UNIVERSIDADE FEDERAL DE SANTA CATARINA
DEPARTAMENTO DE INFORMÁTICA E ESTATÍSTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM MÉTODOS E GESTÃO EM AVALIAÇÃO

CÁCIO FABRÍCIO GOMES DA ROCHA

O IMPACTO DO FUNCIONAMENTO DIFERENCIAL DO ITEM (DIF) EM TESTES COM ITENS DICOTÔMICOS: UM ESTUDO DE SIMULAÇÃO

Florianópolis

2019

CÁCIO FABRÍCIO GOMES DA ROCHA

O IMPACTO DO FUNCIONAMENTO DIFERENCIAL DO ITEM (DIF) EM TESTES COM ITENS DICOTÔMICOS: UM ESTUDO DE SIMULAÇÃO

Dissertação submetida ao Programa de Pós-Graduação em Métodos e Gestão em Avaliação da Universidade Federal de Santa Catarina para a obtenção do Grau de Mestre em Métodos e Gestão em Avaliação.

Orientador: Prof. Dr. Adriano Ferreti Borgatto

Florianópolis

2019

Ficha de identificação da obra elaborada pelo autor através do Programa de Geração Automática da Biblioteca Universitária da UFSC.

ROCHA, CACIO FABRICIO GOMES
O IMPACTO DO FUNCIONAMENTO DIFERENCIAL DO ITEM (DIF) EM TESTES COM ITENS DICOTÔMICOS: UM ESTUDO DE SIMULAÇÃO / CACIO FABRICIO GOMES ROCHA ; orientador, Adriano Ferreti Borgatto, 2019.
63 p.

Dissertação (mestrado profissional) - Universidade Federal de Santa Catarina, Centro Tecnológico, Programa de Pós-Graduação em Métodos e Gestão em Avaliação, Florianópolis, 2019.

Inclui referências.

1. Métodos e Gestão em Avaliação. 2. Avaliação educacional. 3. TRI. 4. DIF. I. Borgatto, Adriano Ferreti. II. Universidade Federal de Santa Catarina. Programa de Pós-Graduação em Métodos e Gestão em Avaliação. III. Titulo.

CÁCIO FABRÍCIO GOMES DA ROCHA

O IMPACTO DO FUNCIONAMENTO DIFERENCIAL DO ITEM (DIF) EM TESTES COM ITENS DICOTÔMICOS: UM ESTUDO DE SIMULAÇÃO

O presente trabalho em nível de mestrado foi avaliado e aprovado por banca examinadora composta pelos seguintes membros:

Prof. Eduardo Carvalho Sousa, Dr.
Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira

Prof. Dalton Francisco de Andrade, Dr.
Universidade Federal de Santa Catarina (UFSC)

Prof. André Wüst Zibetti, Dr.
Universidade Federal de Santa Catarina (UFSC)

Certificamos que esta é a **versão original e final** do trabalho de conclusão que foi julgado adequado para obtenção do título de mestre em Métodos e Gestão em Avaliação.

Marcelo Menezes Reis:88891755915
Assinado de forma digital por Marcelo Menezes Reis:88891755915
Dados: 2019.08.20 16:52:05 -03'00'

Prof. Dr. Marcelo Menezes Reis
Coordenador do Curso

Adriano Ferreti Borgatto:19547506846
Assinado de forma digital por Adriano Ferreti Borgatto:19547506846
Dados: 2019.08.20 16:43:05 -03'00'

Prof. Dr. Adriano Ferreti Borgatto - Orientador
Universidade Federal de Santa Catarina (UFSC)

Florianópolis, 16 de agosto de 2019.

Dedico este trabalho ao meu amado filho,

Pedro Henrique.

AGRADECIMENTOS

Ao Professor e Orientador, Adriano Borgatto, que durante essa trajetória, além das orientações e puxões de orelha, se mostrou ser um grande amigo. Obrigado pelos conselhos.

Ao Cebraspe, pelo incentivo e apoio durante essa trajetória.

À Simone Barros, nossas conversas me ajudaram a chegar aqui. Estou seguindo suas orientações, já retirei o "vou tentar" e inseri o "vou fazer!".

Aos colegas do Cebraspe, em especial a Brunno Thadeu Bittencourt, Elianice Castro e a minha chefinha querida Leticia Santos (Dra. Alsam).

À Professora Girlene Ribeiro, pelo incentivo e apoio durante essa trajetória.

À amiga Ediane Ferreira, pelo apoio nessa reta final.

Aos colegas do programa de pós-graduação em Métodos em Gestão e Avaliação, que me acolheram e compartilharam experiências, conhecimentos e as etapas de suas pesquisas.

Aos Forasteiros Wagner, Adrain, Vanessa e Silvani, pela amizade e pelos momentos na cozinha.

À minha querida amiga, Patrícia Vieira, obrigado por tudo.

Às minhas eternas vó Madalena (*in memoriam*) e dindinha Lisete (*in memoriam*); minha tias e tios: Eugênia e Sérgio, Nicy e Ceslo, Célia.

À minha mãe por ser meu porto seguro; à minha amada irmã Hully e ao cunhado, Neto, pelo carinho e pela minha braquelinda Jhuly. Obrigado pelo sustento e pelo amor.

“TUDO tem o seu tempo determinado, e há tempo para todo o propósito debaixo do céu.”

Eclesiastes 3:1

RESUMO

O Funcionamento Diferencial do Item – *Differential Item Functioning* (DIF) – ocorre quando um item, em uma mesma escala, se comporta de maneira diferente entre dois ou mais grupos de indivíduos com mesmo nível de habilidade. Ou seja, esses grupos têm desempenhos diferenciados ao responderem ao mesmo item. O objetivo do presente estudo foi verificar o impacto no processo de equalização ao inserir itens com DIF. A importância de realizar a análise de DIF nos diferentes grupos de indivíduos se justifica para que as proficiências estimadas sejam fidedignas e os resultados possam ser adequadamente equalizados. Foram simulados no *software R*, versão 3.5.2, dois grupos (referência e focal) com características de uma avaliação em larga escala. O grupo referência foi submetido a um teste com 45 itens novos e o grupo focal a um teste parcialmente diferente com 45 itens (itens comuns com o teste do grupo referência). Para equalização dos parâmetros dos itens e das proficiências dos indivíduos, utilizou-se a Teoria de Resposta ao Item (TRI) com auxílio do *software* Bilog-MG. Na literatura, existem diferentes métodos para identificar o DIF e alguns autores recomendam a utilização de mais de um método no processo de análise de DIF, visto que cada método apresenta sua limitação. Para identificar itens comuns com DIF, foi adotada a metodologia utilizada pelo Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP), que consiste em comparar as proporções esperadas de acertos ao item nos pontos de quadratura para cada grupo compreendida no intervalo P5 (percentil 5) e o P95 (percentil 95), quando a diferença máxima entre os grupos for maior ou igual a 0,15 o item é classificado com DIF. Nos cenários simulados, foram avaliados o número de itens comuns e os diferentes incrementos adicionados no parâmetro de dificuldade com o intuito de gerar DIF uniforme. Os resultados mostraram que o número de itens comuns auxilia no processo de identificação dos itens com DIF uniforme, pois quanto mais itens de ligação no processo de equalização, menor será o impacto sofrido por itens que apresentam algum tipo de viés. Os itens identificados com DIF trazem informações importantes para os sistemas educacionais, pois podem diagnosticar deficiências curriculares, índices de discriminação racial, características econômicas ou diversidades culturais. Contudo, com base nos resultados, recomenda-se a inserção de outros métodos de identificação, pois a atual metodologia mostrou-se sensível a itens que apresentam diferenças elevadas nas proporções de acerto esperadas entre o grupo focal e o grupo de referência, quando associadas a um número menor de itens comuns.

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

ABSTRACT

Differential Item Functioning (DIF) - occurs when an item on the same scale behaves differently between two or more groups of individuals with the same skill level. That is, these groups perform differently when responding to the same item. The objective of the present study was to verify the impact on the equalization process when inserting items with DIF. The importance of performing the DIF analysis in the different groups of individuals is justified so that the estimated proficiencies are reliable and the results can be adequately equalized. Two groups (reference and focal) with characteristics of a large scale evaluation were simulated in software R, version 3.5.2. The reference group was subjected to a test with 45 new items and the focus group to a partially different test with 45 items (common items with the reference group test). For the equalization of the parameters of the items and the proficiencies of the individuals, the Theory of Response to the Item (TRI) was used with the help of the software Bilog-MG. In the literature, there are different methods to identify DIF and some authors recommend the use of more than one method in the DIF analysis process, since each method presents its limitation. To identify common items with DIF, the methodology used by the National Institute of Educational Studies and Research Anísio Teixeira (INEP) was adopted, which consists of comparing the expected proportions of correct answers to the item in the quadrature points for each group included in the P5 interval (percentile 5) and P95 (95th percentile), when the maximum difference between groups is greater than or equal to 0.15 the item is classified as DIF. In the simulated scenarios, the number of common items and the different increments added in the difficulty parameter were evaluated in order to generate uniform DIF. The results showed that the number of common items assists in the process of identifying items with uniform DIF, since the more binding items in the equalization process, the smaller the impact of items with some kind of bias. Items identified with DIF bring important information to educational systems, as they can diagnose curriculum deficiencies, racial discrimination indices, economic characteristics or cultural diversities. However, based on the results, it is recommended to insert other methods of identification, since the current methodology was sensitive to items that present high differences in the expected proportion of accuracy between the focal group and the reference group, when associated with fewer common items.

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

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE SIGLAS

ANA	Avaliação Nacional da Alfabetização
CESBRASPE	Centro Brasileiro de Pesquisa em Avaliação e Seleção e de Promoção de Eventos
CCI	Curva característica do item
EAP	Expected a Posteriori
ENEM	Exame Nacional do Ensino Médio
ETS	Educational Testing Service
EUA	Estados Unidos da América
DIF	Differential Item Functioning
INEP	Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira
FUNBEC	Fundação Brasileira para o Ensino da Ciência
FGV	Fundação Getúlio Vargas
HMV	Habilidade Mental Verbal
MH	Mantel-Haenszel
MEC	Ministério da Educação
MMSE	Mini Exame do Estado Mental
ML1	Modelo Logístico de 1 parâmetro
ML2	Modelo Logístico de 2 parâmetro
ML3	Modelo Logístico de 3 parâmetro
ML4	Modelo Logístico de 3 parâmetro
MVM	Máxima Verossimilhança Marginal
NAEP	National Assessment of Educational Progress
P5	Percentil 5
P95	Percentil 95
SAEB	Sistema Nacional de Avaliação da Educação Básica
SAEP	Sistema de Avaliação do Ensino Público de 1º grau
SARESP	Sistema de Avaliação de Rendimento Escolar do Estado de São Paulo
SIB	Simultaneous Item Bias
TCT	Teoria Clássica dos Testes
TRI	Teoria de Resposta ao Item

SUMÁRIO

1 INTRODUÇÃO

1.1 CONTEXTUALIZAÇÃO

A presente dissertação de Mestrado aqui apresentada está circunscrita ao campo de estudo das avaliações educacionais externas em larga escala. Essa modalidade de avaliação educacional 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. Os Estados Unidos da América (EUA) são um dos precursores nos estudos de avaliação educacional. Dentre os importantes marcos da avaliação educacional externa em larga escala destaca-se a *National Assessment of Educational Progress*[1] (NAEP), tendo em vista sua influência para a elaboração do sistema de avaliação educacional brasileiro. O NAEP foi criado em 1969 para medir o desempenho dos estudantes norte-americanos em diferentes áreas do conhecimento, tais como: artes, ciências, história, geografia, economia, matemática, leitura e escrita. Em conjunto com as provas, são aplicados questionários contextuais aos estudantes, professores e gestores das escolas, cuja as respostas auxiliam os pesquisadores e a comunidade a entenderem melhor os resultados (OLIVEIRA, 2012).

Apesar do pioneirismo norte-americano nos estudos para melhor compreender a realidade educacional, é relevante registrar, como discorre Gontijo (2008), que o Brasil já em meados do século XIX criou a Diretoria Geral de Estatísticas do Brasil que coletava informações sobre a realidade educacional do país. Também Horta Neto (2006) e Souza (2009) destacam que o governo brasileiro a partir de 1907 iniciou a coleta de informações sobre o número de escolas, docentes, matrículas e repetência dos estudantes, para compor o chamado Anuário Estatístico do Brasil.

No que se refere aos levantamentos sobre a Educação Básica, de acordo com Mariani (1982), em 1960 o Censo Escolar iniciou o levantamento de dados nacionais. Esses dados subsidiavam estudos que passam a associar desempenho a fatores contextuais. Segundo Vianna (2005), ainda nos anos 1960, estudos de avaliação foram empreendidos tendo como foco a Educação Básica, como os desenvolvidos pela Fundação Brasileira para o Ensino da Ciência (FUNBEC) e pela Fundação Getúlio Vargas (FGV). Com relação ao estudo da FGV, Viana

[1] Avaliação Nacional do Progresso Educacional.

(2005) registra que se tratou de uma avaliação realizada no estado do Rio de Janeiro, inspirada no teste Iowa Basic Skills[2], o qual era realizado nos EUA desde 1935 com a finalidade de entregar às escolas informações para auxiliar nas ações de melhoria do ensino ofertado (IOWA TEST PROGRAMS, 2017).

A avaliação externa do sistema educacional brasileiro ocorreu, de forma sistematizada, apenas em 1988 com a proposição do Sistema Nacional de Avaliação do Ensino Público de 1º grau – SAEP (BONAMINO, 2002; FRANCO, 2004). A partir dessa experiência, em 1990, consolidou-se o Sistema de Avaliação da Educação Básica (SAEB), que previa a articulação das medidas de desempenho e resultados de estudos de contexto, contando com o apoio operacional das Secretarias de Estado de Educação (BONAMINO, 2002).

Em 1995, a metodologia de correção do 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.

A adoção da TRI no contexto do SAEB teve por princípio a compreensão de modelos para avaliar traços latentes. Esses modelos apresentam formas de representar a relação entre a probabilidade de um estudante acertar determinado item e seus traços latentes ou habilidades, na área de conhecimento a ser avaliada ou verificada, os quais não podem ser observados diretamente (ANDRADE; TAVARES; VALLE, 2000). Após a incorporação da TRI nas avaliações do SAEB, foi possível tornar comparável os resultados dos estudantes entre as edições de aplicação, entre os estados, entre as escolas e entre as séries/anos. Permitiu, ainda, estimar o nível de dificuldade de cada item independente do grupo de respondentes do teste e estabelecer uma escala de proficiência. Nesse sentido, como cada item mede uma única habilidade, em diferentes níveis de dificuldade, as avaliações passaram a fornecer informações mais precisas sobre o nível de desempenho de cada estudante.

Em 2009, outra avaliação brasileira passou a utilizar a TRI, tendo em vista a sua maior precisão em medir o conhecimento de cada estudante: o Exame Nacional do Ensino Médio (ENEM). O ENEM foi criado em 1998 com o objetivo de avaliar o desempenho do estudante

[2] Teste *Iowa* de Habilidades Básicas.

ao fim da escolaridade básica e, a partir de 2009 com a adoção da TRI, tornou-se mais uma forma de ingresso à Educação Superior, seja complementando ou substituindo o vestibular convencional.

Estudos sobre o processo de construção de testes foram ampliados com os avanços de áreas de estudo como a psicometria e a estatística. Essas áreas contribuíram para que os testes fossem construídos de forma mais precisa, de maneira a assegurar a validade da medida gerada. Em meio a tais estudos, estão aqueles que propõem o uso TRI nas avaliações em larga escala, uma vez que essa teoria ofereceu melhores soluções, diante das limitações apresentadas pela Teoria Clássica dos Testes (TCT), que se baseia no escore total obtido pelo respondente, e tem como ênfase o instrumento de medida (teste, prova, questionário) como um todo, ou seja, as análises e as interpretações estão associadas sempre ao instrumento. Muñiz (1997) ressalta que a TRI supõe uma mudança radical com relação à TCT, embora não seja uma teoria oposta, mas complementar ao modelo clássico.

Diante do potencial das informações disponibilizadas pelas avaliações educacionais externas, especialmente a partir do uso de novas metodologias que possibilitam maior precisão da medida, como é o caso da TRI, essas informações ganharam papel de destaque na agenda pública nacional e internacional. Atualmente, os dados gerados pelas avaliações, subsidiam o planejamento estratégico e o monitoramento da política educacional. Assim sendo dada a importância dos resultados gerados pelas avaliações externas para a política educacional, é fundamental a constante preocupação em se garantir a uniformização e padronização dos instrumentos utilizados para aferir o desempenho dos estudantes, tal como assevera Pasquali (2000).

1.2 PROBLEMA DE PESQUISA

Nessa perspectiva, cada vez mais busca-se assegurar a uniformidade na elaboração dos itens utilizados nos testes, padronizar comandos e enunciados, planejar a estrutura do teste, além de controlar as situações relacionadas à aplicação do teste (MARTÍNEZ ARIAS, 1997). Dentre os estudos que avançaram no intuito de assegurar a validade da medida gerada, destacam-se os que abordam o Funcionamento Diferencial do Item (DIF), que está intimamente ligado à padronização das condições de aplicação e à verificação da equidade dos instrumentos (HAMBLETON, 1997). A presença de DIF em um processo avaliativo é um fator que afeta a qualidade do teste, uma vez que o torna injusto, pois, ao não gerar um resultado válido, há o

risco de orientar erroneamente o planejamento e o monitoramento da política educacional. Para compreender melhor o impacto da presença de itens com funcionamento diferencial deveremos conhecer o conceito de DIF, que no âmbito da TRI, o item não apresenta DIF se as curvas características (CCI) é idêntica para os grupos avaliados quando comparados num mesmo nível de habilidade ou magnitude da variável latente medida, como afirma Lord (1980).

Nessa perspectiva, surge a necessidade de estudar formas de detectar os itens que apresentam DIF, portanto, a identificação de itens com DIF é de grande importância no ajuste de modelos da TRI, pois essa diferença sistemática pode comprometer toda a inferência realizada, como o estabelecimento dos parâmetros dos itens e a proficiências dos estudantes, além de violar os processos avaliativos que dependam dos resultados dessas avaliações.

1.3 OBJETIVOS

1.3.1 Objetivo Geral

O objetivo deste trabalho é o de analisar o impacto da utilização de itens, em avaliações educacionais em larga escala, que contenham algum tipo de viés e que possam apresentar um funcionamento diferencial em grupos distintos.

1.3.2 Objetivos Específicos

De forma mais específica, pretende-se:

a) mensurar os graus de impacto de itens com comportamento diferencial, em um cenário hipotético de avaliação educacional em larga escala, levando em consideração o número de itens comuns com DIF, suas posições na composição do teste e seus parâmetros de dificuldade, utilizando a metodologia implementada pelo INEP nas avaliações em larga escala;
b) averiguar a viabilidade de proposição de uma escala de uso de itens com DIF, em cenários hipotéticos de uma avaliação em larga escala, desde que observados os critérios estatísticos pertinentes;
c) verificar o ponto de corte adequado para determinar se o item apresenta DIF.

1.4 JUSTIFICATIVA

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.

Considerando a centralidade dos resultados gerados pelas avaliações tanto para o diagnóstico da qualidade da educação brasileira, como para a proposição, monitoramento e reformulação de políticas públicas educacionais, acredita-se que este estudo pode contribuir para o aprimoramento da seleção dos itens para composição dos testes das avaliações em larga escala.

Este trabalho está estruturado em quatro capítulos, a saber:

O primeiro capítulo aborda a introdução do trabalho, inserindo a contextualização, justificativas e objetivo.

O segundo capítulo aborda o contexto histórico, os pressupostos e os modelos usuais da Teoria de Resposta ao Item e um breve histórico sobre o Funcionamento Diferencial do Item (DIF) e a revisão de literatura dos principais métodos de identificação do DIF.

O terceiro capítulo apresenta a metodologia utilizada na geração da população de estudo, bem como os procedimentos para calibração e equalização dos itens e para identificação do DIF.

O quarto capítulo apresenta os resultados obtidos com a implementação em dados simulados.

Por fim, são apresentadas as considerações sobre o estudo e as recomendações para trabalhos futuros.

2 TEORIA DA RESPOSTA AO ITEM

Neste capítulo apresenta-se o contexto histórico, os pressupostos e os principais modelos matemáticos da Teoria da Resposta ao Item.

2.1 CONTEXTO HISTÓRICO DA TEORIA DA RESPOSTA AO ITEM

A TRI surgiu na década 50 com os trabalhos de Lord (1952), que desenvolveu o primeiro modelo unidimensional de dois parâmetros (dificuldade e discriminação), baseado na distribuição normal acumulada (ogiva normal), incorporando mais tarde um parâmetro que tratava do problema do acerto casual, surgindo o modelo de três parâmetros. Anos mais tarde, Birnbaum (1968) substituiu em ambos os modelos propostos por Lord (1952), a função da ogiva normal pela função logística, matematicamente mais simples, pois é uma função explícita dos parâmetros do item e da habilidade e não envolve integração. Rasch (1960), independentemente do trabalho de Lord (1952), propôs o modelo unidimensional de um parâmetro (dificuldade), usando a função ogiva normal e esse modelo mais tarde foi descrito por um modelo logístico por Wright (1968).

Com a finalidade de obter mais informações das respostas dos indivíduos, Samejima (1969) propôs o modelo de resposta gradual. Bock (1972), Andrich (1978), Masters (1982) e Muraki (1992) também propuseram modelos para mais de duas categorias de resposta, assumindo diferentes estruturas entre essas categorias. Bock e Zimowski (1997) introduziram os modelos logísticos de um, dois e três parâmetros para duas ou mais populações de respondentes.

Embora a TRI já tenha uma longa história (PASQUALI, 1996) no Brasil, o uso da TRI nas avaliações em larga escala dá-se a partir do aperfeiçoamento metodológico no âmbito do SAEB, desde o ano de 1995. A partir de então, a utilização dessa teoria nas avaliações educacionais em larga escala tem se expandido, dada a frequente preocupação em garantir a fidedignidade e validade da medida cognitiva gerada. Esse é o caso do ENEM, da Avaliação Nacional da Alfabetização (ANA), do Sistema de Avaliação de Rendimento Escolar do Estado de São Paulo (SARESP), dentre outros, que são avaliações planejadas e implementadas de modo a serem analisadas com a TRI.

2.2 PRESSUPOSTOS DA TEORIA DA RESPOSTA AO ITEM

Há duas pressuposições básicas para aplicação dos principais modelos matemáticos da TRI: a unidimensionalidade e a independência local.

O pressuposto da unidimensionalidade assume que somente um único traço latente tenha influência na probabilidade de resposta (certo ou errado) do respondente j a um determinado item i. No entanto, a Psicologia atesta que o desempenho humano é motivado por mais de um traço latente; embora, um respondente j possua um conjunto de habilidades $(\theta_1, \theta_2, \dots, n)$ necessárias para responder corretamente ao item i, tal pressuposto assume que existe um $\theta = \theta_j$, para algum j, que será dominante. Ou seja, a habilidade dominante é o que se supõe estar sendo medido pelo teste cognitivo ou questionário (PASQUALI, 1996).

A unidimensionalidade dos dados coletados é verificada por meio de análise fatorial exploratória, como sugerido por Embretsone e Reise (2013). Segundo Brown (2006), a análise fatorial exploratória compreende uma série de técnicas multivariadas para identificar a dimensionalidade do instrumento.

Por sua vez, a independência local pressupõe que a probabilidade de um respondente j responder (corretamente ou incorretamente) ao item i não depende das demais respostas dadas aos outros itens. Em outras palavras, o desempenho do respondente em determinado item não é afetado pelo desempenho nas demais questões do instrumento de análise (HAMBLETON; SWAMINATHAN; ROGERS, 1991). Assim, a probabilidade de resposta a um conjunto de itens, dada a habilidade de um respondente, é igual ao produto das probabilidades das respostas do respondente a cada item separadamente,

$$P(x_{j1}, x_{j2}, \dots x_{jI}|\theta_j) = P_1(x_{j1}|\theta_j) \times P_2(x_{j2}|\theta_j) \times \dots \times P_I(x_{jI}|\theta_j)$$

onde o número total de itens é $I, i \in \{1, 2, \dots, I\}$. A habilidade do respondente j, com e $j \in \{1, 2, \dots, n\}$ é considerada aptidão dominante para a resolução do item i e é denotado por θ_j. A resposta do respondente j dada ao item i é dicotômica, ou seja, $x_{ji} = 1$, resposta correta e, caso contrário, $x_{ji} = 0$. Por fim, a probabilidade de sucesso é $P_i(x_i = 1|\theta_j)$ e a probabilidade de fracasso é $P_i(x_i = 0|\theta_j)$.

Para Hambleton, Swaminathan e Rogers (1991), a unidimensionalidade implica independência local. Além disso, segundo Lord e Novick (1968), a independência local implica também a unidimensionalidade. Assim, chega-se ao ponto em que há somente uma, e não duas, suposição a ser verificada.

2.3 MODELOS MATEMÁTICOS DA TEORIA DA RESPOSA AO ITEM

Os modelos matemáticos da TRI apresentam a relação entre a probabilidade de um indivíduo dar uma certa resposta a um item qualquer e os parâmetros do item e da habilidade do respondente, de forma que, quanto maior a habilidade do respondente, maior a probabilidade de acerto ao item (ANDRADE; TAVARES; VALLE, 2000). Os modelos propostos dependem fundamentalmente de três fatores:

i) da natureza do item – dicotômicos ou não dicotômicos;
ii) do número de populações envolvidas – apenas uma ou mais de uma;
iii) do número de traços latentes que estão sendo medidos – apenas um ou mais de um.

Nas seções seguintes serão apresentados os principais modelos unidimensionais da TRI, considerando multigrupos de respondentes e a natureza do item.

2.3.1 Modelos para Itens Dicotômicos

Itens dicotômicos são itens de múltipla escolha com dois possíveis resultados (certo ou errado), envolvendo geralmente 4 ou 5 opções de respostas. Os modelos para itens dicotômicos são os mais utilizados, tendo basicamente três tipos, que se diferem pelo número de parâmetros utilizados para descrever o item. Estes parâmetros são: discriminação (a_i), dificuldade do item (b_i) e acerto casual (c_i).

2.3.1.1 Modelo Logístico de 1 Parâmetro

O Modelo Logístico de 1 parâmetro (ML1) é considerado o modelo mais simples, pois considera que a probabilidade de resposta correta ao item i de um respondente j, depende somente da diferença entre o nível de proficiência do respondente (θ_j) e a dificuldade do item (b_i). Sua expressão dada por

$$P(X_{ijk} = 1|\theta_{jk}, b_i) = \frac{1}{1 + \exp^{-D(\theta_{jk} - b_i)}}$$

com $i = 1, 2, \ldots, I, j = 1, 2, \ldots, n$ e $k = 1, 2, \ldots, K$, em que:

X_{ijk} é a resposta do estudante j ao item i, que é igual a 1, se o estudante responde corretamente o item i, e igual a zero, caso contrário;

θ_{jk} representa a habilidade (ou proficiência/traço latente) do $j-$ésimo respondente da população k;

$P(X_{ijk} = 1|\theta_{jk}, b_i)$ é denominada de Função de Resposta do Item e mede a probabilidade de um respondente j com habilidade θ_j responder corretamente ao item i.

b_i o parâmetro de dificuldade (ou de locação) do item i, e quanto maior o valor de b, maior tem que ser a habilidade do respondente j para responder corretamente ao item i;

exp representa a função exponencial; e

D é um fator de escala constante e igual a 1. Utiliza-se o valor 1,7 quando se deseja que a função logística forneça resultado semelhante ao da função ogiva normal acumulada.

O modelo de Rasch (1960) supõe que o índice de discriminação seja o mesmo para todos.

2.3.1.2 Modelo Logístico de 2 Parâmetros

O primeiro modelo da TRI de 2 parâmetros (ML2) foi criado por Lord em 1952, baseado na função da distribuição normal. A expressão do ML2 é

$$P(X_{ijk} = 1|\theta_{jk}, b_i, a_i) = \int_{-\infty}^{a_i(\theta_{jk} - b_i)} \frac{1}{\sqrt{2\pi}} \exp^{-\frac{x^2}{2}} dx$$

Birnbaum (1968) substituiu a distribuição normal pela função logística. E, assim, a equação utilizada para avaliar a probabilidade de um respondente j com habilidade θ_j responder corretamente ao item i de um teste é dada por

$$P(X_{ijk} = 1|\theta_{jk}, b_i, a_i) = \frac{1}{1 + \exp^{-Da_i(\theta_{jk} - b_i)}}$$

onde $i = 1, 2, \ldots, I$, $j = 1, 2, \ldots, n, k = 1, 2, \ldots, K$ e a_i é o parâmetro de inclinação ou de discriminação do item i e representa o quanto um item discrimina entre os respondentes de diferentes níveis de proficiência. De acordo com Baker (2004), considerando D=1, itens com $a_i = 0$ não apresentam discriminação; com a_i entre 0,01 e 0,34 apresentam discriminação muito baixa; com a_i entre 0,35 e 0,64 apresentam discriminação baixa; com a_i entre 0,65 e 1,34 apresentam discriminação moderada; com a_i entre 1,35 e 1,69 apresentam discriminação alta; e com $a_i > 1{,}70$ apresentam discriminação muito alta. Os demais parâmetros são os mesmos já apresentados no ML1.

2.3.1.3 Modelo Logístico de 3 Parâmetros

O ML1 e o ML2 não consideram o acerto ao acaso de um respondente j ("chute"). Essa possibilidade implica um novo parâmetro a ser incorporado no modelo. Sabe-se que nos testes com itens de múltipla escolha é possível que um respondente j responda corretamente a um item sem ter conhecimento do assunto. Assim, pensou-se no modelo com mais um parâmetro, e surgiu o Modelo Logístico de 3 parâmetros (ML3) (HAMBLETON; SWAMINATHAN; ROGERS, 1991).

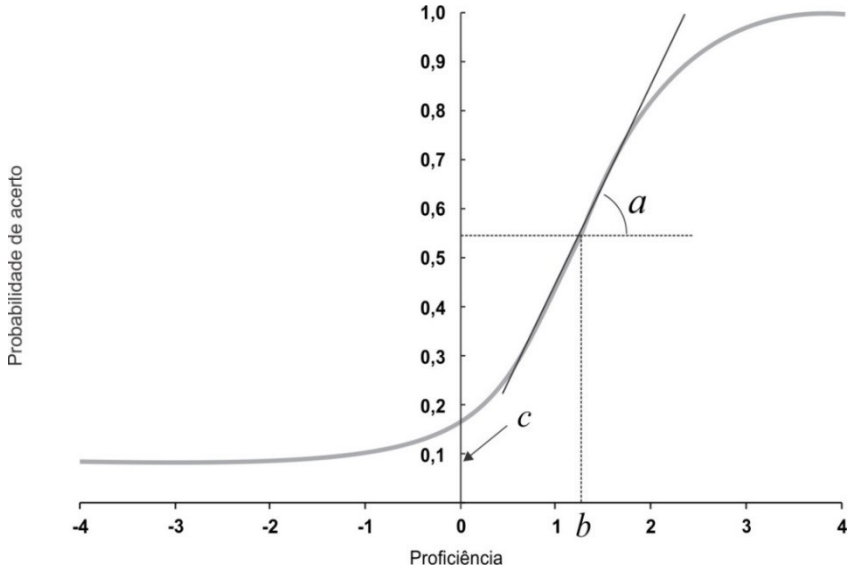
O ML3 foi proposto por Birnbaum em 1968, esse modelo da TRI é atualmente o mais utilizado e sua expressão matemática é dada por:

$$P(X_{ijk} = 1|\theta_{jk}, a_i, b_i, c_i) = c_i + (1 - c_i)\frac{1}{1 + \exp^{-Da_i(\theta_{jk} - b_i)}}$$

em que $i = 1, 2, \ldots, I, j = 1, 2, \ldots, n$, $k = 1, 2, \ldots, K$ e c_ié o parâmetro da assíntota inferior da curva do item i, representando a probabilidade de um respondente j de proficiência extremamente baixa selecionar a opção de resposta correta (probabilidade de acerto casual).

O modelo de Birnbaum (1968) pode ser expresso graficamente pela Curva Característica do Item (CCI), tal como apresentado na Figura 1.

Figura 1 – Exemplo de uma Curva Característica do Item (CCI)



Fonte: Elaborado pelo autor

Na Figura 1, o eixo vertical representa a probabilidade de acertar o item, enquanto o eixo horizontal indica o traço latente dos respondentes. Dessa forma, um respondente com traço latente igual a 1,3 tem probabilidade de 0,55 de acertar o item. O inverso também pode ser observado, isto é, um respondente com traço latente igual a -1,3 tem probabilidade próxima a 0,10 em acertar o item. As propriedades de discriminação, dificuldade e acerto ao acaso são identificáveis a partir da CCI.

A discriminação do item é dada pelo coeficiente angular da reta tangente ao ponto de inflexão (ponto onde a curva muda a concavidade). Quanto maior o valor do parâmetro de discriminação (a), mais íngreme será a curva e mais discriminativo será o item. Valores próximos de zero indicam que o item tem baixo poder de diferenciar os respondentes de alta proficiência dos respondentes de baixa proficiência. Valores negativos para o parâmetro a_i indicam que o item se comporta de maneira inadequada, pois um aumento na proficiência implicaria a diminuição da probabilidade de acerto, assim são esperados apenas valores positivos para o parâmetro de discriminação ($a_i > 0$).

O parâmetro de dificuldade (b) representa a habilidade mínima necessária para um respondente j ter probabilidade alta em acertar o item i (TAVARES, 2014). Quando a habilidade é igual à dificuldade ($\theta - b = 0$) a probabilidade de acertar o item é igual a $\frac{1+c_i}{2}$. E se o valor

do parâmetro c_i for zero, a dificuldade é o valor da habilidade que produz a probabilidade de 0,50 de acertar o item. Outra relação é que quanto maior for o nível de habilidade necessária para o respondente acertar o item, maior será a dificuldade do item. A dificuldade do item está em uma escala contínua que varia no intervalo $(-\infty, \infty)$ que é a mesma escala da proficiência.

O parâmetro de acerto ao acaso (c) mostra a probabilidade de um respondente j com baixa habilidade responder corretamente ao item i. Esse parâmetro é observável a partir do ponto da curva mais próximo ao eixo horizontal, ou seja, o local onde a assíntota inferior corta o eixo das ordenadas (SARTES; SOUZA-FORMIGONI, 2013). Para testes cognitivos com cinco opções de marcação, a probabilidade de marcar cada uma das alternativas é $1/5$, logo o valor ideal para o parâmetro c seria igual a 0,20. Quando o parâmetro de acerto ao acaso é muito superior a esse valor, significa que a alternativa correta atrai os respondentes com pouca habilidade e consequentemente evidência problemas com os distratores na elaboração/construção do item.

Os modelos desenvolvidos por Rasch (1960) e Birnbaum (1968) são facilmente obtidos a partir do ML3. Ou seja, ao substituir o $c_i = 0$ no ML3, temos a fórmula do ML2. Analogamente, para encontramos a fórmula do ML1, basta substituir $c_i = 0$ e $a_i = 1$ no ML3.

Alguns autores propõem o Modelo Logístico de 4 parâmetros (ML4), que visa controlar circunstâncias aleatórias relacionadas com as falhas dos elaboradores no momento da construção dos itens. Entretanto, existem poucas pesquisas sobre esse modelo e aparentemente não existe nenhuma vantagem do ML4 em relação ao ML3 (COUTO; PRIMI, 2011).

2.4 FUNCIONAMENTO DIFERENCIAL DO ITEM

O Funcionamento Diferencial do Item (DIF) emerge no campo de estudo e a importância nas avaliações educacionais em larga escala. O intuito é apresentar o cenário no qual os estudos de DIF ganham proeminência, tendo em vista que objeto de estudo desta dissertação está circunscrito a este campo. Posteriormente realiza-se uma revisão de literatura dos principais métodos de identificação do DIF.

2.4.1 Breve histórico sobre o DIF

A avaliação educacional externa 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 externas, 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 externas, 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 garanta a uniformização e padronização dos instrumentos utilizados para aferir o desempenho dos estudantes, tal como assevera Pasquali (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 como o teste 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 habilidade 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 habilidade 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 habilidade dos indivíduos são vastos e 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.

Essa preocupação com a precisão da medida da habilidade 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.

Pode-se dizer que este campo de estudo sobre o viés é o precursor dos estudos sobre DIF. Na psicometria, ele ganhou maior ênfase nos anos 1960, diante dos movimentos pelos direitos civis que eclodiam nos Estados Unidos da América (EUA). Esse movimento que buscava a igualdade de direitos tem como símbolo Martin Luther King Jr, diante da sua luta pelo fim da discriminação racial. Naquele período, vários grupos sentiam-se tratados de forma discriminatória e, diante disso, eram injustiçados em seleções, sobretudo para vagas de trabalho e de estudo nas universidades, que utilizavam testes psicométricos. As reivindicações eram fundamentadas na diferença dos resultados dos testes usados nessas seleções se comparado os grupos por etnia, sexo, *status* socioeconômico, dentre outros fatores, o que indicava a possível presença de viés nos testes (SISTO, 2006).

2.4.2 Procedimentos estatísticos para detectar o DIF

Whitmore e Shumacker (1999) propuseram dois procedimentos para identificar o viés nos testes (viés de seleção e viés do item). O viés de seleção é verificado pela comparação dos

resultados dos testes com os critérios externos, por exemplo, os escores totais. Quanto ao viés do item, o olhar é voltado para própria estrutura do teste por meio de métodos estatísticos, sendo possível verificar por meio das análises a diferença entre os grupos e, assim, detectar vieses ou o funcionamento diferencial do item.

Outra proposta é apresentada por Mellenbergh (1989), a qual sugere que os métodos de detecção de DIF podem ser divididos em incondicionais e condicionais. A definição implícita dos métodos incondicionais é baseada na interação grupo versus item, em que o item pode diferir em dificuldade e os grupos se distinguem na capacidade em resolver o item, porém isso não aponta o viés no item. O viés do item é contemplado pela relação: a diferença entre os grupos não é constante para todos os itens e os itens que se afastam da tendência geral são considerados tendenciosos.

No método condicional, o conceito de viés do item é pautado nas diferenças de dificuldade do item apresentado entre os grupos, dado o mesmo nível de habilidade. Assim, um item será considerado tendencioso caso essa diferença de dificuldade ocorra. Para Lord (1980) e Mellenbergh (1982), os métodos condicionais devem ser preferidos ao invés dos incondicionais, pois como o próprio método sugere, a probabilidade de acertar um item é condicionada a um certo nível de habilidade. Uma subdivisão na classificação dos métodos condicionais é proposta por Millsap e Everson (1993), que definem os métodos de invariância condicional observada, que consiste em utilizar os escores obtidos no teste, só que na perspectiva da TCT, ou seja, a soma das pontuações dos itens. Os métodos de Mantel-Haenszel, Regressão Logística e Delta Gráfico são exemplos de implementação de invariância condicional observada.

Outra definição proposta pelos autores são os métodos de invariância condicional não observada, que utilizam as habilidades estimadas pelos modelos da TRI, exemplos, a medida da área entre as CCIs, a comparação das probabilidades e o Qui-quadrado de Lord.

Nesse contexto, Andriola (2001) destaca que os métodos condicionais estão fundamentados em um conhecido paradoxo, denominado Paradoxo de Simposon, que em resumo sugere comparar o comparável. No âmbito da avaliação educativa ou psicológica, o autor enfatiza que o paradoxo é importante quando se compara a probabilidade de acerto em um item, considerando os indivíduos com o mesmo nível de habilidade na variável latente medida pelo item.

2.4.2.1 Método Mantel-Hanszel

O método clássico Mantel-Haenszel (MH), proposto por de Mantel e Haenszel (1959), foi introduzido como novo procedimento para estudo de grupos relacionados. Posteriormente, Hollande e Thayer (1988) adaptaram esse método para dados dicotômicos, que se destina a testar a existência de associação entre os grupos e a resposta dada ao item (certo ou errado), condicionada a pontuação total no teste. Assim, o escore total será a variável de estratificação que estabelecerá as comparações da proporção de acerto e erro ao item entre o grupo focal e o grupo de referência. A Tabela 2 apresenta um exemplo dos procedimentos matemáticos do método clássico MH, que consiste em dispor as respostas dos indivíduos dadas aos itens em um determinado teste em *Q* tabelas de contingência 2 x 2, para cada nível de pontuação *h*, contendo as proporções de acerto e erro ao item para o grupo de referência e o grupo focal.

Tabela 1 – Tabela de contingência 2 x 2 para o nível de pontuação h.

Grupo	Acerto	Erro	Total
Referência	A_h	B_h	N_{Rh}
Focal	C_h	D_h	N_{Fh}
Total	$N_{1h} = A_h + C_h$	$N_{0h} = B_h + D_h$	N_h

Fonte: Mantel e Haenszel (1959)

Pelo método de MH, quando a razão do número de indivíduos que acertaram o item e o número de indivíduos que erraram é igual nos dois grupos para todo nível h de pontuação, o item não apresenta DIF (hipótese nula). Em termos matemáticos,

$H_0: (A_h/B_h) = \alpha(C_h/D_h)$, α = 1 para todos os *h* (não apresenta DIF).

$H_1: (A_h/B_h) = \alpha(C_h/D_h)$, α ≠ 1 em algum *h* (apresenta DIF).

Para testar a hipótese nula (H_0), ou seja, a ausência de DIF, Holland e Thayer (1988) propuseram utilizar a estatística X^2_{MH}, que é dada por:

$$\chi^2_{MH} = \frac{\left(\left|\sum_{h=1}^{Q} A_h - \sum_{h=1}^{Q} E(A_h)\right| - 0{,}5\right)^2}{\sum_{h=1}^{Q} Var(A_h)}$$

em que $E(A_h) = \frac{N_{Rh}N_{1h}}{N_h}; Var(A_h) = \frac{N_{Rh}N_{Fh}N_{1h}N_{oh}}{N_h^2(N_h-1)}$ e $0{,}5$ é o valor de correção de continuidade.

Sob H_0, o procedimento de MH segue, assintoticamente, uma distribuição Qui-quadrado com 1 grau de liberdade. Logo, a região crítica (ou de rejeição) do teste MH é dada por:

$$\{MH: MH > \chi^2_{1,\alpha}\}$$

e os valores de referência são os quantis da distribuição Qui-quadrado com 1 grau de liberdade que deixa na cauda superior 100α% da distribuição.

O método MH é um dos mais utilizados para detecção do DIF, no entanto apresenta algumas limitações, como apontado no estudo realizado por Traver, Roma e Benito (2000), que averiguaram a eficácia do método MH para itens com DIF não uniforme. O estudo utilizou dados de aplicação de uma escala de Habilidade Mental Verbal (HMV) e para os parâmetros dos itens comuns foram obtidos após o cruzamento de dois níveis de a (0,50 e 0,75) com três níveis de b (-1,5, 0 e 1,5), em uma escala com média zero e desvio padrão um. A conclusão da pesquisa apontou que o método MH possui baixa eficácia na detecção do DIF não uniforme simétrico. No entanto, quando é feita a variação no DIF não uniforme, o método consegue identificar DIF para itens com baixa dificuldade e alta discriminação, mas para itens com alta dificuldade elevou a taxa de erro tipo I.

Já o estudo proposto por Narayanan e Swaminathan (1994) avaliou o procedimento de MH e o procedimento *Simultaneous Item Bias* (SIB) para dados simulados que refletiam diferentes condições, tais como tamanho de amostra, distribuição de habilidades distintas entre o grupo focal e o grupo referência, proporções distintas de DIF e a magnitude do efeito do DIF, que resultaram em 1296 cenários. Os dois métodos mostraram-se eficazes na detecção de itens com DIF uniforme, o que corrobora o encontrado na literatura. Porém, quando há distribuições de habilidades desiguais entre grupo focal e o grupo referência, o SIB se torna mais eficiente em relação ao MH, dada a alta taxa de falso positivo. Embora os dados encontrados sejam consistentes com outras pesquisas, os pesquisadores apontam a necessidade de aprofundamento, pois ambos os métodos mostraram ter uma certa dependência quando ao tamanho da amostra.

2.4.2.2 Método de Padronização

O método de Padronização, proposto por Dorans e Kulick (1986) é uma abordagem semelhante ao método MH, em que as proporções de uma resposta correta ao item em cada grupo e para cada valor da pontuação total no teste são comparadas. Assim, a diferença p padronizada $(ST - p - DIF)$ é a estatística do teste que representa a média ponderada das diferenças entre as proporções de acerto em cada nível de pontuação no teste do grupo focal e do grupo referência. Sendo representada pela seguinte fórmula:

$$ST - p - DIF = \frac{\sum_j \omega_j \left(P_{Fj} - P_{Rj}\right)}{\sum_j \omega_j}$$

onde, $P_{Fj} = C_j/n_{Fj}$ e $P_{Rj} = C_j/n_{Rj}$ são as proporções de acerto do grupo focal e do grupo referência respectivamente, e ω_j é um sistema de ponderação. Apesar de existirem várias alternativas, usualmente a ponderação de ω_j é feita com os indivíduos do grupo focal (DORANS; KULICK, 1986). A variação dos valores da estatística $ST - p - DIF$ encontra-se no intervalo de -1 a 1, o item não apresenta DIF caso o valor dessa estatística seja zero, como apresentado por Magis (2010). No entanto, com a finalidade de interpretar o tamanho do DIF, Dorans, Schmitt e Bleistein (1992) propuseram os valores absolutos para estatística $ST - p - DIF$, ou seja, valores entre o intervalo 0 e 1, para os itens que apresentam valores abaixo de 0,05 são classificados com DIF insignificante, itens com valores entre 0,05 e 0,10 são classificados com DIF moderado e itens em que a estatística $ST - p - DIF$ seja superior a 0,10 são identificados como DIF grande.

2.4.2.3 Método SIBTEST ou SIB

O método SIBTEST ou SIB, pode ser visto como uma generalização do método de Padronização (SHEALY; STOUT, 1993). O procedimento estima a quantidade de DIF em um item e testa estatisticamente se essa quantidade é diferente de zero. Bolt (2000) relata que o método SIB é semelhante a outros métodos de detecção do DIF, pois avalia as diferenças no desempenho dos itens para os grupos de referência e focal, sujeitos ao nível de habilidade. Esse método apresenta várias vantagens em relação a estatística $ST - p - DIF$, dentre elas o fato de testar o DIF em um conjunto de itens, em vez de realizar a análise em cada item separadamente. Magis (2010) ressalta que ao fazer uso da estatística SIBTEST, um dos pressupostos que deve ser assumido é o de que o grupo de referência e o grupo focal têm o mesmo nível de capacidade média. A estatística SIBTEST apresenta a seguinte fórmula:

$$B = \frac{\hat{\beta}_U}{\hat{\sigma}(\hat{\beta}_U)}$$

em que $\hat{\beta}_U$ é dado por

$$\hat{\beta}_U = \sum_j F_j \left(\bar{Y}_{Rj} - \bar{Y}_{Fj}\right)$$

onde F_j é a proporção de indivíduos do grupo focal com escore total no teste j, $\bar{Y}_{Rj}$ e $\bar{Y}_{Fj}$ são as médias dos indivíduos do grupo de referência e do grupo focal, respectivamente. A estatística $\hat{\beta}_U$ fornece a indicação do tamanho do DIF para efeitos de interpretação do valor da estatística $\hat{\beta}_U$. Roussos e Sout (1996) propuseram uma classificação derivada do escala utilizada pelo Educational Testing Service (ETS), que faz uso do método de MH para qualificar o DIF. Na escala proposta, os itens que apresentam valores de $\left|\hat{\beta}_U\right| \leq 0{,}59$ são classificados com DIF insignificante, $0{,}59 < \left|\hat{\beta}_U\right| \leq 0{,}88$ com DIF moderado e $\left|\hat{\beta}_U\right| > 0{,}88$ com DIF elevado.

2.4.2.4 Método da Regressão Logística

O método de Regressão Logística foi proposto por Swaminathan e Rogers (1990) e supõe que as proficiências dos indivíduos são conhecidas. O modelo para predizer a probabilidade de ocorrência de uma resposta correta a um item tem a seguinte formulação matemática:

$$\mathrm{P(u = 1)} = \frac{\exp(\mathrm{z})}{1 + \exp(\mathrm{z})}$$

em que u é dado por

$$\mathrm{z} = \tau_0 + \tau_1\theta + \tau_2\mathrm{g} + \tau_3(\theta\mathrm{g})$$

onde, τ_0 é o ponto de interseção da reta de regressão com o eixo das abscissas, τ_1 é a inclinação da reta de regressão, θ é a habilidade, τ_2 é a diferença entre o desempenho dos grupos no item, g é o grupo (de referência ou focal) ao qual pertencem os indivíduos e τ_3é o parâmetro indicador da possível interação entre θ e g. Nesse método, um item terá DIF uniforme se $\tau_2 \neq 0$ e $\tau_3 = 0$, e terá DIF não uniforme se $\tau_3 \neq 0$ (seja ou não $\tau_2 = 0$).

Swanson et al. (2002) apontam que, nos últimos 25 anos, uma gama de métodos paramétricos e não paramétricos foram desenvolvidos para analisar o DIF. Contudo, para os autores, esses procedimentos normalmente consideram os itens individualmente ou em um pequeno número, tornando desfavoráveis as análsies de identificação de possíveis fontes de DIF. Para todos esses métodos, as respostas dadas ao item investigado são realizadas comparando dois grupos, um de controle "focal" e outro de "referência", usando uma ou mais medidas adicionais como covariáveis para controle de diferenças individuais, em que o objetivo é identificar se as respostas dadas ao item funcionam de maneira diferente para ambos os grupos.

Nesse sentido, como uma alternativa mais robutsa de análise de DIF, o estudo de Swanson et al. (2002) utiliza modelos de regressão logística hierárquica, combinando os resultados logísticos das análises de regressão entre itens para identificar fontes consistentes de DIF, bem como para calcular a proporção da variação explicada nos coeficientes do DIF, e para comparar a precisão preditiva de explicações alternativas para o DIF. Incorporam-se, assim, características dos itens, fornecendo um nível de análise dentro e entre os itens. Como um dos principais resultados, os autores identificaram que a análise de regressão logística hierarquizada, demonstrou-se uma estrutura simples e fácil para combinar resultados de análises de DIF para itens individuais e para avaliar explicações alternativas para a presença de DIF. No conjunto de dados simulados, houve boa recuperação dos parâmetros utilizados na geração de dados, e as estimativas condicionais de proficiência dos indivíduos e dos índices do DIF mostraram-se mais precisos do que aqueles determinados pelo uso de técnicas de regressão logística padrão.

Crane et al (2006), no artigo "Análise do Funcionamento Diferencial do Item com Técnicas de Regressão Logística Ordinal", apresentam o modelo de regressão logística ordinal para identificação de itens com DIF. Nesse estudo são analisados três modelos e as estimativas com base nas respostas dos itens advindos de um conjunto de dados do Mini Exame do Estado Mental (MMSE). No ajuste do modelo, foram utilizadas variáveis de controle como a linguagem, a autodeclaração de raça, etnia, idade, anos de educação e sexo. Como principais resultados, os autores identificaram que alguns itens apresentaram efeitos de DIF relacionado à linguagem e a outras variáveis de controle. Assim sendo, impulsionou-se a conclusão de que ao adotar a regressão logística ordinal combinada com estimativas de capacidade, é possível fornecer uma análise alternativa capaz de detectar o DIF. Crane et al (2006) ainda alertam sobre

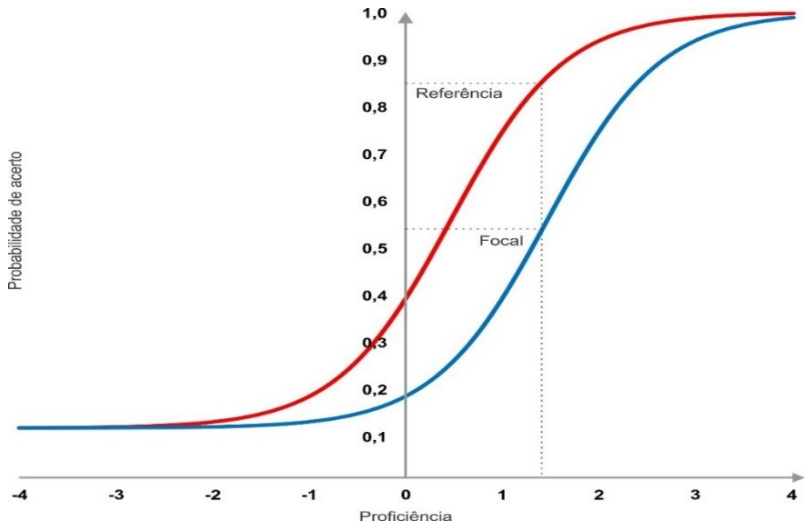
a necessidade de utilizar critérios mais específicos para determinar se um item tem DIF, de modo que os estudos ampliem suas perspectivas de análise para além apenas das técnicas utilizadas.

2.4.2.5 Métodos Baseados na TRI

Os métodos de detecção de DIF, 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 CCI é a mesma para os grupos comparados em um mesmo nível de habilidade medida pelo item. Segundo Magis (2010), existem dois tipos de DIF, o uniforme e o não uniforme.

O DIF uniforme ocorre quando as CCIs do item analisado para o grupo de referência e para o grupo focal são diferentes e não se cruzam em nenhum ponto da escala de habilidade. 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 2 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 2, os indivíduos com proficiências iguais a 1,5, nos dois grupos, têm probabilidades diferentes de acertar o item, o grupo focal tem 55,0% e o grupo de referência, 85,0%, 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.

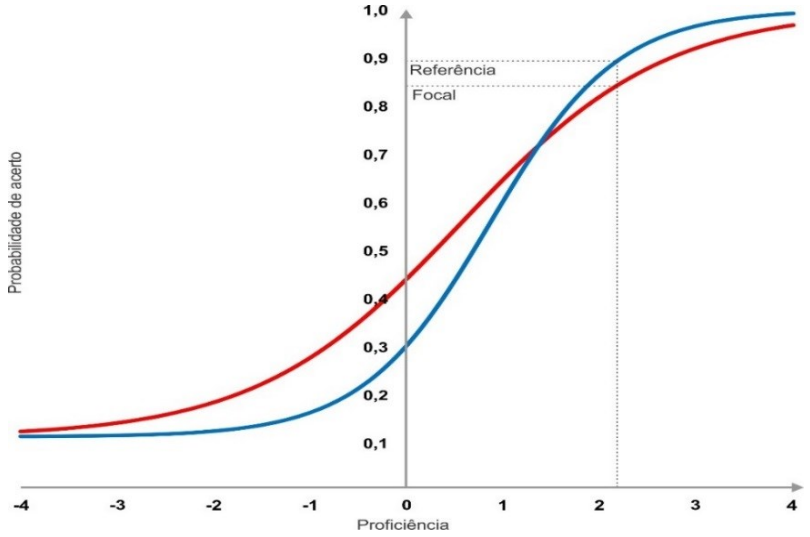
Figura 2 – Representação gráfica de um item com DIF uniforme



Fonte: Elaborado pelo autor

O DIF não uniforme, assim como no caso do DIF uniforme, as CCI's dos grupos em análise são diferentes, mas se cruzam em algum ponto ao longo da escala de habilidade. Isto ocorre quando os parâmetros a, b ou c (parâmetros de discriminação, dificuldade e acerto casual, respectivamente) apresentam valores distintos nas curvas do grupo de referência e do grupo focal. A Figura 3 representa as CCIs de um item com DIF não uniforme na escala (0,1); observa-se que as curvas não são paralelas e no ponto 1,5 as curvas se cruzam. Para os níveis de proficiências mais baixas (abaixo de 1,5), o item favorece o grupo focal. À medida que temos os dois grupos nivelados por proficiências mais altas (acima de 1,5), o DIF se inverte e passa a favorecer o grupo de referência.

Figura 3 – Representação gráfica de um item com DIF não uniforme



Fonte: Elaborado pelo autor

As Figuras 2 e 3 indicam possível existência de DIF, pois se a CCI do grupo referência e a CCI do grupo focal coincidissem a área entre as CCIs teria valor zero e, desse modo, não haveria DIF. Então, é realizado o cálculo da área compreendida nas CCIs com a fórmula proposta por Rudner, Getson e Knight (1980):

$$A = \sum_{\theta=-4}^{\theta=4} |P_{GR}(\theta_j) - P_{GF}(\theta_j)|\Delta\theta$$

onde, $P_{GR}(\theta_j)$ é o valor da probabilidade de acerto ao item do grupo de referência, dado θ_j; $P_{GF}(\theta_j)$ é o valor da probabilidade de acerto ao item do grupo de focal, dado θ_j e $\Delta\theta$ é o valor da base de um retângulo ($\Delta\theta = 0{,}005$) e altura $[P_{GR}(\theta_j) - P_{GF}(\theta_j)]$. Nesse tipo de análise, as áreas são calculadas para os distintos valores de θ que estejam compreendidos no intervalo de -4 a +4.

Após o cálculo da área entre as CCIs, o especialista poderá adotar a decisão a respeito da existência ou não de DIF. Entretanto, deverá ter algum tipo de cuidado já que não existem provas de significância estatística apropriada para confrontar as duas CCIs (MUÑIZ, 1997).

Além do método do cálculo da área entre as CCIs é possível realizar a comparação dos parâmetros dos itens. Ao estimar os parâmetros desse item em cada grupo, considera-se um item com DIF se os parâmetros estimados nos grupos não coincidirem, ou seja, têm diferenças significativas (THISSEN; STEINBERG; WAINER, 1993). No caso ML1, o parâmetro b será comparado nos grupos. A fórmula é:

$$Z = \frac{\hat{b}_R - \hat{b}_F}{\sqrt{S^2(\hat{b}_R) + S^2(\hat{b}_F)}}$$

onde, $\hat{b}_R$ e $\hat{b}_F$ são os parâmetros da dificuldade estimados para o grupo de referência e o grupo focal, respectivamente; $S^2(\hat{b}_R)$ e $S^2(\hat{b}_F)$ são as variâncias do parâmetro de dificuldade para o grupo de referência e o grupo focal, respectivamente, e Z tem distribuição Normal. O valor obtido em Z é comparado com o da distribuição Normal, correspondente ao nível de significância adotado, o que permite rejeitar ou não a hipótese nula $(H_0: \hat{b}_R = \hat{b}_F)$.

Para os ML2 e ML3, são comparados os parâmetros a e b, considerando o valor do parâmetro c invariante (MUÑIZ, 1997). As fórmulas matemáticas são:

$$Z_a = \frac{\hat{a}_R - \hat{a}_F}{\sqrt{S^2(\hat{a}_R) + S^2(\hat{a}_F)}} \quad e \quad Z_b = \frac{\hat{b}_R - \hat{b}_F}{\sqrt{S^2(\hat{b}_R) + S^2(\hat{b}_F)}}$$

onde, $\hat{a}_R$ e $\hat{a}_F$ são os parâmetros de discriminação estimados para o grupo de referência e o grupo focal, respectivamente; $S^2(\hat{a}_R)$ e $S^2(\hat{a}_F)$ são as variâncias do parâmetro de discriminação para o grupo de referência e o grupo focal, respectivamente; $\hat{b}_R$ e $\hat{b}_F$ são os parâmetros de dificuldade estimados para o grupo de referência e o grupo focal, respectivamente; $S^2(\hat{b}_R)$ e $S^2(\hat{b}_F)$ são as variâncias do parâmetro de dificuldade para o grupo de referência e o grupo focal, respectivamente, e Z tem distribuição Normal. A principal limitação desse procedimento é que os parâmetros a e b são comparados separadamente.

Devido à limitação do método de comparação dos parâmetros dos itens, Lord (1980) desenvolveu um método para comparar os parâmetros a e b simultaneamente pelo teste Qui-quadrado de Lord, dado pela fórmula:

$$\chi^2 = V\Sigma^{-1}V'$$

onde, χ^2 tem dois graus de liberdade; V é o vetor de dimensão (1 x 2) das diferenças entre os parâmetros a e b dos grupos de referência e focal; V' é o vetor transporto de V; Σ^{-1}é a inversa da matriz soma de variâncias-covariâncias de V para os grupos de referência e focal, cuja dimensão é 2 x 2. A significância estatística do Qui-quadrado é verificada comparando o valor observado com o valor crítico.

Um ponto importante a ser considerado na investigação do DIF é que a presença de um ou mais itens com DIF pode influenciar nos resultados dos testes de DIF em outros itens, ou seja, dependendo da magnitude do impacto do DIF, alguns itens podem ser identificados erroneamente com funcionamento diferencial. De acordo com Magis (2010), métodos que não utilizam os modelos da TRI para identificar DIF têm sua análise comprometida, pois a inclusão de itens com DIF influenciam o escore total.

Embora nos últimos anos muitos estudos baseados na TRI estejam disponíveis para identificar o funcionamento diferencial de itens entre os grupos, nenhum pode ser considerado

o melhor ou mais eficaz. Cabe ressaltar que, em geral, todos os métodos estatísticos de identificação do DIF compartilham essencialmente a mesma hipótese nula conceitual: isto é, os parâmetros dos itens são invariantes entre os grupos de referência e focal.

Assim como as metodologias apresentadas anteriormente o procedimento utilizado na detecção de itens com DIF pelo Inep assemelha-se ao método de áreas, pois é feita uma comparação entre as CCI's do grupo focal e do grupo de referência. A particularidade no procedimento se dá em compara a diferença na proporção de acerto esperada dos grupos analisados em cada ponto de quadratura, ou seja, o ponto médio de intervalos de θ. A classificação do item com DIF se dá quando a diferença das proporções esperada de acerto é superior a 0,15 em pelo menos um ponto de quadratura.

Na Tabela 1, Magis et al. (2010) listam os principais métodos de identificação de DIF, de acordo com o número de grupos analisados, a abordagem metodológica e o tipo de DIF. Os métodos listados admitem a utilização da purificação dos itens (eliminação iterativa dos itens que apresentam DIF).

Tabela 2 – Principais métodos para detectar Funcionamento Diferencial do Item (DIF).

Método	Tipo do DIF	Número de grupos	
		2	>2
Clássico	Uniforme	Mantel-Haenszel	Comparação por pares
		Padronização	Mantel-HaenszelGeneralizado
		SIBTEST	
		Regressão Logística	
	Não uniforme	Regressão Logística	
		Breslow-Day	
		NU.MH	
		Nu.SIBTEST	
Baseado em modelos da TRI	Uniforme	LRT	Comparação por pares
		Lord	Lord Generalizado
		Raju	
	Não uniforme	LRT	Comparação por pares
		Lord	Lord Generalizado
		Raju	

Fonte: Magis (2010)

3 METODOLOGIA

Nese capítulo, apresenta-se a procedimento utilizado pelo INEP para identificar os itens com DIF. Esse procedimento consiste em verificar a existência de DIF entre os itens comuns dos grupos que fazem parte da equalização, ou seja, consiste em comparar os itens do grupo fixado (grupo referência) com o grupo focal.

É importante salientar que o foco deste trabalho é analisar o impacto do DIF, portanto se limitou a analisar esse impacto utilizando a procedimento adotado pelo INEP, por se tratar da técnica mais empregada na área de avaliação educacional em larga escala no Brasil.

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 nos pontos de quadratura para cada grupo. Essas proporções esperadas são comparadas nos pontos de quadratura com maior concentração de indivíduos. E todo item que apresentar diferença absoluta máxima maior ou igual a 0,15 na intersecção do intervalo (percentil 5 e percentil 95) dos grupos, o item deverá ser considerado com DIF e, consequentemente, excluído ou recalibrado no grupo superior (grupo focal).

Todas as análises que serão apresentadas nos resultados são realizadas utilizando uma escala (0,1) do grupo de referência.

3.1 EQUALIZAÇÃO

Equalização é um procedimento matemático para ajustar a medida de habilidade (θ) entre grupos de indivíduos submetidos a diferentes testes, de forma a obter os parâmetros dos itens e as habilidades na mesma escala e então realizar comparações entre os grupos. De acordo com Araújo, Andrade e Bortolotti (2009), equalizar significa equiparar, tornar comparável, colocar os parâmetros dos itens provenientes de testes diferentes e os traços latentes dos indivíduos de diferentes grupos na mesma escala, tornando os itens e os respondentes comparáveis.

Existem dois principais cenários em que pode ocorrer equalização. O primeiro cenário é via itens, quando indivíduos de grupos diferentes respondem a testes parcialmente diferentes, ou seja, com itens comuns entre os testes, e o percentual mínimo aceitável de itens comuns é de 20%. Nesse cenário, deve-se realizar simultaneamente a equalização dos itens comuns e dos

itens não compartilhados que compõem os testes. O segundo cenário é via população, quando indivíduos da mesma população são submetidos a testes que abordam assuntos similares, entretanto, sem itens comuns (ANDRADE; TAVARES; VALE, 2000).

Os procedimentos de equalização contribuíram significativamente no avanço das avaliações educacionais em larga escala, uma vez que permite que indivíduos avaliados por instrumentos de avaliação parcialmente diferentes (com alguns itens em comum) ou totalmente diferentes (sem itens comuns) 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).

Para o processo de equalização das principais avaliações em larga escala realizadas no Brasil, o INEP utiliza o Modelo Logístico 3 parâmetros, com a estimação dos parâmetros sendo realizada pelo método da Máxima Verossimilhança Marginal (MVM), e as proficiências dos respondentes são estimadas pelo método *Expected a Posteriori* (EAP). Todas as estimativas são obtidas usando o *software* Bilog-MG.

O *software* Bilog-MG foi desenvolvido por Zimowski, Muraki, Mislevy e Bock (1996) com o propósito de estimar os parâmetros dos itens e as proficiências dos indivíduos. Conforme Andrade, Tavares e Valle (2000), esse software realiza análises via TRI para itens dicotômicos ou dicotomizados, possuindo implementações para as variações dos modelos logísticos unidimensionais de 1, 2 ou 3 parâmetros. A interface desse *software* é intuitiva e permite que o usuário desenvolva, passo a passo, a programação para o processamento dos dados, com do preenchimento das especificações nas opções dos menus de acesso.

O processo de estimação do Bilog-MG envolve 3 fases distintas. Na fase 1, são calculadas as estatísticas clássicas dos itens que compõem o teste (número de tentativas, número de acertos, percentual de acerto e a correlação bisserial). Nessa fase, verificam-se principalmente os valores das correlações bisseriais dos itens novos na tabela de múltiplos grupos. Os itens com bisseriais muito baixas (geralmente menor que 0,01) são excluídos da análise, pois são itens que apresentam algum problema pedagógico. Na fase 2, são estimados os parâmetros dos itens pelo método de MVM, supondo uma distribuição normal com média 0 (zero) e desvio padrão 1 (um) para as proficiências. Nessa fase, verifica-se a qualidade das estimativas dos parâmetros dos itens. Caso um determinado item apresente estimativas dos parâmetros absurdas, ou com problemas de convergência, esse item é descartado do processo

de equalização. Em geral, os itens são retirados da análise se apresentarem o parâmetro de discriminação menor que 0,70 na escala logística ou menor que 0,30 na escala ogiva normal, parâmetro de acerto ao acaso maior que 0,45 ou erro padrão muito alto em relação às demais estimativas. Se um item comum apresentar DIF, será excluído ou recalibrado no grupo superior de comparação. Na fase 3, os parâmetros dos itens gerados na fase 2, são utilizados para gerar as habilidades dos respondentes a partir do método EAP.

3.2 IDENTIFICAÇÃO DO DIF

Na segunda fase do processo de equalização do Bilog-MG, é gerado um arquivo denominado "*Expected*", o qual contém o valor esperado da amostra, o valor esperado do número de respostas corretas, a proporção esperada de respostas corretas, o resíduo padronizado a posteriori e a proporção de respostas corretas do modelo. Esses valores são avaliados para cada ponto de quadratura, sendo que o número de pontos determinado pelo INEP, nas análises das avaliações em larga escala, é de 40 pontos de quadratura.

Para os itens comuns, essas estatísticas são calculadas para cada grupo que é apresentado; de posse dos valores, é calculada a diferença da proporção esperada de acerto em cada um dos pontos. Caso o item apresente um valor maior ou igual a 0,15, o item é identificado com DIF. Também pode-se avaliar a qualidade do ajuste ao modelo para os itens novos.

3.3 MODELO ADOTADO NO ESTUDO

Neste estudo, por se tratar de uma simulação nos moldes de uma avaliação em larga escala como o 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:

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

onde, θ_{jk} é a habilidade do indivíduo j da população k, a_io parâmetro de discriminação do item i, b_io parâmetro de dificuldade do item i e c_io parâmetro de acerto ao acaso do item i.

3.4 SIMULAÇÃO DOS DADOS

O processo de simulação dos parâmetros dos itens e das habilidades dos indivíduos foi implementado computacionalmente no *software R*, versão 3.5.2. Os dados do grupo referência foram gerados a partir de N=100.000 respostas, considerando um teste com 45 itens novos e admitiu-se que habilidade θ segue uma distribuição normal com parâmetros $(\mu; \sigma^2) = (0; 1)$.

Para o grupo focal, foi simulado um N=100.000 respostas, considerando um teste parcialmente diferente com 45 itens e com a habilidade θ seguindo uma distribuição normal com parâmetros $(\mu; \sigma^2) = (0; 2)$. 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.

4 ANÁLISE DOS RESULTADOS

Neste capítulo serão apresentados os resultados de calibração e equalização dos itens simulados para o grupo referência e o grupo focal. Para geração das respostas, dos parâmetros (discriminação, dificuldade e acerto ao acaso) e das habilidades dos indivíduos, foi utilizado software R, versão 3.5.2. A calibração e a equalização dos itens foram implementadas no software Bilog-MG. O critério adotado para identificar os itens comuns com DIF uniforme foi o mesmo utilizado pelo INEP.

A Tabela 3 apresenta os valores dos parâmetros originais, os parâmetros estimados e os erros das estimativas associados às estimativas médias obtidas 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.

Figura 3 – Valores originais simulados e os erros de estimação associados às estimativas médias obtidas para os parâmetros dos itens

Valores originais simulados dos parâmetros dos itens								Erros de estimação							
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	-0,01	-0,01	-0,04	24	-0,01	-0,06	-0,06
2	1,520	0,335	0,152	25	1,631	0,474	0,161	2	-0,03	-0,02	-0,03	25	0,01	-0,02	-0,01
3	1,861	0,844	0,165	26	1,974	-0,076	0,221	3	0,03	0,00	-0,01	26	0,01	-0,01	0,04
4	1,920	-1,583	0,196	27	1,598	1,435	0,140	4	0,01	0,04	0,04	27	0,01	0,00	-0,03
5	1,224	1,360	0,242	28	1,718	1,692	0,205	5	0,01	0,01	0,07	28	0,07	-0,02	0,03
6	2,313	1,107	0,145	29	1,752	0,023	0,222	6	0,04	-0,01	-0,03	29	0,01	-0,01	0,04
7	1,509	-0,564	0,246	30	1,398	-1,495	0,128	7	0,04	0,02	0,08	30	0,02	0,00	-0,05
8	0,933	-1,539	0,185	31	2,071	-2,205	0,124	8	0,02	0,06	0,04	31	-0,03	-0,01	-0,05
9	1,560	0,565	0,136	32	1,856	-0,895	0,106	9	0,03	0,01	-0,04	32	0,01	0,00	-0,07
10	0,800	-0,689	0,123	33	2,031	-0,284	0,115	10	0,04	0,13	-0,01	33	0,02	0,02	-0,05
11	2,267	-1,887	0,241	34	2,113	0,500	0,165	11	0,00	-0,02	0,05	34	-0,01	-0,01	-0,01
12	0,830	-0,223	0,204	35	0,816	-0,255	0,225	12	-0,01	-0,06	0,01	35	0,02	0,02	0,06
13	2,060	-1,164	0,122	36	1,443	1,259	0,176	13	-0,01	-0,01	-0,06	36	-0,03	-0,02	0,00
14	2,089	0,442	0,134	37	2,122	-0,318	0,176	14	0,00	-0,02	-0,05	37	0,02	0,00	0,01
15	2,236	0,352	0,180	38	2,134	1,051	0,156	15	-0,03	0,00	0,00	38	-0,02	0,00	-0,02
16	1,757	1,637	0,242	39	0,950	-0,037	0,146	16	0,09	-0,04	0,07	39	0,00	0,03	-0,02
17	1,615	-0,405	0,141	40	2,169	0,487	0,103	17	-0,01	0,01	-0,03	40	0,00	-0,01	-0,07
18	2,436	0,336	0,244	41	2,125	0,418	0,141	18	0,00	0,00	0,07	41	0,03	0,00	-0,03
19	1,644	0,237	0,241	42	0,957	-0,540	0,157	19	0,01	0,01	0,07	42	-0,01	-0,06	-0,04
20	2,265	1,281	0,236	43	1,903	1,988	0,243	20	0,01	-0,01	0,06	43	0,00	-0,03	0,07
21	1,483	-0,248	0,182	44	1,764	-1,724	0,206	21	0,01	0,01	0,01	44	-0,04	-0,07	-0,01
22	2,076	-0,782	0,106	45	0,992	3,067	0,241	22	0,07	0,04	-0,04	45	-0,01	-0,05	0,06
23	2,119	-1,172	0,137	-	-	-	-	23	0,00	0,01	-0,03	-	-	-	-

Fonte: Elaborado pelo autor

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

$$\mathrm{EQM}_{\zeta} = \frac{1}{\mathrm{I}}\sum_{\mathrm{i}=1}^{\mathrm{I}}\left(\hat{\zeta}_{\mathrm{i}} - \zeta\right)^{2}.$$

Para cada parâmetro estimado, foram encontrados $\mathrm{EQM}_{\mathrm{a}} = 0{,}008$, $\mathrm{EQM}_{\mathrm{b}} = 0{,}001$ e $\mathrm{EQM}_{\mathrm{C}} = 0{,}002$, 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.

A Tabela 4 apresenta as estimativas médias dos parâmetros com seus respectivos desvios-padrão.

Tabela 4 – Estimativas médias dos parâmetros dos itens e seus desvios-padrão

Estimativa média (desvio-padrão) dos parâmetros dos itens							
Item	a_i	b_i	c_i	Item	a_i	b_i	c_i
1	0,887(0,02)	-0,468(0,05)	0,185(0,02)	24	1,751(0,02)	-1,408(0,03)	0,216(0,02)
2	1,554(0,02)	0,351(0,01)	0,180(0,00)	25	1,619(0,02)	0,497(0,01)	0,176(0,00)
3	1,835(0,03)	0,847(0,01)	0,175(0,00)	26	1,967(0,02)	-0,066(0,01)	0,177(0,00)
4	1,905(0,03)	-1,623(0,03)	0,153(0,02)	27	1,591(0,03)	1,436(0,01)	0,171(0,00)
5	1,217(0,03)	1,351(0,01)	0,173(0,00)	28	1,644(0,03)	1,708(0,01)	0,171(0,00)
6	2,277(0,03)	1,116(0,01)	0,172(0,00)	29	1,739(0,02)	0,028(0,01)	0,178(0,01)
7	1,466(0,02)	-0,587(0,02)	0,169(0,01)	30	1,381(0,02)	-1,493(0,04)	0,174(0,02)
8	0,915(0,02)	-1,599(0,08)	0,149(0,03)	31	2,102(0,03)	-2,193(0,04)	0,178(0,03)
9	1,532(0,02)	0,560(0,01)	0,175(0,00)	32	1,849(0,02)	-0,892(0,02)	0,175(0,01)
10	0,761(0,02)	-0,819(0,08)	0,135(0,03)	33	2,006(0,02)	-0,301(0,01)	0,164(0,01)
11	2,263(0,03)	-1,872(0,03)	0,193(0,02)	34	2,123(0,03)	0,512(0,01)	0,178(0,00)
12	0,837(0,02)	-0,158(0,06)	0,192(0,02)	35	0,797(0,02)	-0,274(0,06)	0,162(0,02)
13	2,074(0,03)	-1,150(0,02)	0,178(0,01)	36	1,478(0,03)	1,277(0,01)	0,181(0,00)
14	2,091(0,03)	0,462(0,01)	0,181(0,00)	37	2,103(0,02)	-0,322(0,01)	0,171(0,01)
15	2,262(0,03)	0,356(0,01)	0,177(0,00)	38	2,153(0,03)	1,054(0,01)	0,178(0,00)
16	1,668(0,03)	1,674(0,01)	0,172(0,00)	39	0,947(0,02)	-0,064(0,04)	0,168(0,01)
17	1,623(0,02)	-0,412(0,02)	0,176(0,01)	40	2,174(0,03)	0,497(0,01)	0,177(0,00)
18	2,440(0,03)	0,340(0,01)	0,174(0,00)	41	2,095(0,03)	0,417(0,01)	0,174(0,00)
19	1,633(0,02)	0,227(0,01)	0,169(0,00)	42	0,969(0,02)	-0,483(0,05)	0,196(0,02)
20	2,258(0,04)	1,291(0,01)	0,175(0,00)	43	1,898(0,05)	2,022(0,02)	0,177(0,00)
21	1,473(0,02)	-0,262(0,02)	0,174(0,01)	44	1,804(0,03)	-1,657(0,03)	0,221(0,02)
22	2,006(0,02)	-0,823(0,01)	0,147(0,01)	45	1,000(0,05)	3,119(0,07)	0,177(0,00)
23	2,120(0,03)	-1,183(0,02)	0,172(0,01)	-	-	-	-

Fonte: Elaborado pelo autor

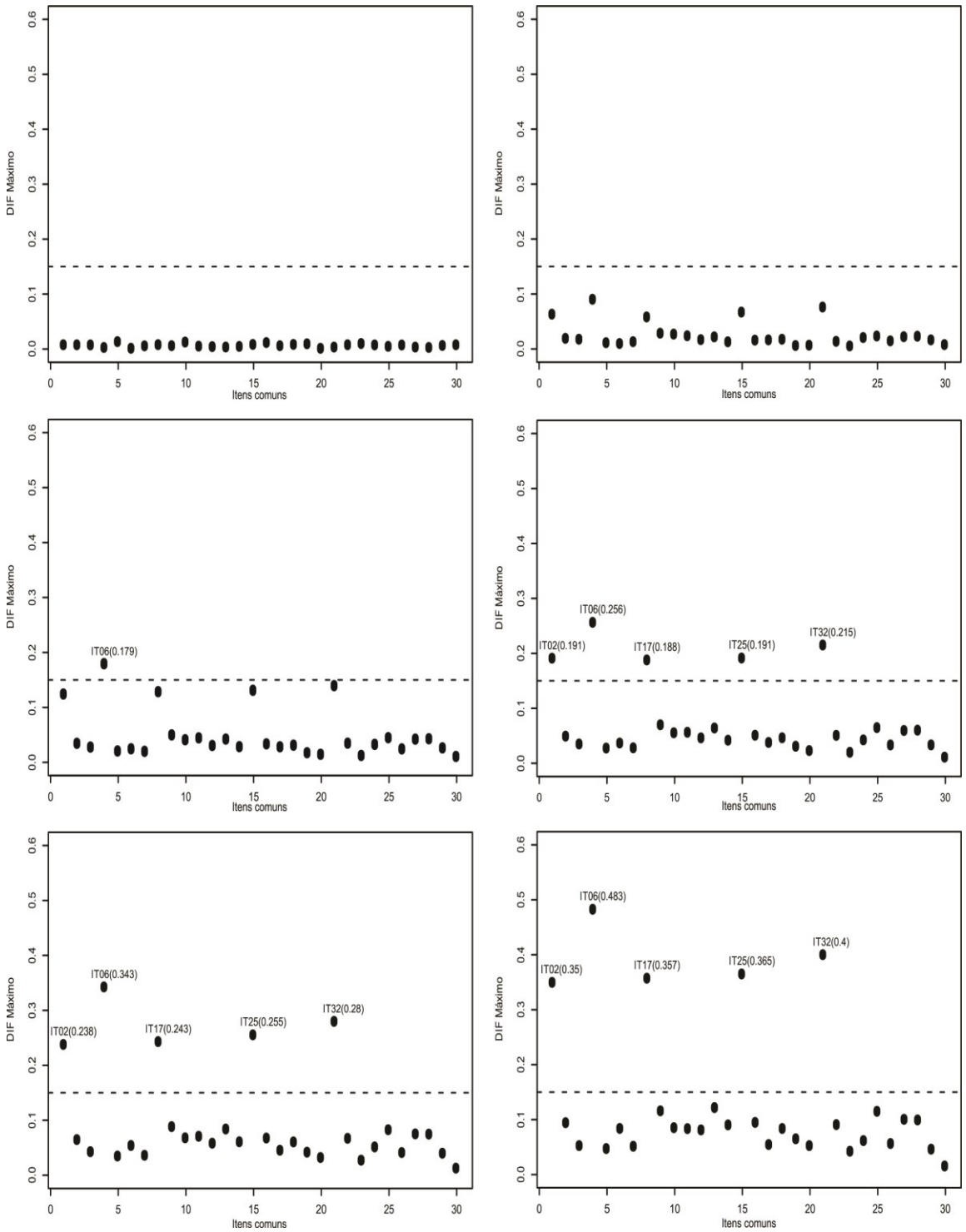
Como as estimativas dos parâmetros simulados retornaram os valores originais de forma satisfatória, procedeu-se às análises para detecção do DIF, nos cenários descritos a seguir.

4.1 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 foi considerado com DIF, a fim de garantir itens com diferentes graus de dificuldade. 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 nos itens ocorreu pela verificação da diferença nas proporções esperadas de acertos para os grupos analisados.

Um item não apresenta DIF uniforme se a diferença máxima nas proporções de acertos não for superior ou igual a 0,15, devendo ser analisada em um intervalo compreendido pelos pontos P5 (percentil 5) e o P95 (percentil 95) da distribuição dos traços latentes. No entanto, 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 4 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 indicando o limite de 0,15.

Figura 4 – 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 pode 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, 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.

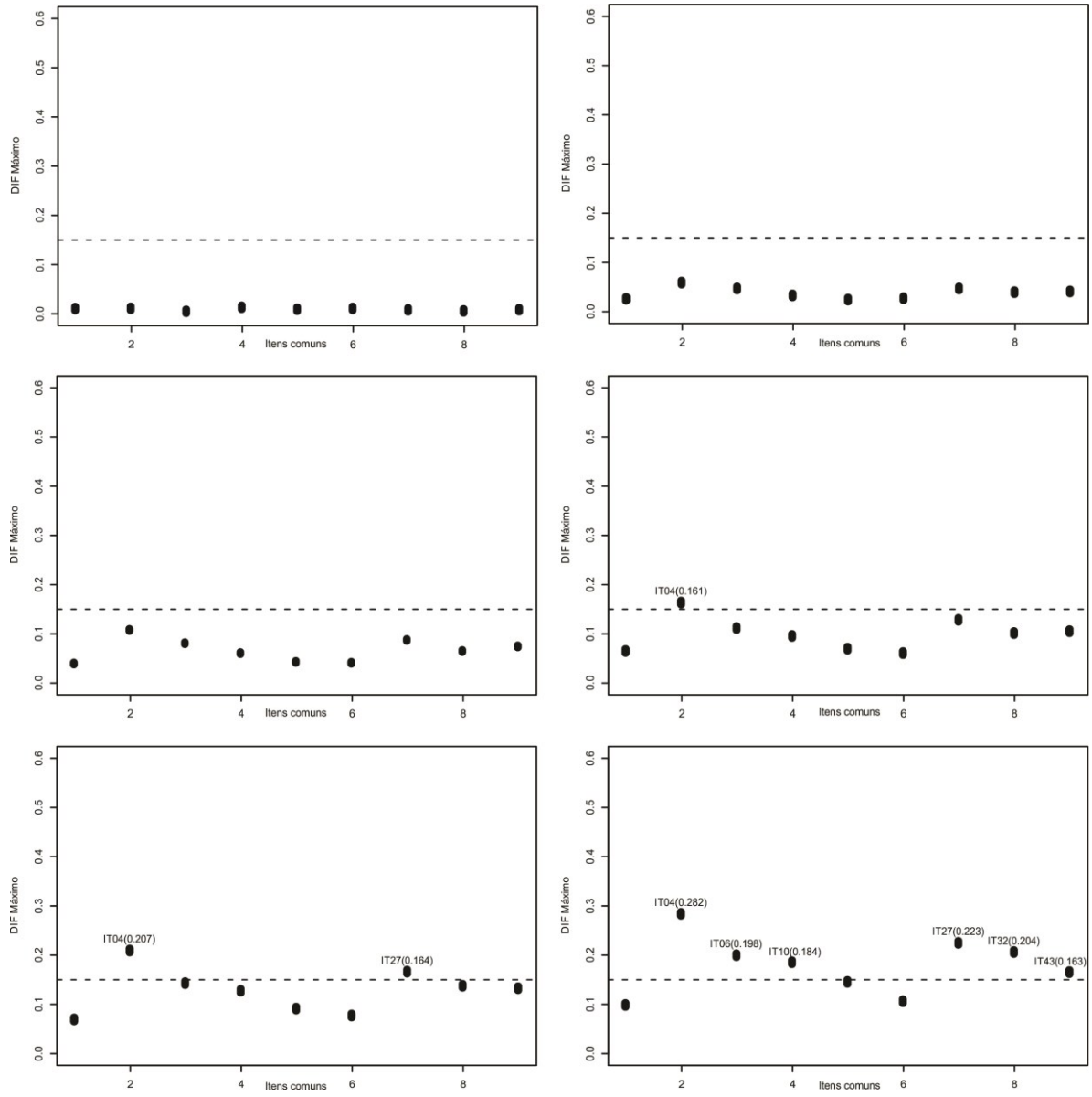
Como forma de minimizar a perda no número de itens comuns, assim como apresentado por Magis (2010) em outros métodos de detecção de DIF, utilizou-se o processo de purificação, ou seja, a retirada iterativa de itens com DIF. Tal procedimento, porém, não apresentou melhora nos resultados. A retirada parcial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

4.2 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 no 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 5, 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 5 – 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% de itens comuns



Fonte: Elaborado pelo autor

O primeiro incremento utilizado reflete uma análise sem a presença de itens com DIF uniforme, 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. 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 e 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.

4.3 IMPACTO DO DIF

Para analisar o impacto nas proficiências do grupo focal em um teste cognitivo que apresenta itens com DIF, foi aplicado 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 utilizando o 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 considerando os parâmetros fixados dos itens com DIF no teste composto por 45 itens (15 novos + 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 novos + 25 comuns.

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

Na Tabela 5 são apresentadas as médias por faixa de proficiência do grupo focal para cada tipo de tratamento.

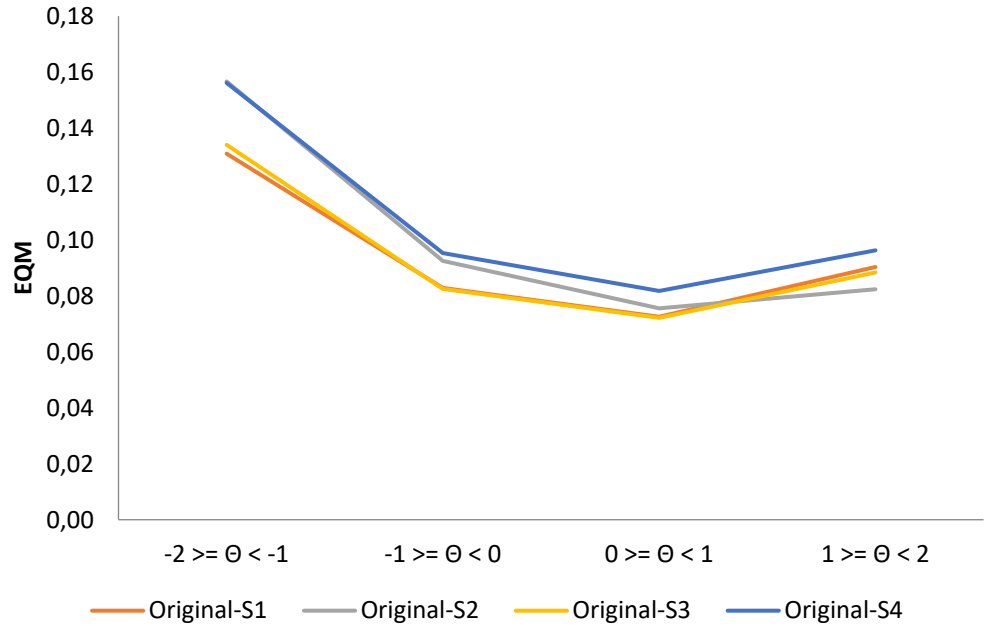
Tabela 5 – Média do grupo focal por faixa de desempenho

Faixa	Habilidade média				
	Habilidade 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 habilidade média do grupo focal nas diferentes situações, para a simulação de um impacto de -0,50 no parametro de dificuldade, o procedimento que mais se aproxima do valor real (habilidade 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 restimados. Foi calcular o EQM nos diferentes tratamentos para cada faixa de proficiência, como apresentado na Figura 6.

Figura 6 – Tratamentos e EQM por faixa de habilidade



Fonte: Elaborado pelo autor

Na Figura 6 a categoria “Original-S1” representa a diferença entre as habilidades originais e as habilidades de um teste sem a presença de itens com DIF, as categorias “Original-S2”, “Original-S3” e “Original-S4” representam a diferença entre as habilidades originais e as habilidades de um teste com a presença de itens com DIF e 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 forma recalibrado 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).

5 CONSIDERAÇÕES FINAIS

O pioneirismo norte-americano em estudos para compreender a realidade e mensurar a qualidade educacional serviu de norte para o Brasil consolidar um sistema de avaliação próprio, a partir dos anos de 1990, chamado SAEB (Sistema de Avaliação da Educação Básica) e, em 1998, criou-se o Exame Nacional do Ensino Médio (ENEM) com o objetivo de avaliar o desempenho dos estudantes egressos do Ensino Médio. Ao longo dos anos, essas avalições sofreram alterações nas metodologias utilizadas para gerar seus resultados, como a inserção da Teoria de Resposta ao Item (TRI).

As informações geradas, fundamentadas nos resultados disponibilizados pelas avaliações, subsidiam o planejamento estratégico e o monitoramento da política educacional. Entretanto, essas informações necessitam da padronização e uniformização dos instrumentos utilizados para aferir o desempenho dos estudantes da rede de ensino.

Segundo Martinez Arias (1997), para uniformidade dos testes são necessários critérios técnicos (matriz de referência, revisão de conteúdo e revisão linguística) adequados no momento da elaboração dos itens. Após os itens atenderem todos os critérios técnicos, são finalmente pré-testados em amostras ou populações de estudantes. A realização do pré-teste é fundamental para calibrar os itens de acordo com TRI. Porém, Hambleton (1997) sustenta que validade da medida gerada e a equidade dos testes são asseguradas após verificação da existência de Funcionamento Diferencial do Item (DIF).

Na literatura, existem vários métodos para identificar item com DIF e alguns autores recomendam a utilização de mais de um método nesse processo, visto que cada método apresenta sua limitação. Sisto (2006) defende o uso dos métodos baseados nos modelos da TRI, uma vez que são superiores aos métodos clássicos, dado que os parâmetros estimados pelo modelo da TRI são invariantes na amostra.

O estudo para detectar itens com DIF nas etapas de equalização se justifica para que as proficiências estimadas sejam fidedignas e os resultados possam ser adequadamente equalizados. Os itens identificados com DIF podem trazer informações importantes para os sistemas educacionais, pois conseguem diagnosticar deficiências curriculares, índices de discriminação racial, características econômicas ou diversidades culturais.

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 geram 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 uma diferença significativa nas habilidades médias do grupo focal que responderam a um teste com itens com DIF em relação as habilidades originais desse grupo. Para minimizar o impacto nas habilidades dos indivíduos foi aplicado três tipos de tratamentos nos itens com DIF e calculado o Erro Quadrático Médio (EQM) para cada faixa de habilidade. 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 habilidades dos indivíduos.

Como possibilidade de pesquisas futuras, sugere-se uma abordagem de identificação utilizando um processo de reamostragem para o cálculo das proporções esperadas de acerto, sendo possível a criação de intervalos de confiança para as proporções esperadas para cada ponto de quadratura. O item apresentaria DIF, caso pelo menos um dos pontos os intervalos do grupo focal e do grupo de referência não se cruzassem. Sugerem-se, ainda, estudos com novos valores de incrementos para o parâmetro de dificuldade e a inserção do DIF não uniforme.

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.

ANDRICH, D. A rating formulation for ordered response categories. *Psychometrika,* 43, 561-573, 1978.

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

ARAÚJO, E. A. C.; ANDRADE, D. F.; BORTOLOTTI, S. L. V. *Teoria de Resposta ao Item*. 2009.

BAKER, Frank B.; KIM, Seock-ho. *Item response theory*: parameter estimation techniques. 2nd edition. Marcel Dekker, Inc, 2004.

BIRNBAUM, A. Some latent trait models and their use in infering an examinee's ability. In: LORD, F. M.; NOVICK, M. R. *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley, 1968.

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

BOCK, R. D.; ZIMOWSKI, M. F. Multiple Group IRT. In: *Handbook of modern item response theory*. New York: Spring-Verlag, 1997.

BONAMINO, Alicia. *Tempos de avaliação educacional*: o Saeb, seus agentes, referências e tendências. Rio de Janeiro: Quartet, 2002.

BOLT, Daniel M. A SIBTEST approach to testing DIF hypotheses using experimentally designed test items. *Journal of Educational Measurement*, v. 37, n. 4, p. 307-327, 2000.

BROWN, T. A. *Confirmatory factor analysis for applied research*. New York: The Guilford Press, 2006.

COUTO, G.; PRIMI, R. Teoria de Resposta ao Item (TRI): conceitos elementares dos modelos para itens dicotômicos. *Boletim de Psicologia*, 2011.

CRANE, Paul K. et al. Differential item functioning analysis with ordinal logistic regression techniques: DIF detect and DIF withpar. *Medical Care*, p. S115-S123, 2006.

DORANS, N. J.; KULLICK, E. Demonstrating the utility of the standardization approach to assessing unexpected differential item performance on the Scholastic Aptitude Test. *Journal of Educational Measurement*, 23, p. 355-368, 1986.

DORANS, N. J.; SCHMITT, A. P.; BLEISTEIN, C. A. The standardization approach to assessing comprehensive differential item functioning. *Journal of Educational Measurement*, 29, p. 309-319, 1992.

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.

_____. *Item response theory*. Hove, United Kingdom: Psychology Press, 2013.

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

FRANCO, C. Quais as contribuições da avaliação para políticas educacionais? In: BONAMINO, A.; BESSA, N.; FRANCO, C. (Org.) *Avaliação da educação básica*. Rio de Janeiro: PUC-Rio; São Paulo: Loyola, 2004.

GONTIJO, R. *Manoel Bomfim, educador e "cientista da educação"*. [S.1.:s.n.], 2008.

HAMBLETON, R. K. Perspectivas futuras y aplicaciones. In: MUÑIZ, J. *Introducción a la teoría de respuesta a los ítems*. Madrid: Ediciones Psicología Pirámide, 1997.

HAMBLETON, R. K.; SWAMINATHAN, H.; ROGERS, H. J. *Fundamental of Item Response Theory*. London: Sage Publications, 1991.

HOLLAND, P. W.; THAYER, D. T. Differential item performance and the Mantel-Haenszel procedure. In: WAINER, H.; BRAUN, H. (Eds.). *Test validity*. Hillsdale, New Jersey: Lawrence Erlbaum Associates, Publishers, 1988.

HORTA NETO, J. L. *Avaliação externa:* a utilização dos resultados do SAEB 2003 na gestão do sistema público de Ensino Fundamental no Distrito Federal. Dissertação de Mestrado. Universidade de Brasília. Brasília, 2006.

IOWA TESTING PROGRAMS. *The Iowa Tests of Basic Skills*. Disponível em: <http://www.education.uiowa.edu/itp/itbs/>. Acesso em: 10 fev. 2017.

LORD, F. M. *A theory of test scores.* Psychometric Monograph, n. 7, 1952.

LORD, F. M. *Applications of item response theory to practical testing problems*. New Jersey: Lawrence Erlbaum, 1980.

LORD, F. M.; NOVICK, M. R. *Statistical theories of mental test scores.* Reading, MA: Addison-Wesley, 1968.

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.

MANTEL, N.; HAENSZEL, W. Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute, 22*, p. 719-748, 1959.

MARIANI, M. C. Educação e ciências sociais: o Instituto Nacional de Estudos e Pesquisas Educacionais. In: SCHWARTZMAN, S. (Org.). *Universidades e Instituições Científicos no Rio de Janeiro*. Brasília: CNPq, 1982.

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

MASTERS, G. N. A Rasch model for partial credit scoring. *Psychometrika*, 47, 149-174, 1982.

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

_____. Item bias and item response theory. *International Journal of Educational Research,* 13, p. 127-143, 1989.

MILLSAP, R. E.; EVERSON, H. T. Methodology review: statistical approaches for assessing measurement bias. *Applied Psychological Measurement*, p. 277-334, 1993.

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

MURAKI, E. A generalized partial credit model: application of an EM algorithm. *Applied Psychological Measurement*, p.159-176, 1992.

NARAYANAN, P.; SWAMINATHAN, H. Performance of Mantel-Haenszel and simultaneous item bias procedure for detecting differential item functioning. *Applied Psychological Measurement*, 18, 315-328, 1994.

OLIVEIRA, A. P. M. *A Prova Brasil como política de regulação da rede pública do Distrito Federal,* 2012.

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.

RASCH, G. *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Danish Institute for Educational Research, 1960.

ROUSSOS, L; STOUT, W. A multidimensionality-based DIF analysis paradigm. *Applied Psychological Measurement*, v. 20, n. 4, p. 355-371, 1996.

RUDNER, L. M.; GETSON, P. R.; KNIGHT, D. L. Biased item detection techniques. *Journal of Educational Statistics*, 5, 213-233, 1980.

SAMEJIMA, F. A. Estimation of latent ability using a response pattern of graded scores. *Psychometric Monograph*, 17, 1969.

SARTES, L. M. A.; SOUZA-FORMIGONI, M. L. O. Avanços na psicometria: da teoria clássica dos testes à teoria de resposta ao item. *Psicologia: Reflexão e Crítica*, Porto Alegre, v. 26, n. 2, 2013.

SHEALY, R.; STOUT, W. A model-based standardization approach that separates true bias/DIF from group ability differences and detects test bias/DIF as well as item bias. *Psychometrika*, 58, p.159-194, 1993.

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

SOUZA, E, R. *Accountability de professores*: um estudo sobre o efeito da Prova Brasil em escolas de Brasília. Dissertação de Mestrado. Universidade Estadual de Campinas. Campinas, São Paulo, 2009.

TRAVER, D. F.; ROMA, V. G.; BENITO, J. G. Mantel-Haenszel statistic and the logistic regression in the detection of DIF in two aptitude tests. *Psicothema*, v. 12, p. 214-219, 2000.

SWAMINATHAN, H.; ROGERS, H. J. Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement*, v. 27, p. 361-370, 1990.

SWANSON, D. B. et al. Analysis of differential item functioning (DIF) using hierarchical logistic regression models. *Journal of Educational and Behavioral Statistics*, v. 27, n. 1, p. 53-75, 2002.

TAVARES, C. M. *A teoria de resposta ao item na avaliação em larga escala*: um estudo sobre o exame nacional de acesso ao mestrado profissional em matemática em rede nacional - PROFMAT. Dissertação de Mestrado. 2014.

THISSEN, D.; STEINBERG, L.; WAINER, H. Detection of differential item functioning using the parameters of item response models. In: HOLLAND, P. W.; WAINER, H. (Orgs.). *Differential item functioning*. New Jersey: Lawrence Erlbaum, 1993.

VIANNA, H. V. *Fundamentos de um programa de avaliação educacional*. Brasília: Líber Livro, 2005.

WHITMORE, M. L.; SHUMACKER, R. E. A comparison of logistic regression and analysis de variance differential item functioning detection methods. *Educational and Psychological Measurement*, v. 59, n. 6, p. 910-927, 1999.

WRIGHT, B. D. Sample-free test calibration and person measurement. *Proceedings of the 1967 Invitational Conference on Testing Problems*. Princeton, N. J.: Educational Testing Service, 1968.

ZIMOWSKY, M.F.; MURAKI, E.; MISLEVY, R.J.; BOCK, R.D. *BILOGMG*: multiple-group IRT analysis and test maintenance for binary items. Scientific Software, Inc., Chicago, 1996.