REVISTA ELETRÔNICA ACTA SAPIENTIA

INDÍCIOS DE INIQUIDADE EM TESTES DE RENDIMENTO EMPREGADOS EM AVALIAÇÕES DE LARGA ESCALA: ESTUDO COMPARATIVO ENTRE ALUNOS DE ESCOLAS PÚBLICAS E PRIVADAS[1]

Wagner Bandeira Andriola[*]
Universidade Federal do Ceará (UFC)
w_andriola@ufc.br

Resumo

O texto aborda a temática da justiça e da equidade no seio de processos avaliativos, sobretudo quando se empregam provas de rendimento compostas por itens objetivos. A área da psicometria que trata do estudo sistemático do Funcionamento Diferencial do Item (DIF) desenvolveu inúmeros procedimentos estatísticos para o tratamento dessa fonte de viés que compromete a justiça de processos avaliativos. Atualmente, com o desenvolvimento de sistema de avaliação que usam testes de rendimento em larga escala, com o fito de obter indícios válidos sobre o desenvolvimento e a qualidade da aprendizagem dos alunos, este tema deve ser incorporado à pauta da avaliação educacional, porquanto sua presença gera distorções graves sobre as inferências estatísticas efetivadas a partir dos resultados obtidos que, por seu turno, baseiam-se no uso dos testes de rendimento.

Palavras-chave: Avaliação Educacional; Testes de Rendimento; Funcionamento Diferencial do Item (DIF); Equidade na Avaliação.

EVIDENCES OF INEQUALITY IN PERFORMANCE TESTS USED IN LARGE-SCALE EVALUATIONS: COMPARATIVE STUDY BETWEEN PUBLIC AND PRIVATE SCHOOL STUDENTS

Abstract

The text addresses the issue of justice and equity within assessment processes, especially when performance tests made up of objective items are used. The area of psychometrics that deals with the systematic study of Differential Item Functioning (DIF) has developed numerous statistical procedures for treating this source of bias that compromises the fairness of evaluative processes. Currently, with the development of assessment systems that use performance tests on a large scale, with the aim of obtaining valid evidence about the development and quality of student learning, this topic must be incorporated into the agenda of educational assessment, since its presence generates serious distortions on the statistical inferences made from the results obtained which, in turn, are based on the use of performance tests.

Keywords: Educational Evaluation; Performance Tests; Differential Item Functioning (DIF); Evaluation Equity.

[1] Pesquisa que originou a Tese para Professor Titular do autor, apresentada e aprovada (*magna cum laude*) em novembro de 2018 na Faculdade de Educação (FACED) da Universidade Federal do Ceará (UFC).

[*] Doutor em Filosofia e Ciências da Educação (*Universidad Complutense de Madrid*), Mestre em Psicologia Social e do Trabalho (Universidade de Brasília), Bacharel e Licenciado em Psicologia (Universidade Federal da Paraíba). Professor Titular da UFC e Pesquisador 1B do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq). *E-mail:* w_andriola@ufc.br

REVISTA ELETRÔNICA ACTA SAPIENTIA

EVIDENCIAS DE DESIGUALDAD EN PRUEBAS DE DESEMPEÑO UTILIZADAS EN EVALUACIONES A GRAN ESCALA: ESTUDIO COMPARATIVO ENTRE ESTUDIANTES DE COLEGIOS PÚBLICOS Y PRIVADOS

Resumen

El texto aborda el tema de la justicia y la equidad dentro de los procesos de evaluación, especialmente cuando se utilizan pruebas de desempeño compuestas por ítems objetivos. El área de la psicometría que se ocupa del estudio sistemático del Funcionamiento Diferencial de Ítems (DIF) ha desarrollado numerosos procedimientos estadísticos para tratar esta fuente de sesgo que compromete la equidad de los procesos evaluativos. Actualmente, con el desarrollo de sistemas de evaluación que utilizan pruebas de desempeño a gran escala, con el objetivo de obtener evidencias válidas sobre el desarrollo y calidad de los aprendizajes de los estudiantes, este tema debe incorporarse a la agenda de la evaluación educativa, ya que su presencia genera serias distorsiones en las inferencias estadísticas realizadas a partir de los resultados obtenidos que, a su vez, se basan en el uso de pruebas de rendimiento.

Palabras clave: Evaluación Educativa; Pruebas de Rendimiento; Funcionamiento Diferencial del Ítem (DIF); Equidad de la Evaluación.

1. INTRODUÇÃO

Nos Estados Unidos da América (EUA), a partir da década de 1960, os resultados de processos de avaliação educacional executados por instituições mundialmente reconhecidas, tais como o *Educational Testing Service (ETS)*, foram discutidos entre intelectuais. Para este público, as diferenças de rendimento educacional 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 (LINN; HARNISCH, 1981; ANDRIOLA, 2002a).

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 de desempenho entre grupos demográficos, pois estavam interessados em buscar explicações a respeito das suas verdadeiras causas (AYALA, 2009; MACHADO; ALAVARSE, 2014).

Martínez Arias (1997) destaca que a investigação sobre o viés dos itens pode remontar aos estudos realizados por Alfred 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, 2002b).

Apesar desses autores pioneiros, o começo da moderna investigação sobre o viés dos itens encontra-se nos trabalhos de K. Eells, A. Davis, R. J. Havighurst, V. E.

Herrick e R. W. Tyler, realizados na Universidade de Chicago, a partir de 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 (ANGOFF; FORD, 1973; ANDRIOLA, 1995b; ADEDOYIN; ADEDOYIN, 2013).

Ademais, como destaca Orden Hoz (1992), as investigações a respeito das influências dos centros educacionais sobre o rendimento acadêmico do alunado 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 (MURPHY, 1992).

Por exemplo, Miles (1974) destacou dez características de um centro educacional eficaz, tais como: 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 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 (CREEMERS, 1996; DE MIGUEL *et al.*, 1994; MORTIMORE, 1996).

Consoante Sammons, Hillman e Mortimore (1995) há onze fatores responsáveis pela alta eficácia de 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.

No entanto, conforme 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 centros educacionais põem ênfase nos processos instrutivos relacionados, fundamentalmente, com o contexto de ensino-aprendizagem (DUNKIN, 1978; ERCIKAN, 1998; HORSBURGH,

1999; SALES et al. 2011; SILVA et al., 2017). Existem estudos a demonstrar que o modo como os estudantes envolvem-se com as atividades educativas é um fator muito importante para explicar a qualidade de sues aprendizados (ADEGOKE, 2013; HARVEY; GREEN, 2013). De acordo com Alexander e Judy (1988), Andriola (2000c), Andriola e Andriola (2009), 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 educacional, a dinâmica utilizada na sala de aula é um dos fatores mais determinantes para que o aprendiz adquira esses processos cognitivos (NUNES et al. 2017; SILVA; LIMA; ANDRIOLA, 2016).

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; o interesse e a motivação pessoal.

Dos dez aspectos destacados por Miles (1974) e dos 11 enunciados por Sammons, Hillman e Mortimore (1995), 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 introdução da inovação nos processos de ensino.

No entanto, Tyler (1949) destacou que 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 os problemas devem ser rapidamente identificados. 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 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 em direção de novos objetivos e procedimentos (VIANNA, 2003; ANDRIOLA, 2008b).

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 (1995), características socioeconômicas, de gênero, etnia e linguagem também exercem influência sobre os resultados acadêmicos. Porém, em termos do progresso estudantil, os efeitos da escola são muito mais importantes do que fatores tais como idade, gênero e classe social. Finalmente, Scheerens (1992), Fuller e Clark (1994) fizeram 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 (BAUER; ALAVARSE; OLIVEIRA, 2015).

Portanto, ademais do fator relativo ao centro educacional (público ou privado; urbano ou rural), algumas características individuais, tais como gênero, etnia e desenvolvimento de competências associadas à língua materna, conformam um rol de variáveis que têm aderência muito forte ao desempenho dos alunos em provas de rendimento (ANDRIOLA, 2000a). Ao se detalhar tais especificidades, em termos de desempenho dos alunos nos itens componentes dessas provas de rendimento, observam-se padrões distintos, que têm associação com o centro educacional e/ou com as características individuais do aluno (ANDRIOLA, 1998; ANDRIOLA, 2000b; ANDRIOLA; BARRETO, 1997). Este é o espaço no qual se insere o problema de pesquisa, conforme a seguir explicitado.

2. DELIMITAÇÃO DO PROBLEMA DE PESQUISA

As diversas investigações referidas ao longo desse trabalho aportam informações válidas sobre a influência que exercem os centros educacionais sobre o rendimento escolar dos aprendizes (ANDRIOLA, 2001a; SOARES, 2002; SAMPAIO; MANCINI, 2007). Nesse âmbito, cabe mencionar que as diferenças entre as escolas públicas e privadas brasileiras, em muitos dos fatores destacados por Miles (1974), Sammons, Hillman e Mortimore (1995), Tobias (1994), Nuthall e Alton-Lee (1995), têm reflexo sobre aspectos facilmente perceptíveis.

Assim, por exemplo, o estudo executado por Barreto, Trompieri Filho e Andriola (1999) atestou que as diferenças no rendimento acadêmico dos alunos desses dois tipos de escolas repercutem sobre as crenças pessoais de êxito acadêmico. Já o estudo executado por Andriola (1997) 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 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 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 (NUTHALL, 1999). No caso brasileiro, as diferenças nessas experiências influem, dentre outras, nas opções de trabalho e no grau de aprendizagem em algumas disciplinas, tais como línguas e matemática (ANDRIOLA, 1995b).

Nesse contexto, é quase inevitável que os itens presentes em 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 de funcionamento diferencial do item (DIF) nas provas para avaliar o desempenho do alunado (ANDRIOLA, 2001b). Por exemplo, 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 (2001d) encontrou seis itens com DIF em um banco de itens organizado para a avaliação do raciocínio verbal de estudantes brasileiros do ensino médio (ANDRIOLA, 1998), 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*, dos EUA. 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 igual habilidade, aqueles da residência médica tiveram maior probabilidade de acertar o mesmo item em até 30% dos casos.

Portanto, os estudos referidos, acerca do DIF, apontam que os sujeitos pertencentes aos grupos minoritários, que não conheciam ou não empregavam cotidianamente certos termos, tinham rendimento mais baixo nas provas empregadas para averiguar o desempenho acadêmico. A partir dessas constatações, surgiu, então, o interesse sistemático pela investigação científica do funcionamento diferencial do item (DIF), nos âmbitos psicológico e educacional (ANDRIOLA, 2001b; ANDRIOLA; LEITE; MAIA, 2013).

Constatada a relevância dos estudos do DIF para garantir-se justiça ao processo avaliativo que emprega provas ou testes de rendimento, bem como diante das lacunas identificadas em estudos sistemáticos do DIF no âmbito brasileiro e cearense, sobretudo quando se empregam testes de rendimento, indaga-se:

> ***As avaliações em larga escala incorreriam em iniquidade contra os respondentes oriundos de escolas públicas?***

A iniquidade referida acima é resultante da presença de itens com DIF nas provas de rendimento que são usadas nas sistemáticas avaliações em larga escala, geralmente enquanto processos de seleção voltados aos candidatos que desejam aceder aos níveis educacionais mais elevados, ou à certificação de alunos egressos de etapas educacionais já cumpridas (ANDRIOLA, 2008a; LIMA *et al.*, 2015; LIMA; ANDRIOLA, 2018).

Portanto, acentua-se, uma vez mais, a relevância pedagógica da sistemática para identificar a presença de DIF em itens componentes de testes ou provas de rendimento, porquanto é um indicador de iniquidade e de injustiça do processo avaliativo que faz uso desses sofisticados instrumentos de medida, conforme opinaram Andriola (2000ab), Andriola (2001b), Andriola (2002ab), Andriola (2008a), Andriola e Pasquali (1995). Realizar estudos acerca da presença de DIF em testes de rendimento, portanto, assegurará a execução de avaliações justas e isentas de vieses que privilegiem grupos específicos de respondentes (ANDRIOLA, 1995a; ANGOFF, 1993; ZUMBO, 1999; KLEIN, 2013).

3. HIPÓTESES DA INVESTIGAÇÃO

Sendo esta uma pesquisa científica, deverá submeter seu objeto de estudo à investigação metódica e sistemática, com o intuito de produzir conhecimento válido e fidedigno. Com efeito, entre os métodos que nos fornecem fundamento lógico para a realização da pesquisa científica encontra-se o método hipotético-dedutivo, cujas matrizes filosóficas baseiam-se no racionalismo e no empirismo (POPPER, 2008).

Proposto pelo filósofo austríaco Karl Raimund Popper, o método hipotético-dedutivo representa a tentativa de superar os problemas dos métodos indutivo e dedutivo, partindo do princípio que para *uma teoria ser considerada científica deverá apresentar a possibilidade de ser falseada*. Assim, para conhecer a realidade provisoriamente, visto que, consoante a lógica de K. R. Popper, a explicação dos fenômenos da realidade será sempre provisória, o cientista deverá, a partir da formulação de problemas, elaborar hipóteses cujas consequências se submeterão à possibilidade de serem testadas empiricamente (POPPER, 1972).

Portanto, seguindo a lógica do método hipotético-dedutivo, *os fatos particulares refutam ou corroboram hipóteses dedutíveis de um corpo teórico*, *enfraquecendo ou fortalecendo o referido corpo teórico* (ou teoria científica), que, sendo assim, *tem sua validade sempre provisória*. Consoante Popper (1972):

> Um dos ingredientes mais importantes da civilização ocidental é o que poderia chamar de 'tradição racionalista', que herdamos dos gregos: a tradição do livre debate – não a discussão por si mesma, mas a busca da verdade. [...] Dentro dessa tradição racionalista, a ciência é estimada, reconhecidamente, pelas suas realizações práticas, mais ainda, porém, pelo conteúdo informativo e a capacidade de livrar nossas mentes de velhas crenças e preconceitos, velhas certezas, oferecendo-nos em seu lugar novas conjecturas e hipóteses ousadas (p.129).

Percebe-se que para Popper (2008) as explicações provisórias acerca do funcionamento dos fenômenos, consideradas *hipóteses*, devem sempre ser contrastadas com a realidade através do teste empírico que tem como base, por seu turno, o método científico. Como resultado, as explicações provisórias (denominadas hipóteses) serão corroboradas e aquilatadas, ou refutadas e enfraquecidas, permitindo, assim, o adensamento do conhecimento científico e, por conseguinte, o progresso da ciência. Por outro lado, o novo rol de conhecimento científico gerado servirá de maneira mais efetiva ao meio social. Sob este prisma teórico, propõem-se, então, as seguintes hipóteses:

- ✓ A primeira hipótese (H1) indica que, *"A maioria dos itens com DIF no Teste de Língua Portuguesa será favorável aos alunos de escolas particulares (GR)".* Expressando a hipótese H1 em notação matemática:

  H0: PGR_i (θs) = PGF_i (θs) [indica a ausência de DIF].

  H1: PGR_i (θs) > PGF_i (θs) [indica DIF benigno ao GR].

  Onde:

  - PGR_i (θs) e PGF_i (θs) são, respectivamente, as probabilidades do grupo de referência – GR (alunos de escolas particulares) e do grupo focal – GF (alunos de escolas públicas) acertar o item i, dado certa

magnitude (s) da variável latente medida (θ), nesse caso específico a competência em Língua Portuguesa.

- ✓ A segunda hipótese (H2) da investigação aponta que: *"A maioria dos itens com DIF no Teste de Matemática será favorável aos alunos de escolas particulares (GR)".* Expressando a hipótese H2 em notação matemática:

  H0: PGR_i (θs) = PGF_i (θs) [indica a ausência de DIF].

  H2: PGR_i (θs) > PGF_i (θs) [indica DIF benigno ao GR].

  Onde:

  - PGR_i (θs) e PGF_i (θs) são, respectivamente, as probabilidades do grupo de referência – GR (alunos de escolas particulares) e do grupo focal – GF (alunos de escolas públicas) acertar o item i, dado certa magnitude (s) da variável latente medida (θ), nesse caso específico a competência em Matemática.

4. MÉTODO EMPREGADO PARA IDENTIFICAR A PRESENÇA DE DIF

Para detectar a presença de itens com DIF nas Provas de Língua Portuesa e de Matemática empregou-se o método Mantel-Haenszel, desenvolvido por N. Mantel e W. Haenszel no ano 1959, e aplicado ao estudo do DIF por P. W. Holland e D. T. Thayer em 1988 (ANGOFF, 1993; DORANS; HOLLAND, 1993). O procedimento consiste na comparação das frequências observadas e esperadas de acertos e erros, considerando-se os grupos de referência (GR) e focal (GF) organizados de acordo com os níveis de habilidade determinados pelo investigador. As respostas dos sujeitos são organizadas em uma tabela de frequências, como a apresentada a seguir.

Tabela 1. Frequências observadas das respostas a um item.

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

Fonte: Adaptada de Dorans e Holland (1993).

Com base nesta lógica, N. Mantel e W. Haenszel propuseram a seguinte fórmula para a comparação das frequências:

Onde:

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

- A_j é a frequência observada das respostas corretas do grupo de referência nos distintos níveis de pontuação escolhidos;
- B_j é a frequência observada das respostas incorretas do grupo de referência nos níveis de pontuação escolhidos;
- C_j é a frequência observada das respostas corretas do grupo focal nos níveis de pontuação escolhidos;
- D_j é a frequência observada das respostas incorretas do grupo focal nos níveis de pontuação escolhidos;
- T_j é o total de erros e acertos de cada grupo nos níveis de pontuação escolhidos.

O coeficiente α_{MH} expressa a magnitude do DIF, no qual $\alpha_{MH} = 1,0$ significará idêntico comportamento do item para os grupos comparados (GR e GF), implicando na inexistência de DIF. Nos casos em que $\alpha_{MH} \neq 1,0$ haverá DIF, conforme expressaram Longford, Holland e Thayer (1993), Fidalgo e Scalon (2012). Quando o α_{MH} for negativo o DIF será benigno ao GR; quando for positivo o DIF será adverso ao GR (DOUGLAS; ROUSSOS; STOUT, 1996; SCHEUNEMAN; GERRITZ, 1990).

4.1. NATUREZA DOS DADOS EMPREGADOS NA PESQUISA

Foram usados dados oriundos das respostas de 29.777 alunos do Ensino Médio candidatos aos cursos de graduação da Universidade Federal do Ceará (UFC) submetidos a duas Provas de Rendimento: Língua Portuguesa e Matemática.

4.2. ORGANIZAÇÃO DOS DADOS EMPREGADOS NA PESQUISA

As respostas dos 29.777 egressos do Ensino Médio foram organizadas em planilhas do Excel e, posteriormente, importadas e adaptadas ao *Software Statistical Package for the Social Sciences* (versión 21.0), com o fito de serem submetidos às análises estatísticas descritivas e inferenciais, bem como para verificar a presença do DIF entre os itens e para testar as hipóteses (RIZOPOULOS, 2006).

4.3. CARACTERÍSTICAS BASILARES DA AMOSTRA INVESTIGADA

No que tange às características mais substanciais da amostra estudada, importa referir que o gênero feminino foi majoritário entre os respondentes, correspondendo a 16.581 casos (55,7%). A significativa maioria vivia na zona metropolitana de Fortaleza (n = 28.224 ou 94,9%), com grupo significativo com idades entre 19 e 24 anos (n = 14.147 ou 47,5%) e valor médio 23,9 anos (desvio-padrão = 3,4 anos).

Com respeito ao tipo de escola da qual eram oriundos, os alunos que estudaram todo o Ensino Fundamental (I e II) e o Ensino Médio em escolas particulares compunham a maioria (n = 17.763 ou 59,7%), enquanto aqueles que o estudaram integralmente em centros educacionais públicos compunham a minoria (n = 4.441 ou 14,9%). Os demais casos conformaram outras categorias não relevantes aos propósitos desse estudo.

5. RESULTADOS DO TESTE DAS HIPÓTESES DA PESQUISA

5.1. RESULTADOS DO TESTE EMPÍRICO DA HIPÓTESE H1

A primeira hipótese (H1) assevera que: *"A maioria dos itens com DIF no Teste de Língua Portuguesa será favorável aos alunos de escolas particulares (GR)".*

5.1.1. ANÁLISE DO DIF NO TESTE DE LÍNGUA PORTUGUESA

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

Tabela 2. Valores de αMH para os 18 itens do Teste de Língua Portuguesa.

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

Fonte: Pesquisa Direta.

Consoante os dados da Tabela 2, observou-se que tão somente os itens 2, 4, 6, 12 e 15 estão ausentes de DIF, o que corresponde a 27,8% do total de itens da Prova de Língua Portuguesa. Como não têm DIF, estes itens apresentam-se justos e livres de viéses que ocasionem iniquidade ou que tragam vantagens indevidas para algum subgrupo dentre os respondentes. Não obstante, nos demais 13 itens, que conformam a maioria, pois representam 72,2% do total de itens da Prova de Língua Portuguesa, detectaram-se a presença de DIF. Esta constatação é grave, pois revela a presença de um problema que ocasiona iniquidade e que imprime vantagem indevida para um subgrupo específico de respondentes.

Dentre os 13 itens com DIF, há seis (itens 3, 5, 10, 14, 17 e 18), que representam 46,2% do total, com DIF benigno ao GR, já que os valores de αMH são negativos. Assim sendo, estes seis itens privilegiam os respondentes de escolas particulares, constituindo-se, assim, em itens injustos para com os componentes do GF (alunos de escolas públicas), causando vantagens indevidas para os respondentes do GR (alunos de escolas particulares). **Para esses seis itens com DIF benigno ao GR, a hipótese H1 [PGRi (θs) > PGFi (θs)] foi corroborada.**

Já nos casos dos itens 1, 7, 8, 9, 11, 13 e 16, que representam 53,8% do total de itens com DIF, o DIF revelou-se adverso ao GR, pois os valores do coeficiente αMH são positivos. Estes sete itens privilegiam os respondentes oriundos de escolas públicas, constituindo-se, assim, em itens injustos para com os componentes do GR (alunos de escolas particulares), pois proporcionam vantagens indevidas para os respondentes do GF (alunos de escolas públicas). **Para estes sete itens que possuem DIF adverso, a hipótese H1 [PGRi (θs) > PGFi (θs)] foi rechaçada.**

Para enriquecer a análise, determinou-se a magnitude do DIF nos itens do Teste de Língua Portuguesa acometidos pelo problema, considerando-se a escala proposta pelo *Educational Testing Service (ETS)*:

- *DIF severo:* $1,0 < \alpha_{MH} < 1,5$ (sendo adotado $\alpha = 0,05$);
- *DIF moderado:* $0,0 < \alpha_{MH} < 1,0$ (sendo adotado $\alpha = 0,05$);
- *DIF desprezível:* itens cujos valores absolutos não estejam em nenhuma das duas categorias anteriores (sendo adotado $\alpha = 0,05$).

A Tabela 3 apresenta os valores do coeficiente α_{MH} para os itens componentes do Teste de Língua Portuguesa que possuem DIF.

Tabela 3. Magnitude do DIF dos itens do Teste de Língua Portuguesa.

Itens	αMH	p	Magnitude do DIF
1	0,30	0,05	Moderada
3	-0,35	0,05	Moderada
5	-0,18	0,05	Moderada
7	0,35	0,05	Moderada
8	0,87	0,05	Moderada
9	0,74	0,05	Moderada
10	-0,58	0,05	Moderada
11	0,20	0,05	Moderada
13	0,50	0,05	Moderada
14	-0,60	0,05	Moderada
16	0,35	0,05	Moderada
17	-0,37	0,05	Moderada
18	-0,31	0,05	Moderada

Fonte: Pesquisa Direta.

Conforme os resultados, todos os itens componentes do Teste de Língua Portuguesa com DIF, o possuem em magnitude moderada. Portanto, com base nas análises efetivadas e para concluir este tópico da Tese que permitiu o teste empírico da hipótese H1, pode-se asseverar que **a Prova de Língua Portuguesa apresenta indícios de ser um instrumento injusto, introdutor de iniquidade no processo de avaliação**, posto que privilegia o desempenho de indivíduos de subgrupos específicos, componentes do universo de respondentes submetidos à Prova de Língua Portuguesa.

Ademais, o referido instrumento de avaliação ora beneficiou alunos oriundos de escolas particulares (no caso dos itens 1, 7, 8, 9, 11, 13 e 16), ora privilegiou alunos de escolas públicas (no caso dos itens 1, 7, 8, 9, 11, 13 e 16). Trata-se, portanto de uma situação inusitada, na qual a Prova de Língua Portuguesa revelou ser composta por uma proporção majoritária de itens com DIF, revelando-se um instrumento de medida extremamente injusto, ora para com os indivíduos do GR (alunos de escolas particulares), ora para com os componentes do GF (alunos de escolas públicas).

5.2. RESULTADOS DO TESTE EMPÍRICO DA HIPÓTESE H2

A segunda hipótese (H2) afirmou que: *"A maioria dos itens com DIF no Teste de Matemática será favorável aos alunos de escolas particulares (GR)".*

5.2.1. ANÁLISE DO DIF NO TESTE DE MATEMÁTICA

A Tabela 4 contém os valores do coeficiente αMH, da prova de contraste (χ^2_{MH}), das probabilidades de que αMH se deva ao acaso (p) e as respectivas classificações para o DIF nos 15 itens componentes do Teste de Matemática.

Tabela 4. Valores do Coeficiente αMH para os 15 itens do Teste de Matemática.

Itens	αMH	χ^2MH	p	Tipo de DIF
1	0,17	3,05	n.s.	Inexistente
2	-0,41	10,76	0,05	Benigno
3	0,74	53,17	0,05	Adverso
4	0,94	80,91	0,05	Adverso
5	0,38	19,11	0,05	Adverso
6	-0,02	0,02	n.s.	Inexistente
7	-0,14	2,00	n.s.	Inexistente
8	0,55	22,63	0,05	Adverso
9	0,27	7,36	0,05	Adverso
10	-0,84	64,38	0,05	Benigno
11	-0,63	49,71	0,05	Benigno
12	-0,37	11,01	0,05	Benigno
13	-0,37	13,46	0,05	Benigno
14	-0,22	4,44	0,05	Benigno
15	-0,25	6,68	0,05	Benigno

Fonte: Pesquisa Direta.

De acordo com os dados da Tabela 4, observou-se que tão somente os itens 1, 6 e 7 estão ausentes de DIF, o que corresponde a 20% do total de itens da Prova de Matemática. Como não têm DIF, estes itens apresentam-se justos e livres de viéses que ocasionem iniquidade ou que promovem vantagens indevidas para algum subgrupo específico dentre os respondentes da Prova de Matemática.

Nos demais 12 itens, que correspondem a 80% da Prova de Matemática, há DIF favorecendo a algum subgrupo de respondente. Percebe-se, assim, que a Prova de Matemática é um instrumento com elevado poder de introduzir injustiça e iniquidade ao processo de avaliação, posta a elevada proporção de itens com DIF.

Há que se ressaltar, por relevante, que, dentre os 12 itens com DIF, observou-se que sete destes (58,3% do total de itens com DIF) foram benignos aos componentes do Grupo de Referência (GR), ou seja, aos alunos de escolas particulares. Assim sendo, os sete itens que privilegiaram os respondentes oriundos de escolas particulares (itens 2, 10, 11, 12, 13, 14 e 15), constituem-se, assim, em itens injustos para com os componentes do GF (alunos de escolas públicas), pois proporcionam vantagens indevidas para os respondentes do GR (alunos de escolas particulares). **Para esses sete itens com DIF benigno ao GR, a hipótese H2 [PGRi (θs) > PGFi (θs)] foi corroborada.**

Os demais cinco itens (76,7% dos itens com DIF) beneficiaram os sujeitos do Grupo Focal (GF), isto é, os alunos de escolas públicas. Assim sendo, estes cinco

itens (3, 4, 5, 8 e 9) privilegiam os respondentes oriundos de escolas públicas, constituindo-se, assim, em itens injustos para com os componentes do GR (alunos de escolas particulares), pois ocasionam vantagens indevidas para os respondentes do GF (alunos de escolas públicas). **Para esses cinco itens que possuem DIF adverso, a hipótese H2 [PGRi (θs) > PGFi (θs)] foi rechaçada.**

Para enriquecer as análises efetivadas, determinou-se a magnitude do DIF nos itens do Teste de Metemática acometidos pelo problema, considerando-se a escala do *Educational Testing Service (ETS).* A Tabela 5 apresenta os valores do coeficiente α_{MH} para os itens do Teste de Matemática que possuem DIF.

Tabela 5. Magnitude do DIF dos itens do Teste de Matemática.

Itens	αMH	p	Magnitude do DIF
2	-0,41	0,05	Moderada
3	0,74	0,05	Moderada
4	0,94	0,05	Moderada
5	0,38	0,05	Moderada
8	0,55	0,05	Moderada
9	0,27	0,05	Moderada
10	-0,84	0,05	Moderada
11	-0,63	0,05	Moderada
12	-0,37	0,05	Moderada
13	-0,37	0,05	Moderada
14	-0,22	0,05	Moderada
15	-0,25	0,05	Moderada

Fonte: Pesquisa Direta.

Conforme os resultados, todos os itens componentes do Teste de Matemática com DIF, o possuem em magnitude moderada. Portanto, com base nas análises efetivadas e de modo a concluir este tópico que permitiu o teste empírico da hipótese H2, pode-se asseverar que **a Prova de Matemática apresenta indícios muito contundentes de ser um instrumento injusto, introdutor de iniquidade no processo de avaliação do alunado**, pois beneficia subgrupos específicos, componentes do universo submetido à Prova de Matemática.

Conforme atestam os dados, ora o referido instrumento de avaliação privilegiou os sujeitos do GR (alunos oriundos de escolas particulares), como se deu no caso dos itens 2, 10, 11, 12, 13, 14 e 15, ora beneficiou os indivíduos do GF (alunos oriundos de escolas públicas), como no caso dos itens 3, 4, 5, 8 e 9. Repete-se, portanto, o padrão similar e pitoresco já identificado na Prova de Língua Portuguesa, na qual a Prova de Matemática também se revelou injusta, ora privilegiando os componentes do GR (alunos oriundos de escolas particulares), ora beneficiando os indivíduos do GF (alunos oriundos de escolas públicas).

5.3. SÍNTESE ANALÍTICA DAS HIPÓTESES H1 E H2

Outrora foi referido por autores especialistas nos estudos do DIF, tais como, Camilli e Shepard (1994), Downing e Haladyna (1997), Muñiz (1997), Martínez Arías (1997), O'Neill e McPeek (1993), que não há provas ou testes cujos itens estejam totalmente isentos e livres de vieses ou de DIF. Assim sendo, é altamente provável

que uma prova ou teste de rendimento quase nunca esteja totalmente isento de itens contendo DIF. Consoante Roznowski e Reith (1999) o problema está na proporção de itens que têm DIF, bem como na magnitude destes.

Assim sendo, é extremamente conveniente a adoção de estratégias estatísticas para detectar o DIF e determinar sua magnitude naqueles itens que comporão as provas ou os testes de rendimento, que, por seu turno, serão empregados em sistemáticas de avaliação em larga escala. Desse modo, se outorgará justiça e equidade aos processos de avaliação em larga escala que empregam esses relevantes instrumentos de medida do nível de aprendizado de alunos. Nesse caso, pode-se empregar, de modo muito contundente e sábio, a assertiva atribuída ao escritor latino Publílio Siro: *qui omnes insidias timet, in nullas incidit* (quem teme todas as ciladas não cai em nenhuma).

Diante do exposto, há que se determinar a presença de DIF entre os itens que comporão as provas ou os testes de rendimento, usados em avaliações de larga escala, de modo a evitar injustiças e iniquidades (ANDRIOLA, 2002a; ANDRIOLA, 2006). Conforme realçou Andriola (2009), a determinação do DIF é, portanto, uma fase absolutamente imprescindível de pré-teste dos itens, que antecede o emprego efetivo destes nas provas ou nos testes de rendimento. Ao mesmo tempo, se evitará essa situação pitoresca, na qual dois instrumentos de medida, quais sejam, uma Prova de Língua Portuguesa e uma Prova de Matemática, revelaram-se injustas para com os usuários.

6. PRINCIPAIS CONCLUSÕES E ENCAMINHAMENTOS

A avaliação educacional possibilita às autoridades educacionais prestar contas acerca dos êxitos e discutir as causas do seus fracassos, ação esta que os norteamericanos denominaram *accountability*, como uma ação que permite o exercício da transparência (*glasnost* para os russos) sobre as Políticas Públicas. A avaliação concebida com atividade científica possui fases hierárquicas e peculiaridades, dentre as quais a valoração dos resultados a partir do uso de critérios escolhidos *a priori*.

Enfatizou-se a complexidade do processo de avaliação da aprendizagem e apresentou-se um poderoso procedimento que é o mais empregado pelos docentes: *a prova ou o teste de rendimento*. Mais adiante, contextualizou-se o surgimento dos estudos sobre o viés dos instrumentos de medida da aprendizagem, enfatizando-se o funcionamento diferencial do item (DIF). Destacou-se a noção de *grupo* como sendo fundamental aos estudos sobre o DIF e notificou-se que, geralmente, é organizado um *grupo de referência* (GR) que é comparado a um *grupo focal* (GF). Com base nessa organização dos grupos, surgiu a categorização de *DIF benigno*, quando é benéfico ao GR, e *DIF adverso*, quando é benéfico ao GF.

Ressaltou-se alguns dos problemas ocasionados pela presença do DIF em testes de rendimento empregados em processos de avaliação educacional em larga escala. Descreveu-se o suposto central dos métodos condicionais: *o paradoxo de Simpson*, que defende a idéia de condicionar a probabilidade de acerto ao item a uma determinada pontuação no teste - invariância condicional observada - ou nível de habilidade no construto latente - invariância condicional não observada. Essa é a premissa central no estudo do DIF.

Finalmente, foram destacados os principais resultados decorrentes do teste empírico das duas hipóteses de estudo, a partir do uso do método escolhido para detectar o DIF nos itens componentes das Provas de Língua Portuguesa e de Matemática. Consoante os resultados obtidos, nos quais foi identificada a presença do DIF em proporções elevadas dentre o total de itens das duas provas analisadas, há que se adotarem estratégias para evitar tal problema nas provas ou testes de rendimento a serem empregados em avaliações de larga escala, de modo a coibirem-se injustiças e iniquidades no processo avaliativo.

Destarte, a determinação do DIF, portanto, se constitui em etapa absolutamente imprescindível de análise dos itens, que antecede ao emprego efetivo destes através das provas ou dos testes de rendimento. Assim sendo, ao proceder-se como aconselhado, se dotará de justiça e equidade os processos de avaliação em larga escala que empregam provas de rendimento como instrumentos de medida do nível de aprendizado do alunado.

Dessa forma, para arrematar o relato da pesquisa, desejamos empregar um último adágio, como foi recorrente ao longo do trabalho. Uma prova, mesmo quando bem concebida, provavelmente conte com itens com DIF, pois como asseverou o filósofo e poeta romano Quinto Horácio Flaco (65 a 8 a. C.): *nam vitiis nemo sine nascitur* (ninguém nasce sem defeitos).

REFERÊNCIAS BIBLIOGRÁFICAS

ADEDOYIN, O. O.; ADEDOYIN, J. A. Assessing the comparability between classical test theory (CTT) and item response theory (IRT) models in estimating test item parameters. **Herald Journal of Education and General Studies**, v. 2, n. 3, p. 107-114, 2013.

ADEGOKE, B. A. Comparison of item statistics of physics achievement test using classical test and item response theory frameworks. **Journal of Education and Practice**, v. 4, n. 22, p. 87-96, 2013.

ALEXANDER, P. A.; JUDY, J. E. The interaction of domain-specific and strategic knowledge in academic performance. **Review of Educational Research**, v. 58, n. 4, p. 375–404, 1988.

ANDRIOLA, W. B. Avaliação do aprendizado discente: estudo com professores de Escolas Públicas. **Educar em Revista (Impresso)**, n. 46, p. 141-158, 2012.

ANDRIOLA, W. B. Psicometria moderna: características e tendências. **Estudos em Avaliação Educacional**, v. 20, p. 319-340, 2009.

ANDRIOLA, W. B. Uso da Teoria de Resposta Ao Item (TRI) para Analisar a Eqüidade do Processo de Avaliação do Aprendizado Discente. **Revista Iberoamericana de Evaluación Educativa, v. 1,** p. 171-189, 2008a.

ANDRIOLA, W. B. Propostas estatais voltadas à avaliação do Ensino Superior brasileiro: Breve retrospectiva histórica do período 1983-2008. **Revista Electrónica Iberoamericana Sobre Calidad, Eficacia y Cambio en Educación, v. 6,** p. 127-148, 2008b.

ANDRIOLA, W. B. Estudo sobre o viés de itens em testes de rendimento: uma retrospectiva. **Estudos em Avaliação Educacional, v. 17,** p. 115-134, 2006.

ANDRIOLA, W. B. Detección del funcionamiento diferencial del ítem (DIF) en tests de rendimiento. Aportaciones teóricas y metodológicas. **Tese de Doutorado Publicada (539 p.)**. Madrid: Universidad Complutense de Madrid, 2002a. Texto disponível em <https://eprints.ucm.es/4841/>. Acesso em 24/10/2022.

ANDRIOLA, W. B. Principais métodos para detectar o funcionamento diferencial do item (DIF) no âmbito da avaliação educacional. **Revista Educação em Debate, Fortaleza,** v. 2, n. 44, p. 83-97, 2002b.

ANDRIOLA, W. B. Factores caracterizadores de centros educativos eficaces. **Bordón: Revista de Pedagogía,** Madrid, v. 53, n. 2, p. 175-183, 2001a.

ANDRIOLA, W. B. Determinación del funcionamiento diferencial de ítems (DIF) destinados a la evaluación del razonamiento verbal considerando el tipo de escuela de los alumnos. **Bordón: Revista de Pedagogía, v. 53, n.4,** p. 473-484, 2001b.

ANDRIOLA, W. B. Descrição dos Principais Métodos para Detectar o Funcionamento Diferencial dos Itens (DIF). **Psicologia: Reflexão e Crítica, v. 14, n.3,** p. 643-652, 2001d.

ANDRIOLA, W. B. Funcionamento diferencial dos itens (DIF): estudo com analogias para medir o raciocínio verbal. **Psicologia: Reflexão e Crítica,** Porto Alegre, v. 13, n. 3, p. 473-481, 2000a.

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

ANDRIOLA, W. B. Calidad educacional y efectividad escolar: conceptos y características. **Educação em Debate**, v. 39, n. 1, p. 7-14, 2000c.

ANDRIOLA, W. B. 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, p. 295-308, 1998.

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

ANDRIOLA, W. B. Os testes psicológicos no Brasil: problemas, pesquisas e perspectivas. In L. S. Almeida; I. S. Ribeiro (Org.). **Avaliação Psicológica. Formas e contextos** (p. 77-82). Braga: Associação dos Psicólogos Portugueses, 1995a.

ANDRIOLA, W. B. Problemas e perspectivas quanto ao uso de testes psicológicos no Brasil. **Psique, V. 6,** 46-57, 1995b.

ANDRIOLA, W. B.; ANDRIOLA, C. G. Avaliação da qualidade educacional da Faculdade de Educação (FACED) da Universidade Federal do Ceará (UFC). **Ensaio: Avaliação de Políticas Públicas em Educação**, v. 17, p. 153-168, 2009.

ANDRIOLA, W. B.; BARRETO, J. A. E. Análise métrica de um instrumento de medida da aprendizagem através da Teoria da Resposta ao Item (TRI). **Ensaio: Avaliação de Políticas Públicas em Educação**, v. 14, n. 5, p. 59-74, 1997.

ANDRIOLA, W. B.; LEITE, R. H.; MAIA, J. L. Análise métrica de questões componentes de testes de rendimento: mecanismo de feedback para aprimorar sua elaboração. **Foro Educacional**, v. 21, p. 13-31, 2013.

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

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

ANGOFF, W. H.; FORD, H. Item-race interaction on a test of scholastic aptitude. **Journal of Educational Measurement**, v. 10, p. 95-105, 1973.

AYALA, R. J. **The theory and practice of Item Response Theory**. New York: The Guilford Press, 2009.

BARRETO, J. A. E.; TROMPIERI FILHO, N.; ANDRIOLA, W. B. Desenvolvimento da estrutura cognitiva dos alunos de 4ª e 8ª séries. **Educação em Debate, v. 37, n.1,** p. 101-113, 1999.

BAUER, A.; ALAVARSE, O. M.; OLIVEIRA, R. P.. Avaliações em larga escala: uma sistematização do debate. **Educação e Pesquisa**, v. 41, p. 1367-1382, 2015.

CAMILLI, G.; SHEPARD, L. A. **MMSS. Methods for Identifying Biased Test Items.** California: SAGE Publications, 1994.

CLAUSER, B. E.; NUNGESTER, R. J.; SWAMINATHAN, H. 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, 1996.

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

CREEMERS, B. P. M. **School effectiveness, effective instruction and school improvement in the Netherlands.** Londres: Cassel, 1996.

DE MIGUEL, M.; MADRID, V.; NORIEGA, J.; RODRÍGUEZ, B. **Evaluación para la Calidad de los Institutos de Educación Secundaria.** Madrid: Editorial Escuela Española, 1994.

DORANS, N. J.; HOLLAND, P. W. DIF Detection and Description: Mantel-Haenszel and Standardization (p. 35-66). In P. W. Holland; H. Wainer (Ed.). **Differential Item Functioning.** New Jersey: Lawrence Erlbaum Associates, 1993.

DOUGLAS, J. A.; ROUSSOS, L. A.; STOUT, W. Item-Bundle DIF hypothesis testing: identifying suspect bundles and assessing their differential functioning. **Journal of Educational Measurement**, v. 33, n. 4, p. 465-484, 1996.

DOWNING, S. M.; HALADYNA, T. M. Test item development: Validity evidence from quality assurance procedures. **Applied Measurement in Education**, v. 10, n. 1, p. 61-82, 1997.

DUNKIN, M. J. Student characteristics, classroom processes, and student achievement. **Journal of Educational Psychology**, v. 70, n. 6, p. 998, 1978.

ERCIKAN, K. Translation effects in international assessments. **International Journal of Educational Research**, v. 29, p. 543-553, 1998.

FIDALGO, A. M.; SCALON, J. D. Uso dos Métodos Mantel-Haenszel para a Detecção do Funcionamento Diferencial dos Itens e Software Relacionado Using

Mantel-Haenszel Methods for Detecting Differential Item Functioning (DIF). **Psicologia: Reflexão e Crítica**, v. 25, n. 1, p. 60-68, 2012.

FULLER, B.; CLARKE, P. Raising school effects while ignoring culture? Local conditions and the influence of classroom tools, rules, and pedagogy. **Review of Educational Research**, v. 64, n. 1, p. 119-157, 1994.

HAMILTON, L. S. Detecting gender-based differential item functioning on a constructed-response science test. **Applied Measurement in Education**, v. 12, n. 3, p. 211-235, 1999.

HARVEY, L.; GREEN, D. Defining quality. **Assessment & Evaluation in Higher Education**, v. 18, n. 1, p. 9-34, 1993.

HORSBURGH, M. Quality monitoring in higher education: the impact on student learning. **Quality in Higher Education**, v. 5, n. 1, p. 9-25, 1999.

KLEIN, R. Alguns aspectos da Teoria de Resposta ao Item relativos à estimação das proficiências. **Ensaio: Avaliação e Políticas Públicas em Educação**, v. 21, n. 78, p. 35–56, 2013.

LIMA, L. A.; ANDRIOLA, W. B. Acompanhamento de egressos: subsídios para a avaliação de Instituições de Ensino Superior (IES). **Avaliação: Revista da Avaliação da Educação Superior**, v. 23, p. 104-125, 2018.

LIMA, L. A.; ANDRIOLA, W. B.; TAVARES, W. A. Melhorando o processo de ensino e aprendizado em cursos de graduação na área de computação por meio da utilização de edublogs. **Revista Ibero-Americana de Estudos em Educação**, v. 10, p. 816-841, 2015.

LINN, R. L.; HARNISCH, D. L. Interactions between item content and group membership on achievement test items. **Journal of Educational Measurement**, v. 18, p. 109-118, 1981.

LONGFORD, N. T.; HOLLAND, P. W.; THAYER, D. T. Stability of the MH D-DIF statistics across populations (pp. 171-196). In P. W. Holland; H. Wainer (Ed.). **Differential Item Functioning.** New Jersey: Lawrence Erlbaum Associates, 1993.

MACHADO, C.; ALAVARSE, O. M. Qualidade das Escolas: tensões e potencialidades das avaliações externas. **Educação & Realidade**, v. 39, n. 2, p. 413–436, 2014.

MARTÍNEZ ARÍAS, R. *Psicometría. Teoría de los Tests Psicológicos y Educacionais.* Madrid: Ediciones Síntesis, 1997.

MILES, R. E. **Theories of management: Implications for organizational behavior and development**. New York: McGraw-Hill, 1974.

MORTIMORE, P. **Issues in school effectiveness.** London: Cassel, 1996.

MOURA CASTRO, C. Escolas feias, escolas boas? **Ensaio: Avaliação e Políticas Públicas em Educação**, v. 7, n. 25, p. 343-354, 1999.

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

MURPHY, J. School effectiveness and school reestructuring: contributions to educational improvement. **School Effectiveness and School Improvement**, v. 3, n. 2, p. 90-109, 1992.

NUTHALL, G. Relating learning to individual differences in ability. **Journal of Educational Research**, v. 31, n. 3, p. 212-255, 1999.

NUTHALL, G.; ALTON-LEE, A. Assessing classroom learning: How students use their knowledge and experience to answer classroom achievement test questions in science and social studies. **American Educational Research Journal**, v. 32, n. 1, p. 185–223, 1995.

O'NEILL, K. A.; McPEEK, W. M. Item and Test Characteristics that are associated with Differential Item Functioning (p. 255-276). In P. W. Holland; H. Wainer (Ed.). **Differential Item Functioning.** New Jersey: Lawrence Erlbaum Associates, 1993.

ORDEN HOZ, A. Calidad y evaluación de la enseñanza universitaria. **Resumenes del Congreso Internacional de Universidades**, julio, p. 531-539. Madrid: Universidad Complutense de Madrid, 1992.

POPPER, K. R. **A lógica da pesquisa científica.** São Paulo, Editora Cultrix, 1972.

POPPER, K. R. **Conjecturas e refutações**. Brasília: Editora da Unb, 2008.

RIZOPOULOS, D. ltm: An R package for latent variable modeling and item response theory analyses. **Journal of Statistical Software**, v. 17, n. 5, p. 1-25, 2006.

ROZNOWSKI, M.; REITH, J. 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, 1999.

SALES, G. L.; BARROSO, G. C.; SOARES, J. M.; ANDRIOLA, W. B.; JOYCE, C. R. Um Indicador de Aprendizagem Não-Linear para EaD Online Fundamentado no Modelo de Avaliação Learning Vectors (LV). **Revista Iberoamericana de Evaluación Educativa**, v. 4, p. 155-180, 2011.

SAMMONS, P.; HILLMAN, J.; MORTIMORE, P. **Key characteristics of effective schools: a review of school efectiveness research.** London: Office for Standards in Education, 1995.

SAMPAIO, R.; MANCINI, M. Estudos de revisão sistemática: um guia para síntese criteriosa da evidência científica. **Revista Brasileira de Fisioterapia**, v. 11, n. 1, p. 83–89, 2007.

SCHEERENS, J. **Effective schooling: research theory and practice**. London: Cassell, 1992. 168 p.

SCHEUNEMAN, J. D.; GERRITZ, K. 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, 1990.

SILVA, F. C. M.; LIMA, A. S.; ANDRIOLA, W. B. Avaliação do suporte de TDIC na formação do pedagogo. Um estudo em universidade brasileira. **Revista Electrónica Iberoamericana sobre Calidad, Eficacia y Cambio en Educación**, v. 14, n. 3, p. 77-93, 2016.

SILVA, T. E. V.; RIBEIRO, G. O.; NUNES, A. O.; LIMA, F. H. V.; ANDRIOLA, W. B.; MOTA, J. C. M. QEOn Questionnaire for Assessing Experiences in Virtual Learning Environments. **IEEE Latin America Transactions**, v. 15, p. 1197-1204, 2017.

SOARES, J. F. (Coord). **Escola eficaz: um estudo de caso em três escolas da rede pública do Estado de Minas Gerais**. Belo Horizonte: Game/FAE/UFMG, Segrac, 2002.

TOBIAS, S. Interest, prior knowledge, and learning. **Review of Educational Research**, v. 64, n. 1, p. 37–54, 1994.

TYLER, R. W. **Basic Principles of Curriculum and Instruction.** Chicago: The University of Chicago Press, 1949.

VIANNA, H. M. Avaliações nacionais em larga escala: análises e propostas. **Estudos em Avaliação Educacional**, n. 27, p. 41–76, 2003.

ZUMBO, B. D. **A Handboock 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, 1999.