Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
ISSN: 1309 - 6575

Değişen Madde Fonksiyonunu Belirlemede Mantel-Haenszel ve Lojistik Regresyon Tekniklerinin Karşılaştırılması *

Comparison of Mantel-Haenszel and Logistic Regression Techniques in Detecting Differential Item Functioning

Devrim ERDEM KEKLİK **

Öz
Bu araştırmada, iki kategorili verilerde değişen madde fonksiyonu (DMF) belirlemede kullanılan Mantel-Haenszel (MH) ve lojistik regresyon (LR) tekniklerinin I. Tip hata oranları ve istatistiksel güç değerlerinin odak ve referans grubun yetenek dağılımı, örneklem büyüklüğü ve örneklem büyüklüğü oranlarının değiştiği çeşitli koşullar altında karşılaştırılması amaçlanmıştır. Araştırmada, yetenek kestirimleri ve cevaplayıcı tepkileri WinGen3 simülasyon programı kullanılarak elde edilmiştir. DMF içeren ve DMF içermeyen madde parametreleri iki parametreli lojistik modele uygun olarak üretilmiştir. Araştırma sonucunda, referans ve odak grup yetenek dağılımları birim normal dağılım gösterdiğinde MH ve LR tekniklerinde benzer ve nominal α düzeyine yakın I. Tip hata oranları ortaya çıkmıştır. Buna karşın, referans ve odak grup yetenek dağılımları farklılaştığında MH ve LR tekniklerinde şişirilmiş I. Tip hataların ortaya çıktığı gözlenmiştir. Ayrıca, en yüksek I. Tip hata oranları grupların yetenek dağılımı arasındaki fark yüksek ve geniş örneklem büyüklükleri söz konusu olduğunda ortaya çıkmıştır. Bu araştırma sonuçları, MH ve LR DMF belirlemede şişirilmiş I. Tip hata oranları gözlendiğinde bu testlerin istatistiksel güçlerinin de büyük ölçüde arttığını göstermiştir.

Anahtar Kelimeler: Değişen madde fonksiyonu, I. tip hata, istatistiksel güç, Mantel-Haenszel, lojistik regresyon.

Abstract
The purpose of this study was to compare Type I error rates and power for Mantel-Haenszel (MH) and logistic regression (LR) tests for reference and focal groups with varying ability distributions, sample size and sample size ratios. In this study WinGen3 was used to simulate ability estimates and to generate response data. The two-parameter logistic item response model was used to generate the item parameters for both non-DIF and DIF items. The results showed that with equal group distribution, Type I error rates of MH and LR techniques had similar and near the nominal α level. On the other hand, with unequal group distribution, inflated Type I error rates were observed in MH and LR procedures. Furthermore, the highest type I error rates were observed when difference between ability distributions was high and sample size was large. Results also showed that conditions leading to type I error rate inflation also result in artificially enhanced power rates.

Key Words: Differential item functioning, type I error, power, Mantel-Haenszel, logistic regression.

GİRİŞ

Test puanlarının geçerliğini tehdit eden ciddi etmenlerden biri yanlılıktır. Bir test üzerindeki performansla yansıtılan, o testle ölçülmek istenen özellikten daha farklı bir bilgi veya beceriye dayalıysa ve bu durum belli bir grubun test puanlarının daha fazla geçerli olmasına neden oluyorsa test yanlılığı söz konusu olur (Camilli ve Shepard, 1994; Holland ve Thayer, 1988). Bir testin yanlılığı genellikle madde bazında yapılan analizlerle belirlenir. Madde yanlılığının belirlenmesi iki aşamada gerçekleştirilir: Öncelikle, yetenek düzeyleri benzer

* Bu araştırma, Prof. Dr. Ezel Tavşancıl danışmanlığında yürütülen ve 2012 yılında Ankara Üniversitesi Eğitim Bilimleri Enstitüsü tarafından kabul edilen doktora tezinin bir kısmından üretilmiştir.
** Yard. Doç. Dr., Niğde Üniversitesi, Eğitim Fakültasi, Niğde-Türkiye, erdem_devrim@yahoo.com

olup da farklı gruplara üye olan bireylerin bir maddeye verdikleri tepkilere ilişkin madde fonksiyonlarının farklılaşıp farklılaşmadığı istatistiksel olarak tespit edilir. Eğer bu aşamada grupların bir maddeye verdikleri tepkilerde manidar düzeyde farklılaşma gözlenirse değişen madde fonksiyonu (DMF) varlığı ortaya çıkarılmış olur. DMF içeren maddelerin saptanmasından sonra, maddelerin niteliksel olarak ele alınıp, gözlenen farkların kaynağı açıklanmaya çalışılır. Bu süreçte madde içeriği dikkate alınır ve farklılaşmanın kaynağına ilişkin öne sürülen açıklamalar eğer testin amaçlarının dışında ise o zaman söz konusu maddenin yanlı olduğuna karar verilir (Camili ve Shepard, 1994; Dorans ve Holland, 1993; Holland ve Wainer, 1993). Bu bağlamda, DMF içeren maddenin belirlenmesi yanlılık için bir ön koşuldur ancak yanlılık kararının verilmesinde tek başına yeterli değildir (Zumbo, 1999).

DMF belirleme çalışmalarında ölçülen özellik bakımından eşit yetenek düzeyinde olan farklı alt gruplardaki bireylerin bir madde üzerindeki performansları karşılaştırılır (Potenza ve Dorans, 1995). Dezavantajlı olduğu düşünülen ve azınlığı oluşturan grup *odak (focal) grup* olarak adlandırılırken, bu grubun performansıyla karşılaştırılan çoğunluk grubu ise *referans (reference) grubu* olarak adlandırılır (Camilli ve Shepard, 1994; Holland ve Wainer, 1993).

Değişen madde fonksiyonu "tek biçimli" (uniform) ve "tek biçimli olmayan" (non-uniform) olmak üzere iki türlüdür (Mellenbergh, 1982). Tek biçimli DMF, bir maddeyi doğru yanıtlama olasılığındaki farklılığın grupların bütün yetenek düzeylerinde tutarlılık göstermesi durumunu belirtir. Tek biçimli olmayan DMF ise yetenek ranjı boyunca bir grubun diğer gruba sağladığı avantaj değiştiğinde söz konusu olur (Camilli ve Shepard, 1994). Ancak tek-biçimli olmayan DMF'nin kuramsal olarak mümkün olsa da gerçek durumda nadiren ortaya çıktığı belirtilmektedir (Camilli ve Shepard, 1994; Gierl ve McEwen, 1998; Mazor, Clauser ve Hambleton, 1994).

İki kategorili puanlanan maddeler için DMF değerlendirme yaklaşımları eşleştirme değişkeni olarak gözlenen puanların mı yoksa örtük değişkenin mi kullanıldığına ve tekniğin parametrik bir teknik olup olmadığına göre değişmektedir (Potenza ve Dorans, 1995). Gözlenen puanlara dayalı olan Mantel-Haenszel (non-parametrik) ve lojistik regresyon (parametrik) teknikleri iki kategorili maddelerde özellikle tek biçimli DMF'in belirlenmesinde öne çıkan güçlü yaklaşımlardır.

Mantel – Haenszel

İki kategorili veriler için kullanılan Mantel-Haenszel (Holland ve Thayer, 1988) modelinde, referans ve odak grupları, genellikle toplam test puanı olarak alınan yetenek düzeyi üzerinde eşleştirilir. Burada, odak ve referans olmak üzere iki grup düzeyi, doğru ve yanlış olmak üzere iki tepki kategorisi ve eşleştirme değişkeninin bölündüğü puan aralığı sayısına göre M adet puan düzeyi bulunur. Böylece, her bir madde için 2x2xM'lik kontincensi tablosu oluşturulur.

Mantel-Haenszel (MH) istatistiğinin hesaplanması, eşleştirme gruplarının her biri için madde puanları ortalamasının karşılaştırılmasına dayalıdır ve yokluk hipotezi "*Toplam puanın belirli bir düzeyinde madde puanı ile grup üyeliği arasında koşullu bir ilişki yoktur."* biçiminde belirtilir. Yokluk hipotezi altında, χ^2_{MH} istatistiği bir serbestlik derecesinde ki-kare dağılımı gösterir ve yokluk hipotezinin red edilmesi, aynı yetenek düzeyindeki referans ve odak gruptaki cevaplayıcıların madde üzerindeki ortalama performanslarının farklı olduğu biçiminde yorumlanır.

Lojistik Regresyon

Lojistik regresyon (LR) analizinin DMF belirlemede kullanılması ilk kez Swaminathan ve Rogers (1990) tarafından önerilmiştir. Belirli bir cevaplayıcının belirli bir test maddesine doğru cevap verme durumu U=1 ile diğer durumlar 0 ile gösterildiğinde; cevaplayıcıların bir maddeye doğru cevap verme olasılığı lojistik regresyon modelinde aşağıdaki eşitlikte verildiği şekliyle tanımlanır (Wiberg, 2007):

$$\pi(\mathrm{U}=1)=\frac{\mathrm{e}^{z}}{1+e^{z}},$$
$$z=\log it(\pi_{ij})=\ln\left[\pi_{ij}/1-\pi_{ij}\right]$$

Burada, aşağıda verilen şu üç model değerlendirilir:

1. $Z=\beta_0+\beta_1\theta$
2. $Z=\beta_0+\beta_1\theta+\beta_2 G$
3. $Z=\beta_0+\beta_1\theta+\beta_2 G+\beta_3(\theta G)$

Bu regresyon denklemlerinde, θ cevaplayıcı yetenek düzeyini *G* ise grup aidiyetini belirtir (Swaminathan ve Rogers, 1990). Eğer, β_1 *sıfır*'dan manidar olarak farklılık gösteriyorsa bu durum toplam test puanı arttıkça bir maddeye doğru cevap vermenin göreli (odds) orantısının da arttığını dolayısıyla o madde ile sözkonusu özelliğin ölçülebildiğini gösterir. Eğer β_2, manidar bir biçimde *sıfır*'dan farklılık gösteriyorsa, bu durum tek biçimli DMF'nin varlığına işaret eder. Eğer, β_3 sıfır'dan manidar biçimde farklılaşıyorsa, bu durum tek biçimli olmayan DMF'nin olduğunu gösterir (Camilli ve Shepard, 1994; Camilli, 2006).

İki kategorili verilerde DMF içeren maddelerin saptanmasında kullanılabilecek modellerin hepsinde temel olarak, her alt grupta işlevsel olan bir maddeyi belirlemede kullanılabilecek madde parametrelerinin ortak bir kümesinin oluşturulup-oluşturulamayacağı değerlendirilir. Eğer her grupta işleyen bir maddeyi tanımlamak için farklı parametrelere ihtiyaç duyuluyorsa, o maddenin DMF içerdiği ifade edilir.

İki kategorili verilerde pek çok DMF belirleme modeli geliştirilmesine ve geniş ölçüde incelenmesine rağmen, literatürde hâlâ bu modellerin sınırlılıklarına, avantajlarına ve hangi özellikteki veriler için hangi koşullar altında kullanılabileceklerine yönelik araştırmalar sürmektedir. Öte yandan, literatürde karşılaştırma araştırmalarının doğası gereği sınırlı koşullar altında gerçekleştirildiği görülmektedir (örn.; Jodoin ve Gierl, 2001; Li, Brooks ve Johanson, 2012; Narayanan ve Swaminathan, 1994; Roussos ve Stout, 1996; Sweeney, 1996; Thissen, Steinberg ve Wainer, 1993; Uttaro, 1992; Woods, 2008). Bu nedenle DMF tekniklerinin performanslarının incelenmesinde özellikle I. Tip hata gibi karşılaştırma ölçütlerinin kullanıldığı ve DMF analizi sonuçlarını etkileyebilecek değişkenler olan yetenek dağılımı, örneklem büyüklüğü ve örneklem büyüklüğü oranları koşulları altında, gerçek durumlarda yaygın olarak ortaya çıkan tek biçimli DMF belirleme süreçlerinin incelenmesi gerekli görülmüştür.

Araştırmanın Amacı

Tek biçimli DMF saptamada güçlü bir teknik olarak öne çıkan ve yaygın olarak kullanılan Mantel-Haenszel ve lojistik regresyon süreçleri araştırma kapsamında ele alınmıştır. Bu bağlamda araştırmanın genel amacı, tek-biçimli DMF varlığında odak ve referans grupların yetenek dağılımları farklılaştığında, örneklem büyüklüğü ve örneklem büyüklüğü oranları değiştiğinde, iki kategorili verilerde kullanılan Mantel-Haenszel (MH) ve lojistik regresyon (LR) tekniklerinin I. Tip hata oranı ve istatistiksel güç ölçütlerine dayalı olarak karşılaştırılarak performanslarının değerlendirilmesi olarak belirlenmiştir. Araştırmanın bu genel amaçları doğrultusunda aşağıdaki sorulara yanıt aranmıştır:

1. İki kategorili verilerde MH ve LR tekniklerinin I. Tip hata oranları nedir?
   a. Odak ve referans grupların yetenek dağılımları eşit olduğunda,
   b. Odak ve referans grupların yetenek dağılımları farklılaştığında,
   c. Örneklem büyüklüğü ve örneklem büyüklüğü oranları değiştiğinde

MH ve LR tekniklerinin I. Tip hata oranları nasıldır?

2. İki kategorili verilerde MH ve LR tekniklerinin DMF saptamadaki istatistiksel gücü nasıldır?
   a. Odak ve referans grupların yetenek dağılımları eşit olduğunda,
   b. Odak ve referans grupların yetenek dağılımları farklılaştığında,
   c. Örneklem büyüklüğü ve örneklem büyüklüğü oranları değiştiğinde

MH ve LR tekniklerinin DMF saptamadaki istatistiksel gücü nasıldır?

YÖNTEM

Simülasyon Deseni

Bu kısımda, simülatif verilerin hangi koşullar altında üretildiği belirtilmiştir. İlk aşamada, MH ve LR tekniklerinin I. Tip hataları DMF içermeyen ilk 17 madde üzerinden değerlendirilmiştir. Bu tekniklerin DMF saptamadaki gücü (power) ise DMF içeren son üç madde üzerinden değerlendirilmiştir. Toplam 18 simülasyon koşulu oluşturulmuş ve her bir koşul için 100 tekrar yapılmıştır.

Sabit Faktörler

İki Parametreli Madde Tepki Modeli

Referans ve odak gruplara ait cevaplayıcı tepkileri, yetenek ve madde parametreleri iki parametreli lojistik modele göre üretilmiştir. Gerçek test koşullarında, 1PL modele kıyasla 2PL modelin model veri uyumunun daha iyi olması (Bergan, 2010; Kline, 2005) nedeniyle 1PL model tercih edilmemiştir. Hambleton, Swaminathan ve Rogers (1991), 3PL modelde *c* parametre kestiriminin sıklıkla zayıf olmasından dolayı kestirimin standart hatasının büyüdüğünü, sonuç olarak da DMF'nin belirlenmesinde güçlü bir test istatistiğinin sağlanamadığını belirtmişlerdir. Bu gerekçelerle, veriler 2PL modele göre üretilmiştir.

Test Uzunluğu

Türkiye'de yapılan geniş ölçekli ulusal sınavlarda alt test uzunlukları genellikle 12 ile 25 madde arasında değişmektedir. Bu nedenle, toplam madde sayısı 20 olarak belirlenmiştir.

DMF İçeren Madde Sayısı

Gerçek test koşullarında, incelenen bir alt testte genellikle birden fazla sayıda DMF içeren madde bulunmasından dolayı, bu çalışmada her bir simülasyon koşulu içinde üç madde DMF içerecek biçimde üretilmiştir.

DMF Türü

Tek-biçimli DMF koşulunu sağlamak için a-parametreleri sabit tutularak, odak ve referans gruplar için *b*-parametreleri değiştirilmiştir.

Eşleştirme Değişkeni

Gözlenen puanlara dayalı süreçlerde eşleştirme değişkeni olarak toplam test puanı kullanılmaktadır. Gerçek test koşullarında DMF'den tamamen arınık eşleştirme değişkeninin sağlanması oldukça zor olduğundan bu çalışmada DMF içeren maddeler de eşleştirme değişkeninde yer almıştır.

Manipüle Edilen Faktörler

Referans (R) ve Odak (O) Grupların Yetenek Dağılımları

Grupların yetenek dağılımları aynı olduğunda ve farklılaştığında (iki düzey) DMF sonuçları değerlendirilmiştir. Karşılaştırma gruplarının aynı yetenek dağılımına sahip olduğu koşulda, referans (R) ve odak (O) grupların yetenek dağılımları birim normal dağılım göstermektedir. Diğer iki düzey için ise referans grubun yetenek dağılımı birim normal dağılıma sahip iken, odak grubun yetenek dağılımının sırayla $q_O \sim N(-.5,1)$ ve $q_O \sim N(-.75,1)$ olduğu koşullar oluşturulmuştur. Karşılaştırma gruplarının yetenek düzeylerinin ortalamaları arasındaki .5, .75 ve 1 standart sapmalık farklar gerçek test uygulamaları sonuçlarında yaygın olarak ortaya çıkan değerler olarak belirtilmiştir (Tian, 1999).

Referans ve Odak Grupların Örneklem Büyüklüğü Oranları

Referans ve odak gruplar arasındaki örneklem büyüklükleri ve örneklem büyüklüğü oranları istatistiklerin kestirim sağlamlığını etkileyebilecek faktörlerdir (Swaminathan ve Rogers, 1990). Simülasyon araştırmaları, MH ve LR tekniklerinin yeterli performansı gösterebilmesi için alınabilecek en küçük örneklem büyüklüğünün her grup için 200–250 arasında olması gerektiğini göstermektedir (Narayanan ve Swaminathan, 1996; Rogers ve Swaminathan, 1993). Literatürde orta örneklem genişliği için her grupta 400, 500, 600; büyük örneklem genişliği için ise 750, 800, 1000, 1200, 1600 kişinin örnekleme alındığı görülmektedir (Atar ve Kamata, 2011; Finch ve French, 2008; French ve Maller, 2007). Bu çalışmada örneklem büyüklüğü, küçük-orta-büyük olmak üzere üç düzeyde incelenmiştir. Odak ve referans grupları arasında 1:1 ve 1:2 olmak üzere iki düzey örneklem büyüklüğü oranı belirlenmiştir. Böylece, küçük örneklem genişliği (N=600) için 300:300 ve 200:400; orta örneklem genişliği (N=1200) için 600:600 ve 400:800; büyük örneklem genişliği (N=2400) için ise 1200:1200 ve 800:1600 olmak üzere altı koşul oluşturulmuştur.

DMF Miktarı

Literatürde simülatif veriler üzerinde yapılan çalışmalarda DMF miktarı olarak 0.10, 0.15, 0.25, 0.30, 0.32, 0.43, 0.53 gibi değerlerin kullanıldığı görülmektedir (Atar, 2007; Fidalgo, Mellenberg ve Muniz, 2000; Kristjansson, 2001; Wang ve Su, 2004; Zwick, Donoghue ve Grima, 1993). Bu çalışmada, üç madde DMF içerecek biçimde simüle edilmiş; küçük-orta-yüksek düzey DMF miktarları için 0.10, 0.25 ve 0.50 değerleri seçilmiştir.

Verilerin Üretilmesi

Veriler, hem iki kategorili hem de çok kategorili madde tepki kümeleri üretmek üzere geliştirilen WinGen3 (Han, 2007) programı kullanılarak simüle edilmiştir. WinGen, gerçek durumlara uygun verilerin üretilebildiği, çoğu madde tepki modelini destekleyen, farklı dağılımlara göre madde ve yetenek parametrelerinin üretilmesine izin veren, sonuçların grafiklerle desteklendiği, BILOG-MG, MULTILOG gibi pek çok program için betik dosyası oluşturabilen ve kullanım kolaylığı olan bir programdır. Bu gerekçelerle, verilerin üretilmesinde WinGen3 tercih edilmiştir. Birim normal dağılım gösterececek şekilde üretilen *a* ve *b* paremetre değerleri Tablo 1'de görülmektedir.

Tablo 1. *Madde Parametre Değerleri*

Madde no	1	2	3	4	5
a	0.244	0.527	1.705	-2.172	-1.273
b	-0.024	-0.958	-0.361	-0.203	-0.251
Madde no	6	7	8	9	10
a	0.185	1.012	0.058	1.013	1.718
b	-0.018	0.227	-0.463	-0.764	1.659
Madde no	11	12	13	14	15
a	-0.681	0.832	0.504	-1.911	-0.270
b	-0.101	0.213	0.923	0.070	-0.839
Madde no	16	17	18	19	20
a	-0.363	-1.708	-0.248	-1.316	0.654
b	0.782	0.023	0.205	0.652	0.135

a: madde ayırıcılığı b: madde güçlüğü

Verilerin Analizi

Bu araştırmada Mantel-Haenszel ve lojistik regresyon analizleri SPSS17 programı kullanılarak yapılmıştır. Logistik regresyon analizi ile DMF belirlemede Zumbo (1999) tarafından geliştirilen SPSS betiği (SPSS syntax for binary DIF with WLS R-squared) kullanılmıştır.

BULGULAR

Yetenek Dağılımları Eşit Olduğunda MH ve LR Tekniklerinin I. Tip Hata Oranları

Grup yetenek dağılımı, örneklem büyüklüğü ve örneklem büyüklüğü oranları koşullarında MH ve LR tekniklerinin I. Tip hata oranları ve istatistiksel güçlerine ilişkin değerler toplu olarak Tablo 2'de verilmiştir.

Tablo 2. *MH ve LR Tekniklerinin I. Tip Hata Oranları ve İstatistiksel Güçleri*

		I. Tip Hata		İstatistiksel Güç	
		MH	LR	MH	LR
R~*N*(0,1), O~*N*(0,1)					
O:R=1:1	300/300	0.029	0.053	0.433	0.366
	600/600	0.041	0.029	0.206	0.203
	1200/1200	0.064	0.082	0.633	0.533
O:R=1:2	200/400	0.058	0.047	0.466	0.532
	400/800	0.041	0.063	0.367	0.433
	800/1600	0.023	0.029	0.333	0.300
R~N(0,1), O~N(-0.5,1)					
O:R=1:1	300/300	0.565	0.529	0.833	0.733
	600/600	0.724	0.694	0.933	0.900
	1200/1200	0.782	0.759	0.933	0.933
O:R=1:2	200/400	0.565	0.506	0.580	0.603
	400/800	0.706	0.665	0.800	0.867
	800/1600	0.759	0.735	0.967	0.967
R~N(0,1), O~N(-0.75,1)					
O:R=1:1	300/300	0.782	0.705	1.000	0.933
	600/600	0.894	0.858	1.000	0.966
	1200/1200	0.917	0.929	1.000	0.966
O:R=1:2	200/400	0.783	0.747	0.900	0.800
	400/800	0.847	0.853	1.000	0.900
	800/1600	0.917	0.923	1.000	1.000

Tablo 2 incelendiğinde grupların yetenek dağılımları aynı olduğunda, MH I. Tip hata oranlarının 0.02-0.064 arasında ve LR için I. Tip hata oranlarının 0.02-0.082 arasında değiştiği görülmektedir. Bu bulgular, odak ve referans grupların yetenek dağılımları aynı olduğunda MH ve LR I. Tip hata oranlarının benzer ve nominal α düzeyine yakın olduğunu göstermektedir. Bu sonuçlar literatürdeki benzer çalışmalarla paralellik göstermektedir (Ankenmann, Witt ve Dunbar, 1996; Atar ve Kamata, 2011; Gierl, Jodoin ve Ackerman, 2000; Rogers ve Swaminathan, 1993; Roussos ve Stout, 1996; Vaunghn ve Wang 2010).

Yetenek Dağılımları Farklılaştığında MH ve LR Tekniklerinin I. Tip Hata Oranları

Odak grubun yetenek dağılımı ortalaması referans grubun yetenek dağılımı ortalamasından -0.5 standart sapma farklılaştığında I. Tip hata oranları MH için 0.565-0.782 arasındayken LR için 0.529-0.759 arasında değişmektedir (Tablo 2). Odak grubun yetenek dağılımı ortalaması referans grubun yetenek dağılımı ortalamasından -0.75 standart sapma farklılaştığı koşulda ise I. Tip hata oranları MH için 0.782-0.917 arasında iken LR için 0.705-0.929 arasında değerler almaktadır (Tablo 2). Kısaca, grup yetenek dağılımları farklılaştığında MH ve LR I. Tip hata oranlarının nominal α düzeyinin 10–18.5 katı kadar değerler aldığı gözlenmiştir.

Bu bulgulara dayalı olarak, odak ve referans grupların yetenek dağılımları farklılaştığında MH ve LR için benzer olmakla beraber şişirilmiş I. Tip Hata oranlarının ortaya çıktığı ifade edilebilir. Literatürdeki benzer araştırma sonuçları da bu bulguları desteklemektedir (örn., DeMars, 2009; Güler ve Penfield, 2009; Li ve Stout, 1996; Magis ve De Boeck, 2012; Narayan ve Swaninathan, 1996; Rogers ve Swaminathan, 1993; Swaminathan ve Rogers, 1990). DeMars (2009, 2010) ortaya çıkan bu sonuçları, MH ve LR tekniklerinin gözlenen puanlara dayalı süreçler olmasından dolayı grupların puan ortalamaları arasındaki fark büyüdükçe gerçek puana bağlı grup eşleştirme değişkeninin I. Tip hata oranlarının şişmesine katkı sağladığı şeklinde açıklamıştır. Clauser, Mazor ve Hambleton (1993) ile Penfield (2000) tarafından da, eğer bir test birden fazla DMF'li madde içeriyorsa bu durumun eşleştirme değişkeni (toplam test puanı) üzerinde büyük ölçüde bozulma meydana getireceği ve böylece gözlenen puanlara dayalı yöntemlerin DMF belirleme performanslarının etkileneceği vurgulanmıştır.

Bu açıklamalara ek olarak, neden I. Tip hata oranlarında şişme gözlendiği kurulan simülasyon deseninin yapısıyla da açıklanabilir. Zira bu çalışmada kısa bir test kullanılmıştır ve eşleştirme değişkeni üç değişik düzeyde DMF'li maddeler içermektedir. Literatürde de DMF belirleme amacıyla yapılan simülasyon çalışmalarında eşleştirme değişkeninin DMF içerdiği durumlarda I. Tip hata oranlarında şişme gözlenebileceği vurgulanmıştır (Stark, Chernyshenko, Drasgow ve Williams, 2006; Wang ve Yeh, 2003; Woods, 2009).

Örneklem Büyüklüğü ve Örneklem Büyüklüğü Oranları ile MH ve LR Tekniklerinin I. Tip Hata Oranları

Grup yetenek dağılımları eşitken MH ve LR I. Tip hata oranlarının örneklem büyüklüğüne göre dalgalanma gösterdiği, buna karşın grup yetenek dağılımları farklı olduğunda bu tekniklerin I. Tip hata oranlarının örneklem büyüklüğü arttıkça arttığı ve büyük örneklem genişliklerinde (N_F=N_R=1200 ve N_F= 800, N_R=1600) I. Tip hata oranlarının en yüksek değerleri aldığı gözlenmiştir (Tablo 2).

Odak grubun yetenek dağılımı çarpıklaştıkça MH ve LR I. Tip hata oranlarının şişmesi ve örneklem büyüklüğü arttıkça da en yüksek I. Tip hata oranlarının gözlenmesi literatürdeki araştırma sonuçları ile de tutarlılık göstermektedir (Clauser, Jodoin ve Gierl, 2001; DeMars, 2009; Li, Brooks ve Johanson; 2012; Li ve Staut; 1996; Roussos ve Stout, 1996). Mazor ve Hambleton (1994) ile Donoghue ve Allen (1993) ortaya çıkan bu durumun

nedenini, MH yönteminin gözlenen puanlara dayalı olduğu ve her bir gözlenen toplam puanın bir puan seviyesini oluşturduğu; ancak kısa testler ve farklılaşan grup yetenek dağılımları söz konusu olduğunda geniş örneklem büyüklüklerine bağlı olarak bazı ham puanların kaba bir biçimde iç içe geçtiği ve bu eğilimin I. Tip hatalarda şişme meydana getirdiği şeklinde açıklamışlardır.

Grup Yetenek Dağılımları Eşit Olduğunda MH ve LR Tekniklerinin İstatistiksel Gücü

Referans ve odak grup dağılımları aynı olduğunda MH ve LR testlerinin DMF içeren madde belirleme güçlerinin 0.64 değerinin altında kalarak DMF belirlemede yeterli performansı göstermedikleri ortaya çıkmıştır.

Grup Yetenek Dağılımları Farklılaştığında MH ve LR Tekniklerinin İstatistiksel Gücü

Odak ve referans grupların yetenek dağılımları farklı olduğunda MH ve LR tekniklerinin istatistiksel gücüne ilişkin bulgular genel olarak incelendiğinde, bu tekniklerin DMF içeren maddelerin belirlenmesinde benzer güçte ve çok yüksek düzeylerde (%83'ünün 0.73'ün üstünde) olduğu görülmektedir (Tablo 2). Araştırma bulguları bir bütün olarak değerlendirildiğinde ise MH ve LR DMF belirleme süreçlerinde şişirilmiş I. Tip hata oranları gözlendiğinde bu testlerin istatistiksel güçlerinin de büyük ölçüde arttığı ortaya çıkmıştır. Literatürdeki çalışmalarda da benzer sonuçların elde edildiği görülmektedir. Jodoin ve Gierl (2001) yaptıkları simülasyon çalışmasında DMF belirlenmesinde lojistik regresyonun I. Tip hata oranları ile istatistiksel gücünü incelemişler ve I. Tip hata oranlarındaki küçük değişimlerin testin istatistiksel gücünü büyük ölçüde etkilediğini belirtmişlerdir. Li ve Stout (1996) da I. Tip hata oranlarındaki şişmeye bağlı olarak LR testinin gücünün oldukça yüksek çıkma eğiliminde olduğunu, buna ek olarak MH ve LR DMF tekniklerinin, I. Tip hata oranlarının büyük olması durumunda bu testlerin istatistiksel güçlerinin yorumlanmasının sorunlu olduğunu vurgulamışlardır. Jodoin ve Gierl (2001) DMF belirleme sürecinde testlerin istatistiksel güçlerinin yorumlanabilmesi için I. Tip hata oranlarının karşılaştırılabilir olması gerektiğini, çünkü ortaya çıkan şişirilmiş I. Tip hata oranlarının bir testin istatistiksel gücünü yapay bir biçimde büyüttüğünü ifade etmişlerdir.

Ankenmann, Witt ve Dunbar (1999) ise bütün analizlerde nominal α düzeyi için %95 güven aralığının oluşturulması gerektiğini ve üst güven sınırları dışında kalan şişirilmiş I. Tip hata oranlarının ortaya çıktığı koşullarda testin istatistiksel gücünün rapor edilmemesi gerektiğini belirtmişlerdir. Finch (2005) de yaptığı simülasyon çalışmasında I. Tip hata oranındaki şişmeye bağlı olarak, I. Tip hata oranı 0.10 değerinin üstünde olan pek çok koşul için, kullandığı testlerin gücünü rapor etmediğini belirtmiştir. Yapılan bu araştırmada da testlerin istatistiksel güçlerine ilişkin bulgularının değerlendirilmesinde literatürdeki bu sonuçlar ve öneriler dikkate alınmıştır. Dolayısıyla, MH ve LR I. Tip hata oranlarının litaratürde belirtilen hata sınırlarının çok üstünde olduğu koşullar için bu testlerin istatistiksel güçlerine ilişkin yorum ve karşılaştırma yapılmamıştır.

Örneklem Büyüklüğü ve Örneklem Büyüklüğü Oranları ile MH ve LR Tekniklerinin İstatistiksel Gücü

Örneklem büyüklüklerine bağlı olarak istatistiksel gücün nasıl değiştiği grup dağılımlarına göre değerlendirilmiştir (Tablo 2). MH ile DMF belirlenmesi sürecinde grup dağılımları aynıyken, odak ve referans grupların örneklem büyüklükleri oranları 1:1 olduğunda istatistiksel güç değerlerinin dalgalanma gösterdiği ve bu değerlerin 0.206 – 0.633 arasında değiştiği, bu oranlar 1:2 olduğunda ise MH tekniğinin istatistiksel gücünün örneklem büyüklüğü arttıkça gittikçe düştüğü ve değerlerin 0.333–0.466 arasında değiştiği görülmektedir.

Grup dağılımları ortalaması arasında 0.5 standart sapmalık fark olduğu durumda odak ve referans grupların örneklem büyüklükleri oranları 1:1 olduğunda istatistiksel güç değerlerinin örneklem büyüklükleri arttıkça arttığı ve bu değerleri 0.833 – 0.933 arasında değiştiği, bu oranlar 1:2 olduğunda ise MH tekniğinin istatistiksel gücünün örneklem büyüklüğü arttıkça gittikçe arttığı ve değerlerin 0.580 – 0.967 arasında değiştiği görülmektedir.

Grup dağılımları ortalaması arasında 0.75 standart sapmalık fark olduğu durumda ise odak ve referans grupların örneklem büyüklükleri oranları 1:1 iken istatistiksel güç değerlerinin bütün örneklem büyüklüklerinde 1.000 değerini aldığı, bu oranlar 1:2 olduğunda ise MH tekniğinin istatistiksel gücünün sadece 200/400 koşulu altında 0.900 olduğu diğer koşullarda 1.000 değerlerini aldığı ortaya çıkmıştır. Bunlara ek olarak, grup yetenek dağılımları aynıyken, MH için en düşük istatistiksel gücün 600/600 koşulunda, en yüksek istatistiksel gücünse 1200/1200 koşulunda ortaya çıktığı görülmektedir. Grup yetenek dağılımları farklılaştığında ise, MH için en düşük istatistiksel gücün 200/400 koşulunda, en yüksek istatistiksel gücünse 800/1600 koşulunda ortaya çıktığı görülmektedir.

Farklı örneklem büyüklüklerinde LR için ortaya çıkan istatistiksel güç değerleri ise grup dağılımlarına göre değerlendirilmiştir. Buna göre, LR ile DMF belirlenmesinde grup dağılımları aynıyken, odak ve referans grupların örneklem büyüklükleri oranları 1:1 olduğunda istatistiksel güç değerlerinin dalgalanma gösterdiği ve bu değerlerin 0.203 – 0.533 arasında değiştiği, bu oranlar 1:2 olduğunda ise LR tekniğinin istatistiksel gücünün örneklem büyüklüğü arttıkça düştüğü ve değerlerin 0.300–0.532 arasında değiştiği görülmektedir.

Grup dağılımları farklılaştığında ise genel olarak örneklem büyüklüğü arttıkça LR tekniğinin DMF belirleme gücünün de arttığı görülmektedir. Grup dağılımları ortalaması arasında 0.5 standart sapmalık fark varken ve odak ve referans grupların örneklem büyüklükleri oranları 1:1 olduğunda istatistiksel güç değerlerinin 0.733 – 0.933 arasında değiştiği, bu oranlar 1:2 olduğunda ise bu değerlerin 0.603 – 0.967 arasında değiştiği görülmektedir. Grup dağılımları ortalaması arasında 0.75 standart sapmalık fark olduğu durumda ise odak ve referans grupların örneklem büyüklükleri oranları 1:1 iken istatistiksel güç değerlerinin 0.933–0.966 arasında, bu oranlar 1:2 olduğunda ise LR tekniğinin istatistiksel gücünün 0.800–1.000 arasında değiştiği bulunmuştur. Bunlara ek olarak, grup yetenek dağılımları aynı iken, LR için en düşük istatistiksel güç değerinin 600/600 koşulunda, en yüksek istatistiksel güç değerinin ise 1200/1200 koşulunda ortaya çıktığı görülmektedir. Grup yetenek dağılımları farklılaştığında ise, LR için en düşük istatistiksel güç değerlerinin 200/400 koşulunda, en yüksek istatistiksel güç değerinin ise 800/1600 koşulunda ortaya çıktığı görülmektedir.

Bu araştırma sonucunda, örneklem büyüklüğü oranları 1:2 olduğunda, odak ve referans grupların örneklem büyüklükleri arttıkça MH ve LR testlerinin istatistiksel güçlerinin de düştüğü dikkati çekmektedir. Benzer bir bulguya Vaughn ve Wang (2010) yaptıkları araştırmada rastlanmaktadır. Bu araştırmacılar MH ve LR testlerinin istatistiksel gücünün örneklem büyüklüğü arttıkça düştüğünü belirtmişlerdir. Bu bugulara paralel olarak Jodoin ve Gierl (2001) de, eşleştirme değişkeninin %10 DMF içerdiği koşul altında MH ve LR testlerinin gücünün genel olarak örneklem büyüklüğü arttıkça arttığını, ancak odak ve referans grup örneklem büyüklüğünün en dengesiz olduğu koşulda (N_R=1000 ve N_F=250) bu testlerin güçlerinin en düşük değeri aldığını belirtmişlerdir.

I. Tip hata oranındaki şişmeye bağlı olarak MH ve LR testlerinin DMF belirleme güçlerindeki yapay büyümeyi meydana getiren grup yetenek dağılımlarının farklı olduğu koşullar dışarıda tutularak bir değerlendirme yapıldığında, bu araştırma sonucunda MH ve LR DMF belirleme güçlerinin 0.64 değerinin altında kalarak yetersiz bir performans gösterdikleri ortaya çıkmıştır. Bu durumun nedeni, elde edilen madde ayırıcılık

parametrelerinin yüksek olmaması ile açıklanabilir; çünkü bu araştırmada madde ayırıcılık parametreleri belirli bir ayırıcılık düzeyini koruyacak biçimde ayarlanmamıştır. Chang, Mazzeo ve Roussos (1996) da madde ayırıcılığı arttıkça buna bağlı olarak kullanılan testlerin DMF belirleme güçlerinin de büyük ölçüde arttığını vurgulamışlardır. Ayrıca, yine belirtilen (yetenek dağılımlarının eşit olduğu) koşul çerçevesinde ele alındığında örneklem büyüklüğü oranları 1:2 iken MH ve LR testlerinin güçlerinin, örneklem büyüklüğü arttıkça düştüğü ortaya çıkmıştır. Bu testlerin belirtilen koşullarda istatistiksel güçlerinin örneklem büyüklüğü arttıkça düşmesi, eşleştirme değişkeninde, örneklem büyüklüğünden dolayı, daha fazla bozulmaya bağlı olarak DMF belirleme gücünün zorlaşması ile açıklanabilir. Bu çalışmada, belirtilen koşullar altında istatistiksel güç değerlerinin düşük çıkması madde sayısının azlığına bağlı da olabilir. Uzun testler cevaplayıcı tepkilerine ilişkin daha fazla bilgi vermekte bu da parametre kestirimini dolayısıyla DMF belirleme gücünü arttırmaktadır (Atar ve Kamata, 2011).

SONUÇ ve TARTIŞMA

Bu araştırmada, odak ve referans grupların yetenek dağılımları, örneklem büyüklüğü ve örneklem büyüklüğü oranlarının değiştiği koşullar altında, tek-biçimli DMF varlığında iki kategorili veriler için kullanılan Mantel-Haenszel ve lojistik regresyon tekniklerinin I. Tip hata oranı ve istatistiksel güç ölçütlerine dayalı olarak performanslarının değerlendirilmesi amaçlanmıştır. Bu araştırmada odak ve referans grupların yetenek dağılımları aynı olduğunda, MH ve LR I. Tip hata oranlarının benzer ve nominal α düzeyine yakın olduğu, grupların yetenek dağılımları farklılaştığında ise şişirilmiş I. Tip hata oranlarının ortaya çıktığı saptanmıştır. Ayrıca**,** en yüksek I. Tip hata oranları grupların yetenek dağılımı arasındaki fark yüksek ve geniş örneklem büyüklükleri söz konusu olduğunda gözlenmiştir. Bu araştırma sonuçları, MH ve LR DMF belirlemede şişirilmiş I. Tip hata oranları gözlendiğinde bu testlerin istatistiksel güçlerinin de büyük ölçüde arttığını göstermiştir.

Bu araştırma sonucunda, referans ve odak grup yetenek dağılımları birim normal dağılım gösterdiğinde MH ve LR tekniklerinde benzer ve düşük I. Tip hata oranları elde edilmiştir. Bu nedenle, gerçek test koşullarında uygulayıcılar grup yetenek dağılımları aynı olduğunda bu iki tekniği de tercih edebilirler. Ancak referans ve odak grup yetenek dağılımları farklılaştığında MH ve LR tekniklerinde şişirilmiş I. Tip hata oranlarının ortaya çıktığı gözlenmiştir. Bu nedenle, ileride yapılacak araştırmalarda normal dağılım bozulumu durumlarında bu tekniklerin ve diğer DMF belirleme tekniklerinin performansının değerlendirilebileceği daha ayrıntılı araştırmalar yapılabilir.

İleride DMF alanında yapılabilecek araştırmalarda bu araştırma kapsamında sabit faktör olarak alınan fakat DMF belirleme tekniklerinin performansını etkileyebilecek test uzunluğu, testteki DMF içereren maddelerin oranı, odak grup sayısı, madde ayırıcılığı gibi farklı değişkenlerle oluşturulacak desenlerle DMF belirleme tekniklerinin performansları incelenebilir. Tek-biçimli olmayan DMF belirlemede kullanılan tekniklerin karşılaştırılacağı araştırmalar yapılabilir. Ayrıca daha farklı örneklem büyüklükleri ve örneklem büyüklüğü oranları kullanılarak benzer değişkenlere ilişkin DMF belirleme tekniklerinin performansları karşılaştırılabilir.

KAYNAKLAR

Ankenmann, R. D., Witt, E. A. and Dunbar, S. B. (1999). An investigation of the power of the likelihood ratio goodness-of-fit statistics in detecting differential item functioning. *Journal of Educational Measurement, 36*, 4, 277–300.

Atar, B. (2007). Differential item functioning analyses for mixed response data using IRT likelihood-ratio test, logistic regression, and GLLAMM procedures (Doctoral Dissertation, Florida State University, Tallahassee).

Atar, B. and Kamata, A. (2011). Comparison of IRT likelihood ratio test and logistic regression DIF detection procedures. *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi, 41*, 36–47.

Bergan, J. R. (2010). Assessing the relative fit of alternative item response theory models to the data.

Camilli, G. (2006). Test fairness. In Brennan, R. L. (Ed). Educational Measurement (p. 221–257). CT: Wesport. American Council on Education.

Camilli, G. and Shepard, L. A. (1994). Methods for identifying biased test items. V 4. Thousand Oaks: Sage Publications, Inc.

Chang, H., Mazzeo, J. and Roussos, L. (1996). Detecting DIF for polytomously scored items: an adaptation of SIBTEST procedure. *Journal of Educational Measurement, 33*(3), 333–353.

Clauser, B. E., Mazor, K. M., and Hambleton, R. K. (1993). The effects of purification of the matching criterion on the identification of DIF using the Mantel-Haenszel procedure. *Applied Measurement in Education, 6*, 269–279.

Clauser, B., Mazor, K., and Hambleton, R. K. (1994). The effects of score group width on the Mantel-Haenszel procedure. *Journal of Educational Measurement, 31*, 67–78.

DeMars, C. E. (2009). Modifcation of the Mantel-Haenszel and logistic regression DIF procedures to incorporate the SIBTEST regression correction. *Journal of Educational and Behavioral Statistics, 34*, 149-170.

DeMars, C. E. (2010). Type I error inflation for detecting DIF in the presence of impact. *Educational and Psychological Measurement, 70*, 961–972.

Dorans, N. J. and Holland, P. W. (1993). Detection of differential item functioning using the parameters of item response models. In P. W. Holland ve H. Wainer (Eds.), Differential item functioning (pp. 25–66). Hillsdale, NJ: Lawrence Erlbaum Associates.

Donoghue, J. R., and Allen, N. L. (1993). Thin versus thick matching in the Mantel- Haenszel procedure for detecting DIF. *Journal of Educational Statistics, 18*, 131–154.

Fidalgo, A. M.; Mellenberg, G. J., and Muniz, J. (2000). Effects of amount of DIF, test length, and purification type on robustness and power of Mantel-Haenszel procedures. *Methods of Psychological Research Online, 5*(3), 43–53.

Finch, H. (2005). The MIMIC model as a method for detecting DIF: comparison with Mantel-Haenszel, SIBTEST, and the IRT likelihood ratio. *Applied Psychological Measurement, 29*, 278–295.

Finch, W. H. and French, B. F. (2008). Anomalous type I error rates for identifying one type of differential item functioning in the presence of the other. *Educational and Psychological Measurement, 68*, 742–759.

French, B. F. and Maller, S. J. (2007). Iterative purification and effect size use with logistic regression for differential item functioning detection. *Educational and Psychological Measurement, 67*(3), 373– 393

Gierl, M. J., Jodoin, M. G., and Ackerman, T. A. (2000). *Performance of Mantel-Haenszel, simultaneous item bias test, and logistic regression when the proportion of DIF items is large.* Paper Presented at the Annual Meeting of the American Educational Research Association (AERA) New Orleans, Louisiana, USA April 24–27.

Gierl, M. J. and McEwen, N. (1998, May). *Differential item functioning on the Alberta Education Social Studies 30 diploma examination.* Paper presented at the annual meeting of the Canadian Society for the Study of Education, Ottawa, ON, Canada.

Güler, N., and Penfield, R. D. (2009). A comparison of logistic regression and contingency table methods for simultaneous detection of uniform and nonuniform DIF. *Journal of Educational Measurement, 46*, 314-329.

Han, K. T. (2007). WinGen: Windows software that generates IRT parameters and item responses. *Applied Psychological Measurement, 31*(5), 457-459.

Hambleton, R. K., Swaminathan, H. and Rogers, J. H. (1991). Fundamentals of item response theory. Sage Publications. Buston.

Holland, P. W. and Thayer, D. T. (1988). Differential item functioning detection and the Mantel-Haenszel procedure. In H. Wainer and H. Braun (Eds.), Test Validty (pp. 129–145). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Holland, P. W. and Wainer, H. (Eds.). (1993). Differential item functioning. Hillsdale, NJ: Erlbaum.

Jodoin, M. G. and Gierl, M. J. (2001). Evaluating type I error and power using an effect size measure with the logistic regression procedure for DIF detection. *Applied Measurement in Education, 14*(4), 329–349.

Kline, T. (2005). Psychological Testing: A practical approach to design and evaluation. SAGE Publications.

Kristjansson, E. (2001). Detecting DIF in polytomous items: an empirical comparison of the ordinal logistic regression, logistic discriminant function analysis, Mantel, and generalized Mantel Haenszel procedures (Unpublished Doctoral Dissertation, University of Ottawa, Ottawa).

Li, H. and Stout, W. (1996). A new procedure for detection of crossing DIF. *Psychometrika, 61*, 647–677.

Li, Y., Brooks, G. P. and Johanson, G. A. (2012). Item discrimination and type I error in the detection of differential item functioning. *Educational and Psychological Measurement, 72*(5), 847-861.

Magis, D. and De Boeck, P. (2012). A robust outlier approach to prevent type I error inflation in differential item functioning. *Educational and Psychological Measurement, 72*(2), 291-311.

Mazor, K. M., Clauser, R. E. and Hambleton, R. K. (1994). Identification of nonuniform differential item functioning using a variation of the Mantel-Haenszel procedure. *Educational and Psychological Measurement, 54*, 284–291.

Mellenberg, G. J. (1982). Contingency table models for assessing item bias. *Journal of Educational Statistics, 7*, 105–118.

Meredith, W. and Millsap, R. E. (1992). On the misuse of manifest variables in the detection of measurement bias. *Psychometrika, 57*, 289–311.

Millsap, R. E. and Meredith, W. (1992). Inferential conditions in the statistical detection of measurement bias. *Applied Psychological Measurement, 16*, 389–402.

Narayanan, P., and Swaminathan, H. (1994). Performance of the Mantel-Haenszel and simultaneus item bias procedures for detecting differential item functioning. *Applied Psychological Measurement, 18*(4), 315–328.

Narayanan, P., and Swaminathan, H. (1996). Identification of items that show nonuniform DIF. *Applied Psychological Measurement, 20* (3), 257–274.

Penfield, R. D. (2000). The effects of matching criterion contamination on the Mantel- Haenszel procedure (Doctoral thesis, University of Toronto).

Potenza, M. T. and Dorans, N. J. (1995). DIF assessment for polytomously scored items: a framework for classification and evaluation. *Applied Psychological Measurement, 19*, 23–37.

Rogers, H. J. and Swaminathan, H. (1993). A comparison of logistic regression and Mantel-Haenszel procedures for detecting differential item functioning. *Applied Psychological Measurement, 17*, 105–116.

Roussos, L. A., and Stout, W. F. (1996). Simulation studies of the effects of small sample size and studied item parameters on SIBTEST and Mantel-Haenszel type I error performance. *Journal of Educational Measurement, 33*, 215-230.

Swaminathan, H., and Rogers, H. J. (1990). Detecting differential item functioning using lojistic regression procedures. *Journal of Educational Measurement, 27*, 361–370.

Sweeney, K. P. (1996). A Monte Carlo investigation of the likelihood-ratio procedure in the detection of differential item functioning (Doctoral dissertation, Fordham University, New York, NY).

Stark, S., Chernyshenko, O. S., Drasgow, F. ve Williams, B. A. (2006). Detecting differential item functioning with confirmatory factor analysis and item response theory: toward a unified strategy. *Journal of Applied Psychology, 91*, 1292–1306.

Thissen, D., Steinberg, L. ve Wainer, H. (1993). Detection of differential item functioning using the parameters of item response models. In P. W. Holland ve H. Wainer (Eds.), Differential item functioning (pp. 67–114). Hillsdale, NJ: Lawrence Erlbaum Associates.

Tian, F. (1999). Detecting DIF in Polytomous Item Responses (Doctoral Dissertation, University of Ottawa, Ottawa).

Uttaro, T. (1992). Factors influencing the Mantel-Haenszel procedure in the detection of differential item functioning (Doctoral dissertation, Graduate Center, City University of New York).

Vaughn, B. K. and Wang, Q. (2010). DIF trees: using classifications trees to detect differential item functioning. *Educational and Psychological Measurement, 70*(6) 941–952.

Wang, W.-C., and Su, Y. H. (2004). Factors influencing the Mantel and generalized Mantel-Haenszel methods for the assessment of differential item functioning in polytomous items. *Applied Psychological Measurement, 28*, 450–481.

Wang, W. C. and Yeh, L. Y. (2003). Effects of anchor item methods on differential item functioning detection with the likelihood ratio test. *Applied Psychological Measurement, 27*, 479-498.

Wiberg, M. (2007). Measuring and detecting differential item functioning in criterion-referenced licensing test: a theoretic comparison of methods. Umea University. EM No 60.

Woods, C. M. (2008a). Likelihood-ratio DIF testing: effects of nonnormality. *Applied Psychological Measurement, 32*, 511–526.

Woods, C. M. (2008b). IRT-LR-DIF with estimation of the focal-group density as an empirical histogram. *Educational and Psychological Measurement, 68*, 571–586

Woods, C. M. (2009). Empirical selection of anchors for tests of differential item functioning. *Applied Psychological Measurement, 33*, 42–57.

Woods, C. M. (2011). DIF testing for ordinal items with Poly-SIBTEST, the Mantel and GMH Tests, and IRT-LR-DIF when the latent distribution is nonnormal for both groups. *Applied Psychological Measurement, 35*(2), 145-164.

Zumbo, B. (1999). A handbook on the theory and methods of differential item functioning (DMF): logistic regression modeling as a unitary framework for binary and likert type (ordinal) item scores. Ottowa, ON: Directorate of Human Resources Research and Evaluation, Department of National Defense.

Zwick, R. (1990). When do item response function and Mantel-Haenszel definitions of differential item functioning coincide? *Journal of Educational Statistics, 15,* 185–197.

Zwick, R., Donoghue, J. R., and Grima, A. (1993). Assessment of differential item functioning for performance tasks. *Journal of Educational Measurement, 30*, 233–251.

Zwick, R., Thayer, D. T., & Mazzeo, J. (1997). Descriptive and inferential procedures for assessing differential item functioning in polytomous items. *Applied Measurement in Education, 10*(4), 321-344.

EXTENDED ABSTRACT

Introduction

Ideally, all tests should not have bias. Put differently, a test should be fair to all applicants and should not be biased for any sub-group of a given population. Bias can result in systematic errors that distort inferences one make from a given test's scores. Test bias results when its items contain sources of difficulty that are irrelevant or extraneous to the construct or ability being measured and when these factors affect performance. Recent decades have witnessed an explosion in the number of studies conducted on test bias. Thus, bias detection is an essential part of test validation. As result of practical and theoretical changes in last decades have transformed ways in which test validity is viewed. The concept, method, and processes of validation are vital to evaluating tests since without validation any inferences made from scores of tests would be meaningless. Examination of statistical determination of differential item performance across groups is the first step toward assessment of bias. This statistical procedure is called differential item functioning (DIF). So far, a rich body of research has well-established DIF procedures in dichotomously scored items. However, great deal of work is needed toward examining performance of these procedures under various circumstances. Therefore, the purpose of this study was to compare Type I error rates and power for Mantel Haenszel (MH) and logistic regression (LR) tests for reference and focal groups with varying ability distributions, sample size and sample size ratios.

Method

In this study WinGen3 was used to simulate ability estimates and to generate response data. The two-parameter logistic item response model was used to generate the item parameters for both non-DIF and DIF items. The same item parameters were used for both reference and focal groups. Uniform-DIF was created by altering *b* parameters between the reference and focal groups. Three variables were manipulated: ability distribution differences between groups, sample size and sample size ratio while four factors were fixed: number of groups (reference and focal), number of items, DIF-conditioning (uniform DIF), and matching variable. Taken together, 18 conditions were studied. A 3 (ability distribution) X 6 (sample size combinations) fully-crossed design was used. Each condition was replicated 100 times. Therefore, 1800 sets of 1-0 data were generated.

Results and Discussion

When focal and reference groups were in equal ability distribution condition, MH and LR type I error rates were closer to the nominal α level. On the other hand, when focal and reference groups were in unequal ability distribution condition, MH and LR had inflated

type I error rates. Various studies have reported parallel results. Why MH and LR DIF detection techniques result in inflated type I error can be attributed to the theoretical premises of these techniques. MH and LR rely on observed scores thus measurements errors may lower reliability estimates which in turn results in deviation of true scores from the mean (Zwick, Thayer ve Mazzeo, 1997). Particularly, in short tests, when item responses are generated through IRT models and data analytic procedures relying on observed score, such as MH and LR, are utilized inflated type I error is observed (Meredith & Millsap, 1992; Millsap & Meredith, 1992; Uttaro, 1992; Woods, 2011; Zwick, 1990).

In addition, in the equal ability distribution condition, power of MH and LR tests was lower than 0.64 thus lacking sufficient performance in determining DIF. On the other hand, in the unequal ability distribution condition, these tests resulted in artificially enhanced power rates. Comparing these results with findings of previous studies, it could be concluded that under conditions leading to type I error rate inflation, comparing or assessing power rates of MH and LR could be misleading.

When focal and reference groups were in equal ability distribution condition, MH and LR type I error rates were closer to the nominal α level. In contrast, when these groups were in unequal ability distribution condition, MH and LR had inflated type I error rates. Furthermore, the highest type I error rates were observed when difference between ability distributions was high and sample size was large. Results also showed that conditions leading to type I error rate inflation also result in artificially enhanced power rates.

Based on these results, a couple recommendations could be made for test administrators. One, they can use MH and LR when groups have equal ability distributions. However, when groups have unequal ability distributions use of such DIF detection techniques should be done with caution. DIF detection performance of these techniques under such conditions needs extensive further examination.