Psicologia: Reflexão e Crítica, 2000, 13(3), pp.475-483

Funcionamento Diferencial dos Itens (DIF): Estudo com Analogias para Medir o Raciocínio Verbal

Wagner Bandeira Andriola[1]
Universidade Federal do Ceará

Resumo

Este estudo objetivou determinar o funcionamento diferencial de 30 analogias destinadas à avaliação do raciocínio verbal, considerando a variável sexo. Utilizou-se uma amostra de 730 alunos do Ensino Médio, com idade média de 17,74 anos (*dp*= 3,12 anos). A maioria procedia de escolas públicas (58,5%) e era do sexo feminino (53,2%). Os grupos organizados para a investigação foram compostos por homens (*n*=342) e mulheres (*n*=388). Os parâmetros métricos dos itens foram determinados pelo modelo TRI de dois parâmetros logísticos. Para a verificação do DIF foram comparados os parâmetros métricos dos itens. Os resultados indicaram a presença de cinco itens com DIF.

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

Differential Items Functioning (DIF): Study with Analogies for Measurement the Verbal Reasoning

Abstract

This research aimed the determination of the differential item functioning (DIF) in 30 analogies used for the verbal reasoning assessment in students, taking into account the sex variable. A sample of 730 high school students, whose average age was 17,74 years (*sd* = 3,12 years) was used. The majority was composed by students from public schools (58,4%) and females (53,3%). The groups which participated in the study of DIF were composed by men (*n*= 342) and women (*n*= 388). The metric parameters of the items were determined according to the TRI model of two logistics parameters. For the determination of the DIF the method of comparation of the metric parameters of the items was used. The results indicated the presence of five items with DIF.

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

No âmbito da Teoria Clássica dos Testes (TCT) o termo *viés* é utilizado para rotular os itens que possuem parâmetros de dificuldade ou discriminação diferentes nos distintos grupos investigados. De acordo com Camilli e Shepard (1994), o viés é uma fonte de invalidez ou de erro sistemático que influencia no modo como um teste ou item mede aos membros de um grupo particular; é sistemático porque produz distorções nos resultados de um teste.

Hambleton (1989a) apresentou uma coletânea de diversas definições de viés, entre as quais destacamos as seguintes:

– O item tem viés se as médias dos escores dos grupos de comparação são significativamente diferentes. De acordo com esta definição um item tem viés se, por exemplo, a média dos escores das mulheres é significativamente superior a média dos escores dos homens no item estudado;

– O item tem viés se existem diferenças significativas entre os índices de dificuldade para grupos ou populações distintas;

– O item tem viés se existe interação entre a pontuação ítem-teste-grupo, isto é, se existe interação entre o rendimento dos grupos de comparação com respeito ao escore total no teste e no item;

– O item tem viés quando os sujeitos de distintos grupos, que possuem a mesma pontuação no teste, têm diferentes probabilidades de responder corretamente ao item. O problema desta definição é que a pontuação total do sujeito em um teste depende do índice de dificuldade do item e este índice pode não ser o mesmo para os distintos grupos comparados.

É interessante observarmos que a idéia de *grupo* é central nestas diversas definições. Por esse motivo, o *viés* tem sido estudado fundamentalmente nas investigações acerca das diferenças relacionadas com algumas características dos grupos, tais como o sexo, idade, classe social, região de moradia ou qualquer outra característica

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

sociodemográfica dos sujeitos. Neste âmbito, sempre que se identifiquem grupos de sujeitos para os quais haja suspeita de diferenças nas pontuações obtidas no teste, se deve aplicar algum procedimento para a análise de potenciais viéses (Martínez Arias, 1997).

Hambleton (1989a), ao finalizar sua exposição sobre as várias definições de viés, constata que todas elas padecem de um problema comum: não consideram a necessidade de controlar a própria capacidade dos sujeitos na variável latente medida pelo item ou teste. Reside aqui a principal diferença entre os conceitos de *viés* e de funcionamento diferencial dos itens (*DIF*). O procedimento DIF trata de controlar a magnitude da variável latente (geralmente expressa pela letra grega θ) no item avaliado, isto é, os grupos são comparados com respeito às suas pontuações no item ou teste, considerando-se que suas magnitudes na variável latente (capacidades, habilidades ou aptidões) têm idêntico valor (Hambleton, 1994).

Alguns autores, entre os quais Camilli e Shepard (1994), insistem que os índices estatísticos utilizados na análise do DIF por si mesmos não proporcionam prova de viés, preferindo denominá-los de *índices de discrepância do item* ou de *funcionamento diferencial do item*. Segundo eles, este último termo engloba os diferentes procedimentos estatísticos para a detecção de um possível funcionamento diferencial, todavia, insistem em que o DIF não é sinônimo de viés, embora alguns autores creiam que sim. Desse modo, os termos *funcionamento diferencial do item* (DIF) e *viés* não deveriam empregar-se como sinônimos.

É que os métodos estatísticos de DIF se utilizarão para identificar itens que exibem um funcionamento diferencial nos distintos grupos e, posteriormente, depois de uma análise lógica ou experimental no contexto da validez de construto dos itens, se determinará quais têm viéses para que, assim, possam ser eliminados do teste ou banco de itens. Em outras palavras, os métodos de DIF são procedimentos estatísticos, enquanto a análise do viés se insere num contexto mais geral da validez do teste.

Não obstante, neste último procedimento se usam, freqüentemente, os resultados obtidos com a aplicação do primeiro. Como afirmam Camili e Shepard (1994) e Mellenbergh (1989), os índices DIF algumas vezes produzem resultados estatisticamente significativos na ausência do viés e em outras vezes não detectam o viés quando este se encontra presente em muitos itens, dada a circularidade do critério interno que utilizam (Whitmore & Shumacker, 1999).

Definição do Termo "Funcionamento Diferencial do Item" (DIF)

Atualmente, quase não se utiliza o termo *viés dos itens*, que foi preterido pelo de *funcionamento diferencial dos itens*[2]. Uma das razões que explicam tal troca é que o uso das técnicas desenvolvidas para a detecção do DIF tem proliferado muito nos últimos anos, sobretudo, nos Estados Unidos de América.

Outra razão deve-se à potência e força das técnicas para o estudo do DIF, já que foram desenvolvidas com o objetivo de detectar se um item funciona igual ou diferentemente para grupos de distintas características sociodemográficas, cujos sujeitos componentes tenham a mesma magnitude na variável medida (Camilli & Penfield, 1997; Jiang & Stout, 1998; Kim & Cohen, 1998; Oshima, Raju & Flowers, 1997; Scheuneman & Grima, 1997; Wainer & Lukhele, 1997; Williams, 1997). Isso é tudo, as técnicas para a detecção do DIF não permitem ir mais além, quer dizer, não possibilitam nenhuma informação acerca de sua natureza ou causa (Cohen, Kim & Baker, 1993; Muñiz, 1997).

Cohen, Kim e Baker (1993) distinguem os estudos de DIF quanto aos objetivos pretendidos, isto é, se pode falar de investigações para a *detecção do DIF* e outras para *verificar o impacto do DIF*. No primeiro grupo estão as investigações que utilizam algum método tradicional para a identificação do DIF. Neste caso, os estudos objetivam somente a detecção do DIF, quer dizer, determinar a possível diferença entre as CCI's dos itens de acordo com os grupos comparados. No segundo grupo estão as investigações para a busca e identificação das causas do DIF. Neste contexto, o objetivo do investigador que utiliza as distintas técnicas para a detecção do DIF, é tentar saber quais são as causas (psicológicas, educativas, culturais, sociais, atitudinais, etc.) que ocasionam o funcionamento diferencial de um item (Downing & Haladyna, 1997).

No âmbito da TRI um item não tem DIF quando a curva característica do item (CCI) é idêntica para os grupos comparados em um mesmo nível ou magnitude da variável latente medida através do item (Mellenbergh, 1989). Geralmente os estudos para a determinação do DIF utilizam dois grupos, denominados *de referência (GR)* e *focal (GF)*.

Como já enfatizamos, em termos da TRI, um item terá DIF se para valores iguais de θ não correspondem valores iguais de P(θ) nas CCI's dos grupos considerados, isto é, quando $T_{jGR}(\theta) \neq T_{jGF}(\theta)$, onde:

[2] A sigla adotada, *DIF*, é originária do termo em Inglês ***d****ifferential* ***i****tem* ***f****unctioning*.

– T_{jGR} é a pontuação verdadeira do sujeito j que pertence ao grupo de referência e que tem uma certa magnitude na variável latente θ;

– T_{jGF} é a pontuação verdadeira do sujeito j que pertence ao grupo focal e que tem uma certa magnitude na variável latente θ.

De acordo com Mazor, Hambleton e Clauser (1998), o uso do número de respostas corretas para a determinação do DIF só é aceitável no caso do teste ser unidimensional e, ademais, quando as respostas são dicotômicas. Como modo de visualizar o DIF em um hipotético item se apresenta a Figura 1.

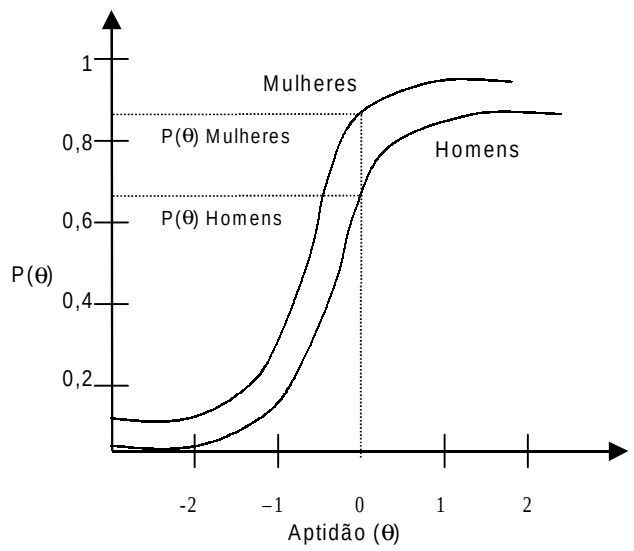

Figura 1. Representação gráfica das CCI's de um item com DIF

Podemos observar que para uma mesma magnitude de θ o valor de P(θ) é sempre superior para as mulheres, ou seja, em níveis iguais de competência na variável medida θ não existe a mesma probabilidade de superar o item. Neste caso, o item tem viés contra os homens (GR) pois os valores de P(θ) para um mesmo nível de θ são sempre maiores para as mulheres (GF). Se adotarmos o valor θ = 0, observamos que a probabilidade de acerto [P(θ)] para os homens é 0,65, enquanto para as mulheres é 0,85.

Como conseqüência de resultados dessa natureza, Douglas, Roussos e Stout (1996) propuseram os conceitos de *DIF benigno* e *DIF adverso*. No caso do DIF beneficiar o grupo de referência, isto é, quando $T_{jR}(\theta) > T_{jF}(\theta)$, é caracterizado o *DIF benigno*. O *DIF adverso* ocorre no caso do DIF beneficiar o grupo focal, isto é, quando $T_{jGR}(\theta) < T_{jGF}(\theta)$. No exemplo da Figura 1 temos o caso de um DIF adverso.

Ainda utilizando o item da Figura 1 tentemos aclarar o que ocorre com um item sem DIF, observando, para tanto, a Figura 2.

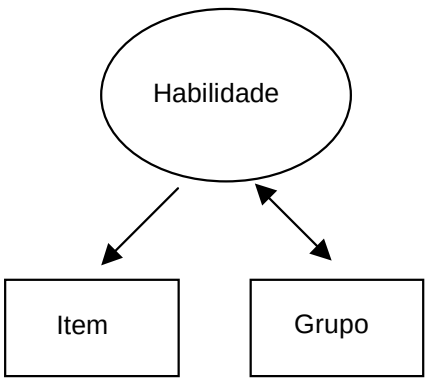

Figura 2. Relação entre habilidade, item e grupo num item sem DIF

O círculo indica a habilidade ou construto latente, que tem uma relação causal com o item. Grupo e variável latente estão associados. Em outras palavras e como modo de exemplificar, poderíamos dizer que as mulheres têm elevada habilidade na variável latente e que esta variável tem uma relação causal sobre o item, isto é, o grupo com maior capacidade (as mulheres) na variável latente têm mais respostas corretas no item. Neste caso, o rendimento no item depende, exclusivamente, da magnitude da variável latente que os indivíduos tenham, ou seja, se trata de um item sem DIF.

Agora, observemos o que ocorre num item com DIF, representado na Figura 3.

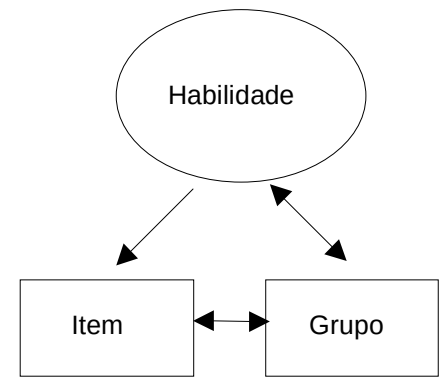

Figura 3. Relação entre habilidade, item e grupo num item com DIF

No caso da Figura 3 temos a mesma situação descrita na Figura 2, porém com o fato de existir uma associação entre grupo e item. Neste segundo caso, a associação

pode favorecer o rendimento de um grupo sobre outro devido a características particulares como sexo, raça, *background* educativo, origem social, etc. Deve ser enfatizado que, neste caso, se supõe que a magnitude da variável latente está sendo controlada, ou seja, os sujeitos são comparados com respeito a seu rendimento tendo em conta que possuem a mesma magnitude no construto. Este segundo exemplo caracteriza o caso no qual o rendimento no item não depende somente da magnitude da variável latente que os indivíduos tenham, senão de características do grupo, isto é, se trata de um item com DIF. Em nosso exemplo, a característica do grupo que influencia o rendimento diferenciado no item é o fato do sujeito ser homem ou mulher. Em síntese, se trata de uma característica de natureza demográfica que influencia o rendimento dos sujeitos com a mesma capacidade (Andriola, no prelo a).

É necessário reconhecer que o DIF ocasiona sérias implicações ao processo de avaliação, já que pode privilegiar um determinado grupo em detrimento de outro (Douglas, Roussos & Stout, 1996), conforme observamos no exemplo do rendimento dos homens e das mulheres. Muñiz (1997) adverte que o problema pode ter repercussões sociais mais graves se é a cultura dominante que elabora os itens para avaliar os sujeitos oriundos de culturas minoritárias. Por exemplo, suponhamos que são construídos itens para avaliar a capacidade de raciocínio verbal em alunos de escolas públicas e privadas. Ocorre que os alunos destes tipos de escolas são, geralmente, oriundos de classes sociais muito distintas, com diferentes *backgrounds* culturais, sociais, econômicos, etc. Todos estes aspectos podem implicar em que um tipo de aluno tenha o vocabulário mais rico que o outro. Dado que o raciocínio verbal é medido através de itens que empregam palavras, muito provavelmente, aquele tipo de aluno que conheça melhor o vocabulário utilizado nos itens terá uma clara vantagem na resolução destes mesmos itens.

Em síntese, o que tentamos dizer é que dada a grande variabilidade dos antecedentes históricos dos sujeitos implicados na avaliação do raciocínio verbal, se o item ou teste, se apoia mais nos antecedentes de uma cultura que nos da outra, terá altíssimas probabilidades de não ser eqüitativo, de possuir algum tipo de viés (Andriola, no prelo a). Em outras palavras, se confunde o efeito da capacidade de raciocínio verbal com o grau de conhecimento do vocabulário, ou seja, se um aluno tem baixa pontuação no teste não saberemos ao certo se pode explicar tal fato à sua baixa capacidade de raciocínio verbal ou ao seu baixo conhecimento vocabular.

Como nos fala Muñiz (1997), a casuística é interminável e pode dizer-se que não existem testes ou itens isentos completamente de viéses. Nesse contexto, se trata de detectar a quantidade de viés que é tolerável em um determinado teste ou item. Finalmente, deve ser enfatizado que neste contexto, a importância dos estudos que objetivam verificar a existência de DIF está plenamente justificada. Cabe ao avaliador verificar se em seu teste existem itens com DIF para que possa buscar as causas explicativas e, assim, evitar sua utilização com o grupo em desvantagem (Hambleton, 1989b; Mislevy, 1996).

Tipos de DIF no Âmbito da TRI

Enfatizamos que no contexto da TRI a lógica para a detecção do DIF consiste na comparação das CCI's dos itens investigados, considerando os grupos de referência e focal. Os distintos métodos para a detecção do DIF foram desenvolvidos baseados em dois tipos de DIF. O primeiro tipo é denominado *DIF uniforme* ou *consistente*, e se observa quando as CCI's do item estudado com respeito aos grupos de referência e focal são diferentes, porém não se cruzam. Em outras palavras, quando há uma vantagem relativa para um dos grupos investigados, cujo valor é constante ao largo de todo o intervalo atitudinal. Este caso ocorre quando o parâmetro *a* (discriminação) tem o mesmo valor nas duas CCI's, isto é, quando as CCI's são paralelas. Este tipo de DIF está representado na Figura 4.

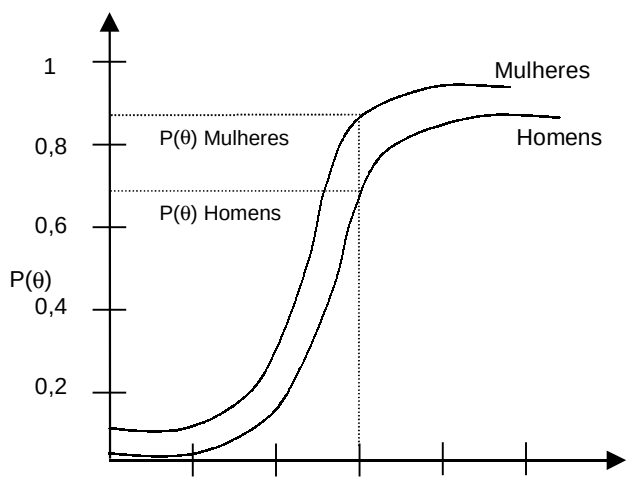

Figura 4. Representação gráfica de um item com DIF uniforme

Na Figura 4, observamos que a CCI das mulheres (grupo focal) está situada mais à esquerda que a dos

homens (grupo de referência), indicando que o item é mais fácil para o grupo focal. Tal diferença no parâmetro *b* (dificuldade) supõe que o item possui DIF. O segundo tipo de DIF é denominado *DIF não uniforme* ou *inconsistente*, e se observa quando as CCI's do item estudado, com respeito aos grupos de referência e focal, são diferentes e, ademais, se cruzam em algum ponto do intervalo atitudinal. Em outras palavras, quando há uma vantagem relativa para um desses grupos investigados, cujo valor é variável ao largo de todo o intervalo atitudinal. Este caso ocorre quando os parâmetros *a*, *b* ou *c* têm valores distintos nas duas CCI's, isto é, quando as CCI's não são paralelas. Este tipo de DIF está representado na Figura 5.

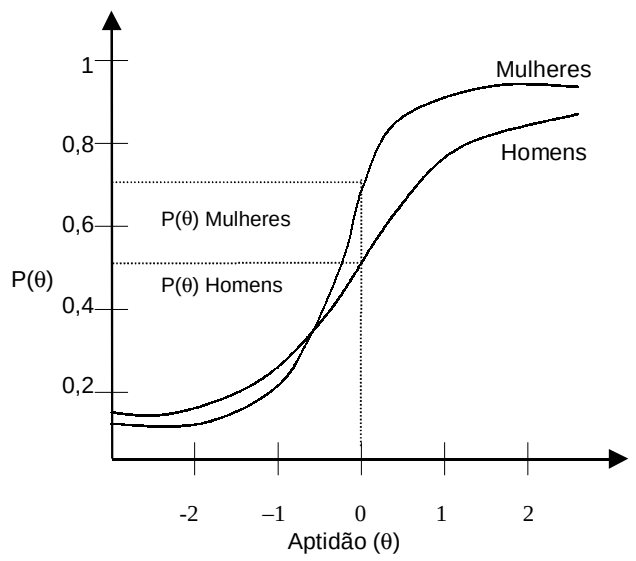

Figura 5. Representação gráfica de um item com DIF não-uniforme

A Figura 5 ilustra o caso de diferenças nos parâmetros *a*, *b* e *c* para os dois grupos investigados. É necessário dizer que neste segundo tipo de DIF é inapropriado um exame global dos dados, dado que tal procedimento poderia ocultar sua presença. É que a natureza variável do DIF em distintas zonas da variável latente pode cancelar total ou parcialmente sua detecção (Martínez Arias, 1997). Nesse caso concreto não se deve utilizar, por exemplo, o procedimento denominado *Differential Bundles Functioning (DBF)*, que estuda o DIF dos itens a partir de sua organização em subconjuntos (*bundles*) com alguma característica comum (Douglas, Roussos & Stout, 1996).

Métodos para Detecção do DIF

A Teoria de Resposta ao Item (TRI) oferece um marco bastante apropriado para o estudo do DIF (Hambleton, Swaminathan & Rogers, 1991). Neste contexto, a lógica subjacente à detecção do DIF consiste em:

– Estimar os parâmetros métricos dos itens para os grupos de comparação;
– Colocá-los em uma mesma escala;
– Representá-los através de suas curvas características (CCI's);
– Comparar os valores dos seus parâmetros entre os grupos escolhidos.

Já dissemos que, segundo a ótica da TRI, um item terá DIF se sua CCI não é a mesma para os grupos comparados, cujos sujeitos tenham a mesma magnitude na variável latente θ (Kim & Cohen, 1998). Como conseqüência desta definição, o procedimento mais adequado para avaliar o DIF é a estimação da CCI para ambos grupos e sua posterior comparação. No caso de que difiram significativamente é possível afirmar que o item tem DIF (Mazor, Hambleton & Clauser, 1998).

Não obstante, o problema central desta área de investigação psicológica e educativa está em como medir com precisão a discrepância entre as CCI's originadas de distintos subgrupos. Assim, existem diferentes procedimentos para o estudo do DIF apresentados por Holland e Wainer (1993), entre os quais destacamos:

– O método das áreas;
– O método das probabilidades;
– O método de comparação dos parâmetros métricos dos itens;
– O método da regressão logística;
– O método do *Qui-quadrado* de Lord;
– O método de Mantel-Haenszel.

As investigações para a detecção e estudo do DIF estão baseadas em uma mesma argumentação: a existência do DIF é um fator que influencia a validez da interpretação, realizada a partir da pontuação obtida pelo sujeito num item ou teste. É que sobre a interpretação da pontuação, seja no âmbito psicológico ou educativo, reside toda a credibilidade e reputação da investigação (Downing & Haladyna, 1997). Assim, está plenamente justificada a relevância das pesquisas a respeito do DIF, sobretudo aquelas que buscam identificar suas causas. Geralmente, as investigações sobre o DIF adotam como variáveis de comparação o sexo, o nível sócio-econômico, o grupo étnico e o *background* educativo (Mellenbergh, 1989; Clauser, Nungester & Swaminathan, 1996).

Neste contexto, o objetivo de nosso estudo foi investigar o funcionamento diferencial dos itens em dois grupos de alunos: homens e mulheres. Para tanto, foram utilizados 30 itens destinados à avaliação do raciocínio verbal de estudantes do Ensino Médio, já calibrados e organizados em um banco de itens por Andriola (1998).

Método

Participantes

Foi composta por 730 estudantes do Ensino Médio, cuja idade média foi 17,7 anos (*dp* = 3,12 anos), sendo a maioria originária de escolas públicas (58,5%) e pertencendo ao sexo feminino (53,2%).

Instrumento

Foram utilizadas 30 analogias verbais componentes de um banco de itens já calibrados anteriormente para o uso com a população de estudantes do segundo grau, através do modelo logístico de dois parâmetros (Andriola, 1998).

Procedimento

Depois dos primeiros contatos com os dirigentes das escolas, para organizar os horários e as turmas que seriam utilizadas no estudo, os 30 itens foram aplicados de modo coletivo na amostra de estudantes. Não houve limitação de tempo para a resolução dos itens.

Resultados

Para a análise estatística dos dados utilizou-se o programa BILOG-MG, desenvolvido por Zimowski, Muraki, Mislevy e Bock (1996). Este programa utiliza o método de comparação dos parâmetros métricos dos itens para verificar a existência de DIF, mais especificamente a dificuldade do item (parâmetro *b*). Todavia, os itens utilizados no estudo foram calibrados através do modelo TRI de dois parâmetros logísticos, ou seja, tiveram determinados sua discriminação (parâmetro *a*) e dificuldade (parâmetro *b*). Nesse caso, o programa assume como iguais os valores do parâmetro *a* para os grupos de comparação (homens e mulheres).

Como afirma Martínez Arias (1997), o uso deste procedimento pretende testar as seguintes hipóteses:

$$H_0: \Delta b = b_F - b_R = 0$$
$$H_1: \Delta b = b_F - b_R \neq 0.$$

Onde:

– Δb é o diferencial do parâmetro *b* entre os grupos de referência e focal;

– b_F é o valor do parâmetro *b* no grupo focal;

– b_R é o valor do parâmetro *b* no grupo de referência.

Para estabelecer a significância da diferença é necessário dispor do erro-padrão da diferença entre os parâmetros *b*, mediante o uso da fórmula:

$$s_{\Delta b} = \sqrt{s^2_{bF} + s^2_{bR}}$$

Onde:

– $s\Delta_b$ é o erro-padrão da diferença dos parâmetros *b* nos grupos focal e de referência;

– s^2_{bR} é a variância do parâmetro *b* no grupo de referência;

– s^2_{bF} é a variância do parâmetro *b* no grupo focal.

Neste âmbito, a prova de contraste é obtida dividindo-se a diferença dos parâmetros *b* pelo seu erro-padrão, isto é:

$$Z = \frac{\Delta b}{s_{\Delta b}}$$

Onde:

– Z é o resultado da prova de contraste;

– Δb é a diferença entre os parâmetros *b* do grupo focal e de referência;

– $s\Delta_b$ é o erro padrão da diferença entre os parâmetros *b* dos grupos focal e de referência.

Como Z tem uma distribuição aproximadamente normal, podem ser usadas as tabelas desta distribuição para comprovar a significância do valor obtido, através da comparação desse valor com o valor tabelado.

No nosso caso, o procedimento inicial para o estudo do DIF consistiu na determinação dos parâmetros métricos dos itens para a amostra total de estudantes. O valor obtido para o teste da bondade de ajuste do modelo (método de máxima verossimilitude) aos dados foi 25.633,96. Em seguida, os parâmetros dos itens foram determinados para os grupos de comparação (homens e mulheres). Deve ser dito que tais parâmetros têm que estar em uma mesma escala. O teste da bondade de ajuste do modelo (método de máxima verossimilitude) aos dados, considerando o grupo de referência (homens), resultou no valor 25.441,77.

A diferença entre os dois valores resultantes do teste da bondade de ajuste do modelo tem uma distribuição como *qui-quadrado* e, se é significativa, há evidência da existência de DIF no conjunto de itens. Assim, a diferença entre o valor inicial (25.633,96) e o final (25.441,77) resultou em 192,19. Dissemos que esse valor se distribui como *qui-quadrado*, com 29 graus de liberdade[3]. Assim, o valor $\chi^2_{(29)} = 192{,}19$, resultou significativo ($p < 0{,}01$). De acordo com esta informação, é plenamente justificável prosseguir o estudo de verificação do DIF no conjunto de 30 itens.

Nesse caso, a etapa seguinte consistiu no estudo do DIF para cada item individualmente, considerando o grupo de referência (homens) e o grupo focal (mulheres). A Tabela 1 apresenta os parâmetros *b* (dificuldade) dos itens, segundo os grupos de referência e focal.

[3] Os graus de liberdade são calculados com *n-1*, onde *n* é o número de itens estudados.

Tabela 1. Valores do Parâmetro *b* (dificuldade) dos 30 itens

Itens	Dificuldade dos itens (Parâmetro *b*)		Diferença entre os grupos	Erro-padrão da diferença
	Homens(GR)	Mulheres (GF)	(GF-GR)	
1	3,469	3,303	- 0,166	0,365
2	1,132	0,988	-0,145	0,174
3	0,004	-1,401	-1,405	0,151
4	-2,113	-2,327	-0,214	0,228
5	-1,818	-1,460	0,358	0,203
6	-1,264	-1,373	-0,109	0,201
7	-1,302	-1,625	-0,323	0,197
8	-0,844	-1,064	-0,220	0,203
9	-1,209	-1,023	0,186	0,204
10	-0,600	-0,145	0,455	0,189
11	-1,066	-0,913	0,153	0,185
12	-0,877	-1,625	-0,748	0,191
13	-0,210	0,035	0,245	0,193
14	-0,349	-0,021	0,328	0,186
15	-1,101	-0,550	0,551	0,182
16	-1,264	-1,105	0,159	0,198
17	-0,827	-0,483	0,344	0,186
18	-0,600	-0,927	-0,327	0,187
19	-0,287	0,428	0,716	0,191
20	-0,537	-0,671	-0,134	0,175
21	-1,014	-1,091	-0,077	0,183
22	-1,283	-1,358	-0,075	0,189
23	-0,088	-0,430	-0,342	0,184
24	-0,272	-0,281	-0,010	0,182
25	-0,134	0,134	0,268	0,187
26	-0,011	-0,159	-0,147	0,181
27	0,019	0,368	0,349	0,199
28	-0,241	-0,254	-0,013	0,174
29	0,126	0,309	0,183	0,177
30	-0,011	0,149	0,160	0,183

Podemos observar, inicialmente, que 16 itens (53,3%) foram mais fáceis para o grupo de referência já que, nesses casos, existem valores negativos no resultado da diferença do parâmetro *b* entre os grupos.

Com as diferenças dos parâmetros *b* entre os grupos e os seus respectivos erros padrões estimados, foi realizado o cálculo de Z para cada item, adotando-se $\alpha = 0,95$. Detectamos a presença de DIF em cinco itens (3, 10, 12, 15 e 19), isto é, em 16,7% das analogias destinadas à avaliação do raciocínio verbal. Destes cinco itens, três (10, 15 e 19) são favoráveis ao grupo focal e dois (3 e 12) ao grupo de referência. Nestes casos a hipótese nula (H_0) foi rejeitada, isto é, as diferenças entre os parâmetros *b* dos grupos de referência e focal foram estatisticamente significativas, adotando-se um intervalo de confiança de 95%. Em linguagem matemática podemos dizer que $b_{GF} - b_{GR} \neq 0$ para $p = 0,05$.

Como havíamos dito, Douglas, Roussos e Stout (1996) propuseram dois conceitos, que caracterizam diferentes tipos de DIF. No caso do DIF beneficiar o grupo de referência, isto é, quando $T_{jR}(\theta) > T_{jF}(\theta)$, é caracterizado o *DIF benigno*. O *DIF adverso* beneficia o grupo focal, isto é, quando $T_{jR}(\theta) < T_{jF}(\theta)$. Em nosso caso, temos dois itens com DIF benigno (3 e 12), ou seja, que proporcionam vantagem ao grupo de referência, e três com DIF adverso (10, 15 e 19), isto é, que favorecem ao grupo focal.

Considerações Finais

O estudo relatado não tinha o objetivo de identificar as causas do DIF, mas verificar sua existência entre 30 analogias componentes de um banco de itens destinados à avaliação do raciocínio verbal em estudantes do ensino médio. Como afirmam Cohen, Kim e Baker (1993), se trata de uma investigação para detectar o DIF, isto é, determinar a possível diferença entre as CCI's dos itens, de acordo com os grupos comparados, adotando, para tanto, o critério de comparação do parâmetro *b* (dificuldade) dos itens.

Assim mesmo, como modo de verificar a plausibilidade de se tratar de itens multidimensionais que, supostamente, é uma causa de DIF (Andriola, no prelo b), observamos as cargas fatoriais destes cinco itens no fator único extraído através de análise fatorial (método de máxima verossimilitude). Ditas cargas se situaram entre 0,328 e 0,582, ou seja, não podemos afirmar que tais itens sejam multidimensionais, dado o elevado valor de suas saturações no fator extraído. Não obstante, haveria que tentar identificar outras possíveis causas do DIF destes cinco itens, adotando outros procedimentos, tais como, a análise de conteúdo de ditos itens por expertos na área.

É necessário afirmar que a principal contribuição desta investigação para a área da avaliação psicológica e educativa foi identificar os itens com DIF, que são componentes de um banco já organizado e pronto para ser utilizado na avaliação psicológica de estudantes do ensino médio. Os resultados possibilitarão, desse modo, que estes cinco itens não sejam utilizados nos processos de avaliação do raciocínio verbal em dita população de estudantes. Finalmente, a modo de conclusão, queremos apresentar algumas palavras de Pasquali (2000), destacando a importância dos estudos sobre o DIF:

> "[...] essas informações sobre cada item de um teste [...] lhe permitem 1) tomar decisões sobre a qualidade do mesmo e 2) colocá-lo num banco de itens sem que ele perca a sua identidade, porque você tem uma série de indicadores individuais ou entradas nesta carteira de identidade do item, os quais o tornam inconfundível." (p. 129)

Referências

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. (no prelo a). Determinación del funcionamiento diferencial de ítems (DIF) destinados a la evaluación del razonamiento verbal considerando el tipo de escuela de los alumnos. *Bordón.*

Andriola, W. B. (no prelo b). Factores explicativos del funcionamiento diferencial de los ítems (DIF) componentes de un Test de Razonamiento Abstracto. *Psicología.*

Camilli, G. & Penfield, D. A. (1997). Variance estimation for differential test functioning based on Mantel-Haenszel statistics. *Journal of Educational Measurement, 34*, 123-139.

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.

Cohen, A. S., Kim, S. & Baker, F. B. (1993). Detection of differential item functioning in the graded response model. *Applied Psychological Measurement, 17*, 335-350.

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.

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

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

Hambleton, R. K. (1994). Item response theory: A broad psychometric framework for measurement advances. *Psichotema, 6*, 535-556.

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

Holland, P. W. & Wainer, H. (1993). *Differential item functioning*. New Jersey: Lawrence Erlbaum.

Jiang, H. & Stout, W. (1998). Improved type I error control and reduced estimation bias for DIF detection using SIBTEST. *Journal of Educational and Behavioral Statistics, 23*, 291-322.

Kim, S. & Cohen, A. S. (1998). Detection of differential item functioning under the graded response model with the likelihood ratio test. *Applied Psychological Measurement, 22*, 345-355.

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

Mazor, K. H., Hambleton, R. K. & Clauser, B. E. (1998). Multidimensional DIF analyses: The effects of matching on unidimensional subtest scores. *Applied Psychological Measurement, 22*, 357-367.

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

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: Psicología Pirámide.

Muñiz, J. & Hambleton, R. K. (1992). Medio siglo de Teoría de Respuestas a los Ítems. *Anuario de Psicologia, 52*, 41-66.

Oshima, T. C., Raju, N. S. & Flowers, C. P. (1997). Development and demonstration of multidimensional IRT-based internal measures of differential functioning of items and tests. *Journal of Educational Measurement, 34*, 253-272.

Oshima, T. C., Raju, N. S., Flowers, C. P. & Slinde, J. A. (1998). Differential Bundle functioning using the DFIT framework: Procedures for identifying possible sources of differential functioning. *Applied Measurement in Educational, 11*, 353-369.

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

Scheuneman, J. D. & Grima, A. (1997). Characteristics of quantitative word items associated with differential performance for female and black examines. *Applied Measurement in Education, 10*, 299-319.

Wainer, H. & Lukhele, R. (1997). Managing the influence of DIF from big items: The 1988 advanced placement history test as an example. *Applied Measurement in Education, 10*, 201-203.

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

Williams, V. S. L. (1997). The "unbiased" anchor: Bridging the gap between DIF and item bias. *Applied Measurement in Education, 10*, 253-267.

Zimowski, M. F., Muraki, E., Mislevy, R. J. & Bock, R. D. (1996). *BILOG-MG. Multiple Group IRT Analysis and Test Maintenance for Binary Items*. Chicago: Scientific Software International (SSI).

Recebido em 14.10.1999
Primiera revisão em 10.01.2000
Aceito em 14.03.2000

Sobre o autor

Wagner Bandeira Andriola é Professor do Mestrado em Avaliação Educacional da Universidade Federal do Ceará (UFC); Psicólogo pela Universidade Federal da Paraíba (UFPB); Especialista em Psicometria pela Universidade de Brasília (UnB); Mestre em Psicologia Social e do Trabalho pela Universidade de Brasília (UnB); Doutorando do Programa *Investigación, Diagnóstico y Evaluación para la Calidad Educativa* da *Universidad Complutense de Madrid (UCM)*; Bolsista da Fundação Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES). Atua nas áreas de Instrumentação e Medida Psicológica e Educativa utilizando as Teorias Clássica dos Testes (TCT) e de Resposta ao Item (TRI), com especial interesse pelo estudo do funcionamento diferencial dos itens (DIF) em testes psicológicos e educativos.

Núcleo de Estudos e Capacitação em Desenvolvimento Humano

Objetivo Geral: Implementar a formação em Desenvolvimento Humano de profissionais, técnicos, agentes de comunidade e estudantes das áreas de Educação e Saúde, através de discussões teórico-temáticas, dinâmicas pedagógicas e produção de material pedagógico.

NECADEH - CEP-RUA
Instituto de Psicologia
Rua Ramiro Barcelos, 2600, Sala 104
90035.003, Porto Alegre, RS
Fones: (51) 3165150/3309507 Fax: (51) 3304797