Denis da Silva Paranhos

PROCESSO DE INCORPORAÇÃO DE ITENS NA ESCALA DO SAEB PARA A CRIAÇÃO DE UM BANCO DE ITENS

Dissertação submetida ao Programa de Pós-Graduação em Métodos e Gestão em Avaliação da Universidade Federal de Santa Catarina para obtenção do Grau de Mestre em Métodos e Gestão em Avaliação.
Orientador: Prof. Dr. Dalton Francisco de Andrade

Florianópolis
2016

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

Paranhos, Denis da Silva

Processo de Incorporação de Itens na Escala do SAEB para a Criação de um Banco de Itens / Denis da Silva Paranhos; orientador, Dalton Francisco de Andrade - Florianópolis, SC, 2016.

96 p.

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

Inclui referências

1. Métodos e Gestão em Avaliação. 2. Teoria de Resposta ao Item. 3. Equalização. 4. Escala. 5. SAEB. I. Andrade, Dalton Francisco de. II. Universidade Federal de Santa Catarina. Programa de Pós-Graduação em Métodos e Gestão em Avaliação. III. Título.

Denis da Silva Paranhos

PROCESSO DE INCORPORAÇÃO DE ITENS NA ESCALA DO SAEB PARA A CRIAÇÃO DE UM BANCO DE ITENS

Esta dissertação foi julgada adequada para obtenção do título de mestre e aprovada em sua forma final pelo Programa de Pós-Graduação em Métodos e Gestão em Avaliação.

Florianópolis, 24 de outubro de 2016.

_____
Prof. Renato Cislaghi, Dr.
Coordenador do Programa

Banca Examinadora:

_____
Prof. Dalton Francisco de Andrade, Dr.
Universidade Federal de Santa Catarina
Orientador

_____
Profª. Maria Silvia Olivi Louzada, Drª.
Universidade Estadual Paulista

_____
Prof. Juliano Anderson Pacheco, Dr.
Federação das Indústrias do Estado de Santa Catarina

_____
Prof. Adriano Ferreti Borgatto, Dr.
Universidade Federal de Santa Catarina

Este trabalho é dedicado à minha amada esposa, sempre presente e compreensiva nas horas de trabalho intenso e à minha gerente, Maria Christina Salerno dos Santos, minha maior incentivadora profissional.

AGRADECIMENTOS

Presto os meus mais sinceros agradecimentos ao meu orientador, Professor Doutor Dalton Francisco de Andrade pela clareza nas explicações e nas direções apontadas, muitas vezes repetidas em várias ocasiões, ao Professor Doutor Adriano Ferreti Borgatto e à sua enorme paciência, sempre muito solícito, disposto e esclarecedor em todas as etapas deste trabalho, à minha gerente e amiga Maria Christina Salerno dos Santos com seu afeto maternal e seu empenho sem fim na minha formação desde o longínquo ano 2000, e ao também amigo e presidente do Instituto Qualidade no Ensino, Sr. Horácio Almendra, pela chancela financeira sem a qual este título não seria possível e do qual eu carinhosamente empresto minha epígrafe. Escrevo que, explicitamente, sem vocês, este trabalho não seria possível.

Agradeço ainda à minha revisora ortográfica, Maria Helena Braga, cuja educação e graça impõem obediência. E aos meus colegas de trabalho cujas lutas e desafios compartilhados são inspiradores.

E agradeço a todos que, direta ou indiretamente, contribuíram ou apoiaram este trabalho.

“Vencer sem riscos é triunfar sem glória”
Pierre Corneille

RESUMO

Esta dissertação introduz as características do processo de incorporação de itens na escala de proficiência do SAEB, por meio da Teoria de Resposta ao Item. Também são descritas técnicas da Teoria Clássica dos Testes, usadas na averiguação de resultados como o uso do coeficiente de correlação bisserial, análise de grupos extremos, uso da variável Normit, uso do Coeficiente Alfa de Cronbach e estudo de desempenho de alternativas.

Junto com as análises realizadas, é apresentada em detalhes a interpretação dos resultados, as decisões tomadas mediante situações encontradas em cada etapa e a relação entre os resultados obtidos e os pressupostos pedagógicos envolvidos antes e depois do processo, que corroboram para alcançar os objetivos propostos.

A rotina de uso do *software* BILOG-MG para uso em múltiplos grupos, assim como seus relatórios, também são discutidos de maneira pormenorizada.

Palavras-chave: Teoria de Resposta ao Item. Equalização. Escala.

ABSTRACT

This paper introduces the characteristics of the items annexation process into SAEB proficiency scale by Item Response Theory. It also describes techniques of Classical Theory of Tests used in the results verification such as biserial correlation coefficient, analysis of external groups, use of Normit variable, use of Cronbach's alpha coefficient and alternative performance study.

Along with the analysis performed, a detailed interpretation of the results is presented. The decisions taken considering the situation in each stage and the relation between the achieved results and the pedagogical assumptions (before and after the process) that support the accomplishment of the objectives set out are also presented.

The routine of use of BILOG-MG software for application in multiple groups as well as the corresponding reports are also discussed in detail.

Keywords: Item Response Theory. Equalization. Scale.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

ANA - Avaliação Nacional da Alfabetização
ANEB - Avaliação Nacional da Educação Básica
ANRESC - Avaliação Nacional do Rendimento Escolar
BIB - Blocos Incompletos Balanceados
CAT - Computerized Adaptive Testing
CCI – Curva Característica do Item
CII - Curva de Informação do Item
EAP - Expected a Posteriori
ENEM - Exame Nacional do Ensino Médio
FII - Função de Informação do Item
FRI - Função de Resposta do Item
INEP - Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira
IQE - Instituto Qualidade no Ensino
ML3 – Modelo logístico de três parâmetros
MVM - Máxima Verossimilhança Marginal
NAEP - National Center for Education Statistics
PISA - Programme for International Student Assessment
SAEB - Sistema de Avaliação da Educação Básica
SARESP - Sistema de Avaliação de Rendimento Escolar do Estado de São Paulo
TCT – Teoria Clássica dos Testes
TOEFL - Test of English as a Foreign Language
TRI – Teoria de Resposta ao Item

SUMÁRIO

1 - INTRODUÇÃO

A consolidação de escalas nacionais de proficiências em disciplinas fundamentais pelo INEP, a partir da década de 1990, pode ser considerada uma grande conquista para a educação brasileira. Além de inserir o Brasil entre os grandes avaliadores mundiais na área da educação, introduz de maneira objetiva a referência na aferição de conteúdos, permitindo comparar desempenhos regionais e nacionais ao longo do tempo.

A partir da criação deste marco, pôde-se comprovar o amplo crescimento do interesse público pelas medidas que resultaram das grandes avaliações que passaram a ser aplicadas em nível nacional. Hoje, a escala do Sistema de Avaliação da Educação Básica - SAEB é amplamente conhecida não somente nos meios acadêmicos e administrativos, mas também por famílias de estudantes de escolas públicas.

A aplicação da Prova Brasil, que integra o SAEB, faz parte um instrumento amplo e valioso, que subsidia a formulação e o monitoramento de políticas educacionais, fazendo parte também de critérios para a distribuição de recursos entre estados e municípios.

Depois de alguns anos após a sua concepção, foi incorporado ao SAEB a Teoria de Resposta ao Item, um conjunto de modelos matemáticos que garante a comparabilidade entre avaliações ao longo dos anos. Esta teoria, que será apresentada e demonstrada mais adiante neste trabalho, permite, seguindo alguns pressupostos, a incorporação de itens de diferentes avaliações de institutos parceiros do INEP, privados e públicos, à escala de proficiência nacional, resultando em medidas comparáveis com as referências nacionais do SAEB. Entre outras características interessantes destes modelos matemáticos, esta em especial permite uma vasta gama de estudos por profissionais da área de educação que enriquecem a análise e o diagnóstico da educação brasileira.

Outra escala largamente utilizada, e que segue os mesmos fundamentos da escala do SAEB, é a escala do ENEM. Essa última muito difundida especialmente por ser atualmente critério de seleção de alunos de diversas universidades públicas e, portanto, um assunto de muito interesse na sociedade brasileira.

A realização desta experiência surge do intuito de se realizarem avaliações em redes de ensino municipais do Brasil nos anos em que não há a aplicação da Prova Brasil, com o propósito de entregar um

diagnóstico preciso do quadro atual da rede em questão, provocando mudanças de estratégias de ensino rumo à melhoria da educação no nosso país.

1.1 - OBJETIVOS

Dentro da Teoria de Resposta ao Item, este trabalho busca avançar na questão da criação de itens de Matemática e Língua Portuguesa de 5º e 9º anos do Ensino Fundamental comparáveis com as referências nacionais, cujos parâmetros estejam na métrica da escala de proficiência do SAEB.

Este trabalho define ainda como objetivos específicos:

a) criar um banco eletrônico de itens equalizados na escala do SAEB;
b) analisar plenamente a integridade das características e parâmetros estimados;
c) identificar itens-âncora durante o processo de equalização;
d) possibilitar o enriquecimento, conforme a análise dos resultados alcançados, da escala do SAEB, com descrições adicionais dos níveis da mesma.

2 - FUNDAMENTOS DA TEORIA DE RESPOSTA AO ITEM - TRI

2.1 - UM POUCO DA TEORIA DE RESPOSTA AO ITEM

A Teoria de Resposta ao Item, também conhecida como TRI, surge, a partir da segunda metade do século XX, da dificuldade de comparar conhecimentos de populações distintas ou de indivíduos que realizaram avaliações diferentes. As principais medidas utilizadas até este período fazem parte da Teoria Clássica dos Testes - TCT, que ainda é usada largamente, e concentra suas análises em seus escores brutos e no instrumento de avaliação ou conjunto de itens, fazendo com que suas medidas de análise dependam muito da população avaliada. Para se comparar populações distintas na TCT é necessário que as populações sejam submetidas às mesmas provas ou a provas paralelas ou equivalentes. No entanto, conforme define GULLIKSEN (1950), avaliações somente podem ser consideradas paralelas quando suas medidas de síntese como média e desvio padrão, entre outras, forem iguais, tornando esse processo muito complexo.

Um dos pioneiros no desenvolvimento e estudos da TRI foi LORD (1952), que introduziu um modelo unidimensional de dois parâmetros para respostas dicotômicas, utilizados em testes de aptidão. Posteriormente, LORD introduziu no modelo o parâmetro de acerto casual e foi sucedido por diversos autores, como BIRNBAUM (1968), que apresenta outra abordagem ao trabalho de LORD usando a função logística, mais conveniente matematicamente, por RASCH (1960), que trabalhou num modelo unidimensional de um único parâmetro, por SAMEJIMA (1969), que discute e apresenta um modelo de resposta gradual, entre outros modelos. Nesta pequena literatura comentada é possível notar que a TRI aborda vários cenários nas mais diferentes variações de instrumentos de avaliação com abordagens individualizadas.

2.2 - A TRI NO BRASIL

ANDRADE & KLEIN (1999) descrevem que a TRI foi primeiramente usada no Brasil em larga escala em 1995, na aplicação das avaliações do SAEB, em 1996, na avaliação do SARESP (Sistema de Avaliação de Rendimento Escolar do Estado de São Paulo) da Secretaria de Educação do Estado de São Paulo e na Avaliação das

Escolas Estaduais do Estado do Rio Grande do Norte, em 1997, onde se incluíram itens do SAEB de 1995, que permitiu uma comparação entre o estado do Rio Grande do Norte e o rendimento nacional daquele ano. Atualmente, de maneira muito mais difundida comparando-se com o final da década de 1990, a TRI é usada no ENEM (Exame Nacional do Ensino Médio), que nas suas últimas edições avaliou milhões de pessoas, e em diversas avaliações educacionais estaduais que se apropriam ou não da escala do SAEB.

2.3 - PRINCIPAIS CONCEITOS

Conforme VALLE (2000), a TRI propõe a aplicação de um conjunto de modelos matemáticos e estatísticos para a análise da medida de um traço latente, ou seja, uma característica que não pode ser observada diretamente, mas que pode ser inferida observando-se variáveis secundárias diretamente relacionadas a ela. VALLE (2000) ainda define que, diferente da TCT, que associa todas suas análises considerando a avaliação como um todo, a TRI introduz um modelo que centraliza as análises nos itens do instrumento. Essa mudança de foco em direção à análise de itens, e não do instrumento como um todo, pode ser considerada uma das grandes vantagens desta abordagem, a qual, seguindo determinados pressupostos, permite elaborar avaliações que podem ser aplicadas em populações distintas e ainda assim permanecerem comparáveis entre si. ANDRADE (2001) descreve também uma das principais características da TRI: o parâmetro de dificuldade dos itens e a proficiência do indivíduo estão na mesma escala, o que contribui para a comparabilidade entre populações e construção de instrumentos de avaliação. Esta é uma característica que abre uma imensa gama de análises como, em um cenário educacional, análises de desempenho de alunos ao longo do tempo, entre séries e anos diferentes ou de diferentes localidades.

A TRI pode ser aplicada em qualquer situação em que se deseja criar ou se apropriar de uma escala para um determinado traço latente. Esse pode ser, por exemplo, uma medida de coragem, de memória ou de infraestrutura. Naturalmente, constata-se que os fundamentos e características da TRI vão ao encontro direto com várias necessidades da área educacional.

A aplicação da TRI se propõe a estimar, segundo KLEIN (2009), a probabilidade de um indivíduo acertar determinado item como função

da sua proficiência e de parâmetros inerentes a cada item. Espera-se que essa probabilidade sempre aumente conforme aumenta sua proficiência.

A TRI, embora mais avançada, não substitui de forma alguma a TCT. Ela, na verdade, a complementa em muitos aspectos, sendo que muitas das medidas de síntese da TCT são etapas importantes do processo de análise.

Um outro aspecto muito relevante que pode ser observado na TRI é que, na medida em que o parâmetro de dificuldade dos itens e a proficiência dos indivíduos estão na mesma escala, permite-se a criação de escalas de conhecimento interpretáveis, cujos segmentos podem ser descritos. Esta característica torna-se extremamente útil e retoma uma consideração importante destacada por ANDRADE, TAVARES e VALLE (2000): o êxito da aplicação da TRI, no contexto educacional, passa necessariamente pela cooperação de especialistas tanto da área da Estatística como especialistas da área da Avaliação Educacional ou da área na qual o estudo está sendo realizado. Esta cooperação valida e enriquece o processo de construção destas escalas.

Cabe ressaltar também que, como será visto posteriormente, uma vez que a TRI permite a equalização de diferentes provas utilizando itens comuns, situando-os assim numa mesma escala, surge a possibilidade da criação de Bancos de Itens. Estas bases de dados permitem, de maneira segura, a armazenagem e o acompanhamento de todas as informações de histórico de avaliações e de itens, garantindo integridade e fácil acesso a ferramentas de criação de avaliações. Neste sentido, é notório que nas últimas décadas todos estes avanços descritos foram acompanhados pelo aumento da capacidade computacional disponível, com a criação de instrumentos de enorme utilidade que fizeram com que os cálculos complexos que envolvem boa parte destes processos pudessem ser consolidados de maneira automatizada, permitindo a aplicação desta teoria em grandes populações de indivíduos. E, na medida em que diminuem a complexidade envolvida, também ajudam a difundir o uso da TRI.

Internacionalmente, a TRI é usada amplamente em vários países da Europa, América do Sul e nos Estados Unidos para composição de seus respectivos índices nacionais e, de maneira mais global, no PISA (Programa Internacional de Avaliação de Estudantes) e no TOEFL (*Test Of English as Foreign Language*), este último já tendo sido aplicado em mais de 25 milhões de indivíduos. Para maiores informações consulte Nota Técnica ENEM (2012).

Embora seja muito utilizada para aferição de proficiência, as aplicações da TRI não se restringem apenas à área educacional, sendo aplicada, por exemplo, na criação de índices de grau de satisfação de clientes, na avaliação de intangíveis de empresas, na avaliação de intenções comportamentais, em orientação profissional, em medidas de qualidade de vida, na avaliação de sintomas depressivos, em avaliações psicológicas diversas, entre outras aplicações. Mais aplicações podem ser encontradas em MOREIRA JUNIOR (2010).

2.4 - A UNIDIMENSIONALIDADE

Segundo PASQUALI e PRIMI (2003), as teorias sobre traços latentes começaram a ser estudadas em meados da década de 30 e propõem que, para qualquer aptidão ou habilidade humana, existe um conjunto destes traços, nos quais um indivíduo qualquer se situa num espaço de *n* dimensões e que seu desempenho nesta referida habilidade pode ser expresso como vetor dos pesos de cada uma destas dimensões, como na equação abaixo:

$$\text{Desempenho} = f(\theta_1,\theta_2,\theta_3, \ldots, \theta_n)$$

com $\theta_1,\theta_2,\theta_3, \ldots, \theta_n$ representando as medidas da habilidade nas diferentes dimensões.

Neste contexto, a multidimensionalidade é aceita em vários estudos e trabalhos da Psicometria. Desta maneira pode-se concluir, de maneira natural, que para a realização de qualquer atividade humana, se requerem várias habilidades ou dimensões. Entretanto, a grande maioria dos modelos da TRI, incluindo o modelo unidimensional logístico de 3 parâmetros aqui utilizado, pressupõe unidimensionalidade.

Para atender ao caráter unidimensional do modelo, supõe-se que exista um único traço latente dominante sobre os demais, que ANDRADE, TAVARES e VALLE (2000) definem como a habilidade que está sendo medida. O traço latente medido neste trabalho, proficiência em Matemática e em Língua Portuguesa no 5º e 9º ano do Ensino Fundamental, medido nas suas respectivas escalas do SAEB, já foi objeto de estudos, apontando unidimensionalidade. Construtos multidimensionais, quando usados em modelos unidimensionais geralmente resultam em problemas de estimação de parâmetros. Maiores detalhes sobre aplicação da TRI para modelos multidimensionais podem

ser encontrados em BARBETTA, TREVISAN, TAVARES e AZEVEDO (2014) e RECKASE (2009).

2.5 - INDEPENDÊNCIA LOCAL

VALLE (2000) também expõe que outro pressuposto do modelo que deve ser respeitado é o da independência local, o qual supõe que, dada uma determinada proficiência, as respostas aos diferentes itens do instrumento de avaliação são independentes. HAMBLETON & SWAMINATHAN (1991) destacam um caráter interessante nestes pressupostos: dado que o traço latente e o modelo são unidimensionais, a unidimensionalidade implica na independência local. Neste sentido, para simplificar a aplicação da TRI neste modelo, o que se deve verificar no instrumento de avaliação é apenas a unidimensionalidade.

2.6 - O MODELO LOGÍSTICO UNIDIMENSIONAL DE TRÊS PARÂMETROS - ML3

Os trabalhos iniciais de LORD (1952) na criação do Modelo Unidimensional de dois parâmetros utilizavam a função da Ogiva Normal ou Gaussiana, baseada na distribuição normal acumulada. Este modelo usa a discriminação (parâmetro **a**) e a dificuldade do item (parâmetro **b**).

Posteriormente, LORD & NOVICK (1968) e BIRNBAUM (1968), observando a necessidade de considerar o acerto casual (parâmetro **c**), incorporaram-no ao modelo, desenvolvendo um modelo com três parâmetros. Uma nova abordagem, baseada nestes trabalhos, foi também proposta por BIRNBAUM que utiliza a Função Logística, matematicamente mais conveniente, tornando o processo de estimação mais simples e adiantando nas décadas seguintes a criação de modelos computacionais para a aplicação da TRI em grandes populações. Este modelo proposto por BIRNBAUM é denominado Modelo Logístico Unidimensional de três parâmetros, ou **ML3**, que é dado por:

$$P(U_{ij}=1|\,\theta_j) = c_i + (1-c_i)\,\frac{1}{1+e^{-a_i(\theta_j-b_i)}}$$

onde temos:

U_{ij}: uma variável dicotômica que assume o valor 1, quando o indivíduo j responde corretamente ao item i, e 0 quando o indivíduo j responde incorretamente ao item i.

θ_j: representa a proficiência (traço latente) do j-ésimo indivíduo.

$P(U_{ij} = 1 \mid \theta_j)$: é a probabilidade de um indivíduo j com proficiência θ responder corretamente ao item i. É também denominada de Função de Resposta do Item - FRI.

$\mathbf{a}_i$: é o parâmetro de discriminação do item i, com valor proporcional à inclinação da Curva Característica do Item - CCI no ponto b_i, que será descrita na próxima seção.

$\mathbf{b}_i$: é o parâmetro de dificuldade (ou de posição) do item i, medido na mesma escala da proficiência.

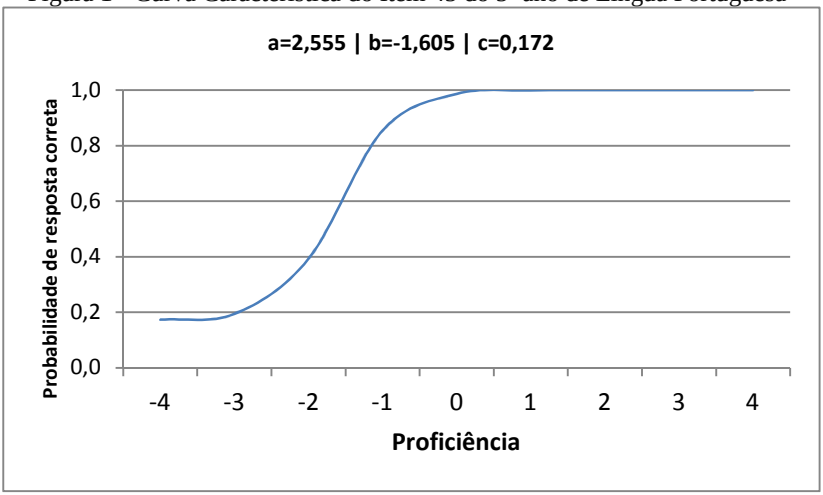
$\mathbf{c}_i$: é o parâmetro do item que representa a probabilidade de indivíduos com baixa proficiência responderem corretamente o item i.

É importante destacar que o ML3 foi desenvolvido para análise apenas de conjunto de itens dicotômicos ou dicotomizados. Existem modelos que tratam de itens não dicotômicos como o Modelo de Resposta Nominal, o Modelo de Resposta Gradual e o Modelo de Crédito Parcial, que medem não apenas respostas corretas, mas também o que foi respondido. Podem-se encontrar mais detalhes sobre estes modelos em ANDRADE, TAVARES e VALLE (2000).

2.7 - CURVA CARACTERÍSTICA DO ITEM - CCI

A CCI é a apresentação gráfica da relação dos três parâmetros do modelo com a FRI. A construção da CCI para todos os itens do instrumento facilita a análise das estimativas de parâmetros de itens. A CCI representa a FRI em cada ponto da escala de proficiência, exibindo a probabilidade de um indivíduo j com proficiência θ responder corretamente o item i, conforme detalhado na seção anterior. Na Figura 1 é representado um exemplo de uma CCI:

Figura 1 - Curva Característica do Item 45 do 5º ano de Língua Portuguesa



Nota-se que se trata de uma função crescente na qual temos na ordenada $P(U_{ii} = 1 \mid \theta_i)$ e na abscissa, proficiência. ANDRADE, TAVARES e VALLE (2000) ainda destacam três aspectos importantes sobre este gráfico:

- A forma de "S" do gráfico se justifica por uma relação não linear entre estas duas variáveis;
- Itens com o parâmetro **a** negativos não são esperados neste modelo, pois indicariam que a probabilidade de responder corretamente o item diminui com o aumento da proficiência, contrariando um dos pressupostos do modelo.
- A curva de probabilidade inicia no valor do parâmetro **c**, justamente a probabilidade de um indivíduo com baixa proficiência acertar o item.

A escala de proficiência utilizada neste gráfico é a escala (0,1) utilizada na TRI onde 0 representa a média de proficiência dos indivíduos e 1 representa um desvio padrão. Esta escala será abordada com mais propriedade na seção 2.9.

Nota-se, ainda, que a região em que a probabilidade de resposta correta para o item começa a tornar-se alta é justamente na região ao redor do parâmetro **b**. Já o parâmetro **a** neste gráfico representa o nível de inclinação da curva.

2.8 - MAIS SOBRE OS PARÂMETROS DO ML3

Os três parâmetros citados anteriormente estão intrinsecamente associados a todos os itens do instrumento de avaliação. Uma descrição mais abrangente e contextualizada destes parâmetros é apresentada a seguir:

- **a**: parâmetro de discriminação. Este parâmetro define em que medida o item consegue diferenciar alunos com proficiências diferentes. Um valor baixo é indicativo que indivíduos com proficiências diferentes terão aproximadamente a mesma probabilidade de acerto, ou seja, ele não discrimina apropriadamente alunos diferentes. Por outro lado, quando a estimativa deste parâmetro é alta, é indicativo de que este item discrimina demais os indivíduos, segregando-os basicamente em dois grupos: os que possuem baixa probabilidade de acertar o item e os que possuem alta probabilidade de acertar o item. Os alunos com proficiências bem próximas terão probabilidades de acerto muito diferentes na região do parâmetro de dificuldade, **b**.
- **b**: parâmetro de dificuldade. Este parâmetro é caracterizado como a proficiência necessária que um determinado indivíduo precisa possuir para começar a ter alta probabilidade de acertar o item. Embora teoricamente este parâmetro possa assumir qualquer valor entre $-\infty$ e $+\infty$ na escala, se espera que a maioria dos itens tenha valores entre -3 e +3, na escala (0,1) Valores fora deste intervalo não são incomuns, porém o erro de estimativa é maior para estimativas mais afastadas da média. Este parâmetro está na mesma escala da proficiência.
- **c**: parâmetro de acerto casual. Trata-se da probabilidade de um aluno com baixa proficiência responder corretamente este item. Pode assumir valores entre 0 e 1, entretanto, espera-se valores em torno de 0,25 para itens com quatro alternativas. Valores altos podem indicar a existência de distratores pouco plausíveis.

Os valores das estimativas destes parâmetros e suas relações entre si são a base de todas as análises da TRI. Estas relações interferem diretamente na representação da proficiência dos indivíduos analisados e em outras medidas. Neste sentido, deve-se destacar que processos minuciosos e bem desenvolvidos de estimação destes parâmetros são de extrema importância para o sucesso da aplicação do modelo.

2.9 - A ESCALA DE PROFICIÊNCIA NA TRI

Conforme abordado anteriormente, a proficiência de um indivíduo pode, teoricamente, assumir qualquer valor real entre $-\infty$ e $+\infty$. Portando é necessário estabelecer uma origem e uma unidade de medida da escala de modo a representar o valor médio (que será o centro da escala) e o desvio padrão das proficiências dos indivíduos em estudo. A maioria dos autores na literatura usa a escala (0,1), na qual 0 é a proficiência média do grupo em estudo e 1 é a unidade de desvio padrão. Desta forma pode-se concluir que um indivíduo situado no ponto 1,15 desta escala está 1,15 desvio padrão acima da média ou ainda que outro indivíduo situado no ponto -2,34 está 2,34 desvios-padrão abaixo da média.

No Brasil temos duas escalas de proficiência que utilizam a TRI e são conhecidas nacionalmente: a escala SAEB (250,50) na qual a média é 250 e o desvio padrão 50 e a escala do ENEM (500,100) na qual a média é 500 e o desvio padrão 100. Estas escalas foram inicialmente construídas como (0,1) e posteriormente transformadas proporcionalmente e arbitrariamente como (250,50) no caso do SAEB e (500,100) no caso do ENEM. Esta transformação respeita todas as relações de ordem entre os itens e entre as proficiências dos indivíduos já estabelecidas na escala (0,1). O principal motivo de se trabalhar com a escala (0,1) em termos práticos é a conveniência computacional, pois os processos de estimação de parâmetros envolvem cálculos complexos que podem resultar em números muito grandes. Por outro lado, o principal argumento que se pode utilizar para a transformação de escalas (0,1) para escalas maiores, principalmente na área educacional, é que os escores dos indivíduos sempre serão positivos. Isto se torna uma questão importante visto que não é natural, para o público em geral, um indivíduo ter uma proficiência negativa.

Ainda existem dois aspectos fundamentais que devem estar claros ao se analisar resultados numa escala que utiliza a TRI:

- em geral, os modelos desenvolvidos expressam a natureza acumulativa do traço latente, ou seja, quanto maior a proficiência do indivíduo, maior a probabilidade dele acertar um determinado item, relação que será matematicamente demonstrada posteriormente. Então pode-se dizer que um aluno posicionado num determinado ponto da escala tem alta probabilidade de dominar os conteúdos da referida região da escala e todos os conteúdos anteriores a este ponto;

- quando se utilizam escalas que sofrem transformações de qualquer natureza, o mais importante a se observar é a relação de ordem entre os itens e não necessariamente a magnitude da escala, e, quando uma escala sofre transformações, as distâncias das posições de todos os itens devem permanecer as mesmas, em termos da unidade de desvio padrão.

3 - PLANEJAMENTO E ORGANIZAÇÃO DA PROVA IQE

3.1 - A PROVA BRASIL E A ESCALA DO SAEB

As primeiras aplicações do SAEB ocorreram em 1990 e 1993 de maneira amostral na 1ª, 3ª, 5ª e 7ª séries do Ensino Fundamental de escolas públicas. Em 1995 adotaram-se os modelos e pressupostos da TRI, o que possibilitou a comparabilidade de resultados e iniciou a série histórica. Neste ano também foi incluída uma amostra da rede privada de ensino. Este modelo permaneceu desta maneira até 2005 quando o SAEB foi reestruturado pela Portaria Ministerial nº 931, de 21 de março de 2005, e passou a ser constituído por duas avaliações: a ANEB e a ANRESC, esta última também denominada de Prova Brasil, que passou a avaliar todas as escolas públicas, desde que atendesse a quantidade mínima de 30 alunos na série/ano estudada. Todas as aplicações destas provas, desde suas respectivas concepções, foram realizadas pelo INEP, Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira, autarquia federal ligada ao Ministério da Educação. Maiores informações sobre o histórico do SAEB podem ser encontradas em www.inep.gov.br.

A abordagem estatística que garante a comparabilidade de resultados das diferentes edições da Prova Brasil ocorre por meio da Equalização de Grupos Não Equivalentes com Itens Comuns, cujas edições a partir do início da série histórica contêm itens calibrados de edições anteriores.

Atualmente o SAEB é composto por três grandes instrumentos de avaliação:

- **Avaliação Nacional da Educação Básica** – ANEB, aplicada bianualmente, de maneira amostral, no 5º e 9º anos do Ensino Fundamental e no 3º ano do Ensino Médio em escolas da rede particular, com, no mínimo, 10 alunos no ano/série em estudo. Na rede pública, no 5º e 9º ano do Ensino Fundamental, é aplicada somente nas escolas com no mínimo 10 alunos e no máximo 19. Para o 3º ano do Ensino Médio, para escolas com no mínimo 10 alunos.
- **Avaliação Nacional do Rendimento Escolar** – ANRESC (também denominada **Prova Brasil**), aplicada de maneira censitária no 5º e 9º anos do Ensino Fundamental em escolas da rede pública com pelo menos 20 alunos matriculados e é aplicada também bianualmente.

- **Avaliação Nacional da Alfabetização** – ANA, aplicada anualmente de maneira censitária em alunos do 3º ano do Ensino Fundamental de escolas públicas, criada em 2013. Em 2015 não foi aplicada por restrições orçamentárias. Esta avaliação possui uma escala própria.

Ao se estudar duas ou mais populações, é necessário estabelecer qual população será referência para atribuição da escala (esta população de referência será descrita mais adiante). No caso da escala do SAEB, a referência é a 8ª série/9º ano da aplicação do SAEB de 1997 e, portanto, todos os resultados das edições a partir deste ano ou de avaliações que utilizam a escada do SAEB são posicionados em relação ao desempenho médio destes indivíduos. A população de 1995 também está nesta escala.

KLEIN (2009) define mais algumas características importantes sobre a escala do SAEB:

- a escala é única para cada disciplina que engloba a 4ª série/5º ano, a 8ª série/9º ano do Ensino Fundamental e o 3º ano do Ensino Médio;
- arbitrou-se, neste ano de 1997, a escala com média 250 e desvio padrão 50, a conhecida escala (250,50);
- neste ano os itens com graduações de resposta além de certo/errado foram omitidos;
- também nesta aplicação foi realizada uma calibração conjunta dos itens dos SAEBs de 1995 e 1997 e foi utilizada a base de dados contendo todas as respostas individuais de todos os alunos de todas as séries destas duas aplicações.

A Prova Brasil, na sua primeira aplicação em 2005, avaliou uma população de alunos da 4ª série/5º ano e da 8ª série/9º ano que responderam a um conjunto de 169 itens nas disciplinas de Língua Portuguesa e Matemática de escolas localizadas na zona urbana da rede pública nacional com no mínimo 30 alunos matriculados em cada série/ano. Em INEP (2015) e HORTA NETO (2007) o leitor pode encontrar maiores informações, além da Portaria Ministerial nº 69 de 4 de maio de 2005.

3.2 - O ESQUEMA DE BLOCOS INCOMPLETOS BALANCEADOS - BIB

Durante o processo de concepção do SAEB e da Prova Brasil, os especialistas envolvidos determinaram que a matriz pedagógica utilizada por este instrumento deveria ser a mais ampla possível. Porém, para tal, é necessário apresentar uma grande quantidade de itens aos alunos, tornando a prova muito extensa e exaustiva de ser realizada. Com o objetivo de elaborar um retrato fidedigno da educação em nosso país e, ao mesmo tempo, não exigir uma prova com uma grande quantidade de itens, foi incorporado um esquema de cadernos e blocos, denominado Blocos Incompletos Balanceados, também conhecido como BIB. Este esquema é muito utilizado em avaliações de larga escala, pois, entre outras vantagens, uniformiza a exposição de itens, blocos e cadernos aos indivíduos, evitando vieses no instrumento.

O BIB pode ser explicado, de maneira simplificada, como um rodízio estruturado de blocos de itens distribuídos em cadernos. Atualmente na Prova Brasil, existem 7 blocos com 11 itens cada para uma das duas disciplinas do 5º ano e 7 blocos com 13 itens cada para cada uma das duas disciplinas (Matemática e Língua Portuguesa) de 9º ano. Cada caderno de prova contém 2 blocos de cada disciplina, totalizando 44 itens para o 5º ano e 2 blocos de cada disciplina, totalizando 52 itens para o 9º ano. Com esta estrutura é possível compor 21 cadernos diferentes de provas para cada ano. Para maiores detalhes veja INEP (2016).

Na aplicação de 2005, fonte da amostra utilizada neste trabalho, a estrutura usada foi diferente: cada caderno de prova foi composto, de forma espiralada, por 3 blocos de 13 itens em cada bloco. Com um total 169 itens distribuídos em 13 blocos, o resultado são 26 cadernos dos quais cada aluno responde no máximo a 39 itens de cada disciplina (veja SAEB (2001)). Os pressupostos deste esquema são os seguintes:

1) Todo caderno deve conter uma mesma quantidade de blocos;
2) Cada bloco se repete uma mesma quantidade de vezes. No caso da Prova Brasil, cada bloco se repete seis vezes no conjunto total de cadernos;
3) Cada dupla de blocos é utilizada uma mesma quantidade de vezes no conjunto total de cadernos.
4) Cada item é utilizado em apenas um bloco;
5) Cada bloco é apresentado no início, no meio e no final de cada caderno uma mesma quantidade de vezes: duas vezes no início, duas no meio e duas no final.

Os cadernos de prova são distribuídos aos alunos de maneira aleatória, garantindo-se que todos os itens e todos os blocos tenham uma quantidade uniformizada de respostas.

Assim, conforme exposto acima, o uso do BIB reduz a quantidade de itens necessária para cobrir matrizes pedagógicas extensas. Outra vantagem de se usar o BIB é o número maior de itens calibrados que se obtém numa mesma prova, o que acaba por enriquecer o Banco de Itens e, por consequência, a escala de proficiência.

3.3 - A PROVA IQE

O objetivo geral deste trabalho é obter itens próprios do IQE calibrados na mesma escala do SAEB, para que seja possível realizar aplicações de provas do Ensino Fundamental comparáveis com os resultados da Prova Brasil. Neste sentido, nas reflexões iniciais, concluiu-se que seria necessária a aplicação de uma prova de ligação, denominada Prova IQE, contendo itens próprios e itens calibrados na escala do SAEB, equalizando, assim, seus parâmetros estimados. Esta prova seria, então, calibrada em conjunto com a base de dados dos itens cedidos pelo INEP.

O modelo de equalização escolhido para este trabalho é o **Modelo para Múltiplos Grupos**, proposto por BOCK & ZIMOWSKI (1997), que será descrito apropriadamente mais adiante.

Definido o modelo de equalização a ser utilizado, foi solicitada ao INEP a cessão de um conjunto de itens já calibrados, juntamente com seus respectivos parâmetros estimados e uma amostra da base de dados de respostas dos indivíduos utilizada no processo de calibração dos itens cedidos. Os dados cedidos são da primeira aplicação da Prova Brasil em 2005.

Os itens da prova de ligação foram organizados em 10 cadernos compostos por dois blocos de itens cada um, utilizando o já ilustrado BIB, assim como na Prova Brasil. Para isso foram criados cinco blocos de itens. A prova do 5º ano foi composta por 11 itens por bloco e a prova do 9º ano por 13 itens por bloco. Em cada bloco do 5º ano foram incluídos três itens do INEP e oito itens próprios do IQE. Na prova de 9º ano foram incluídos três itens do INEP e 10 itens próprios do IQE. O total de itens próprios a serem calibrados e equalizados foi de 40 itens para o 5º ano e de 50 itens para o 9º ano. As tabelas a seguir sumarizam o esquema adotado:

Tabela 1 - Esquema da prova de ligação IQE 2012 para cada disciplina

	5º ano	9º ano
Total de cadernos de prova	10	10
Total de blocos de itens	5	5
Total de blocos por caderno	2	2
Total de itens por bloco	11	13
Itens do INEP em cada bloco	3	3
Itens próprios em cada bloco	8	10
Total de itens por caderno	22	26
Total de itens do INEP	15	15
Total de itens próprios	40	50

Fonte: Elaborado pelo autor

Tabela 2 - Organização de blocos e cadernos da prova IQE 2012 para cada disciplina

Caderno de prova	Blocos	
1	Bloco 1	Bloco 2
2	Bloco 3	Bloco 1
3	Bloco 1	Bloco 4
4	Bloco 5	Bloco 1
5	Bloco 2	Bloco 3
6	Bloco 4	Bloco 2
7	Bloco 2	Bloco 5
8	Bloco 3	Bloco 4
9	Bloco 5	Bloco 3
10	Bloco 4	Bloco 5

Fonte: Elaborado pelo autor

Em cada bloco, os itens do INEP nunca se apresentam no início ou no final do bloco e nunca aparecem na sequência, ficando sempre intercalados entre itens do IQE. Todos os blocos aparecem duas vezes no início do caderno e duas vezes no final. Neste formato, cada aluno do 5º ano responderá a 22 itens e cada aluno do 9º ano responderá a 26 itens. Cada item possui 4 alternativas não ordinais com apenas uma alternativa correta.

Para realizar uma seleção adequada de itens próprios a serem equalizados foi realizado um estudo da matriz pedagógica do IQE e sua equivalência na matriz pedagógica do SAEB pelos especialistas em cada

área, para que a ampla matriz pedagógica do SAEB seja coberta da melhor maneira possível.

Constituído o instrumento de avaliação, a aplicação ocorreu em novembro de 2012 nas cidades de São José dos Campos/SP para o 9º ano e Caruaru/PE para o 5º ano. Em ambas localidades, os alunos realizaram provas de Língua Portuguesa e Matemática.

3.4 - BASE DE DADOS

Depois de tabulados, os dados foram exportados para o formato adequado, formando um arquivo para cada ano/disciplina. Estes arquivos possuem as seguintes informações:

- Identificação exclusiva de cada prova de aluno;
- Identificação do caderno de prova a que o aluno respondeu;
- Identificação do grupo do qual o indivíduo faz parte;
- Um vetor sequencial com as respostas dadas.

As provas das duas disciplinas, em cada localidade, foram aplicadas em dias distintos. O total de respondentes é apresentado na tabela a seguir:

Tabela 3 - Total de respondentes da prova IQE 2012

Prova	Total de respondentes
5º ano Matemática	3153
5º ano Língua Portuguesa	3112
9º ano Matemática	2577
9º ano Língua Portuguesa	2809

Fonte: IQE

4 - ANÁLISES PRELIMINARES DOS DADOS USANDO A TEORIA CLÁSSICA DOS TESTES - TCT

Precedendo a estimação de parâmetros usando a TRI, as análises preliminares dos dados usando a TCT contribuem para garantir a integridade dos resultados alcançados, identificando itens que se destacam por possuírem características que possam influenciar na estimação de parâmetros em alguma etapa adiante.

4.1 - A VARIÁVEL NORMIT

Segundo SAEB 2003 e NAEP 1996, o Normit é uma variável de desempenho obtida por uma transformação não linear a partir dos escores dos examinados fornecidos pela TCT. Considerando que a população é a mesma e que a distribuição dos cadernos é aleatória, o Normit pode ser usado para gerar uma variável mais justa em relação ao escore, pois leva em consideração a dificuldade do caderno que o indivíduo responde, visto que os cadernos da Prova Brasil podem possuir diferentes graus de dificuldade.

O Normit é dado por:

$$\frac{CF(i) + CF(i-1)}{2N}$$

onde temos, para $i>0$, $CF(i)$ como a frequência cumulativa para i itens corretos em um caderno e N é o número de indivíduos que responderam o caderno. Para $i=0$ temos:

$$\frac{CF(0) + \frac{CF(1)}{2}}{2N}$$

BORGATTO e ANDRADE (2012) ainda destacam que o uso do Normit evita situações em que indivíduos com o mesmo escore que realizaram provas diferentes poderiam ter desempenho diferente se tivessem realizado a mesma prova.

4.2 - CORRELAÇÃO BISSERIAL

Foi realizado um estudo sobre a Correlação Bisserial entre o Normit e a resposta de forma dicotomizada de cada alternativa dos itens,

obtendo-se um panorama da relação entre o Normit e as alternativas de cada item. Segundo PASQUALI (2003), o uso deste coeficiente é o mais indicado para situações em que as variáveis correlacionadas são contínuas, mas uma delas foi dicotomizada. Esta análise foi realizada apenas com itens próprios do IQE, observando que os dados fornecidos pelo INEP foram entregues dicotomizados e não contemplam a alternativa respondida pelo indivíduo.

O valor esperado para um item construído de forma adequada neste estudo é um coeficiente positivo e alto justamente na alternativa correta, indicando correlação entre Normit alto e indivíduos que responderam corretamente ao item. Por outro lado, se espera um coeficiente negativo e distante de zero para os distratores, indicando correlação negativa entre Normit alto e os distratores. Maiores detalhes sobre a estimação deste coeficiente pode ser encontrada em LORD & NOVICK (1968).

Os itens que apresentam distratores com um coeficiente positivo (ou, ainda que negativo, mas próximo de zero), que apresentam somente distratores com coeficiente negativo ou que a alternativa correta possui um coeficiente muito baixo devem ser acompanhados com cautela nas análises seguintes, pois podem apresentar problemas de gabarito incorreto, duas respostas corretas ou ainda nenhuma resposta correta.

Nas tabelas a seguir estão reunidos os itens que apresentam algumas das características citadas acima, devidamente comentadas posteriormente.

Tabela 4 - Itens com coeficiente de correlação bisserial atípico do 5º ano de Língua Portuguesa

Item	Gabarito	Coeficiente Bisserial das alternativas
Código 99	B	A: -0,30074
		B: 0,42898
		C: -0,35244
		D: 0,04887
Código 101	B	A: -0,13668
		B: 0,26481
		C: -0,17368
		D: 0,00908
Código 113	C	A: 0,05463
		B: -0,04178
		C: 0,18028
		D: -0,19354

Fonte: IQE

Tabela 5 - Itens com coeficiente de correlação bisserial atípico do 5º ano de Matemática

Item	Gabarito	Coeficiente Bisserial das alternativas
Código 257	A	A: 0,24029
		B: 0,00835
		C: -0,08382
		D: -0,10312
Código 260	C	A: -0,20840
		B: 0,02740
		C: 0,32888
		D: -0,13270
Código 262	C	A: -0,03139
		B: -0,05432
		C: 0,13299
		D: -0,09013

Fonte: IQE

Tabela 6 - Itens com coeficiente de correlação bisserial atípico do 9º ano de Língua Portuguesa

Item	Gabarito	Coeficiente Bisserial das alternativas
Código 134	D	A: -0,41162
		B: 0,18212
		C: -0,44324
		D: 0,13835
Código 146	A	A: 0,30598
		B: 0,01732
		C: -0,36466
		D: -0,17829
Código 151	B	A: 0,05278
		B: 0,24956
		C: -0,25802
		D: -0,11592
Código 178	A	A: 0,17659
		B: 0,12988
		C: -0,18836
		D: -0,18608
Código 187	A	A: 0,31011
		B: -0,18810
		C: -0,35081
		D: 0,06260

Fonte: IQE

Tabela 7 - Itens com coeficiente de correlação bisserial atípico do 9º ano Matemática

9º Matemática		
Item	Gabarito	Coeficiente Bisserial das alternativas
Código 276	C	A:-0,07885
		B: -0,25664
		C: 0,35066
		D: 0,03072
Código 283	D	A: -0,18671
		B: 0,00580
		C: -0,35395
		D: 0,45947
Código 292	D	A: -0,20056
		B: -0,07392
		C: 0,02010
		D: 0,39752
Código 305	C	A: -0,35872
		B: -0,47318
		C: 0,42944
		D: 0,23499
Código 311	A	A: 0,27355
		B: -0,24542
		C: 0,08106
		D: -0,11202
Código 317	C	A: -0,21634
		B: -0,03432
		C: 0,06897
		D: 0,15360
Código 318	C	A: -0,40213
		B: 0,10399
		C: 0,32445
		D: -0,23448
Código 331	C	A: -0,15423
		B: -0,21307
		C: 0,29678
		D: 0,02500

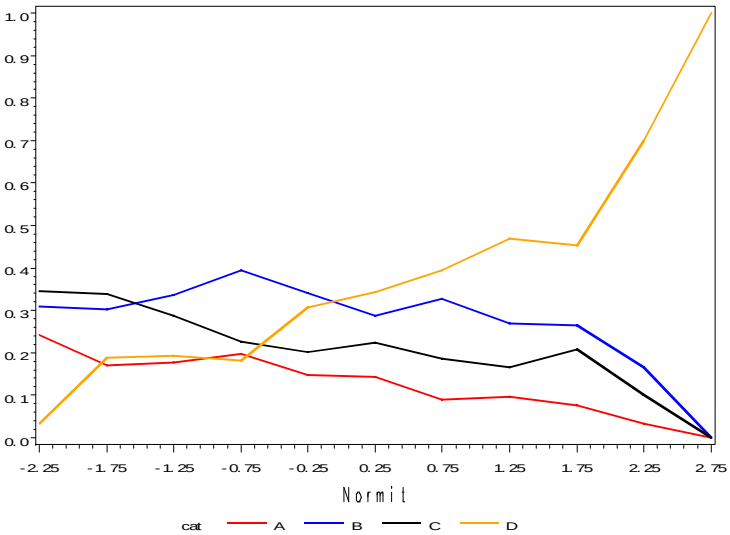
Fonte: IQE

Os itens 99, 101, 113, 260, 146, 151, 187, 276, 283, 292, 311 e 331 possuem distratores com coeficiente positivo ou ao redor de zero, indicando que, além da alternativa correta, estes itens possuem alternativas que, de alguma maneira, atraíram os alunos com bom desempenho.

O item 257, ainda que possua um valor de coeficiente razoável para a alternativa correta, possui coeficientes com valores acima do esperado para todos os distratores. O item 262, além de possuir coeficientes acima do esperado para todos os distratores, também possui um coeficiente baixo para a alternativa correta. Os itens 134, 178, 305 e 318 possuem coeficientes altos para, ao menos, um distrator. O item 317, além de possuir um valor de coeficiente baixo para a alternativa correta, possui um distrator com um alto valor e acima do valor do coeficiente da alternativa correta. Este tipo de situação pode indicar um item com erro de gabarito.

Após a análise dos valores de coeficiente de correlação bisserial dos itens, foi realizado um estudo da distribuição de respostas dos indivíduos entre os valores do Normit. O resultado foi um gráfico representado na Figura 2:

Figura 2 - Distribuição de respostas dos indivíduos do item 358 do 5º ano de Matemática – Gabarito: D

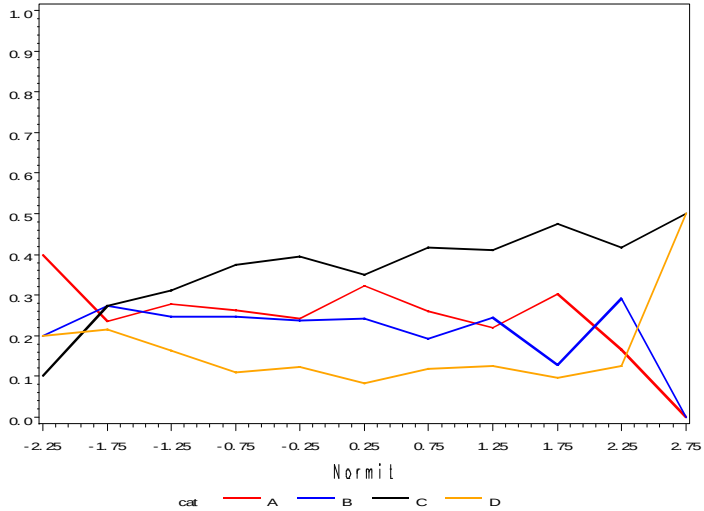


A ordenada representa a proporção de respostas dadas a cada uma das alternativas e a abscissa representa os valores do Normit. Desta maneira é possível acompanhar a proporção de respostas de cada alternativa conforme o valor do Normit aumenta. Quanto antes o desempenho da alternativa correta se sobressair em relação aos distratores, mais este item sugere uma baixa proficiência necessária para acertá-lo e, por outro lado, quando a alternativa correta passa a se destacar a partir de valores mais altos do Normit é indicação de um item que requer maior proficiência para acertá-lo. Para itens coerentes, se espera que o desempenho da alternativa correta seja sempre ascendente e o desempenho dos distratores seja descendente conforme os valores do Normit aumentam. O desempenho da alternativa correta não deve ser descendente, pois este quadro indicaria que, conforme a proficiência dos indivíduos aumenta, a proporção de respostas corretas diminui, indo contra a ideia da natureza acumulativa do conhecimento. Para o exemplo acima, a alternativa correta é a D. Nota-se que, conforme os valores do Normit aumentam, a proporção de indivíduos que responderam corretamente o item também aumenta.

Outro exemplo é o item 262 do 5º ano de Matemática, apresentado na Figura 3, em que a alternativa correta tem um

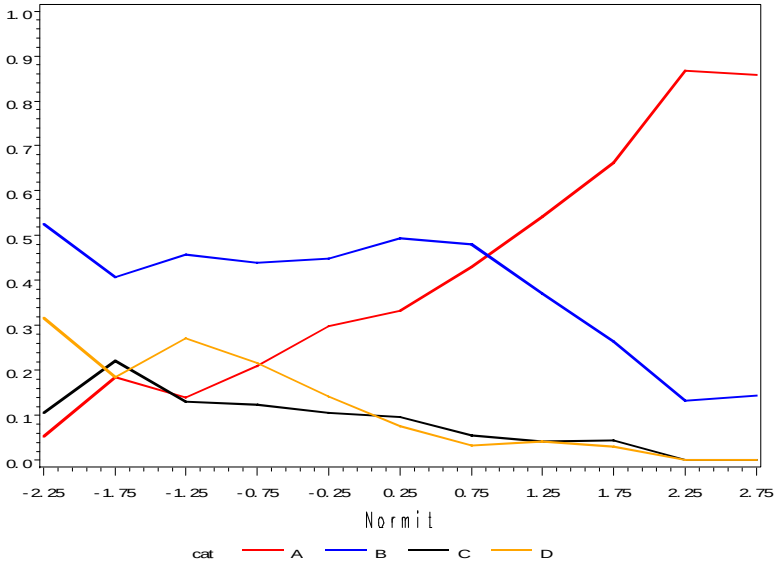
desempenho ascendente, ainda que não tão expressivo. Porém não é esperado que um ou mais distratores tenham uma proporção alta na região em que o Normit é alto, porque este tipo de comportamento sugere que indivíduos com alta proficiência estão em dúvida sobre qual é a resposta correta.

Figura 3 - Distribuição de respostas dos indivíduos do item 262 do 5º ano de Matemática – Gabarito: C



Ainda pode-se ter uma situação em que, em certas regiões do gráfico, nota-se que os indivíduos podem estar atraídos por um determinado distrator, como no caso do item 44 do 5º ano de Língua Portuguesa:

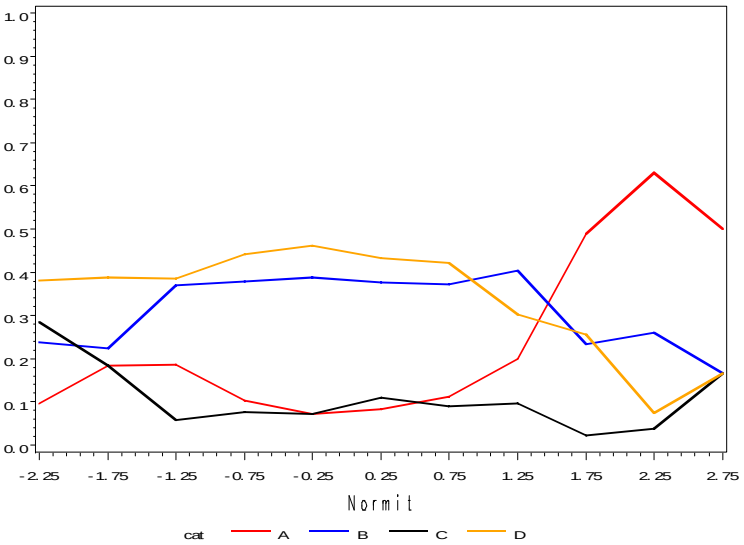
Figura 4 - Distribuição de respostas dos indivíduos do item 44 do 5º ano de Língua Portuguesa – Gabarito: A



É importante ressaltar que, caso um distrator se sobressaia na região em que a alternativa correta ainda não é a alternativa mais respondida, é indicação que indivíduos com baixa proficiência foram atraídos por algum distrator. Isso pode ser de fato uma intenção do especialista que criou o item, e, deliberadamente, se pretende criar esta situação. Por este motivo, destacando-se o caráter de convergência entre todas as áreas na aplicação da TRI, sugere-se apresentar este estudo ao especialista da respectiva área sempre que se notar um comportamento incomum de algum item para uma análise pedagógica mais profunda.

Analisando em conjunto o coeficiente de correlação bisserial e o desempenho destes gráficos podem-se encontrar situações que começam a sugerir problemas que o item poderá revelar quando tiver seus parâmetros estimados, como no caso do item 257, do 5º ano de Matemática:

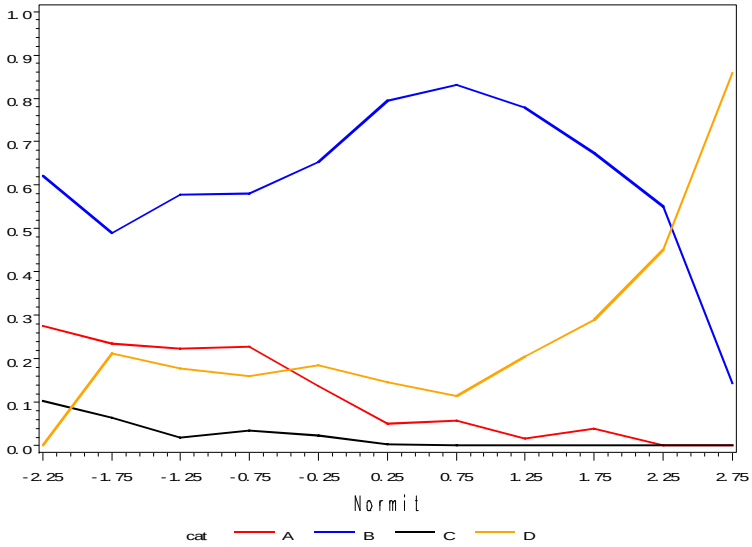
Figura 5 - Distribuição de respostas dos indivíduos do item 257 do 5º ano de Matemática – Gabarito: A



A primeira situação incomum deste gráfico é que a alternativa correta não segue uma linha plenamente ascendente. Ainda nota-se que em uma grande região do gráfico, indivíduos optaram majoritariamente pelas alternativas B e D. Analisando os coeficientes bisseriais dos distratores nota-se que este item também apresenta valores próximos de zero. Cenários como este indicam que o item pode não discriminar alunos com baixa e alta proficiência com eficiência e, portanto, pode ter problemas para estimação do parâmetro **a**.

Outro item que deve ser analisado em conjunto com o coeficiente de correlação bisserial é o item 134 do 9º ano de Língua Portuguesa, representado na Figura 6 abaixo:

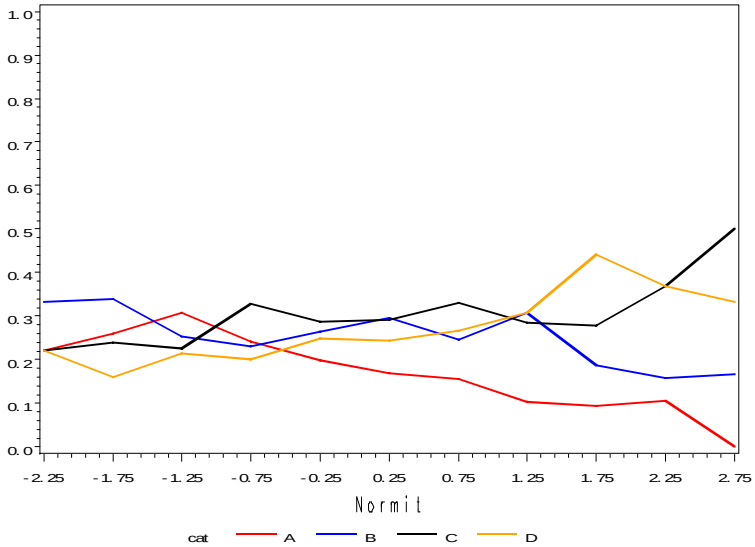
Figura 6 - Distribuição de respostas dos indivíduos do item 134 do 9º ano de Língua Portuguesa –Gabarito: D



Fica claro neste caso que uma grande parte dos respondentes optou pela alternativa B, mesmo na região em que se encontram os indivíduos com alta proficiência. E ainda observa-se que o valor do coeficiente de correlação bisserial para esta mesma alternativa é maior que o gabarito. Este tipo de situação pode resultar num parâmetro **b** com um erro de estimativa alto.

Na Figura 7 temos uma situação que também pode gerar problemas de estimação de parâmetros, na qual o item apresenta um valor de coeficiente de correlação bisserial na alternativa correta, C, menor que o distrator D.

Figura 7 - Distribuição de respostas dos indivíduos do item 317 do 9º ano de Matemática – Gabarito: C



Temos também que, mesmo na região que concentra os alunos com maior proficiência, não se nota a ascendência da alternativa correta.

Além do discutido até agora, deve-se levar em consideração também a distribuição da quantidade de respondentes em cada região da abscissa. Existem situações, nem sempre incomuns, em que um determinado problema é detectado, porém numa região com um baixo número de respondentes e, portanto, não necessariamente relevante.

Enfim, o aspecto mais importante destas análises é verificar posteriormente se as características apresentadas aqui influenciaram na estimação dos parâmetros.

4.3 - ESTUDO DE GRUPOS EXTREMOS

Para cada item também foi realizado um estudo de grupos extremos, no qual os indivíduos são separados em dois grupos:

- grupo superior composto por 27% de respondentes com os maiores escores;
- grupo inferior composto por 27% de respondentes com os menores escores.

Para cada grupo é calculada a proporção de indivíduos que responderam corretamente ao item. Espera-se que a proporção de respostas corretas seja alta no grupo superior, pois se trata de indivíduos com maiores escores, e baixa no grupo inferior, que contém os indivíduos com menores escores. Quanto maior a diferença (que pode assumir valores entre -1 e +1) entre a proporção do grupo superior e do grupo inferior, mais o item discrimina os respondentes. Nesta abordagem também é possível visualizar resultados não esperados como, por exemplo, o grupo superior com proporção de acertos menor que o grupo inferior, ou casos em que a discriminação é pequena. Para estes casos, o item foi marcado para ser checado mais atentamente nas análises seguintes.

Foi realizada também, mas de caráter apenas informativo, uma análise de grupos extremos dos distratores, na qual se pode observar se o grupo superior está sendo atraído por um determinado distrator. Neste caso, a proporção usada é de indivíduos que optaram pelo referido distrator e não a proporção de indivíduos que acertaram o item. Este estudo é útil para observar tendências em casos em que a discriminação do item é baixa ou negativa. Esta análise foi encaminhada aos especialistas de cada área.

O uso do critério de 27% (e não de 10%, 25% ou 50%, por exemplo) foi utilizado com base no estudo de coeficientes de correlação de grupos com vários percentuais, no qual pôde ser verificado que o uso de grupos com 27% dos respondentes possui um maior grau de confiança, conforme descrito no trabalho de KELLEY (1939).

Por fim, destaca-se o caráter preliminar destas análises que utilizam a TCT e que não têm como propósito eliminar itens e sim indicar possíveis situações que vão gerar dificuldades de estimação de parâmetros. Por outro lado, certas características dos itens aqui observadas podem servir como base para uma possível retirada de itens nos estudos seguintes.

4.4 - COEFICIENTE ALFA DE CRONBACH

Embora a finalidade deste trabalho seja calibrar itens numa escala já consolidada por diversos estudos e qualquer indicativo de multidimensionalidade ou incoerência na elaboração das avaliações do IQE seriam reveladas de qualquer maneira por meio dos processos envolvidos, foi calculado o Coeficiente Alfa de Cronbach para todas as

provas, a fim de se incluir na análise uma estimativa de confiabilidade dos instrumentos.

O Coeficiente Alfa de Cronbach foi introduzido por CRONBACH (1951) e se propõe a medir a consistência interna de um instrumento de avaliação, determinando em que medida seus itens se correlacionam entre si. É dado por:

$$\alpha = \frac{K}{K-1}\left(1 - \frac{\sum_{i=1}^{K} \sigma_{Y_i}^2}{\sigma_X^2}\right)$$

onde temos:

- K: total de itens;
- σ_X^2: variância total do teste;
- $\sigma_{Y_i}^2$: variância de cada item.

São esperados valores entre 0 e 1 para este coeficiente, no qual, na medida em que os valores se aproximam de 1, mais confiável é o instrumento. As provas do IQE atingiram os seguintes valores:

Tabela 8 - Valores do Coeficiente Alfa de Cronbach

Ano/Disciplina	Valor do coeficiente
5º ano Matemática	0,676
9º ano Matemática	0,758
5º ano Língua Portuguesa	0,734
9º ano Língua Portuguesa	0,786

Fonte: IQE

O software Iteman versão 3.50 foi usado para estimação dos valores acima.

5 - CALIBRAÇÃO DE ITENS E PROFICIÊNCIA

5.1 - MÉTODOS DE ESTIMAÇÃO

KLEIN (2009) define que o processo de calibrar um item (ou um conjunto de itens) significa estimar seus parâmetros. E quando se trata de estimação de parâmetros e proficiência, a TRI prevê basicamente três situações distintas:

1) Quando se conhece os parâmetros dos itens e se deseja estimar a proficiência dos indivíduos;
2) Quando se conhece a proficiência dos indivíduos e se deseja estimar os parâmetros dos itens;
3) Quando se deseja estimar conjuntamente parâmetros dos itens e proficiências dos indivíduos, com todos ou parte dos parâmetros dos itens conhecidos.

Este trabalho se propõe a estimar os parâmetros dos novos itens utilizados na aplicação da prova de ligação e, posteriormente, estimar as proficiências dos indivíduos.

ANDRADE, TAVARES e VALLE (2000) descrevem que, nos trabalhos iniciais com estimação de parâmetros de itens, o método utilizado era a Máxima Verossimilhança Conjunta, que envolve um número muito grande de parâmetros a serem estimados simultaneamente. BOCK & LIEBERMAN (1970) avançaram nesta questão e introduziram o método da Máxima Verossimilhança Marginal (MVM), estimando parâmetros de itens e proficiência em duas etapas separadas: na primeira estimam-se os parâmetros dos itens e na segunda etapa, com os parâmetros já conhecidos, estimam-se as proficiências dos indivíduos. O SAEB usa o método MVM para a calibração dos itens e o método EAP para a estimação das proficiências.

VALLE (2000) ainda aponta que vários autores defendem que cada indivíduo seja exposto a pelo menos 30 itens e que cada item seja respondido por pelo menos 300 indivíduos para que se obtenham estimativas de proficiência com erros de estimativa razoáveis. De fato, este trabalho utiliza instrumentos em que o indivíduo responde menos de 30 itens, porém, em provas que têm como objetivo a estimação de parâmetros dos itens, esta questão não se caracteriza necessariamente um problema.

5.2 - PRINCIPAIS CARACTERÍSTICAS DO BILOG-MG

Para calibração dos itens foi utilizado o *software* BILOG-MG versão 3 que permite a aplicação da TRI para múltiplos grupos. Para seu funcionamento é necessário entrar com todos os requisitos e características da análise que está sendo realizada. Para isso é necessária a criação de programas escritos numa linguagem específica do *software*. Cada ano/disciplina teve um programa, que contempla as seguintes características de cada prova aplicada:

- Os itens eliminados pelo INEP durante a análise da Prova Brasil de 2005 e, eventualmente, os itens eliminados na prova do IQE. Esta informação é fundamental, pois no processo de estimação de parâmetros dos novos itens devem-se descrever exatamente as mesmas circunstâncias encontradas na ocasião da estimação de parâmetros já calibrados.
- O total de itens da análise (itens próprios do IQE + itens já calibrados do INEP);
- O total de grupos da análise (Grupo 1: SAEB de 2005 e Grupo 2: IQE 2012);
- O total de cadernos da análise (26 cadernos da Prova Brasil de 2005 + 10 cadernos da prova IQE 2012);
- Os parâmetros dos itens já calibrados pelo SAEB;
- A composição, item a item, de cada caderno dos dois grupos.

O BILOG-MG divide o processo de estimação em três fases:

- Fase 1, na qual é realizada a checagem da entrada de dados e uma análise clássica dos itens;
- Fase 2, em que a calibração de itens é realizada;
- Fase 3, quando as proficiências são geradas.

Na primeira fase, os dados são submetidos a uma análise clássica, na qual são gerados parâmetros usados inicialmente no processo de estimação. Nesta fase também é gerado um minucioso relatório de entrada dos dados, que apresenta a forma com que o BILOG-MG leu os dados apresentados, como, por exemplo, posição de cada coluna no arquivo de dados, posição de cada item dentro de cada caderno, composição de cada caderno e cada grupo, quais itens possuem parâmetros já conhecidos, ou seja, toda informação da estrutura do instrumento e da bases de dados. Desta maneira a análise pela TRI somente deverá ter sequência quando for confirmado que os dados

foram lidos corretamente e o processo não avança se houver falha na leitura dos dados ou falha na descrição da composição dos cadernos.

Na fase dois, em que os itens são efetivamente calibrados, o BILOG-MG emite um relatório descrevendo como a calibração foi realizada, incluindo diversas informações juntamente com a estimativa dos parâmetros **a**, **b** e **c** calculados para cada item e seus respectivos erros de medida associados. Valores baixos do parâmetro **a** indicam que o item tem pouco poder de discriminação, podendo apresentar problemas de posicionamento nos níveis da escala, etapa que será realizada posteriormente. Por outro lado, valores altos indicam discriminação excessiva, o que, conforme já discutido anteriormente, segrega os indivíduos em basicamente dois grupos.

No parâmetro **b**, a maioria dos itens possui valores entre -3 e +3. Geralmente, em itens em que este parâmetro está muito distante de zero, o erro de estimativa aumenta. É possível, mas não é esperado, itens com o valor deste parâmetro fora do intervalo citado acima.

Finalmente, em itens com o parâmetro **c** elevado podem demonstrar, por exemplo, que os distratores do item possuem algum tipo de incoerência e que, mesmo indivíduos com baixa proficiência, o acertam por eliminação de alternativas, por exemplo. Para os três parâmetros, existe a dificuldade recorrente de se quantificar os valores de alto e baixo.

Nesta fase do processo, podem existir itens que impedem que a calibração avance de maneira precisa, dado que a TRI gera valores de estimativa de parâmetros baseada nas relações entre todos os itens. Neste sentido, um item com problemas de estimação pode influenciar na calibração de outros itens. Existem casos em que o processo é interrompido, pois tal problema impossibilita o avanço dos cálculos e impede que o processo de estimação de parâmetros atinja convergência. Nestes casos, uma revisão das características do item descritas na TCT ou a revisão do item pela equipe pedagógica pode ajudar a localizar a causa do problema. Eventualmente a eliminação do item no processo de calibração torna-se necessária e o processo se inicia novamente sem o referido item.

A decisão de eliminar um determinado item fica sob a responsabilidade da equipe que analisa os dados, baseada na sensibilidade em determinar, caso a caso, se um item apresenta problemas de estimação que podem afetar a calibração de outros itens. Em todos estes casos torna-se fundamental apresentar de maneira clara

os critérios que foram levados em consideração na eliminação de cada item.

Nos parâmetros **a**, **b** e **c** estimados existe também um erro de estimativa associado (erro padrão), que é apresentado também nesta fase. Itens com erros de estimativa elevados, comparados com seus pares, devem ser estudados com cautela. Uma boa prática é observar itens com parâmetros semelhantes e com número de respostas também semelhantes, comparando seus respectivos erros, onde se espera que devam ser semelhantes também. Visto que, como o processo completo envolve várias camadas de cálculos nem sempre triviais, em que a entrada de dados de um processo é a saída de dados de um processo anterior, mesmo erros em pequena escala podem se acumular e comprometer o resultado final. Por este motivo a análise dos valores de erros é importante.

O método da Máxima Verossimilhança Marginal usado estima parâmetros por meio de processos iterativos, calculados em ciclos que usam os valores obtidos do ciclo anterior para calcular uma nova e, assim se espera, mais precisa estimativa. Este processo se repete até que um critério de parada seja atingido. Neste trabalho foi definido como critério uma diferença menor que 0,01 nas estimativas de dois ciclos consecutivos. Aqui arbitrou-se um limite máximo de 60 ciclos. Caso o processo não obtenha convergência, com este limite, pode-se aumenta-lo. No entanto, para itens que possuam problemas, o processo pode se estender indefinidamente. Conjunto de dados que envolvem itens com problemas, como erro de gabarito ou duas alternativas corretas, por exemplo, podem impedir que estes ciclos avancem e o processo é finalizado. Neste caso devem-se revisar os parâmetros estimados e, novamente, os resultados da TCT podem ser usados para localizar e, eventualmente, eliminar o item.

Para a estimação dos parâmetros no BILOG-MG os seguintes comandos foram usados:

>CALIB NQPT=40, CYCLES=60, NEWTON=0, CRIT=0.01, IDIST=0, DIAGNOSIS=0, REFERENCE=1, PLOT=1, NORMAL, TPRIOR, SPRIOR, GPRIOR, NOFLOAT, NOADJUST, PLOT=1, CHI=(15,9);

Por fim, na fase três, a proficiência de cada indivíduo é gerada com base nos parâmetros estimados na fase dois. Estes escores individuais gerados são utilizados para calcular as médias de

proficiência por escola. Os seguintes comandos foram usados para a estimativa da proficiência:

>SCORE METHOD=2, NQPT=20, IDIST=0, Noprint;

5.3 - ANÁLISE INDIVIDUAL

A primeira análise realizada usando o BILOG-MG é uma rodada individual com apenas um grupo: a prova do IQE 2012. Como se trata de apenas uma prova e uma população, aqui foi utilizado o modelo da TRI para um único grupo. O objetivo nesta etapa é aferir a qualidade inicial dos itens e como se comportam as estimativas iniciais dos parâmetros.

Os resultados mostraram que os itens 262 e 243 do 5º ano de Matemática e o item 317 do 9º ano de Matemática possuem problemas nas estimativas dos parâmetros. Entretanto, como se trata de uma análise preliminar, estes itens não serão eliminados nesta etapa.

Com os parâmetros da prova IQE 2012 estimados separadamente, como um único grupo, foi construído um gráfico de dispersão de pontos utilizando-se regressão linear, que envolveu as estimativas dos parâmetros **b** dos itens do SAEB comuns à prova IQE 2012 e à Prova Brasil 2005 para observar o caráter da correlação entre estas duas provas. Nas Figuras 8 a 11 são apresentados os resultados.

Figura 8 - Distribuição das estimativas do parâmetro **b** dos itens comuns entre a prova IQE 2012 e a prova Brasil 2005 do 5º ano de Língua Portuguesa

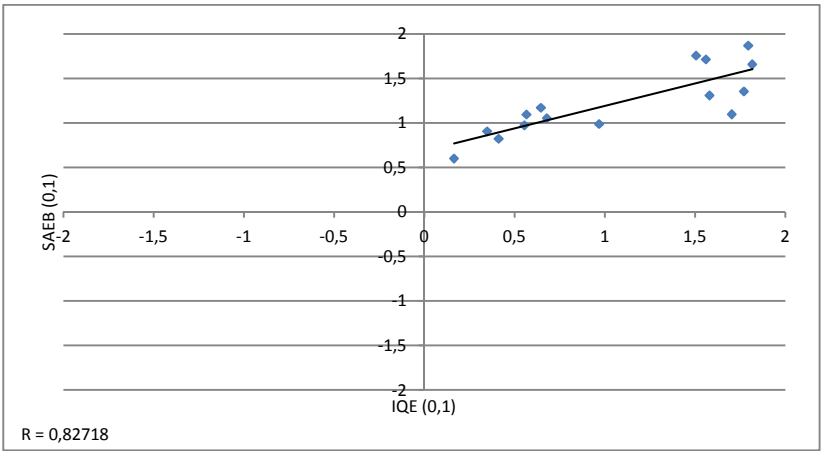


Figura 9 - Distribuição das estimativas do parâmetro **b** dos itens comuns entre a prova IQE 2012 e a prova Brasil 2005 do 9º ano de Língua Portuguesa

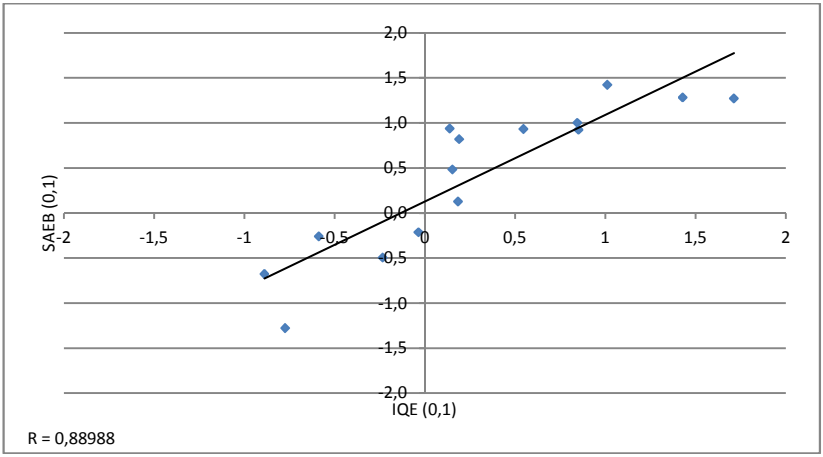


Figura 10 - Distribuição das estimativas do parâmetro **b** dos itens comuns entre a prova IQE 2012 e a prova Brasil 2005 do 5º ano de Matemática

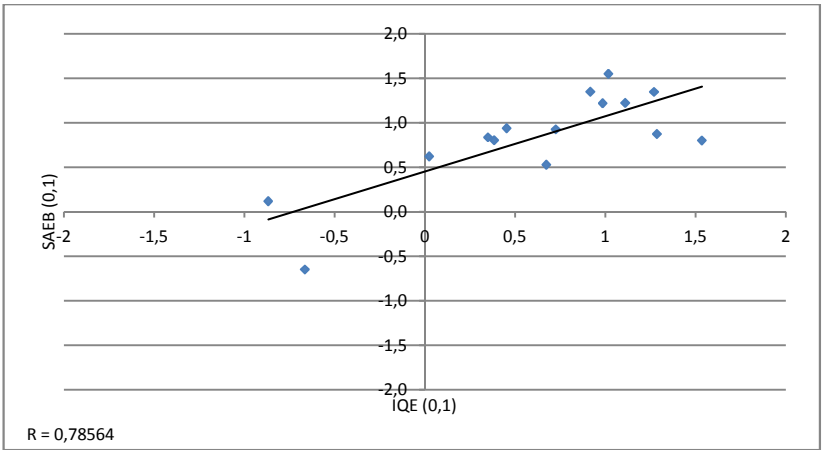
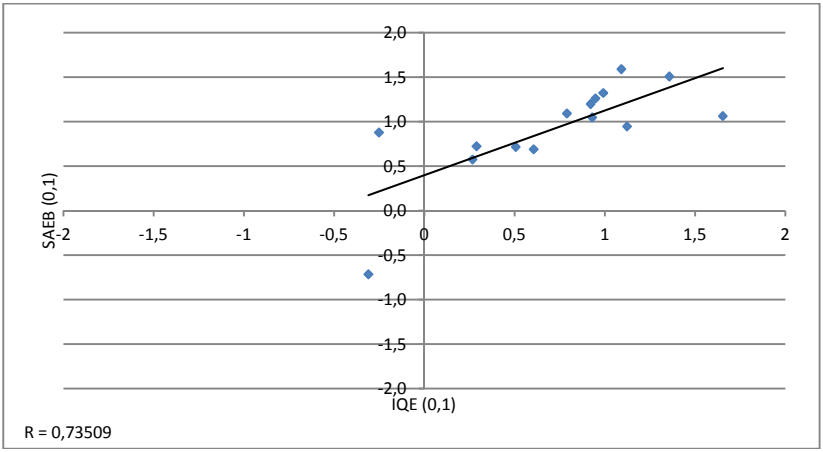


Figura 11 - Distribuição das estimativas do parâmetro **b** dos itens comuns entre a prova IQE 2012 e a prova Brasil 2005 do 9º ano de Matemática



Os gráficos apresentados mostram a linearidade na relação entre as estimativas dos parâmetros, destacada na propriedade da invariância, que será descrita apropriadamente no capítulo 6. O coeficiente de correlação de Pearson entre estas estimativas está representado pela letra R. Este coeficiente mede a correlação linear entre as duas estimativas calculadas dos parâmetros dos itens do SAEB e pode assumir valores entre -1 e 1, onde:

- para R=1, a correlação é perfeita e positiva;
- para R=-1, a correlação é perfeita, porém negativa;
- para R=0 indica que as variáveis não se correlacionam linearmente.

6 - EQUALIZAÇÃO

Conforme dito anteriormente, a TRI abrange uma ampla gama de cenários possíveis quando se deseja medir um determinado traço latente e quando se deseja equalizar resultados numa mesma métrica. Destaca-se novamente que, neste processo, todo trabalho de equalização de itens deve envolver um minucioso estudo de como foi realizado a concepção das análises da escala de referência para uma parametrização consistente e íntegra.

Para deixar mais claro os procedimentos aqui descritos, primeiramente é necessário definir o conceito de equalização, que será bastante usado ao longo das análises.

> "Equalizar significa equiparar, tornar comparável, o que no caso da TRI, significa colocar parâmetros de itens vindos de provas distintas ou habilidades de respondentes de diferentes grupos, na mesma métrica, isto é, numa escala comum, tornando os itens e/ou habilidades comparáveis"
> VALLE (2000)

Para criar instrumentos de avaliação comparáveis com a escala do SAEB é necessário equalizar o instrumento com o qual se trabalha. Na TRI existem duas maneiras de se equalizar instrumentos de avaliação: equalização **via população** e equalização **via itens comuns**.

A equalização via população estabelece que, se uma mesma população é submetida a provas distintas, distribuídas de forma aleatória, os itens destas provas e as proficiências dos indivíduos estarão na mesma métrica. No entanto, a equalização via população somente é possível quando se trata de uma mesma população, calibrando-se as distintas provas num único processo**,** o que não é possível visto que os itens cedidos pelo INEP já foram calibrados e a população não é a mesma.

A equalização via itens comuns é realizada quando duas provas com itens comuns entre elas são aplicadas em populações distintas. Neste caso existem duas abordagens diferentes para se equalizar os instrumentos: equalização ***a posteriori*** e equalização **durante o processo de estimação dos parâmetros dos itens**. Quanto maior a quantidade de itens comuns entre as provas, maior é a precisão da equalização. O melhor cenário possível seria quando todos os itens são

comuns, ou seja, uma mesma prova aplicada a duas populações distintas. Neste caso, usando a metodologia apresentada na seção 5.3, o resultado seria uma correlação elevada com pontos muito mais aderentes à linha de tendência.

Conforme sua denominação, a equalização *a posteriori* via itens comuns é realizada depois de finalizado o processo de calibração dos itens, de forma independente, nas diferentes populações. Então, temos dois conjuntos de parâmetros de itens, cada um em sua própria métrica. O processo consiste em estudar os parâmetros dos itens comuns entre os instrumentos e estabelecer uma relação entre eles de maneira que seja possível colocar os parâmetros de um conjunto A na escala do outro conjunto B. Obtida essa relação, aplica-se o processo para os demais itens. Este processo é mais simples de se realizar, entretanto, geralmente resulta numa precisão menor e acumula erros de estimação de parâmetros.

Um dos métodos que surge naturalmente em equalizações *a posteriori* é o uso de regressões lineares simples. No entanto, existem críticas sobre o uso de regressões por ser um método não simétrico, em que uma regressão de x por y não é igual a uma regressão de y por x. O método mais utilizado na literatura é o Método Média-Desvio. Em ANDRADE, TAVARES e VALLE (2000) podem-se encontrar maiores informações sobre este método.

Nesse trabalho utilizamos a equalização via itens comuns realizado durante a calibração junto com o Modelo para Várias Populações, introduzido por BOCK & ZIMOWSKI (1997), no qual o processo de calibração é feito em conjunto para todas as populações envolvidas. Esse método requer um número menor de itens comuns e é considerado mais preciso em relação à equalização *a posteriori*, principalmente por minimizar a propagação de erros na estimação dos parâmetros.

Em ambas as abordagens é necessário ressaltar que existe o problema de indeterminação da escala, em que temos uma métrica para cada população. Este problema é resolvido definindo-se uma população como referência e todas as outras populações envolvidas no processo serão posicionadas em relação a ela. Este trabalho consiste em situar uma prova aplicada com itens novos na escala do SAEB, a referência utilizada neste trabalho.

Uma característica interessante da TRI, neste contexto, é a propriedade da invariância. Ela estabelece que, dado que o modelo é adequado aos dados, as relações de ordem entre os parâmetros dos itens

aplicados permanecem iguais, mesmo sendo aplicados em populações diferentes e que os parâmetros **a**, **b** e **c** são independentes das proficiências da população estudada, conforme EMBRETSON & HERSCHBERGER (1999). Com base nesta propriedade é que se podem estabelecer relações lineares entre esses itens.

O processo de equalização utilizado nesse trabalho foi dividido em duas rodadas, cada uma delas descrita em detalhes a seguir.

6.1 - PRIMEIRA RODADA DA EQUALIZAÇÃO

O objetivo da primeira rodada é estimar os parâmetros dos itens próprios do IQE na escala (0,1) do SAEB 1997. Para isso, as seguintes informações preliminares são apresentadas como premissas desta etapa:

- os itens fornecidos pelo INEP são da aplicação da Prova Brasil de 2005 e os parâmetros destes itens estão na escala (0,1) do SAEB 1997; serão considerados conhecidos pelo BILOG-MG e não serão estimados novamente;
- a base de dados fornecida é uma amostra aleatória da aplicação da Prova Brasil de 2005;
- o Grupo 1 é o grupo de referência que contém os 169 itens da Prova Brasil de 2005 e o Grupo 2 contém, além dos itens em comum, os itens próprios do IQE de 2012 (40 itens do 5º ano e 50 itens do 9º ano) a serem calibrados e equalizados;
- a Prova Brasil de 2005 do 9º ano/8ª série contém 40 itens comuns do 5º ano/4ª série para que a escala de cada disciplina seja única, conforme descrito anteriormente.

Além das informações acima, a tabela a seguir apresenta a quantidade de itens eliminados à época da calibração da Prova Brasil de 2005 e devidamente informados no programa de cada ano/disciplina:

Tabela 9 - Total de itens eliminados na Prova Brasil de 2005

Série/Ano/Disciplina	Quantidade de itens eliminados
4ª série/5º ano Matemática	5
8ª série/9º ano Matemática	8
4ª série/5º ano Língua Portuguesa	10
8ª série/9º ano Língua Portuguesa	7

Fonte: INEP

No momento da calibração de itens por múltiplos grupos, o processo resulta em valores de parâmetros na escala do grupo de referência, que neste momento é o **SAEB 2005**. No entanto os parâmetros dos itens do INEP estão na escala do (0,1) do SAEB 1997. Torna-se necessário então que, antes da execução do BILOG-MG, seja feita uma transformação dos parâmetros originais dos itens do INEP da escala (0,1) do SAEB 1997 para a escala do (0,1) do SAEB 2005, para que os valores dos parâmetros de todos os itens estejam na mesma métrica, do SAEB 2005. Isto se faz necessário, já que não se tem disponível a base de dados do SAEB 1997.

ANDRADE (2001) define que estas transformações são baseadas nas relações lineares existentes entre os parâmetros de um mesmo item, estimados em escalas diferentes, conforme descrito na seção anterior. ANDRADE argumenta que, dado que o modelo é adequado aos dados, os parâmetros **a** e **b** de um item, com exceção de flutuações amostrais, devem satisfazer as seguintes relações lineares:

$$\mathbf{b}_{G1} = \alpha\mathbf{b}_{G2} + \beta \text{ e } \mathbf{a}_{G1} = (1/\alpha)\mathbf{a}_{G2}$$

Os coeficientes α e β usados acima são obtidos através das seguintes equações:

$$\alpha = \frac{SG1}{SG2}$$

e

$$\beta = M_{G1} - \alpha M_{G2}$$

Onde:

- $\mathbf{b}_{G1}$ e $\mathbf{b}_{G2}$ são os parâmetros de dificuldade dos Grupos 1 e 2;
- $\mathbf{a}_{G1}$ e $\mathbf{a}_{G2}$ são os parâmetros de discriminação dos Grupos 1 e 2;
- S_{G1} e S_{G2} são os desvios-padrão do parâmetro **b** dos Grupos 1 e 2;
- M_{G1} e M_{G2} são as médias do parâmetro **b** dos Grupos 1 e 2.

O parâmetro **c** não é transformado pois trata-se de uma probabilidade invariante e independente de escala.

Como os parâmetros dos itens cedidos pelo INEP estão sendo transformados para a escala (0,1) do SAEB 2005, somente para esta transformação, o Grupo 1 de referência é o SAEB 2005 e o Grupo 2 é o SAEB 1997. As médias e os desvios-padrão da aplicação do SAEB 1997 também foram fornecidos pelo INEP.

Com os parâmetros dos itens do INEP devidamente transformados da escala (0,1) do SAEB 1997 para a escala (0,1) do SAEB 2005, o processo continua até a segunda fase do BILOG-MG, informando ao *software* que os parâmetros dos itens do INEP são conhecidos, e assim estimam-se apenas os parâmetros dos itens próprios do IQE na escala (0,1) do SAEB 2005.

Depois de concluída a segunda fase do BILOG-MG, todos os itens estão com seus respectivos parâmetros na escala (0,1) do SAEB 2005. Neste momento, os parâmetros estimados devem ser analisados e, se for o caso, eliminar da análise itens que comprometem a estimação: itens com parâmetros **a** muito baixo, itens com parâmetros **c** elevados, itens com erros-padrão altos de estimativa e os demais critérios já apresentados anteriormente.

Utilizando estes mesmos critérios, seguem abaixo os itens que foram retirados da análise, juntamente com a respectiva justificativa.

Para o 5º ano de Matemática:

Item 262: este item, além de possuir coeficientes de correlação bisserial acima do esperado para todos os distratores, também possui um coeficiente baixo justamente para a alternativa correta. A calibração deste item resultou num parâmetro **a** muito baixo (0,155) e um parâmetro **b** muito alto (4,372) com um alto erro de estimativa (1,219).

Item 243: embora tenha um coeficiente de correlação bisserial, ainda que positivo, mas próximo de zero na alternativa correta, este item não tem outra característica que justifique uma análise mais profunda na TCT. No entanto, a estimação de seus parâmetros resultou num parâmetro **a** muito baixo (0,136) e um parâmetro **b** ainda maior que o item 262, com 6,494 e um erro de estimativa também maior, no valor de 1,477.

Para o 9º ano de Matemática:

Item 317: além de possuir um valor de correlação bisserial baixo para a alternativa correta, possui um distrator com um alto valor e acima do valor do coeficiente da alternativa correta. Os valores dos seus parâmetros também possuem problemas, com o parâmetro **a** estimado em 0,131 e o parâmetro **b** em 8,534 com um erro de estimativa de 1,935.

Para Língua Portuguesa não houve itens eliminados.

Conforme abordado anteriormente, para considerar um erro de estimativa alto para os parâmetros **a** e **b**, observam-se outros itens com estimativas semelhantes, cujos erros de estimativa devem ser próximos, desde que o número de respostas do item seja razoavelmente aproximado. Grandes diferenças não são esperadas e impactam na estimação de outros itens. No entanto, é importante ressaltar que cada item retirado do processo sempre resulta em perda de informação. A eliminação de itens, embora criteriosa, não deve ser indiscriminada, pois existe também o prejuízo financeiro, visto que a elaboração de itens é um trabalho especializado. Houve situações no decorrer deste trabalho em que foi analisado o impacto da eliminação para se determinar o custo-benefício da sua retirada da análise, comparando calibrações com e sem o referido item. Dado que o custo computacional para realizar estas comparações é muito baixo, é uma opção que foi considerada e utilizada sem ressalvas.

Cada vez que um item é eliminado, deve-se atualizar o programa correspondente e reiniciar todo o processo de estimação, pois com a retirada de itens, as estimativas dos parâmetros de todos os outros itens podem se alterar. Este processo pode ocorrer diversas vezes até que se resulte em processos em que o critério de parada é alcançado e não ocorram problemas de estimação de parâmetros.

Por fim, uma vez revisados os parâmetros estimados, o processo finaliza tendo todos os itens com parâmetros na métrica SAEB 2005. Realiza-se, então, a operação inversa da transformação realizada anteriormente: os parâmetros **a** e **b** dos itens próprios do IQE são transformados de volta para escala SAEB 1997. No entanto, para esta transformação, o SAEB 1997 é o Grupo 1 de referência e o SAEB 2005 é o Grupo 2.

6.2 - SEGUNDA RODADA DA EQUALIZAÇÃO

Como inicialmente os parâmetros dos itens cedidos pelo INEP na escala SAEB 1997 eram conhecidos e os parâmetros dos itens próprios do IQE foram transformados para esta escala, temos, então, todos os itens na escala SAEB 1997.

A segunda rodada é aplicada, porém, apenas com o propósito de gerar escore. Todos as estimativas dos parâmetros dos itens são considerados conhecidos. Nesta rodada o processo avança até a fase três do BILOG-MG, que estima a proficiência dos indivíduos na escala (0,1)

do SAEB 1997. Posteriormente estes escores são transformados para a escala (250,50) do SAEB 1997 utilizando a relação:

$$\theta^{1}_{G2} = \alpha\theta_{G2} + \beta$$

onde θ^{1}_{G2} é o valor da proficiência θ_{G2} na escala do Grupo 1 e α e β constantes fornecidas pelo INEP.

Com as proficiências dos indivíduos na escala final, SAEB 1997 (250,50), já é possível calcular todas as medidas de síntese, incluindo médias por escola, por exemplo. Embora o escore dos alunos não seja o objetivo deste trabalho, as médias por escola são medidas importantes para comparações posteriores com outras provas que se utilizam destes itens.

6.3 - FERRAMENTAS IMPORTANTES

6.3.1 - Função de Informação do Item - FII

Outra característica muito importante do item discutida em ANDRADE, TAVARES e VALLE (2000), e usada em conjunto com a CCI, é a Função de Informação do Item, FII, que representa a quantidade de informação que o mesmo contém em determinado ponto da escala e quanto este contribui para estimação de proficiência. Esta função é dada por:

$$I_i(\theta) = \frac{\left[\frac{d}{d\theta}P_i(\theta)\right]^2}{P_i(\theta)Q_i(\theta)},$$

onde temos:

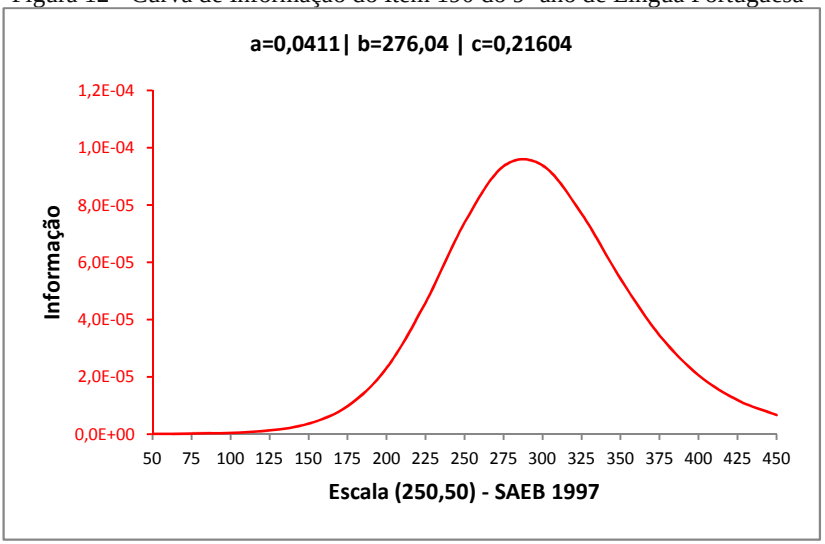
$I_i(\theta)$: a informação fornecida pelo item *i* no nível de proficiência θ
$P_i(\theta)$: $P(U_{ij} = 1|\theta)$
$Q_i(\theta)$: $1 - P_i(\theta)$

Para o Modelo Logístico Unidimensional de três parâmetros, esta função é escrita da seguinte maneira:

$$I_i(\theta) = a_i^2 \frac{Q_i(\theta)}{P_i(\theta)}\left[\frac{P_i(\theta)-c_i}{1-c_i}\right]^2$$

A Figura a seguir mostra, por exemplo, a representação gráfica da informação do item 190 do IQE de 9º ano Língua Portuguesa.

Figura 12 - Curva de Informação do Item 190 do 9º ano de Língua Portuguesa



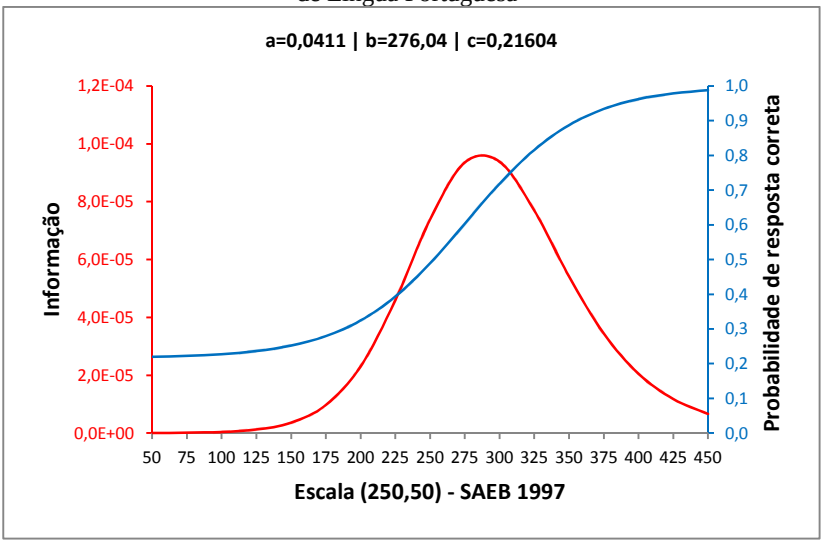
Este gráfico é denominado Curva de Informação do Item - CII, e representa a FII em cada ponto da escala de proficiência. Com esta ferramenta é possível representar e analisar como a relação entre os parâmetros de um item interfere na quantidade de informação estimada no mesmo. E, conforme ANDRADE, TAVARES e VALLE (2000) ponderam, a informação que o item possui aumenta sempre que:

- θ se aproxima do parâmetro $\mathbf{b}_i$;
- parâmetro $\mathbf{c}_i$ se aproxima de zero.

Outro fator que aumenta a quantidade de informação de um item é o parâmetro **a**: quanto maior o valor deste parâmetro, mais informação o item possui. Entretanto, valores muito altos geram uma Curva de Informação com um pico alto, mas apenas numa pequena região da escala de proficiência e, por outro lado, um item com o parâmetro **a** muito baixo gera uma curva justamente ao contrário: baixa e com uma base extensa.

Uma representação gráfica ainda mais rica que também pode ser construída é aquela que reúne a FII e a CCI, como no exemplo da Figura 13:

Figura 13 - Curva de Informação e Curva Característica do Item 190 do 9º ano de Língua Portuguesa



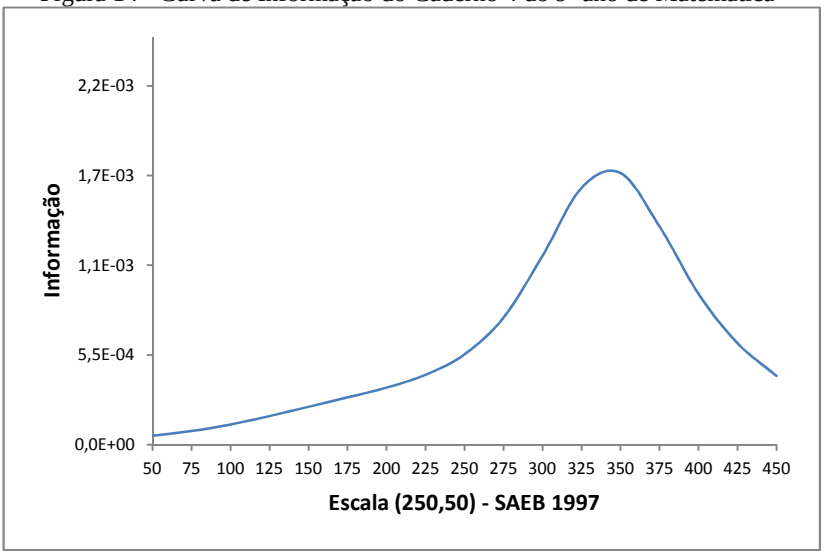
Desta maneira temos duas importantes informações de análise da qualidade de item num mesmo gráfico, simplificando o processo de estudo de parâmetros estimados.

Em conjunto com a FII, pode-se calcular a informação que o teste possui, somando-se as informações de cada item que o compõem, em cada ponto da escala, utilizando:

$$I(\theta) = \sum_{i=1}^{I} I_i(\theta)$$

A Figura 14 representa a quantidade de informação contida no Caderno 4 do 9º ano da prova de Matemática:

Figura 14 - Curva de Informação do Caderno 4 do 9º ano de Matemática

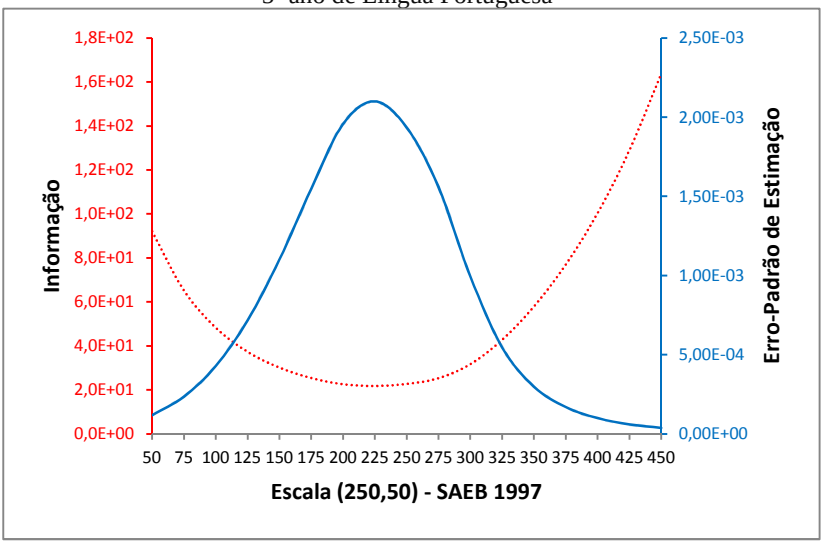


Para cada teste aplicado, pode-se apresentar, em conjunto com a Curva de Informação do Teste, o Erro-Padrão de Estimação, que revela áreas da escala de proficiência em que o teste é mais preciso. O Erro-Padrão de Estimação no método da máxima verossimilhança é dado por:

$$EP(\theta) = \frac{1}{\sqrt{I(\theta)}}$$

Na Figura 15 é mostrada, apenas como exemplo, a Curva de Informação do Caderno 1 do 5º ano da prova de Língua Portuguesa.

Figura 15 - Curva de Informação e Erro-Padrão de Estimação do Caderno 1 do 5º ano de Língua Portuguesa



Como a avaliação do IQE é constituída de dez cadernos de prova em cada ano/disciplina, é possível gerar gráficos que contenham a informação de todos os cadernos de maneira consolidada. Nas Figuras 16 a 19 são apresentadas as curvas de informação de todas as provas:

Figura 16 - Curvas de Informação do 5º ano de Língua Portuguesa

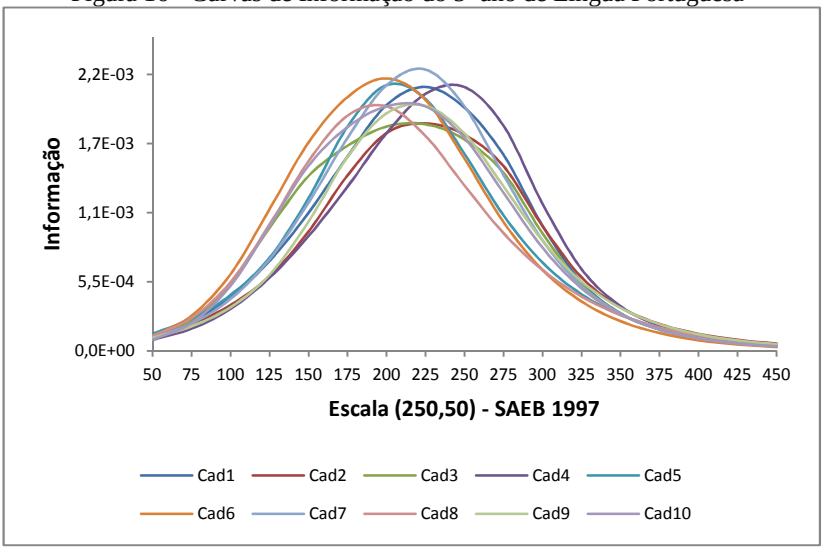


Figura 17 - Curvas de Informação do 5º ano de Matemática

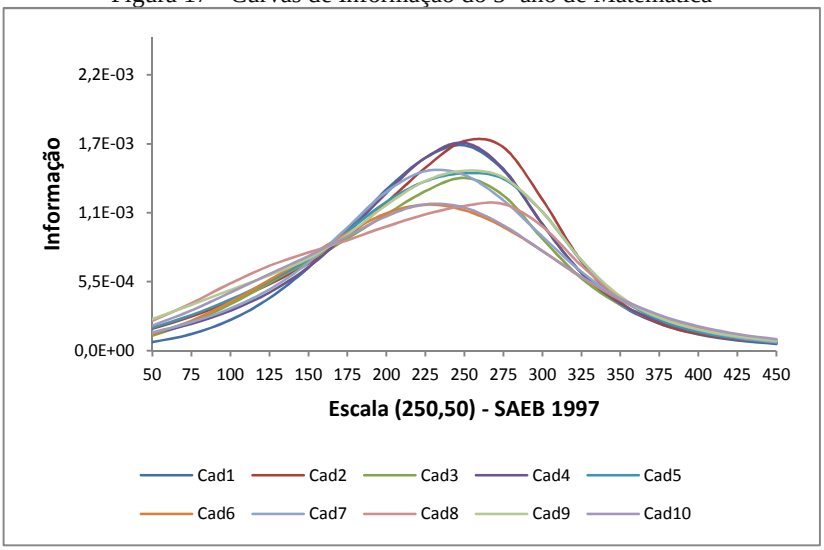


Figura 18 - Curvas de Informação do 9º ano de Língua Portuguesa

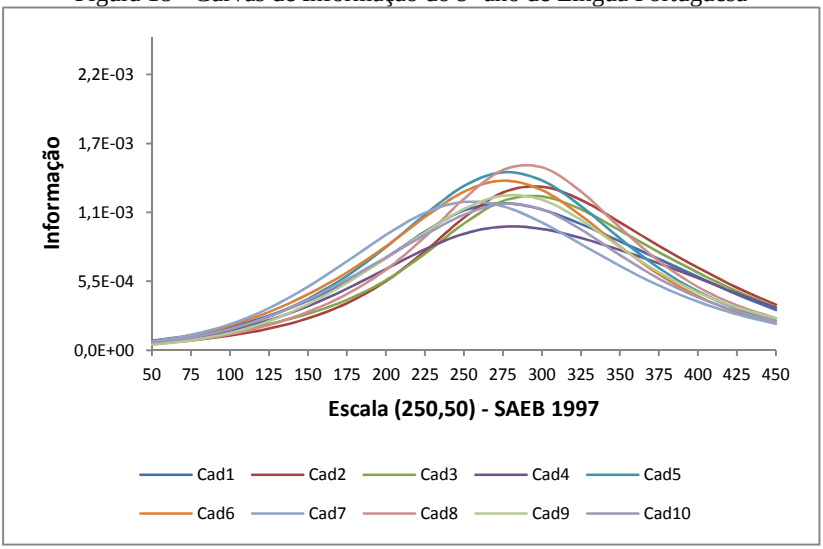
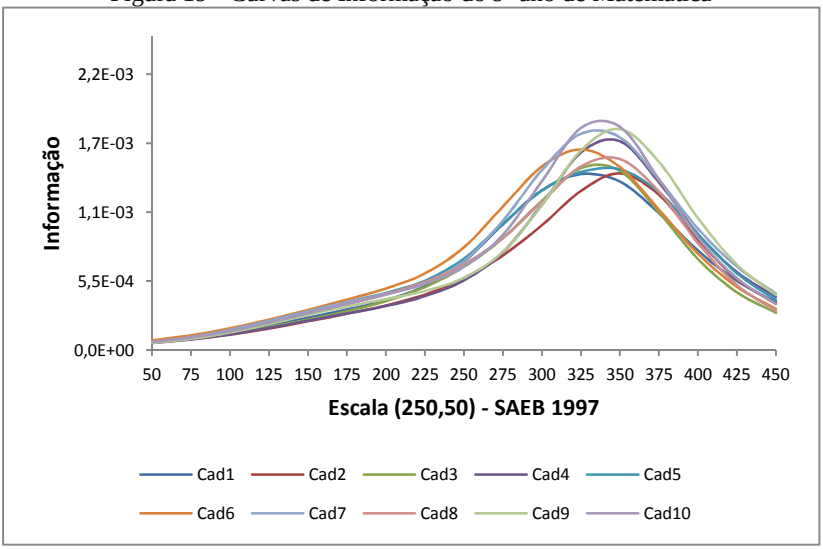


Figura 19 - Curvas de Informação do 9º ano de Matemática



Com a Curva de Informação do Teste pode-se analisar o instrumento por completo e determinar para qual região da escala este teste é mais apropriado e possui menor erro de estimação. A Curva de

Informação de 5º ano e 9º ano de Matemática não contemplam os itens eliminados.

Estas informações e gráficos, além de demonstrar as características importantes dos itens, fazem parte das ferramentas de análise prévia de seleção de itens para que aplicações de prova posteriores sejam mais eficientes. Como por exemplo, com o banco de itens construído, e se deseja avaliar uma população na qual já é esperada, pelo resultado da última Prova Brasil, uma proficiência alta, faz mais sentido optar por itens com bastante informação na parte superior da escala, aumentando a precisão das estimativas. Utilizando os gráficos acima descritos, pode-se realizar então uma seleção de itens mais adequada.

6.4 - MAIS SOBRE A ESCALA DO SAEB E POSICIONAMENTO DE ITENS

Ainda que neste momento do trabalho a estimação dos parâmetros dos itens na escala do SAEB esteja finalizada, a contextualização e apresentação dos resultados obtidos para a equipe pedagógica para a devida discussão e crítica é uma etapa fundamental que deve ser realizada para validação do posicionamento dos itens do IQE na escala do SAEB.

Conforme citada anteriormente, uma grande vantagem do uso da TRI em avaliações educacionais é a construção de escalas de conhecimento interpretáveis, que possibilitam determinar o conteúdo que cada região da escala representa no traço latente. Com uma escala construída e interpretada de maneira adequada, é possível determinar não apenas se um indivíduo possui mais ou menos proficiência que outro, e sim quais habilidades ele possui ou não em relação a outros.

Para tornar uma escala interpretável é necessário dividi-la em níveis. Estes níveis podem ser separados (mas não necessariamente devem) em regiões com a largura de um desvio padrão. Todos os itens que compõem a escala devem ser posicionados em cada um dos níveis definidos. No caso da escala do SAEB o desvio padrão é de 50, mas a escala é dividida em níveis de 25 apenas por conveniência pedagógica.

A escala do SAEB, consolidada nacionalmente ao longo de várias aplicações da Prova Brasil, já possui seus níveis descritos, sendo que, em cada um deles, são expostos pedagogicamente os conhecimentos que se espera que um indivíduo situado na referida posição possua. Como o parâmetro de dificuldade **b** está na mesma escala da proficiência,

também é possível posicionar o item na escala, verificando os conteúdos que se espera que o item exija do indivíduo. Este posicionamento do item na escala de proficiência é de enorme utilidade para a equipe pedagógica que, em conjunto com os parâmetros do item, pode realizar uma seleção aprimorada de itens para compor futuras avaliações.

O posicionamento dos itens foi realizado utilizando-se justamente a equação do ML3 para cada nível da escala do SAEB:

$$P(U_{ij}=1|\,\theta_j) = c_i + (1-c_i)\,\frac{1}{1+e^{-a_i(\theta_j-b_i)}}$$

Desta forma obtém-se uma tabela para cada item, como a que segue, por exemplo:

Tabela 10 - Distribuição da probabilidade de acerto do item 289 do 9º ano de Matemática

Níveis da escala SAEB 1997 (250,50)																
a=0,0306 | **b**=181,4 | **c**=0,097																
50	**75**	**100**	**125**	**150**	**175**	**200**	**225**	**250**	**275**	**300**	**325**	**350**	**375**	**400**	**425**	**450**
0,11	0,13	0,16	0,23	0,34	0,50	**0,68**	0,82	0,91	0,95	0,98	0,99	1	1	1	1	1

Fonte: Elaborado pelo autor

Cada um dos valores acima representa a probabilidade de um indivíduo responder corretamente ao item para cada nível da escala. Este item será posicionado no nível em que a probabilidade de um indivíduo responder corretamente começa a ficar elevada. Dada a dificuldade de definir o conceito, é necessário estabelecer critérios. O critério de posicionamento dos itens na escala de proficiência utilizado neste trabalho foi posicionar o item no nível em que a probabilidade de acerto for maior ou igual a 0,65. O INEP utiliza como critério de posicionamento no nível em que a proporção de acertos for maior ou igual a 0,65. No caso do exemplo acima, o item 289 foi posicionado no nível 200.

6.4.1 - Itens-âncora

Por outro lado, para descrever os níveis da escala, os itens devem apresentar características de posicionamento mais precisas. Portanto, os critérios utilizados foram mais rígidos que o critério de posicionamento:

- $P(U = 1 \mid q = Z) \geq 0{,}65$: A probabilidade de resposta correta do item deve ser maior ou igual a 0,65;
- $P(U = 1 \mid q = Y) \leq 0{,}50$: A probabilidade de resposta correta do nível imediatamente anterior deve ser menor ou igual a 0,50;
- $P(U = 1 \mid q = Z) - P(U = 1 \mid q = Y) \geq 0{,}30$: A diferença entre a probabilidade de resposta correta e a probabilidade de resposta correta do nível imediatamente anterior deve ser maior ou igual a 0,30.

Neste trabalho, os itens que atingem todos os critérios acima são denominados **itens-âncora**, que possuem alta precisão de posicionamento e, portanto, descrevem com mais propriedade cada um dos níveis da escala. Todavia, é possível encontrar outros critérios de identificação destes tipos de itens em outros trabalhos da área. Maiores informações sobre o critério apresentado acima pode ser encontrado em BEATON & ALLEN.

Neste contexto, pode-se destacar que, embora a divisão da escala de proficiência em níveis menores resulta em maior precisão na descrição pedagógica da mesma, esta decisão influencia na quantidade de itens-âncora identificados.

Seguindo a escala do SAEB (250,50) dividida em níveis de um desvio padrão (50), foram identificadas as seguintes quantidades de itens-âncora:

Tabela 11 - Total de itens-âncora identificados por disciplina com níveis de um desvio padrão de largura

Ano	Disciplina	Total de itens equalizados	Número de itens-âncora identificados
5º ano	Língua Portuguesa	40	22
9º ano	Língua Portuguesa	50	12
5º ano	Matemática	40	17
9º ano	Matemática	50	15

Fonte: Elaborado pelo autor

No entanto, para a escala SAEB (250,50) dividida em níveis de meio desvio padrão (25), as seguintes quantidades de itens-âncora foram encontradas:

Tabela 12 - Total de itens-âncora identificados por disciplina com níveis de meio desvio padrão de largura

Ano	Disciplina	Total de itens equalizados	Número de itens-âncora identificados
5º ano	Língua Portuguesa	40	2
9º ano	Língua Portuguesa	50	0
5º ano	Matemática	40	2
9º ano	Matemática	50	3

Fonte: Elaborado pelo autor

Pode-se concluir que, na medida em que a largura do nível na escala aumenta, a quantidade de itens-âncora identificados é maior, mas o detalhamento da escala é menor.

A partir do agrupamento de diversos itens-âncora em seus respectivos níveis, uma equipe de especialistas em cada área avalia o que cada item de cada nível exige do indivíduo e realiza a descrição da proficiência exigida de cada nível. Como a escala do SAEB já possui seus níveis descritos e interpretados, coube aos especialistas do IQE avaliar o posicionamento resultante da aplicação da TRI. Cada item deve corresponder e ser coerente pedagogicamente ao nível no qual foi posicionado, observando outros itens que pertencem à mesma região da escala. Um dos objetivos deste trabalho é propor um enriquecimento da escala do SAEB com descrição pedagógica adicional aos níveis já descritos quando a equipe de especialistas julgar necessário, partindo do princípio que novos itens equalizados podem apresentar novas habilidades necessárias para acertá-lo. Entretanto não se observou nenhuma descrição pedagógica adicional que poderia ser sugerida à escala do SAEB.

Durante esse processo, alguns itens podem ter sua posição na escala criticada pela equipe pedagógica. Por exemplo, a TRI pode posicionar um determinado item no nível 175 e a equipe de especialistas avalia que o item possui características do nível 200. Este tipo de

conflito não é esperado e não deve ser recorrente, pois pode indicar erros de estimação de parâmetros. Enfim, destaca-se novamente o caráter de convergência entre as áreas da Estatística e da Pedagogia para uma equalização eficiente.

Conforme apresentado anteriormente, a questão da natureza acumulativa do conhecimento deve ser observada. Neste sentido, o posicionamento de itens resultantes sempre deve exigir maior proficiência daqueles posicionados abaixo e menor proficiência de itens acima do nível do próprio item. Aqui novamente a análise pedagógica da equipe de especialistas é fundamental.

Outro aspecto interessante no posicionamento de itens em uma escala já interpretada é a possibilidade de enriquecer e refinar a descrição de cada nível da escala de proficiência. Por meio de aplicações de provas em várias ocasiões ao longo do tempo, é possível aprimorar os conteúdos que cada nível exige, não previstos na concepção inicial da escala, e adicioná-los para, então, passar a descrever o nível de maneira mais precisa. Ou pode-se passar a dividir a referida escala em níveis menores e mais detalhados. A escala atual do SAEB pode ser encontrada em INEP (2014).

6.4.2 - Posicionamento final dos itens

A seguir é apresentado o posicionamento final dos itens na escala de proficiência do SAEB, que seguiu a divisão em meio desvio padrão de largura. Como critério de posicionamento, cada item foi posicionado no ponto no qual a probabilidade é próxima de 0,65, marcados na cor azul.

Tabela 13 - Posicionamento dos itens do 5º ano de Língua Portuguesa

	Níveis da Escala																
Código IQE	**50**	**75**	**100**	**125**	**150**	**175**	**200**	**225**	**250**	**275**	**300**	**325**	**350**	**375**	**400**	**425**	**450**
45	0,18	0,19	0,22	0,30	0,48	**0,71**	0,88	0,96	0,99	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00
46	0,06	0,07	0,10	0,15	0,25	0,41	**0,60**	0,77	0,88	0,94	0,97	0,99	0,99	1,00	1,00	1,00	1,00
49	0,29	0,32	0,36	0,41	0,46	0,52	0,58	**0,64**	0,69	0,75	0,80	0,84	0,87	0,90	0,92	0,94	0,95
51	0,29	0,29	0,29	0,30	0,35	0,48	**0,70**	0,89	0,97	0,99	1,00	1,00	1,00	1,00	1,00	1,00	1,00
53	0,38	0,41	0,45	0,50	0,54	0,59	**0,64**	0,69	0,74	0,78	0,82	0,85	0,88	0,90	0,92	0,94	0,95
56	0,42	0,42	0,42	0,42	0,43	0,48	**0,65**	0,86	0,97	0,99	1,00	1,00	1,00	1,00	1,00	1,00	1,00
58	0,09	0,10	0,14	0,30	**0,64**	0,90	0,98	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00
59	0,28	0,33	0,40	0,49	0,59	**0,69**	0,78	0,85	0,90	0,94	0,96	0,98	0,99	0,99	0,99	1,00	1,00
60	0,25	0,36	0,54	**0,74**	0,88	0,95	0,98	0,99	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00
61	0,22	0,22	0,22	0,22	0,23	0,26	0,37	**0,64**	0,88	0,97	0,99	1,00	1,00	1,00	1,00	1,00	1,00
64	0,20	0,20	0,20	0,20	0,21	0,26	0,44	**0,76**	0,94	0,99	1,00	1,00	1,00	1,00	1,00	1,00	1,00
65	0,17	0,17	0,17	0,18	0,18	0,19	0,23	0,38	**0,67**	0,90	0,98	0,99	1,00	1,00	1,00	1,00	1,00
66	0,24	0,26	0,30	0,39	0,55	**0,74**	0,88	0,95	0,98	0,99	1,00	1,00	1,00	1,00	1,00	1,00	1,00
67	0,16	0,19	0,23	0,29	0,38	0,50	**0,62**	0,74	0,83	0,90	0,94	0,97	0,98	0,99	0,99	1,00	1,00
68	0,26	0,26	0,28	0,32	0,42	**0,62**	0,82	0,93	0,98	0,99	1,00	1,00	1,00	1,00	1,00	1,00	1,00
69	0,08	0,09	0,11	0,21	0,45	**0,76**	0,93	0,98	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00
70	0,15	0,15	0,17	0,21	0,34	**0,58**	0,82	0,94	0,98	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00
71	0,38	0,38	0,38	0,38	0,39	0,43	**0,55**	0,75	0,91	0,97	0,99	1,00	1,00	1,00	1,00	1,00	1,00
73	0,18	0,24	0,35	0,54	**0,74**	0,88	0,95	0,98	0,99	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00
84	0,10	0,12	0,16	0,28	**0,51**	0,77	0,92	0,98	0,99	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00
86	0,11	0,11	0,13	0,18	0,28	0,48	**0,71**	0,87	0,95	0,98	0,99	1,00	1,00	1,00	1,00	1,00	1,00
87	0,26	0,26	0,26	0,27	0,31	0,45	**0,69**	0,88	0,97	0,99	1,00	1,00	1,00	1,00	1,00	1,00	1,00
88	0,21	0,21	0,21	0,22	0,24	0,29	0,39	**0,56**	0,74	0,88	0,95	0,98	0,99	1,00	1,00	1,00	1,00
90	0,22	0,22	0,23	0,25	0,28	0,35	0,46	**0,60**	0,75	0,86	0,93	0,96	0,98	0,99	1,00	1,00	1,00
92	0,21	0,22	0,23	0,24	0,28	0,33	0,41	0,53	**0,66**	0,78	0,87	0,93	0,96	0,98	0,99	0,99	1,00
95	0,26	0,26	0,27	0,28	0,32	0,43	**0,62**	0,82	0,93	0,98	0,99	1,00	1,00	1,00	1,00	1,00	1,00
97	0,13	0,16	0,23	0,39	**0,61**	0,81	0,92	0,97	0,99	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00
98	0,14	0,17	0,23	0,37	**0,58**	0,77	0,90	0,96	0,98	0,99	1,00	1,00	1,00	1,00	1,00	1,00	1,00

	Níveis da Escala																
Código IQE	**50**	**75**	**100**	**125**	**150**	**175**	**200**	**225**	**250**	**275**	**300**	**325**	**350**	**375**	**400**	**425**	**450**
99	0,27	0,27	0,27	0,27	0,28	0,29	0,33	0,49	**0,77**	0,94	0,99	1,00	1,00	1,00	1,00	1,00	1,00
100	0,19	0,20	0,21	0,24	0,31	0,46	**0,66**	0,83	0,93	0,97	0,99	1,00	1,00	1,00	1,00	1,00	1,00
101	0,25	0,25	0,25	0,25	0,25	0,27	0,29	0,33	0,42	0,54	**0,69**	0,82	0,90	0,95	0,98	0,99	1,00
103	0,20	0,20	0,20	0,22	0,29	**0,53**	0,84	0,97	0,99	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00
104	0,19	0,19	0,20	0,23	0,33	**0,58**	0,84	0,96	0,99	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00
106	0,10	0,13	0,19	0,32	0,52	**0,72**	0,86	0,94	0,98	0,99	1,00	1,00	1,00	1,00	1,00	1,00	1,00
107	0,19	0,20	0,23	0,29	0,39	0,53	**0,68**	0,81	0,90	0,95	0,98	0,99	0,99	1,00	1,00	1,00	1,00
108	0,21	0,21	0,21	0,22	0,22	0,26	0,35	**0,57**	0,81	0,94	0,98	1,00	1,00	1,00	1,00	1,00	1,00
109	0,26	0,27	0,28	0,30	0,35	0,43	0,56	**0,71**	0,83	0,92	0,96	0,98	0,99	1,00	1,00	1,00	1,00
110	0,13	0,15	0,17	0,21	0,28	0,39	0,53	**0,68**	0,80	0,89	0,94	0,97	0,98	0,99	1,00	1,00	1,00
112	0,21	0,21	0,21	0,21	0,23	0,30	0,51	**0,79**	0,94	0,99	1,00	1,00	1,00	1,00	1,00	1,00	1,00
113	0,18	0,18	0,18	0,18	0,18	0,18	0,18	0,20	0,32	**0,68**	0,94	0,99	1,00	1,00	1,00	1,00	1,00

Fonte: Elaborado pelo autor

Tabela 14 - Posicionamento dos itens do 9º ano de Língua Portuguesa

	Níveis da Escala																
Código IQE	**50**	**75**	**100**	**125**	**150**	**175**	**200**	**225**	**250**	**275**	**300**	**325**	**350**	**375**	**400**	**425**	**450**
134	0,16	0,16	0,16	0,16	0,16	0,16	0,16	0,16	0,16	0,17	0,18	0,22	0,30	0,47	**0,69**	0,86	0,94
135	0,29	0,29	0,29	0,29	0,29	0,30	0,31	0,35	0,41	0,52	**0,66**	0,79	0,89	0,95	0,98	0,99	0,99
136	0,17	0,19	0,21	0,25	0,30	0,37	0,46	0,55	**0,65**	0,74	0,81	0,87	0,91	0,94	0,96	0,98	0,98
137	0,17	0,17	0,17	0,17	0,17	0,17	0,18	0,20	0,24	0,33	0,49	**0,68**	0,84	0,93	0,97	0,99	1,00
138	0,17	0,19	0,22	0,26	0,31	0,38	0,46	0,54	**0,63**	0,72	0,79	0,85	0,89	0,93	0,95	0,97	0,98
139	0,16	0,16	0,17	0,17	0,18	0,19	0,21	0,24	0,29	0,36	0,45	0,56	**0,68**	0,78	0,86	0,91	0,95
140	0,31	0,31	0,31	0,31	0,31	0,32	0,32	0,33	0,35	0,39	0,45	0,55	**0,68**	0,80	0,88	0,94	0,97
141	0,18	0,19	0,20	0,23	0,26	0,31	0,37	0,45	0,55	**0,64**	0,73	0,81	0,87	0,91	0,94	0,96	0,98
143	0,21	0,21	0,22	0,25	0,32	0,43	**0,60**	0,76	0,88	0,95	0,98	0,99	1,00	1,00	1,00	1,00	1,00
144	0,18	0,19	0,21	0,24	0,29	0,36	0,45	0,57	**0,68**	0,78	0,86	0,91	0,95	0,97	0,98	0,99	0,99
145	0,21	0,22	0,22	0,23	0,24	0,26	0,28	0,32	0,37	0,44	0,51	**0,60**	0,69	0,77	0,83	0,88	0,92
146	0,27	0,27	0,27	0,28	0,28	0,28	0,29	0,30	0,32	0,36	0,43	0,52	**0,63**	0,74	0,84	0,90	0,95
147	0,25	0,27	0,29	0,34	0,43	0,55	**0,69**	0,82	0,90	0,95	0,98	0,99	1,00	1,00	1,00	1,00	1,00
149	0,12	0,12	0,13	0,15	0,18	0,23	0,30	0,40	0,51	**0,63**	0,74	0,83	0,89	0,93	0,96	0,98	0,99
150	0,29	0,29	0,29	0,29	0,29	0,31	0,34	0,41	0,54	**0,71**	0,85	0,94	0,98	0,99	1,00	1,00	1,00
151	0,23	0,23	0,23	0,23	0,23	0,23	0,23	0,23	0,24	0,26	0,29	0,37	0,49	**0,65**	0,80	0,90	0,95
153	0,11	0,11	0,13	0,15	0,20	0,28	0,41	**0,58**	0,74	0,86	0,93	0,96	0,98	0,99	1,00	1,00	1,00
154	0,26	0,26	0,26	0,27	0,27	0,29	0,33	0,42	**0,57**	0,74	0,88	0,95	0,98	0,99	1,00	1,00	1,00
157	0,26	0,26	0,26	0,26	0,27	0,28	0,29	0,32	0,38	0,46	0,58	**0,71**	0,82	0,90	0,94	0,97	0,99
158	0,20	0,21	0,21	0,22	0,23	0,25	0,29	0,36	0,46	0,58	**0,71**	0,82	0,89	0,94	0,97	0,98	0,99
160	0,24	0,24	0,26	0,28	0,33	0,43	**0,58**	0,74	0,86	0,93	0,97	0,99	0,99	1,00	1,00	1,00	1,00
161	0,17	0,18	0,19	0,21	0,25	0,29	0,36	0,44	0,53	**0,63**	0,72	0,80	0,86	0,91	0,94	0,96	0,97
162	0,12	0,12	0,13	0,15	0,19	0,24	0,31	0,41	0,53	**0,66**	0,77	0,85	0,91	0,95	0,97	0,98	0,99
163	0,19	0,19	0,19	0,19	0,19	0,21	0,26	0,40	**0,63**	0,84	0,95	0,98	1,00	1,00	1,00	1,00	1,00
166	0,24	0,24	0,24	0,24	0,24	0,25	0,27	0,31	0,37	0,46	0,58	**0,72**	0,83	0,90	0,95	0,97	0,99
167	0,17	0,17	0,17	0,18	0,18	0,18	0,19	0,22	0,29	0,45	**0,68**	0,86	0,95	0,98	0,99	1,00	1,00
168	0,24	0,25	0,25	0,27	0,29	0,34	0,42	0,53	**0,67**	0,79	0,88	0,94	0,97	0,98	0,99	1,00	1,00
169	0,18	0,18	0,18	0,19	0,20	0,21	0,24	0,28	0,35	0,45	0,56	**0,68**	0,78	0,86	0,92	0,95	0,97

	Níveis da Escala																
Código IQE	**50**	**75**	**100**	**125**	**150**	**175**	**200**	**225**	**250**	**275**	**300**	**325**	**350**	**375**	**400**	**425**	**450**
171	0,16	0,17	0,19	0,24	0,34	0,50	**0,69**	0,84	0,93	0,97	0,99	0,99	1,00	1,00	1,00	1,00	1,00
172	0,12	0,12	0,14	0,17	0,22	0,32	0,45	**0,61**	0,75	0,86	0,93	0,96	0,98	0,99	1,00	1,00	1,00
173	0,17	0,19	0,22	0,27	0,35	0,46	**0,59**	0,72	0,82	0,89	0,94	0,96	0,98	0,99	0,99	1,00	1,00
174	0,32	0,32	0,32	0,32	0,33	0,35	0,39	0,49	**0,64**	0,81	0,91	0,97	0,99	1,00	1,00	1,00	1,00
175	0,11	0,11	0,12	0,12	0,15	0,20	0,30	0,48	**0,69**	0,85	0,93	0,97	0,99	1,00	1,00	1,00	1,00
176	0,13	0,13	0,13	0,14	0,14	0,15	0,17	0,21	0,27	0,36	0,49	**0,64**	0,77	0,87	0,93	0,96	0,98
177	0,23	0,23	0,23	0,24	0,24	0,24	0,25	0,27	0,29	0,34	0,40	0,50	**0,61**	0,72	0,82	0,89	0,94
178	0,27	0,27	0,27	0,27	0,27	0,27	0,27	0,28	0,28	0,29	0,31	0,34	0,40	0,49	**0,61**	0,73	0,84
180	0,14	0,15	0,16	0,19	0,22	0,27	0,34	0,44	0,54	**0,65**	0,75	0,83	0,89	0,93	0,96	0,97	0,98
181	0,12	0,12	0,13	0,14	0,16	0,20	0,27	0,37	0,51	**0,65**	0,78	0,87	0,93	0,96	0,98	0,99	0,99
182	0,18	0,20	0,22	0,26	0,31	0,36	0,43	0,51	**0,60**	0,68	0,75	0,81	0,86	0,90	0,93	0,95	0,97
183	0,30	0,32	0,34	0,37	0,41	0,46	0,51	0,56	**0,62**	0,68	0,73	0,78	0,83	0,86	0,90	0,92	0,94
184	0,16	0,18	0,21	0,27	0,37	0,53	**0,70**	0,83	0,92	0,96	0,98	0,99	1,00	1,00	1,00	1,00	1,00
185	0,28	0,28	0,28	0,28	0,30	0,32	0,37	0,46	**0,59**	0,74	0,86	0,93	0,97	0,99	0,99	1,00	1,00
186	0,12	0,14	0,16	0,19	0,25	0,33	0,42	0,54	**0,65**	0,76	0,84	0,90	0,94	0,96	0,98	0,99	0,99
187	0,19	0,20	0,21	0,22	0,24	0,26	0,28	0,31	0,34	0,39	0,43	0,48	0,54	0,60	**0,66**	0,71	0,76
188	0,33	0,33	0,33	0,33	0,33	0,35	0,38	0,47	**0,64**	0,81	0,93	0,97	0,99	1,00	1,00	1,00	1,00
189	0,19	0,19	0,20	0,20	0,22	0,26	0,33	0,46	**0,63**	0,79	0,90	0,95	0,98	0,99	1,00	1,00	1,00
190	0,22	0,22	0,22	0,22	0,22	0,23	0,25	0,30	0,42	**0,60**	0,79	0,91	0,96	0,99	1,00	1,00	1,00
191	0,29	0,29	0,29	0,29	0,30	0,30	0,30	0,31	0,32	0,35	0,38	0,43	0,50	0,59	**0,69**	0,78	0,85
192	0,14	0,16	0,20	0,26	0,36	0,51	**0,66**	0,79	0,88	0,94	0,97	0,98	0,99	1,00	1,00	1,00	1,00
193	0,29	0,29	0,29	0,29	0,29	0,29	0,30	0,32	0,37	0,49	**0,67**	0,84	0,93	0,98	0,99	1,00	1,00

Fonte: Elaborado pelo autor

Tabela 15 - Posicionamento dos itens do 5º ano de Matemática

	Níveis da Escala																
Código IQE	**50**	**75**	**100**	**125**	**150**	**175**	**200**	**225**	**250**	**275**	**300**	**325**	**350**	**375**	**400**	**425**	**450**
358	0,20	0,21	0,22	0,23	0,25	0,28	0,31	0,36	0,41	0,48	0,55	**0,62**	0,69	0,76	0,81	0,86	0,90
263	0,17	0,17	0,17	0,17	0,17	0,18	0,18	0,22	0,41	**0,76**	0,95	0,99	1,00	1,00	1,00	1,00	1,00
243	Eliminado																
255	0,15	0,15	0,15	0,15	0,15	0,17	0,22	0,38	**0,66**	0,88	0,97	0,99	1,00	1,00	1,00	1,00	1,00
239	0,32	0,32	0,33	0,35	0,39	0,48	**0,63**	0,80	0,91	0,97	0,99	1,00	1,00	1,00	1,00	1,00	1,00
225	0,26	0,28	0,31	0,36	0,43	0,53	**0,63**	0,73	0,82	0,88	0,93	0,96	0,97	0,98	0,99	0,99	1,00
231	0,13	0,13	0,13	0,13	0,15	0,20	0,35	**0,61**	0,84	0,95	0,99	1,00	1,00	1,00	1,00	1,00	1,00
267	0,14	0,16	0,21	0,33	0,52	**0,73**	0,88	0,95	0,98	0,99	1,00	1,00	1,00	1,00	1,00	1,00	1,00
264	0,15	0,18	0,24	0,35	0,52	**0,70**	0,84	0,92	0,96	0,98	0,99	1,00	1,00	1,00	1,00	1,00	1,00
238	0,14	0,15	0,19	0,25	0,35	0,51	**0,68**	0,82	0,91	0,96	0,98	0,99	1,00	1,00	1,00	1,00	1,00
262	Eliminado																
244	0,26	0,26	0,26	0,27	0,30	0,39	0,56	**0,77**	0,91	0,97	0,99	1,00	1,00	1,00	1,00	1,00	1,00
210	0,16	0,17	0,19	0,22	0,28	0,36	0,47	**0,60**	0,73	0,83	0,90	0,94	0,97	0,98	0,99	0,99	1,00
228	0,15	0,16	0,17	0,20	0,25	0,34	0,46	**0,61**	0,74	0,85	0,92	0,96	0,98	0,99	0,99	1,00	1,00
241	0,36	0,36	0,36	0,36	0,36	0,36	0,37	0,40	0,49	**0,66**	0,84	0,95	0,98	1,00	1,00	1,00	1,00
256	0,25	0,25	0,25	0,25	0,26	0,30	0,38	0,53	**0,73**	0,88	0,95	0,98	0,99	1,00	1,00	1,00	1,00
229	0,25	0,25	0,26	0,26	0,26	0,27	0,27	0,29	0,31	0,35	0,42	0,50	**0,60**	0,70	0,80	0,87	0,92
359	0,20	0,30	0,46	**0,66**	0,83	0,93	0,97	0,99	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00
266	0,15	0,22	0,34	0,50	**0,68**	0,82	0,91	0,96	0,98	0,99	1,00	1,00	1,00	1,00	1,00	1,00	1,00
257	0,11	0,11	0,11	0,11	0,11	0,11	0,11	0,12	0,18	0,42	**0,79**	0,96	0,99	1,00	1,00	1,00	1,00
242	0,37	0,47	0,58	**0,69**	0,78	0,86	0,91	0,94	0,96	0,98	0,99	0,99	1,00	1,00	1,00	1,00	1,00
224	0,25	0,31	0,39	0,49	0,59	**0,68**	0,77	0,84	0,89	0,93	0,95	0,97	0,98	0,99	0,99	0,99	1,00
245	0,14	0,18	0,27	0,41	**0,60**	0,78	0,89	0,95	0,98	0,99	1,00	1,00	1,00	1,00	1,00	1,00	1,00
265	0,10	0,10	0,11	0,11	0,12	0,15	0,24	0,45	**0,71**	0,89	0,97	0,99	1,00	1,00	1,00	1,00	1,00
360	0,35	0,39	0,44	0,49	0,55	0,60	**0,66**	0,71	0,76	0,81	0,84	0,88	0,90	0,92	0,94	0,95	0,97
250	0,25	0,26	0,27	0,30	0,33	0,39	0,47	0,56	**0,66**	0,76	0,83	0,89	0,93	0,96	0,98	0,99	0,99
270	0,17	0,22	0,34	0,54	**0,75**	0,90	0,96	0,99	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00
240	0,31	0,34	0,39	0,46	0,56	**0,67**	0,77	0,86	0,91	0,95	0,97	0,98	0,99	1,00	1,00	1,00	1,00

	Níveis da Escala																
Código IQE	**50**	**75**	**100**	**125**	**150**	**175**	**200**	**225**	**250**	**275**	**300**	**325**	**350**	**375**	**400**	**425**	**450**
268	0,14	0,20	0,33	0,53	**0,74**	0,88	0,95	0,98	0,99	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00
235	0,20	0,20	0,21	0,22	0,23	0,26	0,29	0,35	0,41	0,50	0,59	**0,69**	0,77	0,84	0,89	0,93	0,95
246	0,35	0,35	0,36	0,37	0,40	0,47	**0,59**	0,74	0,87	0,94	0,98	0,99	1,00	1,00	1,00	1,00	1,00
259	0,32	0,32	0,32	0,32	0,32	0,32	0,33	0,35	0,41	0,53	**0,71**	0,86	0,95	0,98	0,99	1,00	1,00
258	0,10	0,10	0,11	0,12	0,15	0,24	0,42	**0,66**	0,85	0,94	0,98	0,99	1,00	1,00	1,00	1,00	1,00
226	0,27	0,27	0,27	0,28	0,28	0,30	0,32	0,36	0,44	0,56	**0,69**	0,81	0,90	0,95	0,97	0,99	0,99
234	0,29	0,29	0,29	0,30	0,33	0,38	0,50	**0,66**	0,82	0,92	0,97	0,99	1,00	1,00	1,00	1,00	1,00
260	0,16	0,17	0,17	0,18	0,19	0,20	0,23	0,27	0,34	0,42	0,53	**0,63**	0,74	0,82	0,89	0,93	0,96
247	0,25	0,26	0,26	0,27	0,31	0,40	0,57	**0,76**	0,90	0,96	0,99	1,00	1,00	1,00	1,00	1,00	1,00
212	0,18	0,25	0,40	**0,59**	0,78	0,89	0,95	0,98	0,99	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00
237	0,23	0,23	0,24	0,25	0,27	0,34	0,48	**0,68**	0,85	0,94	0,98	0,99	1,00	1,00	1,00	1,00	1,00
269	0,44	**0,59**	0,73	0,85	0,92	0,96	0,98	0,99	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00

Fonte: Elaborado pelo autor

Tabela 16 - Posicionamento dos itens do 9º ano de Matemática

	Níveis da Escala																
Código IQE	**50**	**75**	**100**	**125**	**150**	**175**	**200**	**225**	**250**	**275**	**300**	**325**	**350**	**375**	**400**	**425**	**450**
271	0,16	0,19	0,24	0,30	0,39	0,49	**0,61**	0,71	0,80	0,87	0,92	0,95	0,97	0,98	0,99	0,99	1,00
273	0,19	0,22	0,28	0,35	0,44	0,54	**0,65**	0,74	0,82	0,87	0,92	0,95	0,97	0,98	0,99	0,99	0,99
274	0,27	0,33	0,42	0,53	**0,65**	0,76	0,84	0,90	0,94	0,97	0,98	0,99	0,99	1,00	1,00	1,00	1,00
275	0,28	0,28	0,28	0,28	0,28	0,28	0,29	0,31	0,35	0,44	**0,59**	0,76	0,88	0,95	0,98	0,99	1,00
276	0,10	0,10	0,10	0,10	0,10	0,10	0,10	0,10	0,10	0,11	0,14	0,19	0,29	0,46	**0,66**	0,82	0,92
277	0,11	0,11	0,12	0,13	0,14	0,16	0,18	0,21	0,25	0,29	0,35	0,41	0,48	0,55	**0,62**	0,68	0,74
278	0,22	0,22	0,23	0,23	0,23	0,23	0,24	0,26	0,28	0,31	0,37	0,44	0,54	**0,64**	0,75	0,83	0,89
279	0,16	0,16	0,17	0,18	0,20	0,23	0,26	0,32	0,39	0,48	0,58	**0,68**	0,76	0,83	0,89	0,93	0,95
280	0,16	0,16	0,17	0,19	0,21	0,25	0,29	0,35	0,42	0,51	**0,60**	0,69	0,77	0,83	0,88	0,92	0,95
281	0,23	0,24	0,26	0,27	0,30	0,32	0,36	0,40	0,45	0,51	0,56	**0,63**	0,69	0,74	0,79	0,84	0,87
282	0,27	0,27	0,27	0,27	0,28	0,29	0,30	0,32	0,36	0,42	0,51	**0,61**	0,72	0,81	0,88	0,93	0,96
283	0,18	0,18	0,18	0,18	0,18	0,18	0,19	0,21	0,23	0,28	0,36	0,46	**0,59**	0,72	0,83	0,90	0,95
286	0,19	0,19	0,20	0,21	0,22	0,23	0,25	0,28	0,33	0,39	0,46	0,55	**0,64**	0,73	0,80	0,86	0,90
287	0,21	0,21	0,21	0,21	0,21	0,21	0,22	0,23	0,27	0,36	0,51	**0,69**	0,84	0,93	0,97	0,99	1,00
288	0,25	0,25	0,25	0,25	0,25	0,25	0,25	0,25	0,25	0,26	0,28	0,40	**0,69**	0,91	0,98	1,00	1,00
289	0,11	0,13	0,16	0,23	0,34	0,50	**0,68**	0,82	0,91	0,95	0,98	0,99	1,00	1,00	1,00	1,00	1,00
290	0,30	0,34	0,38	0,42	0,47	0,52	0,57	**0,63**	0,68	0,73	0,77	0,81	0,84	0,87	0,90	0,92	0,94
292	0,08	0,08	0,08	0,08	0,08	0,08	0,08	0,08	0,08	0,08	0,09	0,14	0,34	**0,70**	0,92	0,98	1,00
294	0,14	0,16	0,19	0,23	0,30	0,38	0,47	0,57	**0,67**	0,76	0,83	0,89	0,92	0,95	0,97	0,98	0,99
296	0,08	0,08	0,08	0,08	0,08	0,08	0,08	0,09	0,09	0,13	0,21	0,41	**0,67**	0,87	0,96	0,99	1,00
300	0,16	0,16	0,16	0,16	0,16	0,16	0,16	0,16	0,16	0,19	0,30	**0,56**	0,84	0,96	0,99	1,00	1,00
301	0,17	0,18	0,18	0,19	0,20	0,23	0,26	0,32	0,40	0,50	**0,62**	0,73	0,82	0,88	0,93	0,96	0,98
302	0,18	0,22	0,29	0,39	0,52	**0,66**	0,78	0,87	0,93	0,96	0,98	0,99	0,99	1,00	1,00	1,00	1,00
304	0,22	0,22	0,22	0,22	0,22	0,22	0,22	0,22	0,22	0,25	0,36	**0,63**	0,87	0,97	0,99	1,00	1,00
305	0,16	0,16	0,16	0,16	0,16	0,16	0,16	0,16	0,16	0,17	0,22	0,36	**0,64**	0,87	0,96	0,99	1,00
307	0,31	0,31	0,31	0,31	0,31	0,31	0,31	0,32	0,32	0,35	0,44	**0,62**	0,82	0,94	0,98	0,99	1,00
309	0,38	0,38	0,38	0,38	0,38	0,38	0,38	0,39	0,40	0,41	0,45	0,51	**0,60**	0,72	0,83	0,91	0,95
310	0,16	0,17	0,18	0,21	0,26	0,34	0,47	**0,62**	0,76	0,86	0,93	0,96	0,98	0,99	1,00	1,00	1,00

	Níveis da Escala																
Código IQE	**50**	**75**	**100**	**125**	**150**	**175**	**200**	**225**	**250**	**275**	**300**	**325**	**350**	**375**	**400**	**425**	**450**
311	0,19	0,19	0,19	0,19	0,19	0,19	0,19	0,19	0,19	0,20	0,21	0,25	0,35	0,53	**0,73**	0,88	0,95
313	0,34	0,34	0,35	0,37	0,40	0,44	0,50	0,57	**0,66**	0,74	0,82	0,88	0,92	0,95	0,97	0,98	0,99
314	0,19	0,19	0,19	0,19	0,19	0,19	0,19	0,20	0,24	0,41	**0,73**	0,93	0,99	1,00	1,00	1,00	1,00
315	0,22	0,23	0,26	0,32	0,43	**0,61**	0,78	0,90	0,96	0,98	0,99	1,00	1,00	1,00	1,00	1,00	1,00
316	0,12	0,14	0,17	0,21	0,26	0,32	0,40	0,49	0,59	**0,68**	0,76	0,83	0,88	0,91	0,94	0,96	0,97
317	Eliminado																
318	0,20	0,20	0,20	0,21	0,21	0,22	0,23	0,24	0,26	0,29	0,33	0,39	0,46	0,53	**0,62**	0,70	0,77
319	0,16	0,16	0,16	0,16	0,16	0,16	0,16	0,16	0,17	0,17	0,21	0,35	**0,66**	0,90	0,98	1,00	1,00
320	0,19	0,20	0,20	0,22	0,24	0,27	0,32	0,39	0,49	**0,60**	0,71	0,81	0,88	0,93	0,96	0,97	0,99
322	0,22	0,22	0,22	0,22	0,22	0,23	0,26	0,31	0,42	**0,58**	0,75	0,88	0,95	0,98	0,99	1,00	1,00
325	0,21	0,21	0,22	0,22	0,24	0,27	0,33	0,43	0,55	**0,69**	0,81	0,89	0,94	0,97	0,99	0,99	1,00
326	0,15	0,15	0,16	0,17	0,20	0,25	0,33	0,45	**0,60**	0,74	0,84	0,91	0,95	0,98	0,99	0,99	1,00
327	0,28	0,28	0,28	0,28	0,28	0,28	0,29	0,31	0,38	0,53	**0,74**	0,90	0,97	0,99	1,00	1,00	1,00
328	0,24	0,24	0,24	0,24	0,24	0,24	0,24	0,24	0,25	0,28	0,36	0,52	**0,74**	0,90	0,96	0,99	1,00
329	0,23	0,23	0,23	0,24	0,24	0,25	0,26	0,29	0,33	0,40	0,48	**0,59**	0,70	0,80	0,88	0,93	0,96
331	0,21	0,21	0,21	0,21	0,21	0,21	0,21	0,21	0,21	0,22	0,24	0,28	0,35	0,47	**0,63**	0,78	0,88
334	0,29	0,29	0,29	0,29	0,29	0,29	0,30	0,31	0,32	0,36	0,44	0,56	**0,71**	0,84	0,92	0,96	0,98
337	0,11	0,13	0,16	0,21	0,30	0,43	0,58	**0,72**	0,83	0,90	0,95	0,97	0,99	0,99	1,00	1,00	1,00
338	0,18	0,18	0,18	0,19	0,20	0,21	0,23	0,27	0,32	0,39	0,49	**0,59**	0,70	0,79	0,86	0,91	0,95
339	0,11	0,13	0,17	0,26	0,41	**0,61**	0,79	0,90	0,96	0,98	0,99	1,00	1,00	1,00	1,00	1,00	1,00
340	0,20	0,23	0,28	0,34	0,43	0,54	**0,65**	0,75	0,83	0,89	0,93	0,96	0,97	0,98	0,99	0,99	1,00
361	0,11	0,12	0,15	0,19	0,26	0,36	0,49	**0,63**	0,75	0,85	0,91	0,95	0,97	0,98	0,99	1,00	1,00

Fonte: Elaborado pelo autor

6.5 - SUGESTÃO DE PESQUISA NA ÁREA: COMPUTERIZED ADAPTIVE TESTING

A TRI, permite, respeitando algumas premissas, a aplicação de um instrumento de avaliação muito interessante: o CAT, Computerized adaptive testing. Segundo THISSEN e MISLEVY (2000), trata-se de uma prova na qual os itens são apresentados individualmente, de forma

que, com a geração imediata de escore, o teste em si se nivele com a habilidade do respondente.

Foi descrito na seção 6.3.1 que, na TRI, todo item possui uma determinada carga de informação, e que, na medida em que esta carga aumenta, a precisão da estimativa da proficiência se eleva. E também que o item possui maior informação na região da estimativa do parâmetro **b**. Neste contexto, o CAT se propõe a apresentar inicialmente um item com dificuldade média (cuja proficiência necessária para acertá-lo seja ao redor do ponto zero da escala 0,1). Caso o respondente acerte o item, é lhe apresentado um item com dificuldade maior e, caso ele erre, o computador lhe apresenta um item com menor dificuldade. A cada etapa, a estimativa de proficiência do respondente é estimada com base nos itens que o mesmo respondeu até o momento, gerando diferentes testes para diferentes proficiências. O processo se repete até que algum critério pré-definido seja alcançado, como, por exemplo, um determinado valor de erro de estimativa de proficiência seja obtido. Seguindo este processo, somente são apresentados ao respondente itens com alta carga de informação na região da proficiência do mesmo, gerando alta precisão na estimativa. Além disso, o CAT mostra-se mais preciso também para respondentes com proficiências situadas nos extremos da escala, em que, geralmente em provas tradicionais (com o número fixo de itens) a estimativa carrega uma estimativa de erro maior. Naturalmente, como envolve algoritmos e decisões e estimativas em tempo real, o CAT é realizado de forma exclusivamente eletrônica.

Embora este sistema permita a realização de testes com menor quantidade de itens e, portanto, permite também a redução da exposição desnecessária dos mesmos, o algoritmo que apresenta os itens deverá levar em consideração que, conforme os testes forem aplicados, em algum momento haverá itens mais respondidos que outros. Desta forma, o controle de exposição de itens deve ser cuidadosamente planejado para evitar a fadiga do mesmo.

O grande desafio de usar o CAT é que, tratando-se de estimativas calculadas durante a aplicação da prova, somente o emprego de itens previamente calibrados é admitido, dado que a calibração requer uma estrutura de prova previamente construída e exposição semelhante de itens aos respondentes, a calibração e equalização de itens durante a aplicação torna-se inviável. Portanto, todos os itens devem ser pré-testados numa etapa anterior com um número razoável de respondentes.

Dada estas premissas, futuros pesquisadores podem abordar, por exemplo, as possíveis diferenças entre equalização em dispositivos

eletrônicos (e a influencia de suas particularidades técnicas como tamanho de tela, presença ou ausência de teclado, etc) e equalização em avaliações realizadas no papel.

7 - CONCLUSÃO

A teoria utilizada neste trabalho é fruto de estudos realizados por diversos autores ao longo das últimas décadas e já possui sua credibilidade atestada, sendo utilizada, particularmente na área da Educação, por diversos sistemas ao redor do mundo.

Os resultados produzidos pela TRI, quando aplicados e utilizados de maneira correta, constituem uma poderosa ferramenta de diagnóstico.

Há de se registrar, porém, que a contextualização dos dados e a abstração dos processos aqui utilizados não são triviais, muito embora as ferramentas disponíveis para realização dos cálculos mais complexos estejam consolidadas de maneira automatizada em *softwares*. Neste sentido, destaco o trabalho realizado há décadas pelo INEP, no seu esforço pelo uso destes resultados e principalmente pela maneira que os dados são apresentados ao grande público, estimulando o debate. Iniciativas, como a cessão de itens e parâmetros calibrados para equalizações e estudos por fundações e institutos independentes, colaboram imensamente para a credibilidade e transparência do processo, além de permitir o enriquecimento de análises e processos e difusão do conhecimento.

Com os parâmetros de itens próprios equalizados na escala do SAEB, é possível ao IQE incrementar a quantidade de itens equalizados em outras avaliações, sem a necessidade de cessão de novos itens pelo INEP, o que simplifica imensamente o processo. E também permitirá a realização de avaliações em anos em que a Prova Brasil não é realizada, oferecendo-se como instrumento de correção de rumos, visto que os resultados são apresentados previamente com tempo adequado para a adoção de medidas necessárias.

Ao final deste trabalho, um produto de grande valor também foi concebido: o banco de dados eletrônico de itens que, além de armazenar os dados e as estatísticas de cada item e sua história de aplicação, contém boa parte da automatização dos processos que são usados na equalização, consolidando etapas trabalhosas em procedimentos mais simples, permitindo que as pessoas envolvidas nesta atividade dediquem seu tempo à análise de processos e resultados e não a pormenores e tecnicidades muitas vezes mais trabalhosos que complexos. Também como fruto desta experiência, a análise científica do item fica acentuada na rotina do especialista pedagógico, rumo à instrumentos de avaliação cada vez mais precisos.

O Instituto Qualidade no Ensino já realizou avaliações diagnósticas de aprendizagem em milhões de alunos, em diversos estados brasileiros, desde 1994. O trabalho realizado, descrito nesta dissertação, resultou na criação de bases para outras avaliações nas quais a utilização de uma escala nacional na apresentação de resultados expõe claramente aos envolvidos onde estamos e, principalmente, onde queremos chegar para elevar a qualidade da educação no nosso país.

8 - REFERÊNCIAS BIBLIOGRÁFICAS

ANDRADE, D. F. **Comparando desempenho de grupos de alunos por intermédio da Teoria de Resposta ao Item.** Em Estudos em Avaliação Educacional, n.23. São Paulo: 2001.

ANDRADE, D. F. & KLEIN, R. **Métodos estatísticos para avaliação educacional**: Teoria da Resposta ao Item. Boletim da ABE, n.43, p.21-28. 1999.

ANDRADE, D. F., TAVARES, H. R. & VALLE, R. C. **Teoria da Resposta ao Item:** conceitos e aplicações. 14º SINAPE, Associação Brasileira de Estatística. 2000.

BARBETTA, P. A., TREVISAN, L. M. V., TAVARES, H., AZEVEDO, T. C. **Aplicação da Teoria da Resposta ao Item Uni e Multidimensional.** Em Estudos em Avaliação Educacional, v25, n.57, p.280-302. São Paulo: 2014.

BEATON, A. E. & ALLEN, N. L. **Interpreting scales through scale anchoring**. Journal of Educational Statistics, n.17, p.191-204.

BIRNBAUM, A. **Some Latent Trait Models and Their Use in Infering an Examinee's Ability.** Statistical Theories of Mental Test Scores. Addison-Wesley, 1968.

BOCK, R. D. & LIEBERMAN, M. **Fitting a response model for *n* dichotomously scored items.** In Psychometrika, n.35, p.179-197. Springer-Verlag, 1970.

BOCK, R. D. & ZIMOWSKI, M. F. **Multiple Group IRT.** In Handbook of Modern Item Response Theory. New York: Spring-Verlag, 1997.

BORGATTO, A. & ANDRADE, D. **Análise Clássica de Testes com diferentes graus de dificuldade.** Em Estudos em Avaliação Educacional, v.23, n.52, p.146-156. São Paulo: 2012.

CRONBACH, L. J. 1978. **Coefficient Alpha and the internal structure of tests**. In Psychoketrika, v.16, p.297-334. 1951.

EMBRETSON, S. E. & HERSCHBERGER, S. L. **The new rules of measurement:** What every psychologist and educator should know. New Jersey: Erlbaum, 1999.

GULLIKSEN, H. O. **Theory of Mental Tests.** New York: Wiley, 1950.

HAMBLETON, R. K. & SWAMINATHAN, H. **Item Response Theory: Principles and Applications.** Boston: Kluwer Academic Publishers, 1985.

HORTA NETO, J. L. **Um olhar retrospectivo sobre a avaliação externa no Brasil.** Em Revista Iberoamericana de Educacion, n.42/5. OEI, 2007.

INEP. **Escalas de Proficiência.** 2014. Disponível em <http://portal.inep.gov.br/web/saeb/escalas-de-proficiencia>. Acesso em 07/09/2015.

INEP. **Histórico do SAEB.** 2015. Disponível em <http://portal.inep.gov.br/web/saeb/historico>. Acesso em 17/03/2015.

INEP. **Perguntas frequentes sobre ANEB e ANRESC.** 2016. Disponível em < http://provabrasil.inep.gov.br/perguntas-frequentes>. Acesso em 23/05/2016.

INEP. **Nota Técnica ENEM.** 2012. Disponível em <http://download.inep.gov.br/educacao_basica/enem/nota_tecnica/2011/nota_tecnica_tri_enem_18012012.pdf>. Acesso em 08/05/2015.

KELLEY, T. L. **The selection of upper and lower groups for the validation of test items.** In Journal of Educational Psychology, n.30, p.17-24. 1939.

KLEIN, R. **Utilização da Teoria de Resposta ao Item no Sistema Nacional de Avaliação da Educação Básica (SAEB).** Revista Meta: Avaliação, v.1, n.2, p.125-140. Rio de Janeiro: 2009.

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.** Hillsdale: Lawrence Erlbaum Associates, Inc. 1980.

LORD, F. & NOVICK, M. R. **Statistical theories of mental test scores.** Addison-Wesley, 1968.

MOREIRA JUNIOR, Fernando de Jesus. **Aplicações da Teoria da Resposta ao Item (TRI) no Brasil.** Rev.Bras.Biom., v.28, n.4 , p.137-170. São Paulo: 2010.

PASQUALI, L. **Psicometria: teoria dos testes na psicologia e na educação.** Petrópolis: Vozes, 2003.

PASQUALI, L & PRIMI, R. **Fundamentos da Teoria da Resposta ao Item.** Avaliação Psicológica, v.2, n.2. 2003. Disponível em <http://pepsic.bvsalud.org/scielo.php?pid=S1677-04712003000200002&script=sci_arttext>. Acesso em 18/04/2015.

RASCH, G. **Probabilistic Models for Some Intelligence and Attainment Tests.** Copenhagen: Danish Institute for Educational Research, 1960.

RECKASE, M. D. **Multidimensional item response theory.** New York: Springer, 2009.

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

Sistema Nacional de Avaliação da Educação Básica: **Relatório Nacional**. SAEB, 2001.

Sistema Nacional de Avaliação da Educação Básica. **SAEB 2003. Relatório da Análise Clássica do Teste**. Fundação Cesgranrio, 2004.

THISSEN, D. & MISLEVY, R. J. **Testing Algorithms**. Em Computerized Adaptive Testing: A Primer. Mahwah, NJ: Lawrence Erlbaum Associates, 2000.

U.S. Department of Education. Office of Educational Research and Improvement. National Center for Education Statistics. **The NAEP 1996 Technical Report.** Washington: NCES, 1999.

VALLE, R. C. **Teoria da Resposta ao Item.** Em Estudos em Avaliação Educacional, n.21, p.7-91. São Paulo: Fundação Carlos Chagas, 2000.