BREVE RETROSPECTIVA ACERCA DOS ESTUDOS QUE ABORDAM O VIÉS DOS ITENS DE TESTES PSICOLÓGICOS

A Brief Review of the Studies Concerning the Bias in the Items of Psychological Tests

Wagner Bandeira Andriola[1]

Resumo

O texto retrata a origem dos estudos sistemáticos que abordam a temática dos vieses dos itens e dos testes psicológicos, ressaltando os procedimentos mais freqüentemente utilizados em tal atividade. Destaca o conceito do termo *funcionamento diferencial do item (DIF)*, caracterizando-o como adverso ou benigno a um determinado grupo demográfico. Além do mais, enfatiza a existência do DIF *uniforme* ou *consistente* e do DIF *não-uniforme* ou *inconsistente*. Por fim, destaca a idéia central: a presença do DIF é um fator de injustiça à atividade de avaliação psicológica.

Palavras-chave: Testes psicológicos, funcionamento diferencial do item (DIF), avaliação psicológica.

Abstract

The text treats of the origin of systematic studies concerning the bias of items in psychological tests, giving prominence to the proceedings most frequently used in such activity. It emphasizes the concept of the term "differential functioning of the item" (DFI) characterizing it as adverse or benign to a specific demographic group. It also gives prominence to the existence of uniform or consistent DFI as well as non uniform or inconsistent DFI. It finally emphasizes the central idea: the presence of DFI is a factor of injustice in the work of psychological assessment.

Keywords: Psychological tests, differential functioning of the item (DFI), psychological assessment.

[1] Doutor em Educação pela *Universidad Complutense de Madrid.* Professor Adjunto do Departamento de Fundamentos da Educação da Faculdade de Educação da Universidade Federal do Ceará (UFC). Coordenador do Núcleo de Avaliação Educacional (NAVE) do Programa de Mestrado e Doutorado em Educação da UFC. Coordenador de Avaliação Institucional e Análise da Pró-Reitoria de Planejamento da UFC.
Endereço para contato: Rua Major Tibúrcio Cavalcante, 1222. Ap. 1701, Meireles. Fortaleza - CE. CEP 60125-100.
E-mail: w_andriola@yahoo.com

Origem dos estudos acerca do viés dos testes psicológicos

O viés dos instrumentos de medida psicológica e educacional é um tópico que aparece tardiamente tratado no seio da psicometria moderna (Muñiz, 1997). De acordo com Angoff (1993), seu estudo sistemático iniciou-se nos Estados Unidos, no final dos anos 1960, numa época em que estava em moda os debates acerca dos direitos civis e das desigualdades de oportunidade entre brancos e outras minorias étnicas.

Os resultados dos processos de avaliação educacional, executados por reconhecidas instituições, tais como o *Educational Testing Service (ETS)*, foram discutidos entre diversos intelectuais, tais como sociólogos e pedagogos. Segundo eles, as diferenças de rendimento observadas entre os diversos grupos étnicos e socioeconômicos refletiam, na realidade, disparidades nas oportunidades educacionais e discriminação contra grupos minoritários, tais como negros, hispano-americanos, judeus e árabes (Allen & Wainer, 1989). Podemos observar, desse modo, que foi a discussão social, alheia em grande parte ao círculo psicométrico especializado, que obrigou aos especialistas da área gerar novos procedimentos analíticos, com o objetivo de provar que seus testes ou instrumentos de medida não tinham nenhum tipo de viés (Cole, 1993).

Nessa mesma época, os investigadores começaram a preocupar-se pelo estudo sistemático das diferenças entre os mencionados grupos demográficos, pois estavam interessados em buscar explicações a respeito das suas verdadeiras causas explicativas. Martínez Arias (1997) destaca que a investigação acerca do viés dos itens pode remontar-se aos estudos realizados por A. Binet, em 1910, a respeito das diferenças de *status* socioecomômico no rendimento dos sujeitos submetidos a alguns testes desenvolvidos por ele próprio. Os resultados obtidos possibilitaram a proposição da hipótese de que o rendimento mais baixo destes sujeitos, em alguns itens, poderia dever-se ao efeito da cultura, ao invés de ser fruto de potenciais diferenças na capacidade mental ou no construto latente medido pelo teste (Andriola, 2002). Também W. Stern, o introdutor do termo *Quociente Intelectual*, pode ser considerado como um dos primeiros investigadores da área; estudou as diferenças relacionadas com a classe social, na Alemanha.

Apesar destes autores pioneiros, o começo da moderna investigação sobre o viés encontra-se nos trabalhos de K. Eells, A. Davis, R. J. Havighurst, V. E. Herrick e R. W. Tyler, que foram realizados na Universidade de Chicago, em 1951. Nestes estudos, os citados autores encontraram variações nos itens, em alguns aspectos muito peculiares, tais como conteúdo e formato, que reduziam ou exageravam as diferenças observadas entre os grupos comparados (Hambleton, Swaminathan & Rogers, 1991).

Nesse contexto, surgem os primeiros dados a respeito dos problemas técnicos presentes em alguns itens dos testes então utilizados na avaliação da aprendizagem. O uso indevido da linguagem escrita, que possibilitava certa vantagem de um grupo de sujeitos sobre outro era um desses problemas técnicos. Em suma: *muitos dos termos empregados nos testes eram mais familiares a alguns grupos específicos de estudantes, tais como os norte-americanos brancos, originários da classe média* (Linn & Harnisch, 1981). Em conseqüência, os sujeitos pertencentes aos grupos minoritários, que não conheciam ou não empregavam cotidianamente esses termos, tinham rendimento mais baixo. Surge, então, o interesse pela investigação sistemática do *viés dos itens* (Cole, 1993).

No âmbito da Teoria Clássica dos Testes (TCT), o termo *viés* é utilizado para rotular os itens que têm parâmetros de dificuldade ou de discriminação diferentes, nos distintos grupos estudados. Segundo Camilli & Shepard (1994), o viés é uma fonte de invalidez ou de erro sistemático, que se reflete em como um teste mede aos membros de um grupo particular. É sistemático porque cria uma distorção nos resultados do teste, favorável ou contrário aos membros de um grupo determinado.

Faz-se mister destacar: a idéia de *grupo* é central nas diversas definições de viés e, por esse motivo, este tem sido estudado, fundamentalmente, nas investigações acerca de diferenças relacionadas com algumas características grupais, tais como: classe social, idade, região, *habitat* ou outra característica sociodemográfica relevante (Andriola, 2002).

Procedimentos para a detecção do viés dos itens

No estudo sistemático do viés dos itens utilizam-se duas aproximações estatísticas. Uma

delas utiliza um critério externo ao teste e a outra um critério interno, normalmente, as pontuações ou escores totais obtidos (Whitmore & Shumacker, 1999). De acordo com Osterlind (1979, 1989), o viés externo é o grau em que as pontuações do teste têm correlações com variavéis irrelevantes para sua interpretação e alheias a este. Normalmente, ao falar do viés externo, se faz referência ao teste total e às conseqüências sociais de seu uso; o viés interno se refere às propriedades métricas dos itens dos testes. As técnicas que o detectam podem considerar-se como um tipo particular de análise de itens, que tentam responder à indagação: *itens de testes padronizados têm o mesmo comportamento estatístico para diferentes subgrupos de sujeitos extraídos da mesma população?*

Para Martínez Arias (1997), o termo viés interno tem um significado preciso, único, e se considera como um erro sistemático no proceso de medida; é um termo técnico, sem conotações sociais ou políticas. Os itens são considerados mais ou menos difíceis para um grupo particular, comparativamente ao rendimento de outros grupos extraídos da mesma população e com o mesmo nível de aptidão na variável latente. Os itens do teste se examinam por diferentes procedimentos ou métodos de detecção de viés. O objetivo é observar se conforma-se ou não, um conjunto de regras psicométricas para todas as pessoas de uma população, independentemente de que estas pertençam a grupos particulares dessa mesma população.

O viés dos itens pode ser inserido no contexto da validez de construto dos itens, isto é, o grau em que um item ou conjunto de itens mede uma mesma característica ou construto latente. No âmbito da Teoria de Resposta ao Item (TRI), a probabilidade de que um sujeito responda corretamente a um item se denomina *probabilidade de êxito*. O viés pode estudar-se comparando as probabilidades de êxito para diferentes grupos da mesma população (Angoff, 1993). Desse ponto de vista, um item é considerado não enviesado se a probabilidade de êxito é a mesma para sujeitos com igual aptidão, independentemente do grupo ao qual pertença. Por outro lado, um item enviesado será aquele em que as probabilidades de êxito são diferentes, apesar da igualdade dos sujeitos na capacidade avaliada. Não obstante, se dois sujeitos têm a mesma magnitude na variável latente medida por um item qualquer, poderíamos nos perguntar *a quê se pode dever que um item planejado para medir essa variável possa ter funcionamento diferencial, isto é, possa favorecer a um grupo determinado?*

Ercikan (1998) tenta responder a tal indagação, recordando-nos que na própria atividade de elaboração dos itens, surgem algumas possíveis causas ou fontes de viéses, devido, sobretudo, ao:

- Uso de termos conhecidos por grupos demográficos muito específicos;
- Uso de termos que têm distintas significações, segundo o contexto ou características grupais dos respondentes;
- Tamanho e complexidade da sentença empregada no enunciado ou nas alternativas.

Logicamente, os estudos que buscam as causas do DIF, a partir de variáveis demográficas, partem do suposto de que as fontes propostas por Ercikan (1998) não estão presentes nos itens analisados. No caso de inexistir segurança no cumprimento desse suposto, é aconselhável realizar-se algum estudo qualitativo prévio ao estudo do DIF, que esteja baseado na valoração dos itens por expertos na área (Angoff, 1993).

Muñiz (1997) tem a mesma opinião de Ercikan (1998) e Angoff (1993), já que para ele, o procedimento mais eficiente para evitar o viés dos itens é por meio de uma cuidadosa análise do seu conteúdo, por parte de vários expertos, antes de sua utilização definitiva. Realizada tal revisão e aplicados os itens aos sujeitos, ainda assim se devem executar certas análises estatísticas, que permitam identificar o funcionamento diferencial naqueles itens que *"escaparam"* da detecção pelo uso dos procedimentos prévios.

Alguns autores, entre os quais Camilli e Shepard (1994), insistem em que os índices estatísticos empregados na análise do DIF, por si mesmos, não proporcionam prova de viés, preferindo denominá-los *índices de discrepância* ou de *funcionamento diferencial*. Historicamente, o conceito de *viés* sempre esteve associado ao de DIF, ainda que ambas as definições sejam distintas (Hidalgo Montesinos, López Pina & Sánchez Meca, 1997). Para Cole e Moss (1989), o viés é uma possível causa do DIF, ou seja, as diferenças observadas no funcionamento do item são provocadas por algo irrelevante ao propósito do teste.

Segundo Camilli e Shepard (1994), o DIF engloba os diferentes procedimentos estatísticos para a detecção de um possível funcionamento

diferencial; insistem em que este não é sinônimo de viés, ainda que alguns autores parecem crer que sim. Os métodos estatísticos de DIF serão utilizados para identificar itens que exibem funcionamento diferencial para distintos grupos. Posteriormente, depois de uma análise lógica ou experimental, no contexto da validez de construto dos itens, se determinará quais deles estão enviesados, para que sejam eliminados do teste ou do banco de itens (Andriola, 1998).

Em outras palavras, os métodos DIF são procedimentos estatísticos e as análises de viés se situam no contexto mais geral da validade de construto, ainda que neste último se usem os resultados obtidos com a aplicação do primeiro. Como assinalam Camili e Shepard (1994) e Melenbergh (1989), os índices DIF às vezes produzem resultados estatisticamente significativos na ausência de viés, e às vezes não detectam o viés quando este se encontra presente em muitos itens, dada a circularidade do critério interno que utilizam.

Definição do termo Funcionamento Diferencial do Item (DIF)

Com o recente surgimento do paradigma psicométrico denominado *Teoria da Resposta ao Item (TRI)*, novas áreas de investigação têm proliferado (Andriola, 1998; Hambleton, 1989, 1990). Como opina Hambleton (1997), uma delas tem seu foco dirigido ao estudo do *Funcionamento Diferencial do Item (DIF)*, que caracteriza um dos mais graves problemas presentes nas atividades de avaliação educacional e psicológica, já que se trata de um fenômeno observado em muitos dos itens utilizados em testes de rendimento e psicológicos.

As investigações para a detecção do DIF têm por base uma mesma argumentação: *a existência do DIF é um fator que influencia a validez da interpretação, que é realizada a partir da pontuação obtida pelo sujeito num item ou teste* (Andriola, 2002). Não podemos esquecer que sobre a interpretação da pontuação, seja no âmbito psicológico ou educativo, reside toda a credibilidade e reputação da investigação e da avaliação (Downing & Haladyna, 1997). Assim, está plenamente justificada a relevância das pesquisas que estudam o DIF, sobretudo, aquelas que buscam identificar suas causas (Andriola, 2000).

Nesse âmbito, a necessidade e a 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). 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. É possível dizer, no âmbito da TRI, que o 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 pelo item (Lord, 1980; Melenbergh, 1989). Em linguagem matemática, poderíamos dizer que o item não tem DIF com respeito à variável G (grupo) dado Z (nível de θ) 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 segundo a variável *G*;
- *z* é o valor obtido segundo a variável *Z*.

Nesse contexto, os valores esperados (E) são dados por $E(X \mid g,\theta) = E(X \mid \theta)$ para todo g e θ. No caso de itens dicotômicos, os valores esperados são as probabilidades de acerto ao item, que podem ser expressas nos seguintes termos:

$P(X = 1 \mid g, \theta) = P(X = 1 \mid \theta)$ para todo g e θ. No segundo caso $[P(X = 1 \mid \theta)]$, a equação expressa, em realidade, a curva característica do item (CCI).

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 tem DIF se para valores iguais de q não correspondem valores iguais de P(θ) nas CCI's dos grupos considerados, isto é, quando $T_{jGR}(\theta) \neq T_{jGF}(\theta)$, onde:

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

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

De acordo com Oshima, Raju, Flowers e Slinde (1998), devemos ter em conta que a pontuação verdadeira em um teste unidimensional, composto por *k* itens, é expressada pela fórmula:

$$T_S = \sum_{i=1}^{K} P_i(\theta_S)$$

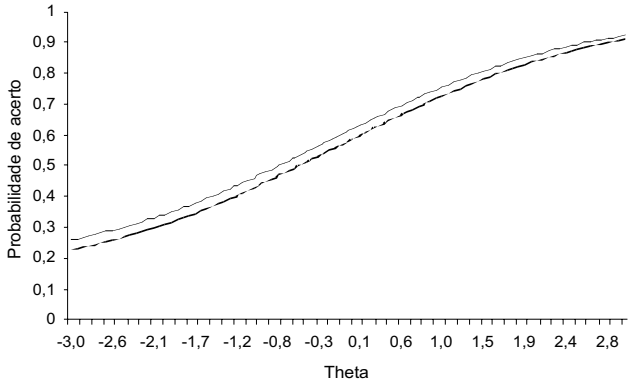
Onde:

• $P_i(\theta_s)$ é a probabilidade de acertar ao item *i* pelo sujeito *s* com a habilidade θ_s.

Segundo Mazor, Hambleton e Clauser (1998), o uso do número de respostas corretas para a determinação do DIF, isto é, a pontuação verdadeira no teste ou item, só é aceitável no caso do teste ser unidimensional e, ademais, se as respostas forem dicotômicas.

Para visualizar o DIF de um hipotético item, apresentamos, a seguir, a figura 1.

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

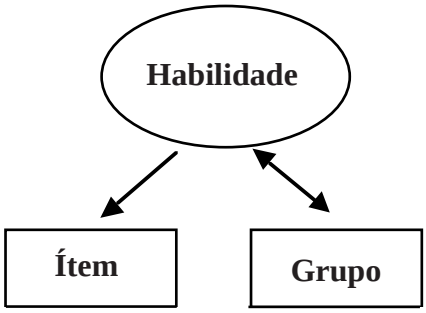


Legenda: Linha superior: CCI das mulheres; Linha inferior: CCI dos homens.

Observamos que, para uma mesma magnitude θ, o valor P(θ) é sempre superior para as mulheres (curva superior), ou seja, em níveis iguais de competência na variável medida θ não correspondem probabilidades iguais de superar o item. Neste caso, o item está enviesado *contra* os homens (GR), pois os valores P(θ) para um mesmo nível θ são sempre maiores para as mulheres (GF). Por exemplo, para θ = 1,4 temos valores aproximados de P(θ) = 0,75 para os homens e P(θ) = 0,80 para as mulheres.

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 $T_{jGR}(\theta) > T_{jGF}(\theta)$, caracteriza-se a existência de *DIF benigno*. O *DIF adverso* ocorre no caso do DIF beneficiar o grupo focal, ou seja, quando $T_{jGR}(\theta) < T_{jGF}(\theta)$. No exemplo da figura 1, temos um caso de DIF adverso. Utilizando o mesmo item, aclaremos o que ocorre na ausência de DIF, observando a figura 2.[2]

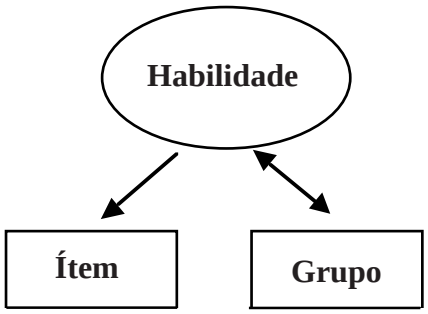
Figura 2. Relação entre habilidade, item e grupo na ausência de DIF.



A elipse indica a habilidade ou construto latente, que tem relação causal com o item. Grupo e variável latente estão associados. Em outras palavras e a título de exemplificação, poderíamos dizer que as mulheres têm elevada habilidade na variável latente e que esta variável tem relação causal com o item, isto é, o grupo com maior capacidade na variável latente æ as mulheres æ 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, trata-se de um item sem DIF. Agora, observemos o que ocorre no caso de um item com DIF, observando a figura 3.[3]

[2] Adaptado de ANDRIOLA (2002).
[3] Adaptado de ANDRIOLA (2002).

Figura 3. Relação entre habilidade, item e grupo na presença de DIF.



No caso da figura 3, temos a mesma situação descrita na figura 2, adicionado do fato de haver associação ou interação entre grupo e item. Neste segundo caso, a associação entre ambas pode favorecer o rendimento superior de um grupo sobre o outro devido, sobretudo, a algumas características demográficas específicas tais como: gênero, raça, *background* educativo, origem social, etc. (Clauser, Nungester & Swaminathan, 1996). Deve ser mencionado que, neste caso, supõe-se que a magnitude da variável latente está sendo controlada, ou seja, os sujeitos são comparados com respeito ao seu rendimento, considerando-se que possuem a mesma aptidão. Este segundo exemplo caracteriza o caso em que o rendimento no item não depende somente da magnitude que os indivíduos tenham na variável latente, senão que também depende das características do grupo, ou seja, trata-se de um item com DIF. Em nosso exemplo, a característica do grupo que afeta o rendimento diferencial no item é o fato do sujeito ser homem ou mulher, isto é, trata-se de uma característica de natureza demográfica, que afeta sistematicamente as respostas dos sujeitos de mesma habilidade.

Portanto, é necessário reconhecer que a presença de 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 comparativo do rendimento dos homens e mulheres. Muñiz (1997) chama a atenção para o fato de que tal problema pode ter repercussões sociais mais graves se é, precisamente, a cultura dominante que elabora os itens, para avaliar os demais sujeitos oriundos de outras culturas. 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 desses tipos de escolas são, geralmente, oriundos de classes sociais distintas, com diferentes bagagens culturais, sociais, econômicas, etc. (Andriola, 1997 a). Todos esses aspectos contribuem para que um tipo de aluno tenha o vocabulário mais rico que o outro. Como o raciocínio verbal é medido pelos itens que utilizam palavras, muito provavelmente aquele tipo de aluno que conheça melhor o vocabulário utilizado nos itens terá uma clara vantagem na resolução destes mesmos itens (Andriola & Pasquali, 1995).

Em síntese, argumentamos que, dada a grande variabilidade de todos esses antecedentes históricos dos sujeitos implicados na avaliação do raciocínio verbal, se o item ou teste, em geral, se apóia mais nos antecedentes de uma cultura que nos da outra, terá altíssima probabilidade de não ser eqüitativo, de estar enviesado. Em outras palavras, se confunde o efeito da capacidade de raciocínio verbal (construto principal) com o conhecimento vocabular (construto secundário), isto é, se um aluno pontua baixo no teste não saberemos, ao certo, se devemos atribuí-lo a sua baixa capacidade de raciocínio verbal ou ao seu baixo conhecimento vocabular. Como nos fala Muñiz (1997), a casuística é interminável e se pode dizer que não existem provas inteiramente isentas de viés. Trata-se, assim, de detectar a quantidade de viés que pode ser aceitável em um teste ou item.

Finalmente, deve ser mencionado que, nesse contexto, a importância dos estudos que objetivam a verificação do DIF está plenamente justificada. Cabe ao avaliador verificar se em seu teste existem itens com DIF, para que (i) possa buscar as causas que o expliquem, (ii) evitar sua utilização com o grupo em desvantagem e, finalmente, (iii) controlar os fatores responsáveis pelo DIF para evitar, desse modo, construir novos itens com o mesmo viés (Hambleton, 1989; Mislevy, 1996).

Anteriormente, destacamos o intensivo uso de testes no contexto norte-americano. Permanece, ainda hoje, a importância desses instrumentos, sobretudo na área que investiga o Funcionamento Diferencial do Item (DIF). Para termos noção dessa relevância, apresentamos parte do discurso proferido por Ree (1993), no renomado *Educational Testing Service (ETS)*:

> *The American military is a unique position because we develop and use our own tests. This year, more than two million young men and women will be tested for enlistment qualification and additional hundreds of thousands of tests will be administered by the service for promotion and certification purposes. These tests will material affect the lives of these military members and the security of our country.* [...] *For the Air Force, which produces both enlisted and officer tests, certain models of DIF detection have become an integral part of the test production and evaluation procedure. Additionally, content and construct validation of our many tests benefit from DIF analyses* (pp. xi-xii)[4].

Obviamente, o estudo do Funcionamento Diferencial do Item (DIF) também é relevante no campo da avaliação educacional, pois como destacam Hartle e Bataglia (1993):

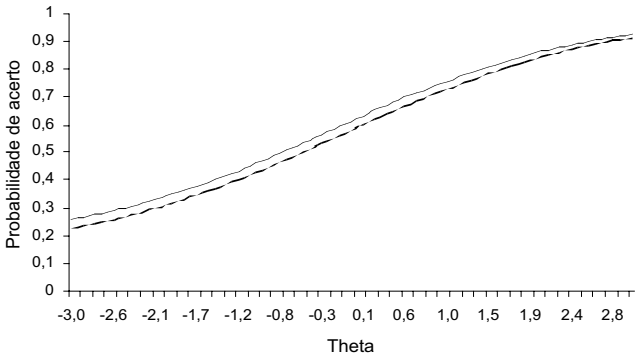
> *The low test scores of minorities and women are a particular problem for federal policymakers. Most federal education programs are designed to increase educational opportunities for disadvantaged groups.* [...] *On the other hand, it is also likely to increase interest in new assessment techniques that do not have a disproportionate racial impact* (pág. 305).[5]

Apresentamos, agora, os dois tipos de Funcionamento Diferencial do Item (DIF) observados no âmbito da Teoria da Resposta ao Item (TRI).

Descrição dos Tipos de DIF no Âmbito da TRI

Enfatizamos que, no contexto da TRI, a lógica para a detecção do DIF consiste em comparar as CCI's dos itens, considerando os grupos de referência (GR) e focal (GF), através da utilização de métodos apropriados. Os distintos métodos para a detecção do DIF foram desenvolvidos com base nos vários tipos de DIF (Bock, 1993). O mais conhecido é denominado *DIF uniforme* ou *consistente*, e é observado quando as CCI's do item estudado para o GR e do GF são diferentes, mas não se cruzam, isto é, são paralelas. Em outras palavras, quando existe uma vantagem relativa para um dos grupos estudados, cujo valor é constante ao largo do contínuo da habilidade. Este caso ocorre quando o parâmetro *a* não tem o mesmo valor nas duas CCI's, isto é, quando são paralelas, conforme está representado na figura 4.

Figura 4. Representação de um item com DIF uniforme.



A figura 4 ilustra o caso de diferenças nos parâmetros *b* e *c* para os dois grupos estudados. Observamos que a CCI do grupo focal está situada mais a esquerda que a do grupo de referência, o que indica que o item é mais fácil para o grupo focal, já que $P_{GF}(q) > P_{GR}(q)$ Ω q. Essa diferença supõe que o item tem DIF.

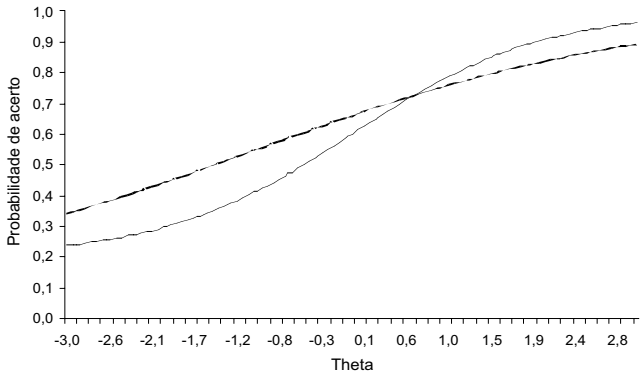
O segundo tipo de DIF é denominado *DIF não uniforme* ou *inconsistente*, e se observa quando as CCI's do item estudado com respeito aos grupos de referência e focal são diferentes e, ademais, se cruzam em algum ponto do contínuo da habilidade, isto é, não são paralelas. Em outras

[4] Nós, militares americanos, temos uma mesma opinião porque desenvolvemos e utilizamos nossos próprios testes. Este ano, mais de 2 milhões de jovens serão testados no processo de seleção e milhares de testes serão utilizados pelo serviço de promoção e certificação dos resultados. Estes testes afetam a vida desses militares e a segurança do nosso país. [...] Para a Aeronáutica, que seleciona os alistados e funcionários, o uso dos modelos de detecção do DIF é conveniente ao processo de produção e avaliação. Ademais, os benefícios da validação de conteúdo e construto de muitos de nossos testes resultam da análise do DIF.

[5] Os baixos resultados obtidos pelas minorias e mulheres são um grave problema para as políticas públicas nacionais. Muitos programas nacionais são propostos para incrementar as oportunidades educativas para os grupos em desvantagem. [...] Por outro lado, também é necessário incrementar o interesse por novas técnicas de avaliação que não causem um desproporcionado impacto racial.

palavras, existe uma vantagem relativa para um dos grupos investigados, cujo valor é variável ao largo de toda a habilidade. Esse caso ocorre quando os parâmetros *a*, *b* ou *c* têm valores distintos nas duas CCI's, ou seja, quando não são paralelas. Esse tipo de DIF está representado na figura 5.

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



É necessário dizer que neste segundo tipo de DIF é inapropriado examinar globalmente os dados, porque tal procedimento poderia ocultar sua presença, pois a peculiar variabilidade do DIF, que se verifica em distintas zonas da variável latente, pode cancelar total ou parcialmente sua detecção (Martínez Arias, 1997). É o caso de não se 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 características comuns (Douglas, Roussos & Stout, 1996).

Autores como Cohen, Kim e Baker (1993) distinguem diferentes pesquisas sobre o DIF, segundo os objetivos pretendidos, isto é, pode-se falar de estudos para a *detecção do DIF* e outros para a *descrição do impacto do DIF*. No primeiro grupo, estão as investigações que empregam algum método tradicional para a identificação do DIF. Nesse caso, os estudos objetivam somente detectar o DIF, ou seja, determinar a possível diferença entre as curvas características dos itens (CCI's) de acordo com os grupos comparados. No segundo grupo, estão as investigações realizadas com o objetivo de identificar as causas do DIF. Nesse âmbito, segundo Clauser, Nungester e Swaminathan (1996), o objetivo do investigador, que utiliza as distintas técnicas para a detecção do DIF, é tentar saber quais são as razões (psicológicas, educacionais, culturais, sociais, atitudinais, etc.) que, teoricamente, acarretam o funcionamento diferencial do item.

À guisa de conclusão

Destacamos ser bastante comum o fato de itens componentes de testes possuírem algum tipo de viés, dentre os quais o DIF. Como enfatizamos anteriormente, o DIF ocasiona sérios problemas às avaliações educacionais. Trata-se de um fator de injustiça para alguns grupos de respondentes, já que os alunos que possuem o mesmo grau de aprendizagem e que provêm de distintos grupos demográficos têm distintas probabilidades de acertar um mesmo item. Portanto, devemos reconhecer a relevância das investigações acerca do DIF, já que podem proporcionar maior eqüidade aos processos de avaliação educacional, pela identificação e não utilização de itens com algum tipo de DIF. Por fim, devemos enfatizar a opinião de Andriola (2002):

> *(...) los estudios para detectar el DIF – mezclando procedimientos cualitativos y cuantitativos - deben ser efectuados inmediatamente después de la elaboración de los ítems que compondrán dichos instrumentos de medida* (pág. 554).

Ademais, é preciso fazer menção ao fato de que a área de investigação do DIF, no âmbito educativo e psicológico, é recente, necessitando de boas hipóteses, fundamentadas em teorias científicas, que tentem "abrir novas perspectivas" aos estudos do DIF (Cole, 1993; Hambleton, 1997; Roznowski & Reith, 1999; Scheuneman & Gerritz, 1990). Como opina Bond (1993): *In general, howewer, theories about why items behave differentially across groups can be described only as primitive* (pág. 278)[6].

Schmitt, Holland e Dorans (1993) acreditam que a área que investiga o DIF não tem progredido no grau desejado em virtude de três fatores:

[6] De um modo geral, teorias acerca do porquê os itens funcionam diferentemente para certos grupos podem ser descritas como primitivas.

• Porque as investigações acerca do DIF são relativamente recentes e, atualmente, a ênfase está no desenvolvimento de métodos estatísticos para sua identificação. Por exemplo, as modernas técnicas para a detecção do *funcionamento diferencial das alternativas (DAF)* têm o mesmo objetivo das técnicas DIF, isto é, compreender as causas da escolha diferenciada das alternativas de um item por sujeitos que têm o mesmo nível de habilidade, mas fazem parte de distintos grupos demográficos (Thissen, Steinberg & Wainer, 1993; Thissen, Steinberg & Fitzpatrick, 1989);

• Porque a identificação do DIF e os fatores a ele relacionados necessitam boas teorias sobre a dificuldade diferencial dos itens, em um campo no qual as teorias sobre os processos cognitivos presentes na resolução dos itens não se encontram, todavia, minimamente avançadas;

• Porque a identificação e descrição dos citados processos cognitivos é muito complexa, já que intervêm múltiplos fatores. Ademais, é um campo de investigação que exige o trabalho multidisciplinar de psicólogos, pedagogos e matemáticos, algo extremamente difícil no estágio atual de desenvolvimento investigativo brasileiro.

Devemos dizer que o processo de criação de boas hipóteses explicativas do DIF deverá, logicamente, ser árduo, difícil e frustrante. As hipóteses deverão sofrer corroborações e rejeições, algo bastante comum à atividade científica (Wilson, 1999). Assim, verificamos que apesar de existir grande variedade de métodos para investigar o DIF, os mesmos padecem limitações. Autores mais críticos aconselham 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.

Ademais, devemos ter claro que a presença do DIF em itens de instrumentos de 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 aos processos avaliativos; permite aos sujeitos que possuem mesmo grau na variável latente ou construto medido pelo item obter melhores resultados, já que esses têm maiores probabilidades de acertá-lo.

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 verificar a presença de DIF e, assim, evitar utilizá-los em processos avaliativos (Andriola, 1998).

Para finalizar, mencionaremos célebre frase latina que é muito sugestiva e sintetiza, na nossa opinião, a importância dos estudos acerca do DIF no âmbito da avaliação psicológica e educacional: *sátius est inítio medéri, quam fini* (é melhor remediar no princípio do que no fim).

Referências

Andriola, W. B. (1997). Avaliação do raciocínio verbal em estudantes do 2º grau. **Estudos de Psicologia** , v.2, n.2, p. 277-285.

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**, v.11, n.2, p. 295-308.

Andriola, W. B. (2000). Funcionamento Diferencial do Item (DIF): estudo com analogias para medir o raciocínio verbal. **Psicologia: Reflexão e Crítica**, v.13, n.3, p. 473-481.

Andriola, W. B. **Detección del funcionamiento diferencial del ítem (DIF) en tests de rendimiento. Aportaciones teóricas y metodológicas.** Madrid, 2002. Tese de Doutorado. Universidad Complutense de Madrid, Madrid.

Andriola, W. B. & Pasquali, L. (1995). A construção de um Teste de Raciocínio Verbal (RV). **Psicologia: Reflexão e Crítica**, v.8, n.1, p. 51-72.

Allen, N. L. & Wainer, H. (1989). **Nonresponse in Declared Ethnicity and the Identification of Differential Functioning Items. Technical Reports Nº 89-89.** New Jersey: Educational Testing Service (ETS).

Angoff, W. H. (1993). Perspectives on Differential Item Functioning. In P. W. Holland & H. Wainer (Ed.), **Differential Item Functioning** (pp. 3-23). New Jersey: Lawrence Erlbaum Associates.

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

Bock, R. D. (1993). Different DIF's: Comment on the Papers Read by Neil Dorans and David Thissen. In P. W. Holland & H. Wainer (Ed.), **Differential Item Functioning** (pp. 115-122). New Jersey: Lawrence Erlbaum Associates.

Camilli, G. & Shepard, L. A. (1994). **MMSS. Methods for Identifying Biased Test Items.** California: SAGE Publications.

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**, v.33, n. 4, p. 453-464.

Cohen, A. S., Kim, S. & Baker, F. B. (1993). Detection of differential item functioning in the graded response model. **Applied Psychological Measurement**, v.17, n.4, p. 335-350.

Cole, N. S. (1993). History and Development of DIF. In P. W. Holland & H. Wainer (Ed.), **Differential Item Functioning** (pp. 25-29). New Jersey: Lawrence Erlbaum Associates.

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**, v.33, n.4, p. 465-484.

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

Hambleton, R. K. (1997). Perspectivas futuras y aplicaciones. In J. Muñiz, **Introducción a la Teoría de Respuesta a los Ítems** (pp. 203-213). Madrid: Ediciones Psicología Pirámide.

Hambleton, R. K. (1990). Item response theory: introduction and bibliography. **Psicothema**, v.II, n.1, p. 97-107.

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

Hambleton, R. K., Swaminathan, H. & Rogers, H. J. (1991). **Fundamentals of Item Response Theory.** North Caroline: Sage Publications.

Hartle, T. W. & Battaglia, P. A. (1993). The Federal Role in standardized Testing. In R. E. Bennett & W. C. Ward (Org.), **Construction versus Multiple Choice Items in Cognitive Measurement** (pp. 291-311). New Jersey: Lawrence Erlbaum Associates.

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

Lord, F. M. (1980). **Applications of Item Response Theory to Practical Testing Problems.** New Jersey: Lawrence Erlbaum Associates.

Martínez Arias, R.(1997). **Psicometría. Teoría de los Tests Psicológicos y Educativos.** Madrid: Ediciones Síntesis.

Mellenbergh, G. J. (1989). Item bias and item response theory. **International Journal of Educational Research,** v.13, n.2, p. 127-143.

Muñiz, J**.** (1997). **Introducción a la Teoría de Respuesta a los Ítems.** Madrid: Ediciones Psicología Pirámide.

Muñiz, J. (1994). **Teoría Clásica de los Tests.** Madrid: Ediciones Pirámide S.A..

Osterlind, S. J. (1979). **Test item bias.** Beverly Hills: Sage Publications.

Osterlind, S. J. (1989). **Constructing test items.** Boston: Kluwe r Publications.

Pasquali, L. (2000). **Psicometria: Teoria dos Testes Psicológicos.** Brasília: Prática Gráfica e Editora Ltda.

Ree, M. J. (1993). Foreword: Differential Item Functioning (DIF): A perspective from the Air Force Human Resources Laboratory. In P. W. Holland & H. Wainer (Ed.), **Differential Item Functioning** (pp. xi-xii). New Jersey: Lawrence Erlbaum Associates.

Roznowski, M. & Reith, J. (1999). Examining the measurement quality of tests containing differentially functioning items: Do biased items result in poor measurement? **Educational and Psychological Measurement**, v.52, n.2, p. 248-269.

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**, v. 27, n.2, p.109-131.

Schmitt, A. P., Holland, P. W. & Dorans, N. J. (1993). Evaluating Hyphoteses about Differential Item Functioning. In P. W. Holland & H. Wainer (Ed.), **Differential Item Functioning** (pp. 281-319). New Jersey: Lawrence Erlbaum Associates.

Whitmore, M. L. & Shumacker, R. E. (1999). A comparison of logistic regresion and analysis de variance differential item functioning detection methods. **Educational and Psychological Measurement**, v.59, n.6, p. 910-927.

Wilson, E. O. (1999). **Consilience. La Unidad del Conocimiento.** Barcelona: Ediciones Galaxia Gutemberg.

Recebido em/ *received in*: 07/07/2005
Aprovado em/ *approved in*: 19/09/2005