Descrição dos Principais Métodos para Detectar o Funcionamento Diferencial dos Itens (DIF)

Wagner Bandeira Andriola [1]
Universidade Federal do Ceará

Resumo

O artigo descreve os principais métodos utilizados, atualmente, na detecção do funcionamento diferencial dos itens (DIF), entre os quais o de Comparação da Área entre as CCI's, Comparação das Probabilidades, Comparação dos Parâmetros dos Itens, Qui-quadrado de Lord, Qui-quadrado de Scheuneman, Qui-quadrado de Pearson ou Total, Mantel-Haenszel, Regressão Logística, Método Padronizado e, finalmente, o Logístico Interativo. Apresentamos as bases matemáticas desses métodos, suas principais vantagens e limitações. Destacamos que a presença do DIF em instrumentos de medida, sejam psicológicos ou pedagógicos, é um problema para o suposto da padronização ou uniformização das condições de aplicação dos testes e que, ademais, acarreta injustiça e falta de eqüidade ao processo avaliativo.

Palavras-chave: Funcionamento diferencial dos itens (DIF); teoria de resposta ao item (TRI); avaliação psicológica.

Main Methods for Detection of the Differential Item Functioning (DIF): A Description

Abstract

This paper aimed at describing the main methods used today in the detection of the differential functioning of items (DIF). We describe the methods of the Area between the ICC's, Comparison of the Probabilities, Comparison of the Items Parameters, Lord's Chi-square, Scheuneman's Chi-square, Pearson's or Total Chi-square, Mantel-Haenszel Method, Logistic Regression, Standardized Method and, finally, the Logistic Interactive Method. We present the mathematical basis, main advantages and limitations of these methods. Finally, we emphasize the presence of the DIF in psychological and educational tests as a problem for the assumption of standardized conditions and also as a cause of injustice and absence of equity for the assessment process.

Keywords: Differential item functioning (DIF); item response theory (IRT); psychological assessment.

A necessidade e relevância da padronização ou uniformização das condições de aplicação dos instrumentos de medida é um dos supostos mais importantes da avaliação, seja no âmbito psicológico ou educativo (Anastasi, 1988; Pasquali, 2000). Para tanto, psicólogos e pedagogos tratam de uniformizar as tarefas ou itens, as instruções, o tempo destinado à resolução das tarefas contidas nos instrumentos, a maneira de corrigir as respostas dos respondentes, as condições de luminosidade, som e a própria atividade de aplicação dos instrumentos de medida, etc. (Martínez Arias, 1997).

Com o recente surgimento do paradigma psicométrico denominado *Teoria da Resposta ao Item (TRI)*, novas áreas de investigação têm proliferado (Andriola, 1998; Hambleton, 1989a, 1990). Como opina Hambleton (1997), uma delas tem seu foco dirigido ao estudo do *Funcionamento Diferencial dos Itens (DIF)*[2], que está intimamente ligado ao suposto da padronização das condições de aplicação dos instrumentos de medida. Devemos ter claro que a presença de DIF num teste é um fator que torna o processo avaliativo injusto.

Para compreendermos essa última afirmação, deveremos conhecer o conceito de DIF. Podemos dizer que, no âmbito da TRI, o item *não* tem DIF quando a sua curva característica (CCI) é idêntica para os grupos comparados num mesmo nível ou magnitude da variável latente medida (Lord, 1980; Melenbergh, 1989). Em linguagem matemática poderíamos dizer que o *item não tem DIF com respeito a variável G (grupo), dado Z (nível de* q*), se, e somente se,* $F(X \mid g,z) = F(X \mid z)$.

Onde:

— X é a pontuação no item;

— g é o valor obtido em X segundo a variável G;

— z é o valor obtido em X segundo a variável Z.

Nesse caso, a associação entre ambas variáveis (g e z) pode favorecer ou prejudicar o rendimento de um grupo sobre o outro. Portanto, é necessário reconhecer que a presença de DIF ocasiona sérios problemas ao processo de avaliação, já que pode privilegiar um determinado grupo em detrimento de outro (Douglas, Roussos & Stout, 1996).

No âmbito da TRI, um item terá DIF se os sujeitos que têm o mesmo grau de aptidão ou habilidade (θ) e compõem distintos grupos demográficos (G_1 e G_2, por exemplo), possuírem diferentes probabilidades (P) de acertar um item (i) utilizado para medir dita aptidão ou habilidade (θ). Em notação matemática podemos afirmar, então, que *existe DIF quando* $P_{iG1}(\theta) \neq P_{iG2}(\theta)$.

[1] Endereço para correspondência: Calle Camino de los Vinateros, 157, Piso 2º, Puerta C, Moratalaz, C.P. 28030, Madrid, España. *E-mail:* w_andriola@yahoo.com

[2] A sigla DIF é originária do termo inglês *Differential Item Functioning*.

Em palavras mais simples e menos técnicas, poderíamos dizer que *a existência de DIF num teste ou item implica reconhecer que sujeitos com a mesma capacidade ou magnitude no construto latente medido possuem diferentes probabilidades de acertá-lo, pelo simples fato de pertencerem a grupos demográficos distintos*. Assim, a presença de DIF num instrumento de medida supõe o desrespeito ao suposto da uniformização das condições de aplicação do mesmo, já que privilegia alguns sujeitos em detrimento de outros, por causas secundárias e irrelevantes ao propósito do teste (Muñiz, 1997). *Ademais, ditas causas podem e devem ser controladas*. Por exemplo, segundo Ercikan (1998), durante o processo de elaboração dos itens o responsável por dita atividade deverá evitar usar:

— Termos ou símbolos conhecidos por grupos demográficos muito específicos;

— Termos ou símbolos que possam ter diferentes significados, segundo o contexto em que se use ou o grupo ao qual se refira;

— Sentenças cujo tamanho seja excessivamente grande;

— Sentenças ou termos pejorativos;

— Elementos secundários para aumentar a complexidade de uma sentença.

Se o psicometrista encontra-se analisando itens já construídos por terceiros deverá, então, recorrer aos procedimentos estatísticos para detectar o DIF e considerar ou supor que os aspectos enumerados por Ercikan (1998) tenham sido controlados adequadamente. Segundo Muñiz (1997), a casuística é interminável e se pode dizer que não existem provas ou testes inteiramente isentos de DIF. Trata-se trata, assim, de detectar a quantidade de DIF aceitável num determinado item ou teste, segundo os objetivos do processo de avaliação.

Como destaca Andriola (2000a), a importância dos estudos que objetivam a verificação do DIF é justificada por que cabe ao avaliador verificar se em seu teste existem itens com DIF para que (1) possa buscar as causas que o expliquem, (2) evitar sua utilização com o grupo em desvantagem e, finalmente, (3) controlar os fatores responsáveis pelo DIF para evitar construir novos itens com o mesmo problema (Ercikan, 1998; Hambleton, 1989b; Mislevy, 1996). Ressaltadas essas idéias fundamentais, é o momento de conhecermos os principais métodos utilizados, atualmente, na detecção do DIF, ademais de sua fundamentação matemático-estatística.

Classificação dos Métodos para Detectar o DIF

Segundo palavras de Hambleton, Swaminathan e Rogers (1991), a TRI oferece um marco apropriado ao estudo do DIF. Como afirma Muñiz (1997) a TRI *"(...) parece venir como anillo al dedo para la evaluación del funcionamiento diferencial de los items (...)"* (p. 165). Neste contexto, a lógica subjacente à detecção do DIF consiste em (1) estimar os parâmetros métricos dos itens para os grupos de interesse (de referência e focal); (2) colocar ditos parâmetros em uma mesma escala; (3) representá-los através de suas CCI's; (4) comparar ditas CCI's nos grupos escolhidos e, finalmente, (5) observar a significação estatística das possíveis discrepâncias entre as CCI's.

Como medir com precisão a discrepância entre as CCI's originárias de distintas subpopulações constitui o problema central desta área de investigação psicológica e educativa. Para tentar solucionar o dito problema, a partir da década de 1950, os estudiosos propuseram vários métodos para a determinação do DIF (Dorans & Holland, 1993). Tal variedade de métodos, segundo Whitmore e Shumacker (1999), pode ser agrupada em duas categorias, de acordo com o critério utilizado na determinação do DIF:

— *Métodos que utilizam um critério interno:* o próprio escore ou a pontuação obtida pelos sujeitos no teste ou grupo de itens estudado;

— *Métodos que utilizam um critério externo:* um critério externo ao teste ou grupo de itens, tal como, a pontuação em outros testes (Clauser, Nungester & Swaminathan, 1996).

Não obstante, existem outras propostas de classificação. Por exemplo, Melenbergh (1989), Van der Flier, Melenbergh, Adèr e Wijn (1984), propuseram a seguinte:

— *Métodos incondicionais*: baseados no suposto de que o grupo de sujeitos e itens tenha algum tipo de interação;

— *Métodos condicionais*: baseados no suposto de que os parâmetros do item sejam diferentes para os sujeitos com a mesma magnitude na variável latente, que são oriundos de distintos grupos demográficos. Tal suposto está baseado na idéia de que a dificuldade de um item tem dois componentes: um intrínseco (as características do item, tais como, tipo — aberto ou fechado; tamanho do enunciado e das alternativas; signos utilizados — verbal, numérico, abstrato, etc.) e um extrínseco (as características dos sujeitos, tais como, gênero, raça, idade, nível sócio-econômico, *background* educativo, etc.). Nesse âmbito, a dificuldade de um item expressa a interação entre ditos componentes (Scheuneman & Gerritz, 1990).

Para Van der Flier, Melenbergh, Adèr e Wijn (1984) são preferíveis os métodos condicionais, já que, como seu próprio nome indica, *condicionam a probabilidade de resposta a um certo nível de habilidade*. Todavia, a categoria *métodos condicionais* conta com uma classificação proposta por Millsap e Everson (1993):

— *Métodos de Invariância Condicional Observada:* utilizam as pontuações observadas no teste, desde a perspectiva da Teoria Clássica dos Testes (TCT), isto é, utilizam o escore total como resultado da soma das pontuações nos itens (p. ex.: métodos de Mantel-Haenszel, Regressão Logística e Delta Gráfico);

— *Métodos de Invariância Condicional Não Observada:* utilizam a habilidade estimada através da TRI (p. ex.: métodos da medida da área entre as CCI's, comparação dos parâmetros dos itens, comparação das probabilidades, Qui-quadrado de Lord).

Nesse contexto, devemos destacar que os métodos condicionais estão fundamentados num conhecido paradoxo, pelo menos no âmbito da literatura estatística, denominado *Paradoxo de Simpson* (Dorans & Holland, 1993).

Paradoxo de Simpson: Fundamento para Detecção do DIF

Dito paradoxo adota uma idéia simples e inteligente: *temos que comparar o comparável*. Para ilustrar a aplicabilidade e relevância de dito paradoxo na investigação do DIF, apresentamos um exemplo citado por Dorans e Holland (1993), cujos dados estão apresentados na Tabela 1.

Tabela 1. Freqüências de Respostas de Dois Grupos a um Item Hipotético

Nível em *m*	Grupo A			Grupo B		
	Nm	*Ncm*	*Ncm/ Nm*	*Nm*	*Ncm*	*Ncm/ Nm*
Baixo	400	40	0,10	1000	200	0,20
Médio	1000	500	0,50	1000	600	0,60
Alto	1000	900	0,90	400	400	1,00
Total	2400	1440	0,60	2400	1200	0,50

Os símbolos *Nm*, *Ncm* e *Ncm/Nm* estão referidos respectivamente:

— Ao número de sujeitos de magnitude *m* na variável latente;

— Ao número de sujeitos de magnitude *m* que acertaram o item;

— A proporção de sujeitos de magnitude *m* que acertaram o item.

Observamos que, dos 2.400 sujeitos do Grupo A, 60% (1.440) conseguiram acertar o item. Por outro lado, somente 50% dos sujeitos do Grupo B (1.200) responderam-no corretamente. Assim, a diferença entre a proporção de acertos nos dois grupos é de 0,10, favorável ao Grupo A.

Entretanto, a diferença entre a proporção de acertos dos dois grupos em cada um dos três níveis de magnitude *m*, que poderíamos chamar de *proporção condicional*, é:

— *No nível mais baixo:* 0,10 para o Grupo A e 0,20 para o Grupo B, ou seja, uma diferença de 0,10 favorável ao Grupo B;

— *No nível médio:* 0,50 para o Grupo A e 0,60 para o Grupo B, ou seja, uma diferença de 0,10 favorável ao Grupo B;

— *No nível alto:* 0,90 para o Grupo A e 1,00 para o Grupo B, ou seja, uma diferença de 0,10 favorável ao Grupo B.

Nesse âmbito, podemos observar que, quando condicionamos a proporção de acerto aos distintos níveis de habilidade, os resultados são bastante distintos da mesma análise realizada com a amostra total, já que esta última não considerou os distintos níveis de habilidade. Surge, então, a necessidade de diferenciar dois aspectos que estão subjacentes a estes diferentes resultados: *o impacto e o DIF*.

No primeiro tipo de análise, verificamos uma diferença quanto ao *impacto*. Dito termo faz referência à diferença entre os resultados de dois grupos de sujeitos; é efeito das diferenças individuais reais entre os grupos com respeito a variável latente ou construto medido através do item ou teste (Dorans & Schmitt, 1993; Zumbo, 1999). Retornando ao nosso exemplo, observamos que os sujeitos do Grupo A têm uma maior capacidade na variável latente medida pelo item, já que obtiveram uma diferença de 0,10 a seu favor.

No segundo tipo de análise, condicionamos as proporções de acerto a três distintos níveis da variável latente, cujos resultados são favoráveis ao Grupo B. Neste caso concreto, temos o exemplo de um item com DIF favorável aos sujeitos do Grupo B, ou seja, *apesar de ter uma menor capacidade na variável latente ou construto medido pelo item estes sujeitos possuem maiores probabilidades de acertar ao item em foco*. Este tipo de análise é o mais adequado ao estudo do DIF (Dorans & Holland, 1993; Van der Flier, e cols. 1984; Scheuneman & Gerritz, 1990).

Podemos dizer que, na prática da avaliação educativa e psicológica, o *Paradoxo de Simpson* enfatiza a importância de comparar a probabilidade de acerto a um determinado item, considerando sempre que os sujeitos têm o mesmo grau ou magnitude na variável latente medida pelo item. Existem diversos métodos ou procedimentos para estudar o DIF, todos eles baseados no Paradoxo de Simpson, alguns dos quais serão descritos a seguir.

Cálculo da Área entre as CCI's

Consiste em estimar as CCI's do item para os grupos de interesse do avaliador ou investigador e, em seguida, realizar o cálculo da área compreendida entre as CCI's (Wainer, 1993). A área entre as CCI's constitui um índice da discrepância entre elas. Em conseqüência, indica a possível existência de DIF, pois se ambas CCI's coincidissem a área entre as mesmas teria valor zero e, desse modo, não haveria DIF. A figura 1 ilustra a lógica do método apresentando as CCI's de um item para dois grupos e a área entre ambas, que deverá ser calculada.

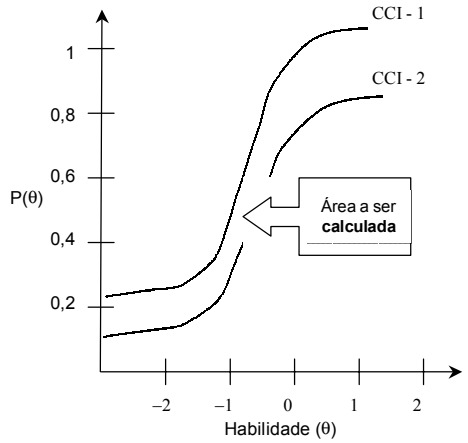

Figura 1. Representação de duas CCI`S e a área que as distingue.

No método das áreas, existem diversos procedimentos para a determinação do valor compreendido entre as CCI's dos grupos estudados. No caso da comparação de dois grupos, Rudner, Getson e Knight (1980) propuseram a seguinte fórmula para seu cálculo:

$$A = \sum_{\theta=-4}^{\theta=4} \left| P_{GR}(\theta_j) - P_{GF}(\theta_j) \right| \Delta\theta$$

Onde:

— $P_{GR}(\theta_j)$ é o valor da probabilidade de acerto ao item do grupo de referência, dado θ_j;

— $P_{GF}(\theta_j)$ é o valor da probabilidade de acerto ao item do grupo focal, dado θ_j;

— $\Delta\theta$ é o valor da base de um retângulo ($\Delta\theta$=0,005) e altura $[P_{GR}(\theta_j)-P_{GF}(\theta_j)]$.

Nesse procedimento as áreas são calculadas para os distintos valores de θ que estejam compreendidos no intervalo -4 a +4, com o incremento $\Delta\theta$. Nesse contexto, quanto menor o valor do incremento mais preciso será o cálculo da área. Linn e Harnisch (1981) propuseram outro procedimento dado por:

$$A = \sum_{\theta=-3}^{\theta=3} \sqrt{\left[P_{GR}(\theta_j) - P_{GF}(\theta_j)\right]^2 \Delta\theta}$$

Os elementos dessa fórmula têm o mesmo significado da proposta por Rudner, Getson e Knight (1980), sendo a única diferença o intervalo adotado para os distintos valores de θ que, neste caso, está compreendido entre -3 e +3. Outro procedimento para o cálculo do DIF foi proposto por Raju (1988):

$$A = (1-c)\left| \frac{2(a_2-a_1)}{Da_1a_2} \ln\left[1+e^{\frac{Da_1a_2(b_2-b_1)}{(a_2-a_1)}}\right] - (b_2-b_1) \right|$$

Onde:

— *a* é o parâmetro de dificuldade;

— *b* é o parâmetro de discriminação;

— *c* é a probabilidade de acerto ao acaso;

— *D* é uma constante de valor 1,7;

— *e* é a base dos logaritmos neperianos, de valor 2,7182.

Para o uso desta fórmula, assume-se que o valor do parâmetro *c* é o mesmo para os grupos analisados. Uma vez calculada a área entre as CCI's, o investigador poderá adotar a decisão a respeito da existência ou não de DIF. No entanto, deverá ter algum tipo de cuidado já que não existem provas de significação estatística apropriadas para a confrontação das duas CCI's comparadas (Muñiz, 1997).

Comparação das Probabilidades de Acertar o Item

Camilli e Shepard (1994) apresentam uma grande vantagem na sua utilização. Segundo eles, dito método permite utilizar unicamente os valores de θ para os quais existem sujeitos do grupo focal. Com este procedimento se pretende dar mais importância ao DIF nos intervalos onde realmente existem indivíduos pertencentes ao grupo focal, e não em outras zonas de θ onde não existem sujeitos. Sua formulação matemática é:

$$DP = \sum_{j=1}^{n_{GF}} \frac{\left[P_{GR}(\theta_j) - P_{GF}(\theta_j)\right]}{n_{GF}}$$

Onde:

— $P_{GR}(\theta_j)$ é a probabilidade que as pessoas do grupo de referência têm de superar o item para o valor θ_j;

— $P_{GF}(\theta_j)$ é a probabilidade que as pessoas do grupo focal têm de superar o item para o valor θ_φ;

— n_{GF} é o número de pessoas do grupo focal.

O somatório pode variar desde um até o número total de pessoas do grupo focal (n_{GF}), ou seja, só se consideram aqueles valores de θ obtidos pelos membros do grupo focal. Em síntese, como assinalam Camilli e Shepard (1994), é uma forma de autoponderação baseada nas pessoas do grupo focal, em que se dá mais peso às zonas de θ onde estas se encontram e, por outro lado, se omitem as zonas de θ onde não existem indivíduos de dito grupo.

No caso de que não existisse DIF, os valores $P_{GR}(\theta_\varphi)$ e $P_{GF}(\theta_j)$ coincidiriam para todos os valores de θ_j e, dessa

maneira, o valor do índice DP seria zero. O DIF aumentará na medida em que o valor do índice DP se distancie de zero, seja positivamente ou negativamente. Se o valor é positivo, quer dizer que os valores $P_{GR}(\theta_j)$ são superiores aos valores $P_{GF}(\theta_j)$, o que indicaria que o item em questão está prejudicando o grupo focal. Em caso contrário, se o valor de DP é negativo, o item está prejudicando o grupo de referência. Como ocorre no método das áreas, tampouco aqui existe uma prova estatística definitiva, que informe a respeito da significância do valor DP.

Comparação dos Parâmetros dos Itens

A lógica deste procedimento é simples: um item terá DIF se os parâmetros estimados nas subpopulações não coincidem, isto é, têm diferenças significativas (Thissen, Steinberg & Wainer, 1993). No caso do modelo logístico de um parâmetro, o que vai ser comparado nas subpopulações é o parâmetro *b*. Sua formulação é:

$$Z = \frac{\hat{b}_R - \hat{b}_F}{\sqrt{S^2(\hat{b}_R) + S^2(\hat{b}_F)}}$$

Onde:

— b_R e b_F são os parâmetros da dificuldade do item, estimados em cada grupo (referência e focal);

— $S^2(b_R)$ e $S^2(b_F)$ são as variâncias de *b* em cada grupo (referência e focal);

— Z tem distribuição normal.

O valor obtido de Z é comparado com o da distribuição normal, correspondente ao nível de confiança adotado, o que permite corroborar ou não a hipótese nula (H_0: $b_1 = b_2$). Para os modelos logísticos de dois e três parâmetros teremos que comparar os parâmetros *a* e *b*, considerando-se o valor do parâmetro *c* é invariante (Muñiz, 1997). As formulações matemáticas para a comparação de *a* e *b* são:

$$Z_a = \frac{\hat{a}_R - \hat{a}_F}{\sqrt{S^2(\hat{a}_R) + S^2(\hat{a}_F)}} \qquad Z_b = \frac{\hat{b}_R - \hat{b}_F}{\sqrt{S^2(\hat{b}_R) + S^2(\hat{b}_F)}}$$

Onde:

— b_R e b_F são os parâmetros de dificuldade do item, estimados em cada grupo (referência e focal);

— $\hat{a}_R$ e $\hat{a}_F$ são os parâmetros de discriminação do item, estimados em cada grupo (referência e focal);

— $S^2(b_R)$ e $S^2(b_F)$ são as variâncias de *b* em cada grupo (referência e focal);

— $S^2(\hat{a}_R)$ e $S^2(\hat{a}_F)$ são as variâncias de *a* em cada grupo (referência e focal);

— Z tem distribuição normal.

A principal limitação desse procedimento é que os parâmetros *a* e *b* têm que ser comparados separadamente.

Qui-quadrado de Lord

Devido à limitação do método de comparação dos parâmetros dos itens, Lord (1980) propôs outro procedimento em que as comparações dos parâmetros *a* e *b* podem ser realizadas ao mesmo tempo, através do uso do teste Qui-quadrado. Sua formulação matemática é:

$$\chi^2 = V \sum^{-1} V'$$

Onde:

— χ^2 tem dois graus de liberdade;

— V é o vetor de dimensão (1 x 2) das diferenças entre os parâmetros *a* e *b* dos grupos de referência e focal;

— V' é o vetor transposto de V;

— Σ^{-1} é a inversa da matriz soma de variâncias-covariâncias de V para os grupos de referência e focal, cuja dimensão é 2 x 2.

No caso de ser aplicado ao modelo logístico de um parâmetro sua formulação matemática é mais parcimoniosa:

$$\chi 2 = \frac{b_F - b_R}{Var(b_F) - Var(b_R)}$$

Onde b_F e b_R são os valores dos parâmetros *b* em cada grupo e $Var(b_F)$ e $Var(b_R)$ as variâncias estimadas de ditos parâmetros. Para observar a significância do qui-quadrado deveremos comparar o valor observado com o crítico (Thissen e cols., 1993).

Qui-quadrado de Scheuneman

O método proposto por J. Scheuneman, no ano 1979, ficou posteriormente conhecido como *Qui-quadrado de Scheuneman*. O autor parte da premissa de que as probabilidades de acerto dos grupos de referência e focal são iguais nos distintos níveis de habilidade, isto é: $p_{iFK} = p_{iRK}$, onde:

— p_{iRK} é a probabilidade do grupo de referência de acertar o item *i* no intervalo *k*;

— p_{iFK} é a probabilidade do grupo focal de acertar o item *i* no intervalo *k*.

Para o estudo dessa hipótese, Sheuneman (1979) propôs a seguinte prova estatística:

$$\chi_s^2 = \sum_{k=1}^{k} \left[\frac{[A_k - E(A_k)]^2}{E(A_k)} + \frac{[C_k - E(C_k)]^2}{E(C_k)} \right]$$

Na qual:

— $E(A_k) = n_{Rk} \cdot m_{1k} / T_k$;

— $E(C_k) = n_{Fk} \cdot m_{1k} / T_k$;

— n_{Fk} é o número de pessoas do grupo focal;

— n_{Rk} é o número de pessoas do grupo de referência;

— m_{1k} é o número de sujeitos que acertaram o item, que estão no nível *k* da pontuação observada;

— T_k é o número de sujeitos dos grupos de referência e focal que existe no nível *k* da pontuação observada.

Scheuneman (1979) assume que o valor do χ^2 segue uma distribuição como qui-quadrado com *(k-1)(r-1)* graus de liberdade, sendo *r* o número de grupos. Segundo ele, a principal vantagem do método está na sua simplicidade de cálculo (Scheuneman, 1981).

Qui-quadrado de Pearson ou Total

Dito método é também conhecido como total ou completo, porque utiliza tanto os totais marginais de respostas corretas como os totais marginais das respostas incorretas. Sua formulação matemática é dada por:

$$\chi^2 = \sum_{k=1}^{k}\left[\frac{T_k}{m_{0k}} x \frac{[A_k - E(A_k)]^2}{E(A_k)} + \frac{[C_k - E(C_k)]^2}{E(C_k)}\right]$$

Onde:

— T_k é o número de sujeitos dos grupos de referência e focal que existe no nível *k* da pontuação observada;

— m_{0k} é o número de sujeitos que erraram o item e que estão no nível *k* da pontuação observada;

— $E(A_k) = n_{Rk} . m_{1k} / T_k$;

— $E(C_k) = n_{Fk} . m_{1k} / T_k$;

— n_{Fk} é o número de pessoas do grupo focal;

— n_{Rk} é o número de pessoas do grupo de referência;

— m_{1k} é o número de sujeitos que acertaram o item, que estão no nível *k* da pontuação observada.

Na opinião de Hidalgo Montesinos, López Pina e Sánchez Meca (1997) a vantagem desse método reside no fato de considerar as freqüências observadas de respostas incorretas.

Regressão Logística

O modelo para predizer a probabilidade de ocorrência de uma resposta correta a um item, mais conhecido como método da regressão logística, tem a seguinte formulação matemática:

$$P(u = 1) = \frac{\exp(z)}{1 + \exp(z)}$$

Onde:

— *u* é a resposta ao item estudado, sendo $z = \tau_0 + \tau_1 q + \tau_2 g + \tau_3(\theta g)$.

Na segunda formulação, temos:

— τ_0 como ponto de interseção da reta de regressão com o eixo das abscissas;

— τ_1 como inclinação da reta de regressão;

— θ como a habilidade ou variável latente medida pelo item;

— τ_2 como a diferença entre o rendimento dos grupos no item em foco;

— *g* como grupo (de referência ou focal) ao qual pertencem os sujeitos;

— τ_3 como parâmetro indicador da possível interação entre θ e *g*.

Para explicar o DIF nos grupos de interesse (de referência e focal), deveremos especificar distintas equações. Assim, um item terá DIF uniforme ou consistente se $\tau_2 \neq 0$ e $\tau_3 = 0$; e terá DIF não uniforme ou inconsistente se $\tau_3 \neq 0$ (seja ou não $\tau_2 = 0$). Como destaca Bock (1975), este é um procedimento estatístico para predizer uma variável dependente, de natureza dicotômica, a partir de algumas variáveis independentes, em nosso caso, habilidade (θ) e grupo (*g*), sendo, ademais, um dos mais utilizados para detectar o DIF (Swaminathan & Rogers, 1990; Rogers & Swaminathan, 1993; Zumbo, 1999).

Método Mantel-Haenszel

Foi desenvolvido por N. Mantel e W. Haenszel no ano 1959, e aplicado ao estudo do DIF por P. W. Holland e D. T. Thayer em 1988 (Angoff, 1993; Dorans & Holland, 1993). Consiste, basicamente, na comparação das freqüências observadas e esperadas de acertos e erros nos grupos de referência e focal, de acordo com os distintos níveis de habilidade (j) escolhidos pelo investigador. Nesse contexto, as respostas dos sujeitos são organizadas em uma tabela de freqüências, como a apresentada a seguir.

Tabela 2. Freqüências Observadas de Respostas a um Item Hipotético

Grupos	*Acertos (1)*	*Erros (0)*	*Total*
De referência	A_j	B_j	n_{Rj}
Focal	C_j	D_j	n_{Fj}
Total	m_{1j}	m_{0j}	T_j

Baseados nesta lógica, N. Mantel e W. Haenszel propuseram a seguinte fórmula para a comparação das freqüências:

$$\alpha_{MH} = \frac{\sum_{j=1}^{S} \frac{A_j D_j}{T_j}}{\sum_{j=1}^{S} \frac{B_j C_j}{T_j}}$$

Onde:

— A_j é a freqüência observada das respostas corretas do grupo de referência nos distintos níveis de pontuação escolhidos;

— B_j é a freqüência observada das respostas incorretas do grupo de referência nos níveis de pontuação escolhidos;

— C_j é a freqüência observada das respostas corretas do grupo focal nos níveis de pontuação escolhidos;

— D_j é a freqüência observada das respostas incorretas do grupo focal nos níveis de pontuação escolhidos;

— T_j é o total de erros e acertos, de cada grupo, nos níveis de pontuação escolhidos.

O coeficiente α_{MH} é uma medida da quantidade de DIF, no qual o valor 1,0 significará idêntico comportamento do item para os grupos; os valores menores que 1,0 significarão maiores possibilidades de êxito no item para o grupo de referência (Longford, Holland & Thayer, 1993). O *Educational Testing Service (ETS)* propôs uma escala hierárquica para os distintos valores do coeficiente α_{MH}, de acordo com sua magnitude (Zwick, Thayer & Lewis, 1999):

— *Categoria C:* itens cujos valores absolutos sejam $1,0 < \alpha_{MH} < 1,5$ (sendo adotado α=0,05) são considerados *itens com DIF severo*;

— *Categoria B:* itens cujos valores absolutos sejam $0,0 < \alpha_{MH} < 1,0$ (sendo adotado α=0,05) são considerados *itens com DIF moderado*;

— *Categoria A:* itens cujos valores absolutos não sejam agrupados em nenhuma das duas categorias anteriores (sendo adotado α=0,05) são considerados *itens com DIF desprezível*.

Existe um estatístico de contraste para o coeficiente α_{MH}, que possibilita confrontar as hipóteses nula (H_0: α_{MH}=1) e alternativa (H_1: α_{MH}>1). O contraste é expresso em termos de:

$$\chi^2_{MH} = \frac{\left\{\left|\sum_{j=1}^{K} A_j - \sum_{j=1}^{K} E(A_j)\right| - 0{,}50\right\}^2}{\sum_{j=1}^{K} Var(A_j)}$$

Onde:

— A_j é a freqüência de respostas corretas do grupo de referência nos distintos níveis de pontuação observada;

— Var (A_j) é a variância de A_j;

— E (A_j) é a freqüência esperada para os distintos valores de A_j em cada nível da pontuação observada.

O valor do χ^2_{MH} se distribui aproximadamente como qui-quadrado com um grau de liberdade. A variância da freqüência das respostas corretas do grupo de referência, nos distintos níveis de pontuação observada, é dada por:

$$Var(A_j) = \frac{n_{Rj} n_{Fj} m_{1j} m_{0j}}{T_j^2 (T_j - 1)}$$

Onde:

— n_{Rj} é o número de sujeitos do grupo de referência no nível j da pontuação observada;

— n_{Fj} é o número de sujeitos do grupo focal no nível j da pontuação observada;

— m_{1j} é o número de sujeitos do nível j da pontuação observada, que acertou ao item;

— m_{0j} é o número de sujeitos do nível j da pontuação observada, que não acertou o item;

— T_j é o número de sujeitos dos grupos de referência e focal, que existe no nível j da pontuação observada.

A freqüência esperada para os distintos valores de A_j, em cada nível da pontuação observada, pode ser calculada mediante o uso da fórmula:

$$E(A_j) = \frac{n_{Rj} m_{1j}}{T_j}$$

Onde:

— n_{Rj} é o número de sujeitos do grupo de referência no nível j da pontuação observada;

— m_{1j} é o número de sujeitos do nível j da pontuação observada, que acertou o item;

— T_j é o número de sujeitos dos grupos de referência e focal existente no nível j da pontuação observada.

Método Padronizado

O método padronizado e o de Mantel-Haenszel são amplamente utilizados pelo *Educational Testing Service (ETS)*. Possibilita o cálculo do índice de discrepância entre os grupos com respeito ao rendimento num item (*p-difference*). Sua formulação matemática é dada por:

$$STD\ pDIF = \frac{\sum_m w_m (P_{fm} - P_{rm})}{\sum_m w_m}$$

Onde:

— w_m são os pesos adotados para os grupos estudados. Segundo Dorans e Holland (1993), alguns valores possíveis para w_m são:

— $w_m = N_{lm}$, isto é, o número total de sujeitos do nível *m* de habilidade;

— $w_m = N_{rm}$, isto é, o número total de sujeitos pertencentes ao grupo de referência, que estão no nível *m* de habilidade;

— $w_m = N_{fm}$, isto é, o número total de sujeitos pertencentes ao grupo focal, que estão no nível *m* de habilidade;

— w_m = a freqüência relativa de sujeitos pertencentes a algum dos grupos, que estão no nível *m* de habilidade;

— P_{fm} e P_{rm} são, respectivamente, as proporções de sujeitos que acertaram ao item, comparadas ao número total dos que contestaram ao mesmo item, no grupo focal e de referência.

O índice *STP pDIF* pode assumir valores entre –1 e +1, sendo que valores positivos indicam que o item favorece ao grupo focal; valores negativos indicam que

o item favorece ao grupo de referência. Para valores de *STP pDIF* entre –0,05 e +0,05 o DIF é irrelevante; para valores entre –0,06 e -0,010 e entre +0,06 e +0,010 o DIF é moderado; para valores superiores a +0,10 e inferiores a -0,10 o DIF é severo (Dorans & Holland, 1993).

Método Logístico Interativo

Foi formulado por F. B. Baker no início da década de 1980, como uma resposta às limitações do Qui-quadrado de Scheuneman (Baker, 1981a, 1981b). O modelo saturado de dito método tem a seguinte formulação matemática:

$ln\ (F_{ij1} / F_{ij2}) = C + S_i + G_j + (SG)_{ij}$, onde:

— *ln* é o logaritmo natural da proporção de respostas corretas (k=1) e incorretas (k=2);

— F_{ij1} e F_{ij2} são, respectivamente, a freqüência esperada de sujeitos (F) com pontuação *i* situados numa determinada categoria (S_i), que pertencem ao grupo *j* (G_j) e que acertarão (k=1) ou não (k=2) o item em foco;

— C representa o parâmetro de dificuldade do item para a amostra total;

— S_i é o efeito principal da pontuação *i* pertencente a uma categoria *S*;

— G_j é o efeito principal do grupo *j*;

— SG_{ij} é o parâmetro para a interação entre a pontuação *i* e o grupo *j*.

O modelo saturado é utilizado para verificar a presença do DIF não uniforme ou inconsistente. Porém, como destacam Van Der Flier e colaboradores (1984), para verificar a presença do DIF uniforme ou consistente, é comum adotar-se um modelo não saturado, cuja formulação matemática é:

$ln\ (F_{ij1} / F_{ij2}) = C + S_i + G_j$.

Já o modelo nulo, que representa a ausência de DIF, vem dado por:

$ln\ (F_{ij1} / F_{ij2}) = C + S_i$.

Nos três casos, temos:

$$\sum_{i=1}^{s} S_i = 0$$

$$\sum_{j=1}^{g} G_j = 0$$

$$\sum_{i=1}^{s} (SG)_{ij} = \sum_{j=1}^{g} (SG)_{ij} = 0$$

Para verificar o ajuste do modelo aos dados foi proposto o seguinte procedimento:

$$G^2 = 2\sum_{i=1}^{s}\sum_{j=1}^{g}\sum_{k=1}^{2} f_{ijk} \ln\left(\frac{f_{ijk}}{F_{ijk}}\right)$$

Onde G^2 tem distribuição como Qui-quadrado, com *g-1* graus de liberdade; f_{ijk} é a freqüência observada de sujeitos que acertaram (k=1) ou não (k=2) o item em foco, que estão na categoria de pontuação *i* e pertencem ao grupo *j*. Para o cálculo da freqüência esperada (F_{ijk}) devemos usar a fórmula:

$$F_{ijk} = \frac{\left(\sum_{j=1}^{g} f_{ijk}\right)\left(\sum_{k=1}^{2} f_{ijk}\right)}{\left(\sum_{j=1}^{g}\sum_{k=1}^{2} f_{ijk}\right)}$$

Principais Limitações dos Métodos para Detectar o DIF

Autores como Camilli e Shepard (1994), O'Neill e McPeek (1993) e Schmitt, Holland e Dorans (1993) assumem, publicamente, suas preocupações pessoais sobre a importância que os investigadores da área dão aos resultados matemático-estatísticos, esquecendo as considerações teóricas sobre as possíveis causas do DIF. Segundo eles, essa é uma tendência muito freqüente nas investigações sobre o DIF. Compartilhamos com tais autores a preocupação pela ausência de hipóteses baseadas em teorias sólidas, que sejam explicativas do DIF e que deveriam estar presentes no âmbito da investigação científica. Mantendo essa visão crítica, precisamos conscientizarmos das limitações da grande variedade de métodos descritos (Andriola, 2000b) entre as quais destacamos as seguintes:

— *Método da Comparação da Área entre as CCI's:* não conta com provas de significância estatística para confrontar o valor empírico da área entre as duas CCI's comparadas (Muñiz, 1997). Ainda que a ausência de uma prova de significação constitua um problema metodológico, na prática é aconselhável revisar o item. Nesse âmbito, é melhor incrementar o erro Tipo I (revisar ou eliminar itens que não tenham DIF), que Tipo II (não revisar ou deixar de eliminar itens que tenham DIF).

— *Método da Comparação das Probabilidades:* como ocorre no método das áreas, tampouco aqui existe uma prova estatística definitiva, que informe sobre a significância do valor *DP*. Assim, é conveniente adotar o mesmo procedimento de revisão dos itens, apresentado no método das áreas.

— *Método da Comparação dos Parâmetros dos Itens:* sua principal limitação reside na comparação por separado dos parâmetros *a* e *b*, para as sub-populações ou grupos estudados.

— *Método do Qui-quadrado de Lord:* a equivalência ou não entre os tamanhos dos grupos focal e de referência pode ocasionar a obtenção de resultados distintos para o DIF.

— *Método do Qui-quadrado de Scheuneman:* padece do mesmo problema do Qui-quadrado de Lord, isto é, os resultados obtidos para o DIF estão associados aos tamanhos amostrais dos grupos focal e de referência. Ademais, segundo Baker (1981a), o fato de considerar-se, unicamente, a proporção de acertos, pode afetar os resultados pela presença de diferenças reais entre os grupos (*diferenças no impacto*).

— *Método do Qui-quadrado de Pearson ou Total:* possui o mesmo problema dos métodos de Lord e de Scheuneman, isto é, a desigualdade dos tamanhos dos grupos focal e de referência pode ocasionar resultados contraditórios, em função da equivalência ou não entre ambos grupos.

— *Método Mantel-Haenszel:* como outro dos métodos que utilizam tabelas de contingência sofre, igualmente, o mesmo problema do Qui-quadrado de Lord, do Qui-quadrado de Scheuneman e do Qui-quadrado de Pearson ou Total, isto é, a desigualdade dos tamanhos dos grupos focal e de referência pode proporcionar resultados distintos para os índices DIF, em função da equivalência ou não entre ditos grupos.

Apesar dessas limitações, Camilli e Shepard (1994) apresentam algumas vantagens dos modernos métodos para detectar o DIF. Segundo os mesmos, parece haver um acordo generalizado sobre a potência e flexibilidade dos métodos baseados na TRI, sempre que (1) os tamanhos amostrais sejam adequados à estimação de parâmetros estáveis dos itens, (2) utilizem-se mais de um método para detectar o DIF e, ademais, (3) utilizem-se procedimentos estatísticos conjuntamente com procedimentos qualitativos, ou seja, opinião de especialistas na área (Allalouf e cols. 1999; Angoff, 1993; Douglas, Rousos & Stout, 1996; Downing & Haladyna, 1997; Zumbo, 1999).

Considerações Finais

Verificamos que apesar de existir grande variedade de métodos para investigar o problema do DIF, os mesmos padecem limitações. Autores mais críticos aconselham a complementar as análises estatísticas obtidas pelo uso de mais de um procedimento de detecção do DIF, com a opinião de especialistas na área e, assim, aumentar a validez dos resultados (Allalouf, Hambleton & Siresi, 1999; Angoff, 1993; Zumbo, 1999).

Tentamos demonstrar que a presença do DIF em itens de instrumentos para medida psicológica e pedagógica é um grave problema que atenta contra o suposto da padronização ou uniformização das condições de avaliação. É uma fonte de injustiça, já que produz falta de eqüidade em processos avaliativos; permite aos sujeitos que possuem mesmo grau ou nível na variável latente ou construto medido pelo item obter melhores resultados, já que esses têm maiores probabilidades de acertá-lo (Douglas e cols. 1996).

Nesse âmbito, caberá aos responsáveis pela construção, administração e comercialização de testes, psicológicos e pedagógicos, verificar a presença de itens com DIF em seus instrumentos, já que a sua existência é um fator de invalidação dos resultados. Também os psicometristas que começam a organizar bancos de itens necessitam realizar estudos para verificar a presença de DIF e, assim, evitar utilizá-los em processos avaliativos (Andriola, 1998).

Para finalizar, mencionaremos uma célebre frase latina que é muito sugestiva e sintetiza, na nossa opinião, a importância dos estudos sobre o DIF no âmbito da avaliação psicológica e educativa: *fiat justítia, péreat mundus.*[3]

Referências

Allalouf, A., Hambleton, R. K. & Siresi, S. G. (1999). Identifying the causes of DIF in translated verbal items. *Journal of Educational Measurement, 36,* 185-198.

Anastasi, A. (1988). *Psychological testing.* New York: MacMillan.

Andriola, W. B. (1998). Utilização da teoria de resposta ao item (TRI) para a organização de um banco de itens destinados à avaliação do raciocínio verbal. *Psicologia: Reflexão e Crítica, 11,* 295-308.

Andriola, W. B. (2000a). Funcionamento diferencial dos itens (DIF): Estudo com analogias para medir o raciocínio verbal. *Psicologia: Reflexão e Crítica, 13,* 473-481.

Andriola, W. B. (2000b). Principales métodos para la determinación del funcionamiento diferencial de los ítems (DIF). *XII Congreso Nacional y I Iberoamericano de Pedagogía. Resúmenes de Comunicaciones, Tomo II,* 49-50.

Angoff, W. H. (1993). Perspectives on differential item functioning. Em P. W. Holland & H. Wainer (Orgs.), *Differential item functioning* (pp. 3-24). New Jersey: Lawrence Erlbaum.

Baker, F. B. (1981a). A criticism of Scheuneman's item bias technique. *Journal of Educational Measurement, 18,* 59-62.

Baker, F. B. (1981b). Log-linear, logit linear models. A didactic. *Journal of Educational Statistics, 6,* 75-102.

Bock, R. D. (1975). *Multivariate statistical methods.* New York: McGraw-Hill.

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

Clauser, B. E., Nungester, R. J. & Swaminathan, H. (1996). Improving the matching for DIF analysis by conditioning on both test score and an educational background variable. *Journal of Educational Measurement, 33,* 453-464.

Dorans, N. J. & Holland, P. W. (1993). DIF detection and description: Man–tel-Haenszel and Standardization. Em P. W. Holland & H. Wainer (Orgs.). *Differential item functioning* (pp. 35-66). New Jersey: Lawrence Erlbaum.

Dorans, N. J. & Schmitt, A. P. (1993). Constructed response and differencial item functioning: A pragmatic approach. Em R. E. Bennett & W. C. Ward (Orgs.), *Construction versus multiple choice items in cognitive measurement* (pp. 137-166). New Jersey: Lawrence Erlbaum.

[3] Faça-se justiça; pereça o mundo.

Douglas, J. A., Roussos, L. A. & Stout, W. (1996). Item-Bundle DIF hypothesis testing: Identifying suspect bundles and assessing their differential functioning. *Journal of Educational Measurement, 33,* 465-484.

Downing, S. M. & Haladyna, T. M. (1997). Test item development: Validity evidence from quality assurance procedures. *Applied Measurement in Education, 10,* 61-82.

Ercikan, K. (1998). Translation effects in international assessments. *International Journal of Educational Research, 29,* 543-553.

Hambleton, R. K. (1997). Perspectivas futuras y aplicaciones. Em J. Muñiz, *Introducción a la teoría de respuesta a los ítems* (pp. 203-213). Madrid: Ediciones Psicología Pirámide.

Hambleton, R. K. (1989a). Principles and selected applications of item response theory .Em R. L. Linn (Org.), *Educational measurement* (pp. 147-200). New York: MacMilan.

Hambleton, R. K. (1989b). Introduction. *International Journal of Educational Research, 13,* 123-125.

Hambleton, R. K. (1990). Item response theory: Introduction and bibliography. *Psicothema, II,* 97-107.

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

Hidalgo Montesinos, M. D., López Pina, J. A. & Sánchez Meca, J. (1997). Error tipo I y potencia de las pruebas chi-cuadrado en el estudio del funcionamiento diferencial de los ítems. *Revista de Investigación Educativa, 15,* 149-168.

Linn, R. L. & Harnisch, D. L. (1981). Interactions between item content and group membership on achievement test items. *Journal of Educational Measurement, 18,* 109-118.

Longford, N. T., Holland, P. W. & Thayer, D. T. (1993). Stability of the MH D-DIF statistics across populations. Em P. W. Holland & H. Wainer (Orgs.). *Differential item functioning* (pp. 171-196). New Jersey: Lawrence Erlbaum.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems.* New Jersey: Lawrence Erlbaum.

Martínez Arias, R. (1997). *Psicometría. Teoría de los tests psicológicos y educativos.* Madrid: Síntesis.

Mellenbergh, G. J. (1989). Item bias and item response theory. *International Journal of Educational Research, 13,* 127-143.

Millsap, R. E. & Everson, H. T. (1993). Methodology review: Statistical approaches for assessing measurement bias. *Applied Psychological Measurement, 17,* 277-334.

Mislevy, R. J. (1996). Test theory reconceived. *Journal of Educational Measurement, 33,* 379-416.

Muñiz, J. (1997). *Introducción a la teoría de respuesta a los items.* Madrid: Pirámide.

O'Neill, K. A. & McPeek, W. M. (1993). Item and test characteristics that are associated with differential item functioning. Em P. W. Holland & H. Wainer (Orgs.). *Differential item functioning* (pp. 255-276). New Jersey: Lawrence Erlbaum.

Pasquali, L. (2000). *Psicometria: Teoria dos testes psicológicos.* Brasília: Prática.

Raju, N. S. (1988). The area between two item characteristic curves. *Psi–chometrika, 42,* 549-565.

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.

Rudner, L. M., Getson, P. R. & Knight, D. L. (1980). Biased item detection techniques. *Journal of Educational Statistics, 5,* 213-233.

Scheuneman, J. D. (1979). A new method for assessing bias in test items. *Journal of Educational Measurement, 16,* 143-152.

Scheuneman, J. D. (1981). A response to Baker's criticism. *Journal of Edu–cational Measurement, 18,* 63-66.

Scheuneman, J. D. & Gerritz, K. (1990). Using differential item functioning procedures to explore sources of item difficulty and group performance characteristics. *Journal of Educational Measurement, 27,* 109-131.

Schmitt, A. P., Holland, P. W. & Dorans, N. J. (1993). Evaluating hypotheses about differential item functioning. Em P. W. Holland & H. Wainer (Orgs.), *Differential item functioning* (pp. 281-315). New Jersey: Lawrence Erlbaum.

Swaminathan, H. & Rogers, H. J. (1990). Detecting differential item func–tioning using logistic regression procedures. *Journal of Educational Measurement, 27,* 361-370.

Thissen, D., Steinberg, L. & Wainer, H. (1993). Detection of differential item functioning using the parameters of item response models. Em P. W. Holland & H. Wainer (Orgs.). *Differential item functioning* (pp. 67-114). New Jersey: Lawrence Erlbaum.

Van der Flier, H., Mellebergh, G. J., Adèr, H. J. & Wijn, M. (1984). An in–teractive item bias detection method. *Journal of Educational Measurement, 21,* 131-145.

Wainer, H. (1993). Model-Based standardized measurement of an item's differential impact. Em P. W. Holland & H. Wainer (Orgs.). *Differential item functioning* (pp. 123-136). New Jersey: Lawrence Erlbaum.

Whitmore, M. L. & Shumacker, R. E. (1999). A comparison of logistic re-gression and analysis de variance differential item functioning detection methods. *Educational and Psychological Measurement, 59,* 910-927.

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.* Ottawa: Directorate of Human Resources Research and Evaluation, Department of National Defense of Canadá.

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

Recebido: 06/11/2000
Revisado: 20/01/2001
Aceite Final: 15/03/2001

Sobre o autor

Wagner Bandeira Andriola é Psicólogo, Especialista em Psicometria (UnB), Mestre em Psicologia Social e do Trabalho (UnB), Doutorando em Avaliação Educativa pela Universidad Complutense de Madrid (UCM), Professor do Curso de Pedagogia da Universidade Federal do Ceará (UFC).