O funcionamento diferencial dos itens

Fermino Fernandes Sisto[1]

Resumo
A dificuldade de um item deveria ser uma propriedade independentemente de uma população em particular, da mesma forma que a habilidade de uma pessoa deveria ser uma característica sem referência a qualquer conjunto particular de itens. Nesse contexto, a análise do *funcionamento diferencial dos itens* (DIF) busca detectar os itens cuja probabilidade de acertos difere entre distintos grupos, cujas pessoas possuem o mesmo nível de habilidade na variável medida. Para a análise dessa questão são discutidos principalmente, os tipos de abordagem para sua detecção e as propostas de procedimentos para identificação de DIF. Além disso, é comentado que o modelo de Rasch fornece uma estrutura que permite a comparação dos resultados observados com os previstos, avaliando a probabilidade do resultado observado, equacionando problemas não solucionados pelos outros modelos.
Palavras-chave: Viés; Modelo Rasch; Habilidade; Item.

Differential item functioning

Abstract
The item difficulty should be a property irrespective of a particular population, as well as the ability of a person should be a characteristic without relation to any particular set of items. In this context, the analysis of the *differential item functioning* (DIF) searches for detecting items whose probability of correctness differs between distinct groups, whose people possess the same level of ability in the variable measured. For the analysis of this question, the types of approaching for its detention and the procedures for DIF identification are mainly argued. Moreover, it is argued that the Rasch model supplies a structure that allows the comparison of the observed results with the predict ones, evaluating the probability of the observed result, equating problems not solved by the other models.
Keywords: Bias; Rasch Model; Ability; Item.

No processo de medição de características ou habilidades psicológicas pelo menos três elementos estão presentes. Um deles se refere à variável que se pretende medir (por exemplo, habilidade mecânica); o outro está consubstanciado no instrumento construído para medi-la (teste), que nada mais é que um conjunto variável de elementos (itens do teste); e, o último, as respostas que as pessoas dão a esses itens.

Um teste proporcionará medidas de pessoas com uma margem de erro pequena quando atender aos requisitos psicométricos de precisão e validade. No contexto das evidências de validade das medidas é possível avaliá-las considerando o teste, ou cada um de seus itens.

Numa avaliação sempre se deseja obter medidas para qualquer pessoa em algum traço ou habilidade (tal como a compreensão da leitura). Mais especificamente, deseja-se conhecer a posição da pessoa em algum traço ou habilidade. Essa circunstância simples tem, pelo menos, uma implicação importante. Significa que se considera teste apenas quando os itens formam um conjunto homogêneo. Isso exclui, por exemplo, os testes que combinam itens de matemática e leitura em uma única pontuação, porque uma pontuação 12 pode ser produto de seis pontos obtidos nos itens de matemática e seis em leitura, ou 10 em leitura e dois em matemática; em outros termos, uma pontuação dessa natureza é difícil de interpretar, pois sua composição não é unidimensional e não se sabe a contribuição de cada um nem o que significam as diferentes combinações das duas variáveis envolvidas (matemática e leitura).

Ao lado disso, um bom modelo de medida deve requerer, pelo menos, que um teste válido satisfaça três aspectos. Um deles é que uma pessoa mais capaz (ou com maior quantidade de um traço) tenha sempre mais possibilidade de acerto em um item do que uma pessoa menos capaz. Um outro seria que toda a pessoa tenha mais possibilidade de acerto em um item fácil (ou comum) do que em um difícil (ou raro). E, finalmente,

[1] Endereço para correspondência:
Universidade São Francisco – USF – Programa de Pós-Graduação *Stricto Sensu* em Psicologia
Rua Alexandre Rodrigues Barbosa, 45 – 13251-900 – Itatiba – SP
E-mail: fermino.sisto@saofrancisco.edu.br

que essas condições devem ser produto da pessoa e da posição do item no traço ou habilidade e não devem se alicerçar na raça, sexo, entre outras características da pessoa analisada.

Está implícita nessas condições a concepção de que a dificuldade de um item é uma propriedade inerente a e exclusiva do item em qualquer circunstância sem referência a qualquer população em particular. Analogamente, a habilidade de uma pessoa é definida como uma característica da pessoa sem referência a qualquer conjunto particular de itens. Isso traz consigo a concepção de uma interação entre a pessoa e um item dependente de dois e somente dois parâmetros, quais sejam, a habilidade da pessoa e a dificuldade do item. Novamente excluem-se os testes que combinam itens de matemática e leitura, por exemplo, em uma única pontuação, porque uma pessoa com alta pontuação em leitura, mas baixa na matemática poderia acertar os itens difíceis de leitura e errar os itens fáceis de matemática, entre outras possibilidades.

Pode-se afirmar que um teste válido fornecerá medidas idênticas para pessoas com níveis semelhantes da habilidade que o teste se propõe medir. Entretanto, a pontuação obtida é função não só do nível dos sujeitos na variável medida, mas também de outras características irrelevantes como pertencer a diferentes grupos étnicos, culturais, entre outros. Nesse caso, fala-se de funcionamento diferencial do teste (FDT). Tal tipo de interferência constitui uma evidente e definida causa de invalidade do instrumento.

A expressão *funcionamento diferencial dos itens* (DIF) ou como denominado em inglês, *differential item functioning* (DIF) se refere ao item. Busca detectar os itens cuja probabilidade de acertos difere entre distintos subgrupos de uma população dada, cujas pessoas possuem o mesmo nível de habilidade na variável medida. Os estudos com vistas a detectar ou neutralizar esse efeito é de importância indiscutível em razão das implicações éticas, sociais e jurídicas envolvidas na utilização de testes que podem subestimar sistematicamente as capacidades de certos grupos em função de características diferenciadoras irrelevantes para a habilidade em questão.

A importância dos estudos que objetivam a verificação do DIF é justificada porque cabe ao autor do teste ou ao autor do sistema de avaliação ou ao pesquisador verificar se existem itens com DIF para (1) que se possam buscar as causas, (2) evitar sua utilização no grupo em desvantagem e, finalmente, (3) controlar os fatores responsáveis pelo DIF para evitar construir novos itens com o mesmo problema (Andriola, 2001; Hambleton, 1989).

O tema não é novo. Binet, em 1910, ao se deparar com crianças de *status* socioeconômico mais baixo com rendimento pior em alguns itens de seu teste, aventou a possibilidade de que esses itens poderiam estar medindo efeitos de aprendizagem cultural em vez de capacidade mental. Acrescente-se, também, que em 1914, William Stern, o introdutor do termo quociente intelectual, apontou que os testes, na Alemanha, poderiam favorecer uma classe social em detrimento de outras.

Entretanto, o começo da moderna investigação sobre o viés se encontra no trabalho de Eells, Davis, Havighurst, Herrick e Tyler (1951), no qual defenderam que as variações dos itens em relação a conteúdo, formato, por exemplo, atenuavam ou exageravam as diferenças entre os grupos. Nesse estudo, analisaram mais de 650 itens de oito testes de inteligência para detectar os itens que, por familiaridade para certos grupos socioeconômicos, poderiam estar refletindo muito mais as oportunidades de aprendizagens do que a variável que pretendia medir. Esse tipo de viés foi denominado de viés cultural. Em decorrência, os testes deixaram de ser vistos como instrumentos neutros e se questionou se as diferenças socioeconômicas e raciais que se encontravam nos testes de rendimento e aptidões eram reais e não produto dos testes utilizados, pois eram construídos pela classe e raça econômica e politicamente dominantes.

O problema do viés nos instrumentos de medida se tornou um importante tema da literatura psicométrica, educacional e de avaliação desde finais dos anos 60 do século passado. A importância adquirida está ligada ao aparecimento dos diversos movimentos pelos direitos civis em EEUU. Grupos fizeram reivindicações pela igualdade de direitos, pois se consideravam tratados injustamente em situações de seleção para postos de trabalho e de admissão para o ensino superior, quando as decisões se baseavam em testes psicométricos. O argumento apresentado tinha como fundamento as diferentes taxas de admissão de grupos diferenciados por etnia, sexo, *status* socioeconômico, dentre outros, o que sugeriria a possibilidade de vieses nos testes utilizados. Foi novamente trazido para discussão se as diferenças nos resultados de testes de aptidões e rendimento entre diferentes grupos refletiriam diferenças reais entre os grupos ou seriam causadas por fontes sistemáticas de variância alheias ao constructo em questão.

Nesse contexto, é necessário citar também a obra de Jensen (1980), na qual além de relatar dados sobre diferenças de gênero e raça no rendimento nos testes de habilidades, apresenta definições claras e precisas do que se entende por viés, distanciando-o do conceito de justiça e colocando-o no campo da ética. Para Jensen, na psicometria, o viés está relacionado a erros

sistemáticos e interferem na validade preditiva ou na validade de constructo para pessoas pertencentes a grupos particulares de uma mesma população.

O viés dos itens pode ser situado no contexto da validade de constructo dos itens, isto é, o grau em que um item ou conjunto de itens mede um traço ou constructo (AERA, APA & NCME, 1985).

Algumas demarcações

O termo *funcionamento diferencial dos itens* (DIF) foi cunhado por Holland e Thayer (1988), em substituição ao termo viés, que vinha causando dubialidade de sentido e discussões éticas e morais, por exemplo. Essa expressão é a usada correntemente na literatura psicométrica e dizer que um item apresenta DIF significa que ele mostra diferentes propriedades estatísticas em razão de diferentes grupos.

Freqüentemente o DIF é confundido com outros conceitos. Um deles é denominado de impacto adverso, ou simplesmente impacto. Camilli e Shepard (1994) explicitaram essa distinção com o exemplo das diferenças favoráveis aos rapazes freqüentemente encontradas em muitos testes de matemática. As estudantes estadunidenses, via de regra, fazem menos cursos de matemática, mas é possível interpretar que essas diferenças se referem muito mais a conhecimentos adquiridos e não a autênticos vieses dos testes. Dessa forma, as diferenças entre os grupos não são em si mesmas evidências de viés do teste.

Angoff (1982, 1993) e Camilli (1993) e Camilli e Shepard (1994) insistiram em que os índices estatísticos empregados na análise do viés por si mesmos não proporcionam prova de viés, preferindo denominá-los índices de discrepância do item (Angoff, 1982) ou de funcionamento diferencial do item (DIF) (Camilli, 1993, Camilli & Shepard, 1994; Holland & Thayer, 1988). Este último termo engloba os diferentes procedimentos estatísticos para a detecção de um possível funcionamento diferencial, mas defenderam que essa informação não seria sinônima de viés. Os índices estatísticos de DIF serviriam para identificar os itens com funcionamento diferencial em distintos grupos e apenas com base em uma análise lógica ou experimental no contexto da validez de constructo dos itens é que se poderia determinar quais deles estariam enviesados.

Entre os investigadores não expertos em psicometria às vezes se confunde o viés com diferenças reais no rendimento dos grupos. Não obstante, é imprescindível diferenciar esses dois aspectos. A separação dessas duas fontes de variância foi o objetivo dos primeiros investigadores sobre o viés e o que os levou a conceitualizá-lo como dificuldade diferencial ou incremental do item. Calculavam-se os índices de dificuldade para cada grupo e se consideravam como potencialmente enviesados aqueles itens com as maiores diferenças entre os índices de dificuldade (Eells & colaboradores, 1951).

O conceito de grupo é central nas definições de viés. Os grupos mais estudados são os relacionados à etnia e ao gênero. No entanto, pode ou deveria ser estudado também em relação a outros grupos como classe social, idade, religião, ou qualquer outra característica sociodemográfica dos sujeitos. Geralmente as pesquisas analisam dois grupos, um denominado grupo focal e, o outro, de referência. O grupo focal normalmente é minoritário e considerado prejudicado pelo teste em relação ao grupo de referência o qual, por sua vez, é o grupo de comparação, normalmente majoritário.

Na análise do viés dos testes há dois tipos de abordagem para sua detecção. Uma delas utiliza um critério externo ao teste e, a outra, um critério interno, normalmente as pontuações totais obtidas no teste.

Com relação ao critério externo, esse tipo de viés é analisado com vistas a detectar se as pontuações do teste fornecem correlações com variáveis irrelevantes para sua interpretação. Em contrapartida, o viés interno se refere às propriedades psicométricas dos itens, que será o foco do estudo aqui apresentado. Nessa análise procura-se responder se os itens de um teste possuem o mesmo comportamento estatístico (ou equivalência de medida) quando comparados subgrupos de sujeitos pertencentes à mesma população. Em outros termos, a equivalência de medida do item será constatada quando o atributo medido por ele é idêntico para as várias subpopulações. Se observado esse fato, a conclusão é que não há funcionamento diferencial dos itens (DIF); quando a equivalência não é constatada, conclui-se pela presença de funcionamento diferencial dos itens (DIF). Dentro desse contexto, na teoria dos testes, a probabilidade de que um examinando responda a um item corretamente se denomina probabilidade de êxito e o viés pode ser estudado comparando as probabilidades de êxito para diferentes subgrupos da mesma população.

Os itens de um teste podem apresentar tipos de DIF classificados como uniforme e não uniforme (Mellenbergh, 1982). Um DIF é considerado uniforme ou consistente quando não se detecta interação entre o atributo medido e o fato de pertencer a um determinado grupo. Por sua vez, considera-se o DIF não uniforme ou inconsistente quando é constatada uma interação, em outros termos, quando pessoas classificadas no mesmo nível de habilidade em um atributo qualquer (por exemplo, inteligência superior) tiverem diferentes

probabilidades de responder corretamente um item, em razão de pertencerem a grupos distintos (por exemplo, sexo masculino e feminino).

Procedimentos para identificação de DIF

Inicialmente, o viés de um item foi operacionalizado como a dificuldade relativa do item que exagerava ou distorcia a diferença de grupo típica ou constante no teste. A idéia consistia em emparelhar os examinandos quanto à pontuação total do teste para ver se sujeitos comparáveis no teste, mas pertencentes a grupos distintos, rendiam de forma similar nos itens individuais; se não era assim, considerava-se o item enviesado.

Partia da concepção de que toda a informação sobre a habilidade de uma pessoa estaria contida em suas respostas a um conjunto de itens, ou seja, na contagem do número dos itens a que respondeu corretamente. Na abordagem tradicional dos testes, a pontuação bruta tem sido considerada suficiente para informar sobre a habilidade. Para a dificuldade do item a estatística suficiente é o número das pessoas que responderam corretamente esse item.

Entretanto, os métodos tradicionais falham em não tirar vantagem do potencial para a medida que está implícita nessa prática, pois a definição tradicional da dificuldade do item é dependente da amostra. É somente relevante a um grupo particular testado e muda se for encontrado um grupo com uma distribuição diferente da habilidade.

Há também evidências de que os critérios de seleção de itens nessa abordagem tradicional não conduzem a um conjunto ótimo dos itens (Birnbaum, 1968). Os critérios usuais para incluir um item são ter uma discriminação elevada (correlação ponto bisserial da pontuação do item com pontuação do teste) e que a dificuldade do item (proporção de acertos) esteja perto de cinqüenta por cento. Já em 1946, Tucker demonstrou que a realização perfeita desses princípios não resultaria em um bom instrumento, pois um teste com certo número de itens não seria melhor do que um teste com apenas um item.

Os fatores que conduzem às discriminações elevadas incomuns são ainda mais perturbadores. Freqüentemente as altas discriminações são devidas muito mais à influência de uma variável estranha do que a uma associação mais forte do item com o traço pretendido. Davis (1948) defendeu que um item que requer o conhecimento da palavra "idiossincrático" pôde ser eficaz para medir algum tipo de inteligência para pessoas de origem socioeconômica elevada. Entretanto, pessoas de alto nível socioeconômico têm mais probabilidade de que o termo "idiossincrático" lhes seja familiar do que os de baixo nível socioeconômico, em razão das diferenças na exposição cultural a essa palavra. Em geral, as pessoas de nível socioeconômico elevado conseguem maior pontuação em testes de inteligência verbal. Os itens como "idiossincrático" que classificam as pessoas em níveis socioeconômicos, classifica-las-ão também em grupos de inteligência. Em outros termos, quanto maior a diferença nos níveis de realização dos grupos, mais eficazes serão os itens culturalmente enviesados. Se os itens forem selecionados na base da discriminação elevada, os itens culturalmente enviesados estarão selecionados, produzindo testes com maiores vieses.

Em geral os procedimentos têm em comum o fato de utilizar os resultados globais do teste como critério para detectar o DIF e se divide em duas categorias, dependendo se o critério de viés é externo ou interno. Em qualquer das situações a preocupação básica concerne distinguir os itens que definem o traço a ser medido e quais itens são enviesados e por quem ou por qual subgrupo.

Com relação ao critério externo podem-se citar como exemplo alguns estudos que investigaram a extensão em que as pontuações do teste predisseram o sucesso na faculdade (Harris & Reitzel, 1967; Cleary, 1968; Bowers, 1970; Temp, 1971) e sucesso no trabalho (Sadacca & Brackett, 1973; entre outros). Cleary (1968), Cardall e Coffman (1964) e Anastasi (1968) consideraram um teste enviesado quando a linha de regressão misturasse sistematicamente os desempenhos dos grupos.

Estudos sobre a abordagem em termos de regressão identificaram limitações (Darlington, 1971), principalmente em termos de interpretação dos resultados. A esse respeito, Einhorn e Bass (1971) e Cole (1973) defenderam que, mesmo que outros padrões, como forma de se conseguir uma seleção mais justa, fossem usados com critério externo, ainda permaneceria um problema fundamental que se refere à técnica ter que confiar em um critério externo e supor que ele é uma medida não enviesada.

A dificuldade em construir um critério não enviesado, segundo Petersen e Novick (1974), incentivou o estudo das técnicas para detectar e corrigir os vieses de testes que usem um critério interno, ou seja, apenas a informação contida nas respostas das pessoas aos itens do teste. Embora alguns (Potthoff, 1966; por exemplo) defendam que qualquer definição estatística objetiva do viés necessita de uma variável critério, foram propostas técnicas para detectar o viés sem qualquer critério externo.

A análise de distratores ou *differential alternative functioning* (DAF) (Holland & Wainer, 1993) consiste em

examinar as respostas às alternativas incorretas para encontrar padrões seletivos nos distintos subgrupos. O uso de modelos *log* linear e do qui-quadrado para analisar respostas aos distratores como uma maneira indireta para detectar itens enviesados foi considerado por Veale e Foreman (1975) intrigante. Mas seu destino é incerto, porque dar peso aos distratores não parece ter muito impacto na ordenação das pessoas (Hakstian & Kansup, 1975).

Em ausência de um critério externo, se desenvolveu uma grande variedade de procedimentos, utilizando como critério a pontuação total no teste ou num conjunto de seus itens. Assumindo que um teste é qualquer coletânea de itens que pretendem medir um único traço ou habilidade, foram usadas algumas abordagens empíricas ou técnicas para determinar se os mesmos itens definem o mesmo traço para diferentes grupos. Os diferentes métodos DIF soem se dividir em quatro grandes tipos segundo a técnica estatística em que se baseiam: análise da variância, dificuldades transformadas do item, tabelas de contingência, Teoria de Resposta ao Item.

Mais especificamente, a abordagem direta ao viés do item começa com a comparação não ajustada das dificuldades do item calculadas na maneira tradicional. Assim, um item em que a proporção de acertos varia entre grupos (por exemplo, masculino e feminino) é considerado suspeito. Entretanto, não se pode deixar de lado o fato de que as habilidades das pessoas podem não estar distribuídas de forma equivalente entre os grupos e, em conseqüência, as diferenças podem ser devidas à diferença de habilidade entre os grupos e não ao sexo das pessoas.

Uma outra forma de estudo compara as estruturas fatoriais para os diferentes grupos que responderam os mesmos itens. Outra usa correlação ponto bisserial item-total para os diferentes grupos que passam pelos mesmos itens para selecionar os itens "eficazes" para cada grupo e constroem "testes" separados e comparam-nos. Estruturas fatoriais similares para ambos os "testes" seriam interpretadas como indicativas de que os grupos respondem similarmente os itens. Mesmo que essas técnicas tentem responder a uma pergunta vital, a saber, quais itens definem o traço, têm limitações, principalmente por confiar na correlação como instrumento indicador de viés. A correlação é vulnerável à variância do traço nas amostras estudadas e, em decorrência, alguns itens podem parecer enviesados quando não o são.

A análise de variância é uma maneira de estudar a dificuldade relativa dos itens valendo-se dos grupos (Cardall & Coffman, 1964; Cleary & Hilton, 1968). Quando a interação entre grupos de um item é significativa, indica que sua dificuldade relativa não permaneceu constante, e pode ser considerada uma indicação do viés. Cleary e Hilton (1968) propuseram uma transformação para estabilizar a variância da dificuldade do item, já que ela é muito heterogênea, mas essa transformação é muito difícil de ser interpretada sob a perspectiva da psicometria.

Um refinamento da abordagem da análise de variância compara dois grupos por meio de tabela da contingência, considerando um grupo de pessoas com habilidades semelhantes. Scheuneman (1975) defendeu que um item poderia ser considerado enviesado se sua probabilidade de uma resposta correta não fosse a mesma para pessoas com a mesma habilidade, não obstante o grupo em que a pessoa estivesse classificada. Apesar de ter sido muito prometedora e interessante, essa proposta necessitava de um desenvolvimento de seus fundamentos matemáticos de modo que seu potencial pudesse ser inteiramente explorado.

Echtemacht (1972) propôs a transformação das dificuldades tradicionais do item em *probits* escalonados para cada um dos grupos e sua representação em um gráfico normal com as diferenças ordenadas para cada grupo. Entretanto, os itens podem ser enviesados e mesmo assim se localizar perto da linha normal, pois as dificuldades do item poderiam estar distribuídas simetricamente, mas com distintas dispersões; em outros termos, um teste pode ser enviesado, mesmo que as diferenças se encaixem em uma linha normal. Aparentemente, o ajuste do item pode ser confundido com ajuste entre grupos e o problema estaria em dispor de um modelo que permitisse que os itens fossem testados enquanto ajuste dentro e entre grupos ao mesmo tempo em que a variância fosse avaliada quanto à inclinação da linha normal.

Um dos procedimentos estatísticos mais utilizados para detectar itens com DIF é o de Mantel-Haenszel (MH). Além de proporcionar poucas dificuldades tanto de cálculo como para sua interpretação, não requer amostras muito grandes e é intuitivamente mais compreensível para pessoas com pouco domínio de estatística (Holland & Thayer, 1988; Mazor, Clauser & Hambleton, 1992).

Esse procedimento compara as respostas (certas e erradas) do grupo de referência e do grupo focal em relação a um item. Como se tem que levar em consideração o nível de habilidade das pessoas que estão sendo comparadas, geralmente usa-se a pontuação observada no teste em questão como critério de equiparação. Suponhamos que foram formados três grupos, os de pontuação baixa (por exemplo, de zero a 10 pontos), os de pontuação média (por exemplo, de 11

a 20 pontos) e os de alta pontuação (por exemplo, de 21 a 30 pontos). Assim, seria possível construir três tabelas para cada item. Supondo que se analise a tabela de pessoas do grupo com pontuações baixas, seria uma tabela de dupla entrada (tabela de contingência 2 x 2). Nessa tabela, as pessoas que conseguiram uma pontuação considerada baixa são classificadas de acordo com o grupo a que pertençam (focal ou referência) e as possíveis respostas ao item (correta ou errada). É calculado um índice que expressa o quociente ou proporção entre a probabilidade de acertar o item no grupo de focal versus a probabilidade de errá-lo em comparação à probabilidade de acertar versus errar no grupo de referência. Uma limitação importante da estatística de MH é sua escassa potência para detectar o DIF não uniforme (Mellenbergh, 1982; entre outros).

Modelo Logístico Rasch

Nenhuma das técnicas propostas para identificar itens enviesados do teste tinha resolvido satisfatoriamente o problema de como medir razoavelmente pessoas não obstante sua raça, sexo, ou origem cultural. Georg Rasch (1960) forneceu um re-pensar sobre o problema da medida ao abordar adequadamente a maioria dos problemas em relação às deficiências da análise tradicional do item. Seu modelo estocástico descreve a probabilidade de acerto de uma pessoa em um item em função somente da habilidade da pessoa e da dificuldade do item. Valendo-se da exigência tradicional de que uma medida deve ser constituída por um conjunto de itens homogêneos, unidimensionalmente relacionada ao traço a ser medido, Rasch derivou seu modelo de medida na forma de uma expressão logística e demonstrou que, nessa forma, os parâmetros do item e da pessoa são estatísticas separáveis. Andersen (1973) elaborou e refinou a base matemática para o modelo e Wright e Panchapakesan (1969), por sua vez, desenvolveram os procedimentos práticos de estimativa que tornaram viável a aplicação do modelo.

Sua aplicação não necessita da comparação de dificuldades do item por meio de grupos, com base nas diferenças de distribuição da habilidade da pessoa dentro dos grupos. A variância de estimativas da dificuldade do item corresponde à situação real em que a informação é máxima no centro e mínima nos extremos da escala. As técnicas de estimativa da probabilidade máxima, aplicáveis ao modelo, conduzem às estimativas assintóticas da variância de estimativas do parâmetro. Isso faz com que possam ser identificados os vieses de maneira tal que não são mudadas as dificuldades relativas dos itens, mas sim sua escala de medida, para separar itens enviesados dos itens com desajuste por outras razões, e para especificar o valor da variância residual esperada quando os itens e as pessoas se ajustarem ao modelo da medida. Nesse contexto, pode-se dizer que 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; Mellenbergh, 1989).

Os métodos baseados no TRI para avaliar DIF são superiores para comparar os clássicos valores *p* por meio dos grupos de referência e focais. Comparações de valores *p* freqüentemente inflam taxas de erro, identificando itens que não indicam DIF real e não detectando os itens com DIF, respectivamente (Lim & Drasgow, 1990). A vantagem principal de TRI, entretanto, é que seus parâmetros são invariantes na amostra, enquanto que os valores *p* são dependentes dela.

O modelo logístico de Rasch pôde incorporar a maioria do trabalho precedente sobre viés porque começou com suposições similares de medida (Rasch, 1960, 1966; Wright, 1968; Wright & Panchapakesan, 1969). Assumindo a definição tradicional de viés do item, no sentido de que um item é enviesado se medir diferentemente distintos grupos, esse modelo avalia tal efeito por meio da análise dos resíduos. Mas como seus procedimentos são extensões racionais do modelo, a análise do viés pode ser averiguada de maneira mais sistemática e mais integrada do que tem sido feito. Em particular, os procedimentos identificam os itens que podem conduzir a uma medida válida para toda a pessoa e podem conseqüentemente ser usados para detectar e corrigir não somente medidas enviesadas para qualquer grupo, mas para detectar também uma medida enviesada para o indivíduo.

Considerações finais

Em uma linguagem mais técnica o funcionamento diferencial do item (DIF) se refere a uma diferença entre um grupo da referência (por exemplo, pessoas do sexo masculino) e um grupo focal (por exemplo, pessoas do sexo feminino) na probabilidade de acertar um item, sendo que os grupos devem possuir um mesmo nível de habilidade no atributo latente medido pelo teste. Assim, um item enviesado será aquele cujas probabilidades de êxito são diferentes, apesar da igualdade da capacidade das pessoas que responderam a ele.

O desenvolvimento bem sucedido dos procedimentos propostos para detectar e corrigir vieses tem implicações para a avaliação das pessoas e a pesquisa

da medida. Permitirão que os estudiosos detectem itens enviesados, itens que definem o traço pretendido para todos os grupos, e avaliar cada pessoa com respeito a viés. Isto permitirá aos avaliadores reconhecerem os indivíduos (e os grupos) que são injustamente medidos por itens (ou testes) e para rejeitarem itens injustos sempre que possível. Somente quando tais procedimentos são aplicados torna-se possível fazer a pesquisa objetiva sobre as causas e as conseqüências do viés do teste ou para avaliar habilidades ou traços consistentemente.

Outros modelos de resposta do item com mais de um parâmetro foram propostos para superar essas deficiências (Lord,1968, 1975; Birnbaum, 1968; Bock, 1972). Entretanto, estimativas consistentes dos parâmetros adicionais, no geral, são deficitárias ou não existem (Neyman & Scott, 1948; Andersen, 1973), nem tentativas de aplicar esses modelos foram bem sucedidos (Lord, 1975). Em contraposição, para o modelo de um parâmetro de Rasch há estimativas consistentes que provaram ser úteis em uma grande variedade de aplicações.

O modelo de Rasch, baseado na mesma exigência da suficiência da pontuação total dos métodos tradicionais, ofereceu oportunidades para ampliar a compreensão em relação à problemática do viés ou do funcionamento diferencial do item. Desde que os parâmetros são separáveis em termos de modelo, foi possível derivar estimativas independentes. Ao lado disso, a transformação logística colocou a habilidade em uma escala de menos infinito a zero e mais infinito a uma pontuação de cem por cento. Com esse procedimento foi possível eliminar os limites da escala e pôr os erros padrão da medida em um razoável relacionamento com a informação fornecida pela pontuação observada. Também, pôde-se trabalhar com discriminações elevadas e baixas, facilitando a seleção dos itens que consubstanciariam uma definição consistente do traço; conseqüência dos procedimentos técnicos de ajuste do item, que são a base para a sua seleção. Finalmente, a explicitação da expressão matemática do modelo facilita afirmações estatísticas sobre o significado de interações individuais da pessoa-item.

O modelo de Rasch fornece uma estrutura que permite a comparação dos resultados observados com os previstos. O resultado previsto é o predito pelo modelo, supondo que o item é justo com respeito à pessoa e que ela esteja motivada para usar sua habilidade para resolvê-lo. Dessa forma, ele permite avaliar a probabilidade do resultado observado e, então, fazer afirmações sobre a adequação de um item específico para uma determinada pessoa. Nesse sentido, sua aplicação engloba uma grande amplitude de constructos, tais como, ansiedade, depressão, distúrbios familiares, distúrbios alimentares, velocidade de raciocínio, dentre outros, que poderiam ser diagnosticados com a análise linear dos resíduos.

Referências

American Educational Research Association [AERA]; American Psychological Association [APA]; and National Council on Measurement in Education [NCME]. (1985). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.

Anastasi, A. (1968). *Psychological testing* (3rd ed.). New York: Macmillan.

Andersen, E. B. (1973). *Conditional inference and models for measuring*. Copenhagen: Mentalhygiejnisk Forlag.

Andriola, W. B. (2001). Descrição dos principais métodos para detectar o funcionamento diferencial dos itens (DIF). *Psicologia: Reflexão e Crítica, 14* (3), 643-652.

Angoff, W. H. (1982).Uses of difficulty and discrimination indices for detecting item bias. Em R. A. Berk (Org.). *Handbook of Methods for detecting item bias*. Baltimore: John Hopkins University Press.

Angoff, W. H. (1993). Perspectives on differential item functioning. Em P. W. Holland & H. Wainer (Orgs.). *Differential Item Functioning* (pp. 3-24). Hillsdale, NJ: LEA.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. Em F. Lord and M. Novick (Eds.). *Statistical Theories of Mental Test Scores*. Reading, Mass.: Addison-Wesley.

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika, 37*, 29-51.

Bowers, J. (1970). The comparison of GPA regression equations for regularly admitted and disadvantaged freshmen at the University of Illinois. *Journal of Educational Measurement, 7*, 219-225 .

Camilli, G. & Shepard, L. A. (1994). *Methods for identifying biased test items*. Thousands Oaks, CA: Sage Publications.

Camilli, G. (1993). The case against item bias detection techniques based on internal criteria: Do item bias procedures obscure test fairness issues? Em P. N. Holland & H. Wainer (Orgs.). *Diferential item functioning* (pp. 397-413). Hillsdale, NJ: LEA.

Cardall, C. & Coffman, W. E. (1964). A method for comparing the performance of different groups on

the items of a test. *Research Bulletin,* 64-61. Princeton, N. J.: Educational Testing Service.

Cleary, T. A. & Hilton, T. L. (1968). An investigation of item bias. *Educational and Psychological Measurement, 28*, 61-75.

Cleary, T. A. (1968). Test bias: Prediction of grades of Negro and White students in integrated colleges. *Journal of Educational Measurement, 5*, 115-124.

Cole, N. S. (1973). Bias in selection. *Journal of Educational Measurement, 10*, 237-255

Darlington, R. B. (1971). Another look at "culture fairness". *Journal of Educational Measurement, 8*, 71-82.

Davis, A. (1948). *Social-class influences upon learning.* Cambridge: Harvard University Press.

Echtemacht, G. (1972). A quick method for determining test bias. *Research Bulletin* RB, 72-17. Princeton, N. J.: Educational Testing Service.

Eells, K.; Davis, A., Havighurst, R. J., Herrick V. E & Tyler, R. W. (1951). *Intelligence and cultural differences.* Chicago: University of Chicago Press.

Einhorn, H. J. & Bass, A. R. (1971). Methodology considerations relevant to discrimination in employment testing. *Psychological Bulletin, 75*, 261-269.

Hakstian, A. & Kansup, W. (1975). A comparison of several methods of assessing partial knowledge in multiple choice tests: II. Testing procedures. *Journal of Educational Measurement, 12*, 231- 239.

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

Harris, J. & Reitzel, J. (1967). Negro freshman performance in a predominantly non-Negro university. *Journal of College Student Personnel, 8*, 366-368.

Holland, P. W. & Thayer, D. T (1988). Differential item performance and the Mantel-Haenszel procedure. Em H. Wainer & H. I. Braum (Orgs.). *Test Validity.* Hillsdale, NJ: Lawrence Erlbaum.

Holland, P. W. & Wainer, H. (Orgs.) (1993). *Differential Item Functioning.* Hillsdale, NJ: LEA.

Jensen, A. R. (1980). *Bias in mental testing.* New York: The Free Press.

Lim, R. G. & Drasgow, F. (1990). Evaluation of two methods for estimating item response theory parameters when assessing differential item functioning. *Journal of Applied Psychology, 75(2),* 164-174.

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

Lord, F. M. (1975). Evaluation with artificial data of a procedure for estimating ability and item characteristic curve parameters. *Research Bulletin,* 75-33. Princeton, N.J.: Educational Testing Service.

Lord. F. M. (1968). An analysis of the Verbal Scholastic Aptitude Test using Birnbaum's three-parameter logistic model. *Educational and Psychological Measurement, 28*, 989-1020.

Mazor, K. M., Clauser, B. E. & Hambleton, R. K. (1992). The effect of sample size on the functioning of the Mantel-Haenszel statistic. *Educational and Psychological Measurement, 52,* 443-451.

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

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

Neyman, J. & Scott, E. L. (1948). Consistent estimates based on partially consistent observations. *Econometrika, 16*, 1- 32.

Petersen, N. S. & Novick, M. R. (1974). An evaluation of some models for test bias. *ACT Technical Bulletin, 23*. Iowa City, Iowa: The American College Testing Program.

Potthoff, R. F. (1966). *Statistical aspects of the problem of biases in psychological tests* (n.° 479). Chapel Hill, NC: Institute of Statistics Mimeo Series, Department of Statistics, University of North Carolina.

Rasch, G. (1966). An item analysis which takes individual differences into account. *British Journal of Mathematical and Statistical Psychology, 19*, 49-57.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Copenhagen, Denmark: Danmarks Paedogogiske Institut.

Sadacca, R. & Brackett, J. (1973). *The validity and discriminatory impact of the Federal Entrance Examination.* A report to the Urban Institute, Washington, DC.

Scheuneman, J. (1975). *A new method of assessing bias in test items.* Paper presented at American Educational Research Association, Washington, DC.

Temp, G. (1971). Validity of the SAT for blacks and whites in thirteen integrated institutions. *Journal of Educational Measurement, 8*, 245-251.

Tucker, L. R. (1946). Maximum validity of a test with equivalent items. *Psychometrika, 11*, 1-14.

Veale, J. & Foreman, C. (1975). *Cultural validity of items and tests: A new approach.* (Statistics Unit Measurement

Research Center Tech Report 1). Iowa City, Iowa: Westinghouse Learning Corporation.

Wright, B. D & Panchapakesan, N. (1969). A procedure for sample-free item analysis. *Educational and Psychological Measurement, 29*, 23-37.

Wright, B. D. (1968). Sample-free test calibration and person measurement. Em *Proceedings of the 1967 Invitational Conference on Testing Problems*. Princeton, NJ: Educational Testing Service.

Recebido em março de 2006
Aprovado em junho de 2006

Sobre o autor:

Fermino Fernandes Sisto é doutor pela Universidad Complutense de Madrid, livre–docente pela Unicamp e docente do curso de Psicologia e do Programa de Pós-Graduação *Stricto Sensu* em Psicologia, da Universidade São Francisco, câmpus Itatiba-SP. Bolsista produtividade do CNPq.