UNIVERSIDADE FEDERAL DO PARÁ
INSTITUTO DE CIÊNCIAS EXATAS E NATURAIS
PROGRAMA DE PÓS-GRADUACÃO EM MATEMÁTICA E ESTATÍSTICA

IMPACTO LONGITUDINAL NO PROCESSO DE EQUALIZAÇÃO DA ESCALA DO SAEB

Thamara Rúbia Almeida de Medeiros

Orientação: **Profa. Dra. Maria Regina Madruga Tavares**
Coorientação: **Prof. Dr. Héliton Ribeiro Tavares**

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Código de Financiamento 001

Belém
2019

Thamara Rúbia Almeida de Medeiros

IMPACTO LONGITUDINAL NO PROCESSO DE EQUALIZAÇÃO DA ESCALA DO SAEB

Dissertação apresentada ao Curso de Mestrado em Matemática e Estatística da Universidade Federal do Pará, como pré-requisito para a obtenção do título de Mestre em Estatística.

Orientação: **Profa. Dra. Maria Regina Madruga Tavares**

Coorientação: **Prof. Dr. Héliton Ribeiro Tavares**

Belém

2019

Dados Internacionais de Catalogação na Publicação (CIP) de acordo com ISBD
Sistema de Bibliotecas da Universidade Federal do Pará
Gerada automaticamente pelo módulo Ficat, mediante os dados fornecidos pelo(a) autor(a)

M488i Medeiros, Thamara Rúbia Almeida
Impacto Longitudinal no Processo de Equalização da Escala do Saeb / Thamara Rúbia Almeida Medeiros. — 2019.
60 f. : il. color.

Orientador(a): Profª. Dra. Maria Regina Madruga Tavares
Coorientador(a): Prof. Dr. Héliton Ribeiro Tavares
Dissertação (Mestrado) - Programa de Pós-Graduação em Matemática e Estatística, Instituto de Ciências Exatas e Naturais, Universidade Federal do Pará, Belém, 2019.

1. Teoria da Resposta ao Item. 2. Equalização. 3. SAEB. I. Título.

CDD 310

Thamara Rúbia Almeida de Medeiros

IMPACTO LONGITUDINAL NO PROCESSO DE EQUALIZAÇÃO DA ESCALA DO SAEB

Esta Dissertação foi julgada e aprovada para a obtenção do grau de Mestre em Estatística, no Programa de Pós-Graduação em Matemática e Estatística da Universidade Federal do Pará.

Belém, 21 de Março de 2019

João Marcelo B Protazio

Prof. Dr. João Marcelo Brazão Protázio
(Coordenador(a) do Programa de Pós-Graduação em Matemática e Estatística - UFPA).

Banca Examinadora

Maria Regina Madruga Tavares

Profa. Dra Maria Regina Madruga Tavares
PPGME/UFPA
Orientador(a)

Heliton

Prof. Dr. Heliton Ribeiro Tavares
PPGME/UFPA
Coorientador(a)

João Marcelo B Protazio

Prof. Dr. João Marcelo Brazão Protázio
PPGME/UFPA
Examinador(a) Externo

Prof. Dr. Dalton Francisco de Andrade
UFSC
Examinador(a) Externo

A minha mãe Rubenita Medeiros.

Agradecimentos

A Deus por me abençoar, ouvir, iluminar minha vida e por realizar os meus sonhos no tempo certo.

A minha orientadora, Profa. Dra. Maria Regina Madruga Tavares e meu coorientador Prof. Dr. Héliton Ribeiro Tavares, pela paciência, profissionalismo, motivação e todos os ensinamentos compartilhados durante minha graduação e mestrado. Sou eternamente grata a vocês por tudo.

A minha mãe, Rubenita Medeiros, por ser minha fonte de inspiração, por todo seu amor, dedicação, apoio, bondade e o grande motivo para nunca desistir. Obrigada mãe!

A minha família, em especial ao meu pai, Theyrimar Medeiros, as minhas avós Terezinha e Anna Maria, ao meu tio José Roberto e meu irmão Theyrimar Júnior, pelo amor incondicional, por entenderem minha ausência durante esse período do mestrado e pelo apoio para que este sonho se tornasse realidade.

Aos colegas e amigos do PPGME, em especial Armando, Alice, Andrey, Fernando, Miguel e Robinson Ortega, por dividirem comigo as fases ruins e boas do mestrado, pela companhia e grande ajuda nessa fase final. Gratidão amigos.

Ao Andrey, pelo companheirismo, paciência e todo apoio emocional e espiritual. Gratidão!

Aos meus amigos, Helen, Rayssa, Rodrigo, Mônica, Erick, Suellainy, Sayuri e Larissa por todo apoio e incentivo para nunca desistir. Gratidão amigos.

Aos professores, Marinalva, Marina, Valcir e João Marcelo pela atenção, carinho e apoio proporcionado durante todos esses anos.

Ao Prof. Dr. Dalton de Andrade, pela disponibilidade para participar da banca, e pelas valiosas contribuições para a pesquisa.

Finalmente, gostaria de agradecer à UFPA pelo ensino gratuito de qualidade, ao PPGME e à CAPES, sem os quais essa dissertação dificilmente poderia ter sido realizada e a todos

mais que eu não tenha citado nesta lista de agradecimentos, mas que de uma forma ou de outra contribuíram não apenas para a minha dissertação, mas também para eu ser quem eu sou.

"Existir é sobreviver a escolhas injustas."

The OΛ

"As nuvens mudam sempre de posição, mas são sempre nuvens no ceu. Assim devemos ser todo dia, mutantes, porém leais com o que pensamos e sonhamos; lembre-se, tudo se desmancha no ar, menos os pensamentos."

Paulo Beleki

Resumo

Em diversas avaliações educacionais a Escala do Sistema de Avaliação da Educação Básica (SAEB) tem sido usada como a referência nacional, com muitas avaliações estaduais sendo criadas e adotando tal escala como padrão. Porém, têm sido observadas diferenças consideráveis entre as proficiências médias dos estados, nas áreas de Língua Portuguesa e Matemática e de alguns níveis escolares (Ensino Fundamental 1, Ensino Fundamental 2 e Ensino Médio), divulgadas pelo Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP) e as respectivas proficiências médias produzidas pelas avaliações estaduais. Este trabalho teve como objetivo avaliar os possíveis motivos dessas variações e, para isso, realizou-se a equalização conjunta das edições da prova SAEB de 2011 a 2017 para verificar se a equalização realizada desta maneira reproduziria as médias das proficiências estimadas pelo INEP, mantidos os mesmos critérios de análise. Os resultados obtidos apresentaram diferenças consideráveis entre as médias das proficiências estimadas na estimação conjunta em relação às proficiências médias estimadas pelo processo de equalização tradicional do INEP.

Palavras-chave: Teoria da Resposta ao Item, Equalização, SAEB.

Abstract

In several educational evaluations the Scale of the Basic Education Assessment System (SAEB) has been used as the national reference, with many state assessments being created and adopting this scale as the standard. However, considerable differences have been observed between the average state proficiency, in the areas of Portuguese Language and Mathematics and levels (Elementary School 1, Elementary School 2 and High School), published by the National Institute of Educational Studies and Research Anísio Teixeira (INEP) and the respective average proficiencies produced by the state assessments. The purpose of this study is to evaluate the possible reasons for these variations and, for this, the joint equalization of the editions of the SAEB test from 2011 to 2017 was carried out to verify if the equalization carried out in this way would reproduce the averages of the proficiencies estimated by the INEP, following the same criteria for analysis. The results obtained presented considerable differences between the averages of the proficiencies estimated in the joint estimation in relation to the average proficiency estimated by the traditional equalization process of the INEP.

Keywords: Item Response Theory, Equating, SAEB.

Sumário

Lista de Tabelas

Lista de Figuras

Capítulo 1

Introdução

1.1 Aspectos gerais

Sabe-se que até pouco antes de 1990 não existia um levantamento bem aprofundado de dados a respeito do sistema de ensino básico brasileiro, somente o monitoramento acerca de números de matrículas, aprovações, reprovações e nível de evasão escolar (Horta Neto, 2006). Na prática, não havia nenhuma estratégia melhor para mensurar o comportamento do sistema educacional no país. À vista disso, na década de 90 foi implementado o Sistema de Avaliação da Educação Básica (SAEB), com a finalidade de verificar a qualidade do ensino e possíveis fatores que interferem no desenvolvimento da aprendizagem do estudante.

O SAEB é uma avaliação em larga escala conduzida pelo Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP), que abrange três níveis: Anos Iniciais do Ensino Fundamental (EF1), Anos Finais (EF2) e Ensino Médio (EM). Ela envolve as áreas de Língua Portuguesa (LP) e Matemática (MT), e a partir da edição de 2019 será cobrado para os alunos do EF2 o conhecimento em Ciências da Natureza e Ciências Humanas.

Nas primeiras aplicações do SAEB seus resultados foram obtidos mediante a Psicometria Clássica, conhecida como a Teoria Clássica dos Testes (TCT). Neste método a avaliação da proficiência e a interpretação dos resultados são relacionados à prova (teste) como o todo. Por esse motivo, proceder uma comparação entre alunos que não realizaram o mesmo teste torna-se inviável. De modo geral, a finalidade da TCT é medir a competência do aluno baseado na soma dos acertos dos itens, ou seja, calculando os escores brutos (Pasquali,2009). Nos trabalhos de Guilford (1954) e Gulliksen (1950) encontram-se mais informações sobre o método.

Atualmente os resultados dessa avaliação são divulgados na conhecida Escala SAEB. Esta escala foi instituída em 1997 quando da implementação da Teoria da Resposta ao

Item (TRI) como método de equalização horizontal e vertical da avaliação. A respeito da TRI, Andrade, Tavares e Valle (2000) afirmam que ela apresenta formas de interpretar a relação entre a probabilidade de um indivíduo (examinado) dar uma certa resposta a um item em função dos parâmetros desse item e dos traços latentes, habilidades ou proficiências do indivíduo na área do conhecimento avaliada. Vale frisar que, conforme Andriola (2009), a TRI tem seu foco no item e não no teste como todo, sendo justamente nesse fato que se encontra a mais significativa distinção entre a TRI e a TCT. Além disso, nesse mesmo ano foi estabelecido uma escala, de forma que as proficiências médias nacionais da 8ª Série (hoje 9º Ano) teriam média 250 e desvio-padrão 50, tanto em LP como em MT, incluindo alunos da rede Regular de Ensino de escolas Públicas e Privadas, Urbanas e Rurais.

Uma das etapas mais determinantes para o sucesso da análise de dados e a consequente garantia de comparabilidade dos resultados do SAEB é a fase de estimação dos parâmetros dos itens envolvidos na avaliação, também conhecida por *calibração*, que antecede a etapa de estimação das proficiências. De acordo com a TRI, se um conjunto de itens já tiver calibrado, novos itens poderão ser calibrados na mesma escala daqueles calibrados. Por conta disso, um processo de calibração de itens do SAEB sempre inclui itens da edição anterior (ou edições anteriores) do SAEB, motivo pelo qual os itens não podem ser divulgados.

1.2 Justificativa e Importância da Dissertação

Durante vários anos a Escala SAEB tem sido usada como a referência nacional, com muitas avaliações estaduais sendo criadas e adotando tal escala como padrão. Para isso, o INEP repassa aos estados e demais instituições interessadas um conjunto de itens calibrados com base no Modelo Logístico de 3 parâmetros da TRI (além de uma base de dados usada para calibrar tais itens), que devem compor as respectivas avaliações juntamente com itens próprios, o que, teoricamente, deve garantir que os itens próprios sejam calibrados na escala SAEB, e assim as PROFICIÊNCIAS estimadas também estarão na escala SAEB. No entanto, nos últimos anos têm sido observadas diferenças consideráveis nas proficiências médias dos estados, considerando as áreas de LP e MT, e níveis (EF1, EF2 e EM) divulgadas pelo INEP e as respectivas proficiências médias produzidas pelas avaliações estaduais. Este fato tem intrigado muitos especialistas da área de TRI, e o

próprio INEP sinaliza para a criação de uma linha de pesquisa tentando identificar quais fatores poderiam estar interferindo na manutenção da escala do SAEB.

1.3 Objetivos

1.3.1 Objetivo Geral

Diante do exposto, esta dissertação de Mestrado tem por objetivo reconstruir o processo de equalização do SAEB com os mesmos critérios atualmente adotados, bem como com métodos alternativos para verificar se a escala SAEB será mantida.

1.3.2 Objetivos Específicos

i) Recuperar as estimativas dos parâmetros dos itens da prova SAEB do ano 2011 por meio do pacote *ReCalibra*;

ii) Realizar estudo da equalização conjunta das provas SAEB de 2011, 2013, 2015 e 2017.

1.4 Sumário da Dissertação

Este trabalho encontra-se organizado em 5 capítulos, a saber:

- No Capítulo 1 é feita uma introdução aos conceitos da prova SAEB e da TRI, são abordados os aspectos gerais, justificativa e importância do trabalho, os objetivos geral e específicos, e o sumário da dissertação.

- No Capítulo 2 é feita uma revisão bibliográfica da TRI apresentando o principal modelo usual em avaliação educacional, métricas, mudança de escala, Múltiplos Grupos, Estimação.

- No Capítulo 3 o SAEB é apresentado em detalhes.

- No Capítulo 4 explora-se o processo de equalização e possíveis fontes de variação de resultados.

- No Capítulo 5 o SAEB será reanalizado com várias edições tratadas conjuntamente.

- No Capítulo 6 serão apresentadas as considerações finais e recomendações para trabalhos futuros.

Capítulo 2

A Teoria da Resposta ao Item

2.1 Introdução

Os primeiros estudos da Teoria da Resposta ao Item (TRI) surgiram nos anos de 1950 e 1960, os principais percursores dessa moderna teoria foram: o sociólogo Lazarsfeld (1959), o psicometrista Lord (1952) e o matemático Rasch (1960). Apesar de inicialmente os modelos da TRI terem sido criados na década de 50, a teoria só começou a ser largamente utilizada no começo do ano de 1980. Na ocasião, o avanço tecnológico possibilitou aos pesquisadores solucionar os modelos matemáticos complexos da TRI, anteriormente difíceis de resolver de forma rápida e eficiente. Ressalta-se, com a ascensão da tecnologia foi viável o melhoramento dos microcomputadores, desse modo, possibilitou a criação de softwares voltados para tal metodologia (Pasquali e Primi,2003).

A referida teoria é aplicada para medir traços latentes, este traço representado matematicamente por θ, que são competências que não podem ser observados diretamente. Na área Educacional, o traço latente que se deseja mensurar é a proficiência ou habilidade do estudante (aluno).

A TRI alcançou um espaço significativo no cenário da avaliação educacional em larga escala, pois sua metodologia permite a comparação entre populações, mesmo que distintas, e ainda é possível coloca-las em uma mesma escala de conhecimento. Em particular, no SAEB a utilização de prova com itens comuns nas séries e anos avaliados, possibilita a criação de escalas de habilidades comuns entre todas as séries ao longo das edições da avaliação. Desse modo é possível realizar comparações e ainda analisar o desenvolvimento do ensino tanto entre as séries, quanto no decorrer dos anos (será melhor detalhado no Capítulo 4).

De modo geral, a Teoria da Resposta ao Item compreende um conjunto de modelos matemáticos que procuram representar a probabilidade de um indivíduo dar uma certa

resposta a um item como função dos parâmetros do item e da habilidade do respondente (Andrade et al., 2000).

Os modelos propostos pela TRI, os mesmos dependem de três importantes fatores : a) da natureza do item (dicotômicos ou não dicotômicos); b) do número de populações envolvidas (apenas uma ou mais de uma); c) da quantidade de traços latentes que está sendo medida (apenas um ou mais de um).

A seguir, não serão descritos todos os modelos propostos pela TRI, mas o principal modelo aplicado na área de Avaliação educacional em larga escala, descrito na Seção 2.2. A Seção 2.3 aborda Adaptação para Múltiplos Grupos, na Seção 2.4 apresenta estimação por Máxima Verossimilhança Marginal(MVM).

2.2 O Modelo Logístico de 3 Parâmetros

O modelo matemático da TRI mais utilizado na área de Avaliação Educacional em Larga Escala, em particular no SAEB, é o conhecido modelo logístico unidimensional de três parâmetros (ML3) para itens dicotômicos, apresentado pela expressão matemática (2.1):

$$P(U_{ji} = 1|\theta_j) = c_i + (1 - c_i)\frac{1}{1 + e^{-Da_i(\theta_j - b_i)}}, \tag{2.1}$$

com $i = 1, 2, ..., I$ e $j = 1, 2, ..., n$, sendo:

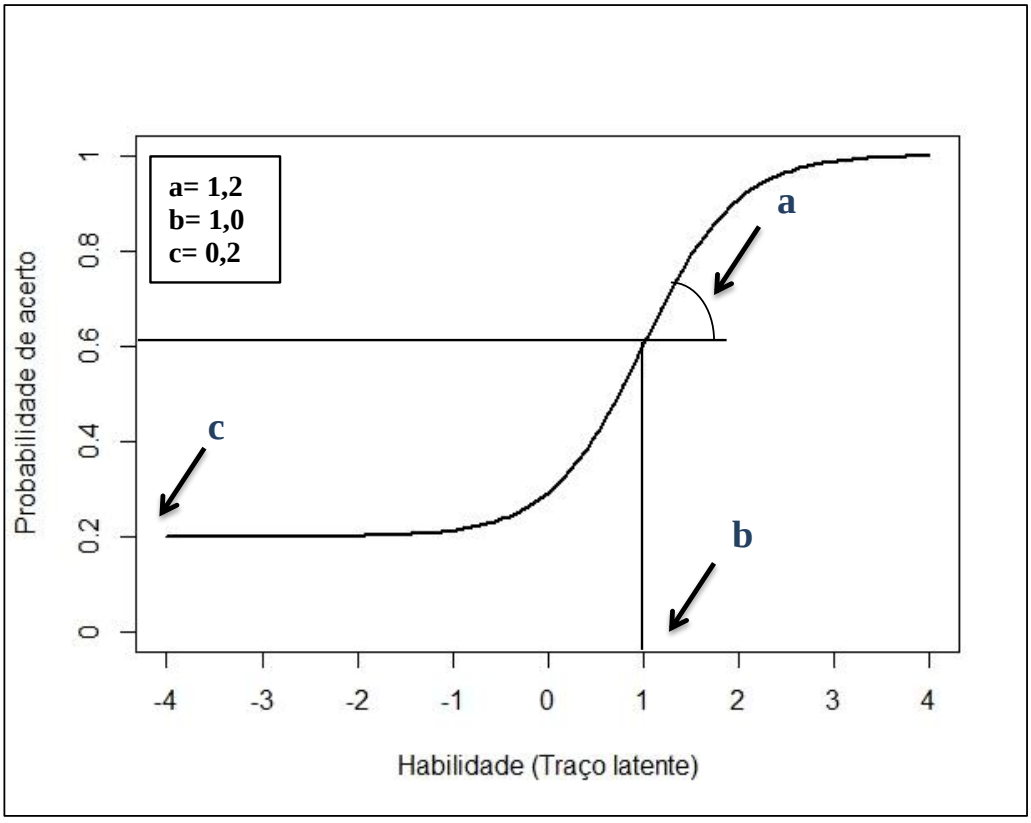
- I é o número de itens do teste ou prova;
- n é o número de indivíduos ou alunos;
- $P(U_{ij} = 1|\theta_j)$ é a probabilidade de um indivíduo j com habilidade θ_j responder corretamente o item i e é chamada de Função de Resposta do Item (FRI).
- U_{ij} é uma variável dicotômica que assume o valor 1 quando o indivíduo j acerta o item i, ou o valor zero, caso contrário;
- θ_j representa o traço latente ou a proficiência do j-ésimo indivíduo;
- a_i é o parâmetro de discriminação do item i;
- b_i é o parâmetro de dificuldade do item i;

- c_i é o parâmetro do item que representa a probabilidade de indivíduos com baixa habilidade responderem corretamente o item i;

- D é um fator de escala, constante igual a um. Emprega-se o valor 1,7 para que a função logística forneça resultado semelhante ao da função ogiva normal.

A representação gráfica da associação existente entre a Função de Resposta do item e os parâmetros do modelo, denomina-se Curva Característica do Item (CCI), esse modelo apresenta um gráfico em forma de "S"(curva sigmoide), como é ilustrado na Figura 2.1.

Figura 2.1 *Exemplo de uma Curva Característica do Item.*

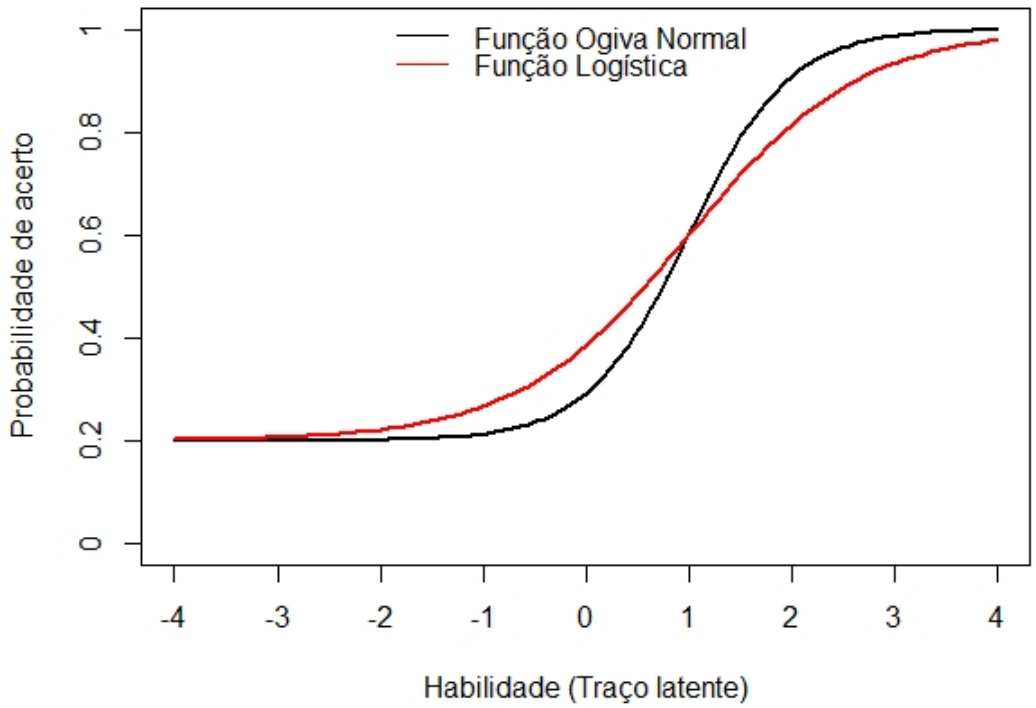


A CCI indica a probabilidade de resposta correta ao item em função do grau de habilidade do examinado. A habilidade (θ) e o parâmetro de dificuldade (b_i) estão medidos na mesma escala, esta é uma escala arbitrária. A inclinação na curva informa a capacidade de discriminação do item (parâmetro a_i) e o parâmetro de acerto casual (c_i) informa a probabilidade de um individuo com baixa proficiência acertar o item, por ser uma probabilidade seus valores estão entre 0 e 1 .

2.2.1 Métricas Normal e Logística

Os primeiros modelos matemáticos da TRI trabalhavam com a função ogiva normal, isto é, a função de distribuição acumulada da distribuição gaussiana (ver F. Lord, 1952). Em 1968, Birnbaum, possibilitou um importante avanço para TRI ao substituir a função ogiva pela função logística, a qual tem um ajuste aproximado ao da acumulada da distribuição normal (gaussiana). A principal vantagem foi evitar trabalhar o uso de integrais, desse modo, facilitando o procedimento matemático. Na Figura 2.2, é apresentado um exemplo quando o parâmetro D do ML3 (2.1) (descrito na subseção 2.2), assume valor igual a um quando se deseja resultados na métrica logística e 1,702 para métrica normal.

Figura 2.2 *Exemplo de uma comparação da Função ogiva normal com a Função logística.*



2.2.2 Mudanças de Escala

Na construção da escala de proficiência (habilidade) da TRI é permitido, teoricamente, que a mesma admita valores em toda reta real $(-\infty, +\infty)$, assim sendo, é preciso admitir para a criação da escala dois valores: (a) medida de posição: representada pela média das habilidades da população do estudo e (b) medida de dispersão: representada pelo desvio-padrão das habilidades da população do estudo.

Normalmente, os programas de análise da TRI disponibilizam os resultados na escala padrão (0,1), isto é, escala com média igual a 0 e desvio-padrão igual a 1. Porém, é comum que esta escala padrão ao final do processo de correção pela TRI seja transformada para uma escala que facilite a análise dos resultados, principalmente em resultados de avaliações educacionais. E, para facilitar a interpretação e divulgação dos resultados, na maioria das vezes, os novos valores para a escala transformada são escolhidos de forma arbitrária, exemplo, a escala do SAEB (250,50). Vale frisar que, independente da escala utilizada os resultados, obtidos serão os mesmos e ainda, a interpretação perante as duas métricas são equivalente.

A expressão a seguir apresenta o modelo matemático para realizar a transformação de escala da proficiência:

$$\theta_t = \theta_{(0,1)} \times \sigma_t + \mu_t \tag{2.2}$$

sendo,

- $\theta_{(0,1)}$ representa a habilidade na escala $(0, 1)$;
- θ_t representa a habilidade na escala desejada;
- μ_t representa a média da escala desejada;
- σ_t representa o desvio-padrão da escala desejada.

A mudança de escala do parâmetro b acontece de forma semelhante a expressão 2.2, basta mudar o elemento $\theta_{(0,1)}$ pelo $b_{(0,1)}$ na escala $(0, 1)$. Para o parâmetro a, a transformação acontece mediante a divisão do parâmetro na escala (0,1) pela medida de dispersão da nova escala deseja. O parâmetro c não sofre modificação, uma vez que trata-se de uma probabilidade.

2.3 Adaptação para Múltiplos Grupos

O modelo para Múltiplos Grupos proposto por Bock & Zimowski (1998) é uma extensão do modelo logístico unidimensional de 3 parâmetros (2.1) quando se tem mais de uma população envolvida. A implementação computacional do modelo foi uma enorme conquista para o avanço da TRI, visto que a comparação entre indivíduos de populações distintas, submetidas a testes diferentes com itens em comum, deu-se de forma mais eficiente. Segue o modelo:

$$P(U_{jik} = 1|\theta_{jk}) = c_i + (1 - c_i)\frac{1}{1 + e^{-Da_i(\theta_{jk} - b_i)}}, \tag{2.3}$$

com $i = 1, 2, \ldots, I$, $j = 1, 2, \ldots, n_k$ e $k = 1, 2, \ldots, K$, sendo:

- $P(U_{ijk} = 1|\theta_{jk})$ é a probabilidade de um indivíduo j da população k, com habilidade θ_{jk} responder corretamente o item i.
- U_{ijk} é uma variável dicotômica que assume o valor 1 quando o indivíduo j da população k acerta o item i, ou o valor zero, caso contrário;
- θ_{jk} representa o traço latente ou a proficiência do j-ésimo indivíduo da população k.

Os demais parâmetros seguem a mesma descrição conforme a equação 2.1. Duas fundamentais suposições são necessárias para o desenvolvimento do modelo, são elas: (a) as respostas vindas de indivíduos diferentes são independentes e (b) independência local.

Segundo Klein (2009), esta adaptação da TRI para várias populações ou grupos não equivalentes é realizada atribuindo distribuições a priori distintas a cada grupo. E de forma semelhante a TRI para uma única população, considera-se a distribuição a priori do grupo de referência. Sendo assim, é realizada a estimação conjuntamente dos parâmetros das distribuições a priori dos demais grupos e dos parâmetros dos itens. Desse modo, este mecanismo possibilita que realize a equalização simultânea, tanto vertical quanto horizontal.

2.4 Estimação por Máxima Verossimilhança Marginal

O processo de estimação dos parâmetros dos itens, também conhecido por *calibração*, e a estimação das proficiências dos indivíduos é uma da etapas mais importantes da Teoria da Resposta ao Item. Na teoria esse processo pode ser dividido em três situações: a) quando os parâmetros dos itens são conhecidos, e pretende apenas estimar as habilidades; b) as habilidades dos indivíduos são conhecidas e deseja-se estimar os parâmetros dos itens, e c) quando deseja-se estimar os parâmetros dos itens e as habilidades dos indivíduos simultaneamente. Nesse capítulo será adotada a situação mais comum, apresentada no item (c).

Existem várias propostas na literatura para estimação dos parâmetros dos itens e do traço latente da TRI (ver F. M. Lord (1974) e Mislevy (1986)) . Nessa secção será abordado o método de estimação por Máxima Verossimilhança Marginal (MVM). Antes de iniciar o método de MVM, conforme Andrade et al. (2000) algumas notações e suposições são necessárias para o desenvolvimento do modelo. Considera-se as seguintes notações: Seja θ_j a habilidade e U_{ji} a variável aleatória que representa a resposta do indivíduo j ao item i, com

$$U_{ji} = \begin{cases} 1, & \text{resposta correta} \\ 0, & \text{resposta incorreta} \end{cases}$$

Ainda,

- n o número total de indivíduos na amostra;
- $\boldsymbol{U}_{j.} = (U_{j1}, \ldots, U_{jI_k})$ o vetor aleatório de respostas do indivíduo j;
- $\boldsymbol{U}_{..} = (U_{1.}, U_{2.}, \ldots, U_{n.})$ o conjunto integral das respostas;
- $u_{ji}, \boldsymbol{u}_{j.}$ e $\boldsymbol{u}_{..}$ representa as respostas observadas.
- $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n$ representará o vetor de habilidades dos n indivíduos;
- $\boldsymbol{\zeta} = (\zeta_1, \ldots, \zeta_n$ o conjunto de parâmetros dos itens.;

As duas principais suposições utilizadas são:(i) as respostas oriundas de indivíduos

diferentes são independentes; (ii) os itens são respondidos de forma independente por cada indivíduo (Independência Local), fixada sua habilidade.

O método de Máxima Verossimilhança Marginal foi proposto por Bock & Lieberman em 1970, o qual propõe fazer a estimação em dois estágios:

1. Primeiro estágio: é realizada estimação dos parâmetros dos itens;

2. Segundo estágio: é realizada a estimação dos traços latentes (habilidade).

O MVM necessita inicialmente de suposições adicionais, a princípio considera-se uma distribuição de probabilidade para o traço latente (não necessariamente no sentido bayesiano). Geralmente, considera-se que as habilidades (θ_j), são relizações de uma variável θ com distribuição contínua e função densidade de probabilidade $g(\theta \mid \boldsymbol{\eta})$. De modo geral, é usual supor que θ segue uma distribuição normal com média igual a zero e desvio-padrão igual a um por consequência os parâmetros dos itens serão estimados em uma métrica (0,1).

2.4.1 Estimação dos Parâmetros dos Itens

Com as definições descritas anteriormente, tem-se que a probabilidade marginal de $\boldsymbol{U}_{j.}$ é dada por

$$P(\boldsymbol{u}_{j.} \mid \boldsymbol{\zeta}, \boldsymbol{\eta}) = \int_{\mathbb{R}} P(\boldsymbol{u}_{j.} \mid \theta, \boldsymbol{\zeta}, \boldsymbol{\eta}) g(\theta|\boldsymbol{\eta}) d\theta = \int_{\mathbb{R}} P(\boldsymbol{u}_{j.} \mid \theta, \boldsymbol{\zeta}) g(\theta|\boldsymbol{\eta}) d\theta,$$

Usando a independência entre as respostas de diferentes indivíduos (ii), podemos escrever a probabilidade associada ao vetor de respostas $\boldsymbol{U}_{..}$ como

$$P(\boldsymbol{u}_{..} \mid \boldsymbol{\zeta}, \boldsymbol{\eta}) = \prod_{j=1}^{n} P(\boldsymbol{u}_{j.} \mid \boldsymbol{\zeta}, \boldsymbol{\eta}) \tag{2.4}$$

Apesar da verossimilhança poder ser escrita como a expressão (2.4), tem -se adotado a abordagem de *Padrões de Respostas*. Dado que tem-se I itens no total, com 2 possíveis respostas para cada item, há portanto $S = 2^I$ padrões de respostas. A medida que o número de indivíduos é grande com relação ao número de itens, podem haver vantagens computacionais em trabalhar com o quantitativo de ocorrências dos diferentes padrões de resposta. Neste sentido, será considerado este raciocínio. Agora, o índice j não representará um indivíduo, mas sim um padrão de resposta.

Seja r_j o número de ocorrências distintas do padrão de resposta j, e ainda $s \leq min(n, S)$ o número de padrões de resposta com $r_j > 0$. Segue que

$$\sum_{j=1}^{s} r_j = n. \tag{2.5}$$

Pela suposição (ii), tem -se que os dados seguem uma distribuição *Multinomial*, conforme a expressão abaixo:

$$L(\boldsymbol{\zeta}, \boldsymbol{\eta}) = \frac{n!}{\prod_{j=1}^{s} r_j!} \prod_{j=1}^{s} P(\boldsymbol{u}_{j.} \mid \boldsymbol{\zeta}, \boldsymbol{\eta})]^{r_j}, \tag{2.6}$$

segue a log-verossimilhança como

$$L(\boldsymbol{\zeta}, \boldsymbol{\eta}) = \log \left\{ \frac{n!}{\prod_{j=1}^{s} r_j!} \right\} + \sum_{j=1}^{s} r_j \log P(\boldsymbol{u}_{j.} \mid \boldsymbol{\zeta}, \boldsymbol{\eta}). \tag{2.7}$$

As equações de estimação para os parâmetros dos itens são obtidas por

$$\frac{\partial \log L(\boldsymbol{\zeta}, \boldsymbol{\eta})}{\partial \boldsymbol{\zeta}_i} = 0, \quad i = 1, \ldots, I. \tag{2.8}$$

2.4.1.1 Abordagem Clássica

Segundo os desenvolvimentos descritos em Andrade et al. (2000), chega-se as seguintes equações de estimação:

$$a_i : D(1 - c_i) \sum_{j=1}^{s} r_j \int_{\Re} [(u_{ji} - P_i)(\theta - b_i)] W_i] g_j^*(\theta) d\theta = 0, \tag{2.9}$$

$$b_i : -Da_i(1 - c_i) \sum_{j=1}^{s} r_j \int_{\Re} [(u_{ji} - P_i)] W_i g_j^*(\theta) d\theta = 0, \tag{2.10}$$

$$c_i : \sum_{j=1}^{s} r_j \int_{\Re} \left[(u_{ji} - P_i) \frac{W_i}{P_i^*} \right] g_j^*(\theta) d\theta = 0, \tag{2.11}$$

onde,

$$g_j^*(\theta) = g(\theta | u_{j.}, \zeta, \eta) = \frac{P\left(u_{j.} | \theta, \zeta\right) g(\theta | \eta)}{\int_{\Re} P(u_{j.} | \theta | \eta) d\theta}. \tag{2.12}$$

A expressão (2.12) representa a função densidade de probabilidade condicional (fdp) da habilidade da população. As equações de estimação (2.9), (2.10) e (2.11) não possuem solução explícita, podendo ser solucionada por algum método numérico, exemplo o algoritmo de *Newton-Rapshson*. Também tem sido muito frequente na TRI aplicar o método *Hemite-Gauss*, conhecido como *método de quadratura gaussiana*.

2.4.1.2 Abordagem Bayesiana

Outra abordagem para a estimação dos parâmetros da TRI, além da MVM, envolve a metodologia Bayesiana. A estimação bayesiana mais utilizada foi proposta por Mislevy (1986) e denominada *Estimação Bayesiana Marginal*.

A metodologia é uma extensão do método proposto de Bock e Aitkin (1981). De modo geral, o método fundamenta-se em determinar distribuições *a priori* para os parâmetros a, b e c dos itens, construir um nova distribuição, chamada de distribuição *a posteriori* e estimar os parâmetros dos itens baseado em alguma característica dessa distribuição.

Conforme este cenário, serão apresentadas a seguir as distribuições *a priori* usuais dos parâmetros, distribuição *a posteriori* e as equações finais de estimação dos parâmetros dos itens:

- **Distribuição *a priori* do parâmetro** a_i

  Em geral, considera-se a distribuição *Log-normal* com parâmetro $\boldsymbol{\tau} = (\mu_a, \sigma_a^2)$ para cada parâmetro de discriminação, em razão de normalmente os a_i serem positivos. Portanto, segue sua densidade:

$$f(a_i \mid \mu_a, \sigma_a^2) = \frac{1}{\sqrt{2\pi} a_i \sigma_a} \exp\left[-\frac{1}{2\sigma_a^2}(\log a_i - \mu_a)^2\right]. \tag{2.13}$$

- **Distribuição *a priori* do parâmetro** b_i

  Como os parâmetros de dificuldade estão na mesma escala da habilidade, geralmente, supõem-se que cada b_i tem distribuição Normal com vetor de parâmetros $\boldsymbol{\tau} = (\mu_b, \sigma_b^2)$. Segue sua densidade:

$$f(b_i \mid \mu_a, \sigma_a^2) = \frac{1}{\sqrt{2\pi}\sigma_b} \exp\left[-\frac{1}{2\sigma_b^2}(b_i - \mu_b)^2\right]. \tag{2.14}$$

- **Distribuição *a priori* do parâmetro** c_i

  Foi proposta por Swaminathan & Gifford (1986) uma priori *Beta* para os parâmetros de acerto casual, dado que o c_i só pode pertencer ao intervalo $[0, 1]$. Segue a densidade da distribuição Beta com dois parâmetros $s + 1$ e $t + 1$:

$$f(c_i \mid s, t) = \frac{\Gamma(s + t + 2)}{\Gamma(s + 1)\Gamma(t + 1)} c_i^s (1 - c_i)^t, \tag{2.15}$$

onde $\Gamma(d)$ é a função Gama, definida por

$$\Gamma(d) = \int_0^\infty x^{d-1}e^{-x}dx.$$

A média desta distribuição é dada por

$$p = \frac{s+1}{s+t+2}$$

Os autores propõem a seguinte reparametrização:

$$\alpha = mp + 1 \quad \text{e} \quad \beta = m(1-p) + 1$$

onde $m = s + t + 2$. Desta forma, $p = (s+1)/m$ e, consequentemente, $s = mp - 1$ e $t = m - s - 2 = m(1-p) - 1$. Segue disso que

$$s = \alpha - 2 \quad \text{e} \quad t = \beta - 2.$$

Dessa forma, retornando a (2.15), temos

$$f(c_i \mid \alpha, \beta) = \frac{\Gamma(\alpha+\beta+2)}{\Gamma(\alpha+1)\Gamma(\beta+1)} c_i^{\alpha-2}(1-c_i)^{\beta-2}, \tag{2.16}$$

Nesse contexto, a média (p) da distribuição passa a ser interpretada como a probabilidade de acerto por examinado com baixa habilidade. Desta maneira, os parâmetros α e β são definidos para que p tenha o valor desejado.

De maneira geral, considera-se que a distribuição da habilidade é função de um vetor de parâmetros, representado por $\boldsymbol{\eta}$, com densidade de probabilidade $g(\theta \mid \boldsymbol{\eta})$, e que a distribuição de $\boldsymbol{\zeta}_i$, $i = 1, \ldots, I$, é função de um vetor de parâmetros $\boldsymbol{\tau}$, com densidade $f(\boldsymbol{\zeta} \mid \boldsymbol{\tau})$. Pode-se estabelecer distribuições a priori para os parâmetros $\boldsymbol{\eta}$ e $\boldsymbol{\tau}$. Com isso, a densidade conjunta desses parâmetros pode ser escrita como :

$$\begin{aligned} f(\boldsymbol{\theta}, \boldsymbol{\varsigma}, \boldsymbol{\eta}, \boldsymbol{\tau}) =& f(\boldsymbol{\varsigma} \mid \boldsymbol{\tau}) g(\boldsymbol{\theta} \mid \boldsymbol{\eta}) \\ =& \left\{ \prod_{i=1}^{I} f(\boldsymbol{\zeta_i} \mid \boldsymbol{\tau}) \right\} \left\{ \prod_{j=1}^{n} g(\theta_j \mid \boldsymbol{\eta}) \right\} f(\boldsymbol{\tau}) g(\boldsymbol{\eta}). \end{aligned} \tag{2.17}$$

Para fazer inferência sobre os parâmetros, deve-se basear na seguinte distribuição a posteriori:

$$f(\boldsymbol{\theta}, \boldsymbol{\zeta}, \boldsymbol{\tau} \mid \boldsymbol{u}..) \propto L(\boldsymbol{u}..; \boldsymbol{\theta}, \boldsymbol{\zeta}) f(\boldsymbol{\zeta} \mid \boldsymbol{\tau}) g(\boldsymbol{\theta} \mid \boldsymbol{\eta}) f(\boldsymbol{\tau}) g(\boldsymbol{\eta}) \tag{2.18}$$

Para fazer inferências com relação aos parâmetros dos itens, é adequado "marginalizar" a posteriori integrando com relação a θ e $\boldsymbol{\tau}$, obtendo a distribuição *a posteriori* de $\boldsymbol{\zeta}$ e $\boldsymbol{\eta}$:

$$\begin{aligned} f^*(\boldsymbol{\zeta}, \boldsymbol{\eta} \mid \boldsymbol{u}..) = & \quad C \int\int L(\boldsymbol{u}..; \boldsymbol{\theta}, \boldsymbol{\zeta}) f(\boldsymbol{\zeta} \mid \boldsymbol{\tau}) g(\boldsymbol{\theta} \mid \boldsymbol{\eta}) f(\boldsymbol{\tau}) g(\boldsymbol{\eta}) d\theta d\boldsymbol{\tau} \\ = & \quad C g(\boldsymbol{\eta}) \left\{ \int f(\boldsymbol{\zeta} \mid \boldsymbol{\tau}) f(\boldsymbol{\tau}) d\boldsymbol{\tau} \right\} \left\{ \int L(\boldsymbol{u}..; \boldsymbol{\theta}, \boldsymbol{\zeta}) g(\boldsymbol{\theta} \mid \boldsymbol{\eta}) d\theta \right\} \\ & \propto L(\boldsymbol{\zeta}, \boldsymbol{\eta}) f(\boldsymbol{\zeta}) g(\boldsymbol{\eta}), \end{aligned} \tag{2.19}$$

sendo C uma constante, $L(\boldsymbol{\zeta}, \boldsymbol{\eta}) \equiv P(\boldsymbol{u}..; \boldsymbol{\theta}, \boldsymbol{\zeta})$ e $f(\boldsymbol{\zeta}) = \int f(\boldsymbol{\zeta} \mid \boldsymbol{\tau}) f(\boldsymbol{\tau}) d\boldsymbol{\tau}$.

Como estimador de $\boldsymbol{\zeta}$ pode-se escolher alguma característica de $f^*(\boldsymbol{\zeta}, \boldsymbol{\eta} \mid \boldsymbol{u}..)$, as mais sugeridas são a média e a moda. Aqui, consideramos a *moda a posteriori* como estimador. Tem-se que:

$$\log f^*(\boldsymbol{\zeta}, \boldsymbol{\eta} \mid \boldsymbol{u}..) = C + \log L(\boldsymbol{\zeta}, \boldsymbol{\eta}) + \log f(\boldsymbol{\zeta}) + \log g(\boldsymbol{\eta})$$

,

Chega-se as equações de estimação dos parâmetros dos itens $\boldsymbol{\zeta}_i$, são dadas por

$$\frac{\partial f^*(\boldsymbol{\zeta}, \boldsymbol{\eta} \mid \boldsymbol{u}..)}{\partial \boldsymbol{\zeta}_i} = \frac{\partial \log L(\boldsymbol{\zeta}, \boldsymbol{\eta})}{\partial \boldsymbol{\zeta}_i} + \frac{\partial \log f(\boldsymbol{\zeta})}{\partial \boldsymbol{\zeta}_i} = 0. \tag{2.20}$$

As equações finais de estimação dos parâmetros dos itens podem ser escritas como:

$$a_i : D(1 - c_i) \int_{\mathbb{R}} (\theta - b_i)[r_i(\theta) - P_i f_{i(\theta)}] W_i d\theta - \frac{1}{a_i} \left[1 + \frac{\log a_i - \mu_a}{\sigma_a^2} \right] = 0, \tag{2.21}$$

$$b_i : - D a_i (1 - c_i) \int_{\mathbb{R}} [r_i(\theta) - P_i f_{i(\theta)}] W_i d\theta - \frac{(b_i - \mu_b)}{\sigma_b^2} = 0, \tag{2.22}$$

$$c_i : \int_{\mathbb{R}} [r_i(\theta) - P_i f_{i(\theta)}] \frac{W_i}{P_i^*} d\theta + \frac{\alpha - 2}{c_i} - \frac{\beta - 2}{1 - c_i} = 0. \tag{2.23}$$

As equações de estimação (2.21), (2.22) e (2.23) não possuem solução explícita.

2.4.2 Estimação das Proficiências

Nesta seção apresenta-se brevemente diferentes métodos de estimação das proficiências dos indivíduos (examinados), considerando os parâmetros dos itens calibrados.

- Estimação por Máxima Verossimilhança (MV)

Conforme a suposição (i) e (ii), pode-se escrever a log-verossimilhança como função de θ, como:

$$\log L(\boldsymbol{\theta}) = \sum_{j=1}^{n}\sum_{i=1}^{I} u_{ji}\log P_{ji} + (1-u_{ji})\log Q_{ji} \tag{2.24}$$

O valor que maximiza a verossimilhança, da próxima equação, é o EMV de θ_j:

$$\frac{\partial \log L(\boldsymbol{\theta})}{\partial \theta_j} = 0 \tag{2.25}$$

Segue então que a equação de estimação de (2.25) para θ_j é :

$$\theta_j : D\sum_{i=1}^{I} a_i(1-c_i)(u_{ji}-P_{ji})W_{ji} = 0. \tag{2.26}$$

Mais uma vez, esta equação de estimação não apresenta solução analítica, por isso, necessita de algum método interativo para solução.

- Estimação por Expected a Posteriori (EAP):

A estimação de θ_j pela média da posteriori $g_j^*(\theta)$ é um método Bayesiano, consiste em obter a esperança da posteriori, dada por:

$$\hat{\theta}_j \equiv E\left(\theta|\boldsymbol{u}_{j.},\boldsymbol{\zeta},\boldsymbol{\eta}\right) = \frac{\int_{\mathbb{R}} \theta g\left(\theta \mid \boldsymbol{\eta}\right) P(\boldsymbol{u}_{j.} \mid \theta,\boldsymbol{\zeta}) d\theta}{\int_{\mathbb{R}} \theta g\left(\theta \mid \boldsymbol{\eta}\right) P(\boldsymbol{u}_{j.} \mid \theta,\boldsymbol{\zeta}) d\theta} \tag{2.27}$$

Este método de estimação da habilidade tem a vantagem de não precisar de nenhum método interativo para a solução, pois pode ser calculada diretamente. Alguns autores (por exemplo, Mislevy e Stocking (1989)) recomendam esta escolha para a estimação das proficiências.

- Estimação por Máximo da Posteriori (MAP)

O MAP (maximum a posteriori), assim como o EAP, é um método bayesiano. Realiza a estimação pela moda da posteriori, consiste em obter o máximo de

$$g_j^*(\theta_j) \equiv g(\theta_j \mid \boldsymbol{u}_{j.}, \boldsymbol{\zeta}, \boldsymbol{\eta}) \propto P(\boldsymbol{u}_{j.} \mid \theta_j, \boldsymbol{\zeta}) g(\theta_j \mid \boldsymbol{\eta}). \tag{2.28}$$

Por facilidade, emprega-se o logaritmo da posteriori, conforme a expressão abaixo:

$$\log g_j^*(\theta_j) = Const + \log P(\boldsymbol{u}_{j.} \mid \theta_j, \boldsymbol{\zeta}) + \log g(\theta_j \mid \boldsymbol{\eta}). \tag{2.29}$$

Após esta etapa técnica, o próximo capítulo apresentará em detalhes a estrutura do SAEB desde a sua criação.

Capítulo 3

Sistema de Avaliação da Educação Básica

3.1 Introdução

Nesta seção será realizada uma revisão histórica acerca do Sistema de Avaliação da Educação Básica (SAEB), as principais informações apresentadas foram baseadas nas informações divulgadas no site do INEP*, em formato de relatórios e arquivos encontrados nos microdados do SAEB.

Na década de 1990 foi fundado o Sistema de Avaliação da Educação Básica (SAEB) com finalidade de avaliar o desempenho da Educação Básica no Brasil, inicialmente nas escolas públicas e no decorrer das edições também nas escolas privadas, no território urbano e rural. A avaliação é conduzida pelo INEP e realizada a cada dois anos. A partir das informações obtidas pela avaliação é possível contribuir com a formulação, reformulação e o monitoramento das políticas públicas nas esferas municipal, estadual e federal, auxiliando a melhoria da qualidade, equidade e eficiência do ensino.

Além de analisar o desempenho dos alunos, por meio de provas, a avaliação também verifica prováveis fatores socieconômicos e contextuais que podem influenciar o desenvolvimento escolar do estudante, os quais encontram-se agrupados em quatro áreas de análise: escola, gestão escolar, professor e aluno.

Com uma amostra das escolas públicas da rede urbana que ofertavam as 1ª, 3ª, 5ª e 7ª séries do Ensino Fundamental no ano de 1990, ocorreu a primeira aplicação da prova do SAEB. Os alunos foram avaliados nas disciplinas de Língua Portuguesa, Matemática e Ciências, porém os estudantes da 5ª e 7ª séries também foram avaliados em redação. Esse padrão se manteve na edição posterior, que ocorreu no ano de 1993.

* http://portal.inep.gov.br/educacao-basica/saeb/historico-do-saeb

Na edição de 1995 foi introduzida uma moderna metodologia de construção do teste e análise de resultados: a Teoria da Resposta ao Item (TRI), o que proporcionou a comparabilidade entre os resultados das avaliações ao longo do tempo. Nesse mesmo ano também foi reformulado o público alvo, o qual se limitou às etapas finais dos ciclos de escolarização: 4ª e 8ª séries do Ensino Fundamental (atualmente correspondem ao 5º e 9º anos do ensino fundamental) e 3º ano do Ensino Médio. Ainda nesse mesmo ano, além das amostras das escolas públicas, também foi incluída uma amostra das escolas particulares.

Nas edições seguintes, de 1997 e 1999, os alunos matriculados no 5º e 9º ano do ensino fundamental foram avaliados em Língua Portuguesa, Matemática e Ciências e os estudantes de 3º ano do Ensino Médio em Língua Portuguesa, Matemática, Ciéncias, História e Geografia.

Vale frisar que o SAEB, a partir da edição de 2001, passou a avaliar somente nas áreas de Língua Portuguesa e Matemática, e no ano de 2005 o exame passou por uma reformulação, no qual sua estrutura ficou composta por duas avaliações: Avaliação Nacional da Educação Básica (Aneb) e Avaliação Nacional do Rendimento Escolar (Anresc), conhecida também por Prova Brasil.

Em 2013 foi inserida outra avaliação no sistema SAEB, chamada de Avaliação Nacional da Alfabetização (ANA). A partir desse momento a avaliação SAEB, passou a ser composta por três avaliações em larga escala: Aneb, Anresc e ANA. Nesse mesmo ano a edição da prova teve outra alteração, de forma experimental, a disciplina de Ciências foi aplicada para estudantes do 9º ano do Ensino Fundamental e do 3º ano do Ensino Médio.

Na edição de 2017 também tiveram algumas modificações, as principais foram: o aumento do grupo de alunos, turmas e escolas avaliadas, a adesão de escolas particulares interessadas em verificar seu desempenho e o estrato do Ensino Médio (EM) passou a ser censitário. A partir de 2019, as siglas ANA, Aneb e Anresc serão retiradas, e todas as avaliações passarão a ser identificadas apenas pela sigla SAEB.

Portanto, espera-se que a cada edição do SAEB os seus resultados cheguem aos gestores, pesquisadores, instituições e interessados na área da educação e possibilite a realização de diagnósticos, estudos e pesquisas que subsidiem o planejamento e a proposição de ações no âmbito escolar e das redes de ensino (INEP microdados, 2015).

3.2 Estrutura do Saeb

Segundo Conde e Laros (2007), a prova SAEB é estruturada conforme o método baseado na amostragem matricial de itens, a qual utiliza o esquema de montagem e aplicação de testes por Blocos Incompletos Balanceados (BIB). Em um contexto geral Bekman (2001) apresenta o BIB como sendo:

> [..]um esquema otimizado para o rodízio de blocos com aplicações em diversas áreas, inclusive educação e agricultura. A necessidade do rodízio se justifica se pressupusermos que possuímos b blocos e só podemos utilizar k deles em cada conjunto. Isto é especialmente útil nos sistemas de avaliação quando desejamos obter informações amplas sobre o ensino, utilizando um grande número de itens, ao passo que precisamos limitar a quantidade de itens submetido a cada aluno num valor aceitável e adequado ao tempo de prova (p.121).

Conforme esse esquema, especificamente no SAEB, são montados 7 blocos para cada uma das áreas do conhecimento. Na avaliação do 5º ano do EF cada bloco contém 11 itens, totalizando 77 itens. Com relação ao 9º ano do EF e 3º ano do EM, cada bloco engloba 13 itens, totalizando 91 itens. De modo geral, são construídos 21 cadernos a partir da combinação de dois blocos de LP e MT, de acordo com a matriz do sistema BIB, ilustrado na Figura 3.1.

Os alunos respondem, cada um, apenas a um único caderno de prova, cada caderno contém itens de Língua Portuguesa e Matemática. Os cadernos do o 5º ano EF possuem 22 itens (11 itens de LP e 11 itens MT), já os cadernos do 9º ano do EF e 3º ano do EM contém 26 itens (13 itens de LP e 13 itens MT).

Os itens que compõem a prova SAEB estão baseados na matriz de referência, que é composta por um conjunto de conteúdos e competências. A matriz compreende as habilidades e processos cognitivos julgados necessários a serem examinados para cada área e série/ano.

É importante frisar que esta matriz não abrange todo currículo escolar. Visto que constitui-se de um recorte do currículo escolar, com conteúdos definidos mais importantes para cada etapa e disciplina, de modo que ficasse um conteúdo em comum a todo território brasileiro. Conforme pesquisas realizadas pelo INEP e profissionais de órgãos educacionais, essas matrizes representam as competências esperadas a serem adquiridas pelos estudantes no período final do 5º, 9º ano do ensino fundamental e 3º ano do ensino médio.

Figura 3.1 *Rodízio de blocos utilizado para composição dos cadernos de teste do Saeb.*

Caderno	Disciplina 1	Blocos		Disciplina 2	Blocos	
		Posição 1	Posição 2		Posição 1	Posição 2
1	Língua Portuguesa	LP1	LP2	Matemática	M1	M2
2	Matemática	M2	M3	Língua Portuguesa	LP2	LP3
3	Língua Portuguesa	LP3	LP4	Matemática	M3	M4
4	Matemática	M4	M5	Língua Portuguesa	LP4	LP5
5	Língua Portuguesa	LP5	LP6	Matemática	M5	M6
6	Matemática	M6	7	Língua Portuguesa	LP6	LP7
7	Língua Portuguesa	LP7	LP1	Matemática	M7	M1
8	Matemática	M1	M3	Língua Portuguesa	LP1	LP3
9	Língua Portuguesa	LP2	LP4	Matemática	M2	M4
10	Matemática	M3	M5	Língua Portuguesa	LP3	LP5
11	Língua Portuguesa	LP4	LP6	Matemática	M4	M6
12	Matemática	M5	M7	Língua Portuguesa	LP5	LP7
13	Língua Portuguesa	LP6	LP1	Matemática	M6	M1
14	Matemática	M7	M2	Língua Portuguesa	LP7	LP2
15	Língua Portuguesa	LP1	LP4	Matemática	M1	M4
16	Matemática	M2	M5	Língua Portuguesa	LP2	LP5
17	Língua Portuguesa	LP3	LP6	Matemática	M3	M6
18	Matemática	M4	M7	Língua Portuguesa	LP4	LP7
19	Língua Portuguesa	LP5	LP1	Matemática	M5	M1
20	Matemática	M6	M2	Língua Portuguesa	LP6	LP2
21	Língua Portuguesa	LP7	LP3	Matemática	M7	M3

Fonte: Daeb/Inep. Banco Nacional de Itens.

3.3 O SAEB como avaliação amostral

O método de amostragem usaso no SAEB, de modo geral, possibilita a obtenção de estimativas de desempenho (habilidade) dos estudantes por série/ano, a nível de Federação, Regiões e País, em todas as disciplinas examinadas. O mesmo se desenha por amostras probabilísticas de estudantes e de amostras de escolas, turmas, professores, diretores, levando em conta o conjunto de estudantes devidamente matriculados no sistema educacional do país. As amostras são estratificadas, considerando os seguintes fatores: escolas por zona, localização e rede de ensino. Seu plano amostral dá-se em três importantes etapas:

1. Seleção de municípios;
2. Seleção de escolas;
3. Seleção de turmas.

Para todos estas etapas a escolha é feita em função da proporção de alunos matriculados. Assim sendo, o conjunto de alunos da turma escolhida participam do processo de avaliação e seus respectivos professores e diretores estão imediatamente participando da avaliação, através dos questionários.

Para finalizar, a escolha das etapas da amostra é aleatória e probabilística, o qual possibilita associar os resultados da amostra com características da população referência (ou população de pesquisa).

3.4 Mudanças no percurso

O Sistema de Avaliação da Educação Básica criado em 1990 correspondia a uma única avaliação, e foi totalmente amostral até 2003. No decorrer das edições houve reformulações em sua estrutura, entre elas, mais precisamente no ano de 2005, esse sistema passou a ser composto por duas avaliações, denominadas de Aneb e Anresc (conhecida como Prova Brasil).

3.4.1 A Prova Brasil e o Saeb

A estrutura da Aneb manteve as mesmas características (forma amostral, escolas e alunos), objetivos principais (avaliar qualidade, equidade e eficiência do ensino) e técnica de avaliação da educação básica do SAEB. Os resultados apresentados pela Aneb envolvem três estratos de interesse: dependência administrativa, localização (urbana e rural) e área (Capital e interior). Também contém os resultados dos fatores socieconômicos e contextuais associados ao desempenho do aluno. Os resultados são representativos do País, das regiões e dos estados.

A Anresc (Prova Brasil) utiliza os mesmos instrumentos da Aneb e é aplicado com a mesma periodicidade, mas tem uma organização diferente, é uma avaliação censitária, envolvendo apenas estudantes da etapa inicial (5º ano) e da etapa final (9º ano) do ensino fundamental das escolas públicas brasileiras que tenham no mínimo 20 alunos matriculados nos anos de interesse. A sua principal finalidade é avaliar a qualidade de ensino das escolas públicas do país, e também apresentar resultados dos fatores socieconômicos e contextuais associados ao desempenho do estudante, para cada unidade escolar participante.

3.4.2 Inclusão da escolas rurais

Na edição de 2007 da Prova Brasil (Anresc) passaram a participar as escolas públicas rurais que ofertavam os níveis iniciais do Ensino Fundamental,isto é, a 4ª série (5º ano), e que tinham um mínimo de 20 estudantes matriculados nessa série. A inclusão dos anos finais do ensino fundamental (8ª série/9º ano) de escolas públicas rurais, ocorreu no ano de 2009, e o critério de avaliação também foi referido ao mínimo de 20 alunos matriculados nessa série.

3.5 A Escala do Saeb

Antes de falarmos da reconhecida escala SAEB, cabe mencionar mais uma vez, que os primeiros resultados do sistema de avaliação ocorreram mediante a teoria clássica dos Testes (TCT). Tais resultados eram relacionados ao cálculo do número e percentual de acertos dos alunos. Por mais que os resultados produzidos por tal método colaborem na identificação das dificuldades apresentadas pelos estudantes em relação aos itens (questões) da prova, não é possível realizar comparações, algo importante na área de avaliação em larga escala, como exemplo, comparação entre estudantes que realizam provas de diferentes Regiões e níveis escolar.

A escala SAEB de proficiência foi estabelecida em 1997, a partir da implementação da importante metodologia estatística, Teoria da Resposta ao Item (TRI), no ano de 1995. Este método, diferente da TCT, permite fazer comparações entre estudantes de níveis distintos e ao longo das edições.

A conhecida escala é a mesma para todos os ciclos escolares e disciplinas (Língua Portuguesa e Matemática) avaliadas. Refere-se a uma escala de proficiência estabelecida de forma arbitraria, para o 9º ano do ensino fundamental, com média 250 e desvio padrão 50.

Neste contexto, uma escala de habilidade gerada pela TRI posiciona os itens conforme seu grau de dificuldade, isto é, a partir do conjunto de respostas dos alunos de diferentes graus de proficiência (habilidade) são calibrados os parâmetros dos itens por meio de critérios probabilísticos. Assim, as proficiências dos estudantes são estimadas nessa mesma métrica (escala), possibilitando a comparação entre os candidatos, níveis de escolarização e anos de edições.

Na prática, tem -se um escala organizada por um conjunto de pontos, com a finalidade de ter a métrica interpretável pedagogicamente, que justifique o posicionamento da proficiência, de maneira a indicar as capacidades de conhecimento e habilidade do estudante. Portanto, considera-se uma separação por intervalos de 25 pontos na escala, denominados de níveis de proficiência, dispondo os itens em grupos cujas características possibilitem serem integrados naquele intervalo.

As Tabelas 3.1 e 3.2 apresentam, respectivamente, as escalas de Língua Portuguesa e Matemática para cada ciclo escolar (5º ano e 9º ano do ensino fundamental e 3º ano do ensino médio) por nível.

Tabela 3.1 *Escala de Língua Portuguesa para 5º ano e 9º ano do ensino fundamental e 3º ano do ensino médio*

Nível	**Ensino Fundamental**		**Ensino Médio**
	5º Ano	**9º Ano**	**3º Ano**
Até o nível 1	0 - 149 pontos	-	-
Nível 1	-	200 - 224 pontos	225 - 249 pontos
Nível 2	150 - 174 pontos	225 - 249 pontos	250 - 274 pontos
Nível 3	175- 199 pontos	250 - 274 pontos	275 - 299 pontos
Nível 4	200 - 224 pontos	275 - 299 pontos	300 - 324 pontos
Nível 5	225 - 249 pontos	300 - 324 pontos	325 - 349 pontos
Nível 6	250 - 274 pontos	325 - 349 pontos	350 - 374 pontos
Nível 7	275 - 299 pontos	350 - 374 pontos	375 - 399 pontos
Nível 8	300 - 324 pontos	375 - 400 pontos	400 - 425 pontos
Nível 9	325 - 350 pontos	-	-

Tabela 3.2 *Escala de Matemática para 5º ano e 9º ano do ensino fundamental e 3º ano do ensino médio*

Nível	**Ensino Fundamental**		**Ensino Médio**
	5º Ano	**9º Ano**	**3º Ano**
Nível 1	125 - 149 pontos	200 - 224 pontos	225 - 249 pontos
Nível 2	150 - 174 pontos	225 - 249 pontos	250 - 274 pontos
Nível 3	175- 199 pontos	250 - 274 pontos	275 - 299 pontos
Nível 4	200 - 224 pontos	275 - 299 pontos	300 - 324 pontos
Nível 5	225 - 249 pontos	300 - 324 pontos	325 - 349 pontos
Nível 6	250 - 274 pontos	325 - 349 pontos	350 - 374 pontos
Nível 7	275 - 299 pontos	350 - 374 pontos	375 - 399 pontos
Nível 8	300 - 324 pontos	375 - 400 pontos	400 - 425 pontos
Nível 9	325 - 350 pontos	400 - 425 pontos	425 - 449 pontos
Nível 10	-	-	450 - 475 pontos

Capítulo 4

O Processo de Análise e Equalização

4.1 Introdução

Entre as principais vantagens do uso da Teoria da Resposta ao Item em avaliações educacionais em larga escala, destaca-se o processo de equalização. Este processo possibilita que habilidades de indivíduos que realizam provas parcialmente distintas sejam colocadas numa mesma escala de conhecimento, tornando-as comparáveis. Além disso, para avaliações realizadas periodicamente, torna-se possível comparar os resultados das provas ao longo do tempo.

De forma geral, o significado de equalizar diz respeito à igualar/uniformizar, segundo Andrade, Tavares e Valle (2000) a sua importância na TRI é de colocar parâmetros dos itens ou proficiências dos indivíduos de grupos distintos na mesma métrica, ou seja, posicionar em uma mesma escala comum, tornando itens e/ou proficiências comparáveis. Ainda, conforme os autores existem duas formas de posicionar os parâmetros, tanto dos itens como das habilidades na mesma métrica no processo de equalização:

1. *Via População*: tem-se apenas um grupo de respondentes submetidos a testes diferentes, para certificar que todos estarão na mesma métrica basta realizar a calibração de todos os itens conjuntamente;

2. *Via Itens*: basta ter duas ou mais populações realizando testes com itens parcialmente distintos; a ligação entre as populações será dada por itens comuns entre os testes, desse modo garantindo que as mesmas terão seus parâmetros em uma única escala.

É comum nas avaliações em larga escala, haver mais de uma população ou grupo envolvido no teste, nesse contexto, normalmente estas populações são caracterizadas em diferentes níveis de escolaridade, exemplo a prova SAEB. Para que os resultados referentes às populações distintas possam ser comparáveis se faz necessário formar um sistema de

relação entre as mesmas. E na maioria das avaliações esse sistema de ligação ocorre por intermédio do processo de equalização *Via Itens*.

Particularmente, na prova SAEB, a equalização é realizada de duas formas: (1) mediante itens comuns entre anos e suas respectivas séries da mesma disciplina, isto é, a edição do ano atual faz uso de um quantitativo de itens (questões) da prova do ano anteriormente aplicada, para cada série (ano) do ensino fundamental e médio da área do conhecimento que planeja ser examinada, esse processo é conhecido como equalização *horizontal*; (2) e pela equalização *vertical*, realizada através de itens comuns entre as séries (5ºano e 9º ano ensino fundamental e o 3º ano do ensino médio) da mesma disciplina do ano atual da realização da prova.

Tanto a equalização *horizontal* quanto a equalização *vertical* ocorrem simultaneamente mediante o modelo de múltiplos grupos (exposto no capitulo 2), sendo que os grupos seriam as séries avaliadas na prova.

4.2 O Ano 1997 como Referência

Na edição do ano de 1997 foi definida a atual escala SAEB, média 250 e desvio padrão 50 da distribuição das habilidades. Segundo Klein (2009), a métrica foi estabelecida na estimação conjunta dos itens de todas as séries das avaliações SAEB 95 e 97, fixando a 8ª série (9º ano) de 97 como grupo referência, excluindo os itens com problemas de mais de uma graduação, isto é, além do certo e errado.

4.3 Procedimentos adotados nos anos posteriores

Para garantir que a escala definida na edição/97 fosse mantida em todas as edições posteriores, ocorreu o processo de equalização horizontal entre as edições e vertical entre as séries por meio de itens comuns. Conforme Klein (2009), no SAEB de 1999 usou-se a equalização horizontal para a mesma disciplina, e a obtenção dos parâmetros na mesma escala estabelecida ocorreu simultaneamente com a equalização vertical, incluindo na calibração uma base de dados da edição de 1997 com os parâmetros dos itens dessa edição considerados conhecidos. Esse procedimento, teoricamente, deve garantir que os itens inéditos sejam calibrados na escala SAEB estabelecida, assim, as habilidades estimadas também estarão na mesma escala. Dessa maneira, ocorreu/ocorre para as demais edições.

4.4 O software BILOG-MG e suas versões *DOS* e *Windows*

Entre os programas computacionais comerciais criados para o desenvolvimento dos modelos da TRI, o mais utilizado atualmente no País é denominado de BILOG-MG (Bock & Zimowski, 1998). O software teve sua primeira versão para o sistema operacional *DOS*, e depois de um período foi criada uma versão disponibilizada para o sistema *Windows*. As duas versões permitem as análises via TRI para itens dicotômicos ou dicotomizados, em ambos são implementados os modelos unidimensionais logísticos de um, dois e três parâmetros, também são disponibilizados os gráficos da curva característica e informação do item e o gráfico de informação do teste. Além de ter implementado o modelo para múltiplos grupos. O BILOG-MG nas suas duas versões desempenha o processo de análise dividida em três fases, de forma resumida são elas:

1. Primeira fase: Realiza a leitura dos dados de entrada, os principais são: o vetor de resposta dos indivíduos dicotomizados ou não dicotomizados e o gabarito quando for necessário. Nesta fase da execução é possível verificar se a leitura dos dados foi realizada de forma correta, e também gera importantes estatísticas da Teoria Clássica dos Testes (TCT), como: percentual de acerto a cada item, correlação bisserial e ponto bisserial. Essas primeiras estatísticas são de grande importância para uma investigação preliminar dos itens, que podem apresentar algum tipo de problema, como erro de gabarito.

2. Segunda fase: ocorre a estimação ou calibração dos parâmetros dos itens e seus erros-padrão. Os métodos de estimação dos parâmetros que são implementados nesse aplicativo são: o método da máxima verossimilhança marginal e o método bayesiano de estimação por maximização da distribuição marginal a posteriori. O aplicativo utiliza duas formas de resolver as equações de máxima verossimilhança marginal, o algoritmo EM ou pelo método de "Scoring" de Fisher.

3. Terceira fase: Acontece a estimação das habilidades dos indivíduos. O programa tem implementados alguns métodos de estimação, como estimação por máxima verossimilhança, estimação por esperança a posteriori (EAP) e estimação por máximo a posteriori (MAP).

4.5 A sintaxe padrão adotada

Nessa seção será descrita a sintaxe do programa BILOG-MG adotada para o processo de equalização. As duas primeiras linhas da Sintaxe são reservadas para um título geral e podem ser deixadas em branco caso não haja necessidade de um título. A seguir as informações da sintaxe:

(1) O primeiro comando, COMMENTS, é opcional, nele pode inserir uma ou mais linhas de observações.

(2) No segundo comando, GLOBAL, é possível fornecer nomes de arquivos de entrada e outras informações usadas nas três fases do programa, como: o comando DFNAME usado para fornecer o arquivo de dados; NPARM=3, indica o número de parâmetros do item no modelo, nesse caso, é adotado o Modelo Logístico de 3 parâmetros;NWGHT=3, é usada para ponderar os registros de resposta, o número três indica que a ponderação está associada à estatísticas e calibrações; PRNAME, usado para fornecer ao programa o arquivo dos parâmetros dos itens que serão fixados; NTEST é usado para indicar o número de subtestes, o Default é igual a um;

(3) O terceiro comando, SAVE, deve seguir o comando GLOBAL para especificar os arquivos a serem salvos, como: os arquivos de estimativas dos parâmetros dos itens, com extensão .PAR e das habilidades, com extensão .SCO;

(4) O quarto comando, LENGTH, é usado para fornecer o número de itens. O NITEMS fornece o número de itens no teste a serem analisados.

(5) O quinto comando, INPUT, tem como objetivo descrever o arquivo de dados brutos, com instruções de formato da variável, descrevendo o layout dos dados. Essas instruções são: NTOTAL, indica o número total de itens; NFMT=1, especifica o número de registros de formato para ler dados do respondente, default é igual a 1; TYPE=2, especifica o tipo de arquivo de dados a ser usado na análise, igual a dois indica dados único com pesos de caso; NALT, especifica o número máximo de alternativas de resposta nos dados brutos; NIDCHAR, especifica o número de caracteres no campo de identificação do respondente; NFORM, especifica o número de formulários (cadernos) de teste; NGROUP, especifica o número de grupos (séries) de respondentes; KFNAME, especifica o nome do arquivo que contém o gabarito do teste.

(6) O sexto comando, ITEMS, é usado para fornecer nomes e números correspondentes para todos os itens nos registros de dados. O INUMBERS especifica uma lista de números

exclusivos fornecidos no NTOTAL; INAMES, especifica uma lista do nome dos itens.
(7) O sétimo comando, TEST, identifica os principais itens de teste. TNAME, fornece um nome para o teste; INUMBERS fornece uma lista de números, conforme especificado no comando ITEMS; FIX especifica se os parâmetros de itens específicos estão livres para serem estimados ou se devem ser fixados em seus valores iniciais.
(8) O oitavo comando, FORM, fornece a ordem das respostas dos itens nos dados. Para cada caderno é especificado um FORM. O LENGTH indica a quantidade de itens em cada FORM; INUMBERS, fornece a lista do números dos itens, conforme especificado no comando ITEMS, na ordem em que a resposta aparece nos dados para cada caderno.
(9) O nono comando, GROUP, especifica as informações sobre os itens em cada grupo (série). O GNAME especifica o nome do grupo; O INUMBERS, fornece uma lista de números de itens, conforme especificado no comando ITEMS, para todos os itens em todos os formulários administrados ao grupo.
(10) O décimo comando controla o procedimento de estimação dos parâmetros dos itens e as especificações das distribuições a priori dos parâmetros do item. O NQPT especifica o número dos pontos de quadratura na estimativa de MVM para cada grupo; CYCLES, define o número máximo de ciclos do algoritmo EM; O NEWTON, especifica o número de iterações de Gauss-Newton após os ciclos de EM.

Normalmente o processo é conduzido em 3 etapas: Individual, Equalização e Scoring.

A seguir apresenta-se a sintaxe padrão para a etapa de equalização entre duas edições do SAEB, genericamente representados por XX e YY para o conjunto de 6 grupos e 126 cadernos (FORMs), pois cada um dos níveis tem 21 cadernos. O número de itens no 5EF no é 22, enquanto para o 9EF e 3EM é 26.

```
SAEB 20XX-20YY: EQUALIZAÇÃO VETICAL E HORIZONTAL: ANOS 5(EF), 9(EF) e 12 (3EM)
AREA: MATEMÁTICA
>COMMENTS
   Grupo 1 = 5o ano EF 20XX (77 itens), 21 cadernos (22 itens);
    Grupo 2 = 9o ano EF 20XX (91 itens, sendo 21 itens em comum
com 5o ano EF 20XX), 21 cadernos (26 itens);
    Grupo 3 = 3o ano EM 20XX (91 itens,sendo 21 itens em comum
com 9o ano EF 20XX), 21 cadernos (26 itens);
    Grupo 4 = 5o ano EF 20YY (77 itens,sendo 21 itens em comum
```

```
com 5o ano EF 20XX), 21 cadernos (22 itens);
    Grupo 5 = 9o ano EF 20YY (91 itens,sendo 21 itens em comum
com 9o ano EF 20XX e 21 itens em comum com 5o ano EF 20YY),
 21 cadernos (26 itens);
    Grupo 6 = 3o ano EM 20YY (91 itens, sendo 21 itens em comum
com 3o ano EM 20XX e 21 itens em comum com 9o ano EF 20YY),
21 cadernos (26 itens).
>GLOBAL  DFNAME='MATXXYY.txt',NPARM=3,PRNAME='PARFIX.PRM', NSUBJECT=999999,NWGHT=3,
>SAVE  PARM='MATXXYY.PAR', SCORE='MATXXYY.SCO',
       EXPECTED='MATXXYY.EXP';
>LENGTH  NITEMS=nnn;
>INPUT   NTOTAL=nnn,NFMT=1,TYPE=2,SAMPLING=999999,
         NALT=5,NIDCHAR=nn, NFORM=nnn, NGROUP=6,
         KFNAME='KFMP.TXT', NFNAME='NFMP.TXT';
>ITEMS  INUMBERS=(1(1)nnn),
        INAMES=();
>TEST   TNAME='SAEXXYYM',INUMBERS=(1(1)nnn),
FIX = (1(0)nnn,0(0)nnn);;
>FORM001  LENGTH= 22, INUMBERS=();
>FORM002  LENGTH= 22, INUMBERS=();
...
>FORM126  LENGTH= 26, INUMBERS=();
>GROUP1 GNAME='A05-XX', LENGTH=nnn, INUMBERS=();
...
>GROUP6 GNAME='A12-YY', LENGTH=nnn, INUMBERS=();
(19A1,7X,I3,1X,I1,1X,F12.6,23X,26A1)
>CALIB  NQPT=40,CYCLES=50,NEWTON=0,CRIT=0.01,IDIST=0,
   DIAGNOSIS=1,REFERENCE=2, NORMAL,TPRIOR,SPRIOR,GPRIOR,READPRI,NOFLOAT;
>SCORE   METHOD=2,NQPT=20,IDIST=0,Noprint;
```

4.5.1 Priori discriminação

Normalmente, a distribuição mais adotada como priori dos parâmetros de discriminação é a distribuição *log-normal*. A escolha dessa distribuição fundamenta-se na teoria,

pois, geralmente os parâmetros a_i são positivos, desse modo, indica que a distribuição do parâmetro de discriminação pode ser modelada por uma distribuição unimodal e com assimetria positiva (Mislevy, 1986).

4.5.2 Fixação de estimativas de parâmetros dos itens

No programa BILOG-MG a fixação de estimativas de parâmetros dos itens necessita de um método de estimação bayesiano, se o objetivo for fixar parâmetros de alguns itens e calibrar o restante, assim como é feito na estimação do SAEB. O procedimento adotado pelo programa para fixar apenas parte dos itens faz uso de distribuições a priori mais adequadas para cada parâmetro. São definidas distribuições a priori na qual as médias são os próprios valores dos parâmetros que deseja-se fixar e os desvios-padrão são bem pequenos, de forma que a distribuição a priori fica aproximadamente degenerada naquele ponto. Na realidade o que acontece é que todos os parâmetros fixados são estimados outra vez, porém a convergência é sinteticamente induzida para os valores que se deseja.

4.6 Os principais pacotes do R em TRI

Na linguagem computacional *R* é possível encontrar vários packages implementados para o uso da técnica estatística TRI. Entre os principais packages, destacam-se: o *ltm* (Rizopoulos, 2006), *irtoys* (Partchev, 2009) e o *mirt* (Chalmers et al., 2012).

O *ltm* (Latent Trait Models under IRT) pode ser usado para dados dicotômicos e politômicos sob a abordagem da Teoria da Resposta ao Item. Engloba o modelo de Rasch, o modelo logístico de 2 e 3 parâmetros, o modelo de resposta gradual e os Modelos de Crédito Parcial Generalizado.

O *irtoys* (A Collection of Functions Related to Item Response Theory) contém um conjunto de funções úteis para aqueles que tem o interesse de aprender e praticar a TRI. Fornece análise das funções com uma interface simples. E ainda, pode ser combinado em programas maiores como o BILOG-MG e o próprio *ltm* mencionado anteriormente.

Por fim, o pacote *mirt* (Multidimensional Item Response Teory) foi criado para análise de dados de respostas dicotômicas e politômicas usando modelos de traços latentes unidimensionais e multidimensionais sob a abordagem da Teoria da Resposta ao Item.

4.7 Diferenças e limitações entre o BILOG-MG e o MIRT

O BILOG-MG é um programa comercial para análise TRI para dados dicotômicos (certo/errado), tem implementado os modelos unidimensionais logísticos de 1,2,3 parâmetros, incluindo ajuste e funcionamento diferencial de itens. Já o mirt é um pacote do programa R, de versão livre, para dados de respostas dicotômicas e politômicas, com modelos unidimensionais e multidimensionais da TRI.

Capítulo 5

Reanalisando o Saeb 2011 a 2017

5.1 Processo de Recalibração: pacote *Recalibra*

Com a finalidade de recuperar as características dos itens da avaliação SAEB da edição de 2011 – as quais não são divulgadas nos microdados – foi implementado um pacote no *software R* denominado *"ReCalibra"*. Para o funcionamento do pacote algumas informações disponibilizadas pelo INEP por meio dos microdados são essenciais, como: o vetor de resposta do estudante, gabarito da prova, identificação do item, identificação do caderno e a proficiência do aluno estimada na escala SAEB.

O pacote implementado tem como propósito encontrar os valores das estimativas dos parâmetros dos itens que geram as proficiências estimadas mais próximas das proficiências conhecidas (SAEB), de forma que o somatório entre a diferença dessas habilidades ao quadrado seja a menor possível, sendo assim, quanto menor for essa diferença maior será o indício que as estimativas dos parâmetros estão próximas das estimativas do INEP. De modo geral, tenta-se aproximar as estimativas dos parâmetros dos itens minimizando a soma dos quadrado desses desvios. A soma dos quadrados dos desvios é uma medida bastante adotada para mensurar a precisão de um processo de estimação, sendo matematicamente representada na seguinte expressão:

$$SQD = \sum_{j=1}^{n} (\theta_j^* - \hat{\theta}_j)^2, \tag{5.1}$$

sendo SQD representando a diferença entre as habilidades (Desvios) ao quadrado; θ_j^* a estimativa da habilidade de cada individuo j SAEB;e $\hat{\theta}_j$ a habilidade estimada de cada individuo j, obtida no *Recalibra*.

O pacote *"ReCalibra"* executa o procedimento de retomar a estimativas dos parâmetros do itens em duas etapas, que se caracterizam da seguinte maneira:

1. Primeira Etapa: é a fase para obter a estimativa inicial dos parâmetros dos itens. Os passos dessa fase são:

- Importa a base de dados do SAEB (microdados) para o programa R com suporte do pacote *data.table*;
- Retira uma amostra aleatória de 100.000 respondentes da área e nível avaliada que responderam a prova regular;
- Realiza o ajuste do ML3 da TRI utilizado no SAEB via pacote *mirt*. Calibra os parâmetros dos itens utilizando o método da MVM com o algoritmo EM e estima as proficiências dos respondentes da amostra pelo método EAP ;
- Compara as estimativas das proficiências divulgadas pelo INEP com as estimativas das proficiências inicias (transformada para escala SAEB). Se na primeira fase as estimativas das habilidades inicias forem muito próximas das divulgadas pelo INEP (analisando o SQD), o processo não continua, caso contrário, o processo segue para a segunda etapa.

2. Segunda Etapa: chamada Refinamento da Calibração, segue os seguintes passos:

- Retirada uma segunda amostra de 5.000 respondentes da área e nível avaliada;
- Utiliza os resultados da primeira etapa como estimativa inicial. Nas estimativas dos parâmetros dos itens inicias são feitas pequenas alterações em cada etapa do processo. As alterações são realizadas por ordem, primeiro para parâmetro a de cada item, depois no parâmetro b e por último o parâmetro c. Depois de alterado os três parâmetros o processo completa um ciclo.
- Estima-se novamente as proficiências e as compara (transformada para escala SAEB) com proficiências divulgadas pelo INEP e se obtem a soma dos quadrados (SQD). O processo continua até que a diferença do SQD do ciclo atual com o SQD do ciclo anterior, seja menor que 1 (critério de parada).

5.2 Resultados por ano

Nesta seção são apresentados os resultados obtidos pelo pacote *"ReCalibra"* para a disciplina de Língua Portuguesa e para todos os níveis de escolaridade da prova SAEB, especificamente da edição de 2011. Os gráficos a seguir apresentam os histogramas dos desvios, que representam a comparação das habilidades do SAEB e as habilidades estimadas pelo *"ReCalibra"* para cada etapa do pacote.

As Figuras 5.1, 5.2 e 5.3 apresentam os histogramas dos desvios das proficiências na Escala SAEB (250,50) do 5º ano e 9º ano do ensino fundamental e o 3º ano do ensino médio, respectivamente. Observa-se nos histogramas que os valores dos desvios apresentam-se relativamente altos para a distribuição de todos os níveis de escolarização, indicando inicialmente que as estimativas inicias dos parâmetros não ficaram próximas as do INEP.

A partir das estimativas iniciais, realizou-se a etapa do refinamento das estimativas dos parâmetros (fase 2). A Figura 5.4 apresenta o resultado da nova distribuição dos desvios (escala SAEB) após a fase 2 do 5º ano do ensino fundamental de língua portuguesa da edição de 2011, nota-se uma redução significativa dos valores dos desvios comparado ao resultado da fase inicial do processo. Também, observa-se uma redução considerável dos valores dos desvios das proficiências do 9º ano do ensino fundamental, como mostra a Figura 5.5. Um resultado ainda melhor na redução dos desvios das habilidades aconteceu para o 3º ano de ensino médio. Esses resultados apresentados indicam que as estimativas recuperadas no procedimento do *"ReCalibra"* estão bem próximas das estimativas dos itens da prova SAEB.

Figura 5.1 *Histograma dos desvios das habilidades do 5º ano do Ensino Fundamental-Língua Portuguesa - edição 2011 - Fase 1.*

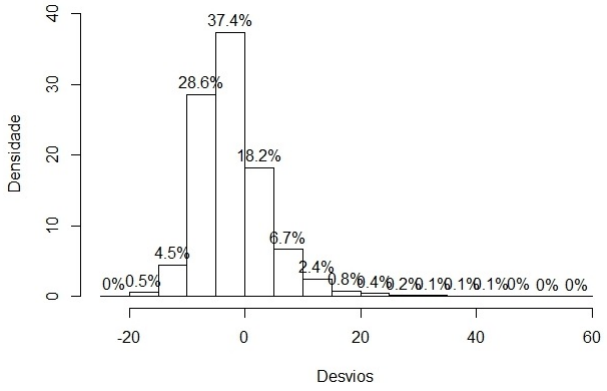


Figura 5.2 *Histograma dos desvios das habilidades do 9º ano Ensino Fundamental - Língua Portuguesa - edição 2011- Fase 1.*

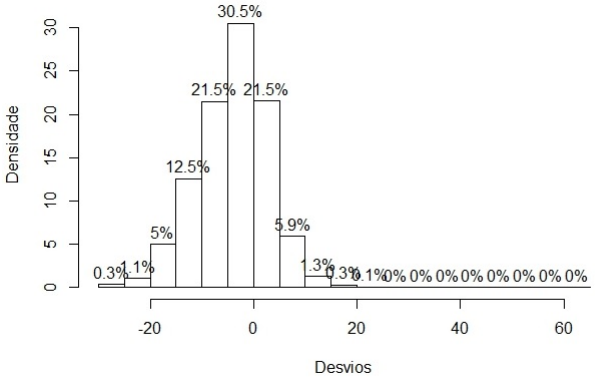


Figura 5.3 *Histograma dos desvios das habilidades do 3º ano do Ensino Médio - Língua Portuguesa - edição 2011- Fase 1.*

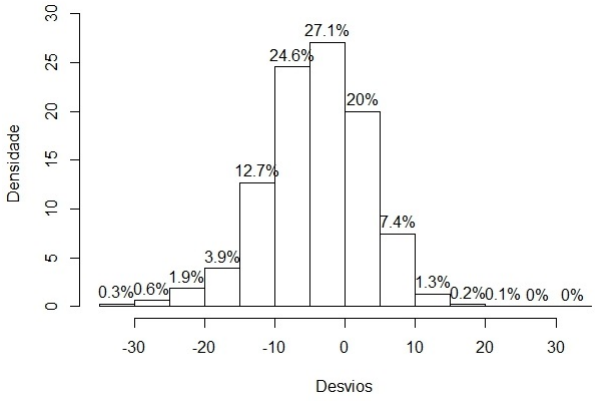


Figura 5.4 *Histograma dos desvios das habilidades do 5º ano do Ensino Fundamental-Língua Portuguesa - edição 2011 - Fase final.*

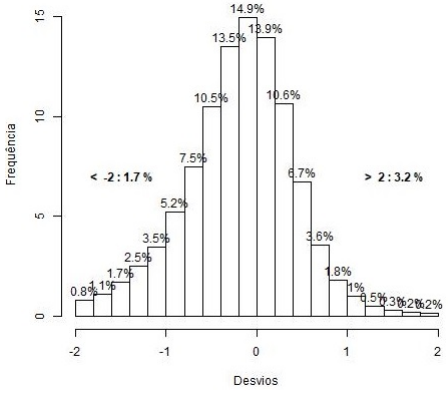


Figura 5.5 *Histograma dos desvios das habilidades do 9º ano Ensino Fundamental - Língua Portuguesa - edição 2011- Fase final.*

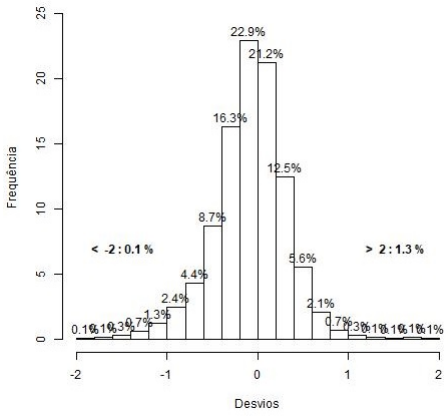
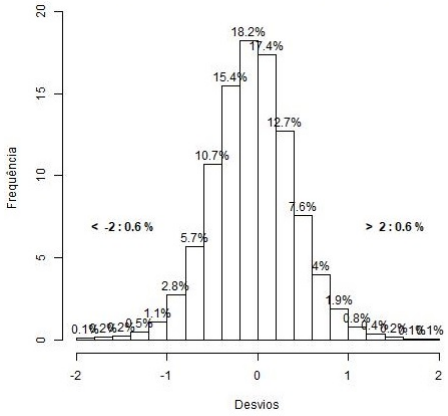


Figura 5.6 *Histograma dos desvios das habilidades do 3º ano do Ensino Médio - Língua Portuguesa - edição 2011- Fase final.*



5.3 O uso de pesos no Saeb

Conforme o relatório técnico (disponível no microdados) da edição da prova SAEB de 2011, 2013 e 2015, foram atribuídos pesos aos alunos para expansão da amostra tanto para área de Leitura quanto para Matemática . Essa ponderação foi considerada devido estudantes realizarem a prova de uma área do conhecimento e da outra não. Por esse motivo, as ponderações são diferentes para cada disciplina. Para esses estudantes, suas habilidades são consideradas somente para análise dos resultados da disciplina que o mesmo realizou. Para a edição de 2017, a nota técnica informou que todos os estudantes realizaram simultaneamente os testes de ambas as disciplinas, de forma que não foi necessário realizar ponderações distintas para cada área do conhecimento, isto é, o peso para a expansão é igual para Língua Portuguesa e Matemática.

5.4 Estimação conjunta 2011-2017

No trabalho foi realizada a estimação conjunta entre as edições de 2011 à 2017 da disciplina de língua portuguesa para o 5º e 9º anos do ensino fundamental (EF) e 3º ano do ensino médio (EM) das escolas estaduais. Realizou-se o processo de equalização horizontal (2011-2017) para cada nível (5º EF, 9º EF e 3º EM), fixando os parâmetros estimados do ano de 2011 pelo pacote *ReCalibra*. Todas as sintaxes, bem como os bancos de dados e estimativas de parâmetros de itens, para cada um dos 3 níveis será disponibilizado no Apêndice A, bem como no site `www.heliton.ufpa.br/saeb`.

Cabe ressaltar que as estimativas obtidas com o ReCalibra estão na escala (0,1)-SAEB. No entanto, como as análises foram feitas por nível, e de forma a evitar um Efeito Prova (Andrade e Borgatto (2012)), as estimativas dos parâmetros dos itens ($a's$ e $b's$) foram padronizados pelos valores dos $b's$, cujas estimativas transformadas tiveram média zero e desvio-padrão 1. O desvio-padrão dos $b's$ foi usado para transformar os valores dos $a's$. Após obtenção das estimativas finais, houve a reversão do processo para a escala (0,1)-SAEB e, posteriormente, para a (250,50)-SAEB.

Na próxima seção serão apresentados os resultados dessa equalização e estimação das proficiências.

5.5 Comparação de resultados

Nesta seção são apresentados os resultados da estimação conjunta das provas SAEB das edições de 2011 à 2017 da disciplina de Língua Portuguesa por nível (5º EF, 9º EF e 3º EM) da rede estadual, as médias das proficiências produzidas pelo INEP e as diferenças em valor absoluto entre as médias das proficiências do INEP e as estimativas médias das proficiências da estimação conjunta.

A Tabela 5.1 apresenta o resultado das médias das proficiências da estimação conjunta do 5º ano do ensino fundamental, observa-se que a média da habilidade estimada da prova do ano de 2011 apresentou resultado aproximado da média da habilidade do SAEB estimada pelo processo de equalização tradicional do INEP, entretanto os resultados das médias das edições de 2013, 2015 e 2017 na estimação conjunta apresentaram diferenças significativas em relação as médias divulgadas pelo INEP. Um resultado similar ocorreu na estimação conjunta do 9º do ensino fundamental, como apresentado na Tabela 5.2, a estimação conjunta das médias também apresentaram diferenças significativas em relação as médias divulgadas pelo INEP nos anos de 2013 à 2017.

Tabela 5.1 *Proficiências médias de Língua Portuguesa da prova SAEB - 5º ano - Rede Estadual*

Ano	n	Proficiência-SAEB	Proficiência-Estudo	Diferença
2011	250000	191,9	191,7	0,2
2013	250000	199,5	193,3	6,2
2015	250000	211,4	201,4	10,0
2017	250000	218,4	208,2	10,2

Tabela 5.2 *Proficiências médias de Língua Portuguesa da prova SAEB - 9º ano - Rede Estadual*

Ano	n	Proficiência-SAEB	Proficiência-Estudo	Diferença
2011	250000	240,8	240,6	0,2
2013	250000	241,3	249,1	7,8
2015	250000	248,3	257,2	8,9
2017	250000	255,7	265,2	9,5

O resultado da estimação conjunta do 3º ano do ensino médio é apresentado na Tabela 5.3, onde pode-se observar que as proficiências médias do SAEB foram bem aproximadas nos anos de 2011 à 2015 em relação às divulgadas pelo INEP, porém o ano de 2017 apresentou resultado da habilidade média bem distante da proficiência média do INEP.

Tabela 5.3 *Proficiências médias de Língua Portuguesa da prova SAEB - 3º ano - Rede Estadual*

Ano	**n**	**Proficiência-SAEB**	**Proficiência-Estudo**	**Diferença**
2011	40207	257,1	257,0	0,1
2013	55927	252,9	253,0	0,1
2015	45333	257,4	259,8	2,4
2017	60000	260,7	270,7	10,0

Na próxima seção serão delineadas as conclusões gerais do estudo e possíveis causas e trabalhos futuros em consequência destes resultados.

Capítulo 6

Conclusões e Considerações Gerais

A conhecida escala SAEB tem sido bastante usada como referência nacional para diversas avaliações estaduais que adotam a referida escala como padrão. Pela necessidade de manter essa escala com critérios bem estabelecidos, o presente estudo teve como finalidade avaliar fatores que poderiam estar interferindo na manutenção da escala do SAEB. Para isso, realizou-se a equalização conjunta entre as edições da prova SAEB de 2011 a 2017 visando verificar se a equalização realizada desta maneira manteria as médias das proficiências próximas às estimadas pelo INEP. Pôde-se averiguar que a estimação proposta apresentou resultados razoavelmente diferentes dos resultados obtidos pela calibração tradicional do SAEB. Abaixo elencamos possíveis causas que podem ter provocado tais diferenças:

1. Talvez os itens fixados de 2011 não tenham sido bem calibrados. Consideramos que isso é algo pouco provável, pois as habilidades para este ano foram bem reproduzidas;
2. Eventuais problemas de posição de itens, mas ressalta-se que as posições dos itens foram confirmadas pelo INEP.
3. A transformação para evitar o Efeito-Prova: procedimento similar já é realizado na análise tradicional.
4. Prioris na etapa de calibração podem ter provocado o Efeito Compressão/Expansão (variabilidade decrescente/crescente entre níveis quando estes aumentam).

6.1 Recomendações para trabalhos futuros

Em trabalhos futuros, pretende-se:

1. refazer para toda a série: 1997 à 2017;

2. incluir todos os estratos do SAEB;

3. realizar estudos de simulação para verificar a origem da divergência.

Bibliografia

Andrade, Dalton Francisco de e Adriano Ferreti Borgatto (2012). "O efeito da prova na estimativa da proficiência através da TRI". Em: *Estudos em Avaliação Educacional* 23(51), pp. 102–114.

Andrade, Dalton Francisco de, Heliton Ribeiro Tavares e Raquel Valle (2000). "Teoria da Resposta ao Item: conceitos e aplicações". Em: *ABE, Sao Paulo*.

Andriola, Wagner Bandeira (2009). "Psicometria moderna: caracterısticas e tendências". Em: *Estudos em Avaliação Educacional* 20(43), pp. 319–340.

Bekman, Roberto M (2001). "Aplicação dos blocos incompletos balanceados na Teoria de Resposta ao Item." Em: *Estudos em avaliação educacional*(24), pp. 119–136.

Birnbaum, ALord (1968). "Some latent trait models and their use in inferring an examinee's ability". Em: *Statistical theories of mental test scores*.

Bock & Zimowski, Michele F (1998). *BILOG-MG: Multiple-group IRT analysis and test maintenance for binary items*. Scientific Software International.

Bock, R Darrell e Murray Aitkin (1981). "Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm". Em: *Psychometrika* 46(4), pp. 443–459.

Bock, R Darrell e Marcus Lieberman (1970). "Fitting a response model forn dichotomously scored items". Em: *Psychometrika* 35(2), pp. 179–197.

Chalmers, R Philip et al. (2012). "mirt: A multidimensional item response theory package for the R environment". Em: *Journal of Statistical Software* 48(6), pp. 1–29.

Conde, Frederico Neves e Jacob A Laros (2007). "Unidimensionalidade ea propriedade de invariância das estimativas da habilidade pela TRI". Em: *Avaliaçao Psicologica: Interamerican Journal of Psychological Assessment* 6(2), pp. 205–215.

Guilford, Joy Paul (1954). "Psychometric methods". Em:

Gulliksen, HAROLD (1950). *Theory of Mental Tests. New York: JohnWiley & sons*.

Horta Neto, J. L. (2006). "Avaliação externa: a utilização dos resultados do Saeb 2003 na gestão do sistema público de ensino fundamental no Distrito Federal." Em: *Repositório UNB*.

INEP, INSTITUTO NACIONAL DE ESTUDOS E PESQUISAS EDUCACIONAIS ANÍSIO TEIXEIRA. (2019). *Microdados da Aneb e da Anresc 2017*. URL: `http://portal.inep.gov.br/basica-levantamentos-acessar`.

Klein, Ruben (2009). "Utilização da teoria de resposta ao item no Sistema Nacional de Avaliação da Educação Básica (Saeb)". Em: *Revista Meta: Avaliação* 1(2), pp. 125–140.

Lord, Frederic (1952). "A theory of test scores." Em: *Psychometric monographs*.

Lord, Frederic M (1974). "Estimation of latent ability and item parameters when there are omitted responses". Em: *Psychometrika* 39(2), pp. 247–264.

Mislevy, Robert J (1986). "Bayes modal estimation in item response models". Em: *Psychometrika* 51(2), pp. 177–195.

Mislevy, Robert J e Martha L Stocking (1989). "A consumer's guide to LOGIST and BILOG". Em: *Applied psychological measurement* 13(1), pp. 57–75.

Partchev, I (2009). "irtoys: Simple Interface to the Estimation and Plotting of IRT Models". Em: *R package version 0.1* 2.

Pasquali, Luiz (2009). "Psychometrics". Em: *Revista da Escola de Enfermagem da USP* 43(SPE), pp. 992–999.

Pasquali, Luiz e Ricardo Primi (2003). "Fundamentos da teoria da resposta ao item: TRI". Em: *Avaliaçao Psicologica: Interamerican Journal of Psychological Assessment* 2(2), pp. 99–110.

Rizopoulos, Dimitris (2006). "ltm: An R package for latent variable modeling and item response theory analyses". Em: *Journal of statistical software* 17(5), pp. 1–25.