Uso dos Métodos Mantel-Haenszel para a Detecção do Funcionamento Diferencial dos Itens e *Software* Relacionado

Using Mantel-Haenszel Methods for Detecting Differential Item Functioning

Ángel M. Fidalgo[*, a] & João D. Scalon[b]
[a]*Universidad de Oviedo, Oviedo, España* & [b]*Universidade Federal de Lavras, Lavras, Brasil*

Resumo
As estatísticas englobadas sob a denominação métodos Mantel-Haenszel (MH), por suas simplicidades, baixo custo computacional e bons resultados, são os métodos mais empregadas para detectar o funcionamento diferencial dos itens (DIF). Os métodos MH podem ser usados para detectar o DIF tanto em itens dicotômicos como em itens politômicos, permitindo comparar dois ou mais grupos simultaneamente, e aplicam-se tanto no âmbito da Teoria Clássica dos Testes (TCT) como da Teoria da Resposta ao Item (TRI). Este artigo fornece uma visão completa e integrada dos métodos MH e apresenta um programa que permite aplicar essas estatísticas no estudo do DIF. O programa é gratuito e está disponível em Espanhol, Inglês e Português.
Palavras-chave: Funcionamento diferencial dos itens, DIF *software*, GMHDIF, métodos de Mantel-Haenszel, estatísticas generalizadas de Mantel-Haenszel.

Abstract
Statistics comprised under Mantel-Haenszel (MH) methods designation constitute one of the most popular and low cost differential item functioning (DIF) detection methods. Mantel-Haenszel methods permit DIF assessment of dichotomous and polytomous items in multiple groups simultaneously, and they can be applied both under Classical Test Theory and Item Response Theory. This paper provides a framework for integrating the different MH statistics used in DIF research, and describes the software that has been developed to provide an easy-to-use program for conducting DIF analyses using the statistics. The program is free of charge and it is available in the following languages: Spanish, English and Portuguese.
Keywords: Differential item functioning, DIF software, GMHDIF, Mantel-Haenszel methods, Generalized Mantel-Haenszel statistics.

A crescente padronização dos métodos de avaliação, o aumento da formação psicométrica e a introdução de novos modelos de medida como a Teoria da Resposta ao Item (TRI), têm levado psicólogos, educadores e pesquisadores brasileiros a publicar um número crescente de revisões (Andriola, 2001; Sisto, 2006a; Valle, 2002) e estudos sobre a detecção do funcionamento diferencial dos itens (*Differential item functioning* [DIF], Andriola, 2000, 2008; Marin Rueda, 2007; Sisto, 2006b; Sisto, Bartholomeu, Angeli dos Santos, Marin Rueda, & Suehiro, 2006; Soares, Gamerman, & Goncalves, 2007; Soares, Genovez, & Galvão, 2005; Traebert, Teline de Lacerda, Thomson, Page, & Locker, 2010). Esta tendência certamente aumentará nos próximos anos. Este artigo tem como meta oferecer aos pesquisadores uma análise integral das estatísticas que compõem o método de referência para a detecção do DIF: a metodologia Mantel-Haenszel (MH). Seus principais objetivos são: (a) fornecer uma visão integral dos métodos MH para a detecção do DIF; (b) Analisar as possibilidades e limitações que esta metodologia tem na análise do DIF; e (c) apresentar um programa de computador que permite avaliar facilmente o DIF com as estatísticas MH. Este artigo é dividido em quatro seções: Funcionamento diferencial do item *versus* impacto e causas do DIF; Métodos Mantel-Haenszel; Possibilidades e limitações; O programa GMHDIF.

[*] Endereço para correspondência: Departamento de Psicología, Universidade de Oviedo, Plaza de Feijoo, s/n, Oviedo, España 33003. E-mail: fidalgo@uniovi.es.
Este trabalho foi financiado pelo Ministério Espanhol de Ciência e Educação (projetos números PR2006-0424, SEJ2006-07491, PCI2006-A7-0553), e tem sido parcialmente escrito na estadia que o primeiro autor realizou no Departamento de Ciências Exatas da Universidade Federal de Lavras, dentro do projeto: "Cooperación internacional entre España, Argentina, Brasil y México para el desarrollo de tecnología y estudios sobre DIF".

Funcionamento Diferencial do Item *Versus* Impacto e Causas do DIF

Diz-se que um item funciona diferencialmente quando a probabilidade de sucesso no item e diferente entre pessoas com o mesmo nível de habilidade na variável medi-

da pelo item, mais que pertencem a diferentes subgrupos de uma população determinada. Por exemplo, se um item de um teste para medir a capacidade espacial estivesse bem construído, todas as pessoas com o mesmo nível de habilidade ou aptidão deveriam ter a mesma chance de acertá-lo. Se, no entanto, para o mesmo nível de aptidão, os homens têm uma probabilidade do sucesso no item mais elevada do que as mulheres, pode-se dizer que o item funciona diferencialmente contra as mulheres. O DIF é, portanto, uma clara ameaça à validade dos itens e do teste.

Do exposto, conclui-se que se um item tem DIF, necessariamente, apresentaria diferentes propriedades estatísticas entre os grupos comparados. No entanto, isso não implica que qualquer teste ou item que mostre diferenças entre os grupos apresente DIF. Deve-se distinguir claramente o termo DIF do termo impacto (*impact*). Suponha que os homens tenham uma maior capacidade espacial do que as mulheres. Se isso fosse verdade, os homens, em média, obteriam maiores escores em testes de habilidade espacial, e também teriam uma probabilidade maior do que as mulheres de acertar os itens desses testes. No entanto, os testes só mostram diferenças reais na habilidade medida. Essas diferenças são chamadas de impacto. Mais formalmente, o impacto, considerando a definição de Ackerman (1992), é a diferença entre os grupos no desempenho em um item causada por uma diferença real na variável medida pelo teste. Se um item apresenta impacto, a probabilidade de respondê-lo corretamente será maior para um grupo do que para outro, refletindo as diferenças entre os grupos na habilidade medida. No entanto, a probabilidade de responder corretamente a este item será a mesma para indivíduos com o mesmo nível de habilidade, independentemente do grupo a que pertençam. Pelo contrário, se a probabilidade de responder corretamente ao item fosse diferente para indivíduos com o mesmo nível de habilidade mais que pertencem a diferentes grupos, então, o item apresenta DIF. O requisito mínimo exigido de qualquer técnica de análise do DIF é distinguir as diferenças reais entre os grupos na variável medida pelo item (impacto) das diferenças espúrias (DIF).

Por que os itens funcionam diferencialmente? A teoria multidimensional do DIF é a explicação mais elegante e consistente teórica e formalmente (Ackerman, 1992; Camilli, 1992; Kok, 1988). Segundo dita teoria, o DIF ocorre quando sob certas condições é violada a suposição de unidimensionalidade do teste. O coração da teoria é a distinção entre a habilidade principal, aquela habilidade que procura medir o teste, e as habilidades espúrias, outras variáveis que não se pretendem medir, mas estão sendo medidas e afetam os resultados do teste. Para simplificar, suponha que temos uma habilidade principal, denotada por Θ, e só uma habilidade espúria, denotada por η. Se um teste contém itens que avaliam dita habilidade espúria, além da habilidade principal, então esses itens podem apresentar DIF. Isso acontecerá se a distribuição condicional da variável espúria difere entre os grupos. Ou seja, se

$$G_1(\eta \mid \Theta) \neq G_2(\eta \mid \Theta) \qquad (1)$$

em que G_i é a distribuição de η para os examinados com valores fixos em Θ, ou seja, a distribuição condicional de η. Portanto, o DIF é causado por diferenças nos parâmetros que definem as distribuições G_1 e G_2. Note-se que o descumprimento da desigualdade (1), implicaria que ainda que o teste fosse multidimensional, a probabilidade de acertar o item seria a mesma para indivíduos com o mesmo nível em Θ, independentemente do grupo a que pertence, ou seja, não haveria DIF. A multidimensionalidade de um item não é *per se* a causa do DIF, se não as diferenças na distribuição condicional das variáveis espúrias (Ackerman, 1992). Um tratamento mais detalhado do tema pode encontrar-se em Fidalgo (1996).

As Estatísticas Mantel-Haenszel

Na literatura há várias estatísticas MH para avaliar o DIF tanto em itens dicotômicos como politômicos. No caso de itens dicotômicos foram Holland e Thayer (1988) quem propuseram analisar o DIF usando a estatística qui-quadrado de Mantel-Haenszel (χ^2_{MH}), desenvolvida por Mantel e Haenszel (1959). Também foi formulada uma abordagem bayesiana ao procedimento MH para itens dicotômicos (Zwick, Thayer, & Lewis, 1999, 2000), mas essas estatísticas não demonstraram nenhuma vantagem adicional em relação à estatística χ^2_{MH} (Fidalgo, Hashimoto, Bartram, & Muñiz, 2007). No caso de itens politômicos, as estatísticas com base no trabalho original de Mantel Haenszel também têm sido utilizadas para a detecção do DIF: o teste generalizado de Mantel-Haenszel (GMH; Mantel & Haenszel, 1959; Zwick, Donoghue, & Grima, 1993) e o teste de Mantel (Mantel, 1963; Zwick et al., 1993). A estatística GMH trata as categorias de resposta do item como uma variável nominal, enquanto o teste de Mantel considera a natureza ordinal das categorias de resposta. No caso da estatística GMH, a hipótese alternativa (H_1) especifica que a distribuição das respostas ao item difere entre os grupos comparados. Por outro lado, o teste de Mantel, ao considerar a natureza ordinal das categorias do item, especifica que a media das respostas difere através da variável de agrupamento. Por tanto, o teste de Mantel pode ser aplicado a itens politômicos com categorias ordenadas. Uma limitação de todas as estatísticas anteriormente expostas é que só permitem a análise do DIF em dois grupos. Felizmente, existem alternativas melhores, embora não sejam suficientemente conhecidas.

Recentemente, Fidalgo e Madeira (2008) formularam um marco unificado para a análise do DIF usando a esta-

tística generalizada de Mantel-Haenszel proposta por Landis, Heyman e Koch (1978). Fidalgo e Madeira (2008) afirmam que dita estatística engloba a estatística GMH e o teste de Mantel, além do mais da estatística χ^2_{MH}. Portanto, pode-se aplicá-la para avaliar o DIF em vários grupos, tanto para itens dicotômicos como para itens politômicos (Fidalgo & Scalon, 2010).

Embora seja possível ignorar a estatística χ^2_{MH}, pois é um caso especial da estatística generalizada de MH, apresentaremos esta estatística por razões pedagógicas: é muito mais fácil apreciar a lógica do procedimento MH no caso mais simples do que na formulação matricial.

Estatística χ^2_{MH}

Conforme o mencionado na primeira seção, os métodos de detecção de DIF deverão estabelecer as comparações entre os grupos, empregando indivíduos com o mesmo nível de competência, na medida em que não queiram confundir o DIF com o impacto. Os métodos MH comumente utilizam o escore total (a soma das pontuações dos itens no teste) como um estimador da variável que pretende medir o teste (Θ). Assim, a escore total será a variável de estratificação que servirá para estabelecer as comparações necessárias entre os grupos.

A primeira coisa que devemos fazer para implementar o procedimento MH é dispor as respostas dos examinados no teste em Q tabelas de contingência de 2 x 2, onde Q é o número de intervalos em que a escore total é dividida ($1 \ldots h \ldots \ldots Q$). Assim, para cada nível de pontuação h, temos uma tabela de contingência 2 x 2, com os membros do grupo (focal / referência) em uma das entradas e a resposta ao item (sucesso/erro) na outra (Tabela 1). Os valores das celas A_h, B_h, C_h e D_h denotam o número de examinados em cada categoria. Os valores marginais N_{Rh} e N_{Fh} representam o número de examinados no grupo de referência e focal, respectivamente, enquanto N_{1h} e N_{0h} representam o número de examinados que responderam ao item corretamente e incorretamente, respectivamente. Finalmente, N_h é o número total de examinados ao nível de pontuação h.

Tabela 1
Tabela de Contingência 2 x 2 para o Nivel de Pontuação h

Grupo	Acertos (1)	Erros (0)	Total
Referência	A_h	B_h	N_{Rh}
Focal	C_h	D_h	N_{Fh}
	N_{1h}	N_{0h}	N_h

A lógica por trás do procedimento MH é a seguinte: se o item não apresenta DIF, a razão entre o número de pessoas que acertam o item e aquelas que o erram deve ser a mesma nos dois grupos comparados em todos os níveis de pontuação. Formalmente

$$H_0 : (A_h / B_h) = \alpha\ (C_h / D_h)$$

sendo $\alpha = 1$ para todos os h (não DIF)

$$H_1 : (A_h / B_h) = \alpha\ (C_h / D_h)$$

sendo $\alpha \neq 1$ em algum h (DIF) .

Holland e Thayer (1988) propuseram utilizar a estatística χ^2_{MH} para testar a hipótese nula de ausência de DIF. Esta estatística é dada por:

$$\chi^2_{MH} = \frac{\left(\left| \sum_{h=1}^{Q} A_h - \sum_{h=1}^{Q} E(A_h) \right| - 0.5 \right)^2}{\sum_{h=1}^{Q} Var(A_h)} \qquad \textbf{(2)}$$

em que $E\ (A_h)$ é o valor esperado de A_h , $Var\ (A_h)$ é a sua variância, que são iguais a:

$$E\ (A_h) = (N_{Rh}\ N_{1h}) / N_h$$

e

$$Var(A_h) = \frac{N_{Rh} N_{Fh} N_{1h} N_{0h}}{N_h^2 (N_h - 1)}$$

A estatística χ^2_{MH} segue uma distribuição χ^2 com um grau de liberdade. Se $\chi^2_{MH} > \alpha \chi^2_1$, então o item estudado mostra DIF com um nível de confiança 1-α.

Uma ampla descrição da estatística χ^2_{MH} pode ser encontrada na entrada que a *Encyclopedia of Statistics in Behavioural Science* dedica aos métodos Mantel-Haenszel (Fidalgo, 2005a). Embora os cálculos necessários para obter a estatística sejam muito simples, quem quiser poupar-se dos incômodos pode solicitar, ao primeiro autor do artigo, uma cópia do programa MHDIF (Fidalgo, 1994).

Estatística Generalizada de MH

Em 1978 Landis et al. propuseram uma estatística generalizada de MH para a análise de tabelas de contingência de dimensão $Q: R \times C$. A estrutura dos dados para esta tabela de contingência geral é mostrada na Tabela 2.

Tabela 2
Tabela de Contingência R x C no h-ésimo Estrato

Níveis do Fator	Categorias da variável de resposta						
	1	2	.	*j*	.	*C*	Total
1	n_{h11}	n_{h12}	.	n_{h1j}	.	n_{h1C}	$N_{h1\cdot}$
2	n_{h21}	n_{h22}	.	n_{h2j}	.	n_{h2C}	$N_{h2\cdot}$
⋮	⋮	⋮	.	⋮	.	⋮	⋮
i	n_{hi1}	n_{hi2}	.	n_{hij}	.	n_{hiC}	$N_{hi\cdot}$
⋮	⋮	⋮	.	⋮	.	⋮	⋮
R	n_{hR1}	n_{hR2}	.	n_{hRj}	.	n_{hRC}	$N_{hR\cdot}$
Total	$N_{h\cdot 1}$	$N_{h\cdot 2}$	.	$N_{h\cdot j}$	·	$N_{h\cdot C}$	$N_{h\cdot\cdot}$

O teste generalizado de Mantel-Haenszel para testar a hipótese nula (H_0) de associação entre o fator (os grupos) e a variável de resposta (as categorias de resposta do item), controlando o efeito da covariável (o nível de competência estimado pelo escore total), é definido em termos de matrizes por Landis et al. (1978) como:

$$Q_{GMH} = \left\{\sum_{h=1}^{Q}(\mathbf{n}_h - \mathbf{m}_h)'\mathbf{A}_h'\right\}\left\{\sum_{h=1}^{Q}\mathbf{A}_h\mathbf{V}_h\mathbf{A}_h'\right\}^{-1}\left\{\sum_{h=1}^{Q}\mathbf{A}_h(\mathbf{n}_h - \mathbf{m}_h)\right\} \quad (3)$$

Onde $\mathbf{n}_h$, $\mathbf{m}_h$, $\mathbf{V}_h$ e $\mathbf{A}_h$ são, respectivamente, o vetor de freqüências, o vetor de freqüências esperadas, a matriz de covariâncias, e uma matriz de funções lineares definidas em conformidade com a hipótese alternativa (H_1) de interesse. A hipótese nula (H_0) de não-associação será testada contra diferentes H_1 que serão funções da escala de medida do fator e da variável de resposta. Assim, teremos uma série de estatísticas que servirão para detectar a associação geral (ambas variáveis são nominais), as diferenças médias (o fator é uma variável nominal e a resposta e uma variável ordinal), e a correlação linear (ambas as variáveis são ordinais). Nós descreveremos apenas as duas primeiras estatísticas por ser ate hoje as únicas que tem sido utilizada para detectar o DIF. A partir da Tabela 2, estes vetores e matrizes são definidos como:

$$\mathbf{n}_h = (n_{h11,}\, n_{h21,}\, \ldots,\, n_{hRC})' \quad (CR \text{ x } 1),$$

$$\mathbf{m}_h = N_{h\cdot\cdot}(\mathbf{p}_{h\cdot *} \otimes \mathbf{p}_{h*\cdot}) \quad (CR \text{ x } 1),$$

$$\mathbf{V}_h = N_{h\cdot\cdot}^2/(N_{h\cdot\cdot}-1)\{(\mathbf{D}_{p_{h\cdot *}} - \mathbf{p}_{h\cdot *}\mathbf{p}_{h\cdot *}')\otimes(\mathbf{D}_{p_{h*\cdot}} - \mathbf{p}_{h*\cdot}\mathbf{p}_{h*\cdot}')\} \quad (CR \text{ x } CR),$$

em que $\mathbf{p}_{h\cdot *}$ e $\mathbf{p}_{h*\cdot}$ são, respectivamente, vetores de dimensões (C x 1) e (R x 1) com as proporções marginais das colunas ($p_{h\cdot j} = N_{h\cdot j}/N_{h\cdot\cdot}$) e as proporções marginais das linhas ($p_{hi\cdot} = N_{hi\cdot}/N_{h\cdot\cdot}$), denotando ⊗ o produto de Kronecker, $\mathbf{D}_{p_{h\cdot *}}$ é uma matriz diagonal com elementos do vetor $\mathbf{p}_{h\cdot *}$ em sua diagonal principal, e $\mathbf{D}_{p_{h*\cdot}}$ é uma matriz diagonal com elementos do vetor $\mathbf{p}_{h*\cdot}$ em sua diagonal principal.

Como foi assinalado anteriormente, a equação 3 será resolvida através da definição da matriz $\mathbf{A}_h$ ($\mathbf{A}_h = \mathbf{C}_h \otimes \mathbf{R}_h$), utilizando uma estatística diferente para a detecção de cada H_1. Resumidamente, estas são:

Q $_{GMH(1)}$ ou a *estatística generalizada nominal de MH*. Quando a variável linha e a variável coluna são nominais, a H_1 especifica que a distribuição da variável resposta difere entre os diferentes níveis do fator. Aqui, $\boldsymbol{R}_h = [\boldsymbol{I}_{R\text{-}1}, -\boldsymbol{J}_{R\text{-}1}]$ e $\boldsymbol{C}_h = [\boldsymbol{I}_{C\text{-}1}, -\boldsymbol{J}_{C\text{-}1}]$, onde $\boldsymbol{I}_{R\text{-}1}$ é uma matriz de identidade, e $\boldsymbol{J}_{R\text{-}1}$ é um vetor de uns. Assim, a dimensão de $\boldsymbol{R}_h$ será (R-1 x R). Da mesma forma, $\boldsymbol{I}_{C\text{-}1}$ é uma matriz de identidade, e $\boldsymbol{J}_{C\text{-}1}$ é um vetor de uns. Sob H_0, $Q_{GMH(1)}$

segue aproximadamente uma distribuição qui-quadrado com graus de liberdade (gl) = (R-1) (C-1). Quando R = C = 2, $Q_{GMH(1)}$ é idêntica a estatística χ^2_{MH}, com exceção da falta da correção de continuidade. Para o caso especial de dois níveis do fator, $Q_{GMH(1)}$ é idêntica a estatística generalizada proposta por Mantel e Haenszel (1959).

$\boldsymbol{Q_{GMH(2)}}$ ou *estatística generalizada ordinal de MH*. Aqui, a hipótese H_1 estabelece que a média das respostas difere entre os níveis do fator, sendo $\boldsymbol{R}_h$ a mesma matriz que foi utilizada no caso anterior e $\boldsymbol{C}_h = (c_{h1}, \ldots, c_{hC})$, sendo um vetor de dimensões (1 x C), em que c_{hj} é uma pontuação que reflete adequadamente a natureza ordinal da categoria da j-ésima resposta no h-ésimo estrato. Na literatura sobre o DIF, os inteiros são a opção mais comum, embora que a seleção dos valores da $\boldsymbol{C}_h$ admita outras possibilidades (Fidalgo & Bartam, 2010). Sob H_0, $Q_{GMH(2)}$ tem aproximadamente uma distribuição qui-quadrado com gl = (R-1). Para o caso especial de dois níveis do fator, $Q_{GMH(2)}$ é idêntico ao teste proposto por Mantel (1963).

Obviamente, quando $C = R = 2$, $Q_{GMH(1)} = Q_{GMH(2)} = \chi^2_{MH}$ (para que esta equivalência seja cumprida, χ^2_{MH} tem de ser calculado sem a correção de continuidade que normalmente inclui).

A diferença entre a aplicação da estatística apresentada na equação 2, que inclui a correção de continuidade, e da estatística generalizada, onde não está incluída, é uma maior potência para detectar o DIF, e um ligeiro aumento na taxa de erro Tipo I (identificar itens que não funcionam diferencialmente como se tivessem DIF), para a última estatística. Estas diferenças ocorrem com tamanhos de amostra pequenos (50 examinados por grupo). Com uma amostra de 500 examinados, as diferenças entre as duas estatísticas são praticamente nulas (Fidalgo et al., 2007). O leitor pode encontrar maiores informações sobre as estatísticas generalizadas de MH e exemplos de seu cálculo em Fidalgo (2005a), Fidalgo e Madeira (2008) e Fidalgo e Scalon (2010).

Possibilidades e Limites dos Métodos MH

Na hora de avaliar o DIF utilizando alguma das estatísticas MH, tem-se que conhecer as vantagens e as limitações que apresentam. A seguir enumeram-se as principais considerações que se deve ter em mente quando se utilizam essas estatísticas na detecção do DIF:

1. Podem ser usadas para detectar o DIF tanto nos testes construídos ou analisados desde a perspectiva da TCT como da TRI. Deve ser conhecido, no entanto, que essas estatísticas se comportam melhor quando os itens do teste se ajustam a família de modelos de Rasch, como o modelo de um parâmetro logístico em itens dicotômicos, ou o modelo de crédito parcial em itens politômicos. Ainda assim, os estudos de simulação mostram as estatísticas MH eficazes em uma ampla variedade de situações, embora os dados não sejam conformes ao modelo de Rasch (para citar apenas alguns estudos de simulação: Roussos & Stout, 1996; Uttaro & Millsap, 1994).
2. Podem ser usadas para detectar o DIF tanto em itens dicotômicos como itens politômicos nominais e politômicos ordinais (Fidalgo, Quintanilla, Fernandez, Pons, & Aguerri, 2010), e tanto em dois grupos como simultaneamente em vários grupos (Fidalgo & Scalon, 2010).
3. Podem ser usadas com tamanhos de amostras muito pequenos. No caso de itens dicotômicos, a maioria dos autores recomenda tamanhos de amostra em torno de 200 pessoas por grupo (Mazor, Clauser, & Hambleton, 1992), mas pode ser útil com amostras tão baixas quanto 50 pessoas por grupo, desde que sejam utilizados níveis de significância mais elevados (Fidalgo, Ferreres, & Muñiz, 2004; Fidalgo et al., 2007).
4. Podem ser usadas para detectar uma grande variedade de tipos de DIF. Além de ser uma das melhores estatísticas para detectar o DIF uniforme (um grupo tem vantagem sobre outro ao largo de todo o nível de habilidade), vários estudos de simulação mostraram que eles também podem detectar o DIF não-uniforme (um grupo têm vantagem em alguns níveis de habilidade, e desvantagem em outros – Fidalgo, Mellenbergh, & Muñiz, 1998; Hidalgo & López-Pina, 2004; Mazor, Clauser, & Hambleton, 1994; Rogers & Swaminathan, 1993). Deve-se notar, contudo, que no caso de itens dicotômicos a detecção do DIF não uniforme requer a modificação do procedimento MH proposta por Mazor et al., (1994), e que implementa o programa MHDIF (Fidalgo, 1994). Aplicando o referido programa Hidalgo e López-Pina (2004) encontraram taxas de detecção do DIF não-uniforme semelhante as obtidas por meio da regressão logística. No caso de itens politômicos, a estatística GMH tem boas taxas de detecção do DIF não-uniforme. No entanto, o teste de Mantel apresenta uma potência muito baixa para detectar este tipo de DIF (Kristjansson, Aylesworth, Mcdowell, & Zumbo, 2005). Finalmente, baseando-se em um amplo estudo de simulação, Fidalgo e Bartram (2010) recomendam aplicar $Q_{GMH(1)}$ no caso de aplicar uma única estatística MH, já que tem maior potência que $Q_{GMH(2)}$ para detectar a maioria dos padrões de DIF.
5. Ao utilizar o escore total no teste como variável de estratificação, as estatísticas devem ser aplicadas em duas etapas, ou de forma iterativa, para evitar que os itens com DIF contaminem a referida variável (Fidalgo, Mellenbergh, & Muñiz, 2000; Miller & Oshima, 1992; Wang & Su, 2004a). Além disso, o item em questão deve ser incluso sempre no cômputo da pontuação total, mesmo que tenha sido identificado com DIF na primeira etapa (Zwick et al., 1993).
6. A presença de diferentes distribuições entre os grupos na variável medida pelo teste, ou seja, aquilo que

chamamos de impacto, implica um aumento na taxa de erro Tipo I. Especialmente quando o procedimento é aplicado para itens que não seguem a família de modelos de Rasch (Penny & Johnson, 1999; Su & Wang, 2005; Wang & Su, 2004b).

7. Deve-se notar que a aplicação do $Q_{GMH(2)}$, ou do teste de Mantel, sempre exige a escolha do sistema de pontuação que melhor represente as categorias de resposta do item, e que escolher um ou outro afetam a potência para detectar diferentes padrões de DIF (Fidalgo & Bartam, 2010). Na ausência de critérios baseados nas características dos itens, e até que não sejam formulados claros critérios estatísticos, podem empregar-se os habituais números inteiros.
8. Além de verificar a significância estatística, mediante algumas das estatísticas apresentadas, os estudos sobre DIF devem ser sempre completados com as estimativas da magnitude de DIF que tem os itens. Quando se tem itens dicotômicos e dois grupos, Mantel e Haenszel (1959) propuseram o bem conhecido estimador da razão de chances comum (*common odds ratio*, $\hat{\alpha}_{MH}$ – o leitor pode encontrar uma descrição desta estatística em Andriola, 2001). Também no caso de ter somente dois grupos, existem generalizações para itens politômicos, sendo a mais recomendável a estatística formulada por Liu e Agresti (1996) (ver Penfield & Algina, 2003, para sua aplicação nos estudos de DIF).
9. O emprego de vários procedimentos para avaliar o DIF acarreta um aumento na taxa de erro Tipo I (identificar itens que não funcionam diferencialmente como se tivessem DIF), ou na taxa de erro Tipo II (não identificar itens que apresentam DIF – Fidalgo et al., 2004). Em Fidalgo e Ferreres (2002) o leitor encontrará uma análise dos custos que, em termos de erro de Tipo I e Tipo II, tem algumas das decisões mais freqüentemente tomadas nos estudos empíricos do DIF e que afetam a os métodos MH: o uso de provas de avaliação do DIF unidimensionais em contextos multidimensionais, a avaliação do DIF numa primeira e única etapa, a ausência de medidas do tamanho de efeito, a escolha dos níveis de significância convencionais, e o uso de diversos procedimentos de avaliação do DIF, entre outras.
10. Finalmente, a Figura 1 mostra um diagrama para determinar, em função das características dos itens e das variáveis relacionadas com o estudo do DIF, que tipo de estatísticas MH devem ser utilizadas.

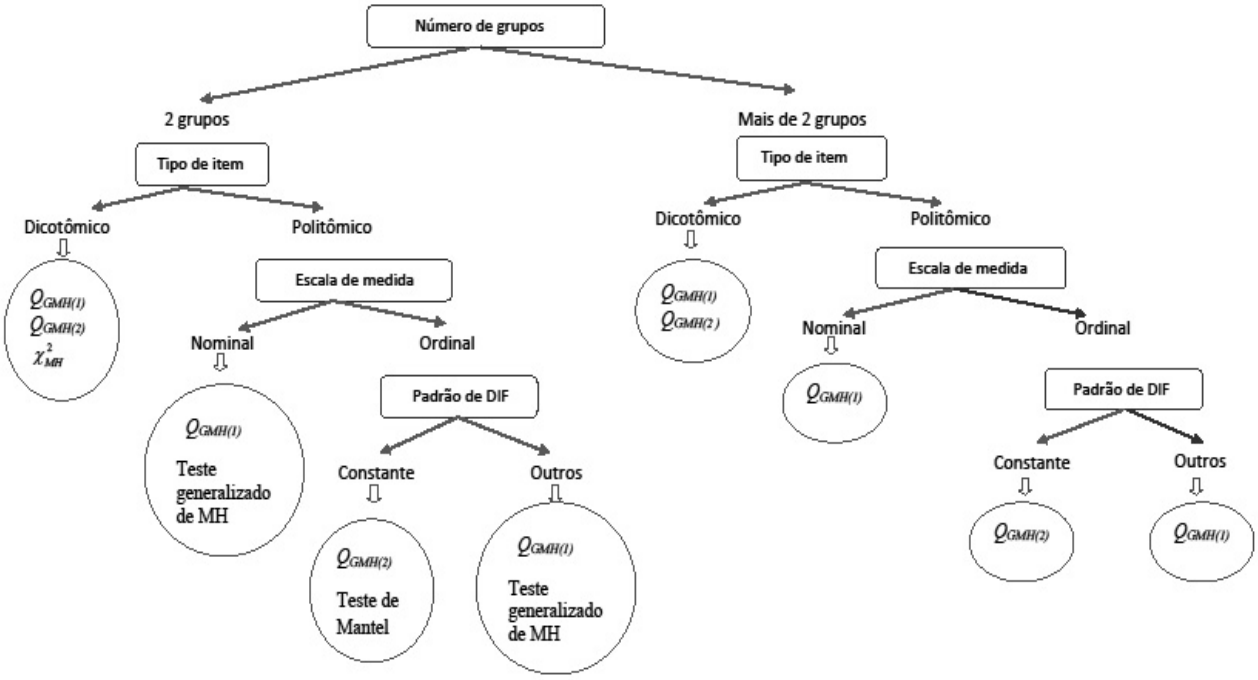

Figura 1. Esquema com os tipos de estatísticas MH que podem ser utilizadas para a detecção do DIF com base nas características do item e dos objetivos da analise do DIF. Quando nos círculos aparece mais de uma estatística o resultado de aplicar uma ou outra é equivalente (Fidalgo, 2010).

O Programa GMHDIF

O programa GMHDIF (Fidalgo, 2011) é um programa baseado em Windows que foi desenvolvido para proporcionar um *software* amigável que permita conduzir análises do DIF a usuários não expertos. O programa permite, através de um simples teste de significância, avaliações simultâneas do DIF em diversos grupos utilizando as es-

tatísticas generalizadas de Mantel-Haenszel, sendo aplicado tanto para itens dicotômicos como para itens politômicos. Além disso, o programa executa análises do DIF em duas etapas, e para os itens identificados com DIF realiza comparações entre os grupos, dois a dois, empregando a correção de Bonferroni para um determinado nível de significância (nível de significância / número de comparações pareadas).

Para realizar análises do DIF, utilizando com o programa GMHDIF, basta seguir os seguintes passos:

1. Importe os dados para serem analisados. Arquivos importados devem ser separados por espaço, tab, vírgula ou ponto e vírgula.
2. Forneça informações sobre as seguintes variáveis: (a) Itens que serão submetidos a uma análise do DIF; (b) Itens que serão usados como variável de estratificação; (c) Variável de agrupamento.
3. Selecione a estatística generalizada de MH desejada: $Q_{GMH(1)}$ ou $Q_{GMH(2)}$.
4. Explore os resultados das análises do DIF. Caso queira, os resultados podem ser salvos como arquivos de texto ou rtf.

A Figura 2 mostra os resultados de uma análise do DIF empregando quatro grupos. Como pode-se ver na figura, o primeiro item (variável 2) não apresenta DIF ao nível de significância de 0,05 em nenhuma das etapas. O segundo item (variável 3) mostra DIF em ambas etapas, tornando-se necessário determinar entre quais grupos o DIF ocorre. Neste caso, pode-se tomar a decisão de comparar todos os grupos entre eles. As comparações pareadas mostram que o DIF existe entre o grupo 2 e todos os outros. Atualmente, está em desenvolvimento uma nova versão do programa que inclui medidas do tamanho do efeito (*effect size*).

GMHDIF

Estatística generalizada de MH = QMH2

Nível de significância = 0,05
Os resultados estatisticamente significativos ao nível de 0,05 são marcados com um asterisco.

VARIÁVEL 2

Etapa 1: QMH = 0,8620 *gl* = 3 *p* = 0,8346
Etapa 2: QMH = 2,0241 *gl* = 3 *p* = 0,5674

VARIÁVEL 3

Etapa 1: QMH = 18,6560 *gl* = 3 *p* = 0,0003*****
Etapa 2: QMH = 20,5184 *gl* = 3 *p* = 0,0001*****

COMPARAÇÕES MÚLTIPAS PAREADAS

As comparações pareadas (CP) estatisticamente significativas são marcadas com um asterisco.
O nível de significância está corrigido pela equação de Bonferroni (0,05/ 6= 0,0083).

CP entre o grupo 1 e o grupo 2 : QMH = 19,8531 gl = 1 *p* = 0,0000*****
CP entre o grupo 1 e o grupo 3 : QMH = 0,5497 gl = 1 *p* = 0,4584
CP entre o grupo 1 e o grupo 4 : QMH = 0,0085 gl = 1 *p* = 0,9264
CP entre o grupo 2 e o grupo 3 : QMH = 8,7135 gl = 1 *p* = 0,0032*****
CP entre o grupo 2 e o grupo 4 : QMH = 10,8825 gl = 1 *p* = 0,0010*****
CP entre o grupo 3 e o grupo 4 : QMH = 0,0048 gl = 1 *p* = 0,9445

Figura 2. Saida com os resultados de uma análise do DIF em duas etapas na que se comparavam quatro grupos.

Disponibilidade

O programa GMHDIF, o manual do usuário, e exemplos com arquivos de dados podem ser obtidos diretamente com o Dr. Angel M. Fidalgo no seguinte e-mail: fidalgo@uniovi.es. O programa e a documentação relacionada estão disponíveis nos seguintes idiomas: Espanhol, Inglês e Português. O uso do programa está limitado ao âmbito acadêmico e a outras aplicações sem fins lucrativos.

Conclusão

A enorme versatilidade dos métodos MH fez deles uma das metodologias de referência na avaliação do funcionamento diferencial dos itens dicotômicos e politômicos.

Como tem sido apresentado, são estatísticas não paramétricas que podem ser aplicadas com tamanhos de amostra pequenas, que permitem detectar um grande número de tipos de DIF, que permitem a avaliação simultânea do DIF em vários grupos, e que, alem dos testes de significância apresentados, dispoem também de estatísticas para avaliar o tamanho do efeito. Além do acima exposto, o comportamento destas estatísticas está bem estabelecido em uma ampla variedade de situações, pela grande quantidade de estudos teóricos e de simulação que tem sido feitos; e contam com *software* especialmente planejado que facilita sua aplicação pelos pesquisadores aplicados (Fidalgo, 1994, 2011; Penfield, 2005). Assim, pode-se concluir que, embora existam muitas outras alternativas para detectar o DIF, especialmente na TRI (Andriola, 2001; Fidalgo, 1996, 2005b), no momento, os métodos MH seguem sendo o padrão-ouro para avaliar o DIF.

Referências

Ackerman, T. A. (1992). A didactic explanation of item bias, item impact, and item validity from a multidimensional perspective. *Journal of Educational Measurement, 29*, 67-91.

Andriola, W. B. (2000). Funcionamento diferencial dos itens (DIF): Estudo com analogias para medir o raciocínio verbal. *Psicologia: Reflexão e Crítica, 13*, 475-483.

Andriola, W. B. (2001). Descrição dos principais métodos para detectar o funcionamento diferencial dos itens (DIF). *Psicologia: Reflexão e Crítica, 14*, 643-652.

Andriola, W. B. (2008). Uso da Teoria de Resposta ao Item (TRI) para analisar a eqüidade do processo de avaliação do aprendizado discente. *Revista Iberoamaericana de Evaluación Educativa, 1*, 171-189.

Camilli, G. (1992). A conceptual analysis of differential item functioning in terms of a multidimensional item response model. *Applied Psychological Measurement*, *16*, 129-147.

Fidalgo, A. M. (1994). MHDIF: A computer program for detecting uniform and nonuniform differential item functioning with the Mantel-Haenszel procedure. *Applied Psychological Measurement, 18*, 300.

Fidalgo, A. M. (1996). Funcionamiento diferencial de los ítems. In J. Muñiz (Ed.), *Psicometría* (pp. 371-455). Madrid, España: Universitas.

Fidalgo, A. M. (2005a). Mantel-Haenszel Methods. In B. S. Everitt & D. C. Howell (Eds.), *Encyclopedia of Statistics in Behavioral Science* (Vol. 3, pp. 1120-1126). Chichester, UK: John Wiley & Sons.

Fidalgo, A. M. (2005b). Enfoque de la Teoría de Respuesta a los Ítems. In J. Muñiz, A. M. Fidalgo, M. A. García-Cueto, R. Martínez, & R. Moreno (Eds.), *Análisis de los ítems* (pp. 79-131). Madrid, España: La Muralla.

Fidalgo, A. M. (2010). GMHDIF: *Manual do usuário* [Manual *software*]. Oviedo, España.

Fidalgo, A. M. (2011). GMHDIF: A computer program for detecting DIF in dichotomous and polytomous items using generalized Mantel-Haenszel Statistics. *Applied Psychological Measurement, 35*, 247-249. doi: 10.1177/0146621610375691

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*, 600-606. doi: 10.1177/0146621610378405

Fidalgo, A. M., & Ferreres, D. (2002). Supuestos y consideraciones en los estudios empíricos sobre el funcionamiento diferencial de los items. *Psicothema, 14,* 491-496.

Fidalgo, A. M., Ferreres, D., & Muñiz, J. (2004). DIF detection using several statistical procedures: Implications on the type I and type II error rate. *The Journal of Experimental Education, 73*, 23-39.

Fidalgo, A. M., Hashimoto, K., Bartram, D., & Muñiz, J. (2007). Application of an empirical Bayes enhancement of the Mantel-Haenszel procedure for detecting DIF under small-sample conditions. *The Journal of Experimental Education, 75*(4), 293-314.

Fidalgo, A. M., & Madeira, J. M. (2008). Generalized Mantel-Haenszel methods for DIF detection. *Educational and Psychological Measurement, 68,* 940-958.

Fidalgo, A. M., Mellenbergh, G. J., & Muñiz, J. (1998). Comparación del procedimiento Mantel-Haenszel frente a los modelos loglineales en la detección del funcionamiento diferencial de los ítems. *Psicothema*, *10*, 209-218.

Fidalgo, A. M., Mellenbergh, G. J., & Muñiz, J. (2000). Effects of amount of DIF, test length, and purification type on robustness and power of Mantel-Haenszel procedures. *Methods of Psychological Research, 5*, 43-53. Retrieved from http://www.psychologie.de/fachgruppen/methoden/mpr-online/issue11/art3/fidalgo.pdf

Fidalgo, A. M., Quintanilla, L., Fernández, R., Pons, F., & Aguerri, M. E. (2010). Detección del DIF en ítems politómicos mediante el uso de los métodos Mantel-Haenszel. *Revista Española de Metodología Aplicada, 15*, 12-18. Retrieved from http://www.psico.uniovi.es/REMA/v15n1/indice.html

Fidalgo, A. M., & Scalon, J. D. (2010). Using Generalized Mantel-Haenszel Statistics to Assess DIF among Multiple Groups. *Journal of Psychoeducational Assessment, 28,* 60-69. doi: 10.1177/0734282909337302

Hidalgo, M. D., & López-Pina J. A. (2004). Differential Item Functioning Detection and Effect Size: A comparison between logistic regression and Mantel-Haenszel Procedures. *Educational and Psychological Measurement, 64*, 903-915.

Holland, W. P., & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel Procedure. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 129-145). Hillsdale, NJ: LEA.

Kok, F. G. (1988). Item bias and test multidimensionality. In R. Langeheine & J. Rost (Eds.), *Latent trait and latent class models* (pp. 263-274). New York: Plenum.

Kristjansson, E., Aylesworth, R., Mcdowell, I., & Zumbo, B. D. (2005). A comparison of Four Methods for Detecting Differential Item Functioning in Ordered Response Items. *Educational and Psychological Measurement, 65*, 935-953.

Landis, J. R., Heyman, E. R., & Koch, G. G. (1978). Average partial association in three-way contingency tables: A review and discussion of alternative tests. *International Statistical Review, 46,* 237-254.

Liu, I.-M., & Agresti, A. (1996). Mantel-Haenszel-type inference for cumulative odds ratios with a stratified ordinal response. *Biometrics, 52*, 1223-1234.

Mantel, N. (1963). Chi-square tests with one degree of freedom; extension of the Mantel-Haenszel procedure. *Journal of the American Statistical Association, 58*, 690-700.

Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute, 22*, 719-748.

Marin Rueda, F. J. (2007). O funcionamento diferencial do item no teste pictórico de memória. *Avaliação Psicológica, 6,* 229-237.

Mazor, K. M., Clauser, B. E., & Hambleton, R. K. (1992). The effect of sample size on the functioning of the Mantel-Haenszel Statistic. *Educational and Psychological Measurement, 52,* 443-452.

Mazor, K. M., Clauser, B. E., & 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.

Miller, M. D., & Oshima, T. C. (1992). Effect of sample size, number of biased items, and magnitude of bias on a two-stage item bias estimation method. *Applied Psychological Measurement*, *16*, 381-388.

Penfield, R. D. (2005). DIFAS: Differential Item Functioning Analysis System. Computer Program Exchange. *Applied Psychological Measurement, 29*, 150-151.

Penfield, R. D., & Algina, J. (2003). Applying the Liu-Agresti estimator of the cumulative common odds ratio to DIF detection in polytomous items. *Journal of Educational Measurement, 40*(4), 353-370.

Penny, J., & Johnson, R. L. (1999). How group differences in matching criterion distribution and IRT item difficulty can influence the magnitude of the Mantel-Haenszel chi-square DIF index. *The Journal of Experimental Education, 67*, 343-366.

Rogers, H. J., & 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., & Stout, W. F. (1996). Simulations 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.

Sisto, F. F. (2006a). O funcionamento diferencial dos itens. *Psico-USF, 11, 35-43.*

Sisto, F. F. (2006b). Estudo do funcionamento diferencial de itens para avaliar o reconhecimento de palavras. *Avaliação Psicológica, 5,* 1-10.

Sisto, F. F., Bartholomeu, B., Angeli dos Santos, A. A., Marin Rueda, F. J., & Suehiro, A. C. B. (2006). Funcionamento diferencial de itens para avaliar a agressividade de universitários. *Psicologia: Reflexão e Crítica, 21*, 474-481.

Soares, T. M., Gamerman, D., & Goncalves, F. B. (2007). Análise bayesiana do funcionamento diferencial do item. *Pesquisa Operativa, 27,* 271-291.

Soares, T. M., Genovez, S. F., & Galvão, A. F. (2005). Análise do Comportamento Diferencial dos Itens de Geografia: Estudo da 4ª série avaliada no PROEB/SIMAVE 2001. *Avaliação Educacional*, *16*, 81-110

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 in polytomous items. *Applied Measurement in Education, 18*, 313-350.

Traebert, J., Telino de Lacerda, J., Thomson, W. M., Page, L. F., & Locker, D. (2010). Differential item functioning in a Brazilian-Portuguese version of the Child Perceptions Questionnaire (CPQ_{11-14}). *Community Dentistry and Oral Epidemiology, 38*, 129-135.

Uttaro, T., & Millsap, R. E. (1994). Factors influencing the Mantel-Haenszel procedure in the detection of differential item functioning. *Applied Psychological Measurement, 18*, 15-25.

Valle, R. C. (2002). Comportamento Diferencial do Item: Uma apresentação. *Estudos em Avaliação Educacional*, 25, 3-21.

Wang, W.-C., & Su, Y.-H. (2004a). Effect of average signed area between two item characteristic curves and test purification procedures on the DIF detection via the Mantel-Haenszel Method. *Applied Measurement in Education, 17*, 113-144.

Wang, W.-C., & Su, Y.-H. (2004b). Factors influencing the Mantel and Generalized Mantel-Haenszel Methods for the assessment of differential item functioning in polytomous items. *Applied Psychological Measurement, 28*, 450-480.

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

Zwick, R., Thayer, D. T., & Lewis, C. (1999). An empirical Bayes approach to Mantel-Haenszel DIF analysis. *Journal of Educational Measurement, 36*, 1-28.

Zwick, R., Thayer, D. T., & Lewis, C. (2000). Using loss functions for DIF detection: An empirical Bayes approach. *Journal of Educational and Behavioral Statistics*, *25*, 225-247.

Recebido: 14/06/2010
Aceite final: 09/12/2010