BAŞKENT UNIVERSITY
JOURNAL OF EDUCATION

2020, 7(1), 62-69 ISSN 2148-3272

Mantel Test ve Olabilirlik Oran Testi'nin Farklı Test Koşullarında Değişen Madde Fonksiyonunu Belirlemedeki Gücünün İncelenmesi*

Investigation of the Power of the Mantel Test and Likelihood Ratio Test in Determining Differential Item Functioning under Different Test Conditions

Safiye Bilican Demir[a], R. Nükhet Çıkrıkçı[b]

[a]*Kocaeli University, Kocaeli, Turkey*
[b]*İstanbul Aydın University, İstanbul, Turkey*

Öz

Bu araştırmada, çok kategorili maddelerde değişen madde fonksiyonu (DMF) belirleme testlerinden Mantel Test ve Olabilirlik Oran Testi (OOT)'nin farklı test koşullarında gerçekte DMF içeren bir maddeyi belirlemedeki performansları karşılaştırılmıştır. Monte Carlo simülasyon yaklaşımıyla farklı örneklem büyüklüğü, DMF miktarı ve DMF örüntüsü içeren toplam 16 test koşulu oluşturulmuştur. Simülasyon koşullarıyla ilgili tutarlı sonuçlar elde etmek üzere her bir koşul için 100 tekrar yapılmıştır. Araştırma sonuçları, değişen test koşullarında OOT'nin, Mantel Test'e göre daha iyi performans gösterdiğini ortaya koymuştur. Düşük DMF örüntüsü koşullarına göre, yüksek DMF örüntüsünde testlerin DMF'yi belirlemedeki performansı yetersiz olmuş; örneklem büyüklüğü ve DMF miktarının artışına bağlı olarak her iki testin de istatistiksel gücü yükselmiştir. İlgili testler en iyi performansı büyük örneklem koşullarında göstermiştir.

Anahtar Kelimeler: Değişen madde fonksiyonu, çok kategorili madde, Olabilirlik oran testi, Mantel Test, istatistiksel güç, simülasyon.

Abstract

The purpose of this study was to investigate the power rate of the likelihood ratio (LR) statistic and Mantel Test in detecting differential item functioning (DIF). A Monte Carlo study design was utilized for simulated data sets. There were 16 test conditions including different sample size, DIF magnitude and DIF pattern. Simulation was replicated for 100 times for each test condition to obtain consistent results. The results showed that LR had higher power rates than Mantel test under all test conditions. In the high shift DIF pattern, the performance of the tests was insufficient. The statistical power of both tests increased due to the increase in sample size and the magnitude of DIF.

Keywords: Differential item functioning, Polytomous items, Mantel test, Likelihood Ratio test, power, Monte Carlo simulation.

© 2020 Başkent University Press, Başkent University Journal of Education. All rights reserved.

1. Giriş

Çok farklı alanlarda bireyler hakkında karar vermek üzere testlerden yararlanılmaktadır. Örneğin bireylerin iş veya akademik performansının belirlenmesi, herhangi bir mesleğe yönlendirme, kurumlara personel seçimi, öğretim kurumlarına yerleştirme gibi farklı amaçlar için testlerden elde edilen sonuçlar kullanılmaktadır. Buna bağlı olarak

---

**Bu makale 1. yazarın 2. yazar danışmanlığında Ankara Üniversitesi, Eğitim Bilimleri Enstitüsü'nde tamamladığı doktora tezinin bir bölümünden üretilmiştir.*

ADDRESS FOR CORRESPONDENCE: Safiye Bilican Demir, Department of Educational Science, Faculty of Education, Kocaeli University, Kocaeli, Turkey, E-mail address: safiyebilican@gmail.com. ORCID ID: 0000-0001-9564-9029.
[b]R. Nükhet Çıkrıkçı, Measurement and Evaluation Research Center, Faculty of Education, İstanbul Aydın University, İstanbul, Turkey, E-mail address: nukhetdemirtasli@aydin.edu.tr. ORCID ID: 0000-0001-8853-4733.
Received Date: April 16th, 2019. Acceptance Date: January 28th, 2020.

testlere dayalı olarak bireyler hakkında verilen kararların ne kadar isabetli olduğu kullanılan testlerin psikometrik nitelikleriyle yakından ilişkilidir. Bu yüzden bu araçlardan elde edilen puanların geçerlik ve güvenirlik düzeylerinin belirli ölçütleri karşılaması gerekmektedir (Horst, 1966). Eğer bu ölçütler sağlanmazsa, ilgili testlere dayalı olarak verilecek kararların geçerliğiyle ilgili kuşkular ortaya çıkacaktır. Bunun yanında özellikle uluslararası iş-ticaret dünyasında farklı dil ve kültür alt yapılarına sahip bireylere ilişkin farklı psikolojik yapıların ölçülmesinin yaygınlaşmasına bağlı olarak uyarlanmış testlerin kullanım sıklığı giderek artmaktadır (Robin, Sireci, & Hambleton, 2003; Sireci & Berberoglu, 2000). Ayrıca, eğitim alanında çok farklı kültür ve dil, eğitim sistemi ve sosyo-ekonomik düzeyden gelen öğrenci, öğretmen ya da okul yöneticilerinin katıldığı ölçme uygulamaları bulunmaktadır. Örneğin OECD tarafından yürütülen PISA 2015 uygulamasına 72 farklı ülkeden 15 yaş grubu öğrencileri katılmıştır (OECD, 2017). Bu noktada, bu tür uygulamalarda kullanılan testlerden elde edilen sonuçların testi alan tüm bireyler için karşılaştırılabilir ve adil olduğunu; test puanlarının geçerli ve güvenilir olduğunu gösteren kanıtlara ihtiyaç vardır (Messick 1995). Ancak uluslararası ölçme uygulamalarında kullanılan testlerin çevirisinin yapılması bu testlerin belirtilen koşulları sağlaması için yeterli olmamaktadır (Robin ve diğ., 2003; Sireci, 1997). Bu süreçte, araştırmacı ve test geliştiricilerin, bu testlerin dilsel ve istatistiksel denkliğinin sağladığını ve bu testlerden elde edilen benzer puanların aynı yeterlik düzeyini temsil ettiğini (karşılaştırılabilirlik) ortaya koymaları beklenir (Sireci & Allalouf, 2003). Uyarlanmış bir testte, bahsedilen denkliklerin sağlanması için, testte yer alan her bir maddenin ölçme değişmezliğinin (measurement invariance) istatistiksel olarak sınanması gerekmektedir. Yani, aynı yetenek düzeyinde ancak farklı alt gruplardan gelen bireylerin bir maddeyi yanıtlama olasılıkları benzer olursa o madde için ölçme değişmezliği sağlanır (Penfield & Camilli, 2007). Eğer ölçme değişmezliği sağlanmazsa, bu durum ölçme sonuçlarında yanlılık (bias) olabileceğine işarettir.

1.1. Yanlılık ve Değişen Madde Fonksiyonu (DMF)

Farklı alt gruplar (cinsiyet, etnik köken vs) için testlerden elde edilen puanların karşılaştırılabilirliğini veya ölçme değişmezliğini olumsuz etkileyen tehditlerden bir de yanlılıktır (Gattamorta, Penfield, & Myers, 2012). Yanlılık, testle ölçülmesi amaçlanmayan başka özelliklerin test sonuçlarına karışarak, bu test sonuçlarına dayalı olarak verilen kararların geçerliğini olumsuz etkilemesi olarak tanımlanabilir. Başka bir anlatımla yanlılık, farklı alt gruplardan gelen bireylere ait test puanlarının ait oldukları gruba bağlı olarak değişmesidir (Zumbo, 1999). Böylece grup özelliği, ölçülmesi amaçlanmayan bir değişim (varyans) kaynağı olarak test puanlarına karışmaktadır. Çoğu durumda araştırmacılar, testlerden elde edilen puanların farklı gruplar bakımından değişmez olup olmadığını test etmek üzere Değişen Madde Fonksiyonu (DMF) analizleri yapmaktadır. Yapılan ilgili analizlere dayalı olarak DMF içeren bir madde olası yanlı madde olarak ele alınır. Test denkliğini sağlamak bakımından, DMF analizleri test uyarlama veya test geliştirme sürecinde, olası yanlı maddelerin kontrol edilmesi ya da testten çıkarılması bakımından önemlidir (Kamata & Vaughn, 2004).

Başarı veya yetenek gibi bir psikolojik yapının ölçüldüğü durumlar için DMF, test ile ölçülen özellik bakımından denk; ancak dil, sosyo-ekonomik düzey, etnik köken veya cinsiyet gibi özellikler bakımından alt gruplarda yer alan bireylerin, bir maddeyi doğru yanıtlama olasılıklarının farklılaşmasıdır (Zumbo,1999). Yani, ilgili maddenin ölçülmek istenen özellikten bağımsız olarak bazı alt gruplarda yer alan bireyler için bir avantaj ya da dezavantaj sağlayıp sağlamadığı istatistiksel olarak test edilir. DMF çalışmalarında temel varsayım, ölçülen özellik bakımından benzer bireylerin, ait oldukları alt gruptan bağımsız olarak test maddelerindeki performanslarının da benzer olması gerektiğidir (Angoff, 1993; Zumbo, 1999). Bu varsayımdan yola çıkılarak, ölçülen özellik bakımından denk ancak cinsiyet, etnik köken vb. özellikler bakımından farklı alt gruplar elde edilir. Bu gruplardan odak grup, referans gruba göre dezavantajlı olduğu düşünülen gruptur. Odak ve referans grup, testle ölçülen özellik bakımından eşleştirilir ve madde düzeyinde grupların performansı karşılaştırılır. Karşılaştırma sonuçlarına göre grupların puan ortalaması arasında fark varsa, maddenin doğru cevaplandırılma olasılığının odak ve referans grup için benzer olmadığına ve maddenin DMF içerdiğine karar verilir (Camilli & Shepard, 1994; Zumbo 1999). DMF çalışmalarında odak ve referans grup, testle ölçülen özellik bakımından eşleştirildiği için, gruplar arasında ortaya çıkan farklılıkların DMF'den kaynaklandığı sonucuna varılır.

1.2. Çok Kategorili Maddeler için DMF

Çok kategorili (polytomus) madde yapısının, bilişsel ve duyuşsal özelliklerin ölçümünde kullanımına bağlı olarak bu tür maddeler için de DMF belirleme çalışmaları önem kazanmıştır (Kim, Cohen, Alagöz, & Kim, 2007; Penfield & Camilli, 2007). Çok kategorili maddeler için birçok DMF belirleme yaklaşımı bulunmaktadır. Bunlardan bazıları bireylerin gözlenen puanlarına dayalı yaklaşımlar (observed score approaches) iken (örneğin: Mantel test, Lojistik regresyon), bazıları ise gizil değişken yaklaşımına (latent-variable approaches) dayalıdır (örneğin: Olabilirlik Oran testi, Poly- SIBTEST, Raju'nun alan ölçümleri) (Dorans & Potenza, 1994). Bu çalışmada, güncel test uygulamalarında sıklıkla kullanılan DMF belirleme testlerinden Olabilirlik Oran Testi (OOT) ve Mantel Test ele alınmıştır.

OOT aynı yetenek düzeyindeki bireylerin, maddelere ait tepki kategorilerinin herhangi birinde yanıt verme olasılıklarının birbirinden farklı olup olmadığını test eder. Bu amaçla dar ve geniş model olarak adlandırılan iki model karşılaştırılır. Bu modeller parametrelerin gruplarda eşitlendiği dar (compact, model C) ve DMF şüphesi olan madde/lerin (studied item), her iki grupta eşit parametrelere sahip olma sınırlığının ortadan kalktığı yani ilgili maddeye ait parametrelerin serbest bırakıldığı geniş (augmented, model A) modeldir. Bu modellerden genel uyum istatistiği, -2log olabilirlik (likelihood) değeri, elde edilir ve bu değerleri karşılaştırmak üzere Formül 1'de verilen bir G^2 istatistiği kullanılır.

$$G^2_{(df)} = -2\log\left[\frac{\text{likelihood [A]}}{likelihood[C]}\right] \qquad (1)$$

Formül 1'de Likelihood [A] ve [C] modeldeki parametrelerin olabilirliğini ve df serbestlik derecesini göstermektedir. G^2 istatistiği ki-kare dağılımı gösterir ve ilgili serbestlik derecesinde tablo değeriyle karşılaştırılarak maddenin DMF içerip içermediğine karar verilir.

Çalışmada ayrıca bireylerin gözlenen puanlarına dayalı bir DMF belirleme testi olan Mantel Test kullanılmıştır. Bu test, ikili puanlanan maddeler için bir DMF belirleme testi olan Mantel-Haenszel Testi'nin bir uzantısı olup Zwick, Dougne ve Grima (1993) tarafından çok kategorili puanlanan maddelerde DMF belirlemek üzere önerilmiştir. Bu testte çapraz tablo yaklaşımıyla DMF'yi test etmek üzere, bir maddenin her bir tepki kategorisinde ve bireylerin eşleştirildiği her bir puan aralık düzeyinde odak ve referans grubun gözlenen puanlarını karşılaştırır ve bir serbestlik derecesinde ki-kare dağılımına ilişkin bir istatistik elde edilir (Zwick, ve diğ., 1993). Mantel Test istatistiği Formül 2'deki gibi hesaplanmaktadır:

$$\chi^2_{Mantel} = \frac{[\sum O_K - \sum E(O_K)]^2}{\sum \text{Var}\,(O_K)} \qquad (2)$$

Formül 2'de, O_K: odak grup beklenen toplam puanlarını, $E(O_K)$ odak grup beklenen puanlarını ve Var (O_K) odak grup varyansını göstermektedir. Mantel Test ile referans ve odak gruba ait satır ortalama puanları arasında bir ilişki olup olmadığı test edilir ve satır ortalama puanları arasında bir ilişki varsa, incelenen maddenin DMF gösterdiği kabul edilir. Bu test, matematiksel olarak hesaplama kolaylığı, analizi için ucuz ve ulaşılabilir yazılım gerektirmesi ve uygulama kolaylığı gibi avantajlarından dolayı sıklıkla tercih edilmektedir (Welch & Miller, 1995; Zwick ve diğ., 1993).

Çok kategorili puanlanan maddelerde DMF belirlemek üzere herhangi bir tekniğin seçilmesi genellikle zordur ve bazı durumlarda oldukça karmaşıktır (Kim ve diğerleri, 2007). Bu karmaşa özellikle DMF belirleme testlerinin aynı maddeler için farklı sonuçlar verdiği durumlarda ortaya çıkmaktadır. Birçok DMF belirleme testi olmasına karşın farklı test koşullarında (örneklem büyüklüğü, grup dağılımları, test uzunluğu vb.) hangi testlerin daha tutarlı ve işe yarar sonuçlar verdiğinin belirlenmesi önemlidir. Bilindiği gibi yanlı olduğuna karar verilen maddelerin nihai test formundan çıkarılması ve o test üzerinde yapılacak sonraki işlemlerde (örneğin: madde haritalama veya grup karşılaştırmaları) kullanılmaması gerekmektedir. Bu bakımdan, DMF belirleme testlerinin olası yanlı maddeleri doğru biçimde belirleyebilmesi beklenmektedir. Test koşullarına uygun olarak seçilmeyen bir DMF belirleme testi, gerçekte DMF içermeyen madde/lerin yanlı olduğu değerlendirmesine neden olabilmektedir. Bu durumda ilgili madde ya da maddelerin testten çıkarılması testin yapı ve kapsam geçerliğini olumsuz etkileyebilecektir. Ancak çoğu durumda hangi DMF belirleme testinin seçileceğine karar vermek kolay değildir (Kim ve diğ., 2007). Bu zorluk, ilgili alanda yapılan araştırmaların büyük kısmında (örneğin: Atar, 2007; Fidalgo & Bartram, 2010; Garrett, 2009; Zimbra, 2018) türetilmiş veri kullanılarak aşılmaya çalışılmaktadır. Çünkü türetilmiş veri seti üzerinden çalışmak araştırmacılara bazı avantajlar sunabilir: Araştırmacılar türetilmiş veriler yoluyla maddelere ilişkin gerçek parametreleri bilmektedir ve farklı test koşullarına uygun olarak parametreler üzerinde değişimleme yapabilmektedir. Böylece hangi maddelerin gerçekte DMF içerdiği kontrol edilebilmektedir. Buna bağlı olarak değişen test koşullarında (örneğin: örneklem büyüklüğü ve dağılım özellikleri, madde sayısı, puanlama biçimi vb.) gerçekte DMF içeren bir maddeyi belirleme konusunda hangi testin daha duyarlı ve işe yarar olduğu konusunda doğru karar verilebilecektir. Bu yüzden bu çalışmada Mantel Test ve OOT'nin değişen test koşullarında performansının türetilmiş veri seti kullanılarak karşılaştırılması amaçlanmıştır.

2. Yöntem

2.1. Araştırma Modeli

Bu çalışmada, Monte-Carlo simülasyon yaklaşımı kullanılarak araştırma amacına uygun veri setleri elde edilmiştir.

2.2. Veriler

Araştırmada, incelenen DMF belirleme testlerinin farklı test koşullarında DMF belirlemedeki gücünü incelemek üzere çok kategorili madde örüntüleri türetilmiştir. Bu amaçla, güncel test koşullarına uygunluk, elde edilen verilerin yorumlanmasında kolaylık ve diğer ilgili araştırma sonuçlarıyla karşılaştırılabilirlik gibi durumlar dikkate alınarak farklı test senaryoları elde edilmiştir. Bu senaryolarda bazı değişkenler araştırmanın tüm koşulları için sabit kalırken bazı değişkenler üzerinde manipülasyonlar yapılmıştır. Araştırmada ilgili test koşullarında, MTK modeli, test uzunluğu, DMF içeren madde sayısı, çoklu puanlama türü, gruplara ait dağılım özellikleri araştırmacılar tarafından belirlenen değerlere sahip sabitlerken; örneklem büyüklüğü, DMF miktarı ve DMF örüntüsü ise farklı test senaryolarında test edilen değişkenler olmuştur. Çalışmada yer alan sabitler ve değişkenler Tablo 1'de verilmiştir.

Tablo 1
Çalışmada yer alan sabitler ve değişkenler

Sabitler	Değişkenler
MTK modeli: Kısmi Puan Modeli	Örneklem büyüklüğü: 250:250 ve 1000:1000
Test uzunluğu: 20 madde	DMF miktarı: 0.43 ve 0.64 lojit birim
DMF'li madde sayısı: 3	DMF örüntüsü: Yüksek ve düşük
Çoklu puanlama türü: 4	
Dağılım özelliği: R~$N(0,1)$, O~$N(0,1)$	

Gruplara ait yanıt örüntüleri, parametrelerinin yorumlanmasının ve hesaplanmasının kolay olmasından dolayı Rasch modelinin uzantısı olan Kısmi Puan Modeli (Partial Credit Model) kullanılarak elde edilmiştir. Daha önce yapılan ilgili araştırmalarla testte yer alan madde sayısının DMF belirleme testlerinin performansı üzerinde çok az etkiye sahip olduğu desteklemiştir (Flowers, Oshima, & Raju, 1999; Oshima, Raju, & Nanda, 2006). Dodeen (2004) ve Dodeen ve Johanson (2003) tarafından yapılan çalışmada güncel test uygulamalarına uygunluğu gerekçe gösterilerek test uzunluğu 20 olarak belirlenmiştir. Bu çalışmada da güncel test koşullarına uygunluk, analizlerin tekrarlanma sayısı ve yorumlama kolaylığı dikkate alınarak test uzunluğu 20 madde olarak belirlenmiştir. Bu maddelerin puanlama kategori sayısı, bilişsel ve duyuşsal özelliklerin ölçülmesine uygunluğu dikkate alınarak dört olarak belirlenmiştir. Testte yer alan 20 maddenin %15'ine karşılık gelen üç madde DMF içerecek biçimde modellenmiştir. Referans ve odak gruba ait yetenek dağılımları birim normal dağılım özelliği gösteren yetenek kestirimlerinden seçilmiş ve buna bağlı olarak yanıt örüntüleri elde edilmiştir.

Araştırma amacına uygun olarak üretilen farklı test senaryolarında, OOT analizleri için kullanılan MULTILOG'da parametre kestirimlerinin doğruluğu dikkate alınarak gruplar için örneklem büyüklüklerinin 250:250 ve 1000:1000 olduğu koşullar oluşturulmuştur. Bu çalışmada yorumlanabilir sonuçlar elde etmek üzere DMF miktarı 0.43 ve 0.64 lojit birim olarak ele alınmıştır. Bu değerler Dorans ve Holland (1992) tarafından yapılan sınıflamaya göre orta (moderate) ve büyük (large) DMF miktarını temsil etmektedir. Bu çalışmada ayrıca gerçek test koşullarına uygunluğu dikkate alınarak düşük DMF örüntüsü (low shift DIF pattern) ve yüksek DMF örüntüsü (high shift DIF pattern) koşulları elde edilmiştir. Düşük DMF örüntüsü koşulunda, birinci adım güçlüğü parametresi, 0.43 ve 0.64 lojit birim odak grup için fazla iken, diğer adım güçlüğü parametreleri aynı kalmıştır ($\delta_{1o}= \delta_{1r}+0.43$, $\delta_{2o}= \delta_{2r}$, $\delta_{3o}= \delta_{3r}$). Yüksek DMF örüntüsü koşulunda, üçüncü adım güçlüğü parametresi, 0.43 ve 0.64 lojit birim odak grup için fazladır. Kalan adım güçlüğü parametreleri gruplarda aynıdır ($\delta_{1o}= \delta_{1r}$, $\delta_{2o}= \delta_{2r}$, $\delta_{3o}= \delta_{3r}+0.43$).

Son durumda, Monte Carlo simülsayonlarında ikişer farklı örneklem büyüklüğü, DMF miktarı ve DMF örüntüsü ve iki farklı DMF testi ile 16 farklı senaryo oluşturulmuştur. Bu senaryolarda, 17 madde DMF göstermezken (ortak madde), üç madde (incelenen madde) için DMF miktarı ve örüntüsüne uygun yanıtlar elde edilmiştir. Bu üç madde, adım güçlüğü parametresi bakımından bir değişkenlik ortaya çıkmaması için, parametre değerleri birbirine yakın olan maddeler arasından seçilmiştir.

Çalışmada ayrıca oluşturulan test koşullarıyla ilgili tutarlı sonuçlar elde etmek için, her bir koşul için 100 tekrar yapılmış ve her biri için DMF analizleri yapılmıştır. Simülasyon çalışmalarında 100 tekrar sayısı tutarlı sonuçlar elde etmek için yeterli olarak kabul görmüştür (Kim, 2010).

2.3. Verilerin Analizi ve Yorumlanması

OOT analizleri için, dar ve geniş model karşılaştırmasından elde dilen G^2 istatistiği ve Mantel Test analizlerinden elde edilen ki-kare istatistiği ilgili serbestlik derecesinde ve 0.05 anlamlılık düzeyinde kritik değerle karşılaştırılmış ve maddenin DMF gösterip göstermediğine karar verilmiştir. İlgili testlerin DMF belirlemede yeterli olduğunu değerlendirmek üzere, istatistiksel güç oranlarının 0.80 ve üzerinde (Cohen, 1992) olması şartı aranmıştır.

Araştırmada ilgili test koşullarına uygun veri setlerini elde etmek üzere WinGen3 (Kyung, 2007), OOT analizleri için MULTİLOG (Thissen, Chen, & Bock, 2002) ve Mantel Test analizleri için DIFAS (Penfield, 2005) programları kullanılmıştır.

4. Bulgular

Değişen örneklem büyüklükleri, DMF örüntüsü ve DMF miktarı koşullarına bağlı olarak Mantel Test ve OOT'nin istatistiksel güç oranları incelenmiştir. Düşük DMF örüntüsü koşulunda, değişen örneklem büyüklüğü ve DMF miktarı için ilgili testlerin istatistiksel güç oranları Tablo 2'de verilmiştir.

Tablo 2
Düşük DMF örüntüsü koşulunda DMF belirleme testlerinin istatistiksel güç oranları

	Örneklem Büyüklüğü		DMF Miktarı	
	Referans	Odak	0.43	0.64
Mantel Test	250	250	0.29	0.54
	1000	1000	0.77	0.99
OOT	250	250	0.31	0.70
	1000	1000	0.93	1.00

Tablo 2 incelendiğinde, düşük DMF örüntüsü ve DMF miktarının 0.43 lojit birim olduğu durumda, istatistiksel güç oranları Mantel Test için 0.29 ve 0.77 olmuştur. Her iki örneklem büyüklüğü koşulları için Mantel Test gerçekte DMF gösteren bir maddeyi DMF'li olarak belirlemede zayıf performans göstermiştir. İlgili koşulda OOT için güç oranları 0.31 ve 0.93 olarak elde edilmiştir. OOT'nin 1000:1000 örneklem büyüklüğünde istatistiksel güç değeri 0.93'tür ve sadece bu koşulda DMF'yi belirlemede yeterli olmuştur. Genel olarak ilgili örneklem büyüklüklerinde istatistiksel güç oranları OOT için daha yüksek değerler almıştır.

Düşük DMF örüntüsü ve 0.64 lojit birim DMF miktarında, istatistiksel güç oranları Mantel Test için 0.54 ve 0.99; OOT için 0.70 ve 1 olmuştur. Bu koşulda her iki DMF belirleme testi de 1000:1000 örneklem büyüklüğünde, DMF'yi belirlemede iyi performans göstermiştir. Ayrıca, DMF miktarı ve örneklem büyüklüğün artması, her iki testin de istatistiksel güç değerlerini artırmıştır.

Araştırmada yüksek DMF örüntüsü koşulu için, değişen örneklem büyüklüğü ve DMF miktarı için ilgili testlerin istatistiksel güç oranları incelenmiş ve bulgular Tablo 3'te verilmiştir.

Tablo 3
Yüksek DMF örüntüsü koşulunda DMF belirleme testlerinin istatistiksel güç oranları

	Örneklem Büyüklüğü		DMF Miktarı	
	Referans	Odak	0.43	0.64
Mantel Test	250	250	0.11	0.14
	1000	1000	0.22	0.38
OOT	250	250	0.20	0.28
	1000	1000	0.45	0.75

Tablo 3 incelendiğinde, DMF miktarının 0.43 lojit birim olduğu yüksek DMF örüntüsü için, Mantel Test'in istatistiksel güç oranları 250:250 örneklem büyüklüğünde 0.11 ve 1000:1000 örneklem büyüklüğünde ise 0.22 olmuştur. Mantel Test, her iki örneklem büyüklüğünde de DMF'yi belirlemede yetersiz kalmıştır. Aynı koşulda OOT için istatistiksel güç oranları 250:250 örneklem büyüklüğünde 0.20 ve 1000:1000 örneklem büyüklüğünde ise 0.45 olmuştur. Bu koşulda Mantel Test'e göre OOT'nin istatistiksel güç oranlarının yüksek olmasına karşın gerçekte DMF içeren bir maddeyi belirlemede hala yetersiz kalmıştır. Yüksek DMF örüntüsü ve 0.64 lojit DMF miktarı koşulu için, Mantel Test her iki örneklem büyüklüğünde de DMF'yi belirlemede yetersiz kalmıştır. Bu koşulda güç oranları sırasıyla 0.14 ve 0.38 olmuştur. Benzer şekilde OOT için istatistiksel güç oranları 250:250 örneklem koşulunda 0.28 ve 1000:1000 örneklem büyüklüğünde 0.75 olmuş ve DMF belirlemede yetersiz kalmıştır.

Genel olarak, her iki MİF belirleme testi için de düşük DMF örüntüsünde elde edilen istatistiksel güç değerlerinin yüksek DMF örüntüsüne göre daha yüksek olduğu görülmektedir. Yüksek MİF örüntüsü için elde edilen istatistiksel güç değerleri oldukça düşüktür. İlgili testlerin istatistiksel gücü grupların örneklem büyüklüğü ve MİF miktarından etkilenmiştir.

5. Sonuç, Tartışma ve Öneriler

Bu araştırmada, çok kategorili maddelerde DMF belirleme testlerinden Mantel Test ve Olabilirlik Oran Testi'nin farklılaşan örneklem büyüklüğü, DMF miktarı ve DMF örüntüsü koşullarında gerçekte DMF içeren bir maddeyi belirlemedeki performansı incelenmiştir. Araştırma bulguları, genel olarak DMF belirleme performansı bakımından OOT'nin, Mantel Test'e göre daha iyi olduğunu göstermiştir. Bu durumda, ilgili testlerin DMF belirlemek üzere kullandıkları yaklaşım etkili olabilir. Bilindiği gibi, DMF analizlerinden önce odak ve referans grubu bireyleri ortak maddeler üzerinden ortak bir ölçek üzerinde eşitlenmektedir. İlgili araştırma bulguları, ortak madde sayısının artmasının ve DMF içermemesinin DMF belirleme testlerinin istatistiksel gücünü arttırdığını göstermektedir (örneğin: Bolt, 2002; Meade & Wright, 2012; Woods, 2009; Zimbra, 2018). Meade ve Wright (2012), 20 madde içeren bir test için ortalama beş ortak maddenin daha yüksek istatistiksel güç için önemli olduğunu vurgulamıştır. Bu çalışmadaki ortak madde sayısının fazla olması ve OOT analizleri için ortak maddelerin gerçekte DMF içermemesi bu testin istatistiksel gücünü olumlu yönde etkilemiş olabilir. Ayrıca MTK'ya dayalı bir test olan OOT için, simülasyon koşulları dikkate alındığında model uyumunun tam olarak sağlanması, bu testin DMF belirlemedeki gücünü desteklemiştir. Araştırma bulguları, MTK'ya dayalı DMF belirleme testleri için, model varsayımlarının sağlanmasının ilgili testlerin DMF belirlemedeki istatistiksel gücünü arttırdığını göstermiştir (Sireci & Rios, 2013; Stout, 1990). Mantel Test için, odak ve referans gruptaki bireyler gözlenen toplam puanlar üzerinden eşitlenmekte ve gerçekte DMF içeren üç madde eşleştirme değişkeni içinde yer almaktadır ve bu durum Mantel Test'in gücünü OOT'ye göre olumsuz etkilemiş olabilir. Araştırma bulguları, gözlenen puana dayalı grup eşleştirmelerinde toplam puanın DMF içeren maddelerden oluşmasının testlerin gücünü önemli ölçüde etkilediğini göstermiştir (Jin, Chen & Wang, 2018; Kopf, Zeileis, & Strobl, 2015).

Düşük DMF örüntüsü koşulunda, Mantel Test her iki örneklem büyüklüğünde de orta düzey DMF'yi belirlemede yetersiz kalmıştır. OOT ise sadece büyük örneklem koşulunda gerçekte DMF içeren bir maddeyi belirlemede iyi performans göstermiştir. Bilindiği gibi, MTK'ya dayalı modellerde hem gerekli varsayımların karşılanması hem de daha tutarlı parametre kestirimleri için daha büyük örneklemler kullanılması önerilmektedir (Clauser & Mazor, 1998). Böylece örneklem büyüklüğü arttıkça OOT'nin DMF belirleme performansı da iyileşmiştir. Büyük DMF miktarını belirleme konusunda ise her iki DMF testi, küçük örneklem koşulunda kötü performans göstermiş; ancak büyük örneklem koşulunda ise her iki test gerçekte DMF içeren bir maddeyi belirlemede benzer ve iyi performans göstermiştir. Düşük DMF örüntüsü koşulunu içeren diğer ilgili çalışmalarda da Mantel Test ve OOT için bu araştırma sonuçlarına benzer sonuçlar elde edilmiştir (Atar, 2007; Stark, Chernyshenko, & Drasgow2006). Benzer şekilde farklı MTK modellerin kullanıldığı ve düşük DMF örüntüsü koşullarını içeren başka çalışmalarda (örneğin: Fidalgo & Bartram, 2010; Thurman, 2009) Mantel Test için kabul edilebilir istatistiksel güç oranları elde edilmemiştir. Ayrıca, Mantel Test ve OOT için farklı DMF örüntülerinin (örneğin: sabit ve ıraksak) kullanıldığı araştırma sonuçları da (örneğin: Bolt, 2002; Garrett, 2009; Kim & Yoon, 2011; Meade & Lautenschlager, 2004; Woods, Cai, & Wang, 2013; Zimbra, 2018) benzer şekilde ilgili DMF testlerinin DMF belirleme performansının artan örneklem büyüklüğü ve DMF miktarına bağlı olarak iyileştiğini göstermiştir.

Araştırma bulguları yüksek DMF örüntüsü bağlamında incelendiğinde, örneklem büyüklüğü ve DMF miktarının artışının, her iki testin de istatistiksel güç oranlarını yükselttiğini göstermiştir. Buna karşın, yüksek DMF örüntüsü için testlerin DMF'yi belirlemedeki performansı tüm test koşullarında yetersizdir. Yüksek DMF örüntüsü koşulunu içeren ilgili araştırma bulguları incelendiğinde, elde edilen sonuçların çalışmalarda kullanılan madde parametrelerinin niteliğine bağlı olarak farklılaştığı dikkat çekmektedir. Örneğin, yüksek DMF örüntüsünde, adım güçlüğü parametrelerinin zor olduğu (örneğin: 0,1 ve 2) ilgili araştırmalarda da Mantel Test için kabul edilebilir güç oranları elde edilememiştir (Fidalgo & Bartram, 2010; Thurman, 2009). Atar (2007) tarafından yapılan çalışmada ise adım güçlüğü parametre değerleri (örneğin: -1,25 ve -0.50), yukarıda bahsedilen ilgili araştırmalarda ve bu çalışmada kullanılanlara göre daha düşüktür. Madde parametrelerindeki farklılaşmaya bağlı olarak Atar (2007) OOT için farklı örneklem büyüklüğünde ve DMF miktarında OOT için kabul edilebilir istatistiksel güç değerleri elde etmiştir. Araştırma bulgularındaki bu farklılık, çalışmalarda kullanılan adım güçlüğü parametre değerleriyle yakından ilgili görünmektedir. Çünkü Atar (2007), kolay olan bir maddenin son kategorisini daha zor hale getirmiş ve bu durum tüm yetenek aralığındaki bireyleri etkilemiştir. Buna bağlı olarak odak ve referans grubun madde parametre kestirimleri arasında fark bulma olasılığı artmıştır. Nispeten daha yüksek adım güçlüğü parametrelerinde ise yüksek DMF örüntüsünde gruplar arası madde parametreleri farkları bulmak daha zor hale gelmiştir. Örneğin, Thurman (2009) da çalışmasında yüksek DMF örüntüsünde yeterli güç oranları elde edememesini adım güçlüğü parametrelerinin nitelikleriyle açıklamıştır.

Araştırma bulguları için dikkat çeken diğer nokta, yüksek DMF örüntüsü için elde edilen istatistiksel güç oranlarının oldukça düşük olmasıdır. Bu durum bu çalışmada kullanılan adım güçlüğü parametre değerleriyle yakından ilgili görünmektedir. Daha önce de vurgulandığı gibi, zor bir maddenin son kategorisinde ortaya çıkan DMF (yüksek DMF örüntüsü) yetenek aralıklarının en üstündeki sınırlı sayıda bireyi etkilediği için, gruplar arasında DMF bulmak zorlaşmakta ve buna bağlı olarak testlerin istatistiksel güç oranları düşmüştür. Buna karşın, zor bir maddeyi ilk adımda daha zor hale getirmek (düşük DMF örüntüsü) daha geniş yetenek aralığındaki bireyleri etkilediği için testlerin istatistiksel gücünün yükselme olasılığı ortaya çıkmıştır (Fidalgo & Bartram, 2010; Thurman, 2009).

Araştırma sonuçlarına dayalı olarak, çok kategorili maddelerle yapılacak DMF analizlerinde, gruplarda yer alan birey sayıları dikkate alınarak, büyük örneklem koşullarında OOT, küçük örneklem koşullarında ise Mantel Test'in kullanılması önerilmektedir. Araştırma sonuçları, Mantel Test ve OOT'nin istatistiksel gücünün grupların örneklem büyüklüğünden etkilendiğini göstermiştir. Bu durumda test geliştiriciler, araştırmacılar veya uygulayıcılar hangi DMF belirleme testini kullanacaklarına karar verirken daha güçlü istatistiksel sonuçlar elde etmek üzere örneklem sayılarını iyi değerlendirmeleri gerekmektedir. Bu çalışmada kullanılan model sadece adım güçlüğü parametresini içermektedir. Farklı güçlükteki adım güçlüğü parametreleri ve ayrıca kategori ayırt edicilik parametresini de içeren modellere uyumlu veriler üzerinden DMF belirleme testleri karşılaştırılabilir. Bunun yanında ikili ve çok kategorili maddelerin bir arada olduğu karma (mixed) test desenleri üzerinden farklı DMF testlerinin performansı incelenebilir.

Kaynakça

Angoff, W. H. (1993). Perspectives on Differential Item Functioning Methodology. In P.W. Holland & H. Wainer (Eds.), *Differential Item Functioning* (pp.3-23). Hillsdale, NJ: Erlbaum.

Atar, B. (2007). Differential Item Functioning Analyses For Mixed Response Data Using IRT Likelihood-Ratio Test, Logistic Regression and Gllamm Procedures (Unpublished doctoral dissertation), Florida State University, U.S.

Bolt, D. M. (2002). A Monte Carlo comparison of parametric and nonparametric polytomous DIF detection methods. *Applied Measurement in Education*, 15(2), 113–141. Doi:10.1207/S15324818AME1502_01

Camilli, G., & Shepard, L.A. (1994). *Methods for identifying biased test items*. California: Sage Publications.

Clauser, B. E., & Mazor, K. M. (1998). Using statistical procedures to identify differentially functioning test items. *Educational Measurement: Issues and Practice*, 17(1), 31–44. Doi: 10.1111/j.1745-3992.1998.tb00619.x

Cohen, J. (1992). A power primer. *Psychological Bulletin*, 112(1), 155.

Dodeen, H. (2004). Stability of differential item functioning over a single population in survey data. *Journal of Experimental Education*,72, 181-193. Doi: 10.3200/JEXE.72.3.181-193

Dorans, N. J., & Holland, P. W. (1993). DIF detection and description: Mantel-Haenszel and standardization. In P. W. Holland and H. Wainer (Eds.), *Differential Item Functioning* (p 35-66). Hillsdale, NJ: Erlbaum.

Dorans, N. J., & Potenza, M. T. (1994). *Equity Assessment for polytomously scored items: A taxonomy of prodecures for assessing differential item functioning*. Web: http://www.eric.ed.gov/PDFS/ED380499.pdf adresinden erişilmiştir.

Fidalgo, A. M., & Bartram, D. (2010). A comparison between some generalized Mantel Haenszel statistics for detecting DIF in data simulated under the graded response model. *Applied Psychological Measurement*, 34(8) 600–606. Doi:10.1177/0146621610378405

Garrett, P. (2009). *A Monte Carlo Study Investigating Missing Data, Differential Item Functioning and Effect Size* (Unpublished doctoral dissertation), Georgia State University, U.S.

Gattamorta, K. A., Penfield, R. D., & Myers, D. N. (2012). Modeling item-level and step-level invariance effects in polytomous items using the partial credit model. *International Journal of Testing*, 12(3), 252-272. Doi: 10.1080/15305058.2011.630546

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. CA: Sage.

Horst, P. (1966). *Psychological measurement and prediction*. Belmont: Wadsworth Pub. Co.

Jin, K- Y., Chen, H-F., & Wang, W, W-C. (2018). Using odds ratios to detect differential item functioning. *Applied Psychological Measurement*, 42(8), 613–629. Doi:10.1177/0146621618762738

Kamata, A., & Vaughn, K. B. (2004). An introduction to differential item functioning analysis. *Learning Disabilities: A Contemporary Journal* 2(2), 49-69.

Kim, J. (2010). *Controlling type 1 error rate in evaluating differential item functioning for four DIF methods: Use of three procedures for adjustment of multiple item testing* (Unpublished doctoral dissertation), Georgia State University, U.S.

Kim, E. S., & Yoon, M. (2011). Testing measurement invariance: A comparison of multiple group categorical CFA and IRT. *Structural Equation Modeling: A Multidisciplinary Journal*, 18(2), 212-228. Doi: 10.1080/10705511.2011.557337

Kim, S., Cohen, A. S., Alagoz, C., & Kim, S. (2007). DIF detection and effect size measures for polytomously scored items. *Journal of Educational Measurement*, 44(2), 93-116. Doi: 10.1111/j.1745-3984.2007.00029.x

Kopf, J., Zeileis, A., & Strobl, C. (2015). Anchor selection strategies for DIF analysis: Review, assessment, and new approaches. *Educational and Psychological Measurement*, 75, 22-56. Doi: 10.1177/0013164414529792

Kyung, T. H. (2013). Windows software that generates IRT parameters and item responses: Research and evaluation programs methods (REMP). *Applied Psychological Measurement*,31(5), 457–459. Doi: 10.1177/0146621607299271.

Meade, A. W. & Lautenschlager, G. J. (2004). A comparison of item response theory and confirmatory factor analytic methodologies for establishing measurement equivalence/invariance. *Organizational Research Methods*, 7(4), 361-388. Doi:10.1177/1094428104268027

Meade, A. W., & Wright, N. A. (2012). Solving the measurement invariance anchor item problem in item response theory. *Journal of Applied Psychology*, 97(5), 1016- 1031.Doi: 10.1037/a0027934

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, 50(9), 741-749. Doi: 10.1002/j.2333-8504.1994.tb01618.x

OECD (2017). *PISA 2015 Technical Report*. OECD Publishing, Paris, France. Web: https://www.oecd.org/pisa/sitedocument/PISA-2015-technicalreport-final.pdf adresinden erişilmiştir.

Penfield, R. D. (2005). DIFAS: Differential item functioning analysis system. *Applied Psychological Measurement*, 29 (2), 150–151. Doi: 10.1177/0146621603260686

Penfield, R. D., & Camilli, G. (2007). Differential item functioning and item bias. In S. Sinharay & C. R. Rao (Eds.), *Handbook of statistics*, Volume 26: Psychometrics (pp. 125–167). New York, NY: Elsevier.

Robin, F., Sireci, S. G., & Hambleton, R. K. (2003). Evaluating the equivalence of different language versions of a credentialing exam. *International Journal of Testing*, 3, 1–20. Doi: 10.1207/S15327574IJT0301_1

Sireci, S. G. (1997). Problems and issues in linking assessments across languages. *Educational Measurement: Issues and Practice*, 16, 12–19. Doi: 10.1111/j.1745 3992.1997.tb00581.x

Sireci, S. G., & Allalouf, A. (2003). Appraising item equivalence across multiple languages and cultures. *Language Testing*, 20, 148–166. Doi: 10.1191/0265532203lt249oa

Sireci, S. G., & Berberoglu, G. (2000). Using bilingual respondents to evaluate translated adapted items. *Applied Measurement in Education*, 13, 229–248. Doi:10.1207/S15324818AME1303_1

Sireci, S., & J. Rios (2013). Decisions that make a difference in detecting differential item functioning. *Educational Research and Evaluation: An International Journal on Theory and Practice*, 19(2-3), 170- 187. Doi: 10.1080/13803611.2013.767621

Su, Y.-H., &Wang, W.-C. (2005). Efficiency of the Mantel, Generalized Mantel-Haenszel, and Logistic Discriminant Function analysis methods in detecting differential item functioning for polytomous items. *Applied Measurement in Education*, 18, 313-350. Doi: 10.1207/s15324818ame1804_1

Stark, S., Chernyshenko, O. S., & Drasgow, F. (2006). Detecting differential item functioning with confirmatory factor analysis and item response theory: Toward a unified strategy. *Journal of Applied Psychology*, 91(6), 1292. Doi: 10.1037/0021-9010.91.6.1292

Stout, W. (1990). A new item response theory modeling approach with applications to unidimensionality assessment and ability estimation. *Psychometrika*, 55, 293-325. Doi: 10.1007/BF02295289

Thissen, D., Steinberg, L., & Wainer, H. (1993). Detection of differential item functioning using parameters of item response models. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 67–113). Hillsdale, NJ: Lawrence Erlbaum Associates.

Thissen, D., Chen, W.-H., & Bock, R. D. (2002). MULTILOG [Computer program]. Lincolnwood,IL: Scientific Software International.

Thurman, C. J. (2009). *A Monte Carlo study investigating the influence of item discrimination, category intersection parameters, and differential item functioning patterns on detection of differential item functioning in polytomous items* (Unpublished doctoral dissertation), Georgia State University, U.S.

Welch, C. J., & Miller, T. R. (1995). Assessing differential item functioning in direct writing assessments: problems and an example. Journal of Educational Measurement, 32(2),163-178. Doi: 10.1111/j.1745-3984.1995.tb00461.x

Woods, C. M. (2009). Evaluation of MIMIC-model methods for DIF testing with comparison to two group analysis. *Multivariate Behavioral Research*, 44(1), 1–27.Doi:10.1080/00273170802620121

Woods, C. M., Cai, L., & Wang, M. (2013). The Langer-Improved Wald Test for DIF testing with multiple groups evaluation and comparison to two-group IRT. *Educationaland Psychological Measurement*, 73(3), 532–547. Doi: 10.1177/0013164412464875

Zimbra, J. D. (2018). *An examination of the MIMIC method for detecting DIF and comparison to the IRT likelihood ratio and wald tests* (Unpublished doctoral dissertation), University of Hawaii, U.S.

Zumbo, B. D. (1999). *A Handbook on the Theory and Methods of Differential Item Functioning (DIF): Logistic Regression Modeling as a Unitary Framework for Binary and Likert-Type (Ordinal) Item Scores*. Web: http://educ.ubc.ca/faculty/zumbo/DIF/handbook.pdf adresinden erişilmiştir.

Zwick, R., Donoghue, J. R., & Grima, A. (1993). Assessment of differential item functioning for performance tasks. *Journal of Educational Measurement*, 30,233–251. Doi: 10.1111/j.1745-398.