (Alkahtani, 2012; Bock, Brennan ve Muraki, 2002; Kim ve Wilson, 2008; Patz, Junker, Johnson ve Mariano, 2002; Verhelst ve Verstralen, 2001; Wilson ve Hoskens, 2001; Zhang ve Roberts, 2013). Ancak yapılan bu çalışmalar MTK'nın yerel bağımsızlık varsayımını ihlal ettiği gerekçesiyle eleştirilmiştir. Glas (1989) farklı puanlayıcıların verdikleri puanların öğrencilerin cevaplarına bağlı olduğunu; bu nedenle bu modelin MTK'nın yerel bağımsızlık varsayımını ihlal ettiğini öne sürmüştür. Bu durumla başa çıkmak için MTK'da başka modeller üzerinde çalışılmıştır. Örneğin ilk olarak Zwinderman (1991) MTK modeli ile yapısal ANOVA modelini birleştirme üzerine çalışmalar yapmıştır. Daha sonra Fox ve Glas (2001) çalışmaların üzerinde durmuş; ancak kesin sonuca ulaşamamıştır. MTK ile GK'nın birleştirilmesi ilk olarak Briggs ve Wilson'ın (2004, 2007) yapmış oldukları çalışmalar ile gerçekleşmiştir.

Briggs ve Wilson (2004, 2007) çalışmalarında örnekleme modeli olan Genellenebilirlik Kuramını ölçekleme modeli olan MTK ile birleştirmiş ve Madde Tepki Modellemesinde Genellenebilirlik (MTMG) isimli yeni bir model ortaya atmışlardır. Yaptıkları çalışmada MTMG yaklaşımının GK'yı MTK'nın içine ilgili madde yüzeyinde dağılımsal varsayımlar yaparak dâhil ettiklerini ileri

sürmüşlerdir. MTMG yaklaşımında, Genellenebilirlik Kuramında geleneksel olarak kullanılan gözlenen puanlar yerine Markov Zinciri Monte Carlo (MZMC) tekniği ile elde edilen beklenen puanlar kullanılmıştır. MTMG'de MZMC yönteminin esnekliğinden de yararlanılarak GK varyans bileşenlerini MTK parametreleri üzerinden kestirmek mümkündür.

Briggs ve Wilson (2004, 2007) çalışmalarında GK ve MTMG arasındaki farkları şu şekilde sıralamışlardır:

- GK varyans bileşenleri kestiriminde gözlenen puanları kullanırken MTMG beklenen puanlar matrisini kullanır.
- Etkileşim varyansı ve hata GK'da birbirinden ayrılamazken MTMG yaklaşımında ayrı ayrı kestirilebilir.

MTMG yaklaşımında beklenen puanlar matrisinin kullanılması daha güvenilir kestirimler yapmaya olanak sağlamaktadır. Diğer yandan GK'nın en büyük dezavantajlarından biri olan hata varyansı sorunu MTMG yaklaşımı ile çözümlenebilir hale gelmektedir.

MTMG yaklaşımının tanıtılması ve örneklendirilmesi bireylerin maddeler ile çaprazlandığı " bxm" deseni üzerinden yapılmıştır. Çalışmada ölçme modeli olarak MTK'nın Rasch modeli, yapısal model olarak GK kullanılmıştır. MTMG yaklaşımı Briggs ve Wilson'ın (2004,2007) yapmış oldukları çalışma ile sınırlı kalmıştır. Yaklaşımın farklı desenler ve farklı çalışma koşulları altında nasıl sonuçlar verdiği henüz bilinmemektedir.

GK ve MTK'nın kullanıldığı diğer çalışma konusu ise madde takımlarıdır. Madde takımları; sınavı alan bireylerin takip edeceği, önceden belirlenmiş belli sayıdaki yolu içeren tek bir konuya ait bir grup ilişkili madde olarak tanımlanır (Wainer, Lewis; 1990). Madde takımları ile kestirilen yetenek; hem bireyin genel yeteneği hem de konu ile ilgili belirli bir yeteneğe bağlıdır (Demars, 2006). Tek uygulama ile birden çok cevap bulan madde takımlarının kullanımı özellikle zamansal açıdan ekonomik olması nedeniyle son zamanlarda artmıştır.

Madde takımları konusu hem GK ile hem de MTK ile ele alınabilir. GK'ya göre incelenen madde takımlarında; madde takımları maddelerin içinde yuvalandığı bir desen olarak ele alınmalıdır. Madde takımları yok sayılıp bir yüzey olarak ele alınmadığında, ölçmenin standart hatası için alt; güvenirlik için ise üst kestirim yapılır (Yen, 1993; Wainer ve Thissen, 1996; Wainer ve Wang, 2000).

Madde takımları maddelerin sahip olduğu karakteristik özelliklerden dolayı Madde Tepki Kuramında yer alan yerel bağımsızlık varsayımını ihlal eder. Yapılan birçok araştırmada madde takımlarının yerel bağımsızlığı bozduğu sonucuna ulaşılmıştır (Rosenbaum, 1988; Yen, 1993; Wainer, 1995; Wainer ve Thissen, 1996; Jiao, Kamata, Wang ve Jin, 2012). Yerel bağımsızlık varsayımının karşılanmadığı durumlarda bireylerin performansları, madde parametreleri ya da test istatistikleri için elde edilen sonuçlar hatalıdır (Yen, 1993; Wainer, 1995; Wainer ve Thissen, 1996; Ferrara, Huynh ve Bagli, 1997; Ferrara, Huynh ve Michaels, 1999; Bradlow, Wainer ve Wang 1999). Madde takımı puanlamada, madde takımını oluşturan maddelerin birbirlerine bağlı olmalarının göz önünde bulundurulması olumlu bir durum iken; bu maddeleri cevaplandıran bireyin cevap deseni ile ilgili bilgi kaybı söz konusudur. Bu olumsuzluğu ortadan kaldırmak adına orijinal MTK modellerine kişiye özgü tesadüfî madde takımı etkisinin de eklenmesi yeni bir strateji olarak ele alınmaktadır (Bradlow, Wainer ve Wang, 1999; Wainer, Bradlow ve Du, 2000; Wang, Bradlow ve Wainer, 2002; Li, Bolt ve Fu, 2006). Bu strateji ise "Madde Takımı Tepki Kuramı" (MTTK) olarak adlandırılmaktadır (Wainer, Bradlow ve Du 2000). Dresher (2004) MTTK modelinin kullanılmasının, madde ayırıcılık ve güçlük parametrelerinin madde takımını oluşturan maddelerin birbirlerine bağımlılıklarını yok sayan madde takımı puanlama ya da tek boyutlu MTK modellerine göre daha iyi kestirim yaptığını bulmuştur (Akt. Chien, 2008). MTTK'da temel olarak kişiye özgü tesadüfî madde takımı etkisini ele alan pek çok madde takımı modeli bulunmaktadır (Bradlow, Wainer ve Wang, 1999; Wainer, Bradlow ve Du, 2000; Wang, Bradlow ve Wainer, 2002; Li, Bolt ve Fu, 2006). Bütün MTTK modelleri, her bir bireydeki yerel madde bağımlılığı miktarını belirtmek için, geleneksel MTK parametrelerinin yanında bir de madde takımı parametresini önermektedir. Genel olarak, geliştirilen bütün MTTK modelleri, çok boyutlu MTK modellerinden ya da daha önce önerilen bir MTTK modelinden uyarlanmıştır.

Briggs ve Wilson'ın (2004,2007) yaptıkları çalışmalar ile ortaya atılan MTMG yaklaşımı birbirinden farklı olan MTK ve GK modelini birbirine bağlar. Bu bağlam bir bireyin tek bir madde için beklenen puanı üzerinden yapılır. Bu durumun, tek yüzeyli desenler için mümkün olduğu Briggs ve Wilson'ın (2004, 2007) yaptığı çalışmalar ile ispatlanmıştır.

MTMG yaklaşımı elde edilen ümit verici sonuçların yanında alanyazında çok az gerçekleştirilen birbirlerinden farklı iki ölçme kuramını bir arada kullanması açısından değerlidir. Ancak MTMG çalışmaları tek yüzeyli desenler ile sınırlı kalmıştır. MTMG yaklaşımının çok yüzeyli durumlarda nasıl işlediğini bilmek gereklidir. Örneğin MTMG çok yüzeyli modellerde çalışmazsa adres gösterilen sorunlarda GK'nın yerine kullanılabileceği söylenemez. Bu nedenle bu çalışmanın temel amacı Briggs ve Wilson'ın (2004, 2007) yaptığı çalışmayı tek yüzeyli desenden çok yüzeyli desene çıkartmaktır. Çalışmada kullanılan çok yüzeyli desen, maddelerin (m) takımlara (t) yuvalandığı ve bireylerin (b) bunlarla çaprazlandığı rastgele dengelenmiş yuvalanmış desendir. Bu desen simgesel olarak bx(m:t) olarak gösterilir.

Madde takımlarından oluşan testlerin kullanımı, beraberinde getirdiği avantajlar nedeniyle hem ulusal hem de uluslararası sınavlarda artış göstermektedir. Ancak madde takımlarının yerel madde bağımsızlığı ihlali göz ardı edildiğinde kestirimde hatalara neden olduğu açıktır (Dresher,2004). Bu çalışmada madde takımlarının farklı koşullar altında elde edilen parametreleri incelenmiştir. Bu koşullar; birey-madde takımlarının etkileşimlerinin varyans büyüklükleri, madde takımları sayısı ve madde takımlarında bulunan madde sayısıdır.

Bu çalışmada birbirinden farklı kuramlarının birleştirmesi amaçlanmıştır. Aynı zamanda madde takımının olduğu farklı durumlar, farklı kuramlar çerçevesinde incelenmiştir. Bu nedenle, hem farklı kuramların bir arada kullanılması hem de madde takımlarının sıklıkla kullanıldığı ulusal çapta yürütülen geniş ölçekli sınavlarda parametre kestirimleri için farklı bir yaklaşım önermesi açısından önemli olduğu düşünülmektedir.

YÖNTEM

Bu araştırmada farklı koşullar için Madde Tepki Modellemesinde Genellenebilirlik yaklaşımı (MTMG) ile sonuçlar elde edilmiş ve elde edilen sonuçlar aynı koşullar için Genellenebilirlik Kuramı (GK) ile elde edilen sonuçlar ile karşılaştırılmıştır. Kontrollü koşulların oluşturularak uygun verilerin türetilmesi bakımından araştırma, bir simülasyon çalışmasıdır. Araştırmada simülasyon verileri ile farklı koşullar oluşturulmakta ve koşulların durumları/sonuçları değerlendirilmektedir. Araştırma bu yönüyle de yöntemlerin geliştirilmesine katkı sağlayacağından temel araştırma olarak kabul edilebilir (Karasar, 2004).

Çalışma Verileri

Çalışmada simülasyon veri kullanılmıştır. Verilerin üretimi için 100 tekrar yapılmış böylelikle hata en aza düşürülmeye çalışılmıştır. İki farklı evren her evrene ait dört farklı koşuldan oluşan 8 farklı veri seti vardır. Her veri seti için simülatif olarak üretilen veriler 10 tekrar üzerinden yürütülmüştür. Ancak bulgular kısmında tüm tekrarların ortalaması paylaşılmıştır.

GK doğrusal veri seti bx (m:t) dengelenmiş rastgele deseni için üretilmiştir. Oluşturulan tüm veri setlerinde birey sayısı ve 1-0 şeklinde puanlanan toplam madde sayısından oluşan n_b x n_m gözlenen puan matrisi elde edilmiştir. Toplam madde sayısı; madde takımı sayısı ve madde takımlarında yer alan maddelerin çarpımı ile elde edilmiştir. Her madde sadece bir madde takımında yuvalanmıştır. n_m ve n_t farklı çalışma koşullarına göre değişiklik gösterir ancak birey sayısı (n_b) bu çalışma için 1000'e sabitlenmiştir.Oluşturulan veriler madde takımı etkisi, madde takımı uzunluğu ve madde takımı sayısı açısından farklılık göstermektedir.

Madde Takımı Etkisi

Veri üretmek için uyarlanan Genellenebilirlik Kuramı doğrusal veri seti bx(m:t) dengelenmiş rastgele deseni için üretilmiştir. GK doğrusal veri setinde madde takımı etkisi olan $\sigma^2_{(t)}$ yerine madde takımlarının anlam ya da zorluk bakımından bireyden bireye farklılık gösterip göstermediğini belirten birey ve madde takımı etkileşiminin ($\sigma^2_{(bt)}$) iki farklı durumu kullanılmıştır. Bu durumlar birey-madde takımı etkileşiminin diğer varyans kaynakları arasında en büyük değere ve en küçük değere sahip olması bakımından değişiklik gösterir. Varyans kaynaklarının değerleri gerçek durumları yansıtması için yapılan birçok çalışma (Lee ve Frisbie, 1999; Lee, Brennan ve Frisbie, 2000; Chien 2008) incelenerek oluşturulmuştur.

Madde Takımı Uzunluğu ve Madde Takımı Sayısı

Madde takımı uzunluğu kullanılan madde takımında bulunan madde sayısını; madde takımı sayısı ise testte bulunan toplam madde takımı sayısını ifade etmektedir. Bu çalışmada bx(m:t) deseni dengelenmiş olarak inceleneceği için madde takımı uzunluğu ve madde takımı sayısı çarpımı toplam madde sayısını vermektedir. Çalışmada gerçek durumlara uygun olması açısından uluslararası (PIRLS; ITBS RC) ve ulusal sınavlar (ALES, KPSS, bazı üniversitelere ait hazırlık geçme sınavları) incelenmiş ve madde takımı uzunluklar 6 ve 9; madde takımı sayıları 3 ve 5 olarak belirlenmiştir. Belirlenen madde takımı uzunluğu ve madde takımı sayılarına göre veride kullanılan madde sayıları sırasıyla 18, 30, 27, 45'tir. Aşağıda yer alan Tablo 1'de çalışma koşulları özetlenmiştir.

Tablo 1. Araştırmada Yer Alan Çalışma Koşulları

	GK doğrusal veri seti
Madde takımı etkisi	Küçük σ^2(bt)
	Büyük σ^2(bt)
Madde takımı uzunluğu - madde takımı sayısı	6-3
	6-5
	9-3
	9-5
Özet Çalışma koşulları	A evreni
	Küçük ve 6-3
	Küçük ve 6-5
	Küçük ve 9-3
	Küçük ve 9-5
	B evreni
	Büyük ve 6-3
	Büyük ve 6-5
	Büyük ve 9-3
	Büyük ve 9-5

Verilerin Analizi

Araştırmanın verileri R programı ile üretilmiştir. A ve B olarak isimlendirilen her evrene ait 4 koşul bulunmaktadır. Toplamda 8 farklı durumunun her biri için 10 farklı veri elde edilmiştir. Üretilen verilerin Madde Tepki Modellemesinde Genellenebilirlik çözümlemeleri WinBUGS, Genellenebilirlik Kuramı çözümlemeleri ise EDUG programı ile yapılmıştır ve her durum için bu kestirimler karşılaştırılmıştır.

BULGULAR

Araştırma problemine ait bulgulardan önce simülasyon verisine ait test güvenirlikleri ve madde güçlükleri ile ilgili bilgilere yer verilmiştir. Tablo 2'de iki evrenden elde edilmiş veri setinin farklı koşullarına ait 10 tekrardan oluşan verilerin ortalama güvenirlik değeri, standart sapma değerleri ve

verilerin almış olduğu maksimum ve minimum güvenirlik değerleri yer almaktadır. Veri setlerine ait güvenirlik katsayıları genellenebilirlik katsayılarıdır ve ölçme objesi olarak bireyler alınmıştır.

Tablo 2. Veri Setlerine Ait Güvenirlik Değerleri

	Evren	*Koşullar*	*SS*	*Max. güvenirlik*	*Min. güvenirlik*	*Ortalama güvenirlik değeri*
GK doğrusal veri seti	A	Küçük ve 6-3	0,011	0,648	0,609	0,624
		Küçük ve 6-5	0,016	0,756	0,699	0,734
		Küçük ve 9-3	0,015	0,723	0,678	0,698
		Küçük ve 9-5	0,009	0,821	0,779	0,800
	B	Büyük ve 6-3	0,018	0,607	0,552	0,579
		Büyük ve 6-5	0,023	0,723	0,659	0,691
		Büyük ve 9-3	0,014	0,664	0,625	0,653
		Büyük ve 9-5	0,008	0,758	0,740	0,750

Tablo 2'de yer alan A ve B evrenlerine ait koşullardan elde edilen güvenirlik katsayıları incelendiğinde madde sayısı arttıkça güvenirliğin arttığı görülmektedir. Bunun yanında eşit madde sayıları için A evreninden kestirilen güvenirlik katsayıları B evrenine göre daha yüksektir. Bu beklenen bir sonuçtur çünkü B evreninde birey-madde takımı etkisi A evrenine göre daha büyük bir değerdedir ve bu değerin büyük olması hataya sebep olmaktadır.

Madde güçlüğü KTK'ya göre maddeye doğru cevap verenlerin yüzdesi; MTK'ya göre ise maddenin P olasılıkla doğru yanıtlanması için gerekli yetenek düzeyi olarak tanımlanır. Tablo 3'te farklı koşullar için elde edilen 10 tekrar verisine ait ortalama madde güçlüğü değerleri; standart sapma değerleri ve alınan maksimum ve minimum madde güçlüğü değerleri yer almaktadır.

Tablo 3. Veri Setlerine Ait Madde Güçlükleri

Evren	*Koşullar*	*SS*	*Max. değer*	*Min. değer*	*Ortalama madde güçlük değeri*
A	Küçük ve 6-3	0,110	0,920	0,385	0,650
	Küçük ve 6-5	0,095	0,915	0,390	0,680
	Küçük ve 9-3	0,105	0,880	0,355	0,650
	Küçük ve 9-5	0,085	0,890	0,355	0,675
B	Büyük ve 6-3	0,120	0,945	0,275	0,740
	Büyük ve 6-5	0,230	0,980	0,324	0,700
	Büyük ve 9-3	0,100	0,930	0,390	0,665
	Büyük ve 9-5	0,095	0,935	0,237	0,675

Tablo 3 incelendiğinde GK doğrusal veri seti için madde güçlüğü ortalamasının 0,6-0,7 aralığında olduğu görülmüştür. Veri setinde yer alan soruların kolay olmasının nedeni veri seti üretilirken kesme noktasının -0,2 seçilmiş olmasıdır. En zor soruya ait madde güçlüğü değeri 0,237 olduğu için çok zor bir sorunun olmadığı görülmektedir.

Simülasyon verisine ait test güvenirlikleri ve madde güçlükleri ile ilgili bilgilerin ardından birey-madde etkileşim varyansının diğer varyans değerleri arasında en küçük, madde takımı sayısı 3 veya 5 ve madde takımlarında yer alan madde sayısı 6 veya 9 olduğu durumlarda (A evreni) Madde Tepki Modelinde Genellenebilirlik yaklaşımına ve Genellenebilirlik Kuramına göre elde edilen; a) varyans bileşenleri arasında fark ve b) bağıl ve mutlak hata varyansı, genellenebilirlik ve Phi katsayısı arasındaki fark incelenmiştir.

Çalışmada varyans değerleri elde edilen değişkenlik kaynakları şunlardır;

Birey değişkenlik kaynağı ($\sigma^2(b)$): Değişkenliğin ilk kaynağı ölçme objesi olan öğrencilerin maddelerden aldıkları farklı puanlardır. Evren puanı için olan bu varyans bileşeni, bireylerin birbirinden ne derece sistematik bir şekilde farklılık gösterdiğini ifade etmektedir. Bu nedenle birey değişkenlik kaynağı değerinin olabildiğince büyük olması istenen bir durumdur.

Madde takımı değişkenlik kaynağı (σ^2(t)): Madde takımı değişkenliği madde takımları arasındaki tutarsızlıktan kaynaklanmaktadır. Diğer bir deyişle bir madde takımının bir bireye kolay gelirken diğer madde takımının aynı birey için zor gelmesi madde takımındaki varyansın sebebidir. Bu değerler, madde takımlarının birbiri arasındaki değişkenliğin derecesini vermektedir. Bu nedenle madde takımı varyans bileşeni değerinin olabildiğince küçük olması istenen bir durumdur.

Birey-madde takımı değişkenlik kaynağı (σ^2(bt)): Madde takımlarının bazı bireyler için kolaylık-zorluk anlamında tutarsızlıkları olabilir. Bu tutarsızlıkların derecesi birey- madde takımı değişkenlik kaynağında incelenir. Bu değerler, birey-madde takımlarının etkileşiminin derecesini vermektedir. Birey - madde takımı değişkenlik kaynağı A ve B evrenleri için manipüle edilen değişkenlik kaynağıdır. A evreninde bu değişkenlik kaynağı diğer değişkenlik kaynakları içerisinde en küçük B evrenine ise en büyük değer olarak elde edilmiştir.

Madde-madde takımı değişkenlik kaynağı σ^2(m:t): Her bir madde takımında yer alan maddelere ilişkin; bireyler bazı maddelerde geçmiş yaşantılarından dolayı daha avantajlı iken bazılarında bu durum söz konusu olmayabilir. Maddelerin güçlük düzeyleri arasındaki farklılıklar, maddeler farklı madde takımlarında yer aldığı için, her bir madde takımında yer alan maddeler üzerinden yorumlanıp, madde takımlarından bağımsız bir yorum yapılamaz. Bu değerler, maddelerin madde takımlarının içinde yuvalanmış olmasından kaynaklanan etki değerlerini verir.

Birey-madde-madde takımı (artık) değişkenlik kaynağı σ^2(bm:t): İki yüzeyli bx(m:t) deseninde yer alan son değişkenlik kaynağı birey, madde ve madde takımı etkileşiminin ve tesadüfî hataların yol açtığı değişkenliktir. GK'da etkileşim varyansı hata varyansından ayrılamazken MTMG'de hata varyansı ayrı olarak elde edilebilir. Bu nedenle birey- madde - madde takımı etkileşimi için elde edilen varyans değerleri için GK ve MTMG kestirimi arasındaki fark MTMG ile elde edilen hata varyansı değerini de kapsamaktadır.

Tablo 4'te A evrenine ait her bir değişkenlik kaynağı için Genellenebilirlik Kuramı (GK) ve Madde Tepki Modellemesinde Genellenebilirlik (MTMG) yaklaşımı ile kestirilen değerler ayrı ayrı verilmiş ve aralarındaki fark hesaplanmıştır.

Tablo 4. A Evrenine Ait Kestirilen Varyans Değerleri

			Birinci koşul	*İkinci koşul*	*Üçüncü koşul*	*Dördüncü koşul*
A evreni	σ^2***(b)***	***GK***	0,01982	0,02004	0,01988	0,02068
		MTMG	0,01980	0,02002	0,01987	0,02066
		FARK	0,00002	0,00002	0,00001	0,00002
	σ^2***(t)***	***GK***	0,00524	0,00527	0,00515	0,00488
		MTMG	0,00523	0,00526	0,00513	0,00487
		FARK	0,00001	0,00001	0,00002	0,00001
	σ^2***(bt)***	***GK***	0,00039	0,00112	0,00100	0,00111
		MTMG	0,00038	0,00111	0,00099	0,00110
		FARK	0,00001	0,00001	0,00001	0,00001
	σ^2***(m:t)***	***GK***	0,00920	0,01080	0,00995	0,01018
		MTMG	0,00919	0,01078	0,00994	0,01016
		FARK	0,00001	0,00002	0,00001	0,00002
	σ^2***(bm:t)***	***GK***	0,18642	0,18801	0,18787	0,19052
		MTMG	0,18582	0,18742	0,18728	0,18995
		FARK	0,00060	0,00059	0,00060	0,00057
	σ^2***(e)***	***MTMG***	0,00059	0,00056	0,00058	0,00055

A evreni için dört farklı koşul vardır. Birinci koşulda birey-madde takımı etkileşimin varyans değerinin diğer varyanslar arasında en küçük, madde takımında yer alan madde sayısı 6, madde takımı sayısı 3'tür. A evreninin ikinci koşulunda birey-madde takımı etkileşimi varyans değeri diğer varyanslar arasında en küçük, madde takımında yer alan madde sayısı 6 ve madde takımı sayısı 5'tir. Birey-madde takımı etkileşimin varyans değerinin diğer varyanslar arasında en küçük olduğu, madde takımında yer alan madde sayısı 9, madde takımı sayısı 3 olması durumu A evreninin üçüncü koşuludur. A evreninin dördüncü koşulunda birey-madde takımı etkileşimin varyans değeri diğer

varyanslar arasında en küçük, madde takımında yer alan madde sayısı 9, madde takımı sayısı 5'tir. Buna göre tüm koşullar için değişkenlik kaynaklarına ait değerlerin MTMG ve GK kestirimi arasındaki farklar incelendiğinde; İki yaklaşım ile elde edilen Birey-madde-madde takımı (artık) değişkenlik kaynağı (σ2(bm:t)) dışında kalan diğer değişkenlik kaynaklarına ait varyans değerleri arasındaki fark her tekrar için 0,00002 ile 0,00001 arasında değişmektedir. Bu durum; A evreni tüm koşullarında Birey-madde-madde takımı (artık) değişkenlik kaynağı (σ2(bm:t)) dışında kalan diğer değişkenlik kaynakları için GK ve MTMG kestirimi arasında fark olmadığı şeklinde yorumlanabilir.

Birey- madde - madde takımı etkileşimi varyans bileşeni için hesaplanan değer GK kestirimi ile tüm tekrarların ortalaması alındığında birinci koşulda GK kestirimi için bu değer 0,18642 MTMG kestirimi için ise 0,18582'dir. İkinci koşulda GK kestirimi için 0,18801 MTMG kestirimi için ise 0,18742'dir. Üçüncü koşulda GK kestirimi için bu değer 0,18787 MTMG kestirimi için ise 0,18728'dir. Dördüncü koşulda ise GK kestirimi için bu değer 0,19052 MTMG kestirimi için ise 0,18995'dir. Elde edilen bu değerler her iki yaklaşım içinde diğer varyans değerleri arasında birinci sırada yer almaktadır. Tablo 4'te yer alan σ^2(e) değerleri incelendiğinde MTMG yaklaşımı ile hesaplanabilen hata varyansının 0,00055 ile 0,00059 arasında değerler aldığı görülmektedir. MTMG ile elde edilen hata varyansı değerleri farktan çıkarıldığında birey- madde - madde takımı etkileşimi için kestirilen değerler arasındaki fark 0,00001 ile 0,00003 arasında olduğu görülür. Bu durum A evrenin her koşulunda birey-madde-madde takımı değişkenlik kaynağı için GK ve MTMG kestirimi arasında fark olmadığı şeklinde yorumlanabilir.

A evreninde birey ve madde takımı etkileşimi ($\sigma^2_{(bt)}$) diğer varyanslar arasında en küçük değerdir ve koşullar içinde madde takımı uzunluğu - madde takımı sayısı sırasıyla 6-3,6-5,9-3,9-5'tir. Yapılan analizler sonucunda A evreninin her koşulu için MTMG yaklaşımı ve GK yaklaşımı arasında bir fark bulunmamıştır.

A evrenine ait GK ve MTMG yöntemleri ile kestirilen bağıl ve mutlak hata varyans değerleri ile G ve Phi katsayılarını incelenmiş ve evrene ait koşullar çerçevesinde değerler ayrı ayrı verilmiş ve aralarındaki fark hesaplanmıştır. Tablo 5'te elde edilen değerler yer almaktadır.

Tablo 5. A Evreni Koşullarına Ait Bağıl ve Mutlak Hata Varyansı, Genellenebilirlik ve Phi Katsayısı Değerleri

			Birinci koşul	*İkinci koşul*	*Üçüncü koşul*	*Dördüncü koşul*
A evreni	***σ2(δ)***	***GK***	0,01208	0,00722	0,00856	0,00517
		MTMG	0,01207	0,00721	0,00854	0,00515
		FARK	0,00001	0,00001	0,00002	0,00002
	Eb²	***GK***	0,62410	0,73403	0,69841	0,80034
		MTMG	0,62408	0,73402	0,69840	0,80032
		FARK	0,00002	0,00001	0,00001	0,00002
	σ²(Δ)	***GK***	0,01292	0,00771	0,00900	0,00563
		MTMG	0,01291	0,00770	0,00899	0,00561
		FARK	0,00001	0,00001	0,00001	0,00002
	Φ	***GK***	0,60426	0,72102	0,68668	0,78625
		MTMG	0,60425	0,72100	0,68666	0,78623
		FARK	0,00001	0,00002	0,00002	0,00002

A evreninin her koşulu için Genellenebilirlik Kuramına göre kestirilen bağıl hata varyansı değerleri ile Madde Tepki Modellemesinde Genellenebilirlik yaklaşımına göre kestirilen bağıl hata varyansı kestiriminde iki yaklaşım arasındaki fark her tekrar için 0,00001 ile 0,00002 arasında değişmektedir bu nedenle iki yaklaşım arasında her koşul altında bağıl hata varyansı kestiriminde fark olmadığı söylenebilir.

Bağıl hata varyansına bağlı olarak genellenebilirlik katsayısı değerleri de hesaplanmıştır. Genellenebilirlik katsayısı kestiriminde iki yaklaşım arasındaki fark tüm koşullarda her tekrar için 0,00001 ile 0,00002 arasında değişmektedir bu nedenle iki yaklaşım arasında genellenebilirlik katsayısı kestiriminde fark olmadığı söylenebilir.

Benzer bir durum mutlak hata değerleri içinde geçerlidir. Mutlak hata varyansı kestiriminde iki yaklaşım arasındaki fark tüm koşullarda her tekrar için 0,00001 ile 0,00002 arasında değişmektedir. Bu nedenle iki yaklaşım arasında mutlak hata varyansı kestiriminde fark olmadığı söylenebilir.

Mutlak hata varyansına bağlı olarak hesaplanan Phi katsayısı değerleri için her koşul altında iki yaklaşım arasındaki fark her tekrar için 0,00001 ile 0,00002 arasında değişmektedir, bu nedenle iki yaklaşım arasında Phi katsayısı kestiriminde fark olmadığı söylenebilir.

A evrenine ait koşullar madde takımı ve madde takımlarında yer alan madde sayıları açısından farklılık göstermektedir. Madde takımı uzunluğu 6 olan birinci ve ikinci koşul arasında madde takımı sayısı daha fazla olan ikinci koşuldan daha yüksek güvenirlik elde edilmiştir. Benzer şekilde madde takımı uzunluğu 9 olan üçüncü ve dördüncü koşullarda madde takımı sayısı fazla olan dördüncü koşuldan daha yüksek güvenirlik elde edilmiştir. Madde takım sayıları eşit olan birinci-üçüncü ve ikinci-dördüncü koşullarda madde takımı sayısı fazla olan koşullardan daha yüksek güvenirlik elde edilmiştir. Tüm bu bulgular her dört koşul incelendiğinde toplam madde sayısı arttıkça güvenirliğin arttığını göstermektedir.

Çalışmada ayrıca Birey-madde etkileşim varyansının diğer varyans değerleri arasında en büyük, madde takımı sayısı 3 veya 5 ve madde takımlarında yer alan madde sayısı 6 veya 9 olduğu durumlarda (B evreni) Madde Tepki Modelinde Genellenebilirlik yaklaşımına ve Genellenebilirlik Kuramına göre elde edilen; a) varyans bileşenleri arasında fark b) bağıl ve mutlak hata varyansı, genellenebilirlik ve Phi katsayısı arasında fark incelenmiştir.

Tablo 6'da B evrenine ait her bir değişkenlik kaynağı için Genellenebilirlik Kuramı (GK) ve Madde Tepki Modellemesinde Genellenebilirlik (MTMG) yaklaşımı ile kestirilen değerler ayrı ayrı verilmiş ve aralarındaki fark hesaplanmıştır.

Tablo 6. B Evrenine Ait Kestirilen Varyans Değerleri

			Birinci koşul	*İkinci koşul*	*Üçüncü koşul*	*Dördüncü koşul*
B evreni	***σ2(b)***	***GK***	0,00956	0,00938	0,01002	0,00945
		MTMG	0,00954	0,00936	0,01000	0,00943
		FARK	0,00002	0,00002	0,00002	0,00002
	σ2(t)	***GK***	0,00517	0,00569	0,00515	0,00543
		MTMG	0,00515	0,00568	0,00513	0,00541
		FARK	0,00002	0,00001	0,00002	0,00002
	σ2(bt)	***GK***	0,01847	0,01784	0,01989	0,01836
		MTMG	0,01845	0,01782	0,01987	0,01834
		FARK	0,00002	0,00002	0,00002	0,00002
	σ2(m:t)	***GK***	0,00942	0,00603	0,00904	0,00894
		MTMG	0,00940	0,00602	0,00902	0,00892
		FARK	0,00002	0,00001	0,00002	0,00002
	σ2(bm:t)	***GK***	0,01808	0,01568	0,01848	0,01512
		MTMG	0,01751	0,01496	0,01732	0,01432
		FARK	0,00057	0,00072	0,00116	0,00080
	σ2(e)	***MTMG***	0,00055	0,00071	0,00115	0,00079

B evreninin birinci koşulunda birey madde takımı etkileşimi varyansı diğer varyans değerleri arasında en büyük, madde takımında yer alan madde sayısı 6 ve madde takımı sayısı 3'tür. B evreninin ikinci koşulunda birey madde takımı etkileşimi büyük, madde takımında yer alan madde sayısı 6 ve madde takımı sayısı 5'tir. B evreninin üçüncü koşulunda birey madde takımı etkileşimi büyük, madde takımında yer alan madde sayısı 9 ve madde takımı sayısı 3'tür. Dördüncü koşulda ise birey madde takımı etkileşimi büyük, madde takımında yer alan madde sayısı 9 ve madde takımı sayısı 5'tir.

Buna göre tüm koşullar için değişkenlik kaynaklarına ait değerlerin MTMG ve GK kestirimi arasındaki farklar incelendiğinde; İki yaklaşım ile elde edilen Birey-madde-madde takımı (artık) değişkenlik kaynağı (σ2(bm:t)) dışında kalan diğer değişkenlik kaynaklarına ait varyans değerleri arasındaki fark her tekrar için 0,00002 ile 0,00001 arasında değişmektedir. Bu durum; B evreni tüm

koşullarında Birey-madde-madde takımı (artık) değişkenlik kaynağı (σ2(bm:t)) dışında kalan diğer değişkenlik kaynakları için GK ve MTMG kestirimi arasında fark olmadığı şeklinde yorumlanabilir.

Birey- madde - madde takımı etkileşimi varyans bileşeni için hesaplanan değer GK kestirimi ile tüm tekrarların ortalaması alındığında birinci koşulda GK kestirimi için bu değer 0,01808 MTMG kestirimi için ise 0,01751'dir. İkinci koşulda GK kestirimi için 0,01568 MTMG kestirimi için ise 0,01496'dir Üçüncü koşulda GK kestirimi için bu değer 0,01848 MTMG kestirimi için ise 0,01732'dir. Dördüncü koşulda ise GK kestirimi için bu değer 0,01512 MTMG kestirimi için ise 0,01432'dir. Elde edilen bu değerler her iki yaklaşım içinde diğer varyans değerleri arasında ikinci sırada yer almaktadır. Tablo 6'da yer alan σ2(e) değerleri incelendiğinde MTMG yaklaşımı ile hesaplanabilen hata varyansının 0,00055 ile 0,00115 arasında değerler aldığı görülmektedir. MTMG ile elde edilen hata varyansı değerleri farktan çıkarıldığında birey- madde - madde takımı etkileşimi için kestirilen değerler arasındaki fark 0,00001 ile 0,00002 arasında olduğu görülür. Bu durum B evrenin her koşulunda birey-madde-madde takımı değişkenlik kaynağı için GK ve MTMG kestirimi arasında fark olmadığı şeklinde yorumlanabilir.

B evreninde birey ve madde takımı etkileşimi ($\sigma^2_{(bt)}$) diğer varyanslar arasında en büyük değerdir ve koşullar içinde madde takımı uzunluğu - madde takımı sayısı sırasıyla 6-3,6-5,9-3,9-5'tir. Yapılan analizler sonucunda B evreninin her koşulu için MTMG yaklaşımı ve GK yaklaşımı arasında bir fark bulunmamıştır.

B evreninin tüm koşulları için Madde Tepki Modelinde Genellenebilirlik yaklaşımına ve Genellenebilirlik Kuramına göre elde edilen bağıl ve mutlak hata varyansı, genellenebilirlik ve Phi katsayısı arasında fark olup olmadığı incelenmiştir.

Tablo 7'de B evreninin dört farklı koşuluna ait bağıl ve mutlak hata varyansları ile genellenebilirlik ve Phi katsayısı değerleri yer almaktadır.

Tablo 7. B Evreni Koşullarına Ait Bağıl ve Mutlak Hata Varyansı, Genellenebilirlik ve Phi Katsayısı Değerleri

			Birinci koşul	*İkinci koşul*	*Üçüncü koşul*	*Dördüncü koşul*
B evreni	***σ2(δ)***	***GK***	0,01361	0,00786	0,00934	0,00613
		MTMG	0,01359	0,00784	0,00933	0,00612
		FARK	0,00002	0,00002	0,00001	0,00001
	Eb^2	***GK***	0,57973	0,69163	0,65353	0,75004
		MTMG	0,57972	0,69162	0,65352	0,75002
		FARK	0,00001	0,00001	0,00001	0,00002
	$\sigma^2(\Delta)$	***GK***	0,01572	0,00974	0,01154	0,00713
		MTMG	0,01571	0,00973	0,01152	0,00711
		FARK	0,00001	0,00001	0,00002	0,00002
	Φ	***GK***	0,54355	0,63148	0,61845	0,72069
		MTMG	0,54354	0,63146	0,61843	0,72067
		FARK	0,00001	0,00002	0,00002	0,00002

B evreninin tüm koşulları için bağıl hata varyansı kestiriminde iki yaklaşım arasındaki fark her tekrar için 0,00001 ile 0,00002 arasında değişmektedir bu nedenle iki yaklaşım arasında bağıl hata varyansı kestiriminde fark olmadığı söylenebilir.

Bağıl hata varyansına bağlı olarak genellenebilirlik katsayısı değerleri hesaplanmıştır. Genellenebilirlik katsayısı kestiriminde iki yaklaşım arasındaki fark her tekrar için 0,00001 ile 0,00002 arasında değişmektedir bu nedenle iki yaklaşım arasında genellenebilirlik katsayısı kestiriminde fark olmadığı söylenebilir.

B evreninin dört koşulu içinde hesaplanan mutlak hata varyansı değerleri incelendiğinde iki yaklaşım arasındaki farkın her tekrar için 0,00001 ile 0,00002 arasında değiştiği görülmektedir. Bu nedenle dört koşul için iki yaklaşım arasında mutlak hata varyansı kestiriminde fark olmadığı söylenebilir.

Mutlak hata varyansına bağlı olarak Phi katsayısı değerleri hesaplanmıştır. Her dört koşul için Phi katsayısı kestiriminde iki yaklaşım arasındaki fark her tekrar için 0,00001 ile 0,00002 arasında değişmektedir. Bu nedenle iki yaklaşım arasında Phi katsayısı kestiriminde fark olmadığı söylenebilir.

Bağıl hata varyansına bağlı olarak hesaplanan genellenebilirlik katsayısı değerleri ise GK kestirimi

B evrenine ait koşullar madde takımı ve madde takımlarında yer alan madde sayıları açısından farklılık göstermektedir. Madde takımında yer alan madde sayısı 6 olan birinci ve ikinci koşul arasında daha fazla madde bulunan ikinci koşuldan daha yüksek güvenirlik elde edilmiştir. Benzer şekilde madde takımında yer alan madde sayısı 9 olan üçüncü ve dördüncü koşullarda madde sayısı fazla olan dördüncü koşuldan daha yüksek güvenirlik elde edilmiştir. Madde takım sayıları eşit olan birinci-üçüncü ve ikinci-dördüncü koşullarda madde takımı sayısı fazla olan koşullardan daha yüksek güvenirlik elde edilmiştir. Tüm bu bulgular her dört koşul incelendiğinde toplam madde sayısı arttıkça güvenirliğin arttığını göstermektedir. A ve B evrenleri birey-madde takımı etkileşimi varyansı açısından farklılık göstermektedir. İki evren güvenirlik değerleri açısından karşılaştırıldığında her koşul için birey-madde takımı etkileşiminin küçük olduğu A evreninde daha yüksek güvenirlik elde edilmiştir.

SONUÇLAR ve TARTIŞMA

Araştırmada yer alan Genellenebilirlik Kuramı veri seti iki farklı evrenden (A ve B) ve her evren dört farklı koşuldan oluşmaktadır. Genellenebilirlik Kuramı (GK) ve Madde Tepki Modellemesinde Genellenebilirlik (MTMG) modeline göre A ve B evrenlerinin her koşulu için kestirilen birey ($\sigma^2(b)$), madde takımı ($\sigma^2(t)$), birey madde takımı etkileşimi ($\sigma^2(bt)$) ve madde-madde takımı ($\sigma^2(m:t)$) varyans bileşenleri için fark 0,00001 ile 0,00002 arasında değerler almıştır. İki yaklaşım arasındaki varyans bileşenleri değerleri farkının en fazla on binde iki seviyesinde olması aralarında fark olmadığı şeklinde yorumlanabilir. Diğer yandan birey-madde-madde takımı etkileşimi ($\sigma^2(bm:t)$) varyans bileşeni değeri her iki evrenin her koşulu altında iki yaklaşım için en fazla farkın elde edildiği varyans bileşeni olmuştur. Ancak bu durum Genellenebilirlik Kuramı ile Madde Tepki Modellemesinde Genellenebilirlik yaklaşımı arasındaki temel farktan kaynaklanmaktadır. Bu fark; Madde Tepki Modellemesinde Genellenebilirlik yaklaşımının hata varyansını etkileşim varyansından ayırmasıdır. Bir diğer deyişle Genellenebilirlik Kuramı ile kestirilen birey-madde-madde takımı etkileşimi ($\sigma^2(bm:t)$) varyans değeri içinde hata varyansını da barındırır. Bu nedenle MTMG ile elde edilen birey-madde-madde takımı etkileşimi ($\sigma^2(bm:t)$) varyans bileşeni ile hata $\sigma^2(e)$ değerleri toplanıp GK'dan elde edilen birey-madde-madde takımı etkileşimi ($\sigma^2(bm:t)$) varyans bileşeni ile karşılaştırıldığında aradaki farkın 0,00001 ile 0,00002 değerleri arasında olduğu ve bu durumun varyans değerleri arasında fark olmadığı şeklinde yorumlanabileceği görülmüştür. Bu bulgu Briggs ve Wilson'ın (2007) yapmış oldukları MTMG çalışması ile örtüşmektedir.

Genellenebilirlik Kuramı veri seti A ve B evrenleri koşulları altında incelenen bağıl ve mutlak hata varyansı, genellenebilirlik ve Phi katsayısı arasında iki yaklaşım ile kestirilen değerler açısından fark 0,00001 ile 0,00002 arasındadır. Oldukça küçük olan bu değer madde tepki modelinde genellenebilirlik modeli ve Genellenebilirlik Kuramına göre bağıl ve mutlak hata varyansı, genellenebilirlik ve Phi katsayısı arasında fark olmadığı şeklinde yorumlanabilir. Elde edilen bu sonuç MTMG yaklaşımını ortaya atan çalışmalar ile uyum göstermektedir (Briggs ve Wilson, 2004,2007).

Ayrıca A ve B evrenlerine ait koşullar kendi içlerinde madde takımı ve madde takımlarında yer alan madde sayıları açısından farklılık göstermektedir. Madde takımı sayısı 6 olan birinci ve ikinci koşul arasında madde takımında daha fazla madde bulunan ikinci koşulda daha yüksek güvenirlik elde edilmiştir. Benzer şekilde madde takımı sayısı 9 olan üçüncü ve dördüncü koşullarda madde sayısı fazla olan dördüncü koşulda daha yüksek güvenirlik elde edilmiştir. Madde takımlarında bulunan madde sayıları eşit olan birinci-üçüncü ve ikinci-dördüncü koşullarda madde takımı sayısı fazla olan koşullarda daha yüksek güvenirlik elde edilmiştir. Tüm bu bulgular her dört koşul incelendiğinde toplam madde sayısı arttıkça güvenirliğin arttığını göstermektedir. Araştırmanın bu bulgusu madde takımlarının güvenirliği için yapılan çalışmaları (Thissen, Steinberg ve Mooney, 1989; Sireci, Thissen ve Wainer, 1991; Yen, 1993; Wainer, 1995; Wainer ve Thissen, 1996; Ferrara, Huynh ve Bagli, 1997;

Ferrara, Huynh ve Michaels, 1999; Bradlow, Wainer ve Wang, 1999; Wang, Bradlow ve Wainer, 2002) destekler özelliktedir.

Bunun yanında A ve B evrenleri birey-madde takımı etkileşimi varyansı açısından farklılık göstermektedir. Birey-madde takımı değişkenlik kaynağı (σ^2(bt)) madde takımlarının kolaylık ve zorluk seviyelerinin bireylere göre farklılık gösterip göstermediğinin incelendiği varyans değerlerine sahiptir. Madde takımlarının zorluk seviyelerindeki tutarsızlıkların bireylere göre değişmesi hatayı artırır. Bu nedenle İki evren güvenirlik değerleri açısından karşılaştırıldığında her koşul altında birey-madde takımı etkileşiminin küçük olduğu A evreninde daha yüksek güvenirlik elde edilmiştir. Elde edilen bu sonuç birey madde etkileşimi üzerine yapılmış çalışmalardan elde edilen sonuçlarla örtüşmektir (Alkahtani, 2012; Güler, Kaya Uyanık, Taşdelen Teker, 2012; Hendrickson, 2001; Lee ve Frisbie, 1999; Lee ve Park ,2012; Zhang ve Roberts, 2013).

Araştırmanın temel amacı Genellenebilirlik Kuramına alternatif olarak gösterilen MTMG yaklaşımının çok yüzeyli desenlerde de kullanılabileceğini göstermektir. Elde edilen bulgular sonucunda iki yüzeyli bx(m:t) deseni için MTMG ve GK yaklaşımları arasında varyans değerlerini, mutlak ve bağıl hata varyanslarını ve güvenirlik katsayılarını kestirmede fark olmadığı ortaya çıkmıştır. Bu durumda daha pratik ve kolay analiz yapılan programları olan Genellenebilirlik Kuramının yerine kestirimlerinin daha zor yapıldığı MTMG yaklaşımını kullanmak önerilmemektedir. Ancak MTMG'nin GK karşısındaki en büyük avantajı hata varyansını etkileşim varyansından ayrı kestirebilmesidir. Bu sonuçlar doğrultusunda hata varyansını etkileşim varyansından ayrı olarak kestirebildiği için bx(m:t) deseni için MTMG yaklaşımının kullanılması önerilmektedir.

Çalışmada kullanılan çok yüzeyli desen, maddelerin takımlara yuvalandığı ve bireylerin bunlarla çaprazlandığı rastgele dengelenmiş bx(m:t) yuvalanmış desendir. Çalışmada kullanılan tüm evrenler ve evrenlere ait koşulların tümünde elde edilen güvenirlik değerleri madde sayısı arttıkça güvenirliğin arttığını göstermiştir. Bu nedenle madde takımı ile yapılan çalışmalarda madde takımları ve madde takımlarında yer alan madde sayılarının mümkün olduğunca fazla olması önerilmektedir.

Yapılan çalışma ile daha yüksek güvenirlik elde edilmesi için birey-madde takımı etkileşiminin küçük olması gerektiği sonucuna varılmıştır. Bezer şekilde birey-madde takımı etkisi arttıkça güvenirliğin düştüğü görülmüştür. Elde edilen sonuçlara dayanarak madde takımlarının yer aldığı durumlarda güvenirliği arttırmak için birey-madde takımı etkileşiminin küçük olması; madde takımı etkisinin düşük olması ve genel yetenek ve madde takımı arasındaki ilişki ve madde takımları arasındaki ilişkinin düşük olması önerilmektedir.

MTMG yaklaşımı elde edilen ümit verici sonuçların yanında birbirlerinden farklı iki ölçme kuramını bir arada kullanması açısından değerlidir. Ancak MTMG çalışmaları tek yüzeyli desenler ile sınırlı kalmıştır. MTMG yaklaşımının çok yüzeyli durumlarda nasıl işlediğini bilmek gereklidir. Örneğin MTMG çok yüzeyli modellerde çalışmazsa adres gösterilen sorunlarda GK'nın yerine kullanılabileceği söylenemez. Bu nedenle bu çalışmanın temel amacını Briggs ve Wilson'ın (2004, 2007) yaptıkları çalışmayı tek yüzeyli desenden çok yüzeyli desene çıkartmak oluşturmaktadır. Bu temel amaç doğrultusunda elde edilen bilgiler MTMG yaklaşımının bx(m:t) iki yüzeyli deseni içinde uygun bir yaklaşım olduğunu göstermiştir. Ancak MTMG yaklaşımının daha fazla gelişmesi ve uygunluğunun test edilmesi için farklı sayıda yüzeyin bulunduğu farklı desenlerin kullanılması önerilmektedir.

Son yıllarda yapılan testlerde madde takımlarının kullanımını artması madde takımlarının nasıl puanlanacağı, nasıl analiz edileceği ve madde takımlarının güvenirlik üzerinde etkisinin ne olacağı önemini arttıran bir konu haline gelmiştir. Bu çalışma ile kullanılan maddelerin madde takımları içinde yuvalandığı bx(m:t) deseni için güvenirlik değerlerinin farklı koşullar altında nasıl değiştiği gözlenmiştir. Gözlem sonuçlarında madde takımı ve madde sayısı uzunluklarının; birey-madde takımı varyans değerinin; madde takımı etkisinin; madde takımları arasındaki ilişkinin güvenirlik üzerinde etkili olduğu görülmüştür. Gelecekte yapılacak çalışmalar için madde takımları üzerinde etkisinin olabileceği düşünülen diğer etmenlerinde araştırılması önerilmektedir.

KAYNAKÇA

Alkahtani, S. F. (2012). *Oral performace scoring using generalizability theory and many-facet Rasch measurement: A comparison study* (Unpublished Doctoral Dissertation). The Pennsylvania State University.

Bock, R. D., Brennan, R. L., & Muraki, E. (2002). The information in multiple ratings. *Applied Psychological Measurement, 26*, 364-375.

Bradlow, E. T., Wainer, H., & Wang, X. (1999). A bayesian random effects model for testlets. *Psychometrika, 64*, 153-168.

Brennan, R. L. (2001). *Generalizability theory.* New-York: Springer-Verlag.

Briggs, D. C., & Wilson, M. (2004, June). *Generalizability theory in item response modeling*. *Presentation at the International Meeting of the Psychometric Society, Pacific Grove, CA.*

Briggs, D. C., & Wilson, M. (2007). Generalizability theory in item response modeling. *Journal of Educational Measurement, 44*(2), 131-155.

Chien, Y. M. (2008). *An investigation of testlet-based item response models with a random facets design in generalizability theory* (Unpublished Doctoral Dissertation). University of Iowa.

DeMars, C. E. (2006). Application of the bi-factor multidimensional item response theory model to testlet-based tests. *Journal of Educational Measurement, 43*(2), 145-168.

Dimitrov, D. M. (2003). Marginal true-score measures and reliability for binary items as a function of their IRT parameters. *Applied Psychological Measurement, 27(6),* 440- 458.

Dresher, A. R. (2004, April). An empirical investigation of LID using the testlet model: A further look. Paper presented at the Annual Meeting of the National Council on Measurement in Education, San Diego, CA.

Feldt, L. S., & Quails A. L. (1989). Reliability. In R. L. Linn (Ed.), *Educational measurement* (3rd Ed.) (pp. 105-146). New York: American Council on Education and Macmillan.

Ferrara, S., Huynh, F. L., & Bagli, H. (1997). Contextual characateristics of locally dependent open-ended item clusters on a large-scale performance assessment. *Applied Measurement in Education, 12*, 123-144.

Ferrara, S., Huynh, F. L., & Michaels, H. (1999). Contextual explanations of local dependence in item clusters in a large-scale hands-on science performance assessment. *Journal of Educational Measurement, 36*, 119-140.

Fox, J. P., & Glas, C. A. W. (2001). Bayesian estimation of a multilevel IRT model using Gibbs sampling. *Psychometrika*, *66*, 271-288.

Glas, C. A. W. (1989). *Contributions to estimating and testing Rasch models* (Unpublished Doctoral Dissertation). Enschede, University of Twente.

Güler, N., Kaya Uyanık, G. ve Taşdelen Teker, G. (2012). *Genellenebilirlik kuramı*. Ankara: Pegem Akademi.

Hendrickson, A. B. (2001,April). *Reliability of scores from tests composed of testlets: A comparison of methods.* Paper presented at the Annual Meeting of the National Council on Measurement in Education, Seatle, WA.

Jiao, H., Kamata, A., Wang, S., & Jin, Y. (2012). A multilevel testlet model for dual local dependence. *Journal of Educational Measurement, 49*(1), 82-100.

Karasar, N. (2004). *Bilimsel araştırma yöntemi* (13. Baskı). Ankara: Nobel Yayınları.

Kim, S. C., & Wilson, M. (2008). A comparative analysis of the ratings in performance assessment using generalizability theory and the many-facet Rasch model. *Journal of Applied Measurement*, *10*(4), 408-423.

Kolen, M., & Harris, D. (1987, April). *A multivariate test theory model based on item response theory and generalizability theory*. Paper presented at the American Educational Research Association, Washington, DC.

Lee, G., & Park, I. Y. (2012). A comparison of the approaches of generalizability theory and item response theory in estimating the reliability of test scores for testlet-composed tests. *Asia Pacific Education Review, 13*(1), 47-54.

Lee, G., Brennan, R. L., & Frisbie, D. A. (2000). Incorporating the testlet concept in test score analyses. *Educational Measurement: Issues and Practice, 19*(4), 9-15.

Lee, G., & Frisbie, D. A. (1999). Estimating reliability under a generalizability theory model for test scores composed of testlets. *Applied Measurement in Education, 12*(3), 237-255.

Li, Y., Bolt, D. M., & Fu, J. (2006). A comparison of alternative models for testlets. *Applied Psychological Measurement, 30(1), 3-21*.

Linacre, J. M. (1989). *Many-facet Rasch measurement*. Chicago: MESA Press.

Linacre, J. M. (1999). *FACETS (Version 3.17) [Computer software]*. Chicago: MESA Press.

Lord, F. M. (1983). Unbiased estimation of ability parameters, of their variance, and of their parallel forms reliability. *Psychometrika, 48*, 233-245

Patz, R., Junker, B., Johnson, M. S., & Mariano, L. (2002). The hierarchical rater model for rated test items and its application to large-scale educational assessment data. *Journal of Educational and Behavioral Statistics, 27,* 341-384.

Raju, N. S., & Oshima, T. C. (2005). Two prophecy formulas for assessing the reliability of item response theory-based ability estimates. *Educational and Psychological Measurement, 65(3),* 361-375.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Copenhagen: Danish Institute for Educational Research.

Rosenbaum, P. R. (1988). Items bundles. *Psychometrika, 53*(3), 349-359.

Samejima, F. (1977). A use of the information function in tailored testing. *Applied Psychological Measurement, 1,* 233-247.

Samejima, F. (1994). Estimation of reliability coefficients using the test information function and its modifications. *Applied Psychological Measurement, 18,* 229-244.

Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: A primer*. USA: SAGE Publications.

Sireci, S. G., Thissen, D., & Wainer, H. (1991). On the reliability of testlet-based tests. *Journal of Educational Measurement, 28,* 237-247.

Thissen, D., Steinberg, L., & Mooney, J. (1989). Trace lines for testlets: A use of multiple-categorical response models. *Journal of Educational Measurement, 26,* 247- 260.

Verhelst N. D., & Verstralen, H. H. F. M. (2001). IRT models for multiple raters. In A. Boomsma, T. Snijders, and M. Van Duijn (Eds.), *Essays in item response modeling* (pp. 89-108). New York: Springer-Verlag.

Wainer, H. (1995). Precision and differential item functioning on a testlet-based test: The 1991 law school admissions test as an example. *Applied Measurement in Education, 8,* 157-186.

Wainer, H., & Kiely, G. L. (1987). Item clusters and computerized adaptive testing: A case for testlets. *Journal of Educational Measurement, 24*(3), 185-201.

Wainer, H., & Lewis, C. (1990). Toward a psychometrics for testlets. *Journal of Educational Measurement, 27*(1), 1-14.

Wainer, H., & Thissen, D. (1996). How is reliability related to the quality of test scores? What is the effect of local dependence on reliability? *Educational Measurement: Issues and Practice, 15*(1), 22-29.

Wainer, H., & Wang, C. (2000). Using a new statistical model for testlets to score TOEFL. *Journal of Educational Measurement, 37,* 203-220.

Wainer, H., Bradlow, E. T., & Du, Z. (2000). *Testlet response theory: An analog for the 3PL model useful in testlet-based adaptive testing*. Dordrecht: Kluwer Academic Publishers.

Wang, X., Bradlow, E. T., & Wainer, H. (2002). A General bayesian model for testlets: Theory and application. *Applied Psychological Measurement, 26*(1), 109-128.

Wilson, M., & Hoskens, M. (2001). The rater bundle model. *Journal of Educational and Behavioral Statistics, 26,* 283-306.

Yen, W. M. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement, 30,* 187-213.

Zhang, X., & Roberts, W. L. (2013). Investigation of standardized patient ratings of humanistic competence on a medical licensure examination using many-facet Rasch measurement and generalizability theory. *Advances in Health Sciences Education, 18*(5), 929-944.

Zwinderman, A. H. (1991). A generalized Rasch model for manifest predictors. *Psychometrika, 56,* 589-600.

EXTENDED ABSTRACT

Introduction

There are three basic theories in education studies which are Classical Test Theory (CTT), Generalizability Theory (GT) and Item Response Theory (IRT). There are claims that these theories are not completely different, they can be used together or complementary to each other. Therefore, the researchers are focused on the studies combining the CTT and GT with the IRT. In general, designs in which each item is rated by multiple raters were studied in the body of literature in which GT and IRT were used together. However, these studies were criticized for the fact that the IRT violated the local independence assumption. In order to cope with this issue, other models were worked on in IRT. The integration of the IRT and the GT was first achieved by the studies of Briggs and Wilson (2004, 2007). In the result of these studies, a model named Generalizability in Item Response Modeling (GIRM) has been introduced.

In addition to the promising results obtained, the GIRM approach is also valuable in terms of using the two different measurement theories together, which is carried out in the body of literature very

rarely. However, GIRM studies have remained limited to single facet designs. It is necessary to know how the GIRM approach works in many facet cases. For this reason, the main purpose of this study is to develop the study of Briggs and Wilson (2004, 2007) from single facet design to many facet design.

The many facet design used in the study is a random balanced sx (i: t) nested design in which the items are nested in the testlets and the students are crossed with them. Due to the advantages of the tests composed of testlets the usage of them is increasing in both national and international exams. However, it is evident that when the violations of local item independency of the testlets are ignored, they may cause mistakes in the estimation. In this study, the parameters of the testlets obtained under different conditions were examined.

This study is important with regards to combining different theories. At the same time, the different situations in which the testlets in the study were present were examined within the framework of different theories. For this reason, it is considered important, as it suggests a different approach to parameter estimations in large-scale national tests where the testlets are frequently used.

Method

In this study, the results that were obtained using the Generalizability in Item Response Modeling (GIRM) for different conditions were compared with the results that were obtained using the Generalizability Theory (GT) for the same conditions. The research is a simulation study with regard to deriving the appropriate data by creating controlled conditions. The GT linear dataset sx (i: t) was generated with the balanced random design. Every item was nested in only one testlet. ni and nt varied with different working conditions, but the number of students (ns) was fixed at 1000 for this study. The different datasets were generated for each level of the response of the testlet, testlet length, and the number of testlets. Testlet lengths were determined as 6 and 9; the numbers of the testlets were determined as 3 and 5. The numbers of items used in the study were respectively 18, 30, 27, 45 according to the testlet lengths and the number of the testlets determined. The datasets of the study were produced by R software. Generalizability in Item Response Modeling analyses of the data produced were performed with the WinBUGS software, and Generalizability theory was performed with EDUG software and these estimations were compared for each case.

Results and Discussion

The Generalizability Theory dataset in the study consists of two different populations (A and B), and each population consists of four different conditions. According to Generalizability Theory (GT) and Generalizability in Item Response Modeling (GIRM) model, the difference for estimated student for each condition of A and B populations($\sigma 2(s)$), testlet($\sigma 2(t)$), student testlet interaction($\sigma 2(st)$) and item-testlet ($\sigma 2(i:t)$) variance components have the values between 0,00001 and 0,00002. The difference of variance components' values can be interpreted as there is no difference between the two values.

The difference is between 0,00001 and 0,00002 in terms of the values estimated with two approaches between generalizability and Phi coefficient, relative and absolute error variances analysed under the conditions of A and B populations of GT dataset. This value, which is quite low, can be interpreted as there is no difference between relative and absolute error variance, generalizability and Phi coefficient according to the GT and the GIRM. Moreover, the conditions of the A and B populations differ within themselves in terms of the testlets and the number of items in the testlets. Between the first condition and the second condition with 6 testlets, higher reliability is obtained in the second condition, in which there are more items in the testlet. Similarly, among the third condition and the fourth condition, where the number of the testlet is 9, higher reliability is obtained in the fourth condition in which the number of the items is higher. In the first-third conditions and the second-fourth conditions, where the number of items is equal, higher reliability is obtained in the conditions where the number of items in the testlets is higher. All these findings show that when all four conditions are analysed, the reliability

increases as the total number of items increases. The A and B populations, however, differ in terms of the variance of the student-testlet interaction. The student-testlet variance source (σ2 (st)) has the variance value that is studied whether the simplicity and difficulty levels of the testlets differ according to the individuals. If the inconsistencies in the difficulty of the testlets vary from person to person, the error increases. For this reason, when the two populations are compared in terms of their reliability values, higher reliability is obtained in the population where the individual-testlet interaction is low under all conditions. This result is consistent with the results obtained from studies on individual item interaction.