Wellington Silva

Proposta de uma metodologia para a produção e interpretação de medidas educacionais em avaliação em larga escala por meio da utilização da modelagem Rasch com duas ou mais facetas

TESE DE DOUTORADO

DEPARTAMENTO DE EDUCAÇÃO

Programa de Pós-Graduação em Educação

Rio de Janeiro

Fevereiro de 2019

PONTIFÍCIA UNIVERSIDADE CATÓLICA
DO RIO DE JANEIRO

Wellington Silva

PROPOSTA DE UMA METODOLOGIA PARA A PRODUÇÃO E INTERPRETAÇÃO DE MEDIDAS EDUCACIONAIS EM AVALIAÇÃO EM LARGA ESCALA POR MEIO DA UTILIZAÇÃO DA MODELAGEM RASCH COM DUAS OU MAIS FACETAS

Tese de Doutorado

Tese apresentada ao Programa de Pós-graduação em Educação da PUC-Rio como requisito parcial para obtenção do grau de doutor em Educação.

Orientadora: Profa. Alicia Maria Catalano de Bonamino

Coorientador: Prof. Joaquim José Soares Neto

Rio de Janeiro

Fevereiro de 2019

Wellington Silva

Possui mestrado em educação, com ênfase em políticas públicas educacionais, pela Universidade Federal de Juiz de Fora (2010) e graduação em Engenharia Elétrica pela Universidade Federal de Juiz de Fora (1987). Atuou como Engenheiro Eletricista e de produção por 15 anos nas áreas de máquinas elétricas rotativas e controle estatístico de processo. A partir de 2012 exerce a função de coordenador de produção de medidas educacionais do Centro de Políticas Públicas e Avaliação da Educação - CAEd/UFJF atuando como psicometrista em análises de bases de dados com ênfase em medidas educacionais pela Teoria da Resposta ao Item - TRI e modelagens multiníveis.

Ficha Catalográfica

Silva, Wellington

Proposta de uma metodologia para a produção e interpretação de medidas educacionais em avaliação em larga escala por meio da utilização da Modelagem Rasch com duas ou mais facetas / Wellington Silva ; orientadora: Alicia Maria Catalano de Bonamino ; co-orientador: Joaquim José Soares Neto. – 2019.
150 f. ; 30 cm

Tese (doutorado)–Pontifícia Universidade Católica do Rio de Janeiro, Departamento de Educação, 2019.
Inclui bibliografia

1. Educação – Teses. 2. Modelos Rasch multifacetas. 3. Avaliação educacional. 4. Escala de classificação. 5. Efeito contexto. I. Bonamino, Alicia Catalano de. II. Soares Neto, Joaquim José. III. Pontifícia Universidade Católica do Rio de Janeiro. Departamento de Educação. IV. Título.

CDD: 370

Agradecimentos

Sair de Juiz de Fora, enfrentar no mínimo 3 horas de viagem até a rodoviária do Rio, em seguida, pegar VLT, metrô e ônibus, aí já se foram mais algumas horas, até finalmente, chegar a PUC-Rio. É como encontrar um oásis: todo o cansaço da trajetória desaparece. Almoçar no "bandejão", no Couve-flor ou nas feirinhas, estar no meio de pessoas com tantas diferentes tendências e contemplar a simplicidade elegante da paisagem da PUC foi uma experiência fascinante. Sou muito agradecido a todos que me apoiaram nessa trajetória, pois, tão importante quanto o lado acadêmico, formalizado nessa Tese, foi a oportunidade de conviver com os professores, colegas de aulas, grupos de pesquisas, funcionários da secretaria e pessoas desconhecidas pelas escadarias dos dez andares do Prédio Cardeal Leme.

Obrigado à Prof. Alicia Bonamino, por ter aceitado o desafio de orientar um engenheiro no doutorado de Educação. Tenho certeza de que não foi uma tarefa fácil, por isso, meus profundos agradecimentos. Sinto-me muito honrado por ter sido orientado pela mesma orientadora da minha esposa. Essa é uma relação de muitos anos, de muita amizade e muitas histórias.

À Fundação CAEd, pela oportunidade de trabalhar ao longo dos últimos 16 anos com as principais políticas de avaliação educacional em praticamente todos os estados brasileiros. Nesse período, tive a felicidade de conviver com profissionais que veem construindo a história da avaliação educacional no país. Essa tese é fruto desse relacionamento, que muito mais do que o envolvimento técnico, tem a sua principal característica a amizade. Nesse contexto, agradeço profundamente à Prof. Lina Kátia Mesquita de Oliveira e ao Prof. Manuel Fernando Palácios da Cunha e Meloa, responsáveis pela condução política e técnica do CAEd, pela oportunidade de poder estudar no Rio de Janeiro nos últimos quatro anos.

Ao Prof. Neto, companheiro de safari na África e passeios de fusca nas ladeiras de Juiz de Fora, meu co-orientador, que me apresentou a MFRM pela primeira vez e me forneceu os recursos técnicos necessários para os estudos realizados nessa Tese.

À pesquisadora Cecilia Alves, que participou ativamente no conteúdo dessa tese, compartilhando sua experiência adquirida nos EUA e Canadá. Foram informações preciosas e que me possibilitaram rodar os *softwares* utilizados nessa tese. Abraços nas meninas e curta muito as delícias de vê-las crescer. Ver nos olhos

dos filhos a felicidade de descobrir algo novo, que para nós, é antigo e que nem nos damos conta, nos renova, nos dá força e nos faz acreditar e em um mundo melhor.

Ao Prof. Thierry Rocher, pelas dicas técnicas, por ter hospedado meu filho na França e por estar sempre disponível a tentar responder minhas perguntas objetivas, que, infelizmente, não têm respostas simples e rápidas.

À minha equipe de produção de medidas no CAEd, pelo apoio na realização das atividades do dia a dia nos meus momentos de idas ao Rio, pelas trocas de ideias e por ouvirem minhas aventuras em terras fluminenses.

À CAPES, pela ajuda de custo envolvendo minha participação e apresentação de trabalho intitulado "A transmissão da tecnologia da avaliação em larga escala no Brasil", no Seminário Internacional: *Penser les nouvelles problématiques dans une perspective internationale*, realizado na Université Paris-Est Créteil, Créteil, France em 2016.

À CAPES e à PUC- Rio, por meio do Decanato do CTCH – Centro de Teologia e Ciências Humanas, pela ajuda de custo para minha participação e apresentação do trabalho intitulado "Análise da Eficácia Escolar das Escolas de tempo Integral no Ensino Médio do Ceará", no VI Congresso Ibero Americano de Política e Administração de Educação realizado na Catalunha – Espanha, 2018

Agradecimento especial à minha revisora de Língua Portuguesa, que além de corrigir e dar fluidez aos escritos dessa tese, escreve comigo, a mais de 30 anos, crônicas, poemas, artigos científicos e listas de compras de supermercado.

Resumo

Silva, Wellington; Bonamino, Alicia Maria Catalano (orientadora). **Proposta de uma metodologia para a produção e interpretação de medidas educacionais em avaliação em larga escala por meio da utilização da modelagem Rasch com duas ou mais facetas**. Rio de Janeiro, 2019. 150p. Tese de Doutorado – Departamento de Educação, Pontifícia Universidade Católica do Rio de Janeiro.

Nesta tese, trabalhou-se com a modelagem Rasch visando a apresentar alternativas mais práticas e de melhor qualidade em termos de medida, para dois cenários distintos. O primeiro está relacionado ao fato de que medir conhecimento é algo muito complexo e de difícil entendimento para profissionais que não são da área da psicometria. Por meio de experimentos envolvendo modelos da "família" Rasch, apresentamos a aplicabilidade e as potencialidades dessa modelagem para atender a novas demandas de avaliação em larga escala no Brasil. O segundo cenário relaciona-se à busca de medir, de modo o mais imparcial possível, itens de produção escrita, em que a nota recebida pelos alunos é influenciada pela subjetividade dos corretores, ou seja, corretores lenientes beneficiam alunos e corretores severos penalizam alunos. Diante desses dois cenários, esta tese tem os seguintes objetivos: (i) trazer para o âmbito das avaliações realizadas no Brasil uma modelagem matemática mais simples que aquela atualmente adotada, visando uma melhor comunicação com os professores, e; (ii) a possibilidade de operar não apenas com itens de múltipla escolha, corrigidos de forma automática, mas também com itens de produção escrita, em que a subjetividade dos corretores (severidade) é controlada pelo modelo psicométrico, gerando medidas de melhor qualidade. Para isso, utilizou-se a modelagem Rasch com multifacetas, abordando, por meio de casos práticos, as vantagens dessa modelagem em relação a outras metodologias atualmente adotadas no país. Assim, para a alcançarmos o primeiro objetivo, confrontamos a modelagem Rasch com multifacetas com a modelagem de três parâmetros logísticos em um estudo de efeito contexto em testes compostos por diferentes modelos de cadernos e com mais de uma disciplina avaliada por caderno e, para o segundo, comparamos as medidas de proficiência através da Rasch com

multifacetas com as notas médias das duplas correções dadas pelos corretores aos alunos em testes do tipo redação. A partir dos resultados encontrados, concluímos que a Rasch com multifacetas pode ser utilizada de forma alternativa ou concomitante com as avaliações que utilizam a modelagem de três parâmetros logísticos, produzindo resultados mais rápidos e de entendimento mais fácil por parte dos professores e que, no caso de redações, as proficiências obtidas pela Rasch com multifacetas apresentaram medidas com melhores indicadores de fidedignidade e validade, quando comparadas com as medidas de notas via Teoria Clássica do Teste, sendo, portanto, uma alternativa mais viável para esse tipo de avaliação. Conclui-se essa tese apresentando situações de empregabilidade das metodologias estudadas.

Palavras-chave

Modelos Rasch multifacetas; avaliação educacional; escala de classificação; efeito contexto.

Abstract

Silva, Wellington; Bonamino, Alicia Maria Catalano (adviser). **Proposal of a methodology for the production and interpretation of educational measures in large-scale assessment by using Rasch modeling with two or more facets**. Rio de Janeiro, 2019. 150p. Tese de Doutorado – Departamento de Educação, Pontifícia Universidade Católica do Rio de Janeiro.

In this thesis, we worked with Rasch modeling, aiming to present more practical alternatives and better quality in terms of measurement, for two different scenarios. The first one is related to the fact that measuring knowledge is something very complex and difficult to understand for professionals who are not in the psychometrics area. Through experiments involving the Rasch "family" models, we present the applicability and the potentiality of this model to adequately comply with the new demands of the large-scale evaluation in Brazil. The second scenario is related to the search of measuring, in the most impartial way possible, written production items which grade received by the subjectivity of the raters (severity), that is, lenient raters benefit students and severe raters penalize them. In view of these two scenarios, this thesis has the following objectives: (i) to bring to the scope of the evaluations carried out in Brazil a simpler mathematical modeling than the currently adopted, aiming at a better communication with the teachers; and (ii) the possibility of operating not only with multiple choice items, corrected automatically, but also with written production items, in which the subjectivity of the raters (severity) is controlled by the psychometric model, generating better quality measures. For this, Many-Facet Rasch Measurement was used, approaching, through practical cases, the advantages of this modeling in relation to other methodologies currently adopted in the country. Thus, in order to reach the first objective, we confronted Many-Facet Rasch Measurement with the modeling of three logistic parameters in a study of context effect in tests composed by different models of test books and with more than one discipline evaluated by test book and, for the second one, we compared the measures of proficiency through the Many-Facet Rasch Measurement with the average scores of the double corrections given

by the raters to the students in tests of the essay type. From the results found, we conclude that the Many-Facet Rasch Measurement can be used in an alternative or concomitant way with the evaluations that use the three logistic parameters model, producing faster results and easier to understand by the teachers and that, in the case of essays, the measures of proficiency obtained by Many-Facet Rasch Measurement presented measures with better reliability and validity indicators, when compared to the grading measures through the Classical Theory of Testing, being, therefore, a more viable alternative for this type of evaluation. This thesis concludes with situations of usability of the methodologies studied.

Keywords

Many-facet Rasch measurement; educational evaluation; rating scale; context effect.

Sumário

Lista de Figuras

Lista de Tabelas

Lista de Quadros

Lista de Abreviaturas e Siglas

ATI	- *Assessment Tchnology Incorporated*
ATS	- *Association of Test Publishe*
ANA	- Avaliação Nacional da Alfabetização
BIB	- Blocos Incompletos Balanceados
CAEd/UFJF	- Centro de Políticas Públicas e Avaliação da Educação da Universidade Federal de Juiz de Fora
CCI	- Curvas Características dos Itens
dp	- Desvio-padrão
2PL	- Dois parâmetros logísticos
GERES	- Estudo Longitudinal da Geração Escolar 2005
Enem	- Exame Nacional do Ensino Médio
Encceja	- Exame Nacional para Certificação de Competências de Jovens e Adultos
DIF	- Funcionamento Diferencial de Itens
Ideb	- Índice de Desenvolvimento da Educação Básica
Inep	- Instituto Nacional de Estudos e Pesquisa Educacionais Anísio Teixeira
MFRM	- Modelagem Rasch com multifacetas
PAEBERS-ALFA	- Programa de Avaliação da Educação Básica do Estado do Espírito Santo
Saeb	- Sistema Nacional de Avaliação da Educação Básica
SPAECE	- Sistema Permanente de Avaliação da Educação Básica do Ceará
TCT	- Teoria Clássica dos Testes
TRI	- Teoria da Resposta ao Item
TMAC	- Teoria das Medidas Aditivas Conjuntas
TRM	- Teoria Representacional da Medida
CAT	- Teste Adaptativo Computadorizado
3PL	- Três parâmetros logísticos

1. Introdução

As metodologias estatísticas na área da Teoria da Resposta ao Item (TRI), adotadas nas avaliações educacionais brasileiras, permanecem praticamente inalteradas desde 1997, ano de sua adoção inicial no Sistema Nacional de Avaliação da Educação Básica (Saeb), pelo Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (Inep).

Podemos dizer que as avaliações em larga escala realizadas no Brasil se caracterizam como: (i) avaliações transversais; (ii) testes formados por itens dicotômicos, e; (iii) utilização do modelo de três parâmetros logísticos da TRI.

Com essas três características, as avaliações nacionais vêm sendo referência para diversas políticas educacionais e práticas pedagógicas voltadas para a melhoria da qualidade e da equidade das escolas. Esse é o caso, por exemplo, das políticas de *accountability*[1], que se baseiam nos resultados das avaliações realizadas por diversos estados brasileiros. E, também, o caso das práticas pedagógicas, que se apoiam na interpretação das habilidades adquiridas pelos estudantes em diversos níveis da escala de proficiência, como acontece, entre outros, com os recursos disponibilizados pelo Inep, através da plataforma digital Devolutivas Pedagógicas, e, ainda, com os trabalhos realizados pelo Centro de Políticas Públicas e Avaliação da Educação da Universidade Federal de Juiz de Fora (CAEd/UFJF) também por meio de plataformas digitais e de oficinas de apropriação de resultados focadas nos resultados de cada estado ou município avaliado.

É inegável que, ao longo desses vinte anos de existência da avaliação em larga escala da educação, se disseminou uma cultura da avaliação no Brasil, a qual se revela não apenas na participação de todas as escolas púbicas no Saeb, mas também e principalmente, nos sistemas próprios de avaliação externa existentes em praticamente todos os estados e em um número significativos de municípios, principalmente capitais.

Embora esses sistemas tipicamente adotem a mesma escala do Saeb, temos observado o surgimento de novos projetos avaliativos, com novas estruturas de testes, novos formatos de itens e a introdução de novas disciplinas, que evidenciam

[1] Responsabilização ou prestação de contas.

a necessidade de uma atualização da tecnologia desenvolvida pelo Brasil na área da TRI, de modo a atender mais adequadamente às crescentes e diversificadas demandas das diversas instâncias educacionais.

São duas as hipóteses que orientam esta tese. A primeira sustenta que a modelagem Rasch com multifacetas, do inglês *Many-Facet Rasch Measurement* (MFRM) minimiza o efeito na proficiência dos alunos provocado pela utilização de diferentes modelos de cadernos de testes em que mais de uma disciplina é avaliada em diferentes posições, por exemplo, cadernos iniciando com Língua Portuguesa e terminado com Matemática e vice-versa.

A segunda hipótese sustenta que a medida de proficiência obtida por meio da MFRM, na correção de redações realizadas dentro das avaliações em larga escala no Brasil, oferece uma qualidade de medida superior e resultados mais justos para os alunos, que os obtidos pela nota média dada por dois corretores distintos.

A primeira hipótese está relacionada ao fato de que as modelagens utilizando duas facetas pela TRI via três parâmetros logísticos não modelam o efeito contexto provocados pelos delineamentos dos cadernos, apresentando resultados em que as proficiências dos alunos que fizeram um disciplina no início do caderno terão uma proficiência maior que os alunos que fizeram a disciplina no final do caderno, provocado pelo efeito cansaço da ordem da disciplina no caderno de teste. Com a utilização da MFRM conseguimos não apenas minimizar esse efeito, tornando os resultados mais justos, como também proporcionar aos professores medidas de desempenho mais simples de se trabalhar no meio educacional em comparação com os resultados obtidos pela modelagem de três parâmetros logísticos. Esse entendimento é uma das finalidades recorrentemente encontradas entre os profissionais envolvidos com avaliações educacionais cujo objetivo, além de monitorar o desempenho dos alunos, é, também, o de proporcionar aos professores uma interpretação pedagógica dos resultados dos estudantes que possa ser utilizada como ferramenta de trabalho em sala de aula e na melhoria do ensino.

A segunda hipótese está relacionada à necessidade de medir, do modo mais imparcial possível, itens de produção escrita em que a nota recebida pelos alunos é influenciada pela subjetividade, isto é, pela severidade ou pela leniência dos corretores, que pode penalizar ou beneficiar os estudantes.

É nesta perspectiva que propomos a utilização de novos procedimentos estatísticos para testes compostos com itens apenas do tipo dicotômico e com testes

compostos com itens dicotômicos e/ou politômicos, por meio da utilização da MFRM.

1.1 Objetivos

Nas modelagens pela TRI, adotadas no Brasil, na área da educação, são levados em consideração apenas dois componentes ou facetas: respondentes e itens do teste, representados nos modelos da TRI pelas variáveis matemáticas proficiência dos alunos e parâmetros dos itens. Por meio da utilização da MFRM, é possível inserir mais facetas na modelagem, de forma a contemplar as seguintes situações: controle do efeito contexto[2] do delineamento dos cadernos e estimativa de indicadores da influência da severidade dos corretores nas pontuações de respondentes, por exemplo, em produções textuais.

Assim, esta tese tem os seguintes objetivos: (i) trazer para o âmbito das avaliações realizadas no Brasil uma modelagem matemática mais simples que aquela que vem sendo adotada desde a criação do Saeb, visando produzir medidas que promovam uma melhor interpretação e uma maior aplicabilidade pedagógica entre professores e gestores, e; (ii) explorar a possibilidade de operar não apenas com itens de múltipla escolha, corrigidos de forma automática, mas também com itens de produção escrita, em que a subjetividade dos corretores (severidade/leniência) é controlada pelo modelo psicométrico, gerando, assim, medidas de melhor qualidade e mais justas. Também se relaciona com a possibilidade de construção de uma escala de redação a ser adotada no país.

Para alcançarmos o primeiro objetivo, confrontamos a MFRM com a modelagem de três parâmetros logísticos (3PL). Mais especificamente, realizamos um estudo com a proposta de uma metodologia para a produção e interpretação de medidas educacionais em avaliação em larga escala por meio da utilização da MFRM que levam em conta os efeitos da utilização de diferentes modelos de cadernos de testes na estimativa das proficiências dos alunos. Isto nos permitiu apresentar outras possibilidades de modelagem da realidade educacional que contribuem para tornar os resultados das avaliações educacionais mais acessíveis

[2] De acordo com Yen (1980), os parâmetros dos itens são influenciados pela localização do item e/ou diferentes agrupamentos em que o item está inserido (natureza dos itens vizinhos).

para os professores. Para atingirmos o segundo objetivo, comparamos as medidas de proficiência por meio da MFRM com as notas médias da dupla correção realizada pelos corretores em provas de redação. As redações realizadas dentro das avaliações em larga escala no Brasil, seguem todas, sem exceção, a estrutura das redações do Exame Nacional do Ensino Médio (Enem) em que a pontuação do aluno é obtida através da nota média dada por dois corretores distintos, ou por uma terceira nota data por um supervisor, caso haja discrepância entre as duas primeiras correções. Com uso de dados reais, foi feita uma comparação entre esse método que produz uma nota e a medida de proficiência obtida através da MFRM.

Em síntese, por meio da utilização da MFRM, foi possível: (i). inserir mais facetas na modelagem, de forma a minimizar os efeitos de modelos de cadernos de testes distintos para estimar as proficiências dos alunos a fim de contemplar o controle do efeito contexto provocado pelo delineamento dos cadernos, e; (ii) a estimativa de indicadores da influência de corretores nas pontuações de respondentes, por exemplo, em itens abertos de resposta graduada.

1.2 Plano de trabalho

Essa tese está estruturada em seis capítulos. Os três primeiros são capítulos teóricos. Neles abordamos os fundamentos das medidas de desempenho educacional envolvendo modelagens pela TRI, com foco nas MFRM, enquanto os três últimos são compostos por dois capítulos empíricos e pelas considerações finais; nestes capítulos, para aplicação da MFRM, utilizamos dados obtidos de avaliações em larga escala realizadas no Brasil.

No primeiro capítulo teórico, apresentamos uma cronologia da evolução da teoria da medida, conforme apresentado por Pasquali (2011), Golino et al. (2015), iniciando no ano de 1927 até os dias atuais. Durante esse período emergiram duas teorias com o objetivo de axiomatizar as medidas não só nas ciências físicas, como também nas Ciências Humanas e Sociais: a Teoria Representacional da Medida (TRM) e a Teoria das Medidas Aditivas Conjuntas (TMAC). Observamos, nessa trajetória, a difícil tarefa de se obter uma medida extensiva para fenômenos latentes, a qual foi obtida através da formulação de Rasch (1956), dando origem aos modelos da TRI, amplamente utilizados a partir dos anos de 1980 com a disseminação de

recursos computacionais. Pensamos ser oportuno esse levantamento histórico, no sentido de nos apropriarmos do embasamento teórico das medidas das avaliações que utilizam a TRI, assim como o conceito de indicadores de qualidade da medida, o ajuste do modelo aos dados, o erro da medida e outros elementos necessários de serem analisados a fim de se verificar a qualidade das medidas obtidas.

Nosso foco, no capítulo 2 teórico, é um levantamento histórico da evolução da TRI no Brasil, com o início de sua adoção no ano de 1995 pelo Saeb até os dias atuais. Ao longo desse período, a modelagem matemática se caracterizou pelos modelos de duas facetas e 3PL (Lord, 1980). Apresentamos as principais características desse modelo, a metodologia adotada no Brasil para sua implementação, a fim de se construir e manter uma escala nacional de conhecimento e as iniciativas adotadas para tornar os resultados dessas avaliações em recursos pedagógicos para os professores usarem em suas práticas em sala de aula com o objetivo de se melhorar a qualidade de ensino. Finalizamos esse capítulo apresentando as tendências das avaliações em função de novos recursos tecnológicos.

No terceiro capítulo teórico, apresentamos os fundamentos das modelagens Rasch e MFRM, suas formulações matemáticas, as características do *software FACETS* (Linacre, 1989) utilizado na parte empírica dessa tese, e a metodologia empregada a fim de se garantir uma medida de qualidade da proficiência dos alunos.

A parte empírica é apresentada em dois capítulos com estudos de modelagens por meio utilização da MFRM. Nesses capítulos, ao abordamos os modelos Rasch no panorama nacional, onde são empregados, basicamente, modelos de 3PL, para avaliações em larga escala, e modelos via Teoria Clássica dos Testes (TCT), para correção de redações, nossa intenção é apresentar outras alternativas para a produção de medidas educacionais.

O capítulo 5 apresenta a análise do impacto da dificuldade dos cadernos na proficiência dos alunos através do uso da MFRM e os benefícios de medidas obtidas por meio dessa modelagem no meio educacional.

No capítulo 6, ao trabalharmos com itens politômicos, entramos em uma área da psicometria que estuda o comportamento dos juízes na medição do constructo através da utilização de escalas de classificação (*rating scales*). Realizamos estudos comparativos que indicam uma melhor estimativa do desempenho dos alunos quando se utiliza a medida de proficiência via MFRM ao

invés da nota via TCT.

No sétimo e último capítulo, apresentamos propostas baseadas nos estudos realizados, que têm foco na utilização de modelos mais parcimoniosos, representados pelos modelos da "família" Rasch. Apontamos suas aplicabilidades em situações ainda pouco exploradas pelas avaliações em larga escala no Brasil, tendo sempre como foco uma medida com qualidade que atenda às necessidades do professor em suas atividades em sala de aula.

2. Teoria da medida

Medir é uma atividade fundamental para tomarmos uma decisão. Realizamos medições com muita naturalidade em praticamente tudo que fazemos em nosso dia a dia como, por exemplo, medir distâncias, temperaturas e velocidade com um rápido olhar para o painel de um carro. Basicamente, segundo Albertazzi Jr & Souza (2018), as medições são realizadas com três propósitos, sendo cada um desses propósitos utilizados de forma independente ou concomitante. São eles: monitorar, controlar e investigar.

Ao recebermos os resultados oriundos de uma avaliação educacional em larga escala, monitoramos quando observamos passivamente os resultados de proficiência dos alunos e quando verificamos a evolução do desempenho de uma escola ao longo do tempo e comparamos esse desempenho com outras escolas de outros municípios. Controlamos, quando, além de comparar, agimos, utilizando os resultados para formular políticas educacionais e estabelecer procedimentos para melhorar os níveis de desempenho medidos. E investigamos para descobrir o novo, para explicar e formular hipóteses sobre uma dada realidade, através, por exemplo, da utilização conjunta dos dados quantitativos medidos e de informações qualitativas advindas da experiência dos professores e demais atores do meio educacional, de forma a ter um melhor entendimento da realidade educacional e, dessa forma, poder mapear as problemáticas dessa realidade tão complexa.

Essas três possibilidades de uso dos resultados das avaliações ocorrem de forma concomitante em um ambiente educacional. Simplesmente monitorar, não justificaria a quantidade de recursos financeiros e logísticos direcionados a uma avaliação em larga escala. Observamos, no país, diversos exemplos de políticas públicas no sentido de fazer os professores se apropriarem dos resultados visando o controle e a investigação.

Normalmente, não temos um conhecimento teórico do que é uma medida, no que se refere ao seu conceito e propriedades, sendo muitas vezes confundida com dar número às coisas. Porém, medir não significa simplesmente quantificar coisas ou processos. O conceito de medida está além da quantificação. Em muitas situações, estamos apenas quantificando e não medindo. O Vocabulário

Internacional de Metrologia (VIM, 2012) define o termo medição como sendo um "Processo de obtenção experimental de um ou mais valores que podem ser, razoavelmente, atribuídos a uma grandeza". Porém, como atribuir valores de forma razoável a uma grandeza e termos uma medida? E, mais complexo ainda, como atribuir valores e termos uma medida para emoções, conhecimentos e demais variáveis latentes das áreas das Ciências Humanas e Sociais?

Mais especificamente, no meio educacional, foco desta tese, é comum pensar que a nota atribuída ao aluno pelo professor é uma medida de conhecimento do aluno na disciplina avaliada. No entanto, essa prática tão comum e consolidada no meio escolar não é uma medida, mas apenas de uma quantificação. Esse é um tema que prejudica muito a utilização dos resultados de uma avaliação em larga escala por parte dos professores e gera muita rejeição e críticas a esse tipo de avaliação. É difícil para os professores quebrar esse paradigma e entrar no ambiente dos modelos matemáticos da TRI, que realmente medem o desempenho dos alunos em uma escala, através do constructo denominado proficiência.

Ao longo desse capítulo, traremos conceitos, marcos históricos, propriedades e indicadores de qualidade da medida, envolvendo a mensuração de constructos nas Ciências Humanas e Sociais, mais especificamente na Educação, no intuito de esclarecer o que é medir, e permitir o entendimento dos capítulos empíricos desta tese, que se iniciam com o capítulo 5.

2.1 A natureza da medida

O uso do número para representar os fenômenos naturais constitui o objeto da teoria da medida. Não iremos aprofundar, nessa tese, os axiomas que envolvem a teoria da medida nas ciências físicas, que por sinal estão bem consolidados, mas sim, apresentar um breve histórico da evolução da medida nas Ciências Humanas e Sociais até se chegar na TRI, que representa, atualmente, a metodologia mais adequada e utilizada mundialmente na medição de variáveis latentes.

Ao se pretender medir algo, independente se a variável em estudo é da ciência física, humana ou social, devemos ter, conforme asseveram Pasquali (2011) e Golino et al. (2015), respostas para três problemas relacionados à medida, a saber:

- Representação ou isomorfismo

É justificável o uso do número para representar o fenômeno natural em estudo?

- Unicidade da representação

A quantificação é a única ou melhor representação da realidade estudada? Esse é um tema que gera grande polêmica na área da medida nas ciências sociais e humanas e que, de certa forma, foi o precursor e fonte motivadora das diversas metodologias desenvolvidas na área levando aos trabalhos de Galton (1880), Binet (1900) Stevens (1946) e Rasch (1960 apud Pasquali, 2011).

- Erro

A toda medida está relacionado um erro. Cabe ao especialista envolvido no processo de medição ter o discernimento das potencialidades da medida e de sua aplicabilidade na interpretação do fenômeno estudado.

2.2 A medida nas Ciências Humanas e Sociais

Na investigação histórica da evolução das medidas em Ciências Humanas e Sociais, ao longo do tempo, torna-se mais que necessária a separação entre a evolução das aplicações de instrumentos de medidas, e a axiomatização desses procedimentos. A primeira ocorre, na maior parte das vezes, por necessidade, já a axiomatização vem bem posteriormente, a fim de enquadrar tais procedimentos dentro de uma teoria matemática mais abrangente. Os axiomas referentes aos procedimentos de medidas são classificados segundo duas perspectivas teóricas: a TRM e a TMAC.

Não temos a intenção de nos aprofundarmos nos trabalhos e discussões envolvendo essas teorias, e muito menos estabelecer uma ordem cronológica de todas as teorias e axiomas desenvolvidos ao longo da história da teoria da medida. Entretanto, traremos para esse tópico os principais fundamentos dessas duas teorias, de modo a termos uma percepção de como evoluíram os esforços na tentativa de se medir os fenômenos latentes até se chegar na TRI e consequentemente na formulação da Psicometria como uma área da Psicometria.

Pasquali (2009,p.1) definiu a Psicometria da seguinte forma:

> Etimologicamente, psicometria representa a teoria e a técnica de medida dos processos mentais, especialmente aplicada na área da Psicologia e da Educação. Ela se fundamenta na teoria da medida em ciências em geral, ou seja, do método quantitativo que tem, como principal característica e vantagem, o fato de representar o conhecimento da natureza com maior precisão do que a utilização da linguagem comum para descrever a observação dos fenômenos naturais.

Na Figura 1, destacamos as principais datas, autores e trabalhos que foram referências na história da psicometria. Iniciaremos nossa abordagem sobre a evolução da Psicometria no ano de 1946. Para estudos mais detalhados sobre pesquisadores e os respectivos trabalhos realizados antes dessa data, consultar Schultz (2019).

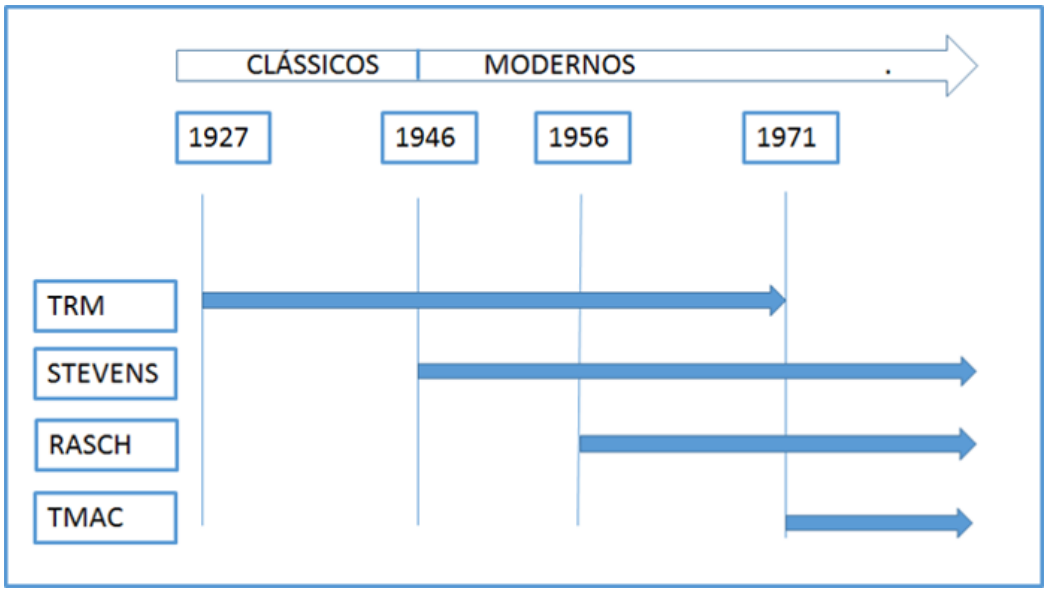


Figura 1 – Cronologia da teoria da medida

Fonte: O Pesquisador

O ano de 1946, data da formulação das escalas de Stevens é considerado como um divisor entre as duas teorias, que separaram os pesquisadores em clássicos e modernos. Antes de Stevens, o que prevalecia era o pensamento clássico, representado por Campbell, segundo o qual a medida era exclusiva para fenômenos em que fossem possíveis concatenações empíricas, o que deixava as Ciências Humanas e Sociais fora do rol das demais ciências.

Stevens tentou caracterizar a medida de tal modo que não dependesse da concatenação empírica, o que possibilitou a medida nas Ciências Humanas e Sociais. O equívoco de Stevens, no entanto, foi não ter especificado a regra para se ter uma medida extensiva, além do fato de que, nessa época, os fenômenos extrínsecos se enquadravam basicamente nas escalas nominal e ordinal, limitando análises mais elaboradas para esses fenômenos.

De acordo com Stevens (1946, p. 677), "medir, no sentido mais amplo, é definido como o ato de atribuir números a objetos ou eventos de acordo com regras".

Para Golino et al. (2015), a crença de que quantificar e medir são a mesma coisa tem origem nesse momento histórico relacionado ao desafio de realizar medidas em ciências humanas, pois, pelo fato de não especificar a natureza da regra a ser utilizada, a medida tornou-se passível de ser alcançada utilizando-se qualquer atividade de quantificação.

Mas, independentemente do fato de Stevens não ter definido as regras para se ter realmente uma medida, as escalas por ele propostas são utilizadas até o momento atual no meio científico, pois são válidas para classificar os tipos de variáveis existentes em qualquer processo de medição.

O fato de que números podem ser atribuídos ao que se está medindo de acordo com diferentes regras, leva a diferentes tipos de escalas e a diferentes tipos de medidas. Stevens elaborou quatro tipos de escalas: nominal, ordinal, intervalar e de razão. A complexidade dessas escalas varia da mais simples, que é a escala nominal, passando em seguida pelas escalas ordinal e intervalar, e terminado com a mais complexa, que é a escala de razão.

Interessante observar que, seguindo essa ordem das escalas, as propriedades de uma escala inferior, seguindo a ordem de complexidade, estão englobadas na escala seguinte, conforme definição apresentada a seguir:

- **Escala nominal**

É o nível mais elementar de representação baseado no agrupamento e classificação de elementos para a formação de conjuntos distintos. As observações são divididas em categorias segundo um ou mais de seus atributos. Assim, tem-se registros essencialmente qualitativos. A única estatística possível para esse tipo de escala é a frequência de dados por categorias.

- **Escala ordinal**

É a avaliação de um fenômeno em termos de onde ele se situa dentro de um conjunto de patamares ordenados, variando desde um patamar mínimo até um patamar máximo. Ou seja, é possível o ranking, mas não se tem a intensidade das diferenças entre as medidas.

- **Escala intervalar**

É uma forma quantitativa de registrar um fenômeno, em que é possível quantificar as distâncias entre medições. Mas, nesse tipo de escala, por exemplo, o valor "10" não significa o dobro de "5", uma vez que não existe um ponto nulo da grandeza medida, ou seja, não existe a referência zero. Diferentemente das duas escalas anteriores, é possível calcular estatísticas descritivas como médias e desvios padrões para as grandezas inseridas nesse tipo de escala.

- **Escala de razão**

É a mais completa e sofisticada das escalas numéricas. Ela é uma quantificação produzida a partir da identificação de um ponto zero que é fixo e absoluto, representando, de fato, um ponto de nulidade, ausência e/ou mínimo. Nela, uma unidade de medida é definida em termos da diferença entre o ponto zero e uma intensidade conhecida. Nessa escala, um valor medido de "10" efetivamente indica uma quantidade duas vezes maior do que o valor "5", o que não necessariamente acontece nas demais escalas.

Apresentamos, no Quadro 1, as funções matemáticas relacionadas a cada uma das escalas de Stevens.

Quadro 1 – Relação entre escalas, tipos de função e medidas

Escala	Função	Análise da medida	
		Posição	Variação
Nominal	Bionívoca	Moda	
Ordinal	Monotônica crescente	Mediana	Percentil
Intervalar	$T = \alpha B + \beta$	Média aritmética	Desvio-padrão
Razão	$T = \alpha B$	Média geométrica e média harmônica	% de variância

Fonte: Hauck Filho (2014).

Após Stevens, inicia-se a era dos pesquisadores modernos, tendo como data marcante o ano de 1956, com o modelo matemático de Rasch, que possibilita a

construção de uma escala intervalar para habilidades latentes, dando origem aos modelos da TRI.

A formulação matemática desse modelo, na concepção de Rasch, se traduz na seguinte expressão:

$$\boldsymbol{X_{pi} \sim \emptyset\ (\beta_p\ \Lambda\ \delta_i\ )}$$

Em que,

A probabilidade de um aluno “p” acertar o item “i” depende da habilidade β do aluno e da dificuldade δ do item. Nessa modelagem, a habilidade do aluno é independente da dificuldade do item, mas a probabilidade de acertar ao item é dependente tanto da habilidade do aluno quanto da dificuldade do item.

Com a axiomatização elaborada por Krantz et al. (1971 apud Golino et al., 2015), tivemos o surgimento da TMAC, possibilitando a entrada das Ciências Humanas e Sociais no rol das demais Ciências.

2.3 Indicadores da qualidade da medida

A qualidade da medida é fundamental nas avaliações educacionais em larga escala cujos resultados são trabalhados em comparações entre diferentes alunos, entre diferentes escolas ou entre avaliações em diferentes períodos de tempo, de forma a garantir subsídios para a tomada de decisões de gestores, professores e alunos.

Por meio dessas comparações, gestores elaboram políticas públicas, professores trabalham pedagogicamente os conhecimentos a serem adquiridos por seus alunos e esses, por sua vez, têm uma percepção do seu nível de conhecimento em relação aos seus pares.

Conforme aponta Pasquali (2007), ter domínio da qualidade da medida produzida ao se aplicar um teste de conhecimento nos remete a uma questão ética, no sentido de como deveríamos utilizar seus resultados. Somente com indicadores apropriados da qualidade da medida, teremos condições de assegurar o nível de apropriação que poderemos realizar com os resultados de uma avaliação.

São muitos os termos utilizados para parâmetros relacionados à qualidade das medidas e, algumas vezes, um mesmo termo refere-se a conceitos totalmente

diferentes quando tratados por diferentes autores. Divergências também são observadas nas nomenclaturas quando esses termos estão relacionados a pesquisas quantitativas e qualitativas, pois não existe uma padronização entre essas áreas no que concerne os parâmetros de qualidade de medida, A pluralidade de termos e definições gera muita dúvida ao tentar entender os indicadores da qualidade da medida.

A qualidade da medida de um constructo, em nosso caso específico, a proficiência do aluno é garantida pela qualidade do instrumento e da metodologia utilizada na medição. Embora sejamos tentados a pensar que não há diferença entre esses dois aspectos, pois, normalmente utilizamos um único método de medição, utilizamos erroneamente os conceitos de qualidade sem fazer distinção entre o que se está medindo e como se está medindo.

Faremos, portanto, uma distinção dos termos a serem utilizados quando pensamos em medida de um constructo e no instrumento/método de medição. Como existe muita confusão na tentativa de definir e de separar as nomenclaturas utilizadas, apresentamos no Quadro 2 nossa versão, baseada em Pasquali (2007).

Quadro 2 – Termos usados para definir a qualidade da medida

Medida de um constructo	**Correspondência**	**Instrumento/Método de medição**	
Termo utilizado na Tese		**Termo utilizado na Tese**	**Termos relacionados**
Legítima	⟷	Validade	Exatidão Acurácia Veracidade Pertinência Relevância
Confiável	⟷	Fidedignidade	Precisão Confiabilidade Consistência Repetibilidade Reprodutibilidade Estabilidade Constância Consistência interna

Fonte: O Pesquisador.

Tendo como referência o Quadro 2, a qualidade da medida é garantida pela qualidade do instrumento/método de medição, onde, instrumentos válidos geram

medidas legítimas e instrumentos precisos geram medidas confiáveis. Conforme pode ser observado, são vários os termos utilizados para a validade e a fidedignidade, que estão apresentados na coluna “termos relacionados” do referido quadro. Tivemos o cuidado de fazer a distinção entre os termos validade e fidedignidade, porque, em alguns artigos eles aparecem misturados, geralmente com a utilização do termo confiabilidade para se referir à validade.

Para verificarmos a qualidade do instrumento de medição são utilizados dois parâmetros, a validade e a confiabilidade, o primeiro, segundo Kelly (1927), ocorre quando o instrumento/método mede aquilo que se pretende medir e está relacionado com o vício[3] da medida. Essa definição, embora antiga, é retomada por Pasquali (2007) em função das muitas ambiguidades do termo ao longo do tempo, que gera muita confusão. O segundo parâmetro tem a ver com a variabilidade das medidas e está relacionado com a calibração do instrumento/método, de forma a minimizar os erros de medição.

A noção de validade e fidedignidade é melhor visualizada com a utilização das imagens da Figura 2, que apresenta quatro situações hipotéticas de uma pessoa atirando quatro vezes em cada alvo. Nessas situações, o centro de cada alvo representa o valor verdadeiro da característica que está sendo medida e os quatro tiros representam as medidas realizadas.

[3] Vício ou tendência é a diferença entre o valor real da característica medida e a média das medidas dessa característica (Werkema, 2006).

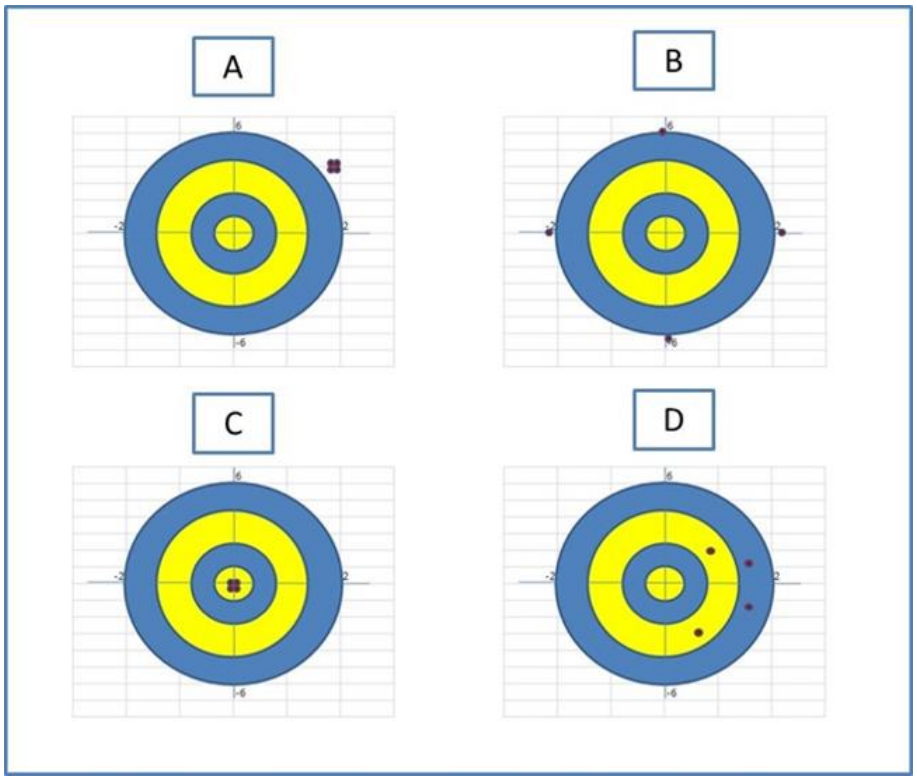


Figura 2 – Situações possíveis de validade e fidedignidade

Fonte: Versão elaborada pelo próprio autor.

Na situação ‘A”, temos um exemplo de processo fidedigno e não válido, uma vez que a variabilidade dos tiros é muito baixa (alta precisão), porém a validade é muito baixa, pois os tiros não acertam o centro do alvo (vício muito alto). Na situação ‘B”, temos um processo não fidedigno e válido; a fidedignidade é muito baixa (alta variabilidade) e a validade é melhor do que na situação ‘A”, pois a diferença do centro do alvo para a média dos tiros é muito baixa (vício próximo de zero). Na situação “C”, temos alta validade e alta fidedignidade, com os tiros acertando o centro do alvo e com pouca variação, o que é a situação ideal ao lidarmos com a medição de qualquer processo, e, na situação “D”, temos um processo não válido e não fidedigno.

2.4 Ajuste dos dados ao modelo psicométrico

A interpretação dos resultados de uma avaliação educacional para ser consistente passa indubitavelmente por uma análise de ajuste dos dados ao modelo psicométrico utilizado no processo de medição. Somente em modelos com um bom

ajuste, teremos como fazer uma análise consistente dos resultados. No caso da modelagem MFRM teremos que analisar os ajustes de todas as facetas utilizadas no modelo, passando naturalmente pelos ajustes das duas facetas fundamentais do modelo: itens e alunos.

No caso dos itens, o ajuste é a comparação entre o que o modelo previu e o que é observado, obtido através do cálculo do resíduo (valor estimado menos valor observado). Apresentamos, nos Gráficos 1 e 2, as curvas características de dois itens, com os valores estimados pelo modelo e os valores observados.

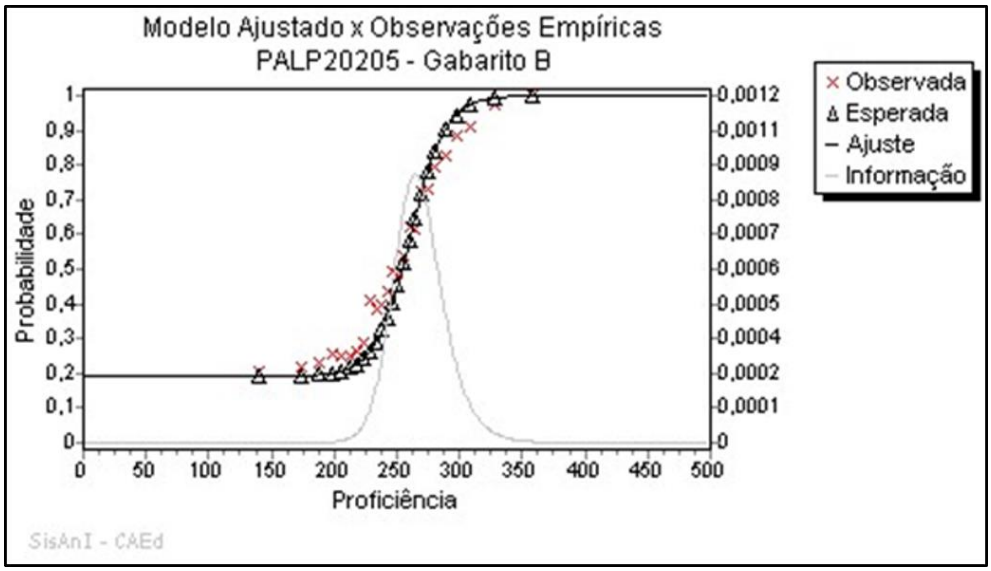


Gráfico 1 – Análise de ajuste do item: item ajustado

Fonte: CAEd/UFJF (2018).

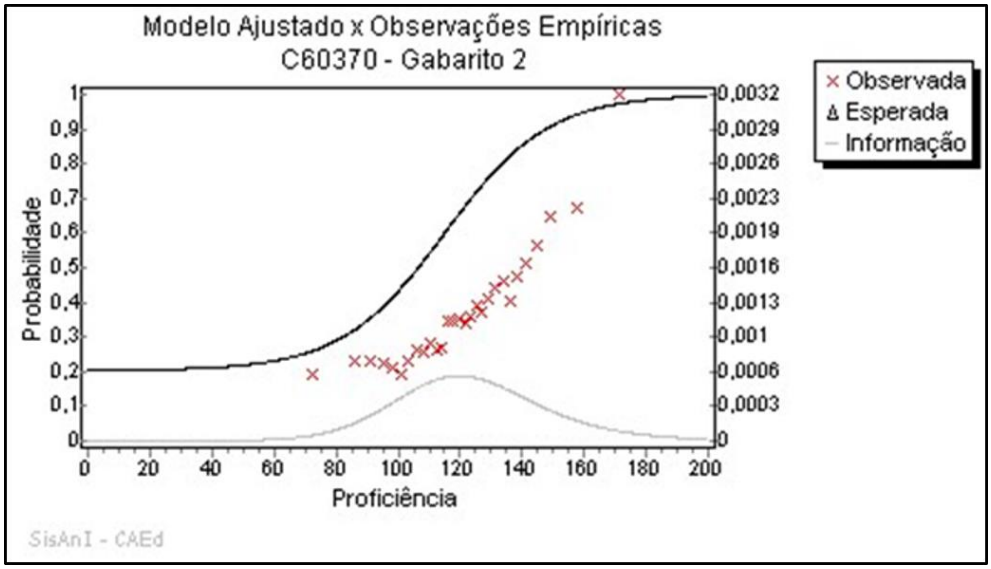


Gráfico 2 – Análise de ajuste do item: item não ajustado

Fonte: CAEd/UFJF (2018).

No Gráfico 1 temos um exemplo de item bem ajustado ao modelo e no Gráfico 2 um exemplo de um item que não se ajustou ao modelo.

No caso de alunos, se levarmos em consideração dois alunos, em que o primeiro acerta os itens fáceis e erra os itens difíceis de um teste, e o segundo aluno erra os itens fáceis e acerta os itens difíceis, temos que o segundo aluno apresenta um comportamento fora do padrão normalmente observado, e, portanto, um desajuste.

Ao analisarmos essas situações, nos deparamos com as seguintes questões: qual deverá ser o padrão de ajuste? E o que fazer se esse padrão não for atingido?

Com relação à primeira indagação, Golino et al. (2015) afirmam que não existe um padrão-ouro a ser seguido, os índices de ajustes devem ser interpretados de maneira relativa e de forma contextualizada.

A segunda indagação remete-nos a dois paradigmas relacionados ao desajuste: o paradigma tradicional e o paradigma Rasch. No paradigma tradicional buscam-se modelos mais complexos fora da família de modelos Rasch, geralmente os modelos de dois parâmetros logísticos (2PL) ou 3PL. Nesse caso, é como se estivéssemos forçando o modelo a se ajustar aos dados. Essa solução é mais bem vista e adotada por estatísticos.

No paradigma Rasch, o desajuste é visto como uma anomalia, o problema não está no modelo, mas sim na qualidade dos itens/ testes que não estão alinhados com o constructo a ser medido. Uma forma de contornar esse problema seria, por exemplo, rever os itens ou fazer uma análise crítica da característica da população em relação ao constructo medido. Nas avaliações educacionais, é comum termos desajustes pelo fato de os alunos não conhecerem o que se está sendo avaliado[4]. Essa forma de tratar o desajuste é normalmente adotada por profissionais das áreas das ciências humanas e naturais.

Vemos, portanto, que o trabalho do psicometrista na modelagem de uma realidade, em nosso caso a realidade educacional, transcende o uso pragmático do melhor modelo, incorporando uma utilização o mais consciente possível do nível de imprecisão do modelo adotado e das inferências possíveis aplicadas à realidade medida.

[4] Nas avaliações realizadas pelo CAEd/UFJF no terceiro ano do ensino médio, tem-se verificado maiores índices de desajustes relacionados à falta de conhecimento dos alunos em disciplinas das áreas de Ciências da Natureza e Matemática do que em Ciências Humanas e Língua Portuguesa.

Esse é um dos pontos centrais dessa tese, ou seja, apresentar aplicações práticas para modelos da família Rasch, cujo interesse principal é uma medida de boa qualidade da proficiência dos alunos e que apresenta a grande vantagem de ter uma interpretação pedagógica mais simples, facilitando o trabalho dos professores ao utilizarem os resultados das avaliações em suas atividades em sala de aula.

2.5 Erro de medida

Temos que ter em mente que toda medida tem erro que está relacionado com o instrumento utilizado e, também, com o próprio operador desse instrumento. Entretanto, essa medida tem que possuir um nível de precisão (fidedignidade) que possibilite sua utilização de forma confiável.

Normalmente não temos muita consciência dos erros das medidas, principalmente aquelas utilizadas no nosso dia a dia, como comprimento, temperatura, velocidade e etc., pois as interpretações que fazemos dessas medidas não exigem precisões elevadas. Para comprarmos uma mesa, definir o tipo de roupa que usaremos em determinada viagem ou mesmo a velocidade do carro, os níveis de precisão não são importantes.

Entretanto, na área educacional, as medidas de desempenho obtidas pelos alunos são muitas vezes utilizadas em processos seletivos ou políticas de bonificação para escolas, municípios ou estados; nessas situações, quanto mais precisas forem as medidas, mais justos serão os resultados fornecidos aos envolvidos no processo em questão.

3. Diferentes possibilidades de modelagem à luz da TRI

Quando nos referimos a um modelo, temos que ter em mente que estamos lidando com a tentativa de representar uma determinada realidade. Em avaliação educacional, ao utilizarmos os modelos da TRI, o objetivo é a construção de uma escala de conhecimento ou escala de proficiência. Para tanto, são utilizados itens representativos do constructo a ser medido. Trabalharemos, nessa tese, com dois tipos específicos de itens: itens dicotômicos e itens politômicos. Os itens dicotômicos estão relacionados com itens do tipo de múltipla escolha em que se tem apenas uma resposta correta. Na construção de uma escala em um teste composto apenas por itens dicotômicos, utilizam-se modelos da TRI constituídos por duas facetas, uma faceta para o item e outra faceta para aluno. À faceta do item estão relacionados os parâmetros do item e à faceta do aluno está relacionada a proficiência.

Já os itens politômicos, não são corrigidos em certos ou errados, mas sim em uma escala do tipo Likert, com categorias variando do totalmente errado ao totalmente certo. A definição da categoria em que se encontra a resposta de um aluno depende da percepção do juiz. Os testes construídos com esse tipo de item são modelados pela TRI segundo três facetas: as facetas para alunos e itens, mencionadas previamente, e a inclusão de uma terceira faceta para os juízes ou corretores de forma a modelar suas severidades. Esse tipo de modelagem que leva em consideração a severidade do corretor foi desenvolvido por Linacre (1989), que estabeleceu os fundamentos teóricos dessa modelagem, considerada como uma extensão dos modelos Rasch, e denominada modelagem Rasch multifacetas. Essa modelagem será mais detalhada no próximo capítulo.

No presente capítulo, apresentaremos a TRI para modelagens com duas facetas em testes unidimensionais em que são utilizados itens dicotômicos. Essas características são encontradas na quase totalidade das avaliações em larga escala no Brasil. Apresentaremos, nos tópicos seguintes, os fundamentos da TRI, os procedimentos para a produção de medidas, a interpretação das escalas de proficiência e experiências brasileiras de utilização dos resultados de uma avaliação em larga escala, como uma ferramenta para a melhoria da qualidade do ensino.

3.1 A modelagem da realidade educacional pela TRI

Existem duas possibilidades de análise dos desempenhos dos alunos em um teste. Uma primeira possibilidade, muito comum e mais empregada nas atividades docentes, por ser uma ferramenta de fácil manuseio, consiste no cálculo do percentual de acerto do aluno no teste, que gera a nota do aluno ou escore. Esse procedimento caracteriza a TCT.

A maioria das análises realizadas a partir da TCT é focada no escore obtido no teste. Assim, um aluno que responde a uma série de itens e recebe um ponto por cada item corretamente respondido, obtém, ao final, um *escore total* (que é a soma destes pontos). Contudo, sob essa perspectiva, é possível, ou até mesmo esperado, que alunos obtenham notas mais altas em testes fáceis e notas mais baixas em testes difíceis. Ou seja, os escores dos examinandos dependem do teste utilizado, são "teste-dependentes".

A segunda possibilidade para se medir o desempenho dos alunos é normalmente utilizada nas avaliações educacionais em larga escala na qual o procedimento de análise da avaliação é obtido por meio da TRI. Nesta metodologia, o desempenho do aluno, denominado proficiência, não é apenas uma nota, é uma medida de conhecimento, estabelecida em função de uma matriz de habilidades construída para o teste como, por exemplo, na matriz apresentada na figura 5.

3.1.1 A TCT e a TRI no ambiente escolar

Nas avaliações realizadas em sala de aula pela TCT, quando um professor elabora um teste, ele tem uma noção de quais são os itens mais fáceis e os mais difíceis e, além disso, de quais alunos terão grandes possibilidades de acertarem ou errarem determinados itens, ou seja, o professor sabe quem são os alunos bons e que acertarão os itens fáceis e difíceis e quem são os alunos mais fracos, que acertarão apenas os itens mais fáceis.

Essa percepção (*feeling*) é fundamental para que ele possa trabalhar com seus alunos, direcionado esforços no sentido de ter toda a turma aprendendo a disciplina, ou seja, alunos que ele classificou como fracos terão mais a sua atenção e alunos que ele classificou como bons não necessitarão tanto de sua atenção, pois

assimilam a disciplina com mais facilidade.

Essa prática, comum na realidade escolar, é exercida por diversos professores, com sensibilidades (severidade) diferentes. Podemos imaginar uma situação hipotética de duas turmas que têm o mesmo desempenho, mas seus professores têm opiniões diferentes, um acha que a turma é muito boa e o outro acha a turma muito fraca.

Diante dessa realidade no ambiente escolar, fica inviável, por exemplo, para um gestor, responsável por uma rede de ensino, dialogar com os professores sobre a qualidade do ensino, pois, não existe uma métrica comum, mas sim, uma métrica individual em função da sensibilidade de cada professor. Ou seja, nas avaliações via TCT, os resultados dos alunos só atendem às necessidades dos professores, não permitem comparações entre escolas e não se têm uma escala interpretável.

A utilização da TRI aproxima a avaliação dessa situação, permitindo realizar a quantificação do conhecimento dos alunos, por meio de uma medida de proficiência, e as características dos itens, por meio dos parâmetros de discriminação, dificuldade e acerto ao acaso (no caso de uma modelagem com três parâmetros). Uma das características dessa modelagem é o fato de se colocar tanto a proficiência dos alunos quanto os parâmetros de dificuldade dos itens em uma mesma métrica ou escala, eliminando a subjetividade oriunda de diferentes percepções dos professores de uma mesma realidade.

Na Figura 3 representamos com bolinhas numeradas de “1” a “20” os itens de um determinado teste, e com colunas, os percentuais de alunos. A linha horizontal variando de “100” a “500”, representa uma escala obtida pela TRI.

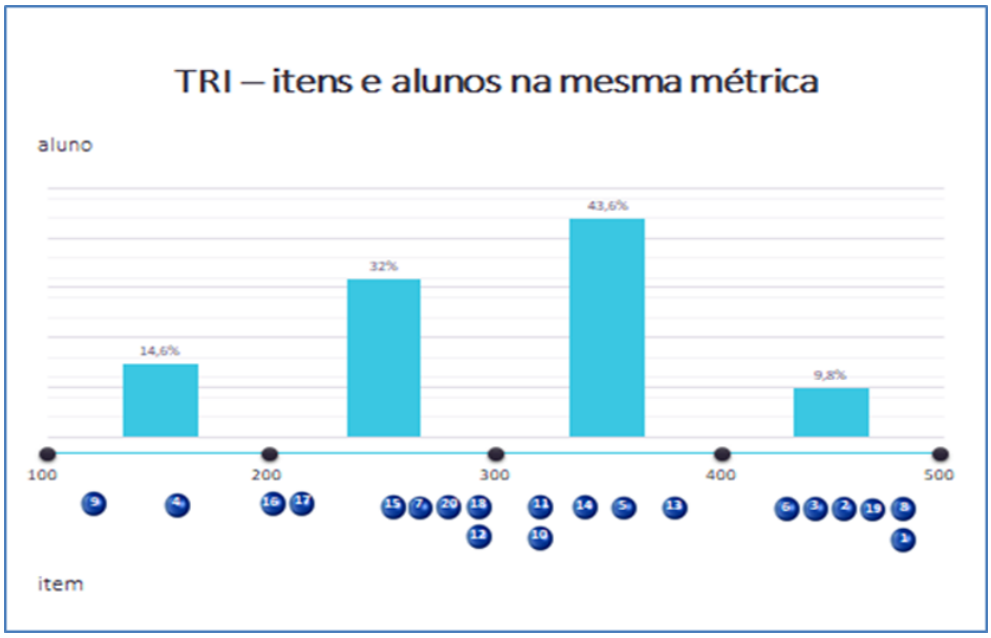


Figura 3 – Representação da TRI (Proficiência e dificuldade na mesma métrica)
Fonte: O Pesquisador.

Como a TRI posiciona em uma mesma escala a dificuldade dos itens e a proficiência do aluno, podemos realizar as seguintes análises:

- ✓ A dificuldade do item e a proficiência dos alunos estão na mesma métrica.
- ✓ Os itens estão ordenados pelas suas dificuldades, o item mais fácil é o "9" e os mais difíceis são os itens "1" e "8".
- ✓ A maioria dos alunos se encontra no meio da escala, temos 14,6% de aluno fracos e 9,8% de alunos bons.
- ✓ Os alunos fracos, posicionados na faia de 100 a 200, conseguem realizar apenas as habilidades relacionadas aos itens "4" e "9" , os alunos medianos, localizados na faixa de "300" a "400", conseguem realizar as tarefas relacionadas aos itens posicionados nessa faixa e nas faixas anteriores, mas não conseguem realizar as tarefas da faixa de "400" a "500", e os alunos bons, faixa de "400" a "500", conseguem realizar todas as tarefas, embora possam ter dificuldade de realizar as tarefas muito difíceis, representadas pelos itens do final da escala. Esse tipo de análise, é a interpretação pedagógica da escala.

Procuramos, com o auxílio de uma situação hipotética, representada pela figura 3 descrever os dois pontos fundamentais da TRI que justificam a sua grande utilização no meio educacional: a comparabilidade e a interpretação pedagógica dos

resultados, o que permite aos professores e gestores elaborarem planos de trabalhos para cada escola.

3.1.2 Aplicabilidade da TRI no ambiente escolar

A TRI é um conjunto de modelos matemáticos em que a probabilidade de acerto a um item é calculada em função da proficiência do aluno e dos parâmetros dos itens. Podemos destacar as seguintes vantagens dessa metodologia: (i) permite a comparação longitudinal de resultados de diferentes avaliações como, por exemplo, os resultados de avaliação dos sistemas estaduais e municipais de ensino com os resultados do Saeb, desde que se incluam itens comuns entre os testes e se conservem os mesmos critérios na construção e organização dos mesmos; (ii) permite avaliar com alto grau de precisão e abrangência uma determinada área do conhecimento, sem que cada aluno precise responder a longos testes; (iii) permite a comparação entre diferentes séries, por exemplo, 5º e 9º ano do ensino fundamental e 3º ano do ensino médio, pela construção de uma escala única de resultados para essas três séries, e; iv) permite a interpretação pedagógica da escala com informações das habilidades e quantidades de alunos por níveis de desempenho. Essa informação, se trabalhada na escola, auxilia nas práticas pedagógicas dos professores e consequentemente na melhoria da qualidade de ensino.

Para entender um pouco mais sobre a TRI no contexto em que essa modelagem é utilizada no Brasil, é necessário conhecer basicamente: como os itens são modelados, como os itens e testes são elaborados, como é construída uma escala de conhecimento, como é possível ter duas avaliações distintas em uma mesma escala, como esta escala é interpretada e como utilizar os resultados de uma avaliação em larga escala para a melhoria da qualidade de ensino.

Nos próximos tópicos, abordaremos as questões apresentadas acima, com o foco na utilização dos resultados das avaliações pelos sistemas de ensino e não nos aspectos mais densos dos modelos matemáticos. Essas explanações também serão úteis para a compreensão dos capítulos empíricos dessa tese.

3.2 Modelagem dos itens pela TRI

Os modelos matemáticos de duas facetas da TRI, para itens dicotômicos, são definidos em função do número de parâmetros utilizados para modelar o item. Apresentamos a equação mais geral, desenvolvida por Lord (1980 apud Pasquali, 2011):

$$\boldsymbol{P(X_{ni} = 1|\theta_n, a_i, b_i, c_i) = c_i + (1 - c_i)\frac{e^{Da_i(\theta_n - b_i)}}{1 + e^{Da_i(\theta_n - b_i)}}}$$

Nessa equação, a probabilidade de um aluno "n" acertar um item "i', depende da habilidade desse aluno representado pela letra grega "θ" e das características do item, representado pelos parâmetros "a", "b" e "C".

Em que,

- Parâmetro *a*: é a capacidade do item de discriminar os alunos que desenvolveram habilidades daqueles que não desenvolveram.
- Parâmetro *b*: está relacionado ao percentual de alunos que respondem corretamente ao item. Assim, quanto menor o percentual de acerto, maior a dificuldade do item.
- Parâmetro *c*: é a probabilidade de acerto ao acaso; leva em consideração a probabilidade de o aluno "chutar" e acertar o item.
- θ : Proficiência do aluno.
- D: fator de escala em que se utiliza o valor 1.7. Desta forma, os resultados da função logística se assemelham aos resultados da função normal

Tomando por base essa equação, os seguintes modelos da TRI:

i) Modelo logístico de três parâmetros (3PL): os três parâmetros variam no processo de modelagem de cada item.
ii) Modelo logístico de dois parâmetros (2PL): os parâmetros "a" e "b" variam no processo de modelagem do item, o parâmetro "c" é igual a zero.
iii) Modelo logístico de um parâmetro (1PL): apenas o parâmetro "b" varia no processo de modelagem do item, o parâmetro "a" é fixo e idêntico para todos os itens e o parâmetro "c" é igual a zero.
iv) Modelo Rasch: apenas o parâmetro "b" varia no processo de modelagem do item, o parâmetro "a" é fixo e idêntico para todos os itens e o parâmetro "c" é igual a zero. Nessa modelagem, a média dos parâmetros "b" de um determinado teste é igual a zero.

Ao se variar os valores dos parâmetros dos itens, teremos diferentes curvas representativas dos mesmos. Essas curvas que possuem uma forma monotônica crescente, são denominadas de Curvas Características dos Itens (CCI). Um exemplo de CCI para um item modelado em 3PL, é apresentado no Gráfico 3.

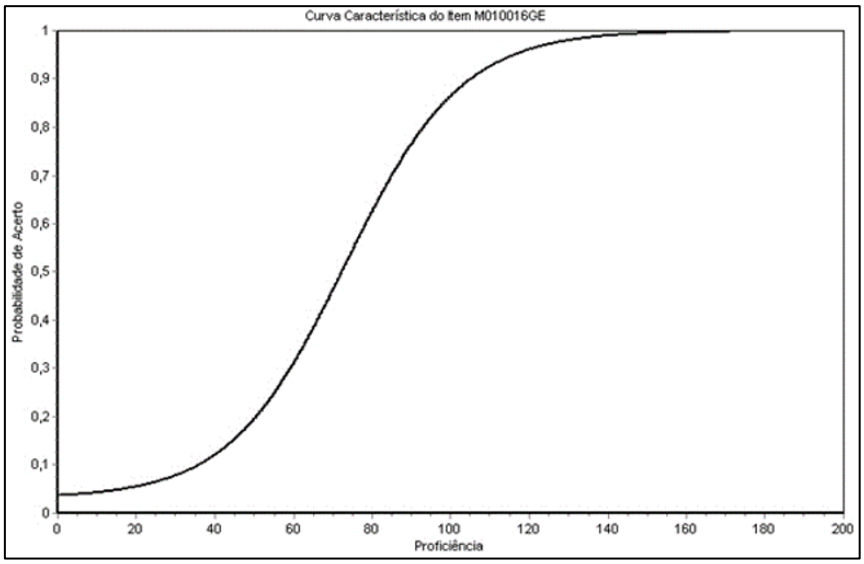


Gráfico 3 – CCI segundo um modelo 3PL

Fonte: CAEd/UFJF (2018).

Nesse gráfico, o eixo horizontal diz respeito à proficiência dos examinados e é representada pela letra grega teta (θ), e o eixo vertical refere-se à probabilidade de um indivíduo com dada habilidade escolher a resposta correta. Observa-se que alunos com maior habilidade possuem maior probabilidade de acertar o item e que essa relação não é linear.

É importante termos uma visualização destes parâmetros na curva característica do item. Assim, o parâmetro ‘a’ seria a inclinação da curva medida no ponto ‘b’. Quanto maior o parâmetro a, mais inclinada é a curva e, portanto, maior o poder de discriminação do item, ou seja, o item separa bem os alunos que dominam determinada habilidade dos alunos que não dominam. O parâmetro ‘b’, que está na mesma métrica da habilidade, representa a dificuldade do item, e por definição é o valor de habilidade que corresponde a uma probabilidade de acerto de 50% mais o parâmetro c dividido por dois. O parâmetro ‘c’ é o valor no eixo vertical, que vai de zero até o início da curva, ou seja, é a probabilidade de acerto para quem tem uma proficiência muito baixa.

Conforme apresentado por Oliveira & Franco Jr. (2008), para uma melhor interpretação da relação entre as características do item e as habilidades dos alunos, podemos nos valer do gráfico da Figura 4, que permite distinguir as três fases dos processos de aprendizagem.

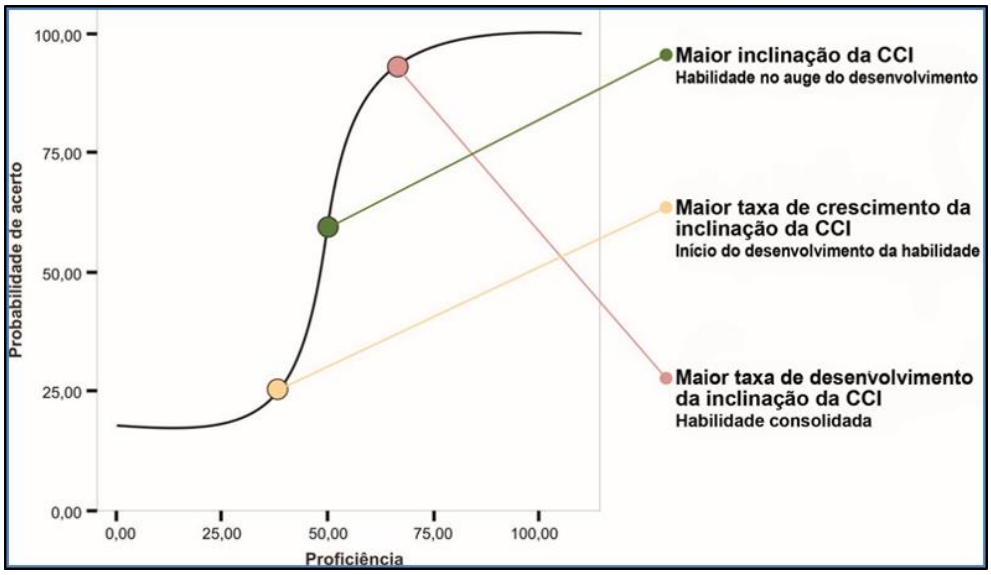


Figura 4 – Estágios do processo de aprendizagem modelados segundo a CCI
Fonte: Oliveira & Franco Jr. (2008).

A primeira fase, assinalada pela cor amarela, indica o nível de proficiência em que os alunos passam a ter maiores condições de desenvolver a habilidade. O segundo ponto, assinalado pela cor verde, indica o parâmetro de dificuldade do item. Em torno desse ponto a habilidade está em rápido desenvolvimento, a CCI atinge a mais elevada inclinação. Esse ponto é um delimitador (Threshold), ou seja, consegue separar os alunos que desenvolveram a habilidade testada (acima do ponto) daqueles que ainda não atingiram essa etapa (abaixo do ponto). O terceiro ponto, rosa, sinaliza a consolidação da aprendizagem, quando observamos a estabilização (não discriminação) da curva.

O entendimento desta relação entre a curva do item a aprendizagem é essencial para a interpretação das escalas de conhecimento

3.3 Elaboração de itens e construção de testes

Basicamente, os testes padronizados usados em avaliação educacional em larga escala são compostos de itens de múltipla-escolha. Os itens são elaborados segundo uma Matriz de Referência, Figura 5, composta por descritores de desempenho por área de conhecimento a ser avaliada.

Matriz de Referência de Língua Portuguesa - SPAECE	
3º ano do Ensino Médio	
I. Procedimentos de Leitura	
D01	Localizar informação explícita.
D02	Inferir informação em texto verbal.
D03	Inferir o sentido da palavra ou expressão.
D04	Interpretar textos não verbais e textos que articulam elementos verbais e não verbais.
D05	Identificar o tema ou assunto de um texto.
D06	Distinguir fato de opinião relativa ao fato.
D07	Diferenciar a informação principal das secundárias em um texto.
II. Implicações do Suporte, do Gênero e/ou do Enunciador na Compreensão do Texto	
D09	Reconhecer gênero discursivo.
D10	Identificar o propósito comunicativo em diferentes gêneros.
D11	Reconhecer os elementos que compõem uma narrativa e o conflito gerador.
III. Relação entre Textos	
D12	Identificar semelhanças e/ou diferenças de ideias e opiniões na comparação entre textos.
D13	Reconhecer diferentes formas de tratar uma informação na comparação de textos de um mesmo tema.
IV. Coerência e Coesão no Processamento do Texto	
D14	Reconhecer as relações entre partes de um texto, identificando os recursos coesivos que contribuem para sua continuidade.
D15	Identificar a tese de um texto.
D16	Estabelecer relação entre tese e os argumentos oferecidos para sustentá-la.
D17	Reconhecer o sentido das relações lógico-discursivas marcadas por conjunções, advérbios, etc.
D18	Reconhecer o sentido do texto e suas partes sem a presença de marcas coesivas.
V. Relações entre Recursos Expressivos e Efeitos de Sentido	
D19	Reconhecer o efeito de sentido decorrente da escolha de palavras, frases ou expressões.
D20	Identificar o efeito de sentido decorrente do uso da pontuação e de outras notações.
D21	Reconhecer o efeito decorrente do emprego de recursos estilísticos e morfossintáticos.
D22	Reconhecer efeitos de humor e ironia.
VI. Variação Linguística	
D23	Identificar os níveis de linguagem e/ou as marcas linguísticas que evidenciam locutor e/ou interlocutor.

Figura 5 – Matriz de referência para Língua Portuguesa 3º ano do ensino médio
Fonte: CAEd/UFJF (2018).

Cabe aos especialistas da disciplina avaliada, elaborar itens representativos de cada descritor, segundo padrões técnicos, que serão testados tanto pela TCT quanto pela TRI de forma a serem validados.

A decisão sobre o número de itens é um ponto importante na composição do instrumento de medida. Por um lado, o teste deve conter muitos itens, pois um

dos objetivos da avaliação em larga escala é medir de forma abrangente as competências essenciais do período de escolaridade a ser avaliado, de forma a garantir a cobertura de toda a matriz de referência adotada. Por outro lado, o teste não pode ser longo, pois inviabiliza sua resolução pelo examinando. Para solucionar essa dificuldade têm-se utilizado um tipo de planejamento de testes denominado Blocos Incompletos Balanceados (BIB).

No BIB, os itens são organizados em blocos e esses, por sua vez, constituem cadernos de teste. Com o uso do BIB, é possível elaborar muitos cadernos de teste diferentes para serem aplicados a alunos de uma mesma série. Podemos destacar duas vantagens na utilização desse modelo de montagem de teste: a colocação de um maior número de itens em circulação no teste, avaliando, assim, uma maior variedade de habilidades (normalmente tem-se de 3 a 4 itens para cada descritor da matriz de referência), bem como um equilíbrio em relação à dificuldade dos cadernos de teste, uma vez que os blocos são colocados em diferentes posições nos cadernos, evitando, dessa maneira, que um caderno fique mais difícil que outro, o que poderia gerar um efeito de cansaço nos alunos ao responderem a testes mais difíceis.

Apresentamos no quarto capítulo, subtópico 5.1.1 um delineamento de BIB utilizado pelo Saeb e por praticamente todas as avaliações estaduais que possuem seus resultados na escala do Saeb.

3.4 Produção de medidas pela TRI

O processo de produção de medidas, envolve, em linhas gerais, a preparação da base de dados, a confecção de uma sintaxe para utilização em *software* específico da modelagem a ser utilizada. Nessa tese, utilizamos o BILOG-MG, o MIRT (Chalmers, 2012) do *Software* R e o *FACETS* (Linacre, 2014) e, controles realizados em pontos críticos ao longo do processo, envolvendo técnicas estatísticas específicas, de forma a obter um bom ajuste do modelo ao constructo avaliado e dessa forma, garantir a qualidade das medidas produzidas

Conforme Denny Borsboom (em “The attack of the Psychometricians”, Psychometrika, 2006 apud Golino et al., 2015), o nível de adequação do modelo à realidade é o desfio que o psicometrista envolvido em um projeto deve procurar ter

sobre controle. Para tanto, técnicas quantitativas e qualitativas devem ser empregadas para se atingir essa finalidade.

Levando essa premissa em consideração, apresentamos no Diagrama 1, o fluxo de atividades que envolve o processo de produção de medidas pela TRI.

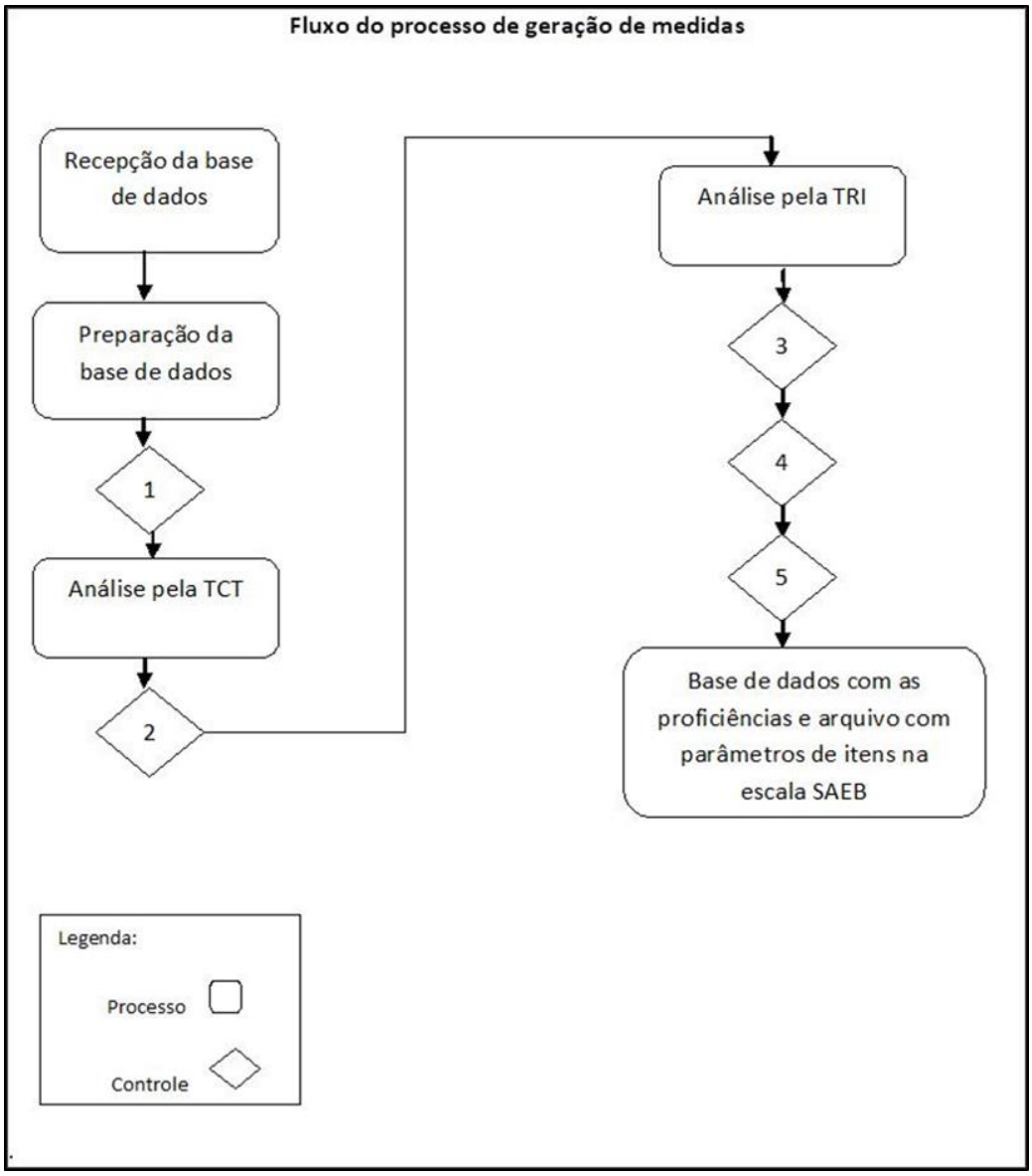


Diagrama 1 – Fluxograma de produção de medidas pela TRI
Fonte: CAEd/UFJF (2018).

Nesse fluxograma, estão identificados os processos e os pontos de controle. Os pontos de controle têm a finalidade de detectar anomalias que trarão como consequências estimativas erradas de proficiências e parâmetros de itens. No Anexo 2 (plano de controle), detalhamos todos os controles e respectivos métodos utilizados, assim como os procedimentos a serem adotados em caso de detecção de

falhas. Essa estrutura de avaliação é utilizada pelo Saeb e por todas as avaliações nacionais em larga escala que têm suas medidas na escala Saeb[5].

Conforme pode ser observado, são procedimentos muito técnicos, da área da psicometria e que, a princípio, não têm como ser replicados dentro do ambiente escolar, ou seja, o professor não tem como calcular a proficiência do aluno. Veremos, no terceiro capítulo, que em determinadas situações, especialmente com a utilização dos modelos Rasch, é possível o professor calcular a proficiência do aluno, devido à relação biunívoca[6] entre proficiência e percentual de acerto existente nesse modelo, possibilitando aos professores se apropriarem dos resultados da avaliação de forma mais rápida.

3.5 A escala de avaliação educacional

Quando lidamos com fenômenos físicos que possuem uma nulidade, ou seja, um zero absoluto, temos uma escala que começa no zero, como por exemplo, a altura, o peso e a idade de uma pessoa.

Conforme já relatamos anteriormente, a nota que um aluno recebe em um teste corrigido por seu professor, não é uma medida, mas uma quantificação que tem o valor zero como referência. Uma das maiores dificuldades ao trabalharmos os resultados de avaliações educacionais pela TRI, é que a medida obtida de desempenho do aluno, não possui o valor zero como origem da escala, pois não existe a nulidade de conhecimento. Temos, nesse caso, um exemplo de escala intervalar.

O que se obtém por meio dos modelos matemáticos da TRI é uma representação do constructo conhecimento em uma escala, através de uma curva normal padronizada variando de menos infinito a mais infinito com média zero e desvio-padrão igual a “1”. Apresentamos na Figura 6 uma representação da escala obtida pela modelagem da TRI.

[5] Para informações mais detalhadas dos métodos estatísticos adotados nas etapas mencionadas, consultar Valle (1999), Klein (2003), Kolen (2004), Silva (2010) e Rocher (2013).
[6] Cada valor de proficiência se associa a um único valor de percentual de acerto.

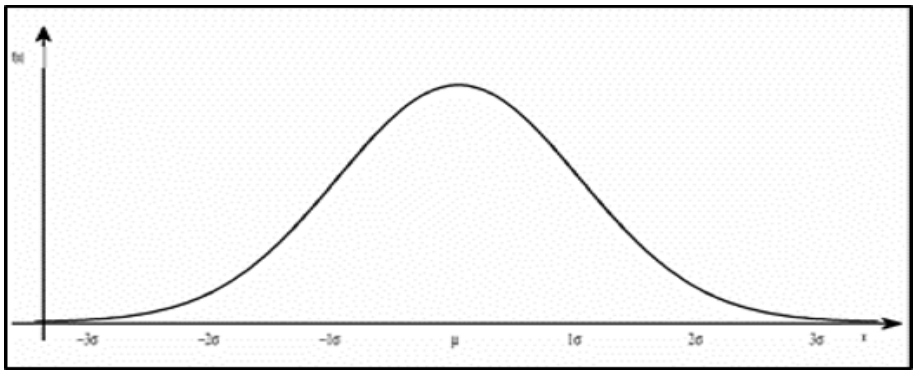


Figura 6 – Escala pela TRI

Fonte: O Pesquisador.

No eixo “x” temos a proficiência do aluno e no eixo “y” a quantidade de alunos. No valor zero, temos a média da população.

Uma escala com essas características não se enquadra bem no meio educacional, pois seria “estranho”, para um aluno receber uma classificação de desempenho negativa, por exemplo “-1”, e o professor falar para esse aluno que sua “nota” é muito boa, pois a média da escola é “-2”. Para não se trabalhar com números negativos, o que se faz na prática é multiplicar e somar a pontuação de cada aluno por duas constantes, de forma a termos uma escala somente com valores positivos.

Dessa forma em nosso exemplo, o professor diria para o aluno que sua “nota” foi 200 e que a “nota” da escola foi 150, o que é uma representação bem mais fácil de assimilar, pois está mais adequada com o que é normalmente utilizado no ambiente escolar. Mesmo com essa transformação, o fato de uma medida não ter uma referência em zero é um dos grandes motivos de dificuldade de entendimento por parte dos professores dos resultados das avaliações em larga escala. Para contornar esse problema, é fundamental, fornecer aos professores noções básicas de medidas e estatística de forma a permitir que eles possam utilizar os resultados da avaliação de forma adequada. Somente dessa forma, é possível fazer os atores envolvidos na análise dos resultados da avaliação deixar de pensar na medida do conhecimento a partir do zero, o que como dissemos não faz sentido, pois não existe a nulidade de conhecimento, e pensar a medida a partir do valor médio obtido pelos alunos. Embora pareça simples, essa mudança de paradigma é um dos grandes entraves para a correta utilização dos resultados de uma avaliação em larga escala.

3.5.1 A equiparação de medidas entre escalas – métodos de equalização

Uma das grandes aplicabilidades da TRI em avaliações de sistemas educacionais é a comparabilidade de resultados entre avaliações distintas e em diferentes períodos de aplicação. Surge então a pergunta: como é possível ter duas avaliações distintas em uma mesma escala?

Após a aplicação de uma avaliação em larga escala, as respostas de cada aluno a cada item do teste são processadas de forma a constituir uma base de dados. Através desta base de dados e a utilização da TRI, são calculados os parâmetros dos itens e as proficiências dos alunos vinculados a uma escala específica dessa avaliação, que chamaremos, por exemplo, escala X. Em seguida, caso se deseje colocar as medidas dessa escala X em uma outra escala, por exemplo, escala Y já existente para fins de comparabilidade entre as duas avaliações, realizam-se procedimentos matemáticos denominados equalizações, de forma a colocar as proficiências e parâmetros dos itens da escala X na escala Y.

A condição fundamental para que se realize uma equalização é que se tenham itens comuns entre as mesmas, dessa forma, é possível optar por diferentes métodos de equalizações, como, por exemplo, os métodos lineares e os métodos não-lineares. Os métodos lineares são compostos pelos métodos do tipo média-sigma, média-média e curva característica; e os métodos não-lineares pelos métodos de calibração simultânea e o método de prefixação dos parâmetros. A decisão de qual método utilizar vai depender das características das avalições. Nas avaliações do Saeb e nas avaliações estaduais que têm seus resultados na escala Saeb, é utilizado o método de prefixação de parâmetros. Para informações mais detalhadas sobre esses métodos, consultar Valle (1999), Klein (2003), Kolen & Brennan (2004), Silva (2010) e Rocher (2013).

Também em Kolen & Brennan (2004) e Silva (2010) são apresentados fatores que afetam a robustez das comparabilidades entre avaliações. Ter apenas itens em comum e adotar métodos de equalização não garantem uma boa comparabilidade. Aspectos como delineamentos diferentes dos testes, motivação dos alunos e posição dos itens comuns, se sofrerem alterações significativas entre duas avaliações a serem equalizadas, podem levar a problemas de comparabilidade.

3.5.2 A escala Saeb para o ensino fundamental e médio

A escala Saeb, construída em 1997 possui as seguintes características:

- ✓ Uma escala para Língua Portuguesa e Uma escala para matemática
- ✓ Escala única da 4ª série/5º ano do Ensino Fundamental ao 3º ano do Ensino Médio. variando em termos práticos de 0 a 500 pontos, obtida através de itens comuns entre a 4ª série e a 8ª série do Ensino fundamental; e itens comuns da 8ª série do ensino fundamental com o 3º ano do ensino médio.
- ✓ A referência para esta escala foi a 8ª série/9º ano ensino fundamental, em uma amostra representativa de todos os alunos brasileiros no ano de 1997, na qual foi estipulado uma média de 250 pontos e um desvio padrão de 50, tanto para Língua Portuguesa quanto para Matemática.
- ✓ O método de equalização utilizado de 1997 até os dias atuais é o de prefixação de parâmetros. Os procedimentos de equalização entre as avaliações do Saeb estão detalhados em Klein (2003).

Apresentamos no Anexo 1 um exemplo de escala de proficiência em Língua Portuguesa, equalizada com o Saeb, com a identificação onde ocorrem as ancoragens[7] das competências avaliadas por faixas de proficiências.

3.6 Avaliação em larga escala e qualidade de ensino

Apresentaremos nesse tópico exemplos de utilização de resultados das avaliações em larga escala com o objetivo de se melhorar a qualidade do ensino tanto em âmbito nacional como estadual. Dentre os quais destacamos: O Índice de Desenvolvimento da Educação Básica (Ideb), Plataformas de devolutivas de resultados e estudos de eficácia escolar.

[7] Ancoragem: Valor de proficiência que corresponde a uma probabilidade de acerto ao item de 65%.

3.6.1 O Índice de Desenvolvimento da Educação Básica (Ideb)

A avaliação educacional no Brasil, com a utilização da TRI, deu-se início com a avaliação do Saeb no ano de 1995 com a criação da escala nacional pra Língua Portuguesa e Matemática. Apresentamos nas Figuras 7 e 8, as séries históricas dos resultados, a partir de 2005 para Língua Portuguesa e Matemática, respectivamente.

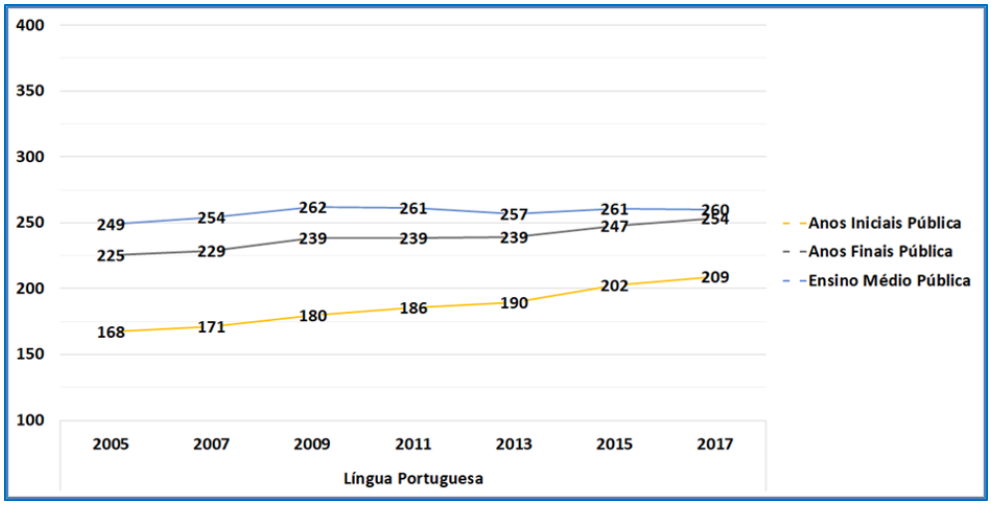


Figura 7 – Série histórica Saeb Língua Portuguesa a partir de 2005

Fonte: Saeb (2018).

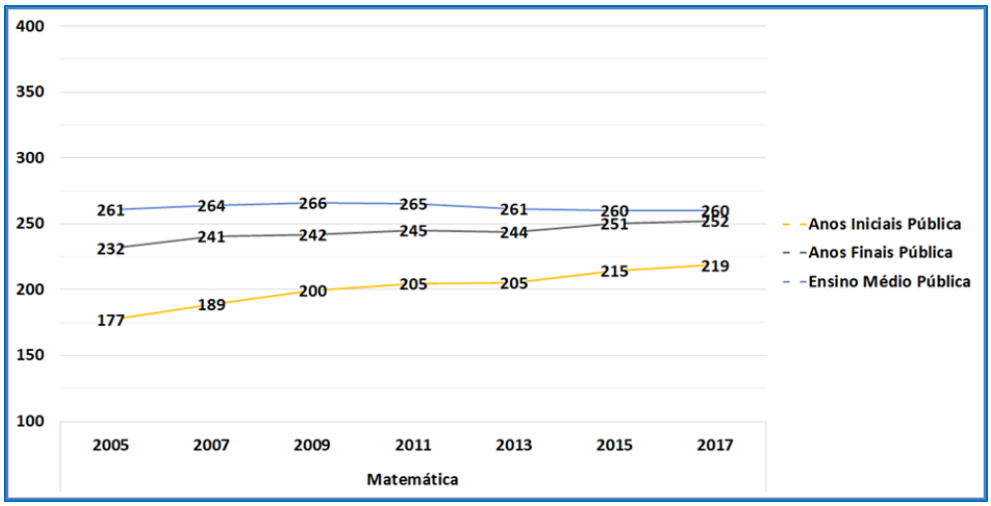


Figura 8 – Série histórica Saeb Matemática a partir de 2005

Fonte: Saeb (2018).

Um fato marcante nessa trajetória, foi a criação do Ideb no ano de 2005, um indicador com a finalidade de se mensurar a qualidade da educação no país,

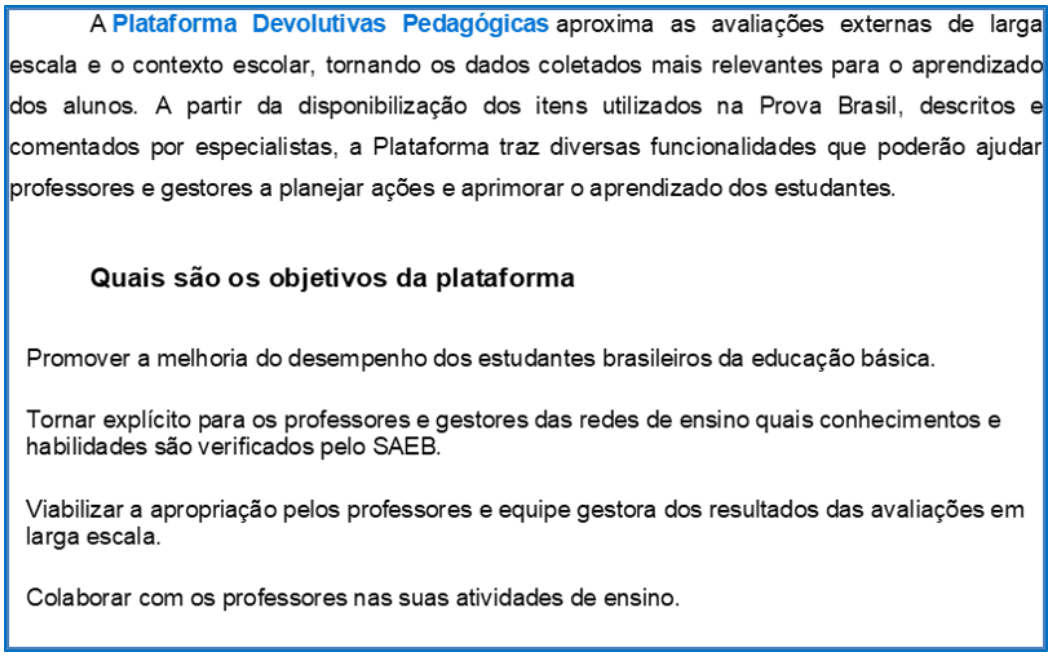
conforme apontado pelo seu idealizador Reynaldo Fernandes (Fernandes, 2007). Para a construção desse indicador, que utiliza o fluxo escolar e a proficiência dos alunos no 5º e 9º ano do Ensino fundamental houve uma mudança muito significativa nas avaliações do Saeb: as avaliações do 5º e 9º ano do Ensino fundamental deixaram de ser amostrais e passaram a ser censitárias, pois o objetivo era construir um indicador e metas para cada escola pública brasileira. A partir dessa data houve uma grande expansão das avaliações em larga escala no país com a finalidade de se monitorar o desempenho das escolas. As secretarias estaduais de educação sentiram necessidade de realizarem suas próprias avaliações, para poderem ter mais domínio dos níveis de desempenho de suas escolas, e dessa forma cumprir as metas do Ideb.

Observam-se, a partir desse período, diversas ações e políticas públicas de accountability vinculadas aos resultados das avaliações desenvolvidos por alguns estados brasileiros, com o objetivo de melhorar o desempenho das escolas e, dessa forma, atingir as metas de desempenho estipuladas pelo Inep. Não é o foco dessa tese discutir as vantagens e/ou desvantagens de tais políticas, para investigações com essa finalidade, consultar Brooke (2006) e Freitas (2013).

3.6.2 Devolutivas

Para que as avaliações em larga escala cumpram sua finalidade de promover a qualidade da educação, é fundamental que os professores compreendam e utilizem seus resultados. Apresentaremos dois programas, um nacional e outro estadual, que foram implementados para cumprir essa finalidade.

O primeiro programa, de nível nacional, foi implementado pelo Inep através do portal na internet, é denominado Devolutivas Pedagógicas das Avaliações Educacionais. Apresentamos na Figura 9 o conceito e objetivos dessa plataforma:



A **Plataforma Devolutivas Pedagógicas** aproxima as avaliações externas de larga escala e o contexto escolar, tornando os dados coletados mais relevantes para o aprendizado dos alunos. A partir da disponibilização dos itens utilizados na Prova Brasil, descritos e comentados por especialistas, a Plataforma traz diversas funcionalidades que poderão ajudar professores e gestores a planejar ações e aprimorar o aprendizado dos estudantes.

Quais são os objetivos da plataforma

Promover a melhoria do desempenho dos estudantes brasileiros da educação básica.

Tornar explícito para os professores e gestores das redes de ensino quais conhecimentos e habilidades são verificados pelo SAEB.

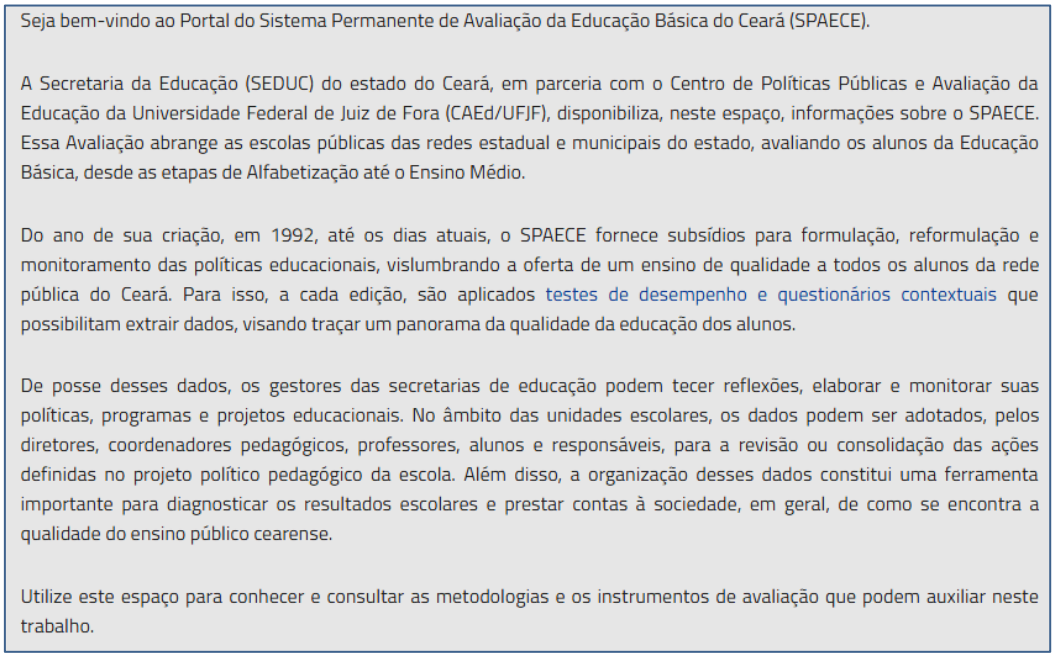
Viabilizar a apropriação pelos professores e equipe gestora dos resultados das avaliações em larga escala.

Colaborar com os professores nas suas atividades de ensino.

Figura 9 – Devolutivas pedagógicas - Inep

Fonte: Inep (2019).

Para o segundo programa, de nível estadual, apresentamos na Figura 10 o Portal do Sistema Permanente de Avaliação da Educação Básica do Ceará (SPAECE).



Seja bem-vindo ao Portal do Sistema Permanente de Avaliação da Educação Básica do Ceará (SPAECE).

A Secretaria da Educação (SEDUC) do estado do Ceará, em parceria com o Centro de Políticas Públicas e Avaliação da Educação da Universidade Federal de Juiz de Fora (CAEd/UFJF), disponibiliza, neste espaço, informações sobre o SPAECE. Essa Avaliação abrange as escolas públicas das redes estadual e municipais do estado, avaliando os alunos da Educação Básica, desde as etapas de Alfabetização até o Ensino Médio.

Do ano de sua criação, em 1992, até os dias atuais, o SPAECE fornece subsídios para formulação, reformulação e monitoramento das políticas educacionais, vislumbrando a oferta de um ensino de qualidade a todos os alunos da rede pública do Ceará. Para isso, a cada edição, são aplicados testes de desempenho e questionários contextuais que possibilitam extrair dados, visando traçar um panorama da qualidade da educação dos alunos.

De posse desses dados, os gestores das secretarias de educação podem tecer reflexões, elaborar e monitorar suas políticas, programas e projetos educacionais. No âmbito das unidades escolares, os dados podem ser adotados, pelos diretores, coordenadores pedagógicos, professores, alunos e responsáveis, para a revisão ou consolidação das ações definidas no projeto político pedagógico da escola. Além disso, a organização desses dados constitui uma ferramenta importante para diagnosticar os resultados escolares e prestar contas à sociedade, em geral, de como se encontra a qualidade do ensino público cearense.

Utilize este espaço para conhecer e consultar as metodologias e os instrumentos de avaliação que podem auxiliar neste trabalho.

Figura 10 – Devolutivas SPAECE

Fonte: SPACE (2019).

Nesses dois exemplos, podemos destacar o foco para que os professores e gestores utilizem os resultados das avaliações para direcionar suas práticas pedagógicas no sentido de melhorar a qualidade do ensino. Essa mesma dinâmica é observada em todos os programas censitários de avaliação em larga escala. Mesmo porque, se os resultados das avaliações não forem utilizados com essa finalidade, não se justificaria a avaliação ser censitária, e se faria um mal-uso de recursos públicos.

A dinâmica de utilização dos resultados, por parte dos professores, passa por duas análises principais. Tomando como exemplo a escola da Figura 11, a primeira análise consiste em comparar as médias e os percentuais de desempenho por padrões de desempenho da escola com os resultados do estado. Nesse exemplo, a média da escola é inferior à média do estado e os percentuais de alunos nos padrões muito crítico e crítico são superiores aos do estado, enquanto os percentuais nos níveis intermediário e avançado são inferiores aos percentuais do estado.

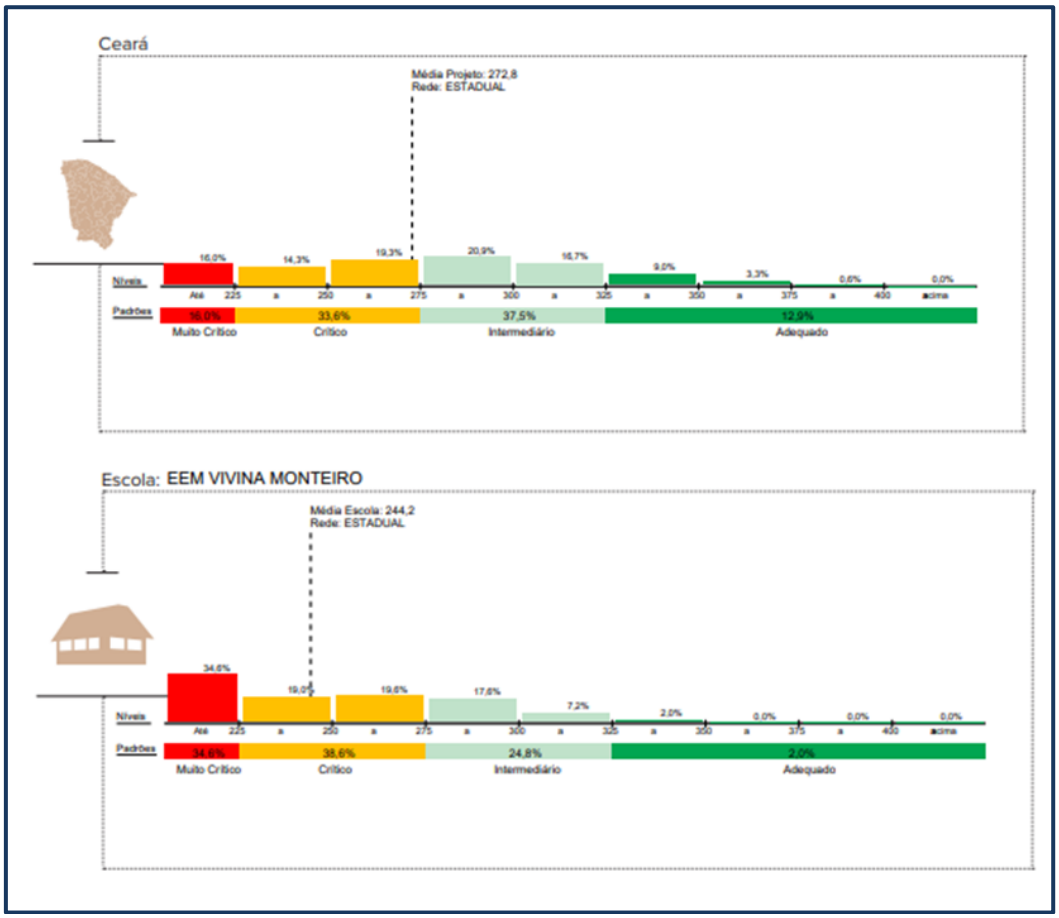


Figura 11 – Resultados de desempenho SPAECE 3EM – 2017

Fonte: SPACE (2019).

Outra análise de comparabilidade de resultados consiste em verificar os desempenhos da escola entre diferentes anos de avaliação. Nesse exemplo, podemos verificar, pela Figura 12, que, entre os anos de 2016 e 2017, a média da escola praticamente não se alterou, o percentual de alunos nos padrões de desempenho teve pouca alteração, com exceção no padrão adequado que passou de 5,7 para 11,4. Mas, de uma maneira geral, a escola não melhorou seu desempenho entre 2016 e 2017.

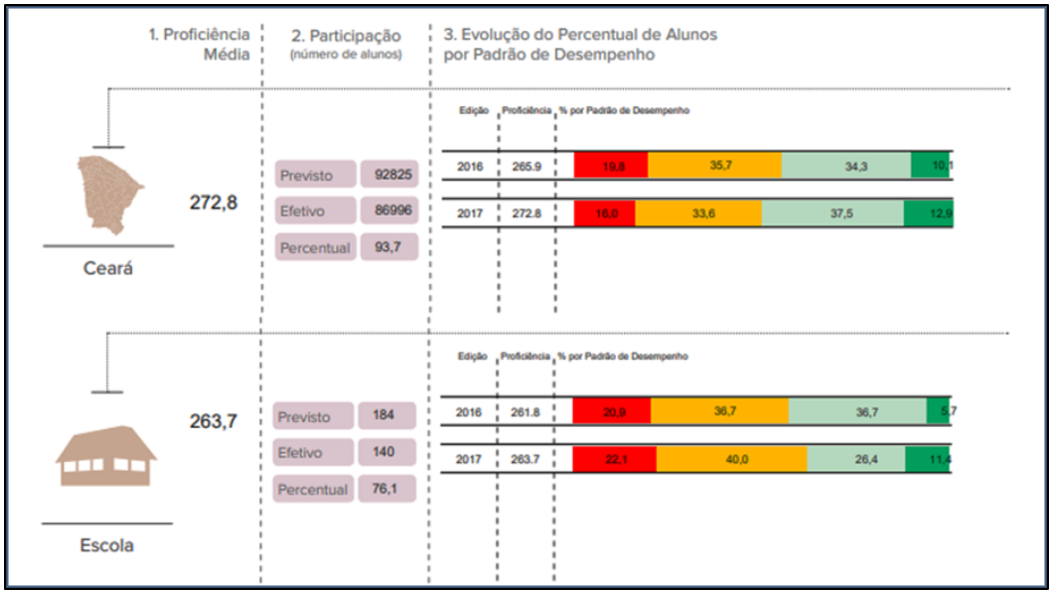


Figura 12 – Resultados de desempenho SPAECE 3EM – 2016/2017

Fonte: SPACE (2019).

A outra análise consiste em o professor verificar quais habilidades estão localizadas em cada um dos padrões de desempenho, principalmente nos padrões abaixo do básico e básico, e se o número de alunos for alto nesses padrões teremos uma situação crítica que necessita o apoio da secretaria para orientar o professor em como trabalhar pedagogicamente esses resultados, no sentido de melhorá-los.

Essas análises dos resultados são fundamentais para o professor sentir como está o seu trabalho e quais habilidades ele deverá trabalhar com seus alunos. Ele não está sozinho, sem um referencial. O fato de se ter uma escala única para todo o sistema proporciona o diálogo entre os professores e gestores no sentido de reverem suas práticas pedagógicas em busca da melhoria da qualidade do ensino.

3.6.3 Estudos de eficácia escolar

As avaliações em larga escala realizada no Brasil, tanto em âmbito nacional como estadual e municipal, têm a característica de serem, na sua totalidade, avaliações transversais[8]. Os indicadores obtidos por essas avaliações, como por exemplo o Ideb e as análises de desempenho em diferentes agregados, são pontuais, indicando apenas se houve melhoras ao longo do tempo para os mesmos períodos de escolaridade.

Segundo Goldstein (2001 apud Silva & Silva, 2016) essa estrutura de avaliação e os modelos da TRI utilizados não modelam de forma adequada a realidade educacional, pois não levam em consideração a sua complexa estrutura, representada por fatores intra e extraescolares. Como uma tentativa de se modelar de forma mais adequada a complexa realidade escolar, Goldstein (2001) propõe a utilização de avaliações longitudinais[9] e a utilização de modelos multiníveis de forma a se obter uma estimativa mais justa dos desempenhos das escolas.

Segundo Teddlie & Reynolds (2000), ao se utilizar modelagens multiníveis em avaliações longitudinais sé possível implementar dois tipos de estudos: (i) estudos de efeito-escola, utilizando-se as proficiências dos alunos e os dados contextuais, e; (ii) estudos de eficácia escolar, ou seja, a análise conjunta dos indicadores de efeito-escola com analises de pesquisas qualitativas (entrevistas e observações do campo).

Como exemplo dos poucos projetos em eficácia escolar no Brasil, temos o Estudo Longitudinal da Geração Escolar 2005 (GERES), realizado entre os anos 2005 a 2008 com o foco em aprendizagem em leitura e matemática durante os primeiros anos do ensino fundamental. Soares et al. (2017) apresentam estudos de efeito escola envolvendo esse projeto.

Como exemplos de outros estudos, citamos:

- Silva (2018) realiza um estudo de efeito escola do 9º ano do Ensino Fundamental ao 3º ano do Ensino Médio e propõe a criação de um indicador de efeito escola considerando o valor agregado de proficiência do 9º ano do Ensino Fundamental para o 1º ano do Ensino Médio, como um indicador com possibilidades de contribuir para a melhoria da qualidade no Ensino

[8] Avaliam sempre as mesmas séries em diferentes períodos de tempo.
[9] Os mesmos alunos são avaliados ao longo dos anos de escolaridade.

Médio, que como observado nas Figuras 7 e 8, permanece estável e muito baixo ao longo da série histórica do Saeb.

- Alves (2006) e Soares & Alves (2008) realizaram estudos de eficácia escolar em escolas públicas de Belo Horizonte com importantes contribuições metodológicas envolvendo pesquisas quantitativas e qualitativas (questionários contextuais e entrevistas), as quais foram utilizadas como variáveis em modelos multiníveis do tipo *piecewise*.

3.7 Tendências na área da avaliação em larga escala

Apresentamos alguns pontos marcantes da história da avaliação em larga escala no Brasil tendo como foco a melhoria da qualidade do ensino. O Brasil desenvolveu ao longo desses anos expertises em produção de medidas e em implantação de políticas públicas vinculadas à avaliação.

Entretanto, nesse cenário de avaliações em larga escala, TRI, políticas de accountability e metas de qualidade, temos os professores em sala de aula e sua tarefa de ensinar e medir o desempenho de seus alunos por meio de provas por eles elaboradas. Como fazer um professor romper com a concepção de que a nota obtida por seus alunos não é uma medida, mas apenas uma quantificação? Como levar ele a compreender que as proficiências de seus alunos obtidas por avaliações externas são realmente medidas? Como capacitá-los a interpretarem pedagogicamente o desempenho de seus alunos, por meio de uma escala, sem que eles tenham acesso aos itens da avaliação? Esses questionamentos são um desafio para os sistemas educacionais. Plataformas de devolutivas, como as apresentadas visam atender a essa necessidade, assim como a realização de treinamentos para os professores e criação de cursos de pós-graduação na área da avaliação.

No cenário internacional, novas técnicas e tendências para o alinhamento da avaliação com a realidade atual, conforme apresentado pela *Association of Test Publishe*[10] (ATS, 2018), já começam a ser estudas e implementadas, como, por exemplo, o Teste Adaptativo Computadorizado (CAT), a elaboração automática de itens, a correção automática de redações, testes eletrônicos, correção de produções

[10]*Association of Test Publisher* (Associação dos Produtores de Testes – EUA).

textuais on line, plataformas eletrônicas para gerenciamento dos sistemas de avaliação, correção da fluência de leitura pela frequência, utilização de inteligência artificial para elaboração de itens/testes e realização de análises psicométricas com o objetivo de agilizar os processos avaliativos.

Diante dessas novas perspectivas, as modelagens matemáticas estarão cada vez mais próximas da complexa realidade escolar, de forma a levar aos professores e redes de ensino informações cada vez mais precisas para que possam orientar suas práticas em direção a uma melhoria da qualidade do ensino, respeitando as particularidades dos entes avaliados.

4. Modelagem Rasch e Modelagem Rasch com Multifacetas (MFRM)

Apresentaremos duas abordagens referentes aos modelos Rasch. Na primeira, discutiremos a modelagem originalmente proposta por Rasch, onde são consideradas duas facetas: a proficiência do aluno e a dificuldade do item. Na segunda abordagem, apresentaremos a modelagem desenvolvida por Linacre (1989), que consiste na utilização da modelagem Rasch para mais de duas facetas.

4.1 Características da modelagem Rasch

A modelagem Rasch, embora pouco utilizada no Brasil, apresenta características que poderiam ser melhor exploradas dentro do contexto das avaliações educacionais em larga escala. Apresentaremos três situações envolvendo a modelagem Rasch, no sentido de nos apropriarmos das vantagens e limitações da utilização de tais modelos.

4.1.1 A Provinha Brasil

Um dos poucos exemplos dessa modelagem nas avaliações nacionais era a Provinha Brasil (2007), que, conforme imagem retirada do site do Inep, apresenta as seguintes características:

A Provinha Brasil é um instrumento pedagógico, sem finalidades classificatórias, que fornece informações sobre o processo de alfabetização e de matemática aos professores e gestores das redes de ensino, e conforme Portaria nº 10, de 24 de abril de 2007, tem os seguintes objetivos:

- Avaliar o nível de alfabetização dos educandos nos anos iniciais do ensino fundamental;

- Oferecer às redes e aos professores e gestores de ensino um resultado da qualidade da alfabetização, prevenindo o diagnóstico tardio das dificuldades de aprendizagem;
- Concorrer para a melhoria da qualidade de ensino e redução das desigualdades, em consonância com as metas e políticas estabelecidas pelas diretrizes da educação nacional.

O delineamento e a construção dessa avaliação preveem, sobretudo, a utilização dos resultados obtidos nas intervenções pedagógicas e gerenciais com vistas à melhoria da qualidade do processo de ensino-aprendizagem.

Para se atingir os objetivos propostos por essa avaliação, que podem ser sumarizados em ter uma escala para análise pedagógica, acesso aos itens do teste pelos professores e acesso rápido aos resultados, a modelagem Rasch foi utilizada, pois, ao corrigir a prova que era realizada pelo próprio professor em sala de aula, a correspondência entre o percentual de acerto e a proficiência era biunívoca. Dessa forma, o professor tinha uma medida que proporcionava um diagnóstico imediato do desempenho de seus alunos, e pelo fato de se ter uma escala, o professor tinha como verificar a dificuldade dos itens, quais alunos acertaram e erraram cada item da prova e, mediante essas informações, realizar suas intervenções pedagógicas de modo a melhorar o desempenho dos alunos.

Essas características não são observáveis nas avaliações em larga escala realizadas no Brasil, que utilizam a modelagem 3PL e uma estrutura de montagem de caderno e equalização que não possibilitam aos professores acesso rápido aos resultados, visualização dos itens e uma correspondência biunívoca entre o percentual de acerto e a medida de proficiência. Essas características, são um grande empecilho para a apropriação dos resultados das avaliações por parte dos professores.

Para uma outra abordagem comparativa, envolvendo aspectos técnicos e metodológicos das propriedades das medidas obtidas pelos modelos de Rasch e pelos modelos de 2PL e 3PL, utilizaremos, como referência, o artigo de Benjamin Wright (1992) que argumenta a favor da utilização do modelo Rasch (*IRT in the 1990s: Which Models Work Best? 3PL or Rasch*), em articulação com o artigo de Bergan (2013), que contesta os argumentos de Wright e apresenta alternativas a fim de se utilizar o melhor modelo (*Rasch Versus Birnbaum: New Arguments In a Old Debate*).

4.1.2 Artigo de Wright (1992)

Em seu artigo, Wright (1992) apresenta algumas das principais diferenças entre essas duas modelagens e enfatiza que a modelagem Rasch, ao contrário das demais modelagens (2PL e 3PL), é a única que possui objetividade específica, que é uma condição básica de qualquer instrumento de medição. Na perspectiva desse modelo, dois alunos que acertarem a mesma quantidade de itens em um mesmo tipo de teste, nunca poderão ter proficiências diferentes. Esse é um dos principais argumentos de Wright a favor de Rasch e contra as modelagens 3PL e 2PL. Apresentamos, no Quadro 3, as principais características desses dois modelos

Quadro 3 – Características dos modelos 3PL e Rasch

Características	Birnbaum Model (3PL) Allan Birnbaum 1957/1968	Rasch Model Georg Rasch 1952/1960
1	Imita dados	Define medidas
2	Aceita o parâmetro “c”	Rejeita o parâmetro “c”
3	Aceita a variação do parâmetro “a”	Rejeita a variação do parâmetro “a”
4	Aceita o cruzamento das CCI	Rejeita o cruzamento das CCI

Fonte: O Pesquisador.

A primeira característica mostrada no Quadro 3 está relacionada ao fato de que a modelagem 3PL, ao imitar dados (no sentido de encontrar um modelo que se ajusta às características do item), praticamente não elimina nenhum item no processo de modelagem, produzindo uma medida mais robusta se comparada à obtida com o uso da modelagem Rasch. Para os estatísticos, essa característica justificaria o uso de uma modelagem mais complexa. Entretanto, para os cientistas sociais e psicólogos, mais importante do que ter um modelo mais complexo que se ajusta aos dados, é ter um modelo que se ajusta ao constructo avaliado. Este seria o caso da modelagem Rasch, que embora seja uma modelagem mais simples, uma vez que se utilize itens bem ajustados, ela fornece uma boa representação da realidade a ser modelada, como sugerido por Wright (1992).

A segunda característica, segundo os seguidores da modelagem Rasch, se baseia no fato de que o acerto ao acaso, modelado pelo parâmetro “c”, não é uma característica do item e sim uma característica da população, não justificando, dessa forma, sua inclusão na modelagem do item.

A terceira e quarta características estão relacionadas entre si. Ao se aceitar a variação do parâmetro “a”, temos, como consequência, o cruzamento das CCI. Podemos observar, no Gráfico 4, uma situação em que dois itens com o mesmo parâmetro “b” e com diferentes parâmetros “a” possuem CCI que se cruzam, indicado que para um respondente com proficiência abaixo do Parâmetro “b” (250) o item 1 é mais fácil que o item 2, e que para um outro respondente com proficiência acima do parâmetro “b” (250) a situação da dificuldade dos itens se inverte. Essa característica da modelagem 3PL, conforme apontado por Wright (1992), não se enquadra como um instrumento de medida, pois não produz uma mesma regra para todos os respondentes e como veremos no subtópico 4.2.2, não atende ao princípio da invariância da medida, de forma a se ter medidas objetivas, como proposto por Rasch (1956).

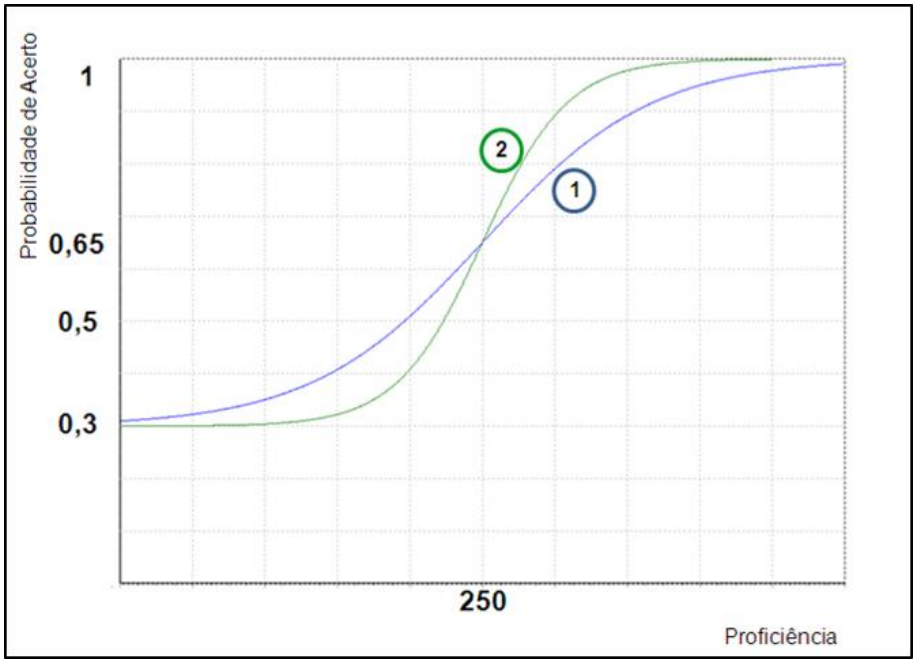


Gráfico 4 – Cruzamento das CCI

Fonte: O Pesquisador.

4.1.3 Artigo de Bergan (2013)

Bergan em 2013, pondera que após vinte anos da apresentação dos argumentos de Wright e dos avanços tecnológicos ocorridos nesse período, a controvérsia sobre qual o melhor modelo a se utilizar ainda existe. Bergan faz uma

abordagem crítica das vantagens apresentadas por Wright do modelo Rasch em relação ao modelo 3PL e enfatiza que embora as abordagens de Wright estejam corretas, as mesmas não tinham evidência empírica e que após vinte anos de evolução tecnológica alguns argumentos de Wright perderam significado.

Bergan utiliza sua experiência na *Assessment Tchnology Incorporated* (ATI), uma empresa americana que realiza anualmente centenas de milhões de testes para sustentar seu argumento de que a escolha do melhor modelo para representar os dados deve ser baseada em dois fatores: parcimônia e se o ajuste do modelo aos dados atende às necessidades de utilização da avalição. De acordo com Bergan (2013), nenhum modelo, a priori, é melhor do que o outro, e apenas por meio de testes empíricos é possível se chegar a uma escolha razoável.

Apresentamos, a seguir as ponderações de Bergan sobre os argumentos de Wright:

- ✓ Objetividade específica: na época de Wright, o modelo Rasch era o único modelo da TRI que relacionava de forma biunívoca o percentual de acerto com a proficiência do aluno, de tal forma que alunos com o mesmo percentual de acerto teriam a mesma medida de proficiência, sendo, portanto, segundo Wright, o único modelo capaz de proporcionar medidas válidas. Após vinte anos, Bergan (2013) apresenta uma inovação tecnológica denominada soma de escores (*Score summing*) que, ao ser aplicada aos modelos de dois e três parâmetros, faz com que ocorra a biunivocidade entre os percentuais de acerto e a proficiência. Embora, com essa inovação tecnológica tenha-se conseguido alinhar o melhor ajuste dos dados, por meio dos modelos de dois e três parâmetros e se obter a objetividade específica, Bergan (2013) reconhece que essa solução não acabará com a controvérsia apontada por Wright, pois a objetividade específica é mais uma questão filosófica do que empírica;
- ✓ Complexidade e Recursos computacionais: com os avanços computacionais, os argumentos de Wright referentes à dificuldade de se trabalhar com modelos complexos que utilizam os parâmetros “a” e “c” não procedem mais. Esse é um argumento que não se justifica na atualidade. Também não se justifica a limitação de modelagem de itens politômicos somente por modelos Rasch;
- ✓ Cruzamento de CCI: na modelagem Rasch, o fato das CCI não se cruzarem

se baseia no argumento de que a habilidade medida por um item é pré-requisito para a habilidade medida por outro item. O cruzamento de curvas é visto, pelos seguidores de Rasch, como um indicativo de viés. Esse é um argumento que simplifica a modelagem, mas na situação real, sabemos que os alunos utilizam diferentes estratégias cognitivas para solucionarem problemas ou itens, para ser mais específico. Não modelar essa característica significa que estamos perdendo informações importantes sobre o comportamento cognitivo dos alunos.

Após apresentar seus argumentos, Bergan (2013, p. 6) sumariza sua opinião:

> Embora a ATI não considera os argumentos de Wright convincentes, a elegância desse modelo é apreciada. A modelagem Rasch proporciona uma excelente representação parcimoniosa dos dados. Como consequência, ela deve ser selecionada quando modelos mais complexos não melhoram significativamente o ajuste do modelo aos dados. Além do mais, ela oferece uma medida em que as habilidades referentes aos itens de um teste estão relacionadas em uma sequência de pré-requisitos.

Nossa intenção não é aprofundar nos aspectos filosóficos e matemáticos destas duas modelagens, mas sim nos aspectos metodológicos e operacionais. Na área da avaliação educacional essas duas modelagens, em função de suas especificidades, podem ser aplicadas de forma isolada ou concomitante, sendo essa escolha baseada em aspectos técnicos e não filosóficos, conforme apontado por Bergan (2013), de forma a proporcionar aos atores envolvidos no sistema educacional informações válidas que os ajudem a medir a qualidade do ensino pelo à qual são responsáveis.

4.2 Modelagem Rasch para duas facetas

Apresentaremos três tipos de modelagem Rasch duas facetas, em função do tipo de item utilizado na produção da medida, ou seja, itens dicotômicos e itens politômicos.

4.2.1 Modelagem Rasch para itens dicotômicos

Esse modelo, originalmente apresentado por Rasch (1956) relaciona a probabilidade de o aluno acertar o item, variável dependente, com duas variáveis independentes: a proficiência do aluno, "θ_n" e a dificuldade do item "b_i"

$$\boldsymbol{P(X_{ni} = 1|\theta_n, a, b_i) = \frac{e^{Da(\theta_n - b_i)}}{1 + e^{Da(\theta_n - b_i)}}}$$

No modelo Rasch, todos os itens de um teste têm o mesmo parâmetro "a" e a somada média dos parâmetros "b" é igual a zero.

4.2.2 Modelagem Rasch para itens politômicos de escala gradual

Esse modelo, desenvolvido por Andrich (1978 apud Linacre, 1989) é apropriado para escalas do tipo Likert, onde o respondente manifesta sua opinião sobre determinado tema escolhendo uma determinada categoria do item. Normalmente essas categorias variam iniciando com o aspecto negativo, passando pelo neutro e terminado com o aspecto positivo. Por exemplo: discordo fortemente, discordo, neutro, concordo e concordo fortemente. Nesse tipo de estrutura não há mensuração entre as categorias.

Formulação matemática:

$$\boldsymbol{P\left(X_{ni} = k\middle|\theta_n, b_i, (\tau_{j=1,k-1}\right) = \frac{e^{\sum_{j=1}^{k-1} \theta_n - (b_i + \tau_j)}}{1 + \sum_{y=0}^{k} e^{\sum_{j=1}^{k-1} \theta_n - (b_i + \tau_j)}}}$$

Em que,

A probabilidade de um aluno "n" responder a categoria "k" do item "i", vai depender da sua habilidade "θ_n", da dificuldade do item "b_i"e do parâmetro de transição entre as categorias "$\tau_{j=1,k-1}$".

Esse modelo apresenta as seguintes características:

- ✓ Todos os itens possuem a mesma quantidade de categorias, os intervalos entre categorias de um item são equidistantes e iguais para todos os itens, ou seja, todos os itens têm o mesmo formato de curva, o que difere é a dificuldade do item;
- ✓ A interseção das curvas de duas categorias adjacentes (K e K-1), representada no modelo pela letra grega "τ", é denominada de Rasch-

Andrich Threshold, ou *step* de dificuldade, nesse ponto, tem-se a mesma probabilidade de estar nas duas categorias;

- ✓ A dificuldade do item "b" é o ponto de interseção da primeira com a última categoria.

4.2.3 Modelagem Rasch para itens politômicos de crédito parcial

O modelo Rasch de crédito parcial foi desenvolvido por Masters (1982 apud Linacre, 1989). Nesse modelo, as distancias entre categorias de um item não são necessariamente iguais às distâncias entre as categorias de um outro item do teste e as distâncias entre as categorias de um mesmo item não são necessariamente iguais.

Esse modelo possui a seguinte formulação:

$$\boldsymbol{P\left(X_{ni} = k \middle| \theta_n, b_{ij}, m_i \ = k-1\right) = \frac{e^{\sum_{j=1}^{x_{vi}} \theta_n - b_{ij}}}{1 + \sum_{y=0}^{k} e^{\sum_{j=1}^{k-1} \theta_n - b_{ij}}}}$$

Em que,

A Probabilidade de uma pessoa "n" estar na categoria" k" de um item "i", depende da proficiência "θ_n" da pessoa e das características do item, representada pela dificuldade de transição, "b_ij", entre as categorias k =1e k. O parâmetro "x_{vi}" varia de 0 a k-1.

4.2.4 Invariância na modelagem duas facetas

O matemático dinamarquês Georg Rasch (1956) deu o nome de Objetividade específica às seguintes propriedades a serem constatadas no ato de mensurar, por exemplo, o conhecimento dos alunos por meio de testes: (i) o desempenho não sofre influência em função das características dos testes a eles submetidos, ou seja, não importa se a proficiência de um mesmo aluno é medida por dois testes com dificuldades distintas, os valores obtidos terão que ser os mesmos e, (ii) a dificuldade dos testes não é afetada pelas características dos alunos, ou seja, a calibração dos itens não é afetada pelo nível de proficiência dos alunos.

Se pensarmos em um procedimento corriqueiro como, por exemplo, medir a altura de uma pessoa com uma fita métrica, temos a convicção de que a pessoa que está sendo medida não altera as características da fita métrica e que se utilizarmos uma outra fita métrica o resultado da medida será o mesmo obtido anteriormente com a primeira fita. Estas duas características tão óbvias quando estamos lidando com fenômenos físicos, se tornam muito complexas e controversas quando medimos variáveis latentes nas ciências sociais e humanas.

Na psicometria, medir é um processo de posicionamento de pessoas e itens em uma escala, que representa um constructo através de uma variável latente, sendo que esse processo de medição está estruturado no conceito de invariância, ou seja, a característica da população a ser medida não afeta o instrumento (*person-free*) e, em contrapartida, o instrumento não altera as características da população (*test-free*). Se para o fenômeno físico isso era óbvio, para variáveis latentes isso se manifesta de outra maneira.

Rasch (1961, p. 331-332, apud Engelhard & Wind, 2018, p. 8-9) definiu esses requisitos como objetividade específica e os detalhou da seguinte forma:

> A comparação entre dois estímulos deve ser independente de quais indivíduos em particular foram utilizados para a comparação; e deve também ser independente de outros estímulos, dentro da classe considerada, que foram ou talvez fossem comparados também. Simetricamente, uma comparação entre dois indivíduos deve ser independente dos estímulos particulares dentro da classe considerada que foi utilizada para a comparação; e também deve ser independente dos outros indivíduos que também foram comparados na mesma ou em alguma outra ocasião.

Apresentamos na Figura 13 os requerimentos para invariância de medidas em modelagens duas facetas.

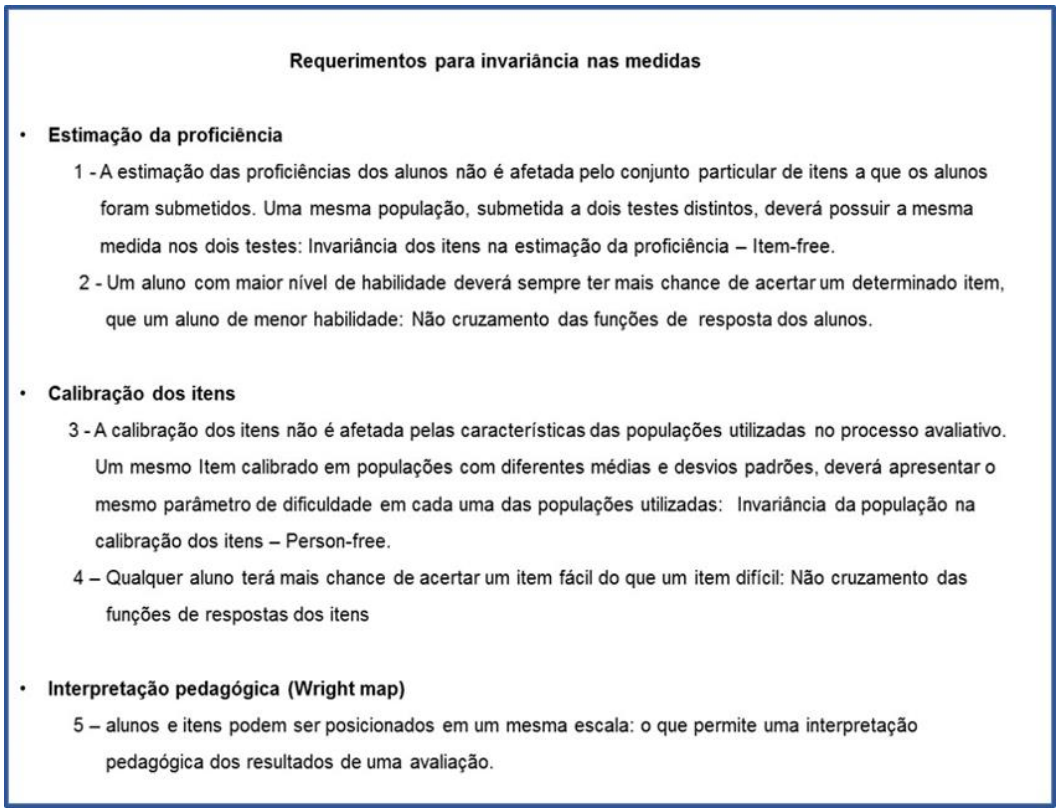


Figura 13 – Requerimentos para invariância de medidas em modelagens duas facetas
Fonte: Engelhard & Wind (2018).

4.3 Modelo Rasch multifacetas (MFRM)

O modelo de MFRM é uma extensão do modelo Rasch para situações em que mais de duas facetas são modeladas para a produção de uma medida de um constructo latente. De acordo com seu criador, Linacre (1989), esse modelo surgiu pela necessidade de tornar as medidas em que são utilizados processos de correções que envolvem julgamentos, mais justas:

O MFRM amplia a aplicação do modelo Rasch para situações em que mais de duas facetas interagem para produzir uma medida. Ele permite a construção de uma estrutura de referência na qual comparações comparativas entre examinandos não dependem de qual juiz ou item utilizado no processo de correção (Linacre, 1989).

A grande versatilidade do MFRM, para avaliações que utilizam juízes é o fato de se colocar na mesma escala a severidade dos corretores com as duas outras facetas normalmente utilizadas nas avaliações pela TRI: a proficiência dos alunos

e dificuldade dos itens. Essa característica torna o processo de medição independentemente de como cada corretor interpreta a categoria de resposta do item em que ele alocará o aluno. Para exemplificar, quando temos uma avaliação pela TCT, não temos como afirmar que o mesmo valor de escore dado a um mesmo aluno por dois corretores diferentes é realmente o mesmo. Ao passo que, ao usarmos a MFRM essa comparação é estatisticamente viável, pois a severidade de cada corretor é utilizada para ajustar a proficiência do aluno ou para calcular o seu escore verdadeiro (*fair score*)[11].

A principal aplicabilidade desse modelo ocorre em testes em que são aplicados itens politômicos de resposta graduada, os quais são modelados segundo a formulação desenvolvida por Linacre (1989):

$$\boldsymbol{log\left(\frac{P_{nijk}}{P_{nijk-1}}\right) = B_n - D_i - C_j - F_k}$$

Em que,

P_{nijk} – Probabilidade do examinando "n" ser classificado (julgado) na categoria "K", pelo juiz "j", no item "i".

P_{nijk-1} – Probabilidade do examinando "n" ser classificado (julgado) na categoria "K – 1", pelo juiz "j", no item "i".

B_n – Habilidade do Examinando "n"

D_i – Dificuldade do item "i"

C_j – Severidade do juiz "j"

F_k – Dificuldade entre as categorias "k-1" e "k" (a categoria "K" varia de 1 a M).

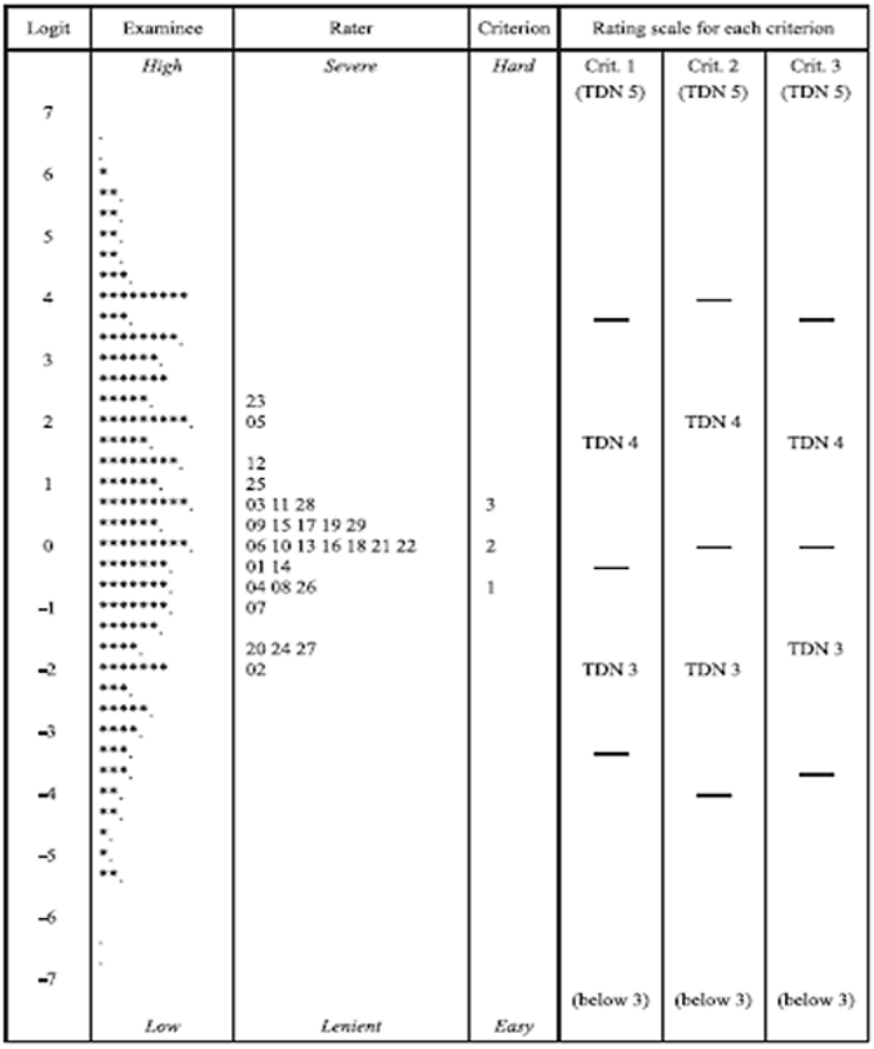
Uma característica do modelo, a fim de aumentar sua identificabilidade e interpretatibilidade, é a imposição da condição de que o somatório das dificuldades associadas à uma mesma faceta seja igual a zero para todas as facetas, com exceção da faceta associada ao item.

Por meio dessa modelagem todas as facetas são referenciadas a uma mesma escala. Uma forma útil de análise das variáveis envolvidas na modelagem é proporcionada pela utilização do mapa de Wright.

Apresentamos, na Figura 14, os resultados de uma modelagem realizada por Eckes (2005), visualizados através da utilização de um mapa de variáveis (também

[11] Escore do aluno ajustado em função da severidade do corretor.

denominado mapa de Wright).



Logit	Examinee	Rater	Criterion	Rating scale for each criterion		
	High	*Severe*	*Hard*	Crit. 1 (TDN 5)	Crit. 2 (TDN 5)	Crit. 3 (TDN 5)
7	.					
	.					
6	*					
	**.					
	**.					
5	**.					
	**.					
	***.					
4	*********					
	***.			—	—	—
	********.					
3	******.					
	*******					
	*****.	23				
2	*********.	05			TDN 4	
	*****.			TDN 4		TDN 4
	********.	12				
1	******.	25				
	*********.	03 11 28	3			
	******.	09 15 17 19 29				
0	*********.	06 10 13 16 18 21 22	2		—	—
	*******.	01 14		—		
	*******.	04 08 26	1			
−1	*******.	07				
	******.					
	****.	20 24 27				TDN 3
−2	*******	02		TDN 3	TDN 3	
	***.					
	*****.					
−3	****.					
	***.			—		
	***.					—
−4	**.				—	
	**.					
	*.					
−5	*.					
	**.					
−6	.					
	.					
−7						
				(below 3)	(below 3)	(below 3)
	Low	*Lenient*	*Easy*			

Figura 14 – Mapa de variáveis (mapa de Wright)

Fonte: Eckes (2005, p. 205).

Nesse mapa de variáveis, a proficiência dos candidatos (*examinee*), a severidade dos corretores (*rater*) e a dificuldade dos itens (*criterion*) estão em uma mesma métrica, que está representada pela primeira coluna do mapa (*logit*). Como pode ser observado, a variabilidade da severidade dos corretores é substancial, o que justifica a inclusão dessa faceta na estimativa da proficiência do candidato.

4.3.1 Invariância da medida na MFRM

Na modelagem de Rasch, a invariância é considerada segundo dois aspectos, ou facetas: aluno (proficiência) e item (dificuldade), pois os itens são do tipo

dicotômico (múltipla escolha). Porém, para itens politômicos, em que a correção dos mesmos está vinculada à presença de um corretor/juiz (severidade) há a necessidade de se considerar também essa faceta, no que se refere à invariância da medida. Apresentamos na Figura 15 os requerimentos para invariância em avaliações que utilizam juízes.

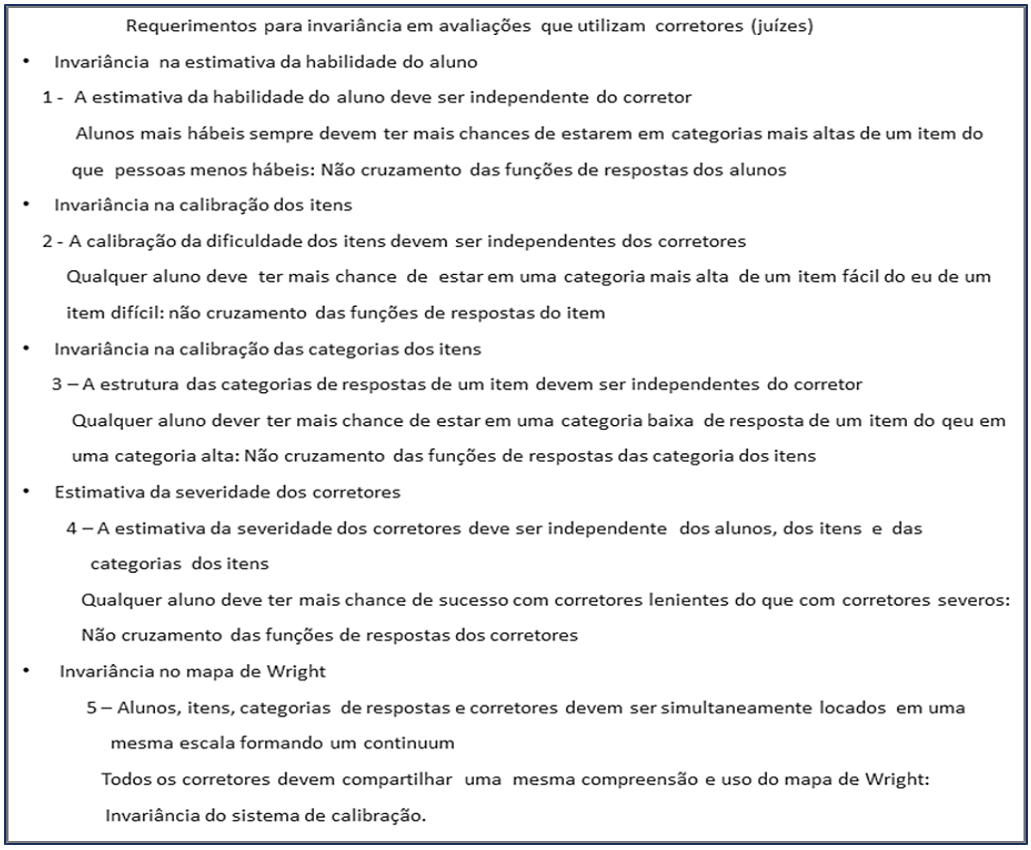


Figura 15- Requerimentos para invariância de medidas em modelagens três facetas
Fonte: Engelhard & Wind (2018).

4.3.2 Escala de classificação

A escala de classificação envolve todo o processo de construção de uma escala intervalar de conhecimento em que itens do tipo politômicos, corrigidos por juízes são modelados segundo três facetas pela TRI, de modo a se obter um *continumm*[12].

A origem da escala de classificação de acordo com Engelhard & Wind

[12] Ordenação de valores em uma escala.

(2018) e com Guilford (1936) se remonta aos estudos de Galton (1936), por meio dos quais esse pesquisador elaborou uma sequência em ordem crescente de intensidade de percepção, para algumas variáveis latentes, de modo a possibilitar que um juiz pudesse localizar a sua percepção para dado fenômeno (características sensoriais), envolvendo: percepção de claridade, intensidade de barulho, odor, sabor, calor, fome, cansaço, sede e etc. Para exemplificar, vamos considerar a classificação da intensidade de claridade às 12:00 horas de um determinado dia, utilizando uma escala do tipo Likert de cinco categorias: (i) muito escuro, (ii) escuro, (iii) moderado, (iv) claro, (v) muito claro.

Embora, nesse exemplo, sejam fornecidos o estímulo e a escala para a sua classificação, teremos, para um mesmo local em um mesmo dia e horário, diferentes classificações em função da sensibilidade do avaliador. Na tentativa de minimizar esse problema são utilizadas referências para cada uma das categorias com o intuito de auxiliar os juízes em suas tarefas de posicionar suas percepções em um contínuo o mais padronizado possível, de forma a podermos ter comparabilidade entre diferentes medições realizadas por diferentes juízes.

Nos processos avaliativos que utilizam questões abertas, sendo representadas nos modelos da TRI por itens politômicos, é de praxe a utilização de referências para julgamentos com o objetivo de orientar e padronizar o processo de correção por parte dos corretores. Normalmente as referências para julgamentos estipulam valores ordenados de pontuações para as categorias de respostas dos itens, em função do nível de conhecimento que os alunos demonstram ao responder o item. Esses níveis variam do totalmente errado ao totalmente certo em uma escala do tipo Likert. Os números de categorias normalmente variam de “3” a “6”. Quanto mais categorias, mais difícil se torna o trabalho do corretor.

As Referências para julgamentos, conforme definição encontrada em Engelhard & Wind (2018) são fontes de informação que os corretores (juízes) utilizam para orientar seus julgamentos de forma a posicionar os alunos em um determinado nível/categoria de um item.

Com o objetivo de ancorar as ideias apresentadas sobre a produção de medida em uma escala gradual, com a utilização de juízes, apresentamos a seguir dois exemplos de testes que utilizam questões que são corrigidas através de referências para julgamentos. Esses dois casos foram escolhidos por representarem a quase totalidade das avaliações educacionais em larga escala realizadas no Brasil

que utilizam corretores no processo de produção de medidas.

No primeiro caso, Anexo 3, o procedimento utilizado como referência para julgamento é denominado de Chave de correção. Trata-se de um tipo de teste aplicado nas avaliações de escrita do terceiro ano do ensino fundamental no projeto do PAEBES-ALFA no Estado do Espírito Santo. O anexo se refere à chave de correção para um item de produção narrativa.

O segundo caso, se refere a redações que utilizam processo de correção conforme o padrão Saeb. Tomamos, como exemplo, a avaliação realizada na Bahia no ano de 2011. Para a correção das redações, os corretores utilizam como referência a matriz de competências para a produção de texto, apresentada no Anexo 4.

4.4 *FACETS*

Atualmente, existem múltiplos programas computacionais voltados para a execução de aplicações da TRI, como *FACETS, ConQuest, BilogMG, IRTPRO, X-calibre, M-plus, Multilog, Parscale e Winstep*. Neste projeto, optou-se utilizar o programa *FACETS* (Linacre, 2013) como ferramenta de análise, pelo fato dele compreender uma ampla variedade de modelos de TRI. Especificamente, a partir do *FACETS,* pode-se trabalhar com itens dicotômicos ou politômicos em modelos multifacetados. A partir desses modelos, pode-se utilizar o *FACETS* em uma série de diferentes aplicações, tais como: realizar a análise de itens em avaliações de larga escala; investigar o Funcionamento Diferencial de Itens (DIF); investigar o efeito do corretor e o comportamento diferencial dos itens.

Os modelos multifacetados do *FACETS* permitem investigar o nível de severidade dos corretores, ou seja, nas correções de itens abertos, quais avaliadores são mais lenientes e quais são mais severos que os demais. Além disso, nos modelos multifacetados, é possível identificar avaliadores que exibem um padrão de correção diferente ou que fazem julgamentos inconsistentes com a opinião dos demais corretores. Tais modelos são chamados multifacetados porque consideram não somente as características dos itens e dos respondentes, como nas análises tradicionais, mas porque possibilitam considerar, também, os efeitos de diversas outras facetas como, por exemplo, a rigidez/leniência do corretor, o tópico

investigado e os critérios de correção utilizados.

Conforme proposto por Eckes (2011), por meio da modelagem com multifacetas, temos como ajustar melhor as informações de testes para mais de uma disciplina ou subescala, assim como a influência dos corretores na correção de itens abertos. Utilizando técnicas estatísticas específicas, é possível calcular um fator de ajuste e levar este fator em consideração para as notas dos alunos que foram alvo de corretores com diferentes níveis de severidade, aumentando ou diminuindo suas pontuações, conforme tenham suas notas oriundas de corretores severos ou lenientes, respectivamente.

4.4.1 Análise de ajuste ao modelo pelo *FACETS*

O objetivo dessa análise é verificar o quanto a estimativa da medida obtida no modelo, para cada elemento, de cada faceta se ajusta aos dados empíricos. Conforme aponta Engelhard & Wind (2018, p. 168), "Quando os dados se ajustam à MFRM estimativas invariantes de cada uma das facetas são obtidas em uma mesma escala". Ou seja, se o ajuste é bom podemos nos assegurar de que o modelo está bem alinhado com os dados e, portanto, as medidas têm um significado interpretável para a realidade (constructo) que está sendo medida. Se o ajuste for ruim, as medidas não retratarão de forma adequada a realidade, e as suas interpretações induzirão os especialistas a análises equivocadas.

As análises de ajuste são realizadas através das estatísticas *outfit* e *infit* média ao quadrado (*Mean-square*), conforme consta no manual do *FACETS*, Linacre (2014). A interpretação dos valores dessas estatísticas é fornecida no Quadro 4.

Quadro 4 – Interpretação das medidas de ajuste

Categoria de ajuste	Valor de *outfit*/*infit*	Interpretação
1	> 2.0	Distorce ou degrada o sistema de medida. São considerados como um sério problema na qualidade da medida. Esses casos devem ser eliminados da análise de forma a obter medidas mais harmoniosas.
2	1.5 a 2.0	Improdutivo para a produção de medida, mas não é degradante
3	0,5 a 1,5	Produtivo para medida – faixa ideal
4	< 0,5	Menos produtivo para geração de medida, mas não é degradante. Pode produzir erros de confiabilidade e separações. Geralmente esses casos não são um problema sério.

Fonte: Linacre (2014).

No caso de a faceta em análise ser item, a estatística outfit detecta desajustes em respostas *outliers*, ou seja, itens muito fáceis ou muito difíceis para o aluno e a estatística infit é sensível a desajustes em itens localizados em torno da média do aluno.

Para dados não ajustados, Linacre (1989) propõe que os mesmos sejam retirados do modelo e que se refaçam as medidas. Em seguida, compara-se as medidas para as duas situações, se as diferenças não forem significativas para o tipo de análise desejada, pode-se considerar os dados não ajustados na análise final.

4.5 Análise da confiabilidade entre corretores (*interrater reliability*)

A análise da confiabilidade das medidas entre corretores é de suma importância em processos avaliativos onde existe a presença de especialistas corrigindo os testes dos alunos. Apenas se tivermos evidências estatísticas de que os corretores atuam de forma independente, o que é um dos pressupostos apresentados por Rasch para a objetividade da medida, teremos como obter medidas de boa qualidade e portanto, confiáveis, do desempenho dos alunos.

Uma avaliação de boa qualidade necessita ter validade (*validity*) e confiabilidade (*reliability*). Validade é definida como "o grau em que todas as evidências acumuladas corroboram a interpretação pretendida dos escores de um

teste para os fins propostos" (AERA, APA, NCME, 1999). Confiabilidade está ligada à consistência e precisão dos resultados do processo de mensuração (Urbina, 2007).

Segundo Eckes (2011), classificações realizadas por seres humanos estão sujeitas a várias formas de erros e de vieses. Este problema vem acompanhado as avaliações de *performance* desde o início de sua concepção e utilização. Por exemplo, corretores com a função de avaliar o desempenho de candidatos em um determinado teste geralmente apresentarão diferentes pontuações e classificações para o mesmo nível de desempenho real dos alunos. Esta situação representa uma ameaça para a validade e confiabilidade da medida, influenciando de forma negativa na imparcialidade da avaliação.

São dois os indicadores de confiabilidade apontados por Eckes (2011), a saber, concordância[13] e consistência[14], que devem ser analisados de foram conjunta, a fim de se verificar o nível de qualidade da medida. O primeiro relaciona-se ao quanto os corretores apresentam pontuações idênticas para os mesmos alunos e o segundo, diz respeito, a quais corretores são mais severos ou mais lenientes, dito de outra forma, as pontuações entre um corretor padrão e um corretor severo ou leniente não são iguais, mas são altamente correlacionadas.

Apresentamos na Tabela 1, as quatro combinações possíveis de concordância e consistência entre corretores em um processo de produção de medidas. Essas combinações estão relacionadas com o nível de severidade inerente a cada corretor.

Tabela 1 – Relação concordância x consistência - severidade dos corretores

Aluno	Situação 1		Situação 2		Situação 3		Situação 4	
	Corretor 1	Corretor 2	Corretor 1	Corretor 3	Corretor 1	Corretor 4	Corretor 1	Corretor 5
A	1	1	1	2	1	1	1	6
B	2	2	2	3	2	2	2	1
C	3	3	3	4	3	3	3	5
D	4	4	4	5	4	3	4	2
E	5	5	5	6	5	3	5	3
Concordância	Alta		Baixa		Alta		Baixa	
Consistência	Alta		Alta		Baixa		Baixa	

Fonte: O Pesquisador.

Nessa tabela, procuramos representar uma avaliação composta por cinco

[13] Ou consenso para alguns autores.
[14] Ou confiabilidade para alguns autores.

corretores, avaliando cinco alunos que responderam a um item politômico de resposta graduada com seis categorias de respostas variando de “1” (totalmente errado) a “6” (totalmente certo). Normalmente, com a finalidade de auxiliar o trabalho dos corretores e minimizar subjetividade inerente a cada corretor, são utilizados documentos com a característica de cada categoria de respostas dos itens. Assim os corretores têm uma referência única para definir em qual categoria o aluno está em determinado item. São exemplos desse tipo de documentos, chaves de correção (Anexo 3) e matrizes de competência para produção de texto (Anexo 4).

Temos, portanto, quatro situações possíveis de combinação de corretores:

- ✓ Situação 1: alta concordância e alta consistência com os corretores trabalhando com o mesmo nível de severidade e com a mesma interpretação das categorias do item
- ✓ Situação 2: baixa concordância e alta consistência, onde o corretor “3” apresenta um menor nível de severidade em relação ao corretor “1”, dando notas maiores aos alunos, salvo pela maior leniência, podemos considerar que os dois corretores apresentam a mesma interpretação das categorias do item;
- ✓ Situação 3: alta concordância e baixa consistência, o corretor “4” apresenta uma concordância alta com o corretor “1” nas categorias baixas do item (1 a 3), mas diverge nas categorias altas dos itens. Podemos considerar que os corretores apresentam a mesma interpretação para as categorias baixas do item, mas divergem para as categorias altas;
- ✓ Situação 4: baixa concordância e baixa consistência, situação extrema e prejudicial para o processo, onde, claramente, o corretor “5” não utiliza de forma adequada a interpretação dos níveis de dificuldade das categorias do item.

Existem diferentes métodos estatísticos para se medir o nível de confiabilidade entre corretores, os quais por uma questão didática, agruparemos em dois grupos distintos em função da forma como foram geradas as medidas de desempenho dos alunos, ou seja, via TCT ou via TRI.

Se as medidas de desempenho dos alunos foram geradas pela TCT, Stemler (2004) enumera os seguintes principais métodos para se verificar a confiabilidade das medidas: Para medir o consenso: índice de concordância entre corretores e estatística de *Cohen’s Kappa,* e para medir a consistência: coeficiente de correlação

de *Pearson* (*r*), *coeficiente de classificação de Spearman* (*rho*) e Coeficiente alfa de *Cronbach*. Cada uma destas estatísticas fornece estimativas do quanto dois ou mais corretores estão aplicando suas correções ou julgamentos de forma previsível e replicável.

A principal desvantagem desses métodos é o fato de serem calculadas pra cada par de corretores em cada item do teste, esta limitação inviabiliza sua aplicabilidade em testes com muitos corretores e muitos itens.

Para as avaliações que utilizam medidas via TRI, Stemler (2004) enumera três métodos para a análise da confiabilidade entre corretores, a saber: análise das componentes principais, Teoria da generalização e por meio da utilização da MFRM.

4.6 Aplicabilidade da MFRM

Robitzsch & Steinfeld (2018) analisaram os artigos dos jornais ‘Language testing” e ‘Language assessment Quartely’ relacionados com testes envolvendo corretores no período de 2007 e 2017 e detectaram que a modelagem Rasch-multifacetas foi utilizada na maioria das avaliações (51,5%), seguida pela modelagem da Teoria da generalização (19,1%) e demais métodos envolvendo análises qualitativas e outros modelos mais complexos da TRI. Corroborando com esses dados, McNamara & Knock (2012) fizeram um levantamento semelhante nas avaliações de linguagem entre 1984 e 2002 e descobriram que a modelagem Rasch é dominante e que o surgimento e aplicação da modelagem através de facetas, levando em consideração a severidade dos corretores, foi um fator decisivo para a popularidade desse tipo de modelagem.

Não iremos descrever os dois primeiros métodos, os quais são trabalhados com maior profundidade em Harman (1967) e Shavelson & Webb (1991), respectivamente. Como o foco dessa tese é a MFRM, realizamos nos capítulos empíricos dessa tese, uma abordagem mais detalhada de como o *software FACETS* realiza esse método, apresentando sua aplicabilidade em avaliações nacionais.

5. Análise do impacto da dificuldade dos cadernos na proficiência dos alunos utilizando a Modelagem Rasch Multifacetas (MFRM)

As avaliações em larga escala são uma riquíssima fonte de aplicação da TRI e, ao mesmo tempo, levantam questões muito pertinentes aos seus pressupostos. É sabido que o foco de uma avaliação em larga escala é a escola, no máximo a turma, portanto é tolerável um grau maior de imprecisão na medida do aluno, em troca de uma melhor interpretação pedagógica dos resultados, através de uma maior abrangência tanto nos descritores quanto nos diferentes níveis de dificuldade dos itens. Nesse sentido, os alunos são submetidos a diferentes modelos de cadernos, através da utilização de BIB, técnica essa a ser apresentada no próximo tópico e a produção de uma medida de conhecimento através da utilização de modelos estatísticos pela TRI que equaliza todos os diferentes cadernos em uma mesma escala. Normalmente, no Brasil, utiliza-se a modelagem com apenas duas facetas, ou seja, dificuldade do item e proficiência do aluno.

Porém, embora tenha-se a preocupação de equilibrar a dificuldade dos diferentes modelos de cadernos na avaliação, esse balanceamento não é perfeito. O que se observa na prática, é a influência da dificuldade dos cadernos no desempenho dos alunos.

Na perspectiva de obter resultados mais equânimes ao nível dos alunos, neste estudo, utilizamos o modelo Rasch multifacetas com o objetivo de minimizar os efeitos da utilização de diferentes modelos de cadernos de testes para diferentes alunos.

5.1 Metodologia

Geralmente, nas avaliações em larga escala utilizamos mais de um instrumento de teste com o objetivo de mensurar a proficiência dos alunos. Nas avaliações que seguem o modelo do Saeb, são utilizados atualmente, 21 modelos de cadernos. Como os cadernos são distribuídos de forma aleatória entre os alunos

participantes, e que são avaliações que envolvem milhares e até milhões de alunos é de se esperar o mesmo valor de proficiência média, em cada um dos modelos de cadernos utilizados.

Tendo como referência os postulados da invariância da TRI, comumente definidas como teste-free e população-free. Para o primeiro postulado, temos que, para um mesmo valor de uma determinada grandeza ou constructo, o valor medido não pode sofrer alterações significativas em função do instrumento adotado, e para o segundo postulado, temos que, a dificuldade dos itens não é afetada por diferentes populações, caracterizadas por diferentes médias e desvios padrões de proficiências, utilizadas para a calibração dos itens.

Nossa metodologia de trabalho foi estruturada no conceito do teste-free, ou seja, diferentes modelos de cadernos distribuídos aleatoriamente em uma grande população devem possuir a mesma média de proficiência. Embora temos ciência que essa é uma propriedade do modelo, e não dos dados empíricos em si, conforme apontado por Wright & Stone (1999), é, portanto, natural de se esperar variações entre os modelos de cadernos. No entanto, três questionamentos sobre essa afirmação de Wright & Stone nos impulsionaram a realizar esse estudo:

1. As variações observadas através da utilização da modelagem 3PL podem ser minimizadas através da utilização da MFRM?
2. Qual o impacto da inclusão de facetas no modelo?
3. Qual o modelo mais parcimonioso para a obtenção de resultados mais justos (menor variação entre os cadernos)?

Com o emprego de uma base de dados que utiliza o mesmo desenho das avaliações do Saeb, ou seja, uma base de dados que caracteriza bem as avaliações em larga escala realizada no Brasil, confrontaremos os resultados dos alunos em diferentes modelos de cadernos por meio de resultados obtidos utilizando-se modelos 3PL e modelos MFRM. Nosso objetivo é verificar o impacto da inclusão de facetas nos modelos MFRM, no sentido de minimizar as variações entre os diferentes modelos de cadernos.

A metodologia está estruturada em quatro etapas: (i) descrição da base de dados, (ii) resultados via modelagem com 3PL, (iii) identificação do problema, (iv) resultados via modelagem MFRM e, (v) ajuste dos alunos e itens ao modelo MFRM.

5.1.1 Descrição da base de dados

Os dados utilizados no presente trabalho são oriundos de uma amostra de dez mil alunos do 5º ano da rede pública de um determinado Estado brasileiro, em 2016, utilizou-se uma base de dados, sem identificação dos alunos, e sorteada por amostragem aleatória simples.

O delineamento do teste utilizado, apresentado no Quadro 5, é formado por 21 modelos de cadernos, sendo cada caderno composto por duas disciplinas, a saber: Língua Portuguesa (P) e matemática (M). Para montagem dos itens nos cadernos, utilizou-se a técnica de BIB.

No BIB, os itens são organizados em blocos e esses, por sua vez, constituem cadernos de teste. Nesse *design*, temos sete blocos de itens para cada disciplina avaliada, blocos P01 a P07 para Língua Portuguesa e blocos M01 a M07 para Matemática, sendo que cada bloco é composto por 11 itens, portanto cada aluno responde a 22 itens de Língua Portuguesa e a 22 itens de Matemática. Onze testes iniciam-se com Língua Portuguesa e dez, com Matemática. Os blocos alternam entre os cadernos, em quatro posições distintas (P1 a P4), tentando-se ao máximo equilibrar o número de vezes que aparecem no início e ao final do caderno.

Quadro 5 – Exemplo de BIB, duas disciplinas 21 modelos de cadernos

Caderno	Posição dos blocos			
	P1	P2	P3	P4
1	P01	P02	M01	M02
3	P03	P04	M03	M04
5	P05	P06	M05	M06
7	P07	P01	M07	M01
9	P02	P04	M02	M04
11	P04	P06	M04	M06
13	P06	P01	M06	M01
15	P01	P04	M01	M04
17	P03	P06	M03	M06
19	P05	P01	M05	M01
21	P07	P03	M07	M03

Caderno	Posição dos blocos			
	P1	P2	P3	P4
2	M02	M03	P02	P03
4	M04	M05	P04	P05
6	M06	M07	P06	P07
8	M01	M03	P01	P03
10	M03	M05	P03	P05
12	M05	M07	P05	P07
14	M07	M02	P07	P02
16	M02	M05	P02	P05
18	M04	M07	P04	P07
20	M06	M02	P06	P02

Fonte: CAEd/UFJF (2016).

Essa técnica de montagem dos cadernos tem por objetivos, além de dificultar a “cola” durante a aplicação dos testes, permitir uma melhor interpretação pedagógica do desempenho dos alunos devido a um maior número de itens em

circulação no teste, avaliando assim, uma maior quantidade de habilidades por meio dos descritores dos itens.

Temos, no entanto, algo que poderíamos denominar de efeito colateral ao adotarmos tal técnica, pois ao submetermos os alunos a diferentes modelos de cadernos, temos uma avaliação amostral por itens, uma vez que os alunos não são submetidos a todos os itens e, além disso, um outro fator de grande impacto na geração das medidas é que a posição de itens que se repetem em diferentes cadernos não é a mesma. Essa última característica produz, conforme definido por Yen (1980) o efeito contexto, segundo o qual, os parâmetros dos itens são influenciados pela localização do item e/ou diferentes agrupamentos em que o item está inserido (natureza dos itens vizinhos). Isso se caracteriza pelo fato de que um item no início do teste tende a ter um parâmetro de dificuldade menor que o observado, caso ele esteja no final do teste, um dos motivos desse comportamento, é o efeito cansaço no aluno, ao se fazer um teste. Ao mudarmos as posições dos itens nos cadernos de testes, tornamos os seus parâmetros calculados em torno de um comportamento médio, minimizando o efeito contexto em equalizações futuras, caso essas avaliações sigam um design alinhado com as avaliações anteriores.

Para exemplificar o efeito contexto, apresentamos na Tabela 2, os percentuais de acerto dos itens do bloco P07 nos cadernos “6” e “7”.

Tabela 2 – Percentuais acerto nos itens do bloco P07 em relação aos cadernos 6 e 7.

Item	Caderno		Diferença
	6	7	(7) - (6)
P67	65,6%	75,1%	9,5%
P68	60,5%	61,0%	0,6%
P69	82,1%	86,7%	4,7%
P70	48,5%	49,0%	0,5%
P71	43,6%	49,6%	6,0%
P72	59,6%	69,0%	9,4%
P73	78,4%	84,1%	5,7%
P74	87,8%	95,1%	7,3%
P75	61,8%	66,3%	4,6%
P76	44,7%	46,1%	1,5%
P77	43,6%	51,4%	7,8%

Fonte: CAEd/UFJF (2016).

Podemos verificar que quando o bloco “7” aparece no final do caderno (caderno 6), os percentuais de acertos nos itens são menores que os percentuais encontrados quando o referido bloco está no início do caderno (caderno 7). A maior diferença, de 9,5% ocorre no item P67.

Entretanto, efeitos contextos presentes em todos os modelos de cadernos de testes utilizados, seja pela ordem das disciplinas, com cadernos ímpares começando com Língua Portuguesa e cadernos pares começando com Matemática e/ou as diferentes posições dos blocos nos cadernos, são alguns dos fatores que provocam as diferenças de médias entre os cadernos.

Com a utilização de facetas, representando diferentes contextos no *design* utilizado, teremos como ajustar as proficiências em função das características dos cadernos utilizados por cada aluno e dessa forma obter resultados com menos variações entre as médias dos 21 modelos de cadernos utilizados.

5.1.2 Resultados via modelagem 3PL

A estratégia utilizada para a análise da variação das proficiências médias entre os diferentes modelos de caderno foi confrontar os resultados obtidos através da modelagem 3PL, que será a nossa modelagem de referência, com os resultados obtidos com a utilização da MFRM. Essa referência se justifica pelo fato da modelagem 3PL, conforme abordado em capítulos anteriores, ser a modelagem oficial utilizada pelo Inep e também a que foi utilizada pelo CAEd na avaliação desse projeto.

Iniciaremos, portanto, com a modelagem da TRI a 3PL para um único grupo, sem fixar nenhum item, nem utilizar a proficiência oficial na calibração dos parâmetros, para que, assim, fosse evidenciada, ao máximo, a relação entre padrão de respostas e desempenho. O pacote utilizado foi o MIRT (Chalmers, 2012) do *Software* R.

5.1.3 Identificação do problema

Por meio da modelagem 3PL, podemos constatar, na Tabela 3, 21 subgrupos da população obtendo médias distintas de proficiência, em função do modelo de caderno que realizaram. Podemos verificar, por exemplo, que a diferença de proficiência entre os alunos que fizeram o caderno 14 (-0,139) e 5 (0,146) é de 0.2854, ou seja, a diferença observada chega a mais que 1/5 de desvio-padrão (dp), em duas subpopulações que deveriam possuir a mesma média e desvio-padrão, admitindo apenas variações amostrais.

Tabela 3 – Proficiência em Língua Portuguesa para alunos submetidos a diferentes modelos de cadernos.

Caderno	Proficiência		Caderno	Proficiência	
	Média	dp		Média	dp
1	0,033	0,989	2	-0,099	1,027
3	0,089	0,989	4	-0,126	1,022
5	0,146	0,971	6	-0,035	1,022
7	0,102	0,971	8	-0,116	1,044
9	-0,020	0,987	10	-0,079	1,003
11	0,079	0,987	12	-0,014	0,955
13	0,069	0,979	14	-0,139	1,022
15	0,141	1,034	16	-0,108	0,987
17	0,066	1,002	18	-0,013	1,010
19	0,085	0,971	20	-0,102	0,993
21	0,024	0,965	Todos os cadernos	0	1

Fonte: CAEd/UFJF (2016).

Nota-se que em todos os cadernos pares (Língua Portuguesa no final do caderno) os valores de proficiência se situam abaixo da média considerando todos os cadernos (valor igual a 0), já entre os ímpares (Língua Portuguesa no início do caderno), com exceção do caderno 9, todos os demais cadernos possuem média acima da média geral (todos os cadernos).Também é possível constatar que sete dos dez cadernos pares possuem desvios-padrões superior ao desvio-padrão geral, enquanto entre os ímpares apenas dois em 11 cadernos manifestam tal situação.

Em nossas simulações, realizadas apenas para a disciplina de Língua Portuguesa, consideramos as características dos diferentes cadernos de testes como facetas adicionais na modelagem via MFRM, no intuito de identificar quais dessas

facetas são mais significativas para um melhor equilíbrio entre as proficiências dos alunos, pois, conforme apresentado anteriormente, através dos resultados obtidos pela modelagem via 3PL, detectamos diferenças significativas entre os mesmos.

5.1.4 Resultados via modelagem MFRM

O modelo de 3PL se torna extremamente complexo no caso de se considerar as inúmeras circunstâncias em torno do item, portanto, optou-se por um modelo mais simples, o de Modelo de *Rasch*, de forma a se poder trabalhar com mais facetas, através da modelagem MFRM, conforme apresentado no modelo a seguir:

$$\boldsymbol{P(Y_{ij} = 1 | \theta_i, \xi_j) = logit(\theta_i - \sum_{k=1}^{F} b_{jk})}$$

$K = 1, 2 \ldots F$ sendo, F número total de facetas utilizadas

Y_{ij} resposta do i-ésimo aluno ao j-ésimo item

θ_i traço latente do i-ésimo aluno

b_{jk} dificuldade da k-ésima faceta para o j-ésimo item

Nota-se que nesse modelo por simplicidade não são adotados, explicitamente, fatores de escala, (por convenção, parâmetro a = 1).

A cada circunstância que envolve o item, bem como ao próprio item pode se associar uma faceta, em nossas simulações consideramos as seguintes facetas:

- Bloco: conjunto de itens sempre aplicados simultaneamente, neste presente trabalho foram utilizados blocos contendo 11 itens;
- Caderno: conjunto de blocos que compõe um teste, no presente estudo são utilizados dois blocos por caderno;
- Posição do bloco: posição na qual ele é exposto no caderno (posições de 1 a 4);
- Posição da disciplina Língua Portuguesa no caderno: início ou final do caderno.

Uma vez identificadas as facetas construímos vários modelos com diversas combinações de facetas no intuito de observar o efeito produzido na variabilidade das médias de proficiências entre os cadernos. Em seguida, conjugamos as

informações do entorno do problema de forma a elaborar nossa estratégia de trabalho: temos uma mesma população, logo qualquer subpopulação que seja amostra aleatória desta, deve possuir a mesma média de proficiência. Nessas subpopulações, que no caso, são os alunos agrupados em cada um dos 21 modelos de cadernos utilizados, temos alunos que foram inicialmente avaliados em Língua Portuguesa ou em Matemática, e sendo submetidos a blocos de itens em diferentes posições nos cadernos. Afinal, quais as consequências que tais práticas podem acarretar?

Visando responder a essas questões, foram realizadas oito simulações, variando de duas facetas a seis facetas conforme apresentado no Quadro 6.

Quadro 6 – Modelos, facetas utilizadas e resultados encontrados nas simulações

Simulação	Modelo	*Software*	Facetas						Amplitude
			Aluno (A)	Item (I)	Caderno (C)	Posição Disciplina (PD)	Bloco (B)	Posição do Bloco (PB)	
Referência	3PL	MIRT	X	X					0,285
1	MFRM	*FACETS*	X	X					0,321
2			X	X			X		0,320
3			X	X	X		X	X	0,169
4			X	X	X		X		0,166
5			X	X	X				0,158
6			X	X	X	X			0,105
7			X	X	X	X	X		0,089
8			X	X	X	X	X	X	0,088

Fonte: o Pesquisador.

As seis facetas consideradas foram: aluno (A) e item (I), que são as duas facetas básicas utilizadas nas modelagens da TRI e as facetas complementares, representadas pelo caderno (C), posição da disciplina (PD), Bloco (B) e posição do bloco (PB).

Duas facetas em particular, merecem uma melhor descrição de suas constituições, a primeira é a faceta posição da disciplina, que assume os valores “1” ou “2”, estando a disciplina de Língua Portuguesa no início ou final do caderno, respectivamente. A segunda faceta em questão é a faceta posição do bloco que assume os valores “1”, “2”, “3” e “4” em função da posição do bloco no caderno.

Nas simulações de “1” a “8” realizadas, optou-se por utilizar o programa *FACETS* (Linacre, 2013) como ferramenta de análise, pelo fato desse *software* compreender uma ampla variedade de modelos da TRI. Especificamente, a partir

do *FACETS*, pode-se trabalhar com itens dicotômicos ou politômicos em modelos multifacetados.

Tais modelos são chamados multifacetados, porque consideram não somente as características dos itens e dos respondentes (alunos), como nas análises tradicionais, mas também, porque possibilitam considerar os efeitos de diversas outras facetas conforme apresentadas no Quadro 6.

Para efeitos comparativos, em todas as simulações realizadas, todas as medidas de proficiência dos alunos foram padronizadas em uma normal (0,1), a fim de garantir a comparabilidade entre os diversos modelos. Os estudos foram ordenados em função dos resultados das amplitudes de proficiências entre os cadernos (ordem decrescente), obtidos no modelo de referência e nos 8 modelos onde utilizamos a MFRM.

Sendo que a amplitude foi calculada pela diferença entre o maior e menor valor de proficiência média encontrada em cada modelo de caderno:

Amplitude = maior proficiência – menor proficiência

Podemos classificar os resultados das simulações pelas amplitudes encontradas, em três grupos com características distintas:

- ✓ O primeiro grupo caracteriza-se pelos modelos MFRM que apresentam as maiores amplitudes entre os cadernos e que se assemelham ao modelo 3PL de referência, são as simulações “1” e “2”, representadas pelos modelos “F_A_I” e “F_A_I_B”, respectivamente. Tal fato já era esperado, uma vez que a simulação 1 tem as mesmas facetas da simulação de referência, diferenciando apenas na mudança do modelo da TRI de 3PL para Rasch e, na simulação 2, a inclusão da faceta bloco não altera de forma significativa os valores encontrados na simulação “1”, pois como cada item pertence a um e somente um bloco, o balanceamento dos blocos pelo BIB, faz com que a inserção do bloco, como mais uma faceta no modelo, seja praticamente sem eficácia no ajuste das diferenças entre os cadernos;
- ✓ No segundo grupo encontram-se as simulações “3”, “4” e “5” representas pelos os modelos “F_A_I_C_B_PB”, “F_A_I_C_B”, e “F_A_I_C”. Nesses modelos, a dificuldade do caderno é a faceta relevante para um maior

equilíbrio entre as proficiências, sendo praticamente irrelevante considerar ou não o bloco, pelo motivo anteriormente citado;

- ✓ No terceiro grupo, na colocação de melhor eficiência, encontram-se as três últimas simulações pelos modelos "F_A_I_C_PD", "F_A_I_C_PD_B" e "F_A_I_C_PD_B_PB", nos quais, além de considerarmos a faceta dos cadernos, inserimos, também, as facetas posição da disciplina e posição do bloco.

Como resultado final tem-se o modelo saturado "F_A_I_C_PD_B_PB" como o mais eficiente por natureza, pois engloba os efeitos de todas as facetas, mas é muito menos parcimonioso. O modelo mais parcimonioso, obtido na simulação "6", representado pelo modelo "F_A_I_C_PD", apresenta resultados praticamente iguais aos encontrados nas simulações "7" e "8", diferenciando-se em centésimos, e demonstrando que o número do caderno e a ordem da disciplina no caderno são as facetas mais relevantes, uma vez que, quando essas facetas são consideradas no modelo, consegue-se atenuar as diferença de amplitude entre os modelos de cadernos para 0,105. Se compararmos esse valor com a amplitude obtida no modelo de referência, que foi de 0.285, podemos constatar que a queda na amplitude entre esses dois modelos foi de quase o triplo (2,71).

5.1.5 Ajuste dos alunos e itens ao modelo MFRM

As análises de ajuste, no *FACETS*, foram realizadas através das estatísticas outfit e infit média ao quadrado (*Mean-square*). Conforme apresentado no tópico 4.4.1.

Na Tabela 4, apresentamos os resultados das estatísticas de ajuste para as facetas consideradas, em que podemos constatar o nível ajuste das facetas ao modelo MFRM.

Tabela 4 – Ajustes das facetas ao modelo

Faceta	Descrição	Casos na faceta	Categorias de ajuste							
			Infit				*Outfit*			
			1	2	3	4	1	2	3	4
1	Aluno	10.000		145	9.854	245	591	8.903	261	
2	Item	76			76			2	74	
3	Caderno	7			7				7	
4	Posição da disciplina	2			2				2	
5	Bloco	7			7				7	
6	Posição do bloco	2			2				2	

Fonte: CAEd/UFJF (2018).

As facetas de “2” a “6” tiveram todos os seus elementos entre as categorias “2” e “3” de *infit* e *outfit*, revelando o excelente ajuste ao modelo. Com relação à faceta “aluno”, tivemos apenas 245 casos, ou seja, 2,5% da base com desajuste no *outfit*, o que é um número muito baixo. Esses valores de ajustes, nos permitiram conduzir nossas análises e validar as nossas conclusões dentro de limites estatísticos aceitáveis.

5.2 Considerações do capítulo

Esse estudo pretendeu demostrar a versatilidade do modelo de facetas, um modelo da classe de Rasch para modelar um problema complexo com grande facilidade, o que seria muito mais difícil em um modelo de 3PL dado sua estrutura complexa por natureza. O modelo proposto torna-se útil em situações onde o foco da avaliação é o aluno em delineamentos de testes compostos por diferentes modelos de cadernos e com mais de uma disciplina avaliada por caderno.

Procuramos demonstrar que, através da inserção de facetas apropriadas no modelo da TRI é possível estimar medidas mais homogêneas para os alunos, minimizando o efeito cansaço provocado pela ordem da disciplina nos cadernos de testes assim como o efeito da dificuldade inerente a cada modelo de caderno de teste.

Como temos nesse estudo 21 modelos de cadernos, a correlação entre os percentuais de acerto e a proficiência pela MFRM não é biunívoca, entretanto, as variações de proficiências por percentual de acerto, considerando essa modelagem, são bem inferiores ao encontrado quando se utilizar a modelagem com três parâmetros, conforme pode ser visualizado através dos Gráficos 5 e 6 (*scatter*

plots), onde, no Gráfico 5, apresentamos a correlação para os percentuais de acerto e proficiência via 3PL e no Gráfico 6, a correlação para os percentuais de acerto e a proficiência MFRM via simulação “6”, que foi a simulação através a MFRM com o melhor desempenho.

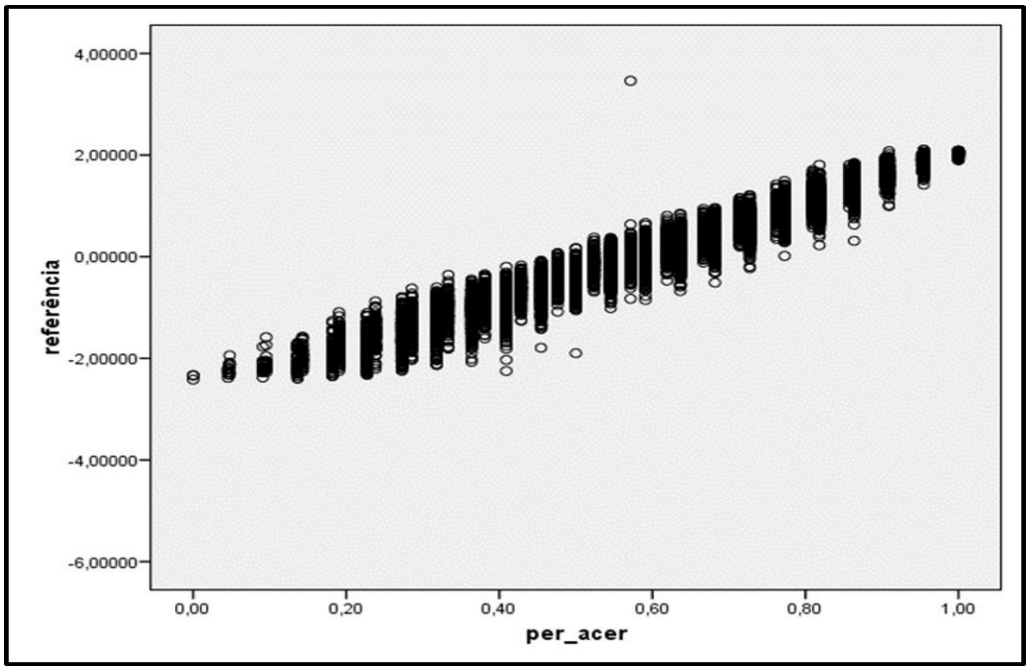


Gráfico 5 – Dispersão entre percentual de acerto e proficiência via 3PL

Fonte: o Pesquisador.

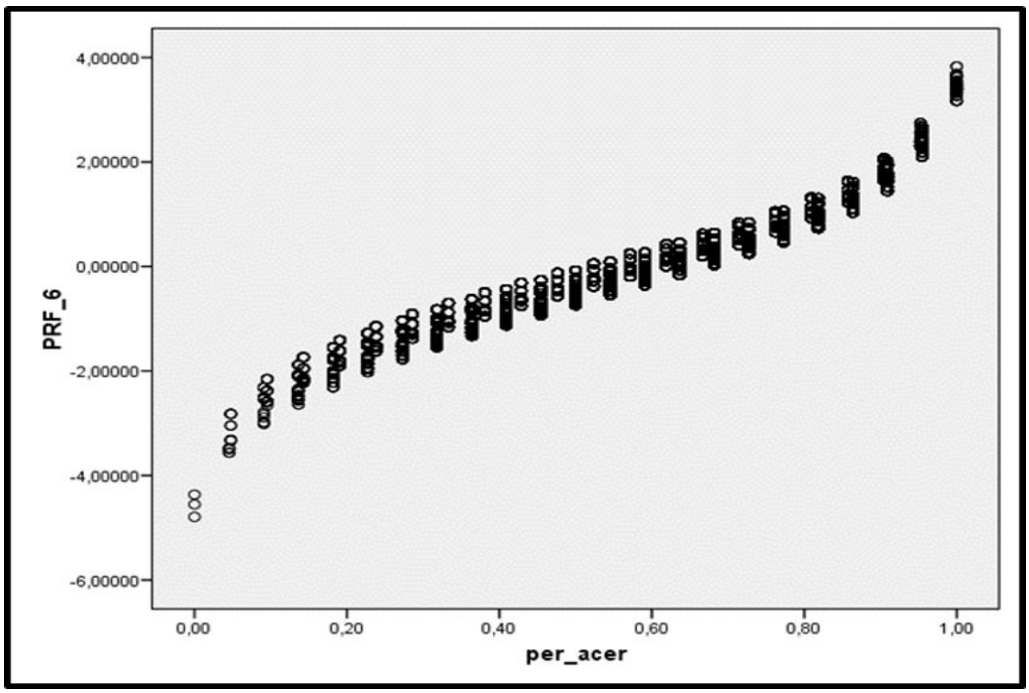


Gráfico 6 – Dispersão entre percentual de acerto e proficiência via MFRM (simulação 6)

Fonte: O Pesquisador.

Mesmo com o melhor ajuste da proficiência via MFRM com os percentuais de acerto, essa correspondência não é biunívoca, o que impede a sua utilização automática por parte dos professores.

Podemos dessa forma, sumarizar a utilização dos resultados da utilização da MFRM em função da aplicação da avaliação, da seguinte forma:

- ✓ Caso de utilização de um único modelo de caderno: a correspondência entre o percentual de acerto e a proficiência é biunívoca e a interpretação dos resultados dos alunos através de uma escala de proficiência é realizada através de uma simples correspondência entre os percentuais de acerto e a proficiência;
- ✓ Caso de utilização de mais de um modelo de caderno: a correspondência entre o percentual de acerto e a proficiência não é biunívoca. A MFRM aproxima bem essa relação, mas para a realização da correspondência entre os percentuais de acerto e a proficiência há a necessidade de se levar em consideração o número do caderno. Uma solução, prática, para que o próprio professor faça essa operação, pode ser através de uma planilha de excel em que o professor fornece o percentual de acerto e o número do caderno que o aluno utilizou. Com essas duas informações, através de uma equação linear obtém-se a proficiência do aluno;
- ✓ A produção de uma medida na escala Rasch, com as vantagens apresentadas anteriormente, pode ainda ser utilizada de forma comparada com outras avaliações que utilizam outras modelagens, como por exemplo a modelagem 3PL. Através, por exemplo do método do equipercentil, conforme apresentado por Livingston (2004), as distribuições das proficiências dos alunos, que realizaram duas avaliações distintas, são emparelhadas e dessa forma, é possível estabelecer a correspondências entre as duas medidas e efetuar a comparabilidade do desempenho dos alunos nas duas avaliações.

Essa análise pode ser realizada, em uma situação em que os alunos que realizaram uma avaliação local com a modelagem Rasch, também realizaram, por exemplo, uma avaliação Nacional da Prova Brasil. Através da comparação das distribuições das proficiências nas duas avaliações, é possível estabelecer uma correspondência entre as mesmas e assim ter um indicativo do quanto o valor da medida em Rasch se assemelha à medida na Prova Brasil, permitindo a professores

e gestores educacionais analisarem pedagogicamente o desempenho de seus alunos de forma comparada, nas duas avaliações.

Nossa intenção não é aprofundar nessa metodologia, a qual pode ser estudada com maior profundidade, permitindo inclusive sua aplicação, em Livinston (2004), onde além do método do equipercentil são apresentados outros métodos de comparabilidade entre avaliações, com suas vantagens e limitações.

6. Correção de redações utilizando a Modelagem Rasch Multifacetas (MFRM)

A produção de medidas de desempenho de alunos onde se utiliza testes de escrita, é afetada pela subjetividade dos corretores envolvidos no processo de correção. Essa subjetividade, está relacionada à severidade de cada corretor e também às suas interpretações dos critérios adotados no processo de correção, critérios esses nem sempre homogêneos entre os corretores.

Trabalharemos nesse capítulo, duas técnicas utilizadas na tentativa de controlar o efeito da subjetividade dos corretores: A dupla correção via TCT e a dupla correção via MFRM. Sendo a primeira a técnica atualmente adotada no Brasil e a segunda, o modelo alternativo proposto nessa tese. Através de análises estatísticas desenvolvidas ao longo desse capítulo, pretendemos demonstrar os ganhos advindos da adoção da modelagem Rasch-multifacetas para esses tipos de avaliações que utilizam itens de escrita. Nossa hipótese de trabalho é que através da utilização da MFRM em avaliações que utilizam itens de escrita, é possível obter medições de melhor qualidade do desempenho dos alunos ao que se obtêm com as medidas via TCT, normalmente utilizadas no Brasil.

Utilizaremos, para esses estudos, os dados das correções de redações (Produção de textos) do Sistema de Avaliação Baiano de Educação (Avalie BA), no ano de 2011, doravante, nos referiremos a esse projeto como SABE-2011. Essa avaliação foi realizada pelo CAEd/UFJF, no qual, 123 corretores trabalharam, através de um sistema *on-line*, na produção de notas nas redações dos alunos. Neste projeto, cada redação foi corrigida pelo menos duas vezes por diferentes corretores e, no caso de discrepância entre as duas notas de um mesmo aluno, foi realizada uma terceira correção por um supervisor. Esta dinâmica no processo de correção, adotada também pelo Enem e pelo Exame Nacional para Certificação de Competência de Jovens e Adultos (Encceja), garante certo controle sobre a subjetividade dos critérios de correção por parte dos corretores.

Realizamos um estudo comparativo entre esse procedimento de produção de pontuação das redações realizadas pelos alunos e os resultados de desempenho obtidos com a utilização da MFRM, por meio das medidas, nota e proficiência,

respectivamente.

O principal objetivo desse estudo é definirmos formas de melhor estimar as pontuações dos alunos nas redações, através do controle dos efeitos subjetivos de cada corretor, assim como a possibilidade de construção de uma escala de redação a ser adotada no país.

6.1 Descrição da avaliação da escrita

Conforme apresentado na Revista Pedagógica do SABE (2011, p. 55), Avalie Ensino Médio 2011:

> A avaliação da escrita por meio de uma produção textual temática, como no caso do Avalie Ensino Médio, permite aferir a proficiência dos estudantes quanto ao uso da língua escrita, desde a leitura da proposta de produção textual até a elaboração do texto, uma vez que, ao produzir um texto que atenda ao solicitado pela proposta, os estudantes estavam demonstrando domínio entre os processos de oralidade, leitura e escrita.

Tendo a diversidade cultural baiana como temática, a proposta de produção escrita foi sustentada por três textos motivadores pertencentes a gêneros distintos. Esse eixo constituiu o tema no qual os estudantes deveriam produzir os seus textos.

Para realizar a tarefa de produção de texto, esses estudantes precisariam ler atentamente os textos motivadores, atentando-se ao fato de que cada um deles apresentava objetivos comunicativos distintos. A produção de texto pode ser verificada no Anexo 5.

6.1.1 Descrição do processo de correção

A avaliação das competências linguístico-textuais, foi estruturada conforme descrito na Matriz de Competências para Produção de Texto (Anexo 4), em quatro competências de escrita divididas, cada uma, em seis níveis de classificação, sendo avaliadas as habilidades mínimas exigidas a um escritor proficiente cursando a primeira série do ensino médio. A Matriz de Referência foi elaborada pelo

CAEd/UFJF, em parceria com o corpo pedagógico da Secretaria de Estado de Educação da Bahia. Essa matriz encontra-se presente no Anexo 2.

A correção das redações do Avalie 2011 foi realizada *online.* Para isso, todas as redações foram digitalizadas, com um código para identificação posterior, e retirados os elementos que possibilitem a identificação do candidato.

Assim, a equipe de correção de redação trabalhou diretamente em um sistema eletrônico, o qual gerou relatórios automáticos acerca do trabalho de cada um dos corretores. De modo a garantir a qualidade e a unificação dos trabalhos de correção, a equipe de correção foi estruturada da seguinte forma:

- ✓ O número de corretores foi calculado tendo em vista que cada redação seria corrigida pelo menos duas vezes por diferentes corretores, e que cada corretor teria uma produção de, em média, 150 correções por dia;
- ✓ Seis Supervisores de Correção foram responsáveis pela equipe de corretores. A cada um dos especialistas envolvidos no processo de correção das redações foi fornecida uma permissão de acesso ao sistema de modo a poderem desenvolver suas atividades;
- ✓ A fim de orientar os procedimentos e critérios de correção das provas de redação o CAEd/UFJF realizou o treinamento dos profissionais envolvidos;
- ✓ Após receberem o treinamento, os Supervisores repassaram as informações e orientações recebidas aos Corretores. Antes da liberação do sistema de correção online, foi corrigida uma amostra significativa das redações, de modo a se padronizar a aplicação dos critérios de correção, bem como os procedimentos necessários à utilização do sistema de correção.

Após esses procedimentos, o trabalho de correção obedeceu à seguinte organização:

- ✓ Cada Supervisor recebeu, via sistema eletrônico, um determinado número de redações e as repassou aos Corretores sob sua supervisão;
- ✓ Cada redação foi encaminhada a dois corretores, os quais estavam sob seis diferentes supervisões;
- ✓ Após a correção, o sistema emitia automaticamente um relatório das redações que apresentavam notas discrepantes. Essas redações discrepantes, foram encaminhadas para uma terceira correção a ser realizada pelo supervisor que, desconhecendo as notas dadas anteriormente, foi o responsável pela nota final da redação;

- ✓ Cada Supervisor acompanhou o trabalho dos Corretores sob sua supervisão e aqueles que apresentaram por três vezes, divergências em suas correções, foram chamados e solicitados a reverem as correções das redações sob suas responsabilidades;
- ✓ Cada Corretor foi responsável, diariamente, pela correção de cem a 150 redações. Esse quantitativo é considerado ideal de forma a não sobrecarregar os corretores e garantir um trabalho de boa qualidade;
- ✓ A seleção e o treinamento dos professores especialistas para constituição da equipe de correção das redações foi de inteira responsabilidade do CAEd/UFJF. Todas as etapas relativas ao trabalho de correção das redações foram pautadas pelos critérios previamente indicados pela Secretaria de Educação da Bahia e acompanhado por representantes indicados por essa Secretaria;
- ✓ O CAEd/UFJF providenciou a correção das redações dos participantes que apresentaram necessidades educacionais especiais, conforme legislação em vigor;
- ✓ Todos os envolvidos no processo de correção das redações assinaram um Termo de Sigilo.

6.1.2 Cálculo da nota do aluno

O projeto teve dupla correção com resolução de discrepâncias por parte do supervisor. As discrepâncias eram geradas quando as correções apresentavam situações de correções distintas, como por exemplo Fuga ao tema, ou desconsiderado (impossibilidade de leitura), ou por diferença superior a três pontos (inclusive) entre as médias finais de cada correção.

Apresentamos a seguir os passos adotados para o cálculo da nota dos alunos:

- 1º passo: identificação do nível do aluno por item.

Cada corretor, com o auxílio da Matriz de Competências para Produção de Texto (Anexo 4), localizava o nível do aluno em cada uma das quatro competências e atribuía uma nota conforme Quadro 7.

Quadro 7 – Correspondência nível x valor da nota

Nível	Nota do corretor	Valor
0	0	0
1	0,1 a 2,0	2
2	3,0 a 4,0	4
3	4,1 a 6,0	6
4	6,1 a 8,0	8
5	8,1 a 10	10

Fonte: O Pesquisador.

- 2º passo: cálculo da nota do aluno por corretor.

O cálculo da nota final de cada aluno por corretor foi obtido pela média aritmética simples dos valores de cada uma das quatro competências.

- 3º passo: cálculo da nota final do aluno.

O cálculo da nota final do aluno depende do nível de discrepância entre as duas correções realizadas pelos corretores:

- ✓ Se a diferença entre as duas notas for menor que três pontos a nota final é obtida pela média aritmética simples das duas notas;
- ✓ Se a diferença entre as duas notas for maior que ou igual a três pontos a nota final é obtida através de uma terceira correção, realizada pelo supervisor, sendo as duas primeiras notas descartadas.

6.1.3 Descrição da base de dados

Após os devidos tratamentos, como retirada de casos em branco, redações eliminadas por fuga ao tema e impossibilidade de leitura, obtivemos uma base de dados com as seguintes características:

- Total de alunos: 74.895
- Redações com duas correções: 70.603 (94,3%)
- Redações com três correções: 4.292 (5,7%)
- Total de supervisores: 6
- Total de corretores: 117

6.2 Modelagem estatística via TRI - MFRM

A modelagem estatística via MFRM foi implementada considerando cada competência e seus respectivos níveis como sendo itens politômicos com 6 categorias de respostas. Assim, a proficiência do aluno foi calculada tendo como referência um teste composto por quatro itens politômicos, conforme Quadro 8.

Quadro 8 –Itens politômicos utilizados na MFRM

Item	Competência	Nº. de categorias
1	Registro	6
2	Tema	6
3	Tipologia Textual	6
4	Coesão/Coerência	6

Fonte: O Pesquisador.

O modelo de crédito parcial Rasch-Masters foi aplicado. Nesse modelo, conforme apontado por Pasquali & Primi (2003) e Primi (2004), quanto maior a habilidade do aluno, maior será a sua probabilidade de acertar um determinado item do teste. Essa probabilidade também depende da complexidade do item. Nesse sentido, a probabilidade de dominar uma determinada habilidade varia de acordo com o nível de conhecimento da pessoa e da complexidade do item.

6.2.1 Utilização do MFRM

Para a produção da proficiência dos alunos via MFRM, utilizamos o *software FACETS* (Linacre 2014), cuja operacionalidade envolve a construção de uma sintaxe e de uma base de dados. Apresentamos a seguir os detalhes técnicos desses dois arquivos.

6.2.1.1 Sintaxe

Por meio da sintaxe apresentada na Figura 16, produzimos medidas pela TRI considerando três facetas: (i) severidade dos corretores, (ii) proficiência dos alunos

e (iii) dificuldade dos itens. Procuramos, por meio dessa modelagem, contornar o problema de termos diferentes alunos tendo suas redações corrigidas por diferentes perfis de corretores (severidade), de tal forma que ao final do processo de produção de medidas pela MFRM é como se cada aluno tivesse tido sua redação corrigida por um único corretor padrão, entendendo como corretor padrão, um corretor com um nível de severidade mediano. Dessa forma, o objetivo principal é tornamos mais justos os resultados de desempenho obtidos pelos alunos.

```
Title = REDAÇÃO  SABE-2011;
Score file = REDAÇÃOSC,sav;
Facets = 3 ; 3 facetas: CORRETORES, ALUNOS e CRITÉRIOS (ITENS)
Inter-rater = 1
Positive = 2 ; A faceta 2 ( alunos ), Quanto maior o  score  (nota) implica
               que maior será a medida (proficiência).
Noncenter = 2 ; apenas a  faceta 2, alunos não terá a medida de
                  proficiência (medida) fixada em zero.
Pt-biserial = Yes ; report the point-biserial correlation
Yard = 112,4 ; Vertical rulers 112 columns wide, with 4 lines per logit
Model = ?,?,#,R6
Labels =
  1,CORRETORES
  1-123;(elementos=123)
  *
  2,ALUNOS
  1-74895;(elementos=74895)
  *
  3,ITENS
  1-4; (elementos=4)
  *
Data = REDAÇÃO.sav;
*
```

Figura 16 – Sintaxe redação SABE-2011

Fonte: O Pesquisador.

Descreveremos a seguir os principais comandos desta sintaxe:

a) Interrater

Esse comando faz com que o programa calcule estatísticas de concordância entre os corretores. São produzidas duas estatísticas por corretor: a concordância exata observada e a concordância exata esperada. A primeira medida é o percentual de concordância das notas de um corretor com as notas dadas pelos demais corretores; e a segunda medida está relacionada com o percentual de concordância

que seria observado se o dado se ajustasse perfeitamente ao modelo Rasch. Através da comparação dessas duas medidas temos como mapear a qualidade de cada corretor e do processo como um todo.

b) Positive

Em educação a convenção normalmente adotada é que a proficiência dos alunos é positiva e todas as demais são negativas (Linacre, 2014). Assim, em nosso modelo, a faceta "2', aluno, é a única com correlação positiva (positive = 2), ou seja, quanto maior a nota, maior a proficiência e as demais facetas (corretores e itens) são negativas, ou seja quanto maior a severidade do corretor e a dificuldade do item, menor será a proficiência do aluno.

c) Noncenter

Para medirmos alguma coisa, temos que especificar uma origem. Normalmente as médias têm uma referência em zero, sendo as mais comuns no dia-a-dia as medidas de distância e temperatura. O que geralmente causa estranheza para quem não é da área de psicometria ou estatística, é que, as medidas das facetas obtidas pela TRI, não estão referenciadas em zero, mas sim no centro de suas distribuições, sendo que estas distribuições, via de regra, são distribuições normais.

Por imposição dos modelos multifacetados de Rasch, todas as facetas possuem distribuições normais e as medidas são referenciadas no centro de suas distribuições, com exceção da proficiência dos alunos (Linacre, 2014). Assim, em nosso modelo as distribuições normais das severidades dos corretores e dificuldades dos itens estão referenciadas em zero, ou seja, a severidade média dos corretores e dificuldade média dos itens está localizada no centro de suas respectivas distribuições. Já a proficiência média dos alunos não está referenciada no centro de sua distribuição (Noncenter = 2), mas sim no centro das distribuições das duas outas facetas.

d) Model = ?,?,#,R6

O modelo da TRI adotado para a modelagem dos itens foi o modelo Rasch-Master de crédito parcial (1982), representado pelo símbolo "#" no comando Model. Esse modelo pode ser considerado como uma extensão do modelo de escala gradual de Andrich, e é adequado para situações em que os itens, além de

apresentarem respostas ordenadas por níveis de dificuldades (categorias), essas respostas também podem variar em quantidades e distancias entre as mesmas.

O modelo de Andrich, em que todos os itens apresentam o mesmo formato de CCI, variando apenas o parâmetro D (dificuldade) é ideal para questionários que utilizam itens do tipo Likert, em que as categorias das respostas, geralmente indicam níveis de concordância do entrevistado em relação a algum tema específico, como por exemplo: Discordo totalmente, discordo parcialmente, neutro, concordo parcialmente e concordo totalmente.

Em nosso estudo, se os itens fossem modelados conforme proposto por Andrich, Cada um dos quatro itens teria um valor distinto do parâmetro de dificuldade “b”, que no caso de um item politômico é a interseção das curvas das primeira e última categorias do item. O que caracteriza esse modelo é o fato das distâncias entre as categorias dos itens (representadas pelas interseções de duas categorias adjacentes) serem equidistantes dentro de um mesmo item e também, que essas distâncias são as mesmas entre os itens. Assim, as curvas de todos os itens de um teste teriam a mesma representação gráfica, variando apenas a posição do parâmetro de dificuldade ‘b”. Apresentamos, no Gráfico 7, a curva CCI, obtida no SABE-2011 adotando-se essa modelagem.

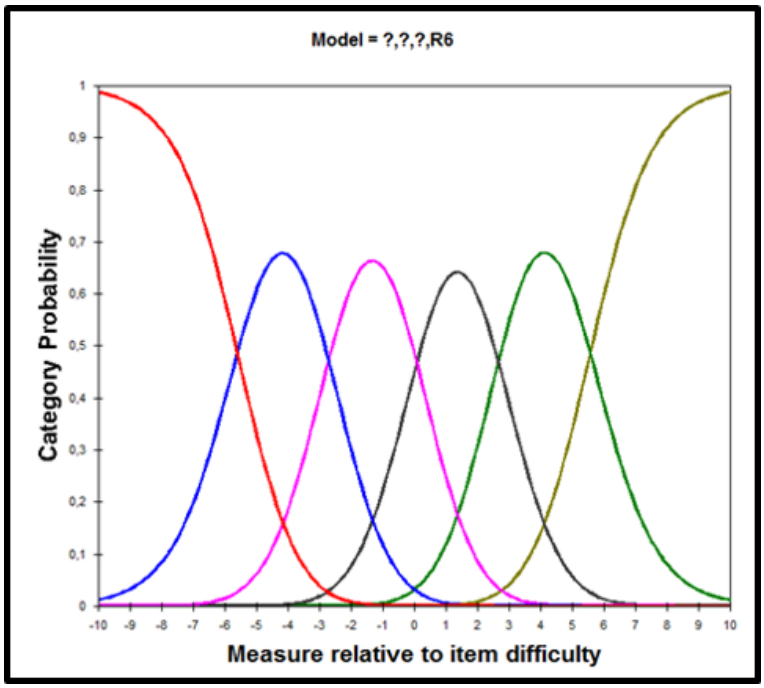


Gráfico 7 – CCI dos itens pelo modelo de resposta gradual de Andrich

Fonte: O Pesquisador através do *software FACETS* 2018.

Conforme pode ser observado pelas características do SABE-2011, esse modelo não seria adequado para o nosso estudo, pois é de se esperar que as distâncias entre as seis categorias de um mesmo item não sejam, necessariamente, equidistantes entre si e muito menos, que essas distâncias sejam, necessariamente, as mesmas para todos os itens.

Em função dessa característica, adotamos o modelo de crédito parcial de Master, onde a probabilidade, P_{nij}, que um aluno n com habilidade B_n seja observado na categoria j de uma escala gradual do item i de dificuldade D_i ao contrário da probabilidade $P_{ni(j-1)}$ de ser observado na categoria (j-1) é dada pela formulação a seguir:

$$\boldsymbol{log_e\,(P_{nij}/P_{ni(j-1)}) = B_n - D_i - F_{ij}}$$

Ou

$$\boldsymbol{log_e\,(P_{nij}/P_{ni(j-1)}) = B_n - D_{ij}}$$

No modelo de crédito parcial, a estrutura da escala gradual, representada pelo fator $\{F_{ij}\}$ é específica para o item *i*, ou seja, o modelo apresenta não apenas diferentes valores dos parâmetros de dificuldade dos itens (D), mas também, diferentes intervalos entre as categorias de respostas de cada item, conforme pode ser constatado no Anexo 6, onde apresentamos as curvas características de cada um dos quatro itens da redação modelados com seis categorias.

6.2.1.2 Base da dados

A base de dados utilizada para processamento foi elaborada utilizando-se o *software* SPSS, conforme estrutura apresentada no Quadro 9.

Quadro 9 – Estrutura da base de dados para processamento

CORRETOR	ALUNO	ITEM	CC1_T	CC2_T	CC3_T	CC4_T
17	1	1-4A	5	4	3	5
48	1	1-4A	4	3	4	3
53	2	1-4A	4	6	5	4
110	2	1-4A	4	4	3	3
27	3	1-4A	4	3	3	2
34	3	1-4A	5	4	4	3
26	4	1-4A	2	3	3	2
35	4	1-4A	2	3	2	2
25	5	1-4A	3	4	3	3
36	5	1-4A	4	5	4	4
71	6	1-4A	4	3	4	4
103	6	1-4A	3	4	4	4
25	7	1-4A	4	5	5	5
36	7	1-4A	5	5	5	4
26	8	1-4A	3	4	4	4
35	8	1-4A	3	5	4	4
4	9	1-4A	4	4	4	3
71	9	1-4A	3	2	3	3
103	9	1-4A	5	4	6	5
17	74894	1-4A	2	2	2	1
49	74894	1-4A	2	3	3	3
58	74895	1-4A	4	4	4	4
71	74895	1-4A	3	2	3	2

Fonte: Avalie Bahia (2011).

Essa estrutura da planilha, está no formato exigido pelo *software FACETS* e foi elaborada no *software* SPSS. A primeira coluna representa a primeira faceta e é composta por 123 corretores e supervisores, representados pelos números de “1” a “6” para os supervisores e de “7” a “123” para os corretores. A segunda coluna representa a segunda faceta do modelo, que são os alunos, os quais foram codificados e “1” a “74.895”. A terceira coluna, é apenas uma indicação que as próximas colunas representam os itens do teste, assim, as colunas de “5” a “7”, representam os quatros itens do teste.

Para exemplificar, temos na primeira linha, que o corretor 17, corrigiu o teste do “aluno 1”, e deu as seguintes pontuações para esse aluno em cada um dos quatro itens do teste:

- ✓ item 1: 5 pontos
- ✓ item 2: 4 pontos
- ✓ item 3: 3 pontos
- ✓ item 4: 5 pontos

Com a sintaxe e base de dados concluídas, passamos para a etapa seguinte que consistiu em "rodarmos" o programa. Naturalmente, para chegarmos a esse ponto do trabalho passamos por diversas tentativas no sentido de ajustarmos esses dois arquivos. Para tanto, tivemos que recorrer por diversas vezes ao manual do *software* e até mesmo enviarmos e-mail para o autor do programa (i.e., Linacre), que se diga de passagem, foi muito atencioso, e nos atendeu de forma rápida e com informações valiosas.

Os arquivos de saída do *software* são um arquivo output em .txt com as análises estatísticas da modelagem e três arquivos, também no formato .txt, com as medidas pela TRI para as três facetas consideradas: a severidade dos corretores, a proficiência dos alunos e a dificuldade dos itens.

Esses arquivos foram utilizados nas análises dos tópicos a seguir.

6.2.2 Análise crítica do modelo

Após rodarmos o *FACETS*, realizamos três análises críticas no intuito de verificarmos a qualidade do modelo adotado:

- ✓ Alinhamento do modelo com a percepção dos especialistas;
- ✓ Análise de ajuste ao modelo;
- ✓ Análise da dificuldade dos itens.

Descrevemos a seguir, os detalhes de cada uma dessas análises.

6.2.2.1 Alinhamento do modelo com a percepção dos especialistas

A primeira análise crítica, consistiu em verificarmos junto aos especialistas que elaboraram a redação e a Matriz de competências para a produção de texto se os resultados estavam condizentes com suas expectativas.

Para essa análise utilizamos o mapa de variáveis, apresentado na Figura 17, obtido no arquivo de output do *FACETS*, com as representações das três facetas utilizadas no modelo.

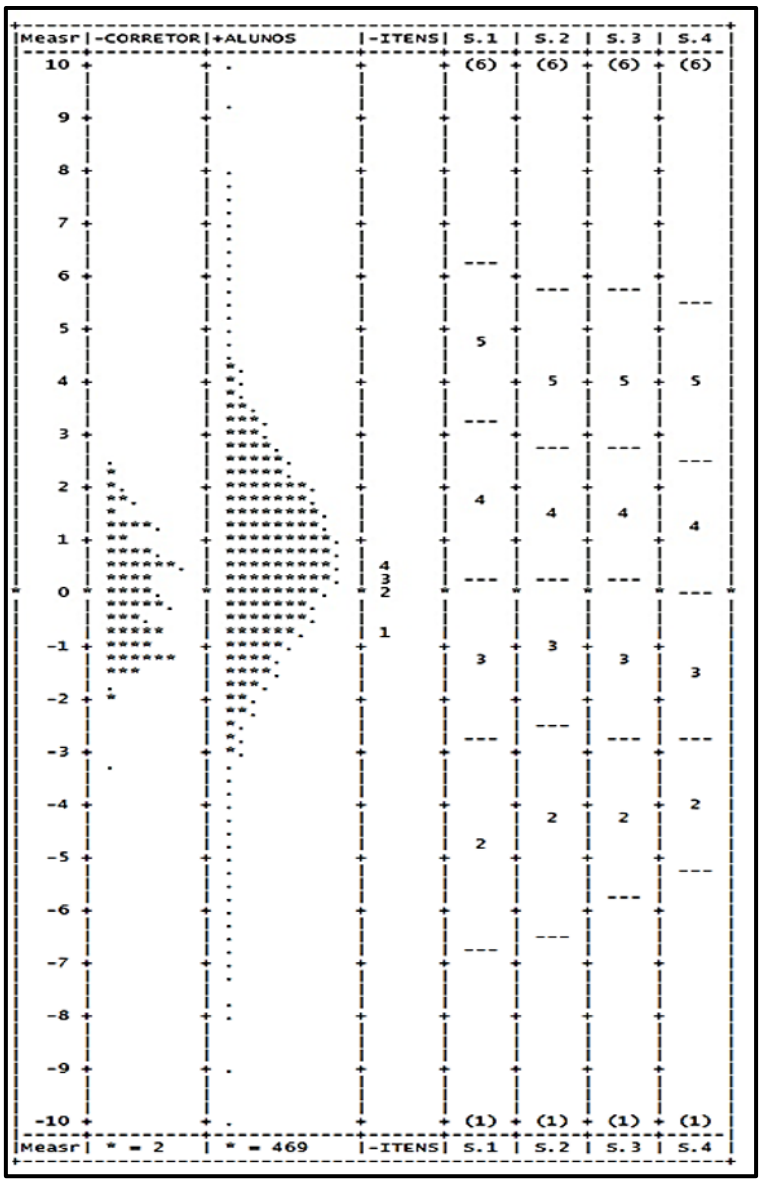


Figura 17 – Mapa de variáveis (*Wright map*) – SABE-2011

Fonte: O Pesquisador através do *software FACETS* 2018.

De acordo com esse mapa de variáveis, para a primeira faceta, observamos a existência de corretores variando de severos a lenientes, na segunda faceta, a distribuição das proficiências dos alunos e a terceira faceta com os diferentes níveis de dificuldade dos quatro itens utilizados, onde o item “1” é o mais fácil, sendo seguido pelos itens “2”, “3” e “4”, esse último, o mais difícil.

As categorias dos itens apresentam dificuldades que estão ordenadas de “1” a “6’ e estão representas nas colunas de S.1 a S.4. Essas ordenações pelos níveis de dificuldades tanto dos itens quanto de suas respectivas categorias foram condizentes com a situação real esperada pelos especialistas que elaboraram a redação e a Matriz de competências para a produção de texto (Anexo 4).

Esse alinhamento das medidas obtidas no modelo com a realidade, no caso, representada pela experiência dos especialistas, nos habilitaram a dar prosseguimento com as análises.

6.2.2.2 Análise de ajuste ao modelo

Realizamos as análises de ajustes nas três facetas utilizadas, conforme procedimentos e critérios exibidos no tópico 4.4.1, cujos resultados são apresentados a seguir.

a) Ajuste dos alunos

Para a faceta aluno, conforme apontado por Nakano & Primi (2014), essa análise é baseada em dois princípios da modelagem Rasch: (i) quanto maior a habilidade do aluno, maior a sua probabilidade de obter altos níveis de acertos nos itens e (ii) itens mais fáceis são mais susceptíveis de serem acertados do que itens difíceis.

Por meio das medidas de *outfit* e *infit* é possível detectar não conformidades com os princípios apresentados acima. No caso da faceta alunos, a estatística *outfit* detecta desajustes em respostas outliers, ou seja, itens muito fáceis ou muito difíceis para o aluno e a estatística *infit* é sensível a desajustes em itens localizados em torno da média do aluno.

Na Tabela 5, apresentamos os quantitativos de alunos por categorias (faixas) de ajuste.

Tabela 5 – Quantitativos em percentuais de alunos por faixa de ajuste

Categorias de ajuste	Valor de *outfit*/*infit*	Alunos			
		Outfit		*Infit*	
		nº de alunos	% de alunos	nº de alunos	% de alunos
1	> 2.0	6.278	8,4	6.156	8,2
2	1.5 a 2.0	6.716	9	6.689	8,9
3	0.5 a 1.5	43.207	57,7	43.497	58,1
4	< 0.5	18.694	25	18.553	24,8
Total		74.895	100	74.895	100

Fonte: O Pesquisador.

A categoria mais crítica para a análise de ajuste é a categoria "1" e, analisando a Tabela 5, podemos constatar um baixo percentual de alunos com resultados não ajustados nessa categoria. Os valores de *outfit* e *infit* são 6.278 (8,4%) e 6.156 (8,2) alunos, respectivamente.

Fizemos uma análise cruzada das categorias de *outfit* pelas categorias de infit, apresentada na Tabela 6, e podemos verificar pela diagonal demarcada na cor cinza, que normalmente os resultados dos alunos se enquadram nas mesmas categorias dos dois indicadores. Assim, temos que dos 6.278 alunos não ajustados por *outfit* e dos 6.156 alunos não ajustados por *infit*, na categoria 1, 6.017 estão não ajustados nos dois indicadores.

Tabela 6 – Análise cruzada de *outfit* por *infit*

Outfit	Quantitativos	*Infit*				Total
		1	2	3	4	1
1	Alunos	6017	253	8	0	6278
	%	95,8	4,0	0,1	0	100
2	Alunos	139	6131	446	0	6716
	%	2,1	91,3	6,6	0	100
3	Alunos	0	305	42445	457	43207
	%	0	0,7	98,2	1,1	100
4	Alunos	0	0	598	18096	18694
	%	0	0	3,2	96,8	100
Total	Alunos	6156	6689	43497	18553	74895
	%	8,2	8,9	58,1	24,8	100

Fonte: O Pesquisador.

Esses baixos quantitativos de alunos com resultados não ajustados em outfit/infit nos habilitam a continuar com nossas análises. Infelizmente, não temos referências nacionais sobre os valores recomendados pra esses indicadores.

b) Ajuste dos corretores

Dos corretores, realizamos dois processos, o primeiro, de nível geral, através do nível de concordância entre corretores (estatísticas interrater), conforme apresentado no terceiro capítulo e o segundo, de forma específica, através das estatísticas *outfit*/*infit* (ver subtópico 4.4.1), para cada corretor.

- Concordância entre corretores (estatísticas *interrate*r)

O modelo estará bem ajustado, quando os corretores executarem os seus trabalhos como corretores independentes, lembrando que a independência é uma das condições para a objetividade da medida. somente sob essa condição, a proficiência do aluno poderá ser modelada de forma independente da severidade do corretor. Essa análise é realizada através da interpretação de dois indicadores fornecidos no final da Tabela 7, do *output* do *FACETS*. Esses indicadores são: a concordância exata observada e a concordância exata esperada. Apresentamos na Figura 18, os valores desses indicadores.

Figura 18 – Valores dos indicadores de concordância entre corretores
Fonte: O Pesquisador através do *software FACETS* 2018.

Temos, portanto:

- ✓ Concordância exata observada = 34,8%
- ✓ Concordância exata esperada = 39,4%

Os valores desses indicadores, nos permitem analisar dois fatores relativos ao processo de correção: (i) se os corretores estão tendo a mesma percepção no que se refere às categorias de respostas dos itens, ou seja, se estão fazendo a mesma interpretação da matriz de competência para a produção de texto (Anexo 2). Esse critério de análise está relacionado ao ajuste do modelo ao processo de correção e, (ii) o nível de concordância entre os corretores. A análise referente ao primeiro fator é obtida na comparação dos valores dos dois indicadores e a segunda, através da magnitude do indicador da concordância exata esperada. Faremos a seguir as análises desses indicadores.

Para o primeiro fator apresentado acima, a concordância exata esperada é ligeiramente superior à concordância exata observada, a própria natureza do processo de correção *on-line* e o grande número de corretores envolvidos no

processo faz com que não haja interação entre os corretores, justificando esse resultado. Entretanto, essa diferença observada é estatisticamente insignificante, e nos permite concluir que o modelo está bem ajustado e que de uma maneira geral, os corretores tiveram a mesma interpretação das categorias de respostas dos itens. Embora no nível individual, retratado com mais detalhes nas análises complementares desse tópico, tivemos corretores com percepções divergentes com relação às categorias de respostas dos itens.

No que se refere ao segundo fator, a concordância entre os corretores é analisada pelo percentual encontrado no indicador de concordância exata esperada. Em nosso caso, o valor de 34,8% é um indicador razoável para esse tipo de avaliação.

Temos que ter em consideração que nos processos de correção, a concordância é vista com bons olhos e incentivada. Mais estudos envolvendo redações são necessários para termos um bom domínio sobre o valor ideal desse indicador, entretanto, vale ressaltar, que em uma situação, hipotética, quando o índice de concordância exata observada se aproximar de 100%, teremos os corretores totalmente alinhados, trabalhando como um único corretor padrão. Nessa situação, não haverá necessidade de coloca-los como uma faceta no modelo, pois os corretores estarão trabalhando como scanners em testes compostos por itens de múltipla escolha.

Para redações corrigidas pela TCT, o perfeito alinhamento dos corretores, ou seja, todos trabalhando com a mesma severidade é uma situação incentivada e vista com bons olhos, entretanto, em avaliações que utilizam a MFRM o fundamental é que os corretores sejam fieis às suas respectivas severidades, e que não as mude durante o processo de correção.

Esse comportamento do corretor durante o processo de correção é analisado através do comportamento diferencial dos corretores (Linacre, 2014). Para a realização dessa análise, teríamos que ter o dia em que as correções foram realizadas, o que infelizmente não temos. Diante desse fato, consideraremos a severidade de cada corretor calculada para todo o processo como um todo.

Na Tabela 7, apresentamos os quantitativos de corretores por categorias de ajuste.

Tabela 7 – Quantitativos de corretores por faixa de ajuste.

Categorias de ajuste	Valor de *Outfit*/*Infit*	Corretores	
		Outfit	*Infit*
1	> 2.0	2	2
2	1.5 a 2.0	4	4
3	0.5 a 1.5	117	117
4	< 0.5	0	0
Total		123	123

Fonte: O Pesquisador.

Apenas dois corretores, representados na base pelos códigos “17” e “69”, apresentaram valores de ajustes na categoria “1”.

c) Ajuste dos itens

A análise do ajuste dos itens foi realizada através das estatísticas de *Outfit*/*Infit* e pela análise da dificuldade dos itens.

Podemos verificar, na Tabela 8, onde apresentamos os quantitativos de itens por categorias de ajuste, que todos os quatro itens ficaram bem ajustados, com seus valores de outfit/infit na categoria “3”.

Tabela 8 – Quantitativos de itens por faixa de ajuste

Categorias de ajuste	Valor de *Outfit*/*Infit*	Corretores	
		Outfit	*Infit*
1	> 2.0	0	0
2	1.5 a 2.0	0	0
3	0.5 a 1.5	4	4
4	< 0.5	0	0
Total		4	4

Fonte: O Pesquisador.

6.2.2.3 Análise da dificuldade dos itens

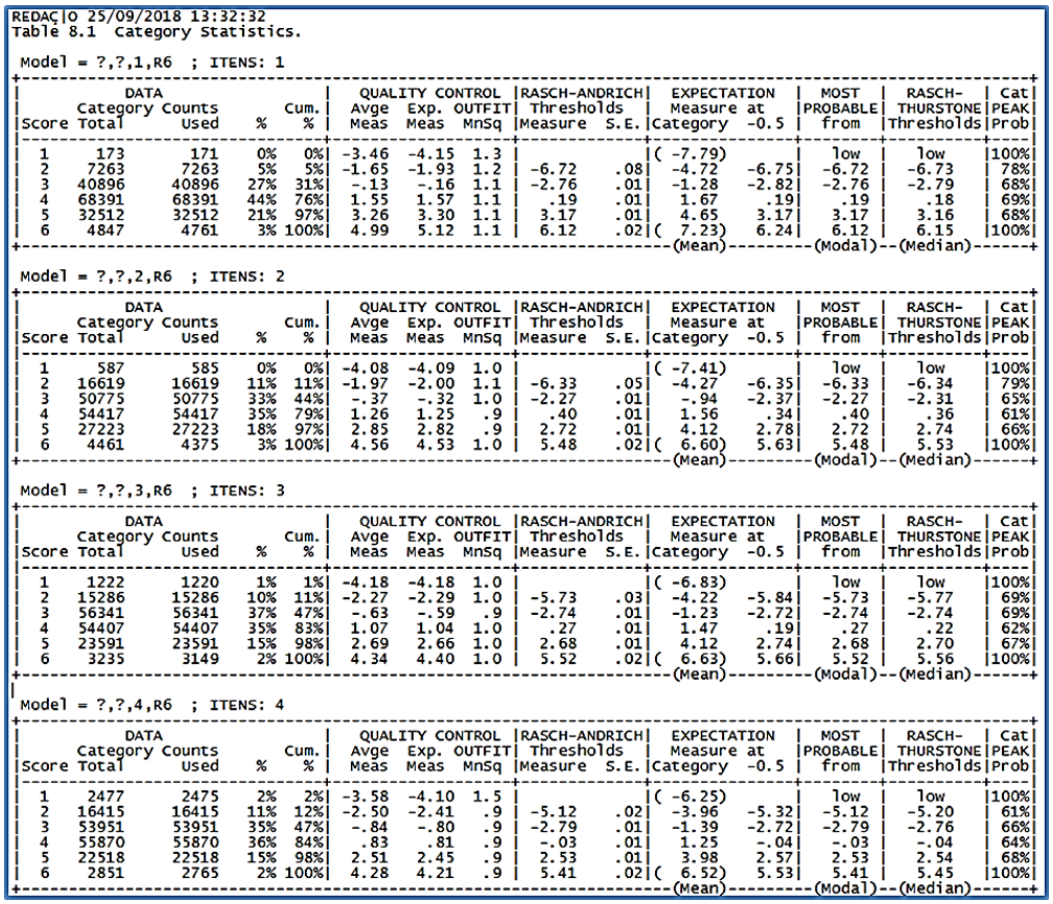
Conforme pode ser observado no mapa de variáveis da Figura 19, os itens, em função de suas dificuldades, apresentam a seguinte ordenação: 1, 2, 3 e 4 (ordem crescente). Apresentamos na Tabela 9 os valores dessas dificuldades.

Tabela 9 – Parâmetro de dificuldade dos itens

Item	Competência	Parâmetro de dificuldade
1	Registro	-0,75
2	Tema	-0,06
3	Tipologia Textual	0,28
4	Coesão/Coerência	0,54

Fonte: O Pesquisador.

Na Figura 19, apresentamos as estatísticas de cada categoria de cada um dos quatro itens, obtidas no *output* do *FACETS*.



REDAÇ|O 25/09/2018 13:32:32
Table 8.1 Category Statistics.

Model = ?,?,1,R6 ; ITENS: 1

Score	DATA Category Total	Counts Used	%	Cum. %	QUALITY CONTROL Avge Meas	Exp. Meas	OUTFIT MnSq	RASCH-ANDRICH Thresholds Measure	S.E.	EXPECTATION Measure at Category	-0.5	MOST PROBABLE from	RASCH-THURSTONE Thresholds	Cat PEAK Prob
1	173	171	0%	0%	-3.46	-4.15	1.3			(-7.79)		low	low	100%
2	7263	7263	5%	5%	-1.65	-1.93	1.2	-6.72	.08	-4.72	-6.75	-6.72	-6.73	78%
3	40896	40896	27%	31%	-.13	-.16	1.1	-2.76	.01	-1.28	-2.82	-2.76	-2.79	68%
4	68391	68391	44%	76%	1.55	1.57	1.1	.19	.01	1.67	.19	.19	.18	69%
5	32512	32512	21%	97%	3.26	3.30	1.1	3.17	.01	4.65	3.17	3.17	3.16	68%
6	4847	4761	3%	100%	4.99	5.12	1.1	6.12	.02	(7.23)	6.24	6.12	6.15	100%

(Mean)--(Modal)--(Median)

Model = ?,?,2,R6 ; ITENS: 2

Score	DATA Category Total	Counts Used	%	Cum. %	QUALITY CONTROL Avge Meas	Exp. Meas	OUTFIT MnSq	RASCH-ANDRICH Thresholds Measure	S.E.	EXPECTATION Measure at Category	-0.5	MOST PROBABLE from	RASCH-THURSTONE Thresholds	Cat PEAK Prob
1	587	585	0%	0%	-4.08	-4.09	1.0			(-7.41)		low	low	100%
2	16619	16619	11%	11%	-1.97	-2.00	1.1	-6.33	.05	-4.27	-6.35	-6.33	-6.34	79%
3	50775	50775	33%	44%	-.37	-.32	1.0	-2.27	.01	-.94	-2.37	-2.27	-2.31	65%
4	54417	54417	35%	79%	1.26	1.25	.9	.40	.01	1.56	.34	.40	.36	61%
5	27223	27223	18%	97%	2.85	2.82	.9	2.72	.01	4.12	2.78	2.72	2.74	66%
6	4461	4375	3%	100%	4.56	4.53	1.0	5.48	.02	(6.60)	5.63	5.48	5.53	100%

(Mean)--(Modal)--(Median)

Model = ?,?,3,R6 ; ITENS: 3

Score	DATA Category Total	Counts Used	%	Cum. %	QUALITY CONTROL Avge Meas	Exp. Meas	OUTFIT MnSq	RASCH-ANDRICH Thresholds Measure	S.E.	EXPECTATION Measure at Category	-0.5	MOST PROBABLE from	RASCH-THURSTONE Thresholds	Cat PEAK Prob
1	1222	1220	1%	1%	-4.18	-4.18	1.0			(-6.83)		low	low	100%
2	15286	15286	10%	11%	-2.27	-2.29	1.0	-5.73	.03	-4.22	-5.84	-5.73	-5.77	69%
3	56341	56341	37%	47%	-.63	-.59	.9	-2.74	.01	-1.23	-2.72	-2.74	-2.74	69%
4	54407	54407	35%	83%	1.07	1.04	1.0	.27	.01	1.47	.19	.27	.22	62%
5	23591	23591	15%	98%	2.69	2.66	1.0	2.68	.01	4.12	2.74	2.68	2.70	67%
6	3235	3149	2%	100%	4.34	4.40	1.0	5.52	.02	(6.63)	5.66	5.52	5.56	100%

(Mean)--(Modal)--(Median)

Model = ?,?,4,R6 ; ITENS: 4

Score	DATA Category Total	Counts Used	%	Cum. %	QUALITY CONTROL Avge Meas	Exp. Meas	OUTFIT MnSq	RASCH-ANDRICH Thresholds Measure	S.E.	EXPECTATION Measure at Category	-0.5	MOST PROBABLE from	RASCH-THURSTONE Thresholds	Cat PEAK Prob
1	2477	2475	2%	2%	-3.58	-4.10	1.5			(-6.25)		low	low	100%
2	16415	16415	11%	12%	-2.50	-2.41	.9	-5.12	.02	-3.96	-5.32	-5.12	-5.20	61%
3	53951	53951	35%	47%	-.84	-.80	.9	-2.79	.01	-1.39	-2.72	-2.79	-2.76	66%
4	55870	55870	36%	84%	.83	.81	.9	-.03	.01	1.25	-.04	-.03	-.04	64%
5	22518	22518	15%	98%	2.51	2.45	.9	2.53	.01	3.98	2.57	2.53	2.54	68%
6	2851	2765	2%	100%	4.28	4.21	.9	5.41	.02	(6.52)	5.53	5.41	5.45	100%

(Mean)--(Modal)--(Median)

Figura 19 – Estatísticas das categorias dos itens

Fonte: O Pesquisador através do *software FACETS* 2018.

No Anexo 5 apresentamos as curvas características dos quatro itens utilizados na redação.

6.3 Análise comparativa nota (TCT) x proficiência em redação (MFRM)

Uma vez concluída a análise crítica da base de dados e do modelo, passamos para as análises exploratórias envolvendo simulações, onde confrontamos as características das duas metodologias de produção de medidas de desempenho dos alunos na base considerada de forma a verificarmos nossa hipótese de trabalho:

- ✓ Nota: metodologia tradicional atualmente adotada, via TCT;
- ✓ Proficiência: metodologia alternativa proposta, via TRI com multifacetas.

Os passos envolvidos pra essa análise foram:

- 1º passo: obtenção dos arquivos com as medidas das facetas;
- 2º passo: classificação dos corretores em função da severidade;
- 3º passo: montagem no SPSS dos arquivos para trabalho;
- 4º passo: análise comparativa nota (TCT) x proficiência em redação (MFRM).

- **1º passo: obtenção dos arquivos com as medidas das facetas**

Através da inserção na sintaxe, do comando "Score file", obtivemos três arquivos:

- ✓ Severidade dos corretores;
- ✓ Proficiência dos alunos;
- ✓ Dificuldade dos itens.

Esses arquivos, originalmente fornecido pelo *software* no formato .txt, foram trabalhados no SPSS de forma a viabilizar as análises desse estudo.

- **2º passo: classificação dos corretores em função da severidade**

De posse da severidade de cada corretor, fizemos uma ordenação crescente dessa severidade (padronizada com média = 0 e dp =1) e dividimos os corretores em três grupos: lenientes (L), moderados (M) e severos (S), adotando como critério de divisão os tercis da severidade dos corretores. No Anexo 4, apresentamos os valores de severidade dos corretores e a respectiva categorização. Esse arquivo está ordenado pela severidade (ordem crescente).

Apresentamos na Tabela 10 os quantitativos de corretores em cada grupo e os respectivos valores de severidade para cada grupo.

Tabela 10 – Classificação dos corretores

FX_Severidade	N	Severidade		
		Mínimo	Máximo	Média
L	41	-2,927	-0,571	-1,093
M	41	-0,553	0,460	-0,031
S	41	0,469	2,338	1,124
Total	123	-2,927	2,338	0,000

Fonte: O Pesquisador.

No Gráfico 8, apresentamos as severidades padronizadas (0,1) de todos os 123 corretores, podemos observar que o corretor mais severo é o corretor "9", e o mais leniente é o corretor "70".

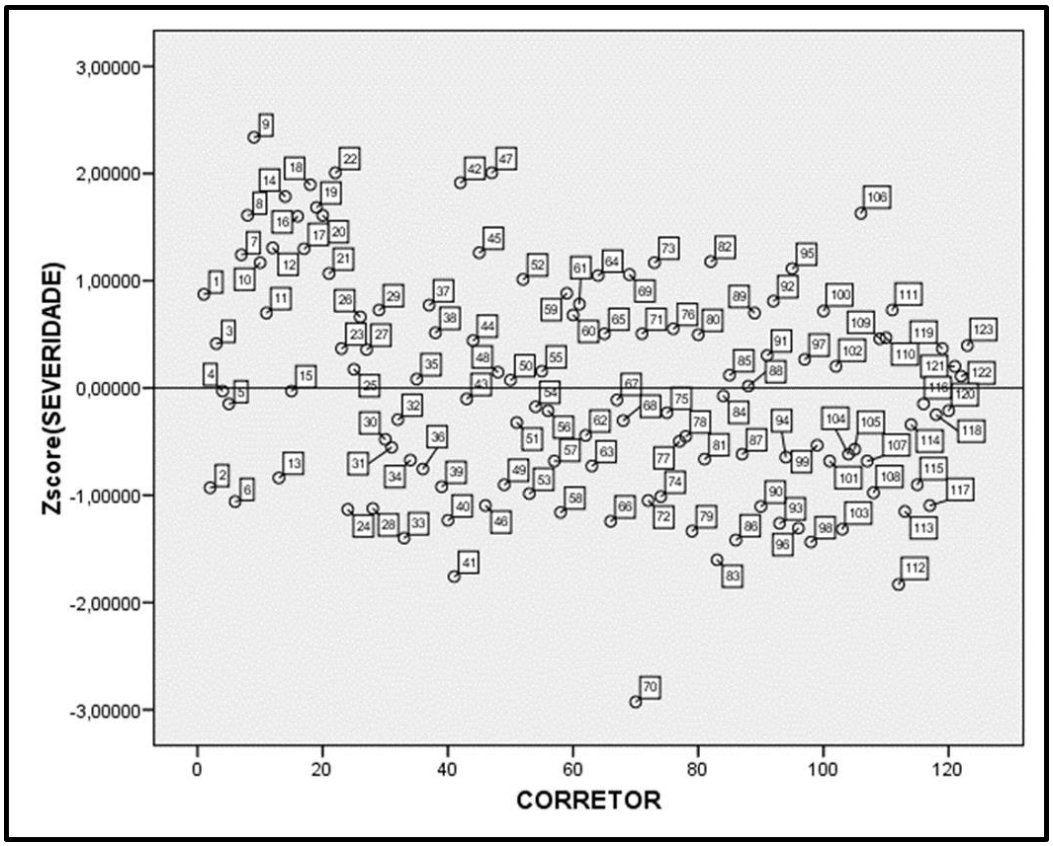


Gráfico 8 – Severidade dos corretores em uma escala padronizada

Fonte: O pesquisador.

- **3º passo: construção no SPSS dos arquivos para trabalho**

Construímos no SPSS, um arquivo com as medidas de nota que cada corretor deu a cada aluno, a proficiência de cada aluno e uma identificação do tipo de dupla de corretores a que o aluno foi submetido, a saber:

- ✓ SS – dois corretores severos;

- ✓ LL – dois corretores lenientes;
- ✓ MM – dois corretores medianos;
- ✓ SL – um corretor severo e um leniente;
- ✓ SM – um corretor severo e um mediano;
- ✓ LM – um corretor leniente e um mediano.

Não retiramos dessa análise os corretores e alunos com indicação de desajuste na análise anterior, por considerarmos o número reduzido de casos problemáticos, o que não é empecilho para as análises desse estudo.

Apresentamos na Tabela 11 a estrutura desse arquivo.

Tabela 11 – Nota e proficiência por aluno em função da dupla de corretores.

Aluno	Corretor			Nota			ZNota	FX_Severidade			Dupla de corretores	ZPRF
	1	2	3	1	2	3		1	2	3		
1	17	49	.	1,5	3,5	.	-2,06178	S	L	.	LS	-2,20739
2	53	110	.	7,5	5	.	0,62529	L	S	.	LS	0,49676
3	27	34	.	4	6	.	-0,2704	M	L	.	LM	-0,37645
4	26	35	.	3	2,5	.	-1,88264	S	M	.	MS	-1,84683
5	25	36	.	4,5	6,5	.	0,08788	M	L	.	LM	-0,07787
6	71	103	.	5,5	5,5	.	0,08788	S	L	.	LS	-0,15111
7	25	36	.	7,5	7,5	.	1,52098	M	L	.	LM	1,42068
8	26	35	-	5,5	6	.	0,26701	S	M	.	MS	0,51366
9	71	103	4	3,5	8	5,5	0,08788	S	L	M	LS	0,0517
10	27	34	.	2	5	.	-1,34523	M	L	.	LM	-1,56515
74894	71	103	4	4,5	9	9	0,44615	S	L	M	LS	0,6714
74895	58	71	.	6	3	.	-0,62868	L	S	.	LS	-0,87221

Fonte: O Pesquisador.

Nessa Tabela, temos a seguinte representação para cada coluna:

- ✓ Coluna 1: identificação do aluno: de 1 a 74895;
- ✓ Colunas de 2 a 4: número de identificação dos corretores (de 1 a 123), na 1ª, 2ª e 3ª correção de cada aluno;
- ✓ Colunas de 5 a 7: nota dada por cada corretor a cada aluno;
- ✓ Coluna 8: variável ZNota – nota padronizada utilizando uma normal (0,1);
- ✓ Colunas de 9 a 11: classificação dos corretores em severos, moderados e lenientes;
- ✓ Coluna 12: dupla de corretores – dupla formada pela 1ª e 2ª correções;
- ✓ Coluna 13: proficiência do aluno padronizada em uma normal (0,1).

Temos por meio desse quadro que, por exemplo, o aluno com número de identificação "1", teve sua redação corrigida por dois corretores: a primeira correção, realizada pelo corretor "17" e a segunda correção pelo corretor "49"; as notas desse aluno foram "1,5" e "3,5" em cada uma das correções; a média padronizada dessas duas notas foi "-2,06178"; o primeiro corretor foi classificado como severo e o segundo como leniente, formando uma dupla "LS" e a proficiência padronizada obtida no *software FACETS* foi de "-2,20739".

- **4º passo: análise comparativa nota (TCT) x proficiência (MFRM) com dupla e terceira correção em 100% da base**

Para verificarmos qual das duas metodologias de produção de medidas, nota (forma tradicional atualmente adotada) ou proficiência via MFRM (forma alternativa) apresentam resultados mais próximos do desempenho real dos alunos, realizamos as análises de fidedignidade e validade das medidas produzidas por essas duas metodologias.

Como em psicometria não se conhece o valor "verdadeiro" do constructo medido por meio de um teste, impossibilitando a comparação dos valores medidos com um padrão de modo a se verificar a qualidade da medida, conforme apresentado no primeiro capítulo, realizamos análises comparativas entre as medidas em diferentes agregados construídos pelas características de severidade dos corretores conforme apresentado no passo anterior. O argumento balizador das análises de fidedignidade e validade realizados a seguir, se baseia no fato de que: Se os valores de severidade, calculados pelo *software FACETS* modelam bem essa característica, (severidade), dos corretores, é de se esperar valores de proficiência muito próximos entre si, entre esses agregados. Ou seja, as médias das proficiências entre esses agregados são próximas entre si (resultados válidos) e, a correlação entre proficiência e nota dentro de cada agregado dever ser superior ao observado na base como um todo (resultados fidedignos). Caso essa modelagem da severidade não seja realizada de forma adequada, teremos resultados semelhantes ao encontrado nas comparações entre TCT e TRI (via modelagem Rasch) duas facetas, não justificando a utilização da MFRM no cálculo de proficiências em redações.

A seguir, detalhamos as análises de fidedignidade e validade do modelo proposto.

6.3.1 Análise da fidedignidade

Conforme Pasquali (1996), a análise de fidedignidade se refere a quanto os escores de um sujeito se matem idênticos em ocasiões diferentes. Temos no Avalie BA um mesmo aluno medido por dois corretores distintos com o objetivo de tornar a medida mais uniforme, ou seja, mais fidedigna. Para a análise da fidedignidade, utilizamos a correlação de *Pearson* em quatro variáveis que representam as medições dos mesmos alunos segundo duas metodologias diferentes. Essas variáveis foram padronizadas em uma normal (0,1) e são as seguintes, a fim de se facilitar a comparação entre as medidas nas diferentes modelagens:

- ✓ ZNota_1 - medida da primeira correção do desempenho do aluno pela TCT.
- ✓ ZNota_2 - medida da segunda correção do desempenho do aluno pela TCT.
- ✓ ZPRF_1 - medida da primeira correção do desempenho dos alunos pela TRI, através da modelagem MFRM (alunos, itens e corretores).
- ✓ ZPRF_2 - medida da segunda correção do desempenho dos alunos pela TRI, através da modelagem MFRM (alunos, itens e corretores).

Apresentamos na Tabela 12 os valores dos coeficientes das correlações de Pearson, obtidos:

Tabela 12 – Correlações de *Pearson* entre notas (ZNota_1 e ZNota_2) e entre proficiências (ZPRF_1 e ZPRF_2)

Categorias	Dupla	*Pearson* Correlation Sig. (2-tailed)	ZNota_1 com ZNota_2	ZPRF_1 com ZPRF_2
1	Todos	Correlação	0,396	0,494
		N	74895	74895
2	LL	Correlação	0,464	0,476
		N	9074	9074
3	LM	Correlação	0,437	0,508
		N	12781	12781
4	LS	Correlação	0,123	0,470
		N	18775	18775
5	MM	Correlação	0,518	0,535
		N	8993	8993
6	MS	Correlação	0,406	0,513
		N	17080	17080
7	SS	Correlação	0,454	0,461
		N	8192	8192

Fonte: O Pesquisador.

A categoria “1” representa a base completa com todas as duplas de corretores e as demais categorias de “2” a “7”, representam agrupamentos em função da severidade da dupla de corretores a que o aluno foi submetido.

Na categoria 4, que representa as redações corrigidas por duplas LS, temos o menor valor de correlação das notas, onde a correlação das notas é de 0,123 e a correlação das proficiência é de 0,470; e na categoria 3, que representa as redações corrigidas por duplas MM, temos o maior valor de correlação das notas, onde a correlação das notas é de 0,518 e das proficiências é de 0,535.

A correlação nas sete categorias é sempre maior entre as medidas de proficiência do que entre as medidas de notas, e que as correlações entre proficiências entre as 7 categorias são mais próximas entre si do que as correlações entre notas, ou seja, além de serem mais fidedignas as medidas de proficiência apresentam resultados mais uniformes entre as categorias.

6.3.2 Análise da validade

Ao incluirmos a faceta severidade do modelo MFRM estamos compensando a proficiência do aluno em função da severidade do corretor. Como a severidade está na mesma métrica da proficiência, os alunos que têm suas redações corrigidas por corretores severos recebem um acréscimo na proficiência, e alunos que têm suas redações corrigidas por corretores lenientes recebem um decréscimo em suas proficiências. Esses valores de acréscimos e decréscimos são proporcionais à severidade dos corretores.

Apresentamos no gráfico 9 a correlação entre as notas médias (Znota) e as proficiências médias (ZPRF_3F) dos alunos corrigidos pelas duplas de corretores lenientes e severos. Como o *software FACETS* compensa a severidade do corretor na medida de proficiência do aluno, para os mesmos valores de notas temos diferentes valores de proficiência em função da dupla de corretores: se a dupla é severa a medida de proficiência é maior e se a dupla é leniente, a medida de proficiência é menor.

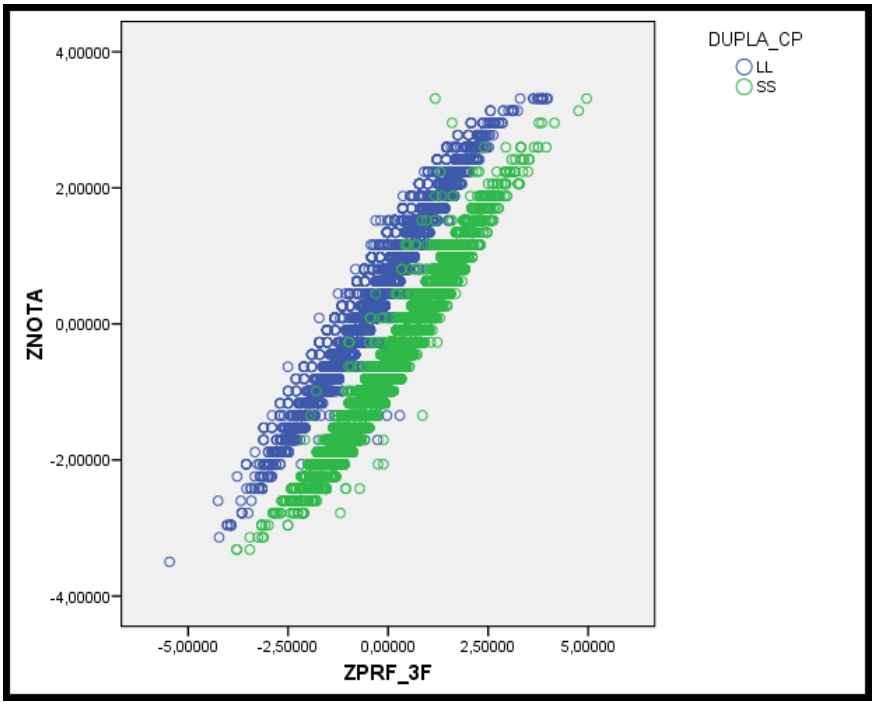


Gráfico 9 – Correlação ente nota média e proficiência média para as dupla de corretores severos e lenientes

Fonte: O Pesquisador.

É de se esperar que esses ajustes feitos pelo *FACETS* tornem os resultados mais válidos. Porém, devido ao fato de não termos o valor verdadeiro do desempenho dos alunos para adotarmos como referência na análise da validade dos dois métodos utilizados, realizamos o seguinte artifício: ao identificarmos os alunos que tiveram suas redações corrigidas por duplas de corretores lenientes e alunos que tiveram suas redações corrigidas por duplas de corretores severos é de se esperar, pela quantidade de alunos existentes nesses dois grupos, 9074 e 8192, respectivamente, que as medidas de desempenho entre esses dois grupos, tanto pela nota média quanto pela proficiência média, sejam bem próximas no que se refere às médias e desvios padrões, admitindo apenas variações amostrais. Assim, o método que apresentar os resultados mais próximos entre esses dois grupos, será o que representa melhor a realidade do constructo medido.

Apresentamos na Tabela 13 as estatísticas descritivas para as duplas de corretores identificadas, assim como as pontuações dos alunos nas variáveis nota e proficiência, respectivamente.

Tabela 13 – Nota e proficiências dos alunos em função da dupla de corretores classificadas em função da severidade.

Dupla de corretores	Estatísticas	ZNota	ZPRF_3F
LL	Média	0,564	-0,107
	Alunos	9074	9074
	dp	1,014	1,083
LM	Média	0,332	0,012
	Alunos	12781	12781
	dp	0,976	1,023
LS	Média	0,017	0,039
	Alunos	18775	18775
	dp	0,943	0,982
MM	Média	-0,049	-0,062
	Alunos	8993	8993
	dp	0,900	0,968
MS	Média	-0,286	-0,002
	Alunos	17080	17080
	dp	0,923	0,981
SS	Média	-0,531	0,081
	Alunos	8192	8192
	dp	0,915	0,966
Total	Média	0,000	0,000
	Alunos	74895	74895
	dp	1,000	1,000

Fonte: O Pesquisador.

Conforme explicado anteriormente, ao dividirmos a população pelas duplas de corretores, é de se esperar que o desempenho dos alunos entre as duplas seja o mesmo, tendo em vista o grande número de estudantes nesses agrupamentos, ou seja, é como se os estudantes fossem divididos de forma aleatória entre as duplas. Podemos assim considerar que qualquer diferença de desempenho entre as duplas deve-se exclusivamente aos diferentes níveis de severidade dos corretores.

Para atingirmos nosso objetivo de verificar a validade das medidas de desempenho, verificamos o desvio, através das diferenças tanto para nota como proficiência, composta pelos alunos corrigidos pelas duplas SS e LL, que representam as situações mais extremas de severidade dos corretores. Assim, os desvios tiveram as seguintes formulações:

- Desvio_nota = ZNota (LL) – ZNota (SS)
- Desvio_proficiência = ZPRF_3F (LL) – ZPRF_3F (SS)

Em que,

✓ Nota (LL) – nota dos alunos corrigidos por duplas de corretores lenientes;

✓ Nota (SS) – nota dos alunos corrigidos por duplas de corretores Severos;

✓ ZPRF_3F (LL) – proficiência dos alunos corrigidos por duplas de corretores lenientes;

✓ ZPRF_3F (SS) – proficiência dos alunos corrigidos por duplas de corretores severos.

Observações:

✓ Cumpre ressaltar que temos os mesmos alunos nos cálculos dos dois indicadores.

A seguir, apresentamos os valores calculados para esses desvios:

✓ Desvio_nota = 0,564 – (-0,531) = 1,095

✓ Desvio_proficiência = -0,107-(0,081) = -0,188

Comparando os resultados dos dois desvios, podemos verificar uma diferença de 1,095 do desvio padrão, quando a correção é realizada pela nota e uma diferença de 0,188 do desvio-padrão ao se considerar os valores de proficiência obtidos pela MFRM.

Na Figura 20, apresentamos uma comparação dos desvios para as duas variáveis consideradas.

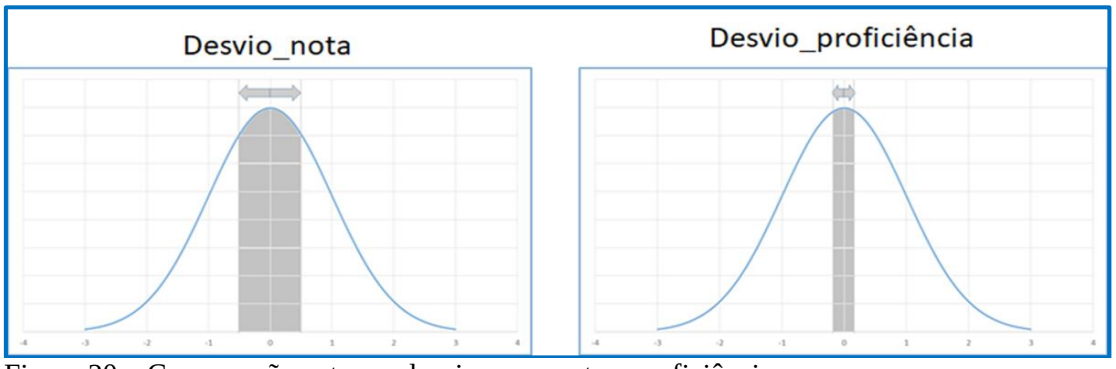


Figura 20 – Comparação entre os desvios para nota e proficiência

Fonte. O Pesquisador.

Podemos concluir que pela modelagem MFRM temos uma medida mais exata do desempenho dos alunos.

6.4 Análises complementares

Apresentamos nesse tópico, quatro análises com diferentes propósitos de utilização da MFRM no sentido de obter maior domínio das possibilidades dessa metodologia. Essas análises são:

- Análise crítica do serviço realizado pelos supervisores;
- Dupla correção com nota (100% da base) x dupla correção com proficiência (em amostras);
- Percepção dos corretores da matriz de competência para a produção de texto;
- Rankeamento dos alunos com utilização da nota e da proficiência.

6.4.1 Análise crítica do serviço realizado pelos supervisores

A primeira análise complementar consistiu em fazermos uma análise crítica dos perfis dos supervisores e o quanto suas severidades poderiam influenciar nas notas atribuídas aos alunos, tendo em vista que no caso de discrepância entre as duas correções, as duas notas são desconsideradas e prevalece a nota do supervisor. Temos, portanto, a figura do supervisor como uma referência inquestionável. De acordo com esse procedimento, seria de se esperar que todos os supervisores apresentassem o mesmo nível de severidade e que essa severidade estivesse localizada no centro da distribuição das severidades de todos os corretores. Qualquer divergência nessas duas situações, a realização da terceira correção trará prejuízos para a pontuação dos alunos.

A 3ª correção, utilizada em casos de discrepância entre dois corretores, foi realizada em 4292 casos, ou seja, 5,7% da base. Podemos verificar na Tabela 14, que os supervisores, responsáveis pela nota definitiva do aluno, possuem níveis diferentes de severidade, o que indica que esse processo pode prejudicar ou beneficiar o aluno, caso sua redação tenha sido corrigida por um corretor severo ou leniente, respectivamente.

Tabela 14 – Classificação da severidade dos supervisores

Supervisor	Nível severidade supervisor			Total
	L	M	S	
1	0	0	594	594
2	275	0	0	275
3	0	621	0	621
4	0	1254	0	1254
5	0	969	0	969
6	579	0	0	579
Total	854	2844	594	4292

Fonte: O Pesquisador.

Podemos verificar que os supervisores “2” e “6” são lenientes, o supervisor “1” é severo e os supervisores “3”, “4” e “5” são medianos. Dessa forma, 854 alunos foram beneficiados e 594 foram prejudicados, e 2844 tiveram suas notas mais próximas da realidade, sem o viés da medida relacionado à severidade.

Por meio da utilização da proficiência obtida pela MFRM a terceira correção deixa de ter utilidade, uma vez que essa técnica ajusta a proficiência do aluno em função da severidade do corretor e esse ajuste produz medidas mais exatas do que a dupla e terceira correções por meio da nota, conforme resultados apresentados na Tabela 13 e também pela não homogeneidade dos supervisores (Tabela 14).

6.4.2 Dupla correção com nota (100% da base) x dupla correção com proficiência (em amostras)

Para essa análise, comparamos a nota obtida com dupla correção em 100% da base com as proficiências obtidas através da MFRM em amostras aleatórias da base formada por alunos com dupla correção e alunos com apenas uma correção, conforme Tabela 15, apresentada a seguir:

Tabela 15 – Simulações com diferentes percentuais de alunos com dupla e uma correção

Simulação	% de alunos com dupla correção	% de alunos com uma correção	Número médio de correções por corretor
1	100	0	130
2	50	50	260
3	40	60	388
4	30	70	518
5	20	80	651
6	10	90	1306

Fonte: O Pesquisador.

A base utilizada foi composta apenas por alunos que tiveram dupla correção, sendo retirados os casos de três correções. Portanto, os supervisores não fizeram parte desse estudo. Também foram retirados da base quatro corretores (70, 95, 97 e 117), por possuírem poucas correções, o que provocaria um desequilíbrio no número médio de correções por corretor. A base final foi formada com 73.171 alunos

O objetivo dessa análise foi verificar as variações das severidades e proficiências em função do número de duplas correções adotadas, no intuito de verificar a qualidade das medidas de proficiência em função da quantidade de alunos com dupla correção em amostras da base total.

Após a construção da base, rodamos o *FACETS* para as seis simulações apresentadas na Tabela 15 e realizamos os mesmos procedimentos de análise apresentados nos tópicos anteriores, através do cálculo dos indicadores “1” e “2”.

Com relação à classificação dos corretores em lenientes, medianos e severos, as variações observadas nessa classificação em relação às amostras utilizadas, foram muito baixas. Ao comparamos as classificações entre as simulações “1” e “6”, verificamos que dos 112 corretores analisados, 99 permaneceram com a mesma classificação e apenas 13 mudaram de categoria. Dessa forma utilizamos para as análises seguintes a classificação obtida na simulação “6’ como referência para todas as simulações.

6.4.2.1 Análise da fidedignidade

Como constatamos no tópico 6.3.1 que a fidedignidade das medidas utilizando a proficiência é maior que a fidedignidade utilizando-se a nota, para analisarmos a fidedignidade nas diferentes simulações deste tópico, realizamos a correlação de *Pearson* entre as variáveis representativas das proficiências nas diferentes simulações, adotando a simulação 1 como referência. Apresentamos na tabela 16, os valores obtidos.

Tabela 16 – Correlação entre a proficiência na simulação 1 com as demais simulações

Duplas		Correlação de *Pearson* Sig. (2-tailed)	ZPRF100 com ZPRF50	ZPRF100 com ZPRF40	ZPRF100 com ZPRF30	ZPRF100 com ZPRF20	ZPRF100 com ZPRF10
D1	Todos	Correlação	0,923**	0,908**	0,894**	0,881**	0,864**
		Alunos	73171	73171	73171	73171	73171
D2	LL	Correlação	0,915**	0,898**	0,886**	0,869**	0,846**
		Alunos	8141	8141	8141	8141	8141
D3	MM	Correlação	0,928**	0,914**	0,896**	0,886**	0,866**
		Alunos	8249	8249	8249	8249	8249
D4	SS	Correlação	0,915**	0,901**	0,883**	0,870**	0,854**
		Alunos	6592	6592	6592	6592	6592
D5	LM	Correlação	0,926**	0,912**	0,900**	0,887**	0,870**
		Alunos	14058	14058	14058	14058	14058
D6	SM	Correlação	0,926**	0,912**	0,900**	0,888**	0,872**
		Alunos	19442	19442	19442	19442	19442
D7	LS	Correlação	0,920**	0,906**	0,891**	0,877**	0,865**
		Alunos	16689	16689	16689	16689	16689

**Correlation is significant at the 0.01 level (2-tailed)

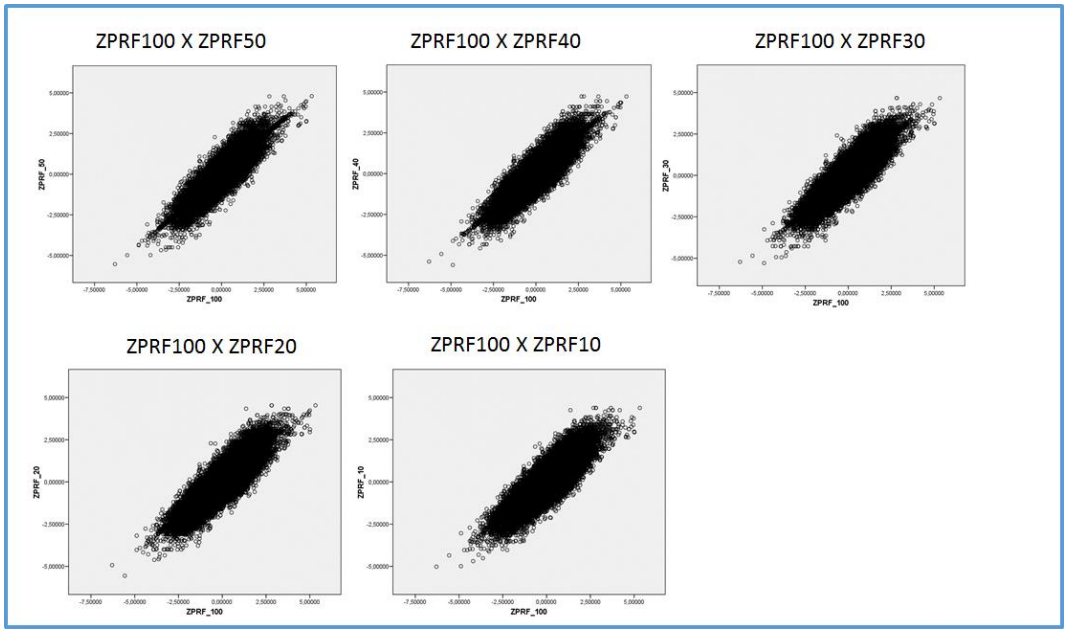
Fonte: O Pesquisador.

A primeira linha (D1), representa a correlação entre as proficiências, sem levar em consideração as características de severidade dos corretores. As linhas de D2 a D7, são as correlações levando-se em consideração as severidades dos corretores.

Podemos observar que a correlação entre as proficiências diminui à medida que diminuímos o número de alunos com duplas correções utilizados para a estimação dos parâmetros das facetas utilizadas nas modelagens.

No entanto, a maior perda de fidedignidade ocorre ao utilizarmos a base com 50% de duplas correções, a partir desse ponto, a diminuição de duplas correções na base provoca pouca perda na fidedignidade conforme pode ser verificado na figura 21.

Figura 21 – Correlação entre proficiências:100% dos alunos com dupla correção por 50%, 40%, 30%, 20% e 10% dos alunos com dupla correção



Fonte: O Pesquisador.

6.4.2.2 Análise da validade

Para a análise do efeito da redução do número de duplas correções na validade das medias, consideramos as mesmas seis simulações apresentadas anteriormente, e calculamos a diferença entre as medidas obtidas para a nota e proficiência entre as duplas de corretores lenientes e severos. Devido à quantidade de alunos dentro das duplas leniente e severos, é de se esperar um desempenho

muito próximo entre as mesmas. Apresentamos na Tabela 17 as medidas obtidas nas simulações realizadas.

Tabela 17 – Nota e proficiências dos alunos, por amostra, em função da severidade da dupla de corretores.

Duplas		Estatística	Simulações S1		S2	S3	S4	S5	S6
			ZNota100	ZPRF100	ZPRF50	ZPRF40	ZPRF30	ZPRF20	ZPRF10
1	Todos	Média	0	0	0	0	0	0	0
		dp	1	1	1	1	1	1	1
		Alunos	73171	73171	73171	73171	73171	73171	73171
2	LL	Média	0,574	-0,112	-0,118	-0,133	-0,081	-0,101	-0,093
		dp	1,020	1,079	1,069	1,072	1,061	1,066	1,054
		Alunos	8141	8141	8141	8141	8141	8141	8141
3	MM	Média	-0,016	-0,045	-0,039	-0,041	-0,073	-0,038	-0,003
		dp	0,926	0,980	0,950	0,939	0,931	0,928	0,916
		Alunos	8249	8249	8249	8249	8249	8249	8249
4	SS	Média	-0,587	0,073	0,065	0,075	0,070	0,057	0,101
		dp	0,880	0,929	0,934	0,933	0,929	0,928	0,935
		Alunos	6592	6592	6592	6592	6592	6592	6592
5	LM	Média	0,303	-0,021	-0,021	-0,031	-0,035	-0,037	-0,063
		dp	1,001	1,038	1,042	1,042	1,048	1,050	1,055
		Alunos	14058	14058	14058	14058	14058	14058	14058
6	SM	Média	-0,267	0,026	0,026	0,034	0,014	0,018	0,028
		dp	0,922	0,980	0,980	0,979	0,978	0,979	0,975
		Alunos	19442	19442	19442	19442	19442	19442	19442
7	LS	Média	0,015	0,036	0,038	0,041	0,061	0,055	0,028
		dp	0,925	0,981	0,996	0,999	1,007	1,004	1,012
		Alunos	16689	16689	16689	16689	16689	16689	16689
DIF = LL - SS			1,161	-0,185	-0,183	-0,207	-0,151	-0,157	-0,194

Fonte: O Pesquisador.

As variáveis, nota e proficiências foram padronizadas em uma normal (0,1). As proficiências foram calculadas para as seis simulações e a nota apenas para a simulação "1", que corresponde à melhor situação para a obtenção da medida. Nossa intenção é comparar as medidas de proficiências nas seis simulações com diferentes percentuais de dupla correção com a nota na situação mais favorável (100% de dupla correção).

Verificamos que a estimação das proficiências nas seis simulações, produzem medidas mais exatas do que a utilização da nota com dupla correção em 100% da base. Na última linha, podemos constatar os valores das diferenças das medidas entre as duplas de corretores lenientes e severos. A correção pela nota com

dupla correção apresentou uma diferença de 1,161 desvios padrões, e a pior estimativa pela MFRM, apresentou uma diferença de 0,207 (amostra com 40% com dupla correção).

As diferenças de proficiências entre as duplas de corretores Severos e Lenientes, parece não sofrer queda com a redução do número de correções utilizadas para a calibração da severidade, os valores oscilaram entre 0,151 e 0,207.

Diante dos resultados obtidos nas análises de fidedignidade e validade, podemos concluir que a utilização de duplas correções em amostras da base total é uma alternativa viável e de melhor qualidade ao comparada com produção de medidas pela nota com dupla correção em toda a base.

Essa alternativa de se utilizar a proficiência e amostras com dupla correção, deve, entretanto, ser utilizada com cautela, não sendo indicada para projetos em que haja seleção de candidatos. Para esse caso, o mais indicado seria a utilização da proficiência via MFRM e dupla correção em 100% da base. Para projetos que não tenham a seleção de candidatos, sendo os resultados dos alunos utilizados para interpretações pedagógicas e cálculos dos níveis de desempenho das escolas, municípios e Estados, vemos que essa alternativa é plenamente viável, principalmente em função da redução dos custos advindos do menor número de correções.

6.4.3 Percepção dos corretores da matriz de competência para a produção de texto

Na terceira análise complementar, constamos algumas divergências entre os corretores com relação à classificação dos alunos nas categorias dos itens. Supondo não haver problema nos critérios de correção utilizados, era de se esperar que todos os quatro itens fossem percebidos em suas seis categorias por todos os corretores, entretanto, conforme apresentado na Tabela 18, isso não acontece. Podemos verificar um grande número de corretores percebendo cinco, quatro e três categorias nos itens, quanto que o esperado seriam seis categorias por item.

Apresentamos na Tabela 18, para cada um dos quatro itens do teste de redação o quantitativo de corretores que observaram de seis a três categorias em todos os itens por eles corrigidos.

Tabela 18 – Número de corretores por categorias x itens

Item	Categorias				
	1 a 6	2 a 6	1 a 5	2 a 5	3 a 5
1	37	66	5	15	
2	68	42	6	6	
3	73	32	10	7	1
4	77	30	9	6	1

Fonte: O Pesquisador.

Temos que para o item "1", primeira linha da Tabela, 37 corretores classificaram alunos nas seis categorias do item, 66 corretores identificaram alunos nas categorias de "2" a "6" (não classificaram nenhum aluno na primeira categoria), cinco corretores não identificaram nenhum aluno na última categoria e 15 corretores não identificaram nenhum aluno na primeira e última categoria do item. Pela grande quantidade de alunos corrigidos pelos corretores, era de se esperar um comportamento em todos os itens mais parecido ao observado nos itens "2", "3" e "4", com a maioria dos corretores observando mais alunos nas seis categorias dos itens.

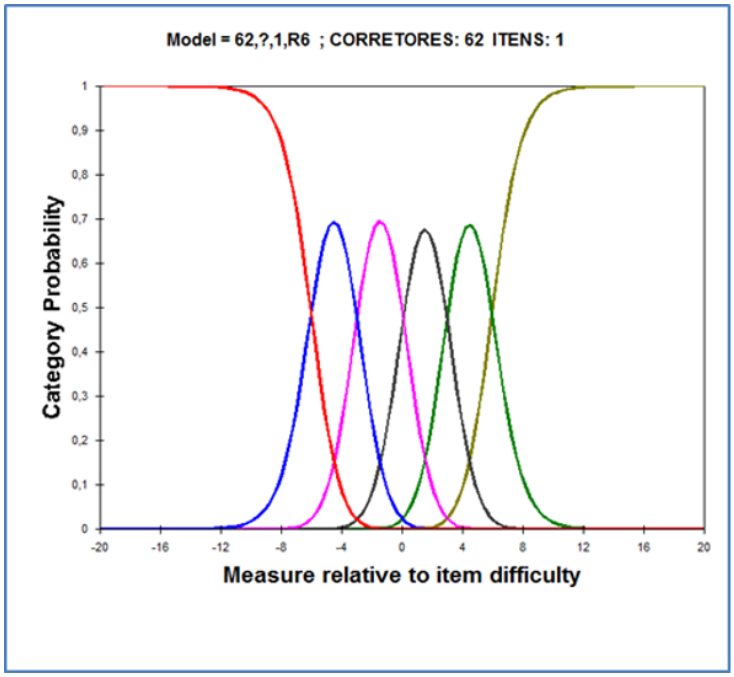
Especificamente para o item "1", tivemos uma incidência muito alta de corretores identificando cinco categorias, não identificado a 1ª categoria, ou seja, a maioria dos corretores num total de 66, que corresponde a 54% de toda a base. Essa alta incidência pode estar relacionada com a característica da matriz de competências para a produção de textos utilizada (Anexo 4), sugerindo uma análise do número de categorias adotadas para esse item, onde parece ser mais factível, utilizar cinco categorias para esse item, juntado em uma única categoria, desde que viável pedagogicamente, as características das categorias "1" e "2".

Já os corretores que identificaram três ou quatro categorias, como por exemplo os corretores 56 e 108 que classificaram todos os quatro itens com três ou quatro categorias, é uma identificação clara que se trata de corretores que não se ajustaram ao padrão de correção.

Nossa intenção ao apresentar essas situações, é mostrar as potencialidades da utilização da MFRM em construir instrumentos mais ajustados com a realidade e ter um maior controle da qualidade dos trabalhos realizados pelos corretores.

Apresentamos a seguir, nos Gráficos 10, 11, 12 e 13 alguns exemplos das situações encontradas. Nesses gráficos, apresentamos o número de categorias

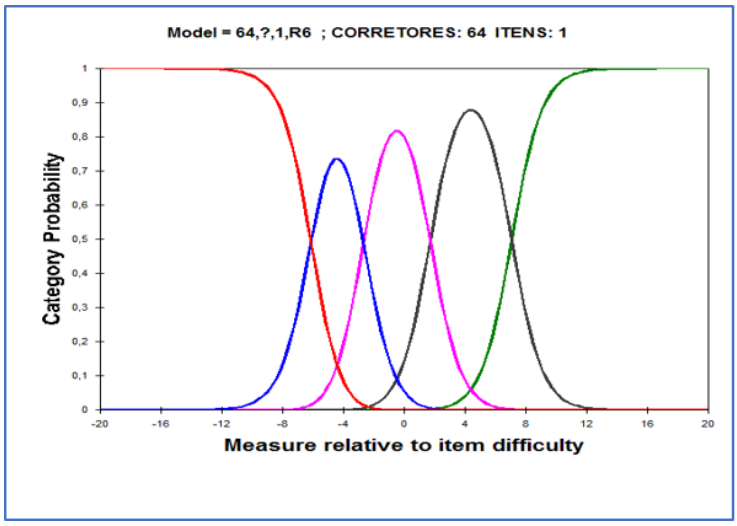
observadas para um mesmo item referente a todos os alunos corrigidos pelo corretor.



CATEGORIA	COR
1	Vermelha
2	Azul
3	Rosa
4	Preta
5	Verde
6	Marrom

Gráfico 10 – Corretor nº. 62 identificou seis categorias no item “1’

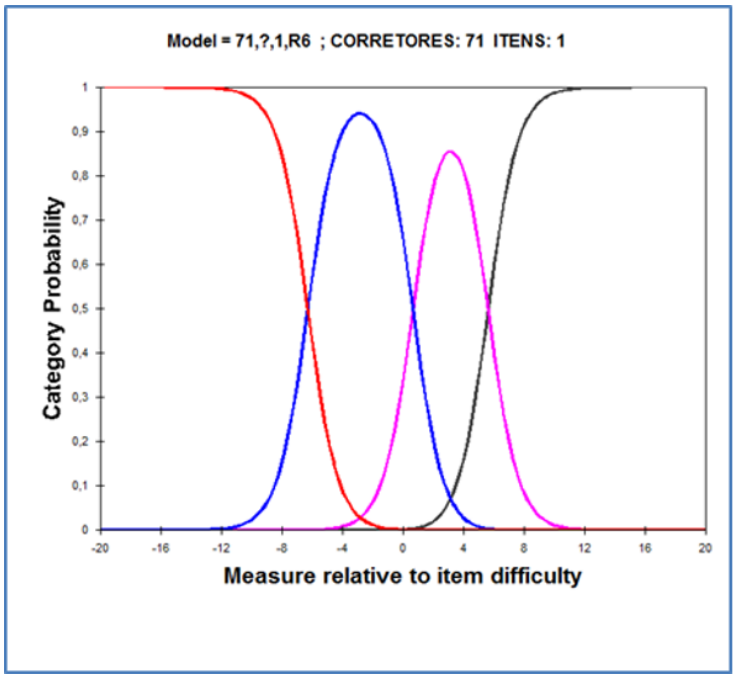
Fonte: O Pesquisador através do *software FACETS* 2018.



CATEGORIA	COR
1	Vermelha
2	Azul
3	Rosa
4	Preta
5	Verde
6	Marrom

Gráfico 11 – Corretor nº. 64 identificou cinco categorias no item “1”

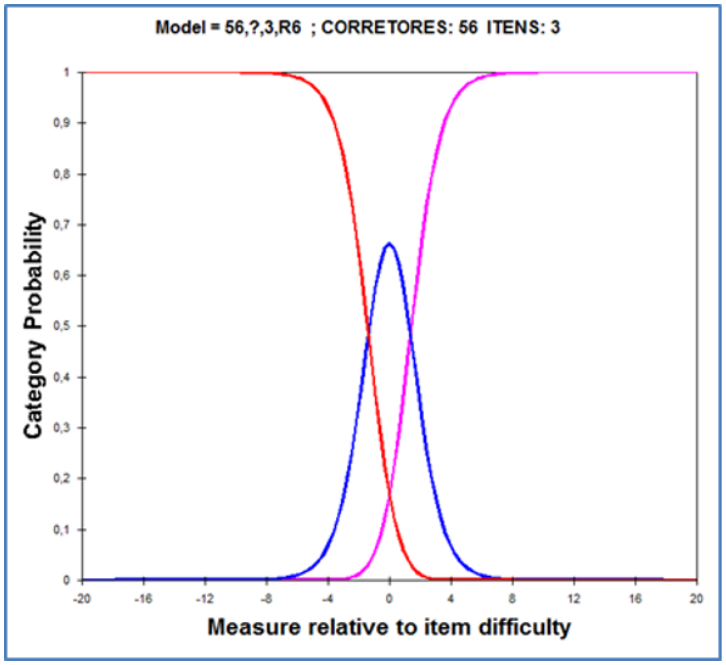
Fonte: Elaboração própria através do *software FACETS* 2018.



CATEGORIA	COR
1	Vermelha
2	Azul
3	Rosa
4	Preta
5	Verde
6	Marrom

Gráfico 12 – Corretor nº. 71 identificou quatro categorias no item “1”

Fonte: O Pesquisador através do *software FACETS* 2018.



CATEGORIA	COR
1	Vermelha
2	Azul
3	Rosa
4	Preta
5	Verde
6	Marrom

Gráfico 13 – Corretor nº. 56 identificou três categorias no item “3”

Fonte: O Pesquisador através do *software FACETS* 2018.

6.4.4 Rankeamento dos alunos com utilização da nota e da proficiência

Embora a correlação ente nota e proficiência seja muito alta, procuramos nessa análise ver o efeito da utilização dessas variáveis para rankeamento de alunos em diferentes situações de seleção. Essas situações variaram na quantidade de alunos selecionados.

Nosso objetivo foi verificar qual o efeito de se utilizar o rankeamento pela proficiência ao invés de se utilizar o rankeamento pela nota.

Para atingirmos nosso objetivo, fizemos inicialmente o rankeamento pela proficiência e criamos a variável Rank_prf (variando de 1 a 74.895), em seguida, ranqueamos pela nota, e criamos a variável Rank_nota, também variando de 1ª 74.895. Tendo como referência o rankeamento pela proficiência, verificamos para cada uma das 12 situações quantos alunos foram excluídos se o critério fosse a nota. Na Tabela 19, apresentamos as doze situações de rankeamentos utilizadas nessa análise.

Tabela 19 – Quantitativos de alunos excluídos, em diferentes percentuais da base, ao se utilizar como critério de seleção a proficiência e não a nota.

Situação	Seleção pela proficiência		Seleção pela nota			
			Nº de alunos		% de alunos	
	% da base	alunos	excluídos	comuns	excluídos	comuns
1	0,1	75	24	51	32,0	68,0
2	0,2	150	59	91	39,3	60,7
3	0,3	225	114	111	50,7	49,3
4	0,4	300	143	157	47,7	52,3
5	0,5	375	205	170	54,7	45,3
6	0,6	450	212	238	47,1	52,9
7	1	750	376	374	50,1	49,9
8	10	7.500	2557	4943	34,1	65,9
9	20	15.000	4015	10985	26,8	73,2
10	30	22.500	4704	17796	20,9	79,1
11	40	30.000	5536	24464	18,5	81,5
12	50	37.500	5545	31955	14,8	85,2

Fonte: O Pesquisador.

Temos para a primeira situação, que ao selecionarmos os 75 primeiros alunos pela variável Rank_prf (proficiência), verificamos que 24 alunos que seriam

selecionados nas 75 primeiras posições se o critério fosse nota, utilizando a variável Rank_nota, ficaram fora do processo de seleção e 51 alunos seriam selecionados pelos dois critérios. Esses quantitativos em termos percentuais são 32% e 68%, respectivamente. As demais situações seguem a mesma linha de interpretação.

Podemos constatar que à medida que aumenta o quantitativo de alunos selecionados, os percentuais de excluídos diminui, o que é óbvio, pois o quantitativo de alunos selecionados tende para o quantitativo da população total.

No que se refere à classificação pelas duas variáveis temos diferenças significativas no rankeamento do aluno, pois um aluno que teve sua redação corrigida por dois corretores severos será prejudicado, como é o caso do aluno 28879 que ficou classificado em 37º lugar pela proficiência e em 500º pela nota, da mesma forma o aluno que teve sua redação corrigida por dois corretores lenientes, será beneficiado com uma nota maior, como é o caso do aluno 2260 que ficou classificado em 1º lugar pela nota e em 127º lugar pela proficiência. Nesses dois exemplos apresentados, em um critério para seleção dos melhores 75 alunos de uma base de 74895 alunos, ou seja, aproximadamente 0,1% da base, teríamos seleções diferentes em função da metodologia utilizada.

Ao utilizarmos o rankeamento pela proficiência por meio da MFRM estaremos sendo mais justos com o processo de seleção, pela maior confiabilidade de validade desse método em relação à nota pela TCT, conforme demonstrado no tópico 6.4.

6.5 Considerações do capítulo

A utilização da MFRM pode ser implementada de várias formas, em função das características das avaliações e também pode ser utilizada como critério de monitoramento do processo de correção com o objetivo de se obter uma maior uniformidade e melhoria dos trabalhos de correção.

Nesse capítulo, realizamos as seguintes análises:

- **Dupla e tripla correção das redações de proficiência pela MFRM**

Nesse procedimento, são mantidos os mesmos critérios adotados para a correção das redações, é fundamental que haja uma boa combinação entre duplas

de corretores de forma permitir uma boa calibração da severidade dos corretores. A medida de desempenho do aluno se dará pela proficiência ajustada pela severidade do corretor. Esse procedimento deverá ser adotado em situações de concurso como, por exemplo, a utilização do resultado como fator de ranking para entrada em uma universidade.

- **Dupla correção das redações e utilização de proficiência com retirada da terceira correção para casos discrepantes**

Nos estudos realizados, demonstramos a pouca eficácia da 3ª correção em função dos diferentes níveis de severidades dos profissionais envolvidos nessa atividade. Esse procedimento também se aplica na situação apresentada anteriormente.

- **Utilização da MFRM para controle do processo de correção**

A calibração da severidade dos corretores quando realizada ao longo do processo de correção, possibilita a detecção do comportamento diferencial da severidade dos corretores (DIF). Essa metodologia tem a finalidade de detectar corretores que mudam sua severidade ao longo do processo e para esses casos problemáticos, com muita variabilidade na severidade, as redações seriam recorrigidas utilizando corretores mais estáveis.

- **Utilização da MFRM para padronização dos itens**

Pela percepção de cada corretor em relação às categorias de respostas dos itens, é possível verificar se as estruturas propostas para os itens estão realmente conforme o esperado pelos especialistas responsáveis pela elaboração dos itens e do teste como um todo. Conforme observamos na Tabela 17, muitos corretores corrigem itens que possuem seis categorias de respostas, com “5”, “4” e “3” categorias. Nesse procedimento, é também possível identificar corretores com características que não estão ajustadas com a correção padrão do projeto.

- **Criação de uma escala de redação nacional**

Por meio da MFRM, é possível, como sugerido por Linacre (2014), a criação de uma escala de redação. Esse procedimento é possível pela utilização de uma equipe de corretores comuns entre redações realizadas em diferentes períodos de tempo. Através dessa equipe comum, é possível realizar equalização entre as redações de forma a ter resultados comparáveis través de uma mesma escala entre as avaliações.

- **Utilização da MFRM com severidade calibrada em amostras de duplas correções**

Em projetos cujo objetivo é produzir medidas para diagnóstico, como por exemplo, testes de alfabetização como Avaliação Nacional da Alfabetização (ANA) e Programa de Avaliação da Educação Básica do Estado do Espírito Santo (PAEBES-ALFA) que possuem avaliação da escrita por meio de itens politômicos; a utilização da MFRM em amostras é uma boa alternativa para a redução de custos de correção e com resultados de melhor qualidade aos comparados com a utilização de notas e dupla correção em toda a base.

Entretanto, ficou evidente que, para redações em que os resultados serão utilizados para ranquear alunos, o mais prudente seria a utilização da dupla correção em 100% da base, onde temos medidas mais fidedignas e válidas.

7. Considerações finais

Procuramos, nessa tese, por meio de procedimentos empíricos, apresentar as funcionalidades de modelos mais parcimoniosos na produção de medidas de desempenho de alunos (proficiência). Essa forma de abordar a produção de medidas vai ao encontro do estipulado por Bergan (2013), para quem, *a priori*, nenhum modelo é melhor do que o outro, devendo, portanto, a decisão de qual modelo a ser utilizado ser baseada em experimentos, na parcimônia e na verificação de se as estimativas obtidas atendem às necessidades de utilização da avaliação.

Observamos, no cenário das avaliações em larga escala realizadas no Brasil ao longo dos últimos vinte anos, uma predominância do método de 3PL da TRI e um direcionamento muito forte para análises de desempenho baseadas em comparações entre medidas. É comum gestores e professores ficarem preocupados com quedas de "1" ponto na proficiência de seus sistemas de ensino, em uma medida cuja escala, em termos práticos, varia de "0" a "500" pontos. É importante salientar que a complexidade dos modelos de 3PL não garante tal nível de precisão e que, além das comparações, os resultados precisam ser trabalhados pedagogicamente pelos professores de forma a justificar os investimentos em avaliações censitárias.

A inclusão do parâmetro de discriminação no modelo da TRI produz realmente medidas mais robustas às comparadas ao modelo Rasch. No entanto, em contrapartida, a interpretação pedagógica dos resultados fica mais complexa, e apesar de todas as iniciativas que envolvem devolutivas voltadas para a utilização dos resultados das avaliações por parte dos professores, ainda se está longe de tornar os resultados das avaliações algo de fácil entendimento por parte dos professores.

A "elegância" dos modelos Rasch, proporcionada pela sua parcimônia e pela objetividade específica (Rasch, 1956), torna os resultados das avaliações mais assimiláveis pelos professores, que podem vir a utilizá-los de forma mais eficaz em suas práticas em sala de aula.

Por meio das análises realizadas no capítulo 5, vemos na modelagem Rasch multifacetas uma alternativa para a produção de medidas em avaliações conduzidas pelo Inep, estados e municípios que utilizam ou utilizaram até então, a TCT, como

por exemplo, as avaliações do Mais Educação, PMALFA, SAERJINHO, PAEBES-TRI, Inova Muriaé e SAEMI, realizadas pelo CAEd/UFJF. Avaliações essas realizadas em âmbito nacional, estadual e municipal, conforme detalhado no Quadro 10.

Quadro 10 – Projetos realizados pelo CAEd/UFJF utilizando a TCT

PROJETO	Responsável	Realizações	periodicidade	Alunos avaliados por edição	disciplinas avaliadas	Séries avaliadas
Mais Educação	Ministério da Educação	2017 e 2018	2 vezes ao ano	1.000.000	Lingua Portuguesa e Matemática	3EF ao 9EF
PMALFA	Ministério da Educação	2018	Anual	2.500.000	Leitura, escrita e Matemática	1Ef e 2EF
SAERJINHO	SEDUC - Rio de Janeiro	2011 a 2015	Bimestral	1.200.000	Língua Portuguesa, matemática, Ciências da Natureza e Ciências humanas	5 EF, 9 EF, 1EM, 2EM e 3EM
PAEBES-TRI	SEDUC - Espírito Santo	2015 a 2018	Trimestral	8.000	Lingua Portuguesa e Matemática	1EM, 2EM e 3EM
Inova Muriaé	Secretaria municipal de Educação de Muriaé - MG	2017 e 2018	Anual	6.800	Lingua Portuguesa e Matemática	3EF, 4EF, 6EF,7EF, 8EFe 9EF
SAEMI	Secretaria municipal de Educação Ipojuca - PE	2014 e 2015	Bimestral	9.000	Lingua Portuguesa e Matemática	3EF ao 9EF

Fonte: CAEd/UFJF (2018).

O que caracteriza todas essas avaliações é o fato de serem diagnósticas e as correções serem realizadas pelos próprios professores, de modo que os resultados de desempenho dos alunos sejam trabalhados o mais rápido possível pela escola. Percebemos uma tendência crescente desse tipo de avaliação em projetos nacionais realizados pelo CAEd/UFJF, utilizando a TCT, em função da simplicidade e rapidez na produção dos resultados, o que agrada aos professores e demais atores envolvidos na avaliação no âmbito das unidades escolares e das redes de ensino.

Para esses tipos de avaliação, a substituição da TCT pela MFRM é uma alternativa plenamente viável e com mais benefícios. Com relação à sua viabilidade, podemos destacar dois fatores, o primeiro advém da própria característica dos modelos Rasch em que as proficiências geradas por esses modelos têm um alinhamento biunívoco com os percentuais de acerto no teste (no caso de avaliações com um único caderno de teste), o que, conforme apresentado no terceiro capítulo, facilita o entendimento por parte dos professores. O segundo fator está relacionado com a rapidez de acesso aos resultados da avaliação, permitindo aos professores trabalharem de forma diagnóstica com seus alunos; além de ser benéfica a possibilidade de se trabalhar os resultados dos alunos em uma escala,

proporcionando comparações de desempenhos entre séries e anos de aplicação com interpretações pedagógicas baseadas nas características dos itens.

Tendo como referência as análises do sexto capítulo, podemos concluir que a MFRM é uma ferramenta que pode ser implementada nos processos de correções de redações adotados no Brasil e que seguem a estrutura utilizada pelo Inep em avaliações como o Enem e Encceja.

Acrescente-se, ainda, a possibilidade da implantação da MFRM na Avaliação de proficiência em Língua Portuguesa para estrangeiros, realizada pelo Certificado de Proficiência em Língua Portuguesa para Estrangeiros (Celpe-Bras). Nesse exame, que utiliza a TCT na produção de notas, os candidatos são avaliados em quatro habilidades: produção oral e escrita e compreensão oral sendo que o processo de correção é realizado por juízes.

Vemos a utilização da MFRM nesses programas de avaliação como uma oportunidade de se construir escalas nacionais de proficiências com interpretações dos níveis de desempenho e medidas comparáveis ao longo do tempo.

Não é nosso objetivo propor a substituição dos modelos 3PL pelos modelos Rasch. O Brasil tem uma escala nacional para Língua Portuguesa e Matemática, construída por meio da modelagem 3PL e as políticas públicas, ao longo desses vinte anos, fizeram com que essa escala se mantivesse como referência para a quase totalidade das avaliações em larga escala no Brasil. Esse é um trabalho que deve ser reconhecido e mantido.

O que propomos é uma utilização de modelos mais simples, representados pelos modelos da "família" Rasch, assim como a implantação de mais projetos envolvendo modelagens multiníveis em estudos longitudinais, testes eletrônicos e demais recursos tecnológicos emergentes, os quais devem ser analisados empiricamente, sem o viés filosófico de que um modelo é melhor do que o outro.

Nossa intenção com as análises realizadas ao longo dessa tese é abrir espaço para novas modalidades de avaliação, de modo a termos modelos que atendam às necessidade das redes de ensino em novos cenários que despontam na atualidade, no sentido de se utilizar, da forma mais ampla e consciente possível, os resultados da avaliação para promover qualidade e equidade do ensino.

8. Referências bibliográficas

ADAMS, R. J.; WU, M.; WILSON, M. **Computer Program Manual ConQuest 2.0**. Hawthorn, Australia: ACER, 2007.

ADAMS, R. J.; WU, M.; WILSON, M. **Computer Program Manual ConQuest 3.0**. Hawthorn, Australia: ACER, 2012.

ALBERTAZZI, A. G. Jr; SOUZA, A. R. **Fundamentos de metrologia científica e industrial**. São Paulo: Manole, 2018. Disponível em: <www.labmetro.ufsc.br/livroFMCI>. Acesso em: 25 ago. 2018.

ALVES, C. B. **Making diagnostic inferences about student performance on the Alberta Diagnostic Matematics Project:** am application of the Attribute Hierarchy Method. Berkley, 2011.

ALVES, M. T. G. **Efeito-escola e fatores associados ao progresso acadêmico dos alunos entre o início da 5ª série e o fim da 6ª série do Ensino Fundamental 6ª série do Ensino Fundamental: uma 6ª série do Ensino Fundamental estudo longitudinal em escolas públicas no município de Belo Horizonte**. 2006. 202p. Tese (Doutorado) - Programa de Pós-Graduação em Educação-FAE, Universidade Federal de Minas Gerais, Belo Horizonte, 2006.

AMERICAN EDUCATIONAL RESEARCH ASSOCIATION, AMERICAN PSYCHOLOGICALASSOCIATION; NATIONAL COUNCIL ON MEASUREMENT IN EDUCATION. **Standards for educational and psychological testing**. Washington, DC: American Educational Research Association, 1999.

ASSOCIATION OF TEST PUBLISHER – ATP. **Innovation in testing**. 2018. <http://www.innovationsintesting.org/atp2018/media/ATP2018_ProgramBook.pdf>. Acesso em: 20 ago. 2018.

BAHIA. Secretaria da Educação do Estado. SABE. **Sistema de Avaliação Baiano de Educação. AVALIE ALFA – 2011**. Universidade Federal de Juiz de Fora, Faculdade de Educação, CAEd, Juiz de Fora, v. 1, annual, 2011.

BAKER, F. B. **The basics of item response theory.** 2. ed. EUA: 2001. 172p.

BERGAN, J. R. **Assessing the relative fit of alternative item response theory models to the data**. Arizona: ATI, 2010.

BERGAN, J. R. **Rach Versus Birnbaum:** new arguments in an old debate. Arizona: ATI, 2013.

BOCK, R. D. A brief history of item response theory. **Educational Measurement: Issues and Practice**, v. 16, n. 4, p. 21-33, 1997. http://dx.doi.org/10.1111/j.1745-3992.1997.tb00605.x

BROOKE, N. O futuro das políticas de responsabilização educacional no Brasil. **Cad Pesqui**, São Paulo, v. 36, n. 128, p. 377-401, maio/ago. 2006.

BRYCK A. S.; RAUDENBUSH, S. W. **Hierarchical Linear Models**. London: Sage Publications, 1992.

CHACHAMOVICH, E. **Teoria da resposta ao item: aplicação do modelo Rasch em desenvolvimento e validação de instrumentos em saúde mental**. 2007. 288p. Tese (Doutorado) – Programa de Medicina, Universidade Federal do Rio Grande do Sul, Porto Alegre, 2007.

CHALMERS, R. P. A Multidimensional Item Response Theory Package for the R Environment. **Journal of Statistical Software**, v. 48, n. 6, p. 1-29, May 2012.

CHAMPLAIN, A.; BOULAIS, A-P.; DALLAS, A. Calibrating the Medical Council of Canada's Qualifying Examination Part I using an integrated item response theory framework: a comparison of models and designs. **J Educ Eval Health Prof**, v. 13, 2016. https://doi.org/10.3352/jeehp.2016.13.6.

ECKES, T. Examining Rater Effects in TestDaF Writing and Speaking Performance Assessments: A Many-Facet Rasch Analysis. **Language Assessment Quarterly**, v. 2, n. 3, 2005. https://doi.org/10.1207/s15434311laq0203_2.

ECKES, T. **Introduction to Many-facet Rasch Measurement**: analyzing an evaluating rater-mediated assessment. Frankfurt: Peter Lang, 2011.

ENGELHARD, G. Jr.; MARK, W. **Objective Measurement: theory into practice**. Papers presented at sucessive International Objective Measurement Workshop (IOMW). USA: Ablex Publishung Corporation, 2000. 325p.

ENGELHARD, G. Jr.; WIND, S. A. **Iinvariant measurement with raters and ratijg scales. Rasch models for rater-mediated assessments**. New York and London: Routledge Taylor and Francis Group, 2018. 351p.

FERNANDES, R. **Índice de Desenvolvimento da Educação Básica (Ideb)**. Brasília: Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira, 2007. 26p. (Série Documental. Textos para Discussão, ISSN 1414-0640; 26).

FEUER, M. J. et al. **Uncommon measures:** equivalence and linkage among educational tests. Washington: National Academy of Sciences, 1999. 135p.

FRANCO, C.; ALVES, M. T. G. A pesquisa em Eficácia Escolar no Brasil: Evidências sobre o efeito das escolas e fatores associados à eficácia escolar". In: BROOKE, N.; SOARES, J. F. (Orgs). **Pesquisa em eficácia escolar:** origem e trajetória. Belo Horizonte: Editora UFMG, 2008. p. 482-500.

FREITAS, L. C. Políticas de Responsabilização: entre a falta de evidência e a ética. **Cad Pesqui**, São Paulo, v. 43, n. 148, p. 348-365, jan./apr. 2013.

GERES 2005. **Estudo Longitudinal da Geração Escolar**. Disponível em: <https://laedpucrio.wordpress.com/projetos/o-projeto-geres/>. Acesso em 20 ago. 2018.

GOLDSTEIN, H. Modelos da Realidade: Novas Abordagens para a Compreensão de Processos Educacionais. In: FRANCO, C. (Org.). **Avaliação, Ciclos e Promoção na Educação**. Porto Alegre: Artmed. 2001, p. 85-99.

GOLINO, H. F. et al. **Psicologia contemporânea:** compreendendo os modelos Rasch. São Paulo: Casa do Psicólogo, 2015. 416p.

HARMAN, H. H. **Modern factor anaysis**. Chicago: University of Chicago Press, 1967.

HAUCK FILHO, N. Medida psicológica: o debate entre as perspectivas conceituais representacionista e realista. **Aval Psicol**, Ttatiba, v. 13, n. 3, p. 399-408, dez. 2014.

INSTITUTO NACIONAL DE ESTUDOS E PESQUISAS EDUCACIONAIS ANÍSIO TEIXEIRA - Inep. **Plataforma Devolutivas**. 2019. Disponível em: <http://portal.inep.gov.br/devolutivas>. Acesso em: 05 jan. 2019.

KLEIN, R. Utilização da teoria da resposta ao item no sistema nacional de avaliação da educação básica (SAEB). **Rev Ensaio**, v. 11, n. 40, p.283-296, jul./set. 2003.

KOLEN, M. J.; BRENNAN, R. L. **Test equating, scaling, and linking:** methods and practices. 2. ed. New York: Springer, 2004. 548p.

LINACRE, J. M. **Computer program manual FACETS 3.71.4**. 2013. Disponível em: <www.winsteps.com>. Acesso em: 20 ago. 2018.

LINACRE, J. M. **Computer program manual FACETS 3.71.4**. 2014. Disponível em: <www.winsteps.com>. Acesso em: 20 ago. 2018.

LINACRE, J. M. **Many-facet Rasch Measurement**. Chicago: MESA PRESS, 1989.

LIVINGSTON, S. A. **Equating Test Scores (Without IRT)**. Educational Testing Service – ETS – USA, 2004.

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

MASTERS, G. N. A Rasch model for partial credit scoring. **Psychometrika**, v. 47, n. 2, p. 149-174, June 1982. https://doi.org/10.1007/BF02296272.

MATOS, DANIEL A. S. Estratégias de Verificação da Confiabilidade e Concordância Entre Juízes: aplicações na área educacional. In: **Anais da VII Reunião da ABAVE**.... Avaliação e Currículo: um diálogo necessário, 2013, p. 345-364.

McNAMARA, T; KNOCH, U. The Rasch wars: the emergence of Rasch measurement in language testing. **Language Testing**, v. 29, n. 4, p. 555-576, 2012. https://doi.org/10.1177/0265532211430367.

NAKANO, T. C.; PRIMI, R. Rasch-Master's Partial Credit Model in the assessment of children's creativity in drawings. **Span J Psychol**, v. 17, p. 35, 2014. https://doi.org/10.1017/sjp.2014.36.

OLIVEIRA, L. K. M.; FRANCO JR, F. C. J. **Três Investigações sobre escalas de proficiência e suas interpretações**. 2008. 203p. Tese (Doutorado) – Pontifícia Universidade Católica do Rio de Janeiro, Rio de Janeiro, 2008.

PASQUALI, L. **Teoria e métodos de medida em ciências do comportamento** /organizado por Luiz Pasquali. — Brasília: Laboratório de Pesquisa em Avaliação e Medida / Instituto de Psicologia / UnB: INEP, 1996. 432p.

PASQUALI, L. **Psicometria:** Rev. esc. enferm. USP vol.43 no.spe São Paulo, 2009.

PASQUALI, L. **Psicometria:** teoria dos testes na Psicologia e na Educação. Petrópolis: Vozes, 2011.

PASQUALI, L. **Teoria da Resposta ao Item – TRI – Manual para Iniciantes**. Brasília: Laboratório de Pesquisa em Avaliação e Medida – LabPAM, 2004. 230p.

PASQUALI, L. Validade dos testes psicológicos: será possível reencontrar o caminho? **Psicologia: Teoria e Pesquisa**, v. 23 n. especial, p. 099-107, 2007.

PASQUALI, L.; PRIMI, R. Fundamentos da teoria da resposta ao item –TRI. **Aval Psicol**, Porto Alegre, v. 2, n. 2, p. 99-110, 2003.

PRIMI, R. Avanços na interpretação de escalas com a aplicação da Teoria de Resposta ao Item. **Aval Psicol**, Porto Alegre, v. 3, n. 1, p. 53-58, 2004.

PROVINHA BRASIL. Conforme Portaria nº. 10, de 24 de abril de 2007. Disponível em: <www.inep.gov.br/provinha-brasil>. Acesso em 15 ago. 2018.

RAUDENBUSH, S. W.; BRYK, A. S. **Hierarchical Linear Models**. 2. ed. Thousand Oaks: Sage Publications, 2002.

RECKASE, M. D. **Multidimensional Item response Theory**. EUA: Springer, 2009.

ROBITZSCH, A.; STEINFIELD, J. Item response models for human ratings: Overview, estimation methods, and implantation in R. **Psychological Test and Assessment Modeling**, v. 60, n. 1, p. 101-109, 2018.

ROCHER, T. **Mesure des competences: les méthodes se valent-elles? Questions de psychométrie danslecadre de lévaluation de la comprehension de lécrit.** 2013. 453p. Tese (Doutorado em Psicologia) – Université Paris Ouest Nanterre La Défense, Paris, 2013.

SCARAMUCCI, M. O exame Celpe-Bras e a proficiência do professor de português para falantes de outras línguas. **Rev DIGILENGUA**, Córdoba, n. 12, p. 48-67, jun. 2012.

SCHULTZ, DUANE P. **História da Psicologia Moderna** – 4ª edição – Cengage Learning. 2019

SHAVELSON, R. J.; WEBB, N. M. **Generalizability theory:** a primer. Newbury Park, CA: Sage Publications, 1991.

SILVA, W. Análise da eficácia escolar das escolas de tempo integral no ensino médio do Ceará. Políticas e práticas de formação dos docentes, dirigentes escolares. planejamento, financiamento e avaliação da educação. In: **Anais do VI Congresso Ibero-Americano de Política e Administração da Educação**... Recife: ANPAE, 2018. p. 387-393.

SILVA, W. **Eficácia dos processos de linkagem na avaliação educacional em larga escala**. 2010, 143f. Dissertação (Mestrado em Educação) – Faculdade de Educação da Universidade Federal de Juiz de Fora, Juiz de Fora, 2010.

SILVA, W.; SILVA, T. J. A transmissão da tecnologia da avaliação em larga escala no Brasil. In: **Colloque International**... Penser les nouvelles problématiques dans une perspective internationale, 16-18 novembre, Paris, 2016.

SISTEMA DE AVALIAÇÃO BAIANO DA EDUCAÇÃO - SABE. **Avalie Ensino Médio 2011, Linguagens, códigos e suas tecnologias**. Rev Pedagógica. Juiz de Fora: CAEd/UFJF, 2011, p. 55.

SISTEMA PERMANENTE DE AVALIAÇÃO DA EDUCAÇÃO BÁSICA DO CEARÁ - SPAECE. Disponível em: <http://www.spaece.caedufjf.net/avaliacao-educacional/o-programa/>. Acesso em: 25 ago. 2018.

SOARES, J. F; ALVES, M. T. G. O efeito das escolas no aprendizado dos alunos: um estudo com dados longitudinais no Ensino Fundamental. **Educação e Pesquisa**, São Paulo, v. 34, n. 3, p. 527-544, set./dez. 2008.

SOARES, T. M. et al. Modelos de valor agregado para medir a eficácia das escolas. **Ensaio: Aval Pol Públ Educ**, Rio de Janeiro, v. 25, n. 94, p. 59-89, jan./mar. 2017.

STEMLER, S. E. A comparison of consensus, consistency, and measurement approaches to estimating interrater reliability. **Practical Assessment, Research & Evaluation**, v. 9, n. 4, 2004.

STEVENS, S. S. On the Theory of Scales of Measurement. **Science New Series,** v. 103, n. 2684, p. 677-680, June 1946.

TEDDLIE, C.; REYNOLDS, D. **The international handbook of school effectiveness research.** London and New York: Falmer Press, 2000.

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

TINSLEY, H. E. A.; WEISS, D. J. Interrater reliability and agreement. In: TINSLEY, H. E. A.; BROWN, S. D. (Eds.) **Handbook of applied multivariate statistics and mathematical modeling**. New York: Academic Press, 2000. p. 95-124.

TOFFOLI, S. F. L.; ANDRADE, F. D.; BORNIA, A. C. Evaluation of open items using many-facet Rasch model. **Journal of Applied Statistics**, v. 43, n. 2, p. 1-18, 2015. https://doi.org/10.1080/02664763.2015.1049938.

TOIT, M. **IRT from SSI:** BILOG-MG, MULTILOG, PARSCALE, TESTFACT. Scientific Software Internacional, Inc., 2003.

URBINA, S. **Fundamentos da testagem psicológica**. Porto Alegre: Artmed, 2007.

VALLE, R. **Teoria da Resposta ao Item**. SP: USP, 1999. Disponível em: <http://publicacoes.fcc.org.br/ojs/index.php/eae/article/view/2225/2183>. Acesso em: 15 ago. 2018.

VIANNA M. H. Avaliações nacionais em larga escala: Análises propostas. **Estudos em Avaliação Educacional**, n. 27, p. 41-76 jan./jun. 2003. http://dx.doi.org/10.18222/eae02720032177.

VOCABULÁRIO INTERNACIONAL DE METROLOGIA – VIM. **Conceitos fundamentais e gerais e termos associados**. Duque de Caxias, RJ: INMETRO, 2012. 81p.

WERKEMA, M. C. C. **Avaliação de sistemas de medição.** Série 6 Sigma. Volume 5. Belo Horizonte: Werkema Editora, 2006, 116p.

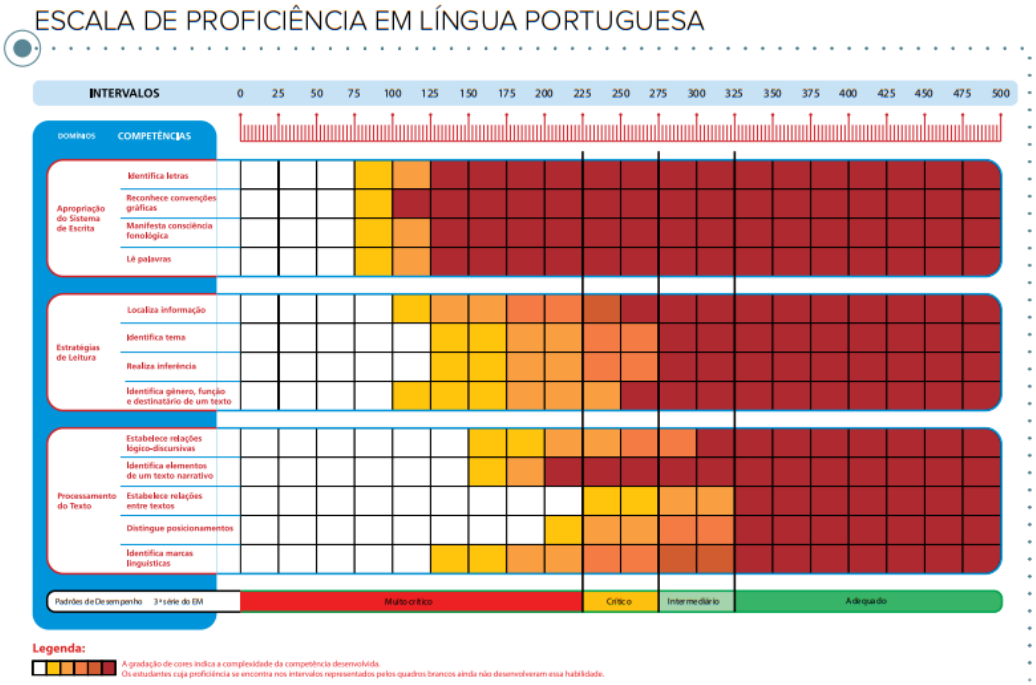
WRIGHT, B. D. **IRT in the 1990s:** Which Models Work Best? 3PL or Rasch? Measurement Transactions, 1992.

YEN, W. M. The extent, causes and importance of context effects on item parameters for two latent trait models. **Journal of Educational Measurement**, v. 17, n. 4, Winter, 1980.

ZIMOWSKI, M. F.; MURAKI, E.; MISLEVY, R. D. **BILOG-MG:** multiple-group IRT anaysis and test maintenance for binary itens. Chicago: Scientific Software, 1996.

9. Anexos

Anexo 1 – Matriz de referência



Anexo 2 – Plano de controle

Nº do Controle	Descrição da Operação	Método de Controle		Análise dos Modos e Efeitos de falhas		Plano de Ação
		Descrição	Modo de Detecção	Tipo de Falha	Efeito da Falha	Ações Recomendadas
1	Preparação da base de dados através da montagem dos cadernos e reposicionamento dos itens na mesma ordem	Verificação das estatísticas dos itens na organização geral e por modelo de caderno	Cálculo dos percentuais das opções de respostas dos itens por modelo de caderno	Discrepância entre os valores das estatísticas entre os diferetens cadernos de uma mesma etapa	Cálculo errado das proficiências	Rever a montagem das bases de dados e conferir a montagem dos itens nos cadernos
2	Análise Clássica	Verificação da correlação bisserial dos itens	Através dos softwares BILOGMG e SisAni	Itens com correlação bisserial negativa	Cálculo errado das proficiências	I) Correção do gabarito do item e gerar novas estatísticas; II) eliminação do item caso gabarito esteja correto.
				Itens com mais de uma opção de resposta com bisserial positiva	Item com baixa qualidade	Rever a estrutura do item e sua permanência no teste
			Através de software SPSS verificando a igualdade entre os valores recalculados e originais dos parâmetros a, b, c dos itens dos grupos de referência	Discrepância entre os valores	Cálculo errado das proficiências	I) Verificação dos valores fixados para os parâmetros dos itens do grupo de referência; II) Conferência da sintaxe do BILOGMG; III)Verificação da base de dados.
3	Análise de DIF	Verificação de anomalias no comporamento do item entre os grupos novos e o grupo de referência	Através de técnicas estatísticas de Mantel-Haenszel	Indicador de comportamento diferencial entre grupos	Perda de precisão no cálculo das proficiências	Verificar duas possibilidades: Eliminação de itens com DIF ou manter os itens no teste, mas consideraã-los como desiguais
4	Análise de Ajuste	Verificação do nível do ajuste entre o modelo teórico do item e o empírico	Através de técnicas estatísticas do qui-quadrado e gráficos das curvas características dos itens (modelo teórico x dados empíricos)	Desajuste da curva característica do item em relação aos dados empíricos	Perda de precisão no cálculo das proficiências	I) Comparação das proficiências mantendo e retirando os itens desajustados; II) analisar a permanência ou não dos itens desajustados no teste
5	Análise de Dimensionalidade	Verificação da unidimensionalidade das escalas	métodos de análise fatorial	itens relacionados a mais de uma escala	Perda de precisão no cálculo das proficiências	Exclusão de itens que não contribuam significativamente para a unidimensionalidade do teste

Anexo 3 – Chave de correção PAEBES ALFA

SITUAÇÃO: IMAGINE QUE VOCÊ DESCOBRIU UMA PASSAGEM SECRETA NO FINAL DA SUA RUA. FOI ATINGIDO POR UM RAIO E GANHOU PODERES MÁGICOS. ESCREVA UMA HISTÓRIA CONTANDO COMO FOI ESSA AVENTURA. NÃO SE ESQUEÇA DE CONTAR AS COISAS QUE VOCÊ VIU E FEZ.	
SITUAÇÃO: IMAGINE QUE VOCÊ ENCONTROU UMA MOEDA VALIOSA E MUITO ANTIGA ENTERRADA NO QUINTAL. ESCREVA UMA HISTÓRIA CONTANDO COMO FOI ESSA DESCOBERTA. NÃO SE ESQUEÇA DE CONTAR AS COISAS QUE VOCÊ VIU E FEZ.	
ASPECTO 1 – ADEQUAÇÃO À PROPOSTA	
A	O estudante escreveu uma história COERENTE com as cenas ou com a situação.
B	O estudante escreveu uma história com POUCA PLAUSIBILIDADE em relação às cenas ou à situação. Isto é, o estudante desenvolve de forma tangencial (superficial) o tema.
C	O estudante escreveu um texto que NÃO é coerente com a cena/situação proposta, ou seja, FUGA AO TEMA.
ASPECTO 2 – TIPOLOGIA TEXTUAL	
A	O estudante escreveu um texto que APRESENTA os elementos constitutivos de uma narrativa: narrador, lugar, tempo, personagens praticando ação e enredo (situação inicial, situação central e desfecho).
B	O estudante escreveu um texto com ausência de 1 (UM) ou MAIS elementos constitutivos de uma narrativa: narrador, lugar, tempo, personagens praticando ação e enredo (situação inicial, situação central e desfecho).
C	O estudante APENAS enumerou aspectos presentes nas cenas ou relativos a uma personagem ou fatos (no caso da narrativa a partir de situação).
D	O estudante escreveu um texto que não se configura como uma narrativa (Exs.: bilhete, lista, receita, poema...).
ASPECTO 3 – USO DA PÁGINA	
A	O estudante escreveu o texto usando o espaço delimitado para sua escrita de forma ADEQUADA, seguindo todas as regras de uso da página, ou seja, respeitando: as direções da escrita (de cima para baixo e da esquerda para a direita), as margens direita e esquerda, linha e sequência da escrita – mudança de linha.
B	O estudante escreveu o texto sem seguir, pelo menos, UMA das regras de uso da página.
C	O estudante escreveu o texto SEM SEGUIR as regras de uso da página.
ASPECTO 4 – ORTOGRAFIA	
A	O estudante escreveu o texto com uma ESCRITA ORTOGRÁFICA.
B	O estudante escreveu o texto com desvios que ainda permitem configurar sua escrita como ALFABÉTICA.
C	O estudante escreveu o texto com desvios que ainda permitem configurar sua escrita como SILÁBICO-ALFABÉTICA, sendo possível compreender o texto sem maior cooperação.

Anexo 4 – Matriz de competências para a produção de texto

MATRIZ DE COMPETÊNCIAS PARA A PRODUÇÃO DE TEXTO DO AVALIE BA 2011		COMPETÊNCIAS			
		I – Demonstrar domínio da norma padrão da língua escrita	*II – Compreender a proposta de produção textual e aplicar conceitos das várias áreas de conhecimento para desenvolver o tema, dentro dos limites estruturais do texto argumentativo*	*III – Selecionar, relacionar, organizar e interpretar informações, fatos, opiniões e argumentos em defesa de um ponto de vista.*	*IV – Demonstrar conhecimento dos mecanismos linguísticos necessários para a construção da argumentação.*
NÍVEL	**NÍVEL 0** (0 zero)	Demonstra desconhecimento da norma padrão, de escolha de registro e de convenções da escrita.	Apresenta informações desconexas, que não se configuram como texto.	Não defende ponto de vista e apresenta informações, fatos, opiniões e argumentos incoerentes.	Não articula as partes do texto.
	NÍVEL I (0,1 a 2,0)	Demonstra domínio insuficiente da norma padrão, apresentando graves e frequentes desvios gramaticais, de escolha de registro e de convenções da escrita.	Desenvolve de maneira tangencial o tema ou apresenta inadequação ao tipo textual argumentativo.	Não defende ponto de vista e apresenta informações, fatos, opiniões e argumentos pouco relacionados ao tema.	Articula as partes do texto de forma precária e/ou inadequada.
	NÍVEL II (2,1 a 4,0)	Demonstra domínio mediano da norma padrão, apresentando muitos desvios gramaticais, de escolha de registro e de convenções da escrita.	Desenvolve de forma mediana o tema a partir de argumentos do senso comum, cópias dos textos motivadores ou apresenta domínio precário do tipo textual argumentativo.	Apresenta informações, fatos e opiniões, ainda que pertinentes ao tema proposto, com pouca articulação e/ou com contradições, ou limita-se a reproduzir os argumentos constantes na proposta de produção textual em defesa de seu ponto de vista.	Articula as partes do texto, porém com muitas inadequações na utilização dos recursos coesivos.
	NÍVEL III (4,1 a 6,0)	Demonstra domínio adequado da norma padrão, apresentando alguns desvios gramaticais e de convenções da escrita.	Desenvolve de forma adequada o tema, a partir de argumentação previsível e apresenta domínio adequado do tipo textual argumentativo.	Apresenta informações, fatos, opiniões e argumentos pertinentes ao tema proposto, porém pouco organizados e relacionados de forma pouco consistente em defesa de seu ponto de vista.	Articula as partes do texto, porém com algumas inadequações na utilização dos recursos coesivos.
	NÍVEL IV (6,1 a 8,0)	Demonstra bom domínio da norma padrão, com poucos desvios gramaticais e de convenções da escrita.	Desenvolve bem o tema a partir de argumentação consistente e apresenta bom domínio do tipo textual argumentativo.	Seleciona, organiza e relaciona informações, fatos, opiniões e argumentos pertinentes ao tema proposto de forma consistente, com indícios de autoria, em defesa de seu ponto de vista.	Articula as partes do texto, com poucas inadequações na utilização de recursos coesivos.
	NÍVEL V (8,1 a 10)	Demonstra excelente domínio da norma padrão, não apresentando ou apresentando escassos desvios gramaticais e de convenções da escrita.	Desenvolve muito bem o tema com argumentação consistente, além de apresentar excelente domínio do tipo textual argumentativo, a partir de um repertório sociocultural produtivo.	Seleciona, organiza e relaciona informações, fatos, opiniões e argumentos pertinentes ao tema proposto de forma consistente, configurando autoria, em defesa de seu ponto de vista.	Articula as partes do texto, sem inadequações na utilização dos recursos coesivos.

Anexo 5 – Produção de texto

PRODUÇÃO DE TEXTO

Leia os textos motivadores abaixo e, em seguida, faça uma produção textual que atenda à proposta de redação.

Texto 1

Você já foi à Bahia?

Você já foi à Bahia, nêga?
Não?
Então vá!
Quem vai ao "Bonfim", minha nêga,
Nunca mais quer voltar.
Muita sorte teve,
Muita sorte tem,
Muita sorte terá
Você já foi à Bahia, nêga?
Não?
Então vá!
Lá tem vatapá
Então vá!
Lá tem caruru,
Então vá!
Lá tem munguzá,
Então vá!
Se "quiser sambar"
Então vá!

Nas sacadas dos sobrados
Da velha São Salvador
Há lembranças de donzelas,
Do tempo do Imperador.
Tudo, tudo na Bahia
Faz a gente querer bem
A Bahia tem um jeito,
Que nenhuma terra tem! [...]

Então vá...!

CAYMMI, Dorival. Disponível em: <http://letras.terra.com.br/dorival-caymmi/45590/>. Acesso em: 10 set. 2011. Fragmento.

Texto 2

Diversidade

[...] Valorizar a diversidade cultural significa valorizar a diferença. A diferença de ideias, de opções religiosas e sexuais, de matrizes culturais e etnias, de ideologias, saberes, práticas. A promoção da diversidade cultural depende de uma compreensão profunda do valor da diferença e da capacidade que uma sociedade tem de aceitar e conviver com o diferente.

Da música à culinária, da religião ao artesanato, a diversidade cultural é o nosso maior patrimônio. Não podemos, portanto, amesquinhar esse patrimônio na tentativa de encontrar uma só identidade cultural para a Bahia.

A cara da Bahia não pode ser apenas a cara do Recôncavo. A cara da Bahia tem que ser a cara da Bahia inteira: do Recôncavo, do Oeste, do São Francisco, do Sertão, do Sul, da Chapada e de todas as outras regiões do estado. Temos que assumir "ao mesmo tempo agora" toda diversidade baiana, as diferenças que, combinadas e recombinadas, misturadas, mestiças, fazem do povo baiano o que ele é. [...]

Disponível em: <http://www.secult.2201.com.br/linhasdeacao/diversidade>. Acesso em: 10 set. 2011. Fragmento. *Adaptado: Reforma Ortográfica.

Texto 3

ARTE E CULTURA REGIONAL BAIANA

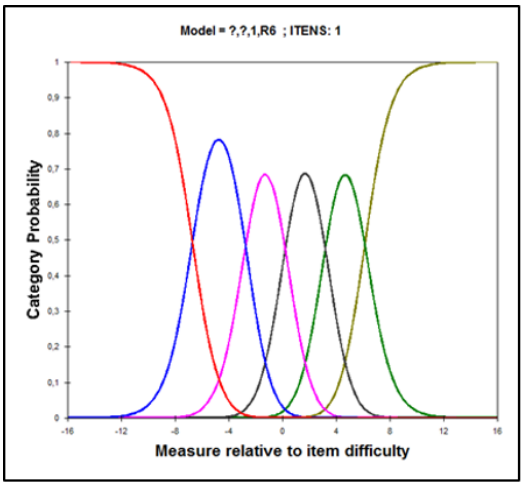
A cultura da Bahia é uma das mais ricas e diversificadas do Brasil, sendo o estado considerado um dos mais ricos centros culturais do país, [...]. A Bahia tem seus expoentes, suas características próprias, resultado da rica miscigenação entre o índio nativo, o português colononizador e o negro escravizado. [...]

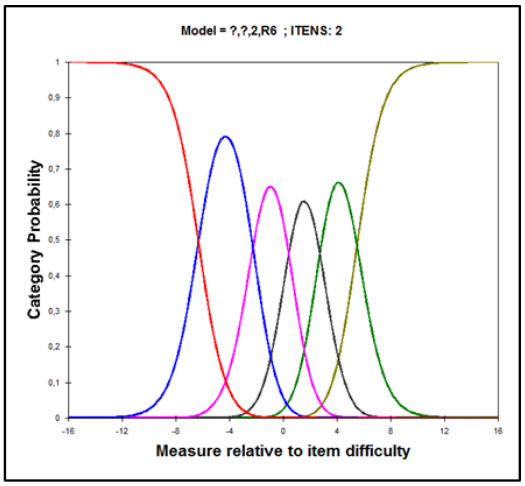
Na Bahia, ainda há espaço para um provérbio, a um tempo jocoso e sério, que retrata a índole do seu povo: "O baiano não nasce, estreia". [...] Do *rock* ao tropicalismo, de Raul Seixas a Caetano Veloso, infinitos nomes desfilam mundo afora, como João Gilberto, Gilberto Gil, Carlinhos Brown... Foi no Carnaval que o baiano encontrou-se com o mundo: Em 1950, Dodô e Osmar inventam o Trio Elétrico, e atrás dele "só não vai quem já morreu". [...] O negro reconquista sua identidade e ganha força nos Filhos de Gandhi, o Olodum une música ao trabalho social. Do Candomblé ou do tabuleiro da baiana, brotam o acarajé, o abará, o vatapá e tantos pratos temperados pelo azeite de dendê, festejando aos santos, como o caruru ou festejando a vida, como a moqueca, a Bahia tem sempre um quindim a despertar o paladar.

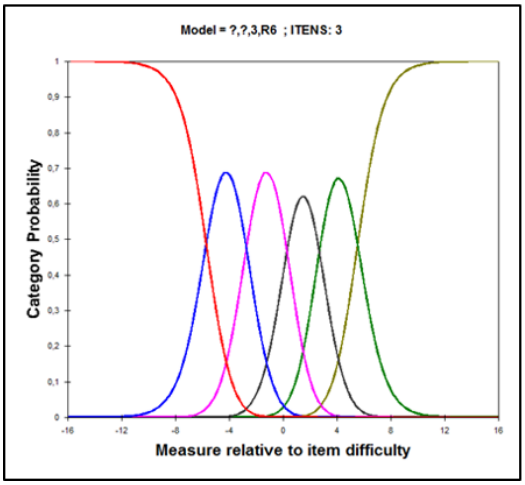
SOUZA, Jeovaci. Disponível em: <http://www.guajenet.com/2010/08/arte-e-cultura-regional-baiana.html>. Acesso em: 10 set. 2011. Fragmento.

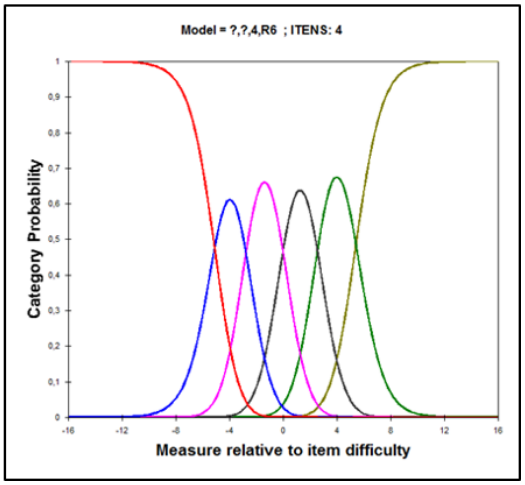
Com base nos textos motivadores apresentados e nos seus conhecimentos individuais, redija um texto argumentativo na modalidade culta da Língua Portuguesa, sobre o tema **A diversidade cultural baiana**, discutindo a importância cultural da Bahia. Não se esqueça de organizar o seu pensamento de forma coesa e coerente, com argumentos e fatos que o auxiliem na defesa do seu ponto de vista.

Anexo 6 – Curvas Características dos Itens (CCI)









Anexo 7 – Classificação dos corretores pela severidade

LENIENTES			MEDIANOS			SEVEROS		
CORRETOR	Z_SEV	FX_SEV	CORRETOR	Z_SEV	FX_SEV	CORRETOR	Z_SEV	FX_SEV
70	-2,92748	1	31	-0,55254	2	110	0,46924	3
112	-1,83206	1	99	-0,53413	2	80	0,49686	3
41	-1,75842	1	77	-0,49731	2	65	0,50606	3
83	-1,60193	1	30	-0,47890	2	71	0,50606	3
98	-1,43624	1	78	-0,45128	2	38	0,51527	3
86	-1,41783	1	62	-0,44207	2	76	0,55209	3
33	-1,39942	1	114	-0,34082	2	26	0,66255	3
79	-1,33498	1	51	-0,32241	2	60	0,68096	3
103	-1,31657	1	68	-0,30400	2	11	0,69937	3
96	-1,30736	1	32	-0,29479	2	89	0,69937	3
93	-1,26134	1	118	-0,24877	2	100	0,71778	3
66	-1,24293	1	75	-0,23035	2	29	0,72699	3
40	-1,23372	1	56	-0,21194	2	111	0,72699	3
58	-1,16008	1	120	-0,21194	2	37	0,77301	3
113	-1,15088	1	54	-0,17512	2	61	0,78222	3
24	-1,13246	1	5	-0,14751	2	92	0,80983	3
28	-1,12326	1	116	-0,14751	2	1	0,87427	3
90	-1,10485	1	67	-0,11069	2	59	0,88348	3
46	-1,09564	1	43	-0,10148	2	52	1,01235	3
117	-1,09564	1	84	-0,07387	2	64	1,04917	3
6	-1,05882	1	4	-0,02784	2	69	1,05837	3
72	-1,04962	1	15	-0,02784	2	21	1,06758	3
74	-1,01280	1	88	0,01819	2	95	1,11361	3
53	-0,98518	1	50	0,07342	2	10	1,16884	3
108	-0,97598	1	35	0,08262	2	73	1,16884	3
2	-0,92995	1	122	0,11024	2	82	1,17804	3
39	-0,92075	1	85	0,11944	2	7	1,24248	3
49	-0,90233	1	48	0,14706	2	45	1,26089	3
115	-0,90233	1	55	0,15626	2	17	1,29771	3
13	-0,83790	1	25	0,17467	2	12	1,30691	3
36	-0,75505	1	102	0,20229	2	16	1,60148	3
63	-0,72744	1	121	0,20229	2	8	1,61069	3
57	-0,68141	1	97	0,26673	2	20	1,61069	3
101	-0,68141	1	91	0,30355	2	106	1,62910	3
107	-0,68141	1	27	0,35878	2	19	1,68433	3
34	-0,67220	1	23	0,36798	2	14	1,78559	3
81	-0,66300	1	119	0,36798	2	18	1,89605	3
94	-0,64459	1	123	0,39560	2	42	1,91446	3
87	-0,61697	1	3	0,41401	2	22	2,00651	3
104	-0,61697	1	44	0,44163	2	47	2,00651	3
105	-0,57095	1	109	0,46004	2	9	2,33790	3