



USO DA TEORÍA DE RESPONSA AO ÍTEM (TRI) PARA ANALIZAR A EQÜIDADE DO PROCESSO DE AVALIAÇÃO DO APRENDIZADO DISCENTE*

Wagner Bandeira Andriola

Revista Iberoamericana de Evaluación Educativa 2008 - Volumen 1, Número 1

<http://www.rinace.net/riee/numeros/vol1-num1/art12.pdf>

* Este texto é parte de un capítulo da tese doutoral do autor, intitulada Detección del Funcionamiento Diferencial del Ítem (DIF) en Tests de Rendimiento. Aportaciones Teóricas y Metodológicas, defendida publicamente em 27 de novembro de 2002 no Departamento de Métodos de Investigación y Medidas en Educación (MIDE) da Universidad Complutense de Madrid (UCM).

J á que os fins educativos consistem, essencialmente, em determinar as mudanças qualitativas por que passam os aprendizes, em termos de aquisição de aprendizagens, podemos garantir, portanto, que *a avaliação da aprendizagem é o processo de determinar em que medida ou grau se conseguem tais mudanças, possibilitando, assim, um juízo de valor acerca da qualidade dessas supostas mudanças*. Nesta definição, há duas idéias básicas:

- *O objeto da avaliação é o comportamento*, que é aqui entendido como sendo determinado por atributos intrapsíquicos ou organísmicos, tais como: motivação, compreensão, raciocínio, memorização, entre outros, pelo qual a tarefa avaliadora consiste na busca das manifestações externas que sirvam de indicadores dos aspectos internos, não avaliados diretamente (Andriola, 1998a; 2002). As relações entre os atributos internos e as manifestações externas —p. ex. as respostas aos itens de um teste de rendimento— têm por base o modelo correlacional;
- *A avaliação deve ser um processo sistemático de coleta e análise de informações*, que deve proporcionar pelo menos duas medidas: uma no início do processo de ensino e outra, no final, pois sua finalidade última é a emissão de um juízo de valor acerca das mudanças de comportamento decorrentes da aprendizagem. É interessante dizer que todo *testemunho* —p.ex.: conversas informais, debates e seminários em sala de aula, freqüência e participação dos alunos, dentre outras formas— é válido para avaliar a aprendizagem (Mc Donald, 2003). Os testes de rendimento constituem uma dessas inumeráveis formas de *dar testemunhos*, que permitem a obtenção de informações úteis, válidas e fidedignas sobre a aprendizagem de estudantes. Não obstante, em qualquer *testemunho*, existem algumas etapas que devem ser seguidas:
 - a) seleção de comportamentos ou ações que sejam indicadores dos objetivos curriculares que se deseja avaliar —tipo de objetivo educativo a ser avaliado;
 - b) apresentação de tarefas pedagógicas que possibilitem aos alunos expressar o que sabem sobre um determinado conteúdo —tipo de tarefa pedagógica a ser utilizada.

Como em toda e qualquer atividade científica, porém, a avaliação do rendimento possui vulnerabilidades ou limitações, que são descritas a seguir.

1. ORIGEM DOS ESTUDOS SOBRE O VIÉS DOS TESTES

O viés dos instrumentos de medida psicológica e educacional é um tópico que aparece tratado tardiamente, no seio da psicometria moderna (Muñiz, 1994). De acordo com Angoff (1993), seu estudo sistemático se iniciou nos Estados Unidos, no final dos anos 1960, numa época em que estavam em moda os debates sobre direitos civis e desigualdades de oportunidade entre brancos e outras minorias étnicas (Rossi, Freeman e Lipsey, 1999).

Os resultados dos processos de avaliação educacional, executados por instituições reconhecidas, tais como o *Educational Testing Service (ETS)*, foram discutidos entre diversos intelectuais, para quem as diferenças de rendimento educativo 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. Foi, portanto, a discussão social, alheia em grande parte ao círculo psicométrico especializado, que obrigou os especialistas da área a produzirem outros procedimentos estatísticos, 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 com o estudo sistemático das diferenças entre grupos demográficos. Estavam interessados em buscar explicações a respeito das suas verdadeiras causas. Martínez Arias (1997) destaca que a investigação sobre o viés dos itens pode remontar aos estudos realizados por A. Binet, em 1910, a respeito das diferenças de *status* socioeconômico no rendimento dos sujeitos submetidos a alguns testes desenvolvidos por ele próprio. Os resultados possibilitaram a proposição da hipótese de que o rendimento mais baixo destes sujeitos, em alguns itens, poderia decorrer do efeito do treinamento cultural, em vez de reais diferenças na capacidade mental ou construto latente medido pelo teste. Também W. Stern, o introdutor da expressão *quociente intelectual*, pode ser considerado como um dos primeiros investigadores da área, visto que ele estudou as diferenças relacionadas com a classe social, na Alemanha (Andriola, 2002).

Apesar desses 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, realizados na Universidade de Chicago, em 1951. Nesses estudos, os investigadores 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. Surgiram, assim, os primeiros dados a respeito dos problemas técnicos que possuíam alguns itens dos testes de rendimento, então utilizados na avaliação da aprendizagem. Eram informações sobre os problemas referentes ao uso indevido da linguagem escrita, que possibilitava certa vantagem de um grupo de sujeitos sobre outro, isto é, 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. Em consequência, os sujeitos pertencentes aos grupos minoritários, que não conheciam ou não empregavam cotidianamente esses termos, tinham rendimento mais baixo. Apareceu, então, o interesse pela investigação sistemática pelo *funcionamento diferencial do item* (DIF).

2. PRESENÇA DE DIF: FENÔMENO QUE SE CHOCA COM A EQUIDADE DA MEDIDA

No âmbito da Teoria de Resposta ao Item (TRI), o item não tem DIF, quando a sua curva característica (CCI) é idêntica para os grupos comparados, considerando-se um mesmo nível ou magnitude da variável latente medida (LORD, 1980; MELENBERGH, 1989). Em linguagem matemática podemos dizer que o item não tem DIF com respeito à variável G (grupo) dado Z (nível de θ) se, e somente se, $F(X | g, z) = F(X | z)$, onde:

- X é a pontuação no item;
- g é o valor obtido dado a variável G ;
- z é o valor obtido dado a variável Z .

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

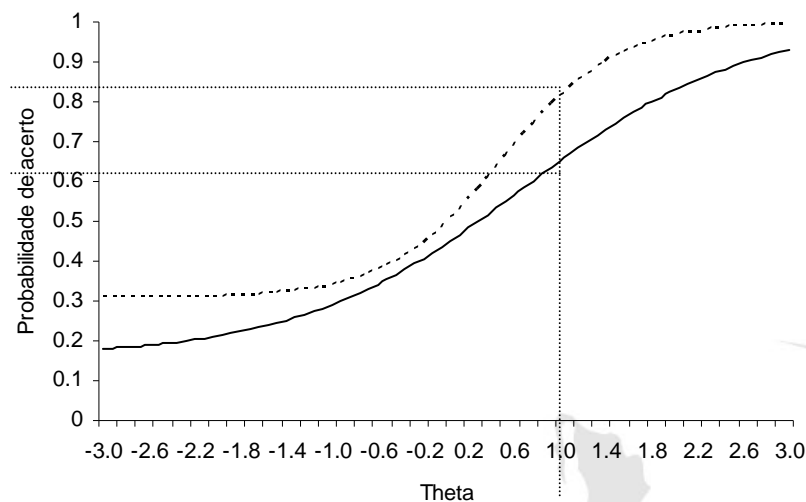
expressas nos seguintes termos: $P(X=1 | g, \theta) = P(X=1 | \theta)$ para todo g e θ . No segundo caso, a equação expressa, em realidade, a curva característica do item (CCI).

Geralmente, os estudos para a determinação do DIF utilizam dois grupos, denominados: *grupo de referência (GR)* e *grupo focal (GF)*. Como já enfatizamos, no âmbito da TRI um item tem DIF se para valores iguais de θ não correspondem valores iguais de $P(\theta)$ 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 certa magnitude na variável latente θ ;
- T_{jGF} é a pontuação verdadeira do sujeito j pertencente ao grupo focal e que possui certa magnitude na variável latente θ .

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

FIGURA 1. REPRESENTAÇÃO DAS CCI'S DE UM SUPOSTO ITEM COM DIF



Legenda: Linha descontinua - CCI das mulheres (GR); Linha contínua - CCI dos homens (GF).

Na Figura 1, observamos que, para uma mesma magnitude de θ , a probabilidade de acerto ao item $[P(\theta)]$ é sempre superior para as mulheres. Isto indica que em níveis iguais de competência na variável medida θ não há igualdade na probabilidade de acertar o item. Neste caso, o item está enviesado *contra* os homens (GR), pois os valores de $P(\theta)$ para um mesmo nível de θ são sempre maiores para as mulheres (GF). Por exemplo, para $\theta = 1,1$ temos valores aproximados de $P(\theta) = 0,69$ para os homens e $P(\theta) = 0,87$ para as mulheres.

Como conseqüência de resultados dessa natureza, Douglas, Roussos e Stout (1996) propuseram os conceitos DIF benigno e DIF adverso. No caso do DIF beneficiar o grupo de referencia (GR), isto é, quando $T_{jGR}(\theta) > T_{jGF}(\theta)$, caracteriza-se a existência de DIF benigno (ANDRIOLA, 2006). O DIF adverso ocorre no

caso do DIF beneficiar ao grupo focal (GF), ou seja, quando $T_{jGR}(\theta) < T_{jGF}(\theta)$. No exemplo da Figura 1, temos um caso de DIF adverso.

3. DIFERENÇAS ENTRE CENTROS EDUCACIONAIS: CONTROVÉRSIAS E DADOS

Nessa mesma época, os estudiosos da área educacional demonstraram que as pessimistas conclusões de J. S. Coleman e colaboradores, em 1966, e de C. S. Jencks e colaboradores, em 1972, a respeito da pouca ou nula influência dos centros educacionais sobre os resultados acadêmicos dos estudantes, são totalmente equivocadas (Andriola, 2002). Os dados de estudos executados nos últimos 20 anos permitem-nos anotar: o efeito que um centro educacional tem sobre a aprendizagem dos seus estudantes pode ser identificado e, até certo ponto, medido (OCDE, 1995; Nuttall, Goldstein, Prosser e Rabash, 1989).

Ademais, como destaca Orden Hoz (1993), as investigações a respeito das influências dos centros educacionais sobre o rendimento acadêmico têm um papel preponderante em muitos países. Vários estudos foram executados com o objetivo de tentar caracterizar os fatores que diferenciam uns centros educacionais de outros, no que se refere a associação dos mesmos ao rendimento acadêmico dos estudantes (Soares, 2002).

Por exemplo, Miles (1974) destacou dez características de um centro educacional são: possuir objetivos claros; ter um bom sistema de comunicação; regras claras de hierarquia; utilização racional dos recursos; coesão entre seus membros; moral elevada de seus membros; preocupação com a inovação; autonomia; adaptação e equilíbrio nas técnicas de resolução de problemas. Estas características têm elevado grau de interdependência e, ademais, definem um marco apropriado de índices indiretos de qualidade, pois são aspectos que, indubitavelmente, condicionam em um sentido ou em outro o funcionamento do centro educacional (De Miguel, 1989).

Sammons, Hillman e Mortimore (1998) destacaram onze fatores responsáveis pela alta eficácia de alguns centros educacionais: capacidade de liderança dos diretores; visão e objetivos comuns; ambiente adequado à aprendizagem; ênfase no processo ensino-aprendizagem; preocupação pela qualidade do ensino; existência de expectativas positivas elevadas; uso de reforços explícitos; acompanhamento regular do progresso dos alunos; explicitação dos direitos e deveres dos alunos; colaboração lar-centro educacional; e, por fim, preocupação pela qualificação profissional.

Segundo Moura Castro (1999), os principais fatores que caracterizam os centros educativos de qualidade são:

- Destinar a maior quantidade de horas possíveis ao envolvimento dos estudantes com as suas tarefas ou atividades escolares;
- Selecionar de bons professores;
- Preocupar-se pela formação e qualificação dos professores;
- Fazer com que os professores sintam-se responsáveis pelos êxitos dos seus alunos;
- Fazer com que os professores utilizem metodologias adequadas às características sociais e cognitivas dos seus aprendizes;
- Valorar o papel social do educador.

Muitas outras investigações sobre as diferenças entre os centros educativos põem ênfase sobre os processos instrutivos relacionados, fundamentalmente, com o contexto de ensino-aprendizagem (Dunkin, 1978; Horsburgh, 1999; Moura Castro, 1999; Rego, 2000, 2001). Nesse âmbito, existem estudos a demonstrar que o modo como os estudantes estejam envolvidos nas atividades educativas é um fator muito importante para explicar suas aprendizagens. De acordo com Alexander e Judy (1988), os estudantes mais envolvidos nas atividades escolares demonstram maior capacidade para organizar e associar as novas informações com as antigas, gerando, assim, outros conhecimentos. Nuthall (1999) destaca o fato de que o conhecimento é resultado da utilização de um conjunto de processos cognitivos reforçados nos âmbitos social, cultural e educativo. No caso do contexto educativo, a dinâmica utilizada na sala de aula é um dos fatores mais determinantes para que o aprendiz adquira esses processos cognitivos.

Tobias (1994), Nuthall e Alton-Lee (1995) identificaram quatro fatores primários, que têm grande poder explicativo para as diferenças individuais do rendimento:

- A compreensão dos objetivos das disciplinas;
- A participação nas atividades acadêmicas de grupo;
- Os conhecimentos anteriores e as crenças no êxito pessoal; e
- O interesse e motivação pessoal.

Dos 10 aspectos destacados por Miles (1974) e dos 11 enunciados por Sammons, Hillman e Mortimore (1998), alguns têm grande similitude com os *fatores primários* pinçados por Tobias (1994), Nuthall e Alton-Lee (1995), quais sejam: *objetivos claros e compartilhados, boa comunicação entre os membros do centro educativo, coesão, moral elevada* e, finalmente, *preocupação com a inovação*.

Os *objetivos escolares* devem possuir as desejáveis características de serem claros e aceitos pelos membros do centro educativo. Ademais, devem ser alcançáveis com os recursos disponíveis e apropriados para as demandas do contexto. As *boas comunicações* devem sofrer o mínimo de distorção no percurso que vai do emissor ao destinatário, isto é, as tensões e problemas devem ser rapidamente identificados em virtude do uso de uma boa comunicação. A comunicação tem efeito sobre a *coesão*, já que este último tem efetiva ligação com o autoconhecimento do centro educativo, em seu conjunto e sobre as partes constituintes. A *moral elevada*, no âmbito organizativo, está associada à idéia de *soma de sentimentos individuais de satisfação*, que apóiam os desejos de realizar esforços para alcançar os objetivos planejados. Finalmente, a *preocupação com a inovação* é a característica desejável de mover-se em direção de novos objetivos e procedimentos.

Várias características da escola são fundamentais para lograr que seus alunos alcancem os propósitos educativos que lhes permitam continuar desenvolvendo-se e aprendendo com autonomia. Não obstante, de acordo com Sammons, Hillman e Mortimore (1998), características socioeconômicas, de gênero, etnia e linguagem também exercem influência sobre os resultados acadêmicos. Mortimore et al. (1988) demonstraram que, em termos do progresso estudantil (valor agregado pela instituição educacional), os efeitos da escola são muito mais importantes do que fatores como idade, gênero e classe social.

Finalmente, Scheerens (1992), Fuller e Clark (1994) fazem algumas reflexões a esse respeito, ressaltando que os efeitos das escolas podem variar para diferentes tipos de conteúdos, sendo maiores para Matemática e Ciências, que são ensinadas basicamente no ambiente educacional, do que para leitura ou línguas estrangeiras, mais suscetíveis a influências do lar.

As diversas investigações aportam informações válidas sobre a influência que exercem os centros educativos sobre o rendimento escolar dos aprendizes (Soares, 2002). Nesse âmbito, cabe mencionar que as profundas diferenças entre as escolas públicas e privadas brasileiras, em muitos dos fatores destacados por Miles (1974), Sammons, Hillman e Mortimore (1998), Tobias (1994), Nuthall e Alton-Lee (1995), têm reflexo sobre aspectos facilmente perceptíveis. Assim, por exemplo, Barreto, Trompieri Filho e Andriola (1999) assinalam que as diferenças no rendimento acadêmico dos alunos desses dois tipos de escolas podem repercutir sobre suas crenças pessoais de êxito escolar.

Já o estudo executado por Andriola (1997b) demonstrou que os alunos desses dois tipos de escolas têm distintas expectativas sobre a universidade brasileira e isso reflete, em certo sentido, as crenças no êxito pessoal. Enquanto 70% dos alunos das escolas particulares crêem que cursar uma carreira universitária lhes garantirá obter uma profissão de qualidade, apenas 11% dos alunos de escolas públicas concordam com essa opinião. De forma análoga, 89% dos alunos de escolas públicas crêem que outras variáveis, tais como: o esforço pessoal e o fato de cursar uma carreira universitária, lhes permitirão obter uma profissão de qualidade. Por outro lado, tão-somente 30% dos estudantes de escolas particulares compartilham essa mesma opinião.

Com base no exposto, pode-se inferir que existem distinções, qualitativas e quantitativas, nas experiências sociais e educacionais dos alunos de escolas particulares e públicas (Ramos, 1999; Nuthall, 1999). No caso brasileiro, as diferenças nessas experiências influem, entre outras coisas, nas opções de trabalho e no grau de aprendizagem em algumas disciplinas, tais como línguas e matemática (Andriola, 1995; 1997ab; 1998c, 1999; Ramos, 1999). Nesse contexto, é quase inevitável que alguns itens presentes em ambos os testes reflitam essas experiências em seus conteúdos e, desse modo, venham a possibilitar a vantagem de um grupo sobre o outro, isto é, determinem a presença do DIF.

Hamilton (1999) observou que, em alguns casos de DIF favorável às mulheres são exigidos determinados conhecimentos obtidos fora da escola, sobretudo em itens utilizados na avaliação do conhecimento em ciências. Também Andriola (2001 d) encontrou seis itens com DIF em um banco de itens organizado para a avaliação do raciocínio verbal de estudantes brasileiros do ensino médio (Andriola, 1998b), sendo três deles benignos aos alunos de escolas públicas (GR) e os outros três benignos aos alunos de escolas particulares (GF).

Por outro lado, Clauser, Nungester e Swaminathan (1996) estudaram a presença de DIF em 440 itens de um teste para medir a aptidão para a Medicina Clínica utilizado pelo *National Board of Medical Examiners*. Os autores supuseram que os itens com DIF poderiam medir alguma aptidão secundária que, de algum modo, pudera possibilitar que certos grupos de indivíduos, com características demográficas muito específicas avantajassem outros indivíduos de grupos com características demográficas distintas. O fato de haverem realizado residência médica pode, supostamente, proporcionar aos estudantes a aquisição de habilidades que lhes possibilita terem vantagem na resolução das tarefas presentes nos itens utilizados. Assim, em comparação com os alunos com a mesma habilidade, os que fizeram residência médica tiveram maior probabilidade de acertar o mesmo item em até 30% dos casos.

4. HIPÓTESES DA INVESTIGAÇÃO

Com fundamento nos dados e nas diversas informações resultantes dos estudos relatados, elaboramos duas hipóteses para serem testadas na presente investigação. A primeira (H_1) indica que,

“Considerando-se o tipo de escola dos alunos, existirão itens, nos Testes de Português e Matemática, com DIF favorável aos alunos de escolas particulares (GR)”.

Expressando a hipótese em notação matemática, temos:

$H_0: P_{GRi}(\theta_s) = P_{GF_i}(\theta_s)$ [indica a ausência de DIF].

$H_1: P_{GRi}(\theta_s) > P_{GF_i}(\theta_s)$ [indica a presença de DIF benigno al GR].

Onde:

$P_{GRi}(\theta_s)$ e $P_{GF_i}(\theta_s)$ são, respectivamente, as probabilidades do grupo de referência – GR (alunos de escolas particulares) e focal – GF (alunos de escolas públicas) de acertar o item i , dada certa magnitude (s) da variável latente medida (θ).

Nesse contexto, há que se destacar que para detectar a presença de itens com DIF foi utilizado o método Mantel-Haenszel, que foi desenvolvido por N. Mantel e W. Haenszel no ano 1959, e aplicado ao estudo do DIF por P. W. Holland e D. T. Thayer em 1988 (Angoff, 1993; Dorans y Holland, 1993). Consiste, basicamente, na comparação das frequências observadas e esperadas de acertos e erros nos grupos de referência e focal, de acordo com os distintos níveis de habilidade (j) escolhidos pelo investigador.

A segunda hipótese (H_2) da investigação aponta que

“Os alunos de escolas particulares (GR) terão melhores rendimentos que os de escolas públicas (GF) nos itens dos Testes de Português e Matemática”.

Em notação matemática, temos:

$H_0: \mu_1 = \mu_2$ [indica a ausência de diferença entre o GR e o GF].

$H_2: \mu_1 > \mu_2$ [indica a existência de diferença entre o GR e o GF, favorável ao GR].

Onde:

μ_1 e μ_2 são, respectivamente, as pontuações médias dos alunos de escolas particulares (GR) e de escolas públicas (GF) em cada um dos testes.

Os alunos que estudaram todo o ensino fundamental (I e II) e médio em escolas particulares compõem o GR ($n = 17.763$), enquanto aqueles que o estudaram em centros educacionais públicos compõem o GF ($n = 4.441$).

5. MÉTODO EMPREGADO

A investigação executada foi do tipo correlacional (*ex post-facto*), já que os dados foram coletados *in situ* e, ademais, não houve nenhum tipo de manipulação de variáveis (Bryman e Cramer, 1992; Kvanli, 1988).

5.1. Descrição da amostra de aprendizes

Com as respostas dos 29.777 candidatos inscritos no processo seletivo da Universidade Federal do Ceará (UFC) nas provas de Língua Portuguesa e Matemática, foram efetuadas as análises para verificar a existência do DIF nos itens dos mencionados testes de rendimento. Não obstante, antes disso, é aconselhável descrever as principais características demográficas da amostra de estudantes.

O gênero feminino foi majoritário entre os respondentes, correspondendo a 16.581 casos (55,7%), sendo que 14.147 tinham idades entre 19 e 24 anos; 94,9% viviam na zona metropolitana de Fortaleza (n = 28.224); 64,2% haviam estudado todo o ensino fundamental em escolas particulares (n = 19.119) e tão-somente 15,9% em escolas públicas (n = 4.719). Esse dado nos faz constatar que um reduzido percentual de alunos do ensino fundamental público chega a tentar ingressar em uma universidade pública. A diferença entre o percentual de egressos procedentes de escolas particulares (64,2%) e de escolas públicas (15,9%) permite-nos obter uma imagem que demonstra a baixa qualidade do ensino público no Ceará.

5.2. Instrumentos utilizados

Foram empregados dois testes de rendimento das disciplinas de Língua Portuguesa e Matemática, ambos destinados a avaliar o grau de aprendizagem dos candidatos a ingressar na Universidade Federal do Ceará para o ano acadêmico de 2000. Essas provas foram aplicadas na primeira etapa do processo seletivo (vestibular) e compunham-se, respectivamente, de 18 e 15 questões fechadas, com cinco opções propostas como respostas. Para identificar as características demográficas dos sujeitos da amostra, foram utilizados os dados resultantes da aplicação do questionário para a caracterização socioeconômica e cultural dos candidatos.

6. APRESENTAÇÃO E DISCUSSÃO DOS RESULTADOS

6.1. Análise do DIF no Teste de Português segundo o tipo de escola

A Tabela 1 contém os valores do coeficiente α_{MH} , da prova de contraste (χ^2_{MH}), das probabilidades de que α_{MH} deva-se ao acaso (p) e as respectivas classificações para o DIF, nos 18 itens do Teste de Português.

TABELA 1. VALORES α_{MH} PARA OS ITENS DO TESTE DE PORTUGUÊS

ITENS	α_{MH}	χ^2_{MH}	p	TIPO DE DIF
1	0,30	7,51	0,05	Moderado
2	-0,02	0,07	n.s.	Inexistente
3	-0,35	16,07	0,05	Moderado
4	-0,11	2,03	n.s.	Inexistente
5	-0,18	3,76	0,05	Moderado
6	-0,07	0,37	n.s.	Inexistente
7	0,35	11,93	0,05	Moderado
8	0,87	45,94	0,05	Moderado
9	0,74	49,55	0,05	Moderado
10	-0,58	42,01	0,05	Moderado
11	0,20	4,74	0,05	Moderado
12	-0,18	3,04	n.s.	Inexistente
13	0,50	29,62	0,05	Moderado
14	-0,60	29,77	0,05	Moderado
15	-0,07	0,62	n.s.	Inexistente
16	0,35	15,91	0,05	Moderado
17	-0,37	9,03	0,05	Moderado
18	-0,31	12,70	0,05	Moderado

Fonte: pesquisa própria.

Os dados apresentados na Tabela 1 demonstram que tão só os itens 2, 4, 6, 12 e 15 não possuem DIF (27,8% do total); nos demais 13 itens o DIF é moderado (72,2% do total), sendo que nos itens 3, 5, 10, 14, 17 e 18 o DIF é benigno ao GR (alunos de escolas particulares), pois os valores do índice α_{MH} são negativos. Para esses casos a hipótese H_1 [$P_{GRI}(\theta_s) > P_{GFI}(\theta_s)$] foi corroborada. Nos itens 1, 7, 8, 9, 11, 13 e 16 o DIF é adverso ao GR, já que os valores do índice α_{MH} são positivos. Para esses itens, a hipótese H_1 [$P_{GRI}(\theta_s) > P_{GFI}(\theta_s)$] não foi corroborada.

6.2. Análise do DIF no Teste de Matemática segundo o tipo de escola

A Tabela 2 contém os valores do coeficiente α_{MH} , da prova de contraste (χ^2_{MH}), as probabilidades de que os valores dos α_{MH} decorrem do acaso (p) e as respectivas classificações para o DIF.

TABELA 2. VALORES α_{MH} PARA OS ITENS DO TESTE DE MATEMÁTICA

ITENS	α_{MH}	χ^2_{MH}	p	TIPO DE DIF
1	0,17	3,05	n.s.	Inexistente
2	-0,41	10,76	0,05	Moderado
3	0,74	53,17	0,05	Moderado
4	0,94	80,91	0,05	Moderado
5	0,38	19,11	0,05	Moderado
6	-0,02	0,02	n.s.	Inexistente
7	-0,14	2,00	n.s.	Inexistente
8	0,55	22,63	0,05	Moderado
9	0,27	7,36	0,05	Moderado
10	-0,84	64,38	0,05	Moderado
11	-0,63	49,71	0,05	Moderado
12	-0,37	11,01	0,05	Moderado
13	-0,37	13,46	0,05	Moderado
14	-0,22	4,44	0,05	Moderado
15	-0,25	6,68	0,05	Moderado

Fonte: pesquisa própria.

Os valores obtidos para o coeficiente α_{MH} e para a prova de contraste (χ^2_{MH}) apresentados na Tabela 2, demonstram que tão-só os itens 1, 6 e 7 não possuem DIF; nos demais 12 itens o DIF é moderado, entre os quais há cinco itens com DIF benigno ao GR (3, 4, 5, 8 e 9), pois possuem valores positivos para o coeficiente α_{MH} . Para esses itens, a hipótese $H_1 [P_{GRi}(\theta_s) > P_{GFi}(\theta_s)]$ foi corroborada. Para os demais sete itens (2, 10, 11, 12, 13, 14 e 15) o DIF é adverso ao GR, já que os valores negativos do coeficiente α_{MH} assim o indicam. Nesses casos, a hipótese $H_1 [P_{GRi}(\theta_s) > P_{GFi}(\theta_s)]$ não foi corroborada.

6.3. Análise do impacto no Teste de Português segundo o tipo de escola

O estudo do impacto em cada um dos 18 itens da prova de Português é importante para que possamos identificar e compreender as possíveis diferenças entre os alunos de tipos distintos de escolas, em termos de suas pontuações médias. Assim, poderemos saber se há alguma relação entre possíveis diferenças no impacto e a presença do DIF nos itens 3, 5, 10, 14, 17 e 18 (que possuem DIF benigno ao GR). A Tabela 3 apresenta os resultados da Análise de Variância (*one-way*) para os 18 itens do teste de Português.

TABELA 3. COMPARAÇÃO ENTRE AS PONTUAÇÕES MÉDIAS DO GR E DO GF NOS 18 ITENS DO TESTE DE PORTUGUÊS

Itens	GR			GF			F	p
	μ	(σ)	n	μ	(σ)	n		
1	112,18	(19,84)	6.103	101,12	(21,06)	965	254,97	0,000
2	111,76	(17,43)	9.884	101,87	(18,12)	1.751	473,66	0,000
3	106,38	(19,80)	9.856	95,09	(19,75)	2.055	563,16	0,000
4	107,87	(19,67)	6.917	95,76	(19,69)	1.324	421,04	0,000
5	108,34	(18,03)	12.104	98,58	(17,78)	2.388	587,69	0,000
6	110,91	(17,00)	11.593	101,59	(17,28)	2.119	536,12	0,000
7	108,89	(16,80)	14.207	99,96	(17,06)	2.655	627,87	0,000
8	105,70	(17,96)	16.572	97,09	(16,89)	3.316	646,90	0,000
9	107,61	(17,36)	14.839	99,50	(16,69)	2.759	513,46	0,000
10	108,14	(18,90)	8.086	97,49	(19,23)	1.708	444,43	0,000
11	110,14	(17,51)	12.341	100,30	(18,56)	2.224	583,89	0,000
12	105,04	(20,62)	4.424	92,85	(21,49)	898	257,34	0,000
13	109,32	(18,34)	9.377	99,92	(18,76)	1.550	350,38	0,000
14	108,41	(21,04)	3.611	96,51	(21,04)	779	204,72	0,000
15	112,48	(20,69)	6.269	98,27	(21,90)	1.101	433,46	0,000
16	110,34	(17,52)	11.349	100,46	(18,28)	1.951	523,48	0,000
17	110,37	(22,32)	2.964	95,19	(21,41)	589	230,20	0,000
18	106,31	(19,06)	9.250	95,23	(19,11)	1.915	535,49	0,000

Fonte: pesquisa própria.

Observando a Tabela 3, nos damos conta de que os 18 itens possuem diferenças significativas entre os alunos do GR e os do GF, já que os valores de F assim o indicam. Nesses itens as médias das pontuações dos componentes do GR foram superiores às médias dos componentes do GF. Em outras palavras, há indícios de que os alunos do GR têm maior grau de conhecimento de Língua Portuguesa do que os do GF.

Nesse contexto, a segunda hipótese da investigação ($H_2: \mu_1 > \mu_2$), afirmando que "os alunos de escolas particulares (GR) terão melhores rendimentos que os de escolas públicas (GF) nos itens dos testes de Português e Matemática" foi corroborada.

6.4. Análise do impacto no Teste de Matemática segundo o tipo de escola

Apresentamos, a seguir, a Tabela 4, contendo os resultados do teste da Análise de Variância (*one-way*) para cada um dos 15 itens do teste de Matemática.

TABELA 4. COMPARAÇÃO ENTRE AS PONTUAÇÕES MÉDIAS DO GR E DO GF NOS 15 ITENS DO TESTE DE MATEMÁTICA

Itens	GR			GF			F	p
	μ	(σ)	n	μ	(σ)	n		
1	101,01	(20,43)	9.349	101,68	(21,22)	1.661	1,51	n.s.
2	101,26	(20,44)	3.302	102,05	(21,18)	593	0,74	n.s.
3	100,79	(20,32)	6.321	101,42	(20,24)	879	0,75	n.s.
4	101,09	(20,59)	6.279	100,75	(21,19)	828	0,19	n.s.
5	100,95	(20,32)	10.540	101,70	(21,23)	1.899	2,15	0,000
6	100,43	(19,95)	4.273	102,49	(22,58)	753	6,55	0,011
7	100,65	(20,38)	4.671	100,45	(20,14)	917	0,75	n.s.
8	101,10	(20,36)	4.649	102,11	(21,39)	672	1,44	n.s.
9	100,47	(20,20)	6.204	100,92	(21,14)	1.060	0,44	n.s.
10	101,54	(20,36)	3.568	101,08	(21,33)	859	0,35	n.s.
11	100,82	(20,28)	5.912	100,90	(21,51)	1.376	0,02	n.s.
12	100,73	(20,43)	3.547	101,97	(21,07)	710	2,14	n.s.
13	101,07	(20,64)	4.635	101,79	(21,43)	973	0,96	n.s.
14	101,03	(20,27)	4.661	101,24	(20,99)	887	0,08	n.s.
15	101,66	(20,46)	5.264	101,61	(20,93)	1.063	0,01	n.s.

Fonte: pesquisa própria.

Como podemos observar na Tabela 4, tão só os itens 5 e 6 possuem diferenças significativas entre as pontuações médias do GR (alunos que estudaram todo o ensino fundamental e médio em escolas particulares) e o GF (alunos que estudaram todo o ensino fundamental e médio em escolas públicas), já que os valores *F* assim o indicam; nos demais 12 itens, não há qualquer diferença entre as pontuações médias de ambos os grupos.

Nos dois únicos casos de diferenças significativas entre as pontuações médias do GR e do GF (itens 5 e 6), os valores das pontuações dos componentes do GF foram superiores às médias dos componentes do GR, ou seja, considerando-se os alunos que acertaram os itens 5 e 6 do Teste de Matemática, existem

indícios de que os sujeitos do GF têm maior grau de conhecimento sobre os conteúdos de matrizes, determinantes e equações do 2º grau (item 5) e sistemas de equações, fatorização e equação modular (item 6) que os alunos do GR.

Há uma explicação plausível para essas diferenças favoráveis aos alunos de escolas públicas em ambos os itens. Existe um considerável número de alunos de escolas públicas que estudaram em dois colégios que lhes proporcionam excelente formação acadêmica e profissional, que se chamam *Centro Federal de Ensino Tecnológico do Ceará (CEFET/CE)* e *Colégio Militar*, cujo ensino nas áreas de Matemática, Física e Química é reconhecido como de alta qualidade. Assim, são compreensíveis as diferenças detectadas em ambos os itens, favoráveis aos que estudaram todo o ensino fundamental e médio em escolas públicas (GF).

Nesse contexto, a segunda hipótese da investigação ($H_2: \mu_1 > \mu_2$), afirmando que "os alunos de escolas particulares (GR) terão melhores rendimentos que os de escolas públicas (GF) nos itens dos Testes de Português e Matemática", não foi corroborada.

6.5. Ponderações finais acerca dos resultados

Considerando-se o *tipo de escola dos estudantes*, detectamos 13 casos de DIF (72,2% do total) no teste de Português (itens 1, 3, 5, 7, 8, 9, 10, 11, 13, 14, 16, 17 e 18), sendo que nos itens 3, 5, 10, 14, 17 e 18, o DIF é benigno ao GR (alunos de escolas particulares). Identificamos 12 casos de DIF (80% do total) no teste de Matemática (itens 2, 3, 4, 5, 8, 9, 10, 11, 12, 13, 14 e 15), tendo cinco deles DIF benigno ao GR (itens 3, 4, 5, 8 e 9).

Esses dados demonstram que a afirmação de Muñiz (1997) é verdadeira. Segundo o autor, não existem provas inteiramente isentas de vieses. Nesse âmbito, temos que detectar a quantidade de vieses que pode ser aceitável em um determinado teste ou item.

Com respeito ao impacto em ambos os testes, foram detectadas diferenças significativas entre as pontuações médias dos grupos em todos os 18 itens do teste de Português favoráveis, em sua totalidade, aos alunos de escolas particulares. Em apenas dois itens do teste de Matemática (5 e 6), as diferenças verificadas foram favoráveis aos alunos de escolas públicas.

Como destacamos antes, muitas investigações aportam informações válidas sobre a influência que exercem os centros educativos sobre o rendimento escolar dos aprendizes, dando-nos conta de que as profundas diferenças entre as escolas públicas e privadas brasileiras têm reflexo sobre aspectos facilmente perceptíveis, tais como o grau de aprendizagem dos discentes. Efetivamente, pelo menos quanto ao domínio da língua materna, os alunos de escolas particulares demonstraram ser muito superiores aos aprendizes de escolas públicas. Quanto ao domínio da Matemática, podemos acentuar que houve um "empate técnico".

Através da análise qualitativa dos itens com DIF, constatamos que termos "pouco conhecidos" ou "pouco utilizados" pelos estudantes estiveram presentes em 25% desses casos no teste de Português, ascendendo a 36,4% no teste de Matemática. Esses resultados indicam certa "pobreza no vocabulário" dos estudantes, apontando para uma importante disfunção das instituições de ensino. Como afirmou o Apóstolo São Mateus: *arbor ex fructu cognoscitur*.

6.6. Ponderações finais acerca da investigação

O propósito de nosso trabalho foi contextualizar o estudo do funcionamento diferencial do item (DIF) no amplo e complexo campo da avaliação da qualidade educacional. Para isso, tentamos demonstrar o aspecto que une as investigações sobre o DIF aos processos de avaliação educacional. Afirmamos que a qualidade educacional, independentemente do seu significado, é um construto intimamente associado ao uso de procedimentos sistemáticos de avaliação (Andriola, 1999a, 2000, 2001). Ademais, sabemos que um poderoso indicador da qualidade de uma instituição educacional é o grau em que esta consegue alcançar os objetivos curriculares propostos *a priori*, isto é, o grau de efetividade ou eficácia escolar que possui (Sammons, Hillman e Mortimore, 1998). Posteriormente, destacamos que o procedimento mais adequado para medir o grau de efetividade escolar é através da comparação dos resultados educativos observados com aqueles esperados (Orden Hoz, 1992), isto é, comparar o grau de aprendizagem obtido pelos alunos com os objetivos educativos propostos. Para tanto, devemos usar os conhecidos testes de rendimento ou referidos a um critério, já que são considerados os instrumentos mais adequados a esse tipo de estudo.

Não obstante, é bastante comum que os itens utilizados nesse tipo de teste tenham uma característica conhecida como *DIF*. Como enfatizamos nos tópicos anteriores, o DIF ocasiona sérios problemas às avaliações educacionais (Hartle e Bataglia, 1993). Trata-se de um fator de injustiça para alguns grupos de respondentes, já que os alunos que possuem o mesmo grau de aprendizagem, mas que provêm de grupos demográficos distintos, têm diferentes probabilidades de acertar um mesmo item. Portanto, devemos reconhecer a relevância dessas investigações, pois podem proporcionar maior equidade aos processos de avaliação educacional, através da identificação e não utilização daqueles itens que possuam algum tipo de DIF. Por fim, devemos enfatizar que *o estudo do DIF deve preceder qualquer processo de avaliação educacional sério, ou seja, esse tipo de estudo deve, necessariamente, compor a fase de pré-teste dos itens antes de sua definitiva utilização* (Andriola, 2001, 2002, 2006).

7. À GUIA DE CONCLUSÃO

Constatamos que a área de investigação do DIF no âmbito educativo e psicológico é recente e, ademais, necessita de boas hipóteses, fundamentadas em teorias científicas, que tentem "abrir perspectivas" aos estudos do DIF (Hambleton, 1997; Roznowski e Reith, 1999; Scheuneman & Gerritz, 1990). Como opina Bond (1993):

In general, however, theories about why items behave differentially across groups can be described only as primitive¹ (pág. 278).

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

- Pelo fato de as investigações acerca do DIF são relativamente recentes e, atualmente, a ênfase está no desenvolvimento de métodos estatísticos para sua identificação. Por exemplo, o desenvolvimento das modernas técnicas para a detecção do *funcionamento diferencial das*

¹ De um modo geral, as teorias sobre o porquê de alguns itens funcionarem diferentemente para certos grupos podem ser descritas somente como primitivas.

alternativas (DAF) tem o mesmo objetivo das técnicas DIF (Thissen, Steinberg e Wainer, 1993; Thissen, Steinberg e Fitzpatrick, 1989);

- Porque a identificação do DIF e os fatores a ele relacionados necessitam boas teorias sobre a dificuldade diferencial dos itens. Há que destacar que, esse é um campo, no qual as teorias acerca dos processos cognitivos presentes na resolução dos itens não se encontram, todavia, minimamente avançadas;
- Porque a identificação e a 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 difícil de ser obtido 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 ser corroboradas ou de rejeitadas, algo bastante comum à atividade científica. Como nos lembra Júlio Verne, a ciência é composta de erros, que são os passos em direção à verdade. Mais recentemente, o economista norte-americano Paul Samuelson, ganhador do prêmio Nobel de Economia, saiu-se com a seguinte frase: de funeral em funeral a ciência avança (Wilson, 1999). Todavia, muito tempo antes, na Roma antiga, o orador Cícero sentenciava: *Vivere est cogitare*.

Devemos ter claro o fato de 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 equidade aos processos avaliativos; permite aos sujeitos que possuem o mesmo grau na variável latente ou construto medido pelo item obter melhores resultados, já que esses têm maiores probabilidades de acertá-lo (Douglas, Roussos e Stout, 1996).

Nesse âmbito, caberá aos responsáveis pela construção, administração e comercialização de testes psicológicos e pedagógicos, a verificação da presença do 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 do DIF e, assim, evitar utilizá-los em processos avaliativos (Andriola, 1998b, 2001, 2002).

Para finalizar, mencionaremos uma célebre frase latina, que é muito sugestiva e sintetiza em nossa opinião, a importância dos estudos sobre o DIF no âmbito da avaliação psicológica e educacional: *fiat justitia, pereat mundus*.

REFERENCIAS BIBLIOGRÁFICAS

- Allalouf, A., Hambleton, R.K. e Siresi, S.G. (1999). Identifying the causes of DIF in translated verbal items. *Journal of Educational Measurement*, 36 (3), pp. 185-198.
- Alexander, P.A. e Judy, J.E. (1998). The interaction of domain-specific and strategic knowledge in academic performance. *Review of Educational Research*, 58, pp. 375-404.
- Andriola, W.B. (1995). Avaliação do raciocínio numérico em estudantes do 2º grau. *Educação em Debate*, 29/32, pp. 95-99.

- Andriola, W.B. (1997a). Avaliação do raciocínio verbal em estudantes do 2º grau. *Estudos de Psicologia*, 2(2), pp. 277-285.
- Andriola, W.B. (1997b). Expectativas de estudantes do 2º grau sobre a Universidade. *Educação em Debate*, 33, pp. 39-45.
- Andriola, W.B. (1998a). Apresentação de um Modelo Teórico destinado à avaliação dos Programas Estaduais de Qualificação Profissional (PEQ's). *Ensaio: Avaliação de Políticas Públicas em Educação*, 19(6), pp. 259-266.
- Andriola, W.B. (1998b). Utilização da teoria da resposta ao item (TRI) para a organização de um banco de itens destinados à avaliação do raciocínio verbal. *Psicologia: Reflexão e Crítica*, 11(3), pp. 295-308.
- Andriola, W.B. (1998c). Inteligência, aprendizagem e rendimento escolar segundo a Teoria Triárquica da Inteligência (TTI). *Educação em Debate*, 35, pp. 75-80.
- Andriola, W.B. (1999a). Evaluación: La vía para la calidad educativa. *Ensaio: Avaliação e Políticas Públicas em Educação*, 25, pp. 355-368.
- Andriola, W.B. (1999). Avaliação do raciocínio abstrato em estudantes do ensino médio. *Estudos de Psicologia*, 4(1), pp. 23-37.
- Andriola, W.B. (2000). Calidad educativa y efectividad escolar: conceptos y características. *Educação em Debate*, 39(1), p. 7-14.
- Andriola, W.B. (2002). Determinación del funcionamiento diferencial de los ítems (DIF) destinados a la evaluación del razonamiento verbal a partir del tipo de escuela. *Bordón: Revista de Pedagogía*, 53(4), pp. 473-484, 2002.
- Andriola, W.B. (2002). Detecção del funcionamiento diferencial del ítem (DIF) em tests de rendimento. Aportaciones teóricas e metodológicas. *Tese Doutoral*. Madrid: Universidad Complutense de Madrid.
- Andriola, W.B. (2006). Estudos sobre o viés de itens em testes de rendimento: uma retrospectiva. *Estudos em Avaliação Educacional*, 17(35), pp. 115-134.
- Angoff, W.H. (1993). Perspectives on differential item functioning. En: P. W. Holland e H. Wainer (Eds.), *Differential Item Functioning*. New Jersey: Lawrence Erlbaum Associate.
- Barreto, J.A.E., Trompieri Filho, N. e Andriola, W.B. (1997). Desenvolvimento da estrutura cognitiva de alunos da 4ª e 8ª séries. *Educação em Debate*, 37, pp. 101-113.
- Bond, L. (1993). Comments on the O'Neill and McPeck paper. En: P. W. Holland e H. Wainer (Eds.), *Differential Item Functioning*. New Jersey: Lawrence Erlbaum Associates.
- Bryman, A. e Cramer, D. (1992). *Análise de dados em Ciências Sociais*. Oeiras: Celta Editora.
- Clauser, B.E., Nungester, R.J. e 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(4), pp. 453-464.
- Cole, N.S. (1993). History and development of DIF. En: P.W. Holland e H. Wainer (Eds.), *Differential Item Functioning*. New Jersey: Lawrence Erlbaum Associates.
- De Miguel, M. (1989). Modelos de investigación sobre organizaciones educativas. *Revista de Investigación Educativa*, 1(13), pp. 21-56.
- Douglas, J.A., Roussos, L.A. e Stout, W. (1996). Item-Bundle DIF hypothesis testing: identifying suspect bundles and assessing their differential functioning. *Journal of Educational Measurement* 33(4), pp. 465-484.
- Dunkin, M. J. (1998). Student characteristics, classroom process and student achievement. *Journal of Educational Psychology*, 70 (6), pp. 998-1009.
- Fuller, B. y Clark, P. (1994). Raising school effects while ignoring culture? Local conditions and the influence of

- classroom tools, rules and pedagogy. *Review of Educational Research*, 64, pp. 119-157.
- Hambleton, R. K. (1997). Perspectivas futuras y aplicaciones. En: J. Muñiz, *Introducción a la Teoría de Respuesta a los Ítems*. Madrid: Ediciones Psicología Pirámide.
- Hamilton, L. S. (1999). Detecting gender-based differential item functioning on a constructed-response science test. *Applied Measurement in Education*, 12(3), pp. 211-235.
- Hartle, T.W. Y Battaglia, P.A. (1993). The Federal Role in standardized Testing (p. 291-311). En: R. E. Bennett y W. C. Ward (Org.), *Construction versus Multiple Choice Items in Cognitive Measurement*. New Jersey: Lawrence Erlbaum Associates.
- Horsburgh, M. (1999). Quality monitoring in higher education: the impact on student learning. *Quality in Higher Education*, 5(1), pp. 9-25.
- Kvanli, A.H. (1988). *Statistics. A Computer Integrated Approach*. Saint Paul: West Publishing Company.
- 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.
- Mc Donald, B.C. (2003). (org.). *Esboços em avaliação educacional*. Fortaleza: Editora da UFC.
- Mellenbergh, G.J. (1989). Item bias and item response theory. *International Journal of Educational Research*, 13(2), pp. 127-143.
- Miles, M.B. (1974). Las diez características del centro docente "sano". *La Educación Hoy*, 5, pp. 197-198.
- Moura Castro, C. (1999). Escolas feias? Escolas boas? *Ensaio: Avaliação e Políticas Públicas em Educação*, 7(25), pp. 343-354.
- Muñiz, J. (1994). *Teoría Clásica de los Tests*. Madrid: Ediciones Pirámide.
- Muñiz, J. (1997). *Introducción a la Teoría de Respuesta a los Ítems*. Madrid: Pirámide.
- Nuthall, G. (1999). Relating learning to individual differences in ability. *Journal of Educational Research*, 31(3), pp. 212-255.
- Nuthall, G. y Alton-Lee, A.G. (1995). Assessing classroom learning: how students use their knowledge and experience to answer achievement test questions in science and social studies. *American Educational Research Journal*, 32, pp. 185-223.
- Nuthall, D.L., Goldstein, H., Prosser, R. y Rabash, J. (1989). Differential scholl effectiveness. *International Journal of Educational Research*, 13(7), pp. 769-776.
- OCDE – Organización para la Cooperación y el Desarrollo Europeo (1995). *Schools under scrutiny*. Paris: Head of Publication Service.
- Orden Hoz, A. (1992). Calidad y evaluación de la enseñanza universitaria. *Resúmenes del Congreso Internacional de Universidades*. Madrid, julio, pp. 531-539.
- Orden Hoz, A. (1993). La escuela en la perspectiva del producto educativo. Reflexiones sobre evaluación de centros docentes, *Bordón*, 45(3), pp. 263-270.
- Ramos, E.A. (1999). Aprendizagem humana. *Cadernos de Educação*, 23, pp. 37-49.
- Rego, A. (2000). Impactos dos comportamentos de cidadania docente sobre os alunos universitários. A perspectiva dos estudantes e dos professores. *Linhas Críticas*, 6(10), pp. 9-30.
- Rego, A. (2001). Eficácia comunicacional dos docentes do Ensino Superior: evidência confirmatória do constructo. *Psicologia: Reflexão e Crítica*, 14(3), pp. 563-568.

- Rossi, P.H., Freeman, H.E. y Lipsey, M.W. (1999). *Evaluation. A systematic approach*. London: Sage Publications.
- Roznowski, M. e Reith, J. Examining the measurement quality of tests containing differentially functioning items: Do biased items result in poor measurement? *Educational and Psychological Measurement*, 52 (2), p. 248-269, 1999.
- Sammons, P., Hillman, J. e Mortimore, P. (1998). *Características Clave de las Escuelas Efectivas*. México: Secretaria de Educación Pública.
- Scheerens, J. (1992). *Effective schooling: research, theory and practice*. Londres: Cassell.
- Scheuneman, J. D. Y Gerritz, K. (1990). Using differential item functioning procedures to explore sources of item difficulty and group performance characteristics. *Journal of Educational Measurement*, 27(2), pp. 109-131.
- Schmitt, A.P., Holland, P.W. y Dorans, N.J. (1993). Evaluating hypotheses about differential item functioning (p. 281-319). En: P. W. Holland y H. Wainer (Eds.), *Differential Item Functioning*. New Jersey: Lawrence Erlbaum Associates.
- Soares, J.F. (Coord.) (2002). *Escola eficaz: um estudo de caso em três escolas da rede pública do Estado de Minas Gerais*. Belo Horizonte: GAME/UFMG.
- Thissen, D., Steinberg, L. Y Fitzpatrick, A.R. (1989). Multiple-choice models: the distractors are also part of the item. *Journal of Educational Measurement*, 26(2), pp. 161-176.
- Thissen, D., Steinberg, L. y Wainer, H. (1993). Detection of differential item functioning using the parameters of item response models (p. 67-113). En: P. W. Holland y H. Wainer (Eds.), *Differential Item Functioning*. New Jersey: Lawrence Erlbaum Associates.
- Tobias, S. (1994). Interest, prior knowledge and learning. *Review of Educational Research*, 64(1), pp. 37-54.
- Wilson, E.O. (1999). *Consilience. La Unidad del Conocimiento*. Barcelona: Ediciones Galaxia Gutemberg.
- Zumbo, B.D. (1999). *A Handbook on the theory and methods of differential item functioning (DIF). Logistic regression modeling as a unitary framework for binary and Likert-type (ordinal) item scores*. Ottawa: Directorate of Human Resources Research and Evaluation, Department of National Defense of Canadá.