ARTIGOS

articles

ESTUDO DO FUNCIONAMENTO DIFERENCIAL DO ITEM EM TESTES COM ITENS DICOTÔMICOS

STUDY OF ITEM DIFFERENTIAL FUNCTIONING ON TESTS WITH DICHOTOMOUS ITEMS

ESTUDIO DEL FUNCIONAMIENTO DIFERENCIAL DEL ÍTEM EN PRUEBAS CON ÍTEMS DICOTÓMICOS

Cácio Fabrício Gomes da Rocha[1]

Adriano Ferreti Borgatto[2]

RESUMO

O objetivo do presente estudo foi verificar o impacto do Funcionamento Diferencial do Item (DIF) na estimação da proficiência. Para identificar itens comuns com DIF, foi adotada a metodologia utilizada pelo INEP, que consiste em comparar as proporções esperadas de acertos ao item para cada grupo no intervalo P5 (percentil 5) e o P95 (percentil 95). Para tal comparação, foram simulados dois grupos (referência e focal), em que o grupo referência foi submetido a um teste com 45 itens e o grupo focal a um teste parcialmente diferente aos itens do grupo referência com 45 itens, sendo introduzido DIF em itens em diferentes pontos da escala. Os resultados mostraram que o número de itens comuns auxilia no processo de identificação dos itens com DIF, de modo que quanto mais itens de ligação (comum), menor será o impacto dos itens que apresentam algum viés.

Palavras-chave: Teoria de Resposta ao Item, Funcionamento Diferencial do Item, Avaliação educacional.

1 Bacharelado em Estatística pela Universidade Federal do Pará e Mestrado em Métodos em Gestão e Avaliação pela Universidade Federal de Santa Catarina. Atualmente, é analista técnico terceirizado no INEP/MEC.

2 Graduação em Estatística pela UNESP, mestrado em Estatística e Experimentação Agropecuária pela UFLA, doutorado em Agronomia (Estatística e Experimentação Agronômica) pela ESALQ/USP. Atualmente é professor titular da Universidade Federal de Santa Catarina. Sua principal área de pesquisa é a Teoria da Resposta ao Item, atuando principalmente em temas como estilo de vida, qualidade de vida e ensino.

ABSTRACT

The objective of the present study was to verify the impact of Differential Item Functioning (DIF) in estimating proficiency. To identify common items with the DIF, it was adopted as a methodology used by INEP, which consists of comparing the expected proportions to the item for each group without interval P5 (percentile 5) and P95 (percentile 95). For comparison, two groups were simulated (reference and focus), in which the reference group was subjected to a test with 45 items and the focus group to an isolated test different from the items in the reference group with 45 items, with DIF being introduced in items at different points on the scale. The results have shown that the number of common items assists in the process of identifying items with DIF, since the more link items (common) the less the impact of the items that will be displayed at some point.

Keywords: Item Response Theory, Differential Item Functioning, Educational Assessment.

RESUMEN

El objetivo del estudio fue verificar el impacto del funcionamiento diferencial de elementos (DIF) en la estimación de la competencia. Se adoptó la metodología utilizada por el INEP, que consiste en comparar las proporciones esperadas de recursos con el elemento para cada grupo sin intervalo P5 (percentil 5) y P95 (percentil 95). Para la comparación, se simularon dos grupos (referencia y enfoque), en los que el grupo de referencia se sometió a una prueba con 45 ítems y el grupo de enfoque a una prueba aislada diferente de los ítems en el grupo de referencia con 45 ítems, introduciéndose DIF en artículos en diferentes puntos de la escala. Los resultados mostraron que la cantidad de elementos comunes auxilia en el proceso de identificación de elementos con DIF, puesto que cuantos más elementos de enlace (comunes) menor será el impacto de los elementos que se mostrarán en algún momento.

Palabras clave: Teoría de respuesta al ítem, Funcionamiento diferencial del ítem, Evaluación educativa.

1. Introdução

A avaliação educacional em larga escala, no decorrer do século XX e, sobretudo, no XXI, tem produzido resultados que viabilizam a compreensão dos mais diversos aspectos em torno da qualidade da educação ofertada à população de diferentes países. No Brasil, diante do potencial das informações disponibilizadas pelas avaliações, essas informações ganharam papel de destaque na agenda pública, visto que subsidiam o planejamento estratégico e o monitoramento da política educacional.

Um dos principais instrumentos utilizados pelas avaliações educacionais, coordenadas pelo Governo Federal, são os testes de desempenho nas áreas de conhecimento em Língua Portuguesa e Matemática, consideradas basilares no decorrer do processo de escolarização das crianças e jovens. Dada a importância das avaliações no âmbito da política educacional, é fundamental que se garantam a uniformização e a padronização dos instrumentos utilizados para aferir o desempenho dos estudantes, tal como assevera Pasquali (2000).

Em 1995, a metodologia de correção do Sistema de Avaliação da Educação Básica (SAEB) foi alterada com a adoção da Teoria de Resposta ao Item (TRI) e, em 1997, a construção dos itens dos testes pautava-se em uma Matriz de Referência, na qual estavam listadas as habilidades a serem avaliadas. Os resultados do SAEB eram divulgados para região e dependência administrativa, desagregados até o nível das redes de ensino, tendo em vista seu desenho amostral.

Para se comparar avaliações de diferentes anos escolares e diferentes épocas, um procedimento adotado nas avaliações em larga escala que utilizam a TRI é a equalização. Os procedimentos de equalização contribuíram significativamente no avanço das avaliações educacionais em larga escala, uma vez que propicia que indivíduos avaliados por instrumentos de avaliação parcialmente diferentes (com alguns itens em comum) sejam colocados numa mesma escala, o que permite compará-los e acompanhar a sua evolução ao longo do tempo (ANDRADE; TAVARES; VALE, 2000; EMBRETSON; REISE, 2000).

Diante do avanço nas áreas de conhecimento, como a psicometria e a

estatística, surgem estudos preocupados com o processo de construção e aplicação dos testes, de maneira a assegurar a validade da medida realizada. Nesse sentido, evidenciou-se a necessidade de buscar uniformidade na elaboração dos itens utilizados no teste, padronizar os comandos e enunciados, planejar a estrutura do teste, além de controlar as situações relacionadas à aplicação do teste (MARTÍNEZ ARIAS, 1997).

De acordo com Hambleton (1997), um dos campos de investigação que emerge com foco na padronização das condições de aplicação de instrumentos de medida é o estudo do DIF. Esse estudo identifica os itens em que a probabilidade de acerto dos indivíduos que apresentam o mesmo nível de uma determinada proficiência ou aptidão medida é diferente, a depender do subgrupo da população-alvo da avaliação em que estes se inserem.

Nesse sentido, como apontam Hambleton (1997), Andriola (2001), Douglas, Roussos e Stout (1996), a existência de DIF em um item torna o processo de avaliação injusto, pois indica que determinados grupos estão sendo privilegiados. De acordo com Muñiz (1997), um item que apresenta DIF indica falhas na padronização e nas condições de uniformização da aplicação do teste, que tem como propósito captar a aptidão ou proficiência do sujeito. Diante disso, é possível dizer que, dada a importância dos resultados gerados pelas avaliações para a política educacional, a existência de DIF, nos itens de um teste, acarreta prejuízos do ponto de vista dos recursos aplicados e do planejamento das ações.

Os estudos que se dedicam à investigação das possíveis interferências na aferição de uma aptidão ou proficiência dos indivíduos são vastos, sendo que os primeiros datam do início do século XX. Segundo Martínez Arias (1997) e Sisto (2006), já em 1905, Alfrad Binet, em seus estudos sobre inteligência, averiguou que crianças de baixo nível socioeconômico tinham menor rendimento em alguns itens dos testes a que eram submetidas, o que o levou a levantar hipóteses de os itens não estarem medindo de fato a aprendizagem das crianças, mas sim questões de ordem cultural. William Stern, considerado um dos pioneiros da psicologia da personalidade e inteligência, apontou que os testes aplicados na Alemanha poderiam favorecer um determinado grupo de pessoas, tendo em vista sua classe social.

Na área de avaliação, Valle (2002) utiliza dados do SAEB com o objetivo de explicitar a importância da utilização do DIF para educadores nos resultados das avaliações em larga escala. Mais recentemente, Andriola (2018) ressalta que a presença do DIF em sistemáticas de avaliação do aprendizado discente pode supor iniquidade deste processo, sobretudo nas avaliações em larga escala.

Essa preocupação com a precisão da medida da proficiência ou aptidão do indivíduo em si, sem a interferência de outros fatores, suscitou a emergência de estudos sobre o viés do item. De acordo com Sisto (2006), os estudiosos Eells, Havighurst, Herrick e Tyler (1951) são representantes da moderna investigação sobre o viés. O autor demarca que esses estudiosos, ao analisarem testes de inteligência, identificaram que as variações nos itens referentes a formato e conteúdo, por exemplo, poderiam atenuar ou aumentar a diferença entre os grupos que respondiam ao teste. Nesse sentido, os testes poderiam estar medindo muito mais as diferenças de oportunidades de aprendizagem do que, necessariamente, a aptidão dos sujeitos, como pretendia.

A relevância em se analisar o DIF de itens em avaliações de larga escala advém do pensamento de Hambleton (1997) e de Andriola (2001), que defendem a possibilidade de validar a medida gerada pelos testes, indicando a não reutilização de itens com DIF em outras avaliações. Dessa forma, é viável investigar as possíveis causas do DIF, bem como viabilizar o controle dos fatores que acarretam o problema.

Faz-se relevante assinalar que Andriola (2001), pautado em Muñiz (1997), ressalta que não há itens ou testes totalmente isentos de DIF. Na realidade, ao se detectar o grau do DIF, é possível analisar se o número de itens com DIF é aceitável diante dos objetivos da avaliação realizada.

Dessa forma, surge a necessidade de estudar técnicas para detectar os itens que apresentam DIF. Portanto, a identificação de itens com DIF é de grande importância no ajuste de modelos da TRI que são utilizados nas avaliações de larga escala no Brasil, pois essa diferença sistemática pode comprometer toda a inferência realizada, como o estabelecimento dos parâmetros dos itens e a proficiência dos estudantes, além de violar os processos avaliativos que dependem dos resultados dessas avaliações.

Assim sendo, o objetivo deste trabalho é o de analisar o impacto na estimação da proficiência ao utilizar itens que contenham funcionamento diferencial em grupos distintos. Para responder ao objetivo, as simulações foram realizadas com diferentes quantidades de itens comuns entre os testes e diferentes tamanhos (amplitude) do DIF entre os grupos. O estudo de simulação limitou-se a utilizar o modelo logístico de 3 parâmetros da TRI, sendo esse o modelo mais comum utilizado em avaliação em larga escala com itens de múltipla escolha.

ESTUDO DE SIMULAÇÃO

Neste estudo, por se tratar de uma simulação nos moldes de uma avaliação em larga escala, como o Exame Nacional do Ensino Médio (ENEM), será utilizado o Modelo Logístico de 3 parâmetros, que é o modelo mais utilizado em avaliação em larga escala para itens dicotomizados (ANDRADE; TAVARES; VALE, 2000). Segue a fórmula matemática:

$$P\left(X_{kij}=1\middle|\theta_{kj},a_i,b_i,c_i\right)=c_i+(1-c_i)\frac{1}{1+\exp^{-a_i(\theta_{kj}-b_i)}}$$

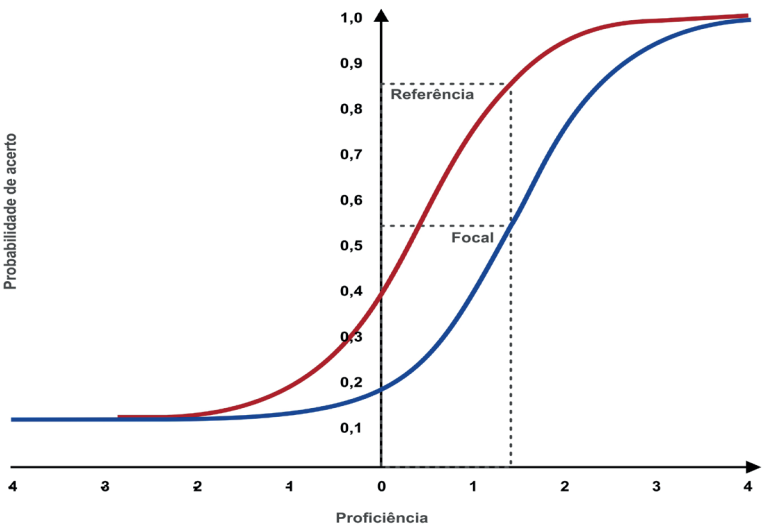
onde, θ_{kj} é a proficiência do *j*-ésimo indivíduo da população *k*, a_i o parâmetro de discriminação do item i, b_i o parâmetro de dificuldade do item i e c_i o parâmetro de acerto ao acaso do item i.

Existem diversos métodos de detecção de DIF, os métodos baseados nos modelos da TRI fornecem uma abordagem abrangente na investigação dos parâmetros dos itens de um teste entre os grupos avaliados (EVERSON; OSTERLIND, 2009). No âmbito da TRI, é possível dizer que o item não apresenta DIF quando a curva característica do item (CCI) é a mesma para os grupos comparados em um mesmo nível de proficiência. Segundo Magis (2010), existem dois tipos de DIF, o uniforme e o não uniforme.

O DIF uniforme, que será o foco do estudo de simulação por se tratar do tipo de DIF mais comum na literatura, ocorre quando as CCIs analisadas para o grupo de referência e para o grupo focal são diferentes e não se

cruzam em nenhum ponto da escala de proficiência. Esse caso ocorre quando o valor do parâmetro *a* (discriminação do item) é o mesmo nas duas CCIs, ou seja, as curvas são paralelas. A Figura 1 representa as CCIs de um item com DIF uniforme ou consistente em uma escala (0,1). Observa-se que as curvas são paralelas e a CCI do grupo referência está situada mais à esquerda que a CCI do grupo focal, o que indica que o item é mais difícil para o grupo focal em todos os níveis de proficiência. Essa diferença aponta que o item apresenta DIF, sendo que, nesse caso, é favorável ao grupo de referência. De acordo com a Figura 1, os indivíduos com proficiências iguais a 1,5, nos dois grupos, têm probabilidades diferentes de acertar o item, o grupo focal tem 55% e o grupo de referência, 85%, o que caracteriza um comportamento anômalo desse item. Em outros pontos da escala, além do valor 1,5, também essa diferença na probabilidade entre os dois grupos apresenta-se muito grande.

FIGURA 1 - REPRESENTAÇÃO GRÁFICA DE UM ITEM COM DIF UNIFORME.



Fonte: elaborado pelo autor

O INEP adota o critério de que, se a diferença na probabilidade esperada (calculada pelo modelo) entre dois grupos é maior do que 15%, o item apresenta DIF. É importante salientar que o foco deste estudo é analisar o impacto do DIF na estimação da proficiência, portanto se limitou a analisar esse impacto utilizando o procedimento adotado pelo INEP, por se tratar da técnica empregada nas avaliações em larga escala de maior impacto no Brasil, que é o SAEB e o ENEM.

Resumidamente, o processo de análise utiliza um arquivo gerado no *software* Bilog-MG chamando *"Expected"*. Esse arquivo fornece as proporções esperadas de respostas corretas ao item para cada grupo em alguns pontos da escala. Essas proporções esperadas são comparadas em pontos da escala com maior concentração de indivíduos, ou seja, na intersecção do intervalo percentil 5 (P5) e percentil 95 (P95) dos grupos.

Todas as análises que serão apresentadas nos resultados são realizadas por meio de uma escala (0,1) do grupo de referência. O processo de simulação dos parâmetros dos itens e das proficiências dos indivíduos foi implementado computacionalmente no *software R*, versão 3.5.2. Os dados do grupo referência foram gerados a partir de 100.000 respostas, considerando um teste com 45 itens novos e admitiu-se que a proficiência (θ) segue uma distribuição normal com parâmetros $(\mu;\sigma^2)=(0;1)$. A quantidade de respostas simuladas neste estudo é suficientemente grande para reproduzir uma avaliação em larga escala e a quantidade de itens foi estipulada considerando uma avaliação em larga escala nos moldes do ENEM.

Para o grupo focal, foi simulada 100.000 respostas, considerando um teste parcialmente diferente com 45 itens e com a proficiência $\boldsymbol{(\theta)}$ seguindo uma distribuição normal com parâmetros $(\mu;\sigma^2)=(0;2)$. O grupo focal foi simulado com uma variabilidade maior do que a do grupo referência, a fim de analisar o comportamento do DIF com proficiências em pontos mais extremos na escala (0,1).

Para os dois grupos, os parâmetros de discriminação, dificuldade e acerto casual foram gerados a partir de uma distribuição uniforme:

I) parâmetro a: [0,8 ; 2,0];

II) parâmetro b: [-4,0 ; 4,0];

III) parâmetro c: [0,10 ; 0,25].

Os intervalos de variação dos parâmetros dos itens foram definidos de acordo com os valores mais plausíveis para cada parâmetro. A escolha desse intervalo foi determinada por valores mais comuns dos itens de múltipla escolha encontrados nas avaliações de larga escala.

Para verificar a qualidade da estimação dos parâmetros dos itens, foram calculados os erros quadráticos médios (EQM), a partir da fórmula matemática:

$$EQM_{\varphi} = \frac{1}{I}\sum_{i=1}^{I}(\hat{\varphi} - \varphi)^2.$$

onde φ representa o valor original do parâmetro do item, $\hat{\varphi}$ o valor estimado do parâmetro do item e I o número de itens.

RESULTADOS

Nesta seção, serão apresentados os resultados de calibração e equalização dos itens simulados para o grupo referência e o grupo focal. A calibração e a equalização dos itens foram implementadas no *software* Bilog-MG.

A Tabela 1 apresenta os valores dos parâmetros originais e os parâmetros estimados via Máxima Verossimilhança Marginal dos 45 itens novos do teste do grupo referência. Nota-se que, de forma geral, as estimativas dos parâmetros dos itens foram satisfatórias, consolidando a consistência dos dados simulados.

TABELA 1 - VALORES ORIGINAIS E ESTIMADOS DOS PARÂMETROS DOS ITENS

VALORES ORIGINAIS SIMULADOS DOS PARÂMETROS DOS ITENS								VALORES ESTIMADOS DOS PARÂMETROS DOS ITENS							
Item	a_i	b_i	c_i	Item	a_i	b_i	c_i	Item	a_i	b_i	c_i	Item	a_i	b_i	c_i
1	0,872	-0,479	0,142	24	1,738	-1,469	0,152	1	0,887	-0,468	0,185	24	1,751	-1,408	0,216
2	1,520	0,335	0,152	25	1,631	0,474	0,161	2	1,554	0,351	0,180	25	1,619	0,497	0,176
3	1,861	0,844	0,165	26	1,974	-0,076	0,221	3	1,835	0,847	0,175	26	1,967	-0,066	0,177
4	1,920	-1,583	0,196	27	1,598	1,435	0,140	4	1,905	-1,623	0,153	27	1,591	1,436	0,171
5	1,224	1,360	0,242	28	1,718	1,692	0,205	5	1,217	1,351	0,173	28	1,644	1,708	0,171
6	2,313	1,107	0,145	29	1,752	0,023	0,222	6	2,277	1,116	0,172	29	1,739	0,028	0,178
7	1,509	-0,564	0,246	30	1,398	-1,495	0,128	7	1,466	-0,587	0,169	30	1,381	-1,493	0,174
8	0,933	-1,539	0,185	31	2,071	-2,205	0,124	8	0,915	-1,599	0,149	31	2,102	-2,193	0,178
9	1,560	0,565	0,136	32	1,856	-0,895	0,106	9	1,532	0,560	0,175	32	1,849	-0,892	0,175
10	0,800	-0,689	0,123	33	2,031	-0,284	0,115	10	0,761	-0,819	0,135	33	2,006	-0,301	0,164
11	2,267	-1,887	0,241	34	2,113	0,500	0,165	11	2,263	-1,872	0,193	34	2,123	0,512	0,178
12	0,830	-0,223	0,204	35	0,816	-0,255	0,225	12	0,837	-0,158	0,192	35	0,797	-0,274	0,162
13	2,060	-1,164	0,122	36	1,443	1,259	0,176	13	2,074	-1,150	0,178	36	1,478	1,277	0,181
14	2,089	0,442	0,134	37	2,122	-0,318	0,176	14	2,091	0,462	0,181	37	2,103	-0,322	0,171
15	2,236	0,352	0,180	38	2,134	1,051	0,156	15	2,262	0,356	0,177	38	2,153	1,054	0,178
16	1,757	1,637	0,242	39	0,950	-0,037	0,146	16	1,668	1,674	0,172	39	0,947	-0,064	0,168
17	1,615	-0,405	0,141	40	2,169	0,487	0,103	17	1,623	-0,412	0,176	40	2,174	0,497	0,177
18	2,436	0,336	0,244	41	2,125	0,418	0,141	18	2,440	0,340	0,174	41	2,095	0,417	0,174
19	1,644	0,237	0,241	42	0,957	-0,540	0,157	19	1,633	0,227	0,169	42	0,969	-0,483	0,196
20	2,265	1,281	0,236	43	1,903	1,988	0,243	20	2,258	1,291	0,175	43	1,898	2,022	0,177
21	1,483	-0,248	0,182	44	1,764	-1,724	0,206	21	1,473	-0,262	0,174	44	1,804	-1,657	0,221
22	2,076	-0,782	0,106	45	0,992	3,067	0,241	22	2,006	-0,823	0,147	45	1,000	3,119	0,177
23	2,119	-1,172	0,137					23	2,120	-1,183	0,172				

Fonte: elaborado pelo autor

Para cada parâmetro estimado, foram encontrados EQM_a=0,008, EQM_b=0,001 e EQM_C=0,002, sugerindo que os resultados obtidos foram adequados ou satisfatórios, pois quanto mais próximos de zero os valores dos erros quadráticos médios, melhor é a precisão da estimação dos parâmetros dos itens.

Como as estimativas dos parâmetros simulados retornaram os valores originais de forma satisfatória, procederam-se as análises para detecção do DIF, nos cenários descritos a seguir.

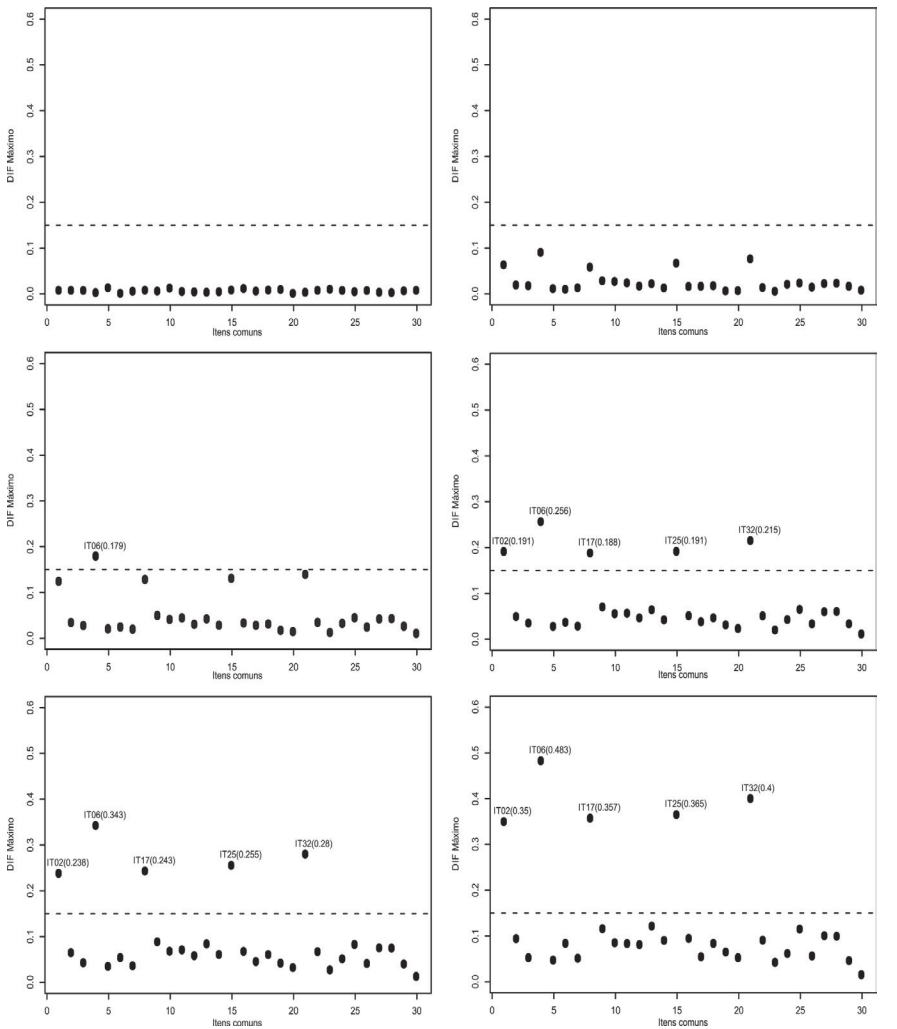
SIMULAÇÃO: CENÁRIO 1

Para o estudo do cenário 1, no teste do grupo focal, foram utilizados 30 itens comuns, ou seja, 66,7% itens comuns com o teste do grupo referência. Dentre os itens comuns, foram selecionados 5 itens (2, 6, 17, 25 e 32) em diferentes pontos da escala, em que foi introduzido o DIF. Para cada um dos 5 itens, foram adicionados incrementos no parâmetro b na magnitude de: 0,0; 0,25; 0,50; 0,75; 1,0 e 1,5, sendo que quanto maior o incremento, maior será o DIF entre grupos. O processo de identificação do DIF uniforme (diferenças somente no parâmetro b) nos itens ocorreu pela verificação da diferença nas proporções esperadas de acertos para os grupos analisados.

Como nas análises das avalições em larga escala, o item apresenta DIF uniforme se a diferença máxima em algum ponto da escala no intervalor P5 e P95 for maior que 0,15. Na Figura 2, é apresentada a diferença máxima encontrada para cada um dos incrementos citados. Para uma melhor identificação dos itens com DIF uniforme, foi inserida uma linha que indica o limite de 0,15.

.

FIGURA 2 - DIFERENÇA MÁXIMA DAS PROPORÇÕES DE ACERTOS DOS ITENS COMUNS, COM OS INCREMENTOS (0,0; 0,25; 0,50; 0,75; 1,0; E 1,5), RESPECTIVAMENTE



Fonte: elaborado pelo autor

O primeiro incremento utilizado reflete uma análise sem a presença de itens com DIF uniforme. Quando se adiciona 0,25 no parâmetro de dificuldade, percebe-se um aumento das diferenças máximas nas proporções de acerto não só dos itens inicialmente sinalizados com DIF, mas nos

demais itens comuns; no entanto, todos apresentam diferenças máximas abaixo de 0,10, o que poderia, em situações reais, ser considerado apenas como flutuação aleatória.

Para análise com a adição de 0,50 no parâmetro b, apenas um item é identificado com DIF uniforme, de modo que os demais itens têm suas proporções alteradas, mas não é identificado como DIF, segundo os critérios adotados no trabalho. Nas análises utilizando incrementos a partir de 0,75, os 5 itens selecionados inicialmente com DIF são identificados na análise.

Como forma de minimizar a perda no número de itens comuns, assim como apresentado por Magis (2010) em outros métodos de detecção de DIF, utilizou-se o processo de purificação, ou seja, a retirada iterativa de itens com DIF. Tal procedimento, porém, não apresentou melhora nos resultados. A retirada parcial ou total dos itens com DIF fez-se necessária para melhorar o processo de equalização.

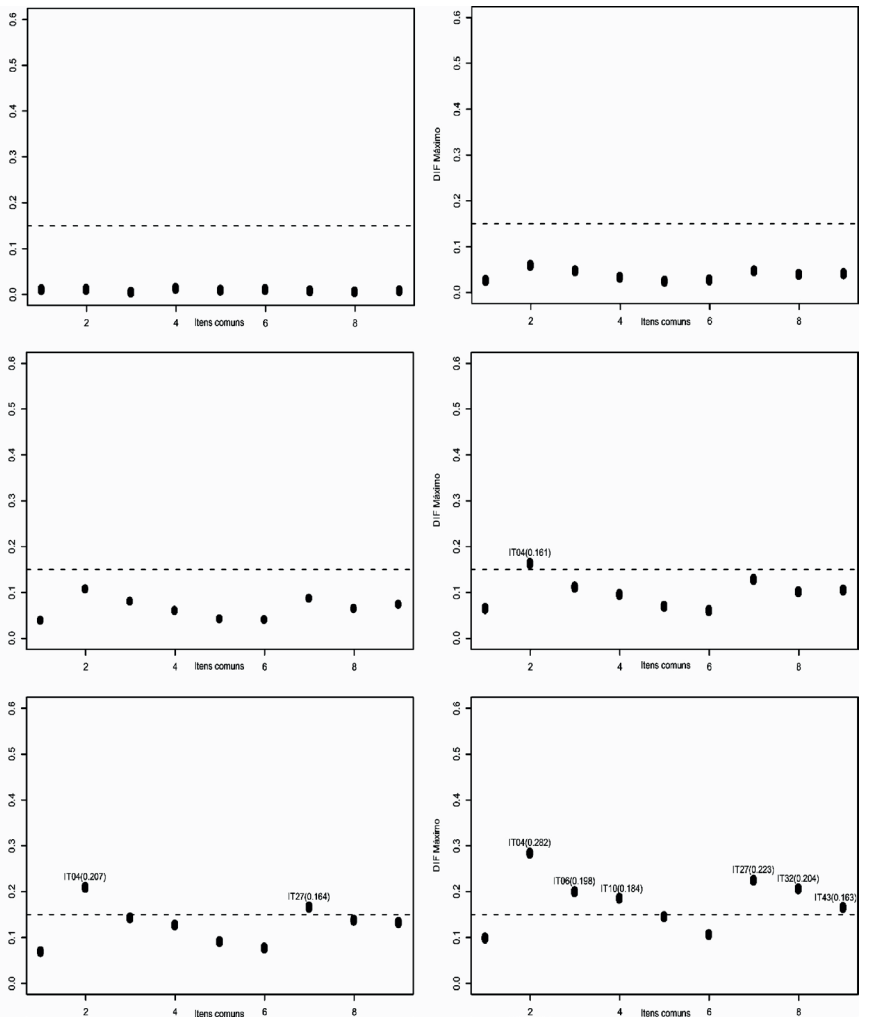
SIMULAÇÃO: CENÁRIO 2

Para o estudo do cenário 2, no teste do grupo focal, foram utilizados 9 itens comuns, ou seja, 20% dos itens comuns com o teste do grupo referência. Esse percentual de itens comuns foi sugerido por alguns autores, conforme apontam Andrade, Tavares e Vale (2000), no entanto as simulações feitas para chegar ao número mínimo de itens comuns foram com itens sem a presença de DIF. No processo de análise desse cenário, por conveniência, os 5 itens selecionados com DIF foram os mesmos itens utilizados no cenário 1 (2, 6, 17, 25 e 32) para receber os incrementos no parâmetro b. Para cada um dos 5 itens, foram adicionados os mesmos incrementos do Cenário 1, ou seja, 0,0; 0,25; 0,50; 0,75; 1,0 e 1,5. O processo de identificação do DIF uniforme nos itens comuns ocorreu pela verificação da diferença nas proporções esperadas de acertos para os grupos analisados.

Na Figura 3, é apresentada a diferença máxima encontrada para cada um dos incrementos utilizados nas análises. Para uma melhor identificação

dos itens com DIF uniforme, foi inserida uma linha com o limite de 0,15.

FIGURA 3 - DIFERENÇA MÁXIMA DAS PROPORÇÕES DE ACERTOS DOS ITENS COMUNS, COM OS INCREMENTOS (0,0; 0,25; 0,50; 0,75; 1,0; E 1,5), RESPECTIVAMENTE, COM 20,0% DE ITENS COMUNS



Fonte: elaborado pelo autor

O primeiro incremento utilizado reflete uma análise sem a presença de itens com DIF uniforme, visto que, como no primeiro cenário, a diferença se deu no número de itens comuns utilizados para equalizar o teste do Grupo Focal. Utilizando o número mínimo de itens comuns, a metodologia adotada consegue captar a presença de um item (Item 4) com DIF, a partir do incremento de 0,75; no entanto, esse item, identificado com funcionamento diferencial, originalmente, não apresentava DIF uniforme, ocasionando uma identificação falsa, ou seja, ocorreu o erro tipo I por conta dos itens problemáticos. Define-se como erro tipo I quando o método identifica DIF nos itens que não possuem DIF. Em um teste real, o Item 4 seria excluído ou considerado como item novo no Grupo Focal. Vale ressaltar que, ao excluir ou retirar a ligação desse item, a qualidade da equalização estaria comprometida devido ao quantitativo final de itens comuns.

A quantidade de itens comuns no cenário 1 ajudou a manter a qualidade no processo de equalização, pois os itens que não sofreram incremento no parâmetro b não foram afetados de forma significativa, ou seja, não foram classificados com falso DIF. No cenário 2, o número reduzido de itens sem incremento apresentou diferenças máximas acima do ponto de corte estabelecido, de modo que, com o aumento no valor do parâmetro b, a taxa de itens com erro tipo I cresceu. Sendo assim, quanto maior o número de itens comuns no processo de equalização, mais sensível se torna a identificação de itens com DIF.

IMPACTO DO DIF NA ESTIMAÇÃO DA PROFICIÊNCIA

Para analisar o impacto na estimação das proficiências do grupo focal em um teste cognitivo que apresenta itens com DIF, foram utilizados três tratamentos nos dados simulados no cenário 1. Como abordado anteriormente, para o cenário 1, as respostas dos indivíduos que compõem o grupo focal foram simuladas por meio do incremento de 0,50 no parâmetro de dificuldade dos itens 2, 6, 17, 25 e 32, que são comuns com o grupo de referência.

O primeiro tratamento consistiu em não retirar das análises os itens com DIF do teste simulado no cenário 1, ou seja, a presença dos itens com DIF

foi ignorada. Logo, as proficiências dos indivíduos do grupo focal foram estimadas com base nos parâmetros fixados dos itens com DIF no teste composto por 45 itens (15 diferentes + 30 comuns).

No segundo tratamento, os itens com DIF foram considerados novos para o grupo focal, ou seja, a ligação com o grupo de referência foi excluída, e, para esses itens, foram estimados novos parâmetros no processo de equalização. Nesse tratamento, as proficiências dos indivíduos do grupo focal foram calculadas considerando 20 diferentes + 25 comuns.

Para o terceiro tratamento, os itens que apresentaram DIF foram excluídos do teste do grupo focal e, consequentemente, a ligação com o teste do grupo de referência foi retirada, de modo que, para geração das proficiências, fixou-se apenas 40 itens (15 diferentes + 25 comuns).

Na Tabela 2, são apresentadas as médias por faixa de proficiência do grupo focal para cada tipo de tratamento.

TABELA 2 -MÉDIA DO GRUPO FOCAL POR FAIXA DE PROFICIÊNCIA **(θ)**

Faixa	Proficiência média				
	Proficiência Original	Sem DIF	Tratamento		
			Não Excluído	Recalibrado	Excluído
θ < -2	-3,0504	-2,1680	-2,1423	-2,1733	-2,1348
-2 >= θ < -1	-1,4703	-1,3362	-1,2692	-1,3190	-1,2897
-1 >= θ < 0	-0,4881	-0,4726	-0,3809	-0,4409	-0,4333
0 >= θ < 1	0,4872	0,4561	0,5460	0,4870	0,4844
1 >= θ < 2	1,4673	1,3810	1,4615	1,4101	1,3979
θ >= 2	3,0568	2,3036	2,3387	2,3072	2,2904

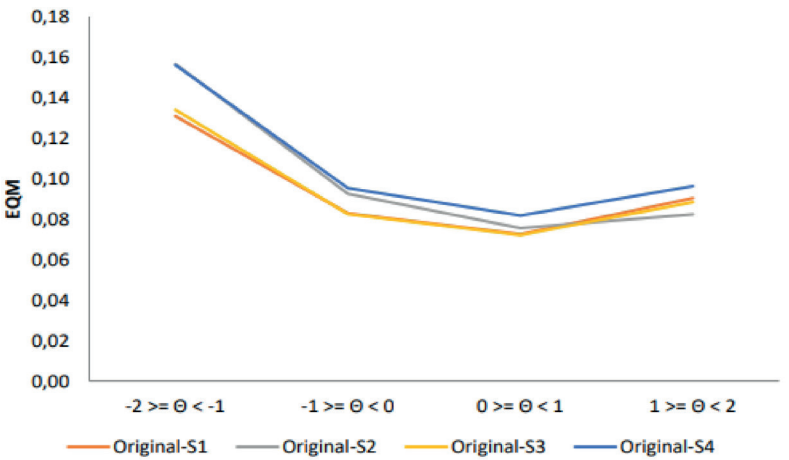
Fonte: elaborado pelo autor

Ao comparar a proficiência média do grupo focal nas diferentes situações, para a simulação de um impacto de 0,50 no parâmetro de dificuldade, o procedimento que mais se aproxima do valor real (proficiência original) é quando o item com DIF passa pelo processo de recalibração,

ou seja, a ligação do item com o teste anterior é retirada e seus parâmetros são reestimados.

Outro procedimento adotado para analisar o impacto do DIF foi calcular o EQM nos diferentes tratamentos para cada faixa de proficiência, como apresentado na Figura 4.

FIGURA 4 - TRATAMENTOS E EQM POR FAIXA DE PROFICIÊNCIA



Fonte: elaborado pelo autor

Na Figura 4, a categoria “Original-S1” representa a diferença entre as proficiências originais e as proficiências de um teste sem a presença de itens com DIF, as categorias “Original-S2”, “Original-S3” e “Original-S4” representam a diferença entre as proficiências originais e as proficiências de um teste com a presença de itens com DIF, de modo que, para cada categoria, foi aplicado um tipo de tipo de tratamento nos itens com DIF. Para a categoria “Original-S2”, não foi aplicado nenhum tipo de tratamento nos itens identificado com DIF, ou seja, os itens com DIF foram ignorados; para categoria “Original-S3”, os itens com DIF foram recalibrados; e, para categoria “Original-S4”, os itens com DIF foram excluídos. O EQM foi calculado para cada faixa de proficiência e verificou-se que o tratamento “Original-S3” causa menor impacto nas proficiências dos indivíduos quando comparado ao cenário de ideal (teste sem itens com DIF).

2. Considerações finais

Os objetivos do presente estudo foram alcançados, pois, com base na metodologia adotada pelo INEP, foi possível verificar a partir de qual magnitude do incremento adicionado no parâmetro de dificuldade do item é possível identificar o DIF e o impacto que provoca nos demais itens comuns.

No decorrer das análises, verificou-se que o número de itens influenciava no processo de identificação dos itens com DIF. Andrade, Tavares e Valle (2000) ressaltam que quanto maior o número de itens comuns, melhor será a qualidade da equalização. Para o grupo focal do cenário 1, o teste continha 66,7% de itens comuns com o teste do grupo referência, e observou-se nesse cenário que os itens comuns não sofreram impacto significativo dos itens com incrementos no parâmetro b (item imputado DIF uniforme). No entanto, para o cenário 2, foi utilizado o número de itens comuns recomendado na literatura, ou seja, 20% de itens comuns; nesse cenário, o impacto da diferença máxima foi significativo para os itens comuns sem DIF.

Quanto ao tamanho do incremento inserido no parâmetro b, a metodologia adotada foi capaz de identificar itens com DIF para os valores de incremento acima de 0,50 para o cenário 1 e valores acima de 0,75 para o cenário 2. Vale ressaltar que, no segundo cenário, foram encontrados itens com falso DIF (erro Tipo 1), ou seja, nesse cenário identificou-se itens com DIF, dado que os itens não apresentavam DIF. Após as análises dos dois cenários, percebe-se que o procedimento adotado pelo INEP consegue identificar itens com DIF até determinado ponto, já que fatores como a quantidade reduzida de itens de ligação podem elevar a taxa do erro Tipo I.

A partir da análise do impacto do uso de itens com DIF nos testes cognitivos, foi possível observar que existe diferença significativa nas proficiências médias do grupo focal que responderam a um teste com itens com DIF em relação às proficiências originais desse grupo. Para minimizar o impacto nas proficiências dos indivíduos, foram aplicados três tipos de tratamentos nos itens com DIF e calculado o Erro Quadrático Médio (EQM) para cada faixa de proficiência. Os tratamentos considerados nas análises foram: i) não excluir os itens com DIF; ii) recalibar os itens com DIF no grupo focal; e iii) excluir os itens com DIF do grupo focal. O tratamento de recalibrar

os itens com DIF no grupo focal apresentou o menor EQM, ou seja, menor impacto nas proficiências dos indivíduos.

Destaca-se, ainda, que existe uma carência na literatura em relação ao estudo do tamanho do viés na comparabilidade entre grupos considerando itens de ligação com DIF. Os autores não encontraram nenhum estudo que tivesse realizado a comparabilidade do tamanho do viés entre grupo relacionado à quantidade de itens com DIF. Dessa forma, esse trabalho vem contribuir em alguns aspectos para essa lacuna que existe na literatura.

Referências

ANDRADE, D. F.; TAVARES, H. R.; VALLE, R. C. Teoria de resposta ao item: conceitos e aplicações. Associação Brasileira de Estatística, 4º SINAPE, 2000.

ANDRIOLA, W. B. Descrição dos principais métodos para detector o funcionamento diferencial dos itens (DIF). Psicologia: Reflexão e Crítica, 2001.

ANDRIOLA, W. B. Funcionamento diferencial do item (DIF): indicador de justiça das avaliações em larga escala. 018. 192f. TESE (Para a Classe de Professor Titular) – Universidde Federal do Ceará, Faculdade de Educação, Departamento de Fundamentos da Educação, Fortaleza (CE), 2018.

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, 33, 465-484, 1996.

EELLS, K.; DAVIS, A.; HAVIGHURST, R. J.; HERRICK, V. E; TYLER, R. W. Intelligence and cultural differences. Chicago: University of Chicago Press, 1951.

EMBRETSON, S. E.; REISE, S. P. Item Response Theory for Psychologists. New Jersey, USA: Lawrence Erlbaum Associates, 2000.

EVERSON, H.; OSTERLIND, S. Differential Item Functioning. London: Sage, 2009.

MAGIS, D.; BELAND, S.; TUERLINCKX, F;DE BOECK, P. A general framework and an R package for the detection of dichotomous differential item functioning. Behavior Research Methods, 42, p. 847-862, 2010.

MARTÍNEZ ARIAS, R. Psicometría: teoría de los tests psicológicos y educativos. Madrid: Síntesis, 1997.

MUÑIZ, J. Introducción a la teoría de respuesta a los ítems. Madrid: Ediciones Psicología Pirámide, 1997.

PASQUALI, L. Teoria da resposta ao item - IRT: uma introdução. In: PASQUALI, L. (Org.). Teoria e métodos de medida em ciências do comportamento. Brasília: INEP, 1996._____. Psicometria: teoria dos testes psicológicos. Brasília: Prática, 2000.

SISTO, F. F. O funcionamento diferencial dos itens. Psico-USF, 11, p. 35-43, 2006.

VALLE, R. C. Comportamento Diferencial do Item (DIF): uma apresentação. Estudos em Avaliação em Larga Escala, 25, p. 167-184, 2002.