UDESC
UNIVERSIDADE DO ESTADO DE SANTA CATARINA – UDESC
CENTRO DE CIÊNCIAS TECNOLÓGICAS - CTT
PROGRAMA DE PÓS GRADUAÇÃO PROFISSIONAL EM MATEMÁTICA EM REDE NACIONAL

DISSERTAÇÃO DE MESTRADO

ANÁLISE DE UM SIMULADO DE MATEMÁTICA UTILIZANDO A TEORIA DE RESPOSTA AO ITEM E A TEORIA CLÁSSICA DOS TESTES

MARCOS ELIAS NUNES

JOINVILLE, 2018

ANÁLISE DE UM SIMULADO DE MATEMÁTICA UTILIZANDO A TEORIA DE RESPOSTA AO ITEM E A TEORIA CLÁSSICA DOS TESTES

Dissertação submetida ao Programa de
Mestrado Profissional em Matemática
para obtenção do grau de mestre.
Universidade do Estado de Santa Catarina.
Orientadora: Profa. Dra. Elisa Henning

Joinville

2018

Nunes, Marcos Elias
Análise de um Simulado de Matemática Utilizando a Teoria de Resposta ao Item e a Teoria Clássica dos Testes / Marcos Elias Nunes. - Joinville , 2018.
81 p.

Orientadora: Elisa Henning
Dissertação (Mestrado) - Universidade do Estado de Santa Catarina, Centro de Ciências Tecnológicas, Programa de Pós-Graduação , Joinville, 2018.

1. Teoria Clássica dos Testes. 2. Teoria de Resposta ao Item. I. Henning, Elisa. II. Universidade do Estado de Santa Catarina. Programa de Pós-Graduação. III. Título.

Item e a Teoria Clássica dos Testes

por

Marcos Elias Nunes

Esta dissertação foi julgada adequada para obtenção do título de

MESTRE EM MATEMÁTICA

Área de concentração em "Ensino de Matemática"
e aprovada em sua forma final pelo

CURSO DE MESTRADO PROFISSIONAL EM MATEMÁTICA EM REDE NACIONAL
DO CENTRO DE CIÊNCIAS TECNOLÓGICAS DA
UNIVERSIDADE DO ESTADO DE SANTA CATARINA.

Banca Examinadora:

Profa. Dra. Elisa Henning
CCT/UDESC (Orientadora/Presidente)

Profa. Dra. Andrea Cristina Konrath
UFSC

Profa. Dra. Regina Helena Munhoz
CCT/UDESC

Joinville,SC, 31 de julho de 2018.

Este trabalho é dedicado...

Aos meus pais: Manoel de Souza Nunes e Pedra Elias Nunes pelos valores ensinados com muito amor e dedicação;

As minhas irmãs:Cristiane Elias Nunes, Cristina Elias Nunes e Gisele Elias Nunes por serem exemplos de honestidade e competência;

A minha esposa: Rejeane de Lima por permanecer ao meu lado em todos os momentos, sejam eles bons ou ruins.

Agradecimentos

A Deus, à minha família e à minha esposa; A todos os meus amigos que me ajudam a manter o equilíbrio emocional; Aos meus colegas do PROFMAT pelo auxílio nos momentos de dificuldade e descontração nos momentos de mais tensão; A professora Elisa Henning, pela dedicação e competência ao me auxiliar em todas as etapas deste trabalho; Aos demais professores da UDESC pelo aprendizado proporcionado; As professoras Andréa Cristina e Regina Helena Munhoz pelas considerações que auxiliaram na finalização desta dissertação; A esta universidade pela oportunidade de crescimento profissional e pessoal;

Resumo

A aplicação de simulados é uma prática comum quando se deseja avaliar um grupo de estudantes. Porém, para que os resultados apresentados sejam válidos, é necessário que os itens que compõem o simulado sejam capazes de discriminar os participantes que possuem as habilidades que estão sendo medidas dos estudantes que não possuem. Desta forma, o presente trabalho se propôs a analisar a qualidade dos itens de um simulado de Matemática aplicado na rede municipal de educação de Jaraguá do Sul. Esta análise foi feita utilizando e comparando os resultados da Teoria Clássica dos Testes (TCT) e da Teoria de Resposta ao Item (TRI). Concluiu-se que a maior parte dos itens necessita de revisão. Os resultados obtidos com as análises da TCT e da TRI foram confrontados, garantindo uma maior fidedignidade nas conclusões sobre os itens. Os resultados das duas teorias apresentaram divergências em relação aos itens com maior poder de discriminação. Por outro lado, o item considerado mais difícil do teste foi o mesmo nas duas teorias. Além disso, foi realizada uma análise pedagógica nos itens considerados deficientes, com o intuito de contribuir com futuros trabalhos similares. Algumas sugestões de alteração no enunciado da questão e das alternativas foram feitas neste trabalho.

Palavras-chave: Teoria Clássica dos Testes. Teoria de Resposta ao Item. Matemática. Avaliação.

Abstract

The application of simulation is a common disease when one has a diagnosis about a certain population. However, for results to be displayed, performance criteria need to be able to discriminate participants who possess skills that can be done by students who do not. In this way, the present work has properly an analysis of the results of a mathematical questionnaire in the municipal education network of Jaraguá do Sul. This research was done using the results of Classical Theory of Tests (CTT) and Item Response Theory (IRT). It was concluded that a larger part of the items needed for revision. The results of the CTT and IRT tests were confronted, guaranteeing greater reliability in the conclusions about the items. The data of the two historical series divergences in relation to the items with greater power of discrimination. On the other hand, the item was more difficult to apply in both theories. In addition, a pedagogical analysis was set up in the following previous issues, to achieve the next similar works. Some suggestions for change do not address the issue and the alternatives are made in this work.

Key-words: Classical Theory of Tests. Item Response Theory. Mathematics. Assessment.

Lista de Figuras

Lista de Tabelas

Lista de Abreviações e Siglas

CCI	Curva Característica do Item
ENEM	Exame Nacional do Ensino Médio
ENADE	Exame Nacional de Desempenho de Estudantes
IBGE	Instituto Brasileiro de Geografia e Estatística
IDEB	Índice de Desenvolvimento da Educação Básica
INEP	Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira
ML1	Modelo Logístico de um Parâmetro
ML2	Modelo Logístico de dois Parâmetros
ML3	Modelo Logístico de três Parâmetros
OBM	Olimpíada Brasileira de Matemática
OBMEP	Olimpíada Brasileira de Matemática das Escolas Públicas e Privadas
OPA	Oportunidade para Aprender
SAEB	Sistema de Avaliação da Educação Básica
SEMED	Secretaria Municipal de Educação
TCT	Teoria Clássica dos Testes
TRI	Teoria de Resposta ao Item

Sumário

1 Introdução

A primeira pergunta que surge ao nos depararmos com o título deste trabalho é: porque utilizar um simulado como método de avaliação educacional? Pois bem, quando surge a necessidade de avaliar um determinado grupo de estudantes, o que se deseja é obter um resultado que apresente, com a maior proximidade possível da realidade, as habilidades adquiridas pelos analisados. Portanto, uma maneira de se obter resultados é submeter este grupo a uma "simulação" em que as resoluções exijam certas habilidades. Esta simulação, ou teste, pode ser formado por dois tipos diferentes de itens: discursivos ou objetivos. A nossa escolha por um teste objetivo se dá pela fidedignidade das análises das respostas. Um teste com itens discursivos pode apresentar diferentes interpretações quando analisados por pessoas diferentes. Além disso, quando a população é consideravelmente grande e os itens do teste são discursivos, a correção se torna muito mais trabalhosa.

Por ser uma etapa importante no processo de avaliação, a elaboração de um teste pelo professor costuma ser cuidadosamente construída e sofre constante aprimoramento. Segundo Klein e Fontanive (1995), em geral, o processo de planejamento dos testes combina os conteúdos curriculares e as habilidades hierarquizadas em níveis de complexidade a partir do que se espera que o aluno saiba e seja capaz de fazer.

A avaliação, de maneira geral, é uma forma de obter dados que permitam uma análise dos objetivos a serem atingidos (LUCKESI, 2008). Apesar de termos muito a discutir sobre as formas de se avaliar, é difícil pensar na evolução de um trabalho sem submeter os sujeitos do mesmo a alguma avaliação. Para Luckesi (2008), a avaliação da aprendizagem é uma prática de investigação do professor, cujo sentido é intervir na busca dos melhores resultados do processo de aprendizagem dos nossos educandos, em sala de aula. Em seu conceito, Luckesi (2008), afirma que a avaliação é um juízo de qualidade sobre dados relevantes para uma tomada de decisão. Elaborar uma avaliação que traduza as habilidades que se deseja analisar não é uma tarefa simples, mas, sem dúvidas, partir dos resultados obtidos em avaliações bem elaboradas, a chance de sucesso em nossas tomadas de decisões é maior. Isso ocorre porque o processo de educação é constante e uma etapa serve de pré-requisito para a próxima etapa. Se os estudantes apresentam um bom desempenho em uma

avaliação que foi mal elaborada, isso pode gerar problemas em etapas futuras que dependam desta, pois os estudantes podem, na realidade, não possuir as habilidades necessárias.

Amplamente utilizada entre os professores, no momento da avaliação, a Teoria Clássica dos Testes (TCT) consiste em atribuir a mesma pontuação para cada questão, independente do grau de dificuldade e de outros parâmetros envolvidos. Este tipo de avaliação apresenta dificuldades para comparar dois respondentes distintos que foram submetidos ao mesmo teste (RABELO, 2013), pois dois indivíduos que atingiram a mesma nota não necessariamente acertaram as mesmas questões e, consequentemente, não possuem as mesmas habilidades. Por exemplo, em um teste de nove questões que tenha sido aplicado para uma turma do 9º ano do Ensino Fundamental. Com este teste, deseja-se medir as habilidades dos respondentes em Trigonometria no triângulo retângulo. Das nove questões, três envolvem seno, três envolvem cosseno e três envolvem tangente. Dois estudantes acertaram seis questões, porém um deles errou as três que envolem tangente enquanto o outro errou uma de cada relação trigonométrica. Se estes estudantes forem avaliados pela TCT terão notas iguais, porém um deles não atingiu o conhecimento necessário sobre tangente. Para resolver estes e outros problemas em potencial, uma opção a ser utilizada como método de avaliação é a Teoria de Resposta ao Item (TRI). Andrade, Tavares e Valle (2000), afirmam que a TRI vem sendo progressivamente introduzida em nosso meio, e é um instrumento poderoso nos processos quantitativos de avaliação educacional, pelo fato de permitir, inclusive, a construção de escalas de habilidades calibradas. Este constante crescimento da utilização da TRI se dá pelo fato de que as avaliações em larga escala possuem, de modo geral, itens objetivos, o que permite a aplicação desta teoria. Dentre estas provas podemos destacar, aqui no Brasil, o ENEM (Exame Nacional do Ensino Médio), a OBMEP (Olimpíada Brasileira de Matemática das Escolas Públicas e Privadas), o SAEB (Sistema de Avaliação da Educação Básica), a Prova Brasil, além de questões de concursos, vestibulares, entre outros testes aplicados em larga escala.

Uma avaliação muito importante para a educação nacional, aplicada bienalmente é a Prova Brasil. Essa importância está relacionada ao resultado dos estudantes nesta prova. A nota da Prova Brasil serve de base para o cálculo do IDEB (Índice de Desenvolvimento da Educação Básica). Infelizmente este índice não é completamente fidedigno, pois a taxa de aprovação escolar é levada em consideração para o cálculo da nota final. Isso significa que se a média geral de duas escolas for a mesma, mas uma das escolas obtiver um índice de aprovação de 100% enquanto a outra

obtiver um índice de aprovação de 80%, a escola que obteve uma taxa de aprovação maior terá um IDEB maior.

Por gerar um índice fundamental no desempenho de uma unidade escolar, a Prova Brasil motivou o município de Jaraguá do Sul a criar uma prova local, a ser aplicada bienalmente no município, denominada a partir daqui de "prova Jaraguá". Aplicada desde 2015, esta prova é formada por itens objetivos e aplicada em larga escala. Por estes motivos, foi escolhida a Teoria da Resposta ao Item na elaboração e análises desta prova. Com isso, buscamos aumentar a efetividade dos testes e, consequentemente, obter resultados que possam traduzir de maneira mais eficaz a qualidade do trabalho que vem sendo feito.

O objetivo geral deste trabalho é analisar um simulado de matemática utilizando a Teoria Clássica dos Testes (TCT) e a Teoria de Resposta ao Item (TRI). Como objetivos específicos deste trabalho vamos analisar analisar os itens de acordo com a TCT utilizando os coeficientes bisseriais, discriminação, dificuldade e distratores; analisar os itens de acordo com a TRI utilizando os coeficientes bisseriais e a dificuldade do item, discriminação, acerto ao acaso, Curva Característica do Item (CCI), e curva de informação do item; fazer uma análise pedagógica das questões que apresentam deficiências, apontando os possíveis fatores que influenciaram no resultado, além de propor melhorias; comparar os resultados das análises dos itens obtidos com as duas teorias, TCT e TRI; determinar a dimensionalidade e confiabilidade do teste e apontar o que pode ser melhorado para que futuras avaliações similares possam discriminar melhor os respondentes, apresentado resultados mais satisfatórios.

Este trabalho justifica-se e é relevante pois irá auxiliar a responder perguntas que aparecem naturalmente durante o processo de aprendizagem, como: Estamos fazendo um bom trabalho? Os estudantes estão em constante evolução? Em que podemos melhorar?

Estas foram algumas das questões que motivaram o município de Jaraguá do Sul a criar a prova Jaraguá. Como a prova é aplicada bienalmente, existe a necessidade de que o nível da prova e as habilidades que se deseja medir, não se alterem entre uma edição e outra. Por este motivo, na elaboração e análise destas provas, a secretaria municipal de educação de Jaraguá do Sul utiliza a Teoria de Resposta ao Item, o que permite, antes mesmo de avaliar o desempenho dos alunos, avaliar a confiabilidade dos testes.

Por ser um professor do quadro efetivo dos servidores do município de Jaraguá do Sul e querer contribuir no processo de aprendizagem, surgiu a ideia de fazer este trabalho que me proporcionou aprofundar os conhecimentos sobre a TRI e analisar estatisticamente e pedagogicamente os itens de um dos simulados que já foi aplicado, colaborando com futuras aplicações da prova Jaraguá.

Além disso, os resultados e discussões apresentados neste trabalho podem auxiliar o dia a dia do professor, fazendo-o refletir sobre o processo de elaboração das suas provas, corrigindo as suas falhas e confirmando as suas certezas.

No capítulo 2 deste trabalho apresentamos a Teoria Clássica dos testes, discutindo as suas limitações e definindo os conceitos utilizados para analisar a prova Jaraguá. De maneira semelhante, no capítulo 3 apresentamos a Teoria de Resposta ao Item. No capítulo 4 estão em destaque alguns trabalhos que executaram pesquisas e análises similares a esta, contendo alguns resultados obtidos. Os procedimentos metodológicos, materiais e métodos utilizados na aplicação da prova Jaraguá e justificativa para a aplicação da mesma se encontram no capítulo 5. Os resultados obtidos estão apresentados no capítulo 6, incluindo a comparação entre os resultados obtidos com a TCT e com a TRI. No capítulo é realizada a análise pedagógica dos itens, principal objetivo das análises deste trabalho. No capítulo 8 há uma discussão que envolve experiências da prática docente do autor. Por fim, no capítulo 9, apresentamos algumas considerações, limitações e sugestões de continuidade do trabalho.

2 A Teoria Clássica dos Testes

Neste capítulo vamos falar um pouco sobre os tipos de avaliação e também sobre a avaliação em larga escala no Brasil. Além disso, vamos descrever um pouco sobre a Teoria Clássica dos Testes. Entre os principais conceitos que envolvem esta teoria, destacamos o resultado que é apresentado com base nos escores brutos, ou seja, dois estudantes que fizeram a mesma prova e acertaram a mesma quantidade de questões terão a mesma nota, mesmo que as questões não sejam as mesmas. Além disso, falamos também das limitações que esta teoria apresenta e como corrigir ou melhorar essas limitações.

2.1 Avaliação

Em todas as atividades que envolvem algum tipo de aprendizagem e deseja-se medir o nível desta aprendizagem, deve-se submeter os participantes a algum tipo de avaliação. A partir da década de 60 a literatura sobre avaliação aumentou consideravelmente (VIANNA, 1995). O termo avaliação vem sendo discutido em vários trabalhos acadêmicos com o passar dos anos, repensando-se estratégias e revendo-se conceitos. Segundo Luckesi (2007) a avaliação, tanto no geral quanto no caso específico da aprendizagem, não possui uma finalidade por si só, mas está inserida em um processo que visa construir um resultado previamente definido. Isso significa que a avaliação não deve terminar apenas na aplicação e obtenção do resultado de uma prova, mas sim que este resultado sirva para que o professor possa repensar as suas estratégias, confirmando ou reformulando os processos didáticos utilizados.

Segundo Cortesão (2002) há três tipos de avaliação: avaliação diagnóstica, avaliação formativa e avaliação somativa. A avaliação diagnóstica é utilizada quando se deseja identificar competências do público alvo antes de se iniciar um trabalho. Em cursos de inglês, por exemplo, costuma-se fazer uma prova para identificar em que nível se encontra o estudante. Posteriormente o estudante é encaminhado para a turma que condiz com o nível que o mesmo se encontra. A avaliação formativa tem por característica contribuir constantemente com o processo de aprendizagem. As provas que os alunos de uma escola fazem durante o ano, por exemplo, tem caráter formativo, desde que haja retomada dos pontos necessários. Por fim, a avaliação somativa tem por objetivo apresentar um sumário de resultados obtidos em uma situação educativa. Este tipo de avaliação costuma ser

aplicada, por exemplo, no final de um ano, no final de um curso, ou de qualquer período letivo de uma unidade de ensino (CORTESÃO, 2002).

A prova Jaraguá, objeto de estudo deste trabalho, tem um caráter tanto de uma avaliação diagnóstica como de uma avaliação somativa, visto que o objetivo da aplicação desta prova é avaliar as competências dos estudantes antes de iniciar o próximo ano letivo para que se possa trabalhar as deficiências, além de avaliar a qualidade da educação durante os dois anos anteriores a mesma.

Quando se fala em avaliação, a maneira mais utilizada para mensurar o domínio de um certo indivíduo sobre determinado conteúdo é a aplicação de provas. Para que se possa comparar tal domínio entre os respondentes destas provas deve-se, de alguma maneira, padronizar a análise dos resultados. A conduta do professor no processo de aferição do aproveitamento escolar tem sido a nota ou coneito (LUCKESI, 2007). Como a prova Jaraguá é aplicada para um número relativamente grande de estudantes (aproximadamente 10.000), necessita-se de uma estratégia eficiente de avaliação em larga escala.

No Brasil, existem alguns testes aplicados periodicamente como o Sistema de Avaliação da Educação Básica (SAEB), Exame Nacional do Ensino Médio (ENEM) e o Exame Nacional de Desempenho de Estudantes (ENADE). Mas a experiência brasileira não se restringe às avaliações de abrangência nacional. Segundo Rabelo (2013), Estados e municípios têm criado os seus próprios sistemas, sendo que muitos deles escolhem metodologias que permitem comparar os resultados obtidos com os nacionalmente estabelecidos.

A Teoria Clássica dos Testes foi criada para cumprir este papel de análise dos resultados. Esta teoria, largamente utilizada antes da criação da Teoria de Resposta ao Item, baseia-se em resultados obtidos em provas através de escores brutos ou padronizados. O que é levado em consideração aqui é o resultado final do teste de cada respondente, quem acerta mais questões possui mais habilidades, conforme descrito por Vianna (2014).

Segundo Rabelo (2013) podemos perceber algumas deficiências na TCT. Em especial, destacam-se a discriminação dos itens, fidedignidade dos testes e, por fim, a impossibilidade de comparação de dois indivíduos que fizeram testes distintos. Digamos que queremos comparar as habilidades de dois alunos que tiraram a mesma nota em um teste. Basicamente, queremos determinar qual dos dois possui mais competências. Porém, por mais que eles tenham atingido a

mesma nota, isso não significa que eles tenham alcançado as mesmas competências, pois eles podem ter acertado questões distintas que exigem habilidades diferentes do respondente, por exemplo.

Outro fator que impossibilita a comparação entre indivíduos através da TCT é que para tal comparação estes alunos deveriam ter feito o mesmo teste (RABELO, 2013). Portanto, para uma prova que seja aplicada periodicamente para um público com o mesmo nível de escolaridade, as análises a partir da TCT não seriam precisas, pois não seria possível aplicar a mesma prova.

Apesar das limitações que a TCT apresenta, podemos analisar vários parâmetros importantes que envolvem um teste, como confiabilidade, dificuldade e discriminação. Antes de tudo, devemos analisar a unidimensionalidade do teste, isto é, apenas uma dimensão está sendo medida pelo teste. Por exemplo, se medirmos a proficiência em matemática de modo geral, provavelmente será multidimensional, mas se medirmos a proficiência em geometria do 7º ano do Ensino Fundamental, provavelmente será unidimensional. Segundo Junker (2012) fazemos isso por dois motivos: primeiro, estes itens aumentarão a confiabilidade do teste; e segundo, itens unidimensionais são mais fáceis de descrever e interpretar.

2.1 Discriminação do Item

O poder de discriminação de um item é a sua capacidade de diferenciar com clareza os respondentes que possuem altas habilidades dos que possuem baixas habilidades. Na TCT, Pasquali (2003) diz que dentre as formas existentes para o cálculo do índice de discriminação, a dos grupos-critério e o da correlação item-total são as mais utilizadas pelos psicanalistas. O método dos grupos-critério trabalha com valores de referência, utilizando como base para tais valores os resultados do próprio teste ou resultados externos, enquanto o método da correlação item-total relaciona o escore total do teste com o escore obtido no item.

> *[...] Quando o cálculo do Coeficiente Bisserial é efetuado para cada uma das alternativas, tem-se a correlação da opção de respostas do indivíduo ao item com o seu desempenho no teste como um todo. Assim, espera-se que alunos que se desempenham bem no teste, tenham feito a opção pela alternativa correta de um determinado item. Caso esses alunos tenham sido atraídos a responder qualquer uma das alternativas que não a certa, o item não é discriminativo e não consegue diferenciar os alunos que construíram proficiências, daqueles que as não construíram.* (FERREIRA, 2009, p. 23)

Neste trabalho, para analisar o parâmetro de discriminação dos itens por meio da TCT, vamos utilizar a correlação ponto bisserial.

O coeficiente de correlação ponto bisserial (ρ_{pb}) é dado por um índice que varia no intervalo $[-1,1]$. Quanto mais próximo de 1, mais discriminativo será o item, consequentemente quanto mais próximo de -1, menos discriminativo será o item (MAIA, 2009). Um coeficiente de correlação ponto-bisserial negativo significa que o escore médio dos respondentes que acertaram o item é menor do que o escore médio total, ou seja, os indivíduos com baixas habilidades acertaram o item enquanto indivíduos com altas habilidades erraram o item, o que é indesejável. O coeficiente de correlação ponto bisserial é expresso pela Equação 2.1.

$$\rho_{pb} = \frac{\overline{S_p} - \bar{S}}{\sigma_S} \cdot \sqrt{\frac{p}{q}}, \tag{2.1}$$

em que $\overline{S_p}$ representa o escore médio no teste para os que acertaram o item, $\bar{S}$ é o escore médio para todos, σ_S é o desvio padrão não nulo dos escores obtidos no teste pelos respondentes, p é a proporção de indivíduos que acertaram o item no teste, ou seja, o índice de dificuldade, e q é o complementar de p.

Não podemos deixar de observar que a correlação ponto bisserial não está definida para o caso em que todos os indivíduos acertaram o item, pois teríamos $q = 0$. Segundo Rabelo (2013), itens que apresentam coeficiente de correlação inferiores a 0,30 são considerados de baixa discriminação e devem ser rejeitados. Ensinam Leite (2003), Vianna (1982) e Arias Lloreda e Lloreda (2006) que a escala apresentada, criada por Ebel (1965), é uma boa referência para a classificação da qualidade discriminativa de um item:

Tabela 2.1 - Classificação do item do teste, por nível do índice de discriminação

Índice de discriminação	**Classificação do item**
Abaixo de 0,19	Ineficiente, devendo ser eliminado ou revisado totalmente
Entre 0,20 e 0,29	No mínimo, necessita de revisão.
Entre 0,30 e 0,39	Aceitável, não requerendo revisão
Acima de 0,40	Satisfatório, devendo permanecer no teste

Fonte: INEP 2014

A escala apresentada na tabela 2.1 é uma boa referência para analisar os resultados da prova Jaraguá, identificando os itens que devem ser excluídos do teste.

2.2 Dificuldade do Item

Um parâmetro importante a ser analisado é a dificuldade dos itens que compõem o teste. Na TCT, o índice de dificuldade de um item é dado pela razão entre o número de candidados que responderam corretamente o item e o número total de candidatos que responderam ao teste, Equação 2.2.

$$I_{df} = \frac{n_a}{n_t} \tag{2.2}$$

em que I_{df} representa o índice de dificuldade, n_a representa o número de candidatos que responderam corretamente o item e n_t representa o número total de candidatos. Portanto a dificuldade do item no âmbito da Teoria Clássica dos Testes varia no intervalo [0,1], sendo 0 um item extremamente difícil (nenhum acerto) e 1 um item extremamente fácil (todos os respondentes acertaram este item).

Com base na tabela apresentada por alguns autores para classificar os itens no que diz respeito a dificuldade na TRI (ver Tabela 2.2) e observando a necessidade de uma análise similar utilizando a TCT, criamos uma tabela para classificar os itens do teste no que diz respeito ao seu nível de dificuldade. Esta tabela pode ter uma importância significativa em futuras análises similares às apresentadas neste trabalho, ampliando as discussões e aumentando a confiabilidade dos resultados.

Tabela 2.2 - Classificação e percentual esperado para os índices de dificuldade da TCT

Classificação	Valor	% esperado
Muito fáceis	0,81 ou mais	10%
Fáceis	de 0,61 a 0,80	20%
Medianos	de 0,41 a 0,60	40%
Difíceis	de 0,21 a 0,40	20%
Muito difíceis	até 0,20	10%

Fonte: do autor

Esta classificação, quando satisfeita, representa um teste equilibrado em relação ao nível de dificuldade.

3 A Teoria de Resposta ao Item

Diferentemente da TCT, que leva em consideração o escore bruto do teste, as análises da Teoria de Resposta ao Item se concentram nas respostas a cada item. Como forma de sanar algumas limitações da TCT, utilizam-se atualmente técnicas provenientes da Teoria de Resposta ao Item (TRI), cujo interesse reside em entender a maneira como as pessoas respondem aos itens (REVELLE, 2015). Assim, por mais que duas provas sejam distintas, se os itens forem semelhantes no que diz respeito a nível de dificuldade e habilidade específica, os resultados obtidos pelos indivíduos podem ser comparados.

Apesar de ser considerada atualmente a melhor maneira de avaliar e comparar as habilidades dos estudantes, é importante destacar que o processo de avaliação é muito complexo para ser completamente contemplado apenas com a aplicação de uma prova. Em relação a isso, Machado (1996), diz que "um processo de avaliação nunca se esgota em um processo de medida, porém vai além dele". Mesmo sabendo que temos que aprimorar o modo de avaliar, não podemos confundir essa necessidade de melhorias com frustração por tudo o que é feito.

3.1 Modelos para itens dicotômicos

Nesta seção serão apresentados modelos utilizados para análise de itens dicotômicos, isto é, itens que possuem duas possibilidades de correção, certo ou errado. Na prática, os modelos logísticos para itens dicotômicos são os modelos de resposta ao item mais utilizados, sendo que há basicamente três tipos: modelo logístico de um parâmetro, modelo logístico de dois parâmetros e modelo logístico de três parâmetros (ANDRADE; TAVARES; VALLE, 2000).

O modelo logístico de dois parâmetros (dificuldade e discriminação do item, modelo ML2) foi o primeiro a ser utilizado, já no final da década de 50 do século passado. A equação para este modelo é dada pela Equação 3.1.

$$P(\theta) = \frac{1}{1 + e^{-a(\theta - b)}}, \tag{3.1}$$

na qual e é o número de Euler, cujo valor é aproximadamente 2,718, b é o parâmetro de dificuldade do item, a é o parâmetro de discriminação do item e θ é o nível de habilidade.

Na década de 60, surgiu o modelo logístico de um parâmetro (modelo ML1), proposto pelo matemático Georg Rasch. Neste modelo, o parâmetro de discriminação tem seu valor fixado em $a = 1{,}0$, para todos os itens, enquanto o parâmetro de dificuldade assume diferentes valores dependendo do item (BAKER e KIM, 2004). O parâmetro de discriminação é dado pela Equação 3.2.

$$P(\theta) = \frac{1}{1 + e^{(b-\theta)}}, \tag{3.2}$$

em que e é o número de Euler, cujo valor é aproximadamente 2,718, b é o parâmetro de dificuldade do item e θ é o nível de habilidade

Criado em 1968 por Allan Birnbaum e utilizado até os dias de hoje, o modelo logístico de três parâmetros (modelo ML3) leva em consideração o acerto ao acaso. Representado pela variável c, o parâmetro que representa o chamado "chute", até então não era considerado. O parâmetro c representa a probabilidade de um aluno com baixa habilidade responder corretamente o item. Este parâmetro também é conhecido como a probabilidade de acerto ao acaso (ANDRADE; TAVARES; VALLE, 2000).

A Equação 3.3 é dada para o modelo 3LP.

$$P(\theta) = c + (1 - c)\frac{1}{1 + e^{-a(b-\theta)}}, \tag{3.3}$$

em que e é o número de Euler, cujo valor é aproximadamente 2,718, b é o parâmetro de dificuldade do item, a é o parâmetro de discriminação do item e θ é o nível de habilidade e c representa o parâmetro de acerto ao acaso.

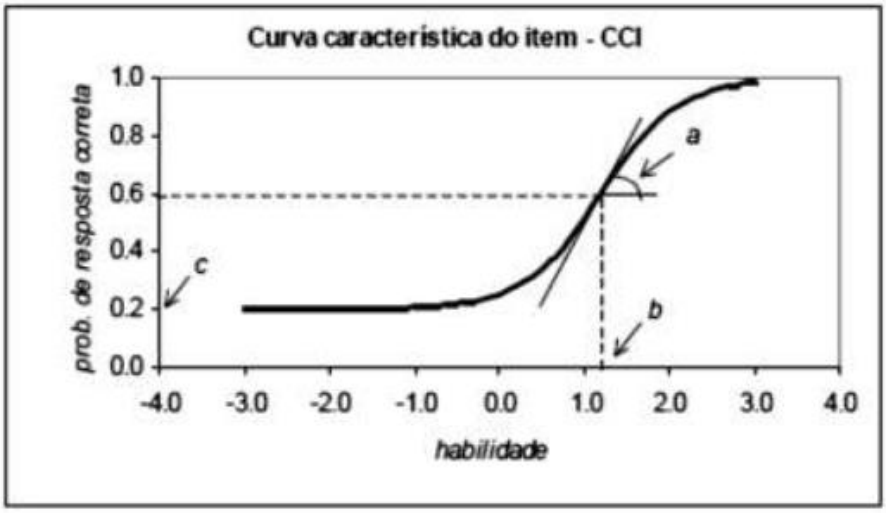
Quando um indivíduo responde aos itens de um teste gera uma série de valores iguais a 1 (no caso de acerto) ou 0 (no caso de erro). Podemos pensar nestes dados na forma de uma tabela com *j* linhas, referente a quantidade de respondentes e *i* colunas, referentes a quantidade de itens. No caso da TRI, deseja-se descobrir qual o valor do traço latente (de habilidade) do indivíduo que melhor explica o acerto ou o erro em cada item individualmente (RABELO, 2013). Para fazer isso, a pergunta a ser respondida é: qual é a probabilidade o j-ésimo acertar o i-ésimo item? A resposta para essa pergunta está relacionada ao nível de habilidade θ que o indivíduo possui e dos parâmetros do modelo ML3, definida pela Equação 3.4.

$$P(X_{ij} = 1/\theta_j) = c_i + \frac{1 - c_i}{1 + \exp[-Da_i(\theta_j - b_i)]}, \qquad (3.4)$$

em que X_{ji} é a resposta do indivíduo *j* ao item *i* (igual a 1 se o indivíduo responde corretamente ao item e, igual a 0, caso contrário), $a_i > 0$ é o parâmetro de discriminação do item *i*, b_i é o parâmetro de dificuldade do item *i*, $0 < c_i < 1$ é o parâmetro da assíntota inferior do item, ou seja, a chance de um respondente com baixa habilidade responder corretamente o item *i*, θ_j representa o traço latente (habilidade) do j-ésimo indivíduo e D é um valor de escala, que é igual a 1 na métrica logística e igual a 1,7 na métrica normal. Segundo Klein (2013) o uso da métrica normal vem do fato de que os primeiros modelos utilizavam a função ogiva normal e de que a função de distribuição cumulativa normal com média 0 e desvio padrão 1 é bem aproximada pela função logística com parâmetro $b = 0$ e parâmetro $a = 1{,}7$, no sentido de que o máximo da diferença pontual entre as duas funções é menor do que 0,01.

Quando estimamos os valores que a função $P(\theta)$ assume para o i-ésimo item, se os resultados estiverem dentro do esperado para termos um item cumprindo bem a sua função avaliadora, o gráfico é uma sigmoide chamada Curva Característica do Item (CCI), ilustrada na Figura 3.1. A função $P(\theta)$ assume, no eixo vertical, valores no intervalo (0,1), que representam a probabilidade de acerto de 0% a 100%. No eixo horizontal a habilidade θ assume valores que estão em uma escala de média igual a 0 e um desvio padrão igual a 1.

Figura 3.1 - Curva Característica do Item - CCI



Fonte: Andrade, Tavares e Valle (2000, p.11)

Vale mencionar que existem estudos de um modelo logístico de quatro parâmetros (ML4). Além de discriminação, dificuldade e acerto ao acaso, o modelo logístico de quatro parâmetros leva em consideração as possíveis falhas na elaboração de um item diante de uma resposta errada de um respondente com altas habilidades. Segundo Muñiz (1990) alguns autores propõem um modelo logístico de quatro parâmetros, que visa controlar circunstâncias aleatórias relacionadas com falhas do construtor no momento da elaboração dos itens. Desde o final da década de 90 este modelo é pouco usado, Muñiz (1997) explica que o mesmo não apresenta vantagens significativas comparativamente aos outros três modelos e, ademais, os problemas que trata de solucionar podem ser muito bem controlados durante a elaboração dos itens.

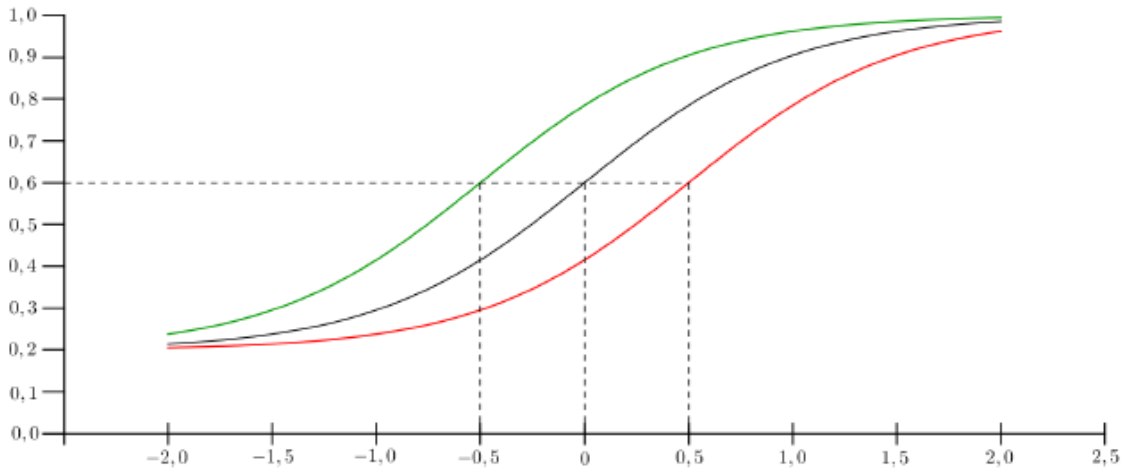
3.2 Dificuldade do item

O parâmetro b (dificuldade do item) na TRI representa o nível de conhecimento (habilidade) que o indivíduo deve possuir para responder a um item. De uma forma mais específica, a dificuldade é o valor da habilidade θ necessário para que se tenha uma probabilidade de acerto igual a $\frac{(1+c)}{2}$. Obtemos tal resultado observando o valor de θ no ponto de interseção entre a CCI e a reta horizontal que passa pelo ponto $\frac{(1+c)}{2}$.

Pelo fato da dificuldade ser medida em uma escala padronizada, seus valores podem variar de $-\infty$ a $+\infty$, porém, na prática este intervalo se reduz a -3 (item muito fácil) e $+3$ (item muito difícil), pois esta escala abrange mais de 99% das ocorrências (TORRES, 2015).

Na Figura 3.2 temos as curvas de três itens. A probabilidade de acerto ao acaso é $c = 0{,}2$, ficando a reta horizontal em $0{,}6$. Podemos ver que a curva em verde representa um item mais fácil ($b = -0{,}5$), a curva em preto representa um item com uma dificuldade média ($b = 0$) e a curva em vermelho representa um item mais difícil ($b = 0{,}5$).

Figura 3.2 - Comparação das CCIs de itens com diferentes níveis de dificuldade e mesma probabilidade de acerto ao acaso.



Fonte: Torres (2015)

Quando não é considerado o acerto ao acaso (modelo ML2) ou simplesmente temos $c = 0$ (não considerar acerto de item por "chute"), a dificuldade é dada pelo valor da habilidade que gera uma probabilidade de 50% de acerto do item (MAIA, 2009).

A Tabela 3.1 representa a distribuição e a classificação adotada pela maioria dos autores da área de avaliação e psicometria, de acordo com a dificuldade dos itens (RABELO, 2013).

Tabela 3.1 - Classificação e percentual esperado para os índices de dificuldade da TRI

Classificação	**Valores de *b***	**% esperado**
Muito fáceis	até -1,28	10%
Fáceis	de -1,27 a -0,52	20%
Medianos	de -0,51 a 0,51	40%
Difíceis	de 0,51 a 1,27	20%
Muito difíceis	1,28 ou mais	10%

Fonte: Rabelo (2013)

Essa classificação pode ser utilizada para determinar o equilíbrio entre o nível de dificuldade das questões, de modo que o teste não seja muito difícil e nem muito fácil.

3.3 Discriminação do item

Quando falamos da discriminação de um item nos referimos a capacidade do item de diferenciar respondentes com alta habilidade de respondentes com baixa habilidade. Na TRI,

podemos observar a discriminação do item através da inclinação da CCI no ponto de inflexão (ponto em que a curva muda de concavidade), representada pelo parâmetro a. Vimos anteriormente que para determinar o ponto de inflexão fazemos $\frac{1+c}{2}$. Teoricamente, o parâmetro de discriminação varia no intervalo $(-\infty, +\infty)$, porém na prática os valores de a variam no intervalo $(0,2)$. Segundo Rabelo (2013) valores negativos da discriminação indicam que o item se comporta de uma maneira estranha, pois estariam indicando que a probabilidade de acerto do item diminui com o aumento da aptidão do sujeito. Em geral, são considerados como discriminativos itens com valores de a superiores a 0,70. De acordo com Rabelo (2013) podemos categorizar os itens de acordo com a Tabela 3.2.

Tabela 3.2 - Classificação do item de acordo com a discriminação pela TRI.

Valores	**Discriminação**
a = 0	nenhuma
0 < a ≤ 0,35	muito baixa
0,35 < a ≤ 0,65	Baixa
0,65 < a ≤ 1,35	moderada
1,35 < a ≤ 1,70	Alta
a > 1,70	muito alta

Fonte: Rabelo (2013)

A Tabela 3.2 servirá de base para a classificação dos itens de acordo com a sua discriminação, o que será feito no capítulo 6. Estes resultados serão confrontados com os dados obtidos na análise de discriminação da TCT.

3.4 Acerto ao acaso

Representado pelo parâmetro c, o acerto ao acaso representa as respostas corretas dadas aleatoriamente, o chamado "chute". Em relação à CCI, o parâmetro c é dado pelo ponto em que a assíntota inferior ao gráfico intercepta o eixo das probabilidades. Em uma prova em que cada item possua quatro alternativas, espera-se valores ligeiramente inferiores a 0,25 para o acerto ao acaso Rabelo (2013). Caso o valor seja muito superior a 0,25, isso indica que a resposta correta atrai tanto os respondentes com altas habilidades quanto os respondentes com baixas habilidades, ou seja, o parâmetro c é o mesmo independentemente da habilidade do respondente.

3.5 Critério dos distratores

Outro componente importante a ser considerado ao avaliar um teste é o estudo dos distratores. Esta análise complementa os parâmetros de dificuldade e discriminação do item.

Conforme Urbina (2007), para que se tenha um item de múltipla escolha ideal, a alternativa correta deste item deve ser óbvia para o testando que possui as habilidades necessárias para resolvê-lo, e os distratores parecem igualmente plausíveis para aqueles que não possuem as habilidades necessárias para resolver o item. Portanto elaborar um item com bons distratores não é uma tarefa simples, pois há uma linha tênue entre um bom distrator e uma "pegadinha", ou seja, uma alternativa incorreta que atrai os testandos que possuem o conhecimento necessário para resolver o item (TORRES, 2015).

Segundo Urbina (2007), depois da aplicação de um teste, deve-se realizar uma análise dos distratores, começando pela quantidade de testandos que selecionou cada distrator. Para a autora, "o exame cuidadoso da frequência com que os vários distratores foram escolhidos por testandos de diferentes níveis de habilidade serve para detectar possíveis falhas nos itens". Desta forma, seguindo a orientação de Urbina, faremos um estudo dos distratores da prova Jaraguá.

3.6 Confiabilidade do Teste

Um fator importante na elaboração de um teste é sua confiabilidade. Vários fatores influenciam nos resultados de um teste, como a motivação dos respondentes, condições do local em que está sendo aplicado, clima, entre outros. Um teste totalmente confiável seria aquele que não possui erros sistemáticos, ou seja, erros que podem ser corrigidos após serem detectados. Por exemplo: se foi detectado que um termômetro mede 1º C a mais em todas as suas medições, basta corrigir esta diferença no momento de apresentar os resultados.

Segundo Vianna (1982, p.157-160), são vários os fatores que afetam a fidedignidade de um teste e podem ser relacionados ao próprio teste ou ao examinado. Com relação ao teste, ele argumenta que;

i) quanto maior o número de itens, maior a fidedignidade;
ii) quanto menor a amplitude da dificuldade dos itens, maior a fidedignidade;
iii) quanto maior a interdependência dos itens, menor a fidedignidade;
iv) quanto mais objetiva a correção, maior a fidedignidade;
v) quanto mais homogêneo o teste, maior a fidedignidade; e

vi) quanto maior a introdução de elementos estranhos e/ou capciosos no teste, menor a sua fidedignidade.

3.7 Alfa de Cronbach

Diante de tantos fatores influenciáveis na confiabilidade de um teste, podemos nos perguntar: como obter um índice aceitável de confiabilidade? Uma das maneiras de responder essa pergunta e que também servirá de base para os estudos deste trabalho é calcular o alfa de Cronbach (α).

Para Anjos e Andrade (2012, p. 9) o coeficiente alfa de Cronbach é utilizado para medir a consistência interna do instrumento de medida. Proposto por Cronbach (1951), é o que gera o menor valor, considerado como limite inferior dos coeficientes de fidedignidade de um teste, conforme Arias, Lloreda& Lloreda (2006, p. 54). Para Muñiz (2003, p. 54), mais que a estabilidade das medidas, o coeficiente α reflete o grau em que covariam os itens que constituem o teste, sendo, portanto, um ótimo indicador de sua consistência interna, cuja estatística é dada pela Equação 3.5.

$$\alpha = \frac{n}{n-1}\left(1 - \frac{\sum_{i=1}^{n}\sigma_i^2}{\sigma_X^2}\right), n \neq 1 \; e \; \sigma_X^2 \neq 0 \qquad (3.5)$$

em que σ_i^2 representa a variância do item i $(i = 1, 2, \dots, n)$ e, σ_X^2, a variância total dos escores do teste.

Conforme σ_i^2 diminui, ou seja, conforme a variância entre os itens diminui, maior será valor de α, o que implica uma maior consistência interna, aumentando a confiabilidade do teste. O valor de α varia no intervalo [0,1]. Segundo Vianna (1982) uma fidedignidade mínima de 0,70 é considerada aceitável para fins de decisão.

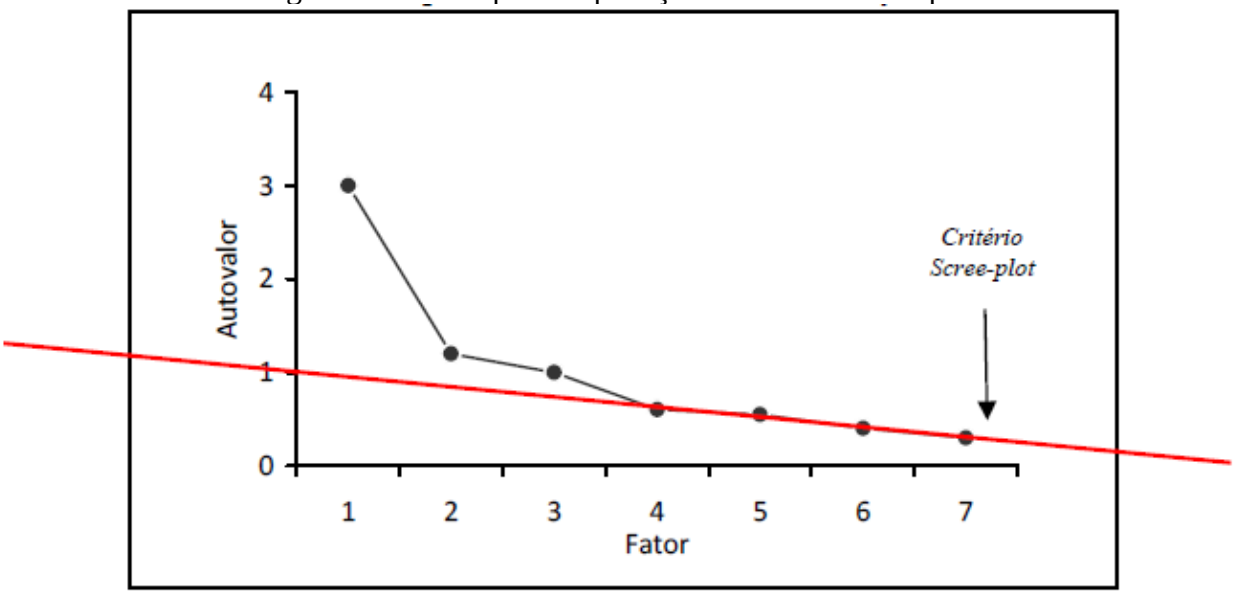
3.8 Unidimensionalidade do teste

Analisar a unidimensionalidade do teste significa analisar se apenas uma habilidade está sendo medida pelo teste. Por exemplo, um teste de matemática que contém alguns itens que são estritamente computacionais e outro que envolve material verbal provavelmente não são unidimensionais (KOLEN; BRENNAN, 2013). Segundo Junker (2012) fazemos isso por dois motivos: primeiro, estes itens aumentam a confiabilidade do teste; e segundo, um teste composto por itens unidimensionais são mais fáceis de expor e explicar.

Existem vários métodos para determinar a dimensionalidade de um teste. Pasquali (2003), comenta que os métodos que apresentam maiores propriedades estão baseados na análise fatorial e no traço latente (TRI). Para Andriola (2009), os seguintes métodos merecem destaque na literatura mundial: o procedimento de Bejar; o contraste de Gustaffson; o método de McDonald; o contraste Q1 e Q2 de Van den Wollenber; a análise de procedência modificada; o método Hattie para a comparação de autovalores reais e simulados; e o método da equação de regressão.

Com fundamento na análise fatorial, neste trabalho vamos utilizar o método das cargas fatoriais para estudar a dimensionalidade do item. Carga fatorial é a correlação entre a variável e o fator. A carga ao quadrado é a quantia de variância total da variável explicada pelo fator (Hair et. al., 2005, p.109). Conforme Pasquali (2003), um critério mínimo para que um item seja unidimensional é uma carga fatorial maior ou igual do que 0,30. Também analisamos a dimensionalidade do teste por meio do método scree-plot. Idealizado por R. B. Cattel em 1996, o método scree-plot é utilizado para identificar o número de fatores a ser extraído com base na representação gráfica dos autovalores da matriz. Para Andriola (2009) o procedimento consiste em traçar uma reta paralela aos fatores que possuem autovalores mais baixos, até que a mesma "corte" o eixo das ordenadas. São detidos tantos fatores quanto o número de autovalores que estejam na parte superior da reta (Figura 3.3).

Figura 3.3 - Exemplo da aplicação do método scree-plot



Fonte: Adaptado de Hair. et. al. (2005)

No exemplo da Figura 3.3 temos um teste considerado multidimensional, pois há 3 fatores com autovalores na parte superior da reta. Este teste também é conhecido como “regra do cotovelo”, o que significa que se conseguimos enxergar um “cotovelo” ao realizar o procedimento de traçar a reta, isso indica que apenas um fator possui autovalor acima desta reta, ou seja, o teste pode ser considerado unidimensional.

4 Revisão de Literatura

Neste capítulo apresentamos os resultados de alguns trabalhos que realizaram análises similares às que propomos neste trabalho, considerando a confiabilidade do teste, discriminação dos itens, nível de dificuldade dos itens, análise dos distratores e acerto ao acaso. Estes critérios vão servir de base para analisarmos a qualidade de um item e se ele deve ou não permanecer no teste.

Um dos trabalhos que nos auxiliou nas análises foi Diferenças nas Realizações Matemáticas de Acordo com a Oportunidade para Aprender (OPA): Um Estudo Utilizando um Modelo de Quatro Parâmetros da Teoria de Resposta ao Item, (BARNARD-BRAK *et al*, 2018).

Neste trabalho foi analisado o desempenho dos estudantes nos itens do Programa Internacional de Avaliação dos Estudantes (PISA) em 2012, nos Estados Unidos. Para este estudo foi utilizado um modelo de quatro parâmetros em que o quarto parâmetro (d) corresponde ao "descuido", ou seja, uma explicação para que um estudante com bom desempenho tenha respondido de forma errada a um determinado item.

O objetivo desta pesquisa é saber se a probabilidade de um estudante com alta proficiência responder errado um determinado item está relacionada com o fato de este estudante ter recebido mais ou menos oportunidades de aprender (OPA).

O resultado foi uma pequena, porém estatisticamente significativa relação entre OPA e o desempenho em testes de matemática. Com o aumento da oportunidade para aprender dos estudantes, o seu desempenho em matemática aumentou também.

Já na tese de doutorado de Maia (2009), Uso da Teoria Clássica dos Testes – TCT e da Teoria de Resposta ao Item – TRI na Avaliação da Qualidade Métrica de Testes de Seleção, foi utilizada a TCT como instrumento na avaliação da qualidade métrica da prova do concrso vestibular da Universidade Estadual do Ceará – UECE, de 2007, envolvendo as disciplinas de Português e Matemática. Participaram 20.016 candidatos a 38 cursos de graduação, que deveriam responder 14 questões de Português e 10 questões de Matemática.

De maneira geral, a prova de matemática que foi analisada neste trabalho apresenta um índice médio de dificuldade, tanto pela TCT quanto pela TRI. Além disso, tanto o item considerado mais fácil (item 03) quanto o item considerado mais difícil (item 06) foram os mesmos em ambas as teorias.

No que diz respeito a discriminação dos itens, os resultados apresentados pela TRI foram mais significativos do que os ocasionados na TCT. Num intervalo prático de 0 a 3, os resultados se mostraram variando de 1,418 para o item 06 a 2,603, para o item 04. Assim, pela TRI, foi concluído pelo autor que essa prova se mostrou com um ótimo comportamento discriminativo.

A análise pedagógica dos itens considerados o mais fácil e o mais difícil da prova, também serviu de referência para as nossas análises. Dentre estas análises destacam-se as sugestões para mudanças no enunciado que o tornariam mais claro, e discussão das estratégias de resolução do item.

Outro trabalho com proposta similar à nossa foi feito por Silva (2015), intitulado Teoria de Resposta ao Item – TRI em Avaliações de Matemática na EEM Professor Gabriel Epifânio dos Reis.

Este trabalho teve como foco principal a avaliação, discutindo o ato de avaliar e os instrumentos que podem ser utilizados para que os resultados atingidos sejam os esperados. De uma forma mais específica, neste trabalho priorizou-se como objetivo analisar a aplicação da Teoria de resposta ao Item (TRI) em avaliações de Matemática do Ensino Médio da escola Gabriel Epifânio.

Para um público de 61 estudantes da 1ª série do Ensino Médio foi aplicada uma prova com 9 itens objetivos, todos dicotômicos, ou seja, com respostas classificadas em certas ou erradas. As notas dos estudantes foram calculadas inicialmente utilizando a TCT, proporcional ao número de acertos, tendo cada questão o mesmo peso e depois recalculada através do modelo logístico unidimensional de 3 parâmetros (ML3) da TRI em que foram utilizados os parâmetros de dificuldade *b*, discriminação *a*, e acerto ao acaso *c*.

O software utilizado para estimar os resultados foi o ICL - (IRT Command Language). Segundo Mendonça (2012), este é um software criado por Brad Hanson e que pode fazer estimativas dos parâmetros dos modelos logísticos 1, 2 e 3 parâmetros de itens dicotômicos.

Analisando as curvas característica dos itens o autor apresenta os itens 1 e 8, considerados o mais difícil e o mais fácil, respectivamente. O item 1 exigia do respondente conhecimentos de Semelhança de Triângulos, enquanto o item 8 exigia conhecimentos sobre a influência dos coeficientes no gráfico de uma equação do 2º grau.

Observamos uma comparação feita pelo autor entre os resultados obtidos na TCT e na TRI. De modo geral, neste trabalho, as notas dos estudantes na TRI foram maiores do que na TCT,

enquanto que em outras situações as notas na TCT são maiores. Um gráfico de dispersão foi utilizado para comparar os resultados dos estudantes nas duas teorias, apontanto, por exemplo, que em algumas situações, dois indivíduos atingiram a mesma nota na TCT (acertaram o mesmo número de questões), mas quando o método avaliativo foi a TRI, atingiram notas distintas.

Na busca por trabalhos que utilizassem a aplicação de simulados, encontramos o trabalho de Tôrres (2015): Uma aplicação da Teoria de Resposta ao Item em um Simulado de Matemática no Modelo ENEM. Além da TRI, a análise dos itens também foi feita utilizado a Teoria Clássica dos Testes, com o objetivo de fornecer um *feedback* aos professores elaboradores do simulado e possibilitar a criação de um banco de itens na escola.

O simulado contendo 45 itens da área de Matemática foi aplicado para 165 estudantes do Ensino Médio e pré-vestibular de um colégio do Distrito federal, respeitando as mesmas condições de aplicação da prova propostas pelo ENEM. Para a análise destes itens utilizou-se o programa R, após os dados estarem tabelados no Excel.

A análise dos itens foi feita em quatro subseções: análise dos itens deficientes, itens sujeitos a reelaboração, itens sujeitos a aprimoramento e itens bons. Apresentamos algumas das principais informações de cada uma dessas subseções.

Na análise de itens deficientes foram descartados quatro itens, sendo três deles por não terem atingido o valor mínimo de 20% no índice de discriminação proposto por Ebel (1965). O quarto item (item 145) descartado foi um pouco mais curioso. Se este item fosse avaliado apenas pela TRI, teria resultado satisfatório, com os parâmetros $a = 1{,}737, b = 2{,}142$ e $c = 0{,}179$, sendo considerado um item muito difícil. Porém, além da discriminação ter sido igual a 19% na TCT, duas das alternativas incorretas apresentam um considerável bisserial positivo, logo há um indicativo de que o item não foi bem formulado, pelo fato destas alternativas serem marcadas com mais frequência por estudantes com bom desempenho.

Em relação aos itens sujeitos a reelaboração foram classificados 7 itens. O autor sugere mudanças no enunciado das questões além de aplicações em um grupo maior de estudantes, para verificar se o comportamento dos distratores se mantém. O autor chama a atenção para o item 166 em que uma simples diferença no arredondamento do π para 3,14 ou 3 faria com que a resposta do estudante fosse diferente. Como este arredondamento não foi esclarecido no enunciado da questão, este provavelmente foi o motivo pelo qual o item não apresentou um resultado melhor.

No grupo dos itens sujeitos a aprimoramento apareceram 8 questões e no grupo dos itens considerados bons foram classificadas 26 questões. Para os itens sujeitos a aprimoramento o autor também sugere mudanças no enuciado das questões além de sugerir novas aplicações para grupos maiores.

Os trabalhos citados, juntamente com as suas referências, auxiliaram muito em nossas certezas provisórias e dúvidas temporárias. Tanto para o embasamento teórico quanto para as análises e discussões dos resultados, estudar o que foi realizado nestes trabalhos, incluindo limitações e sugestões, foi de fundamental importância na realização da nossa pesquisa.

5 Metodologia

Neste capítulo falamos sobre a cidade de Jaraguá do Sul, mencionando algumas características relevantes no que diz respeito a educação. Além disso, apresentamos a prova Jaraguá, como essa prova é aplicada, o público envolvido, além dos procedimentos que serão utilizados para avalia-la.

Situado no norte do estado de Santa Catarina, o município de Jaraguá do Sul, segundo o Instituto Brasileiro de Geografia Estatística – IBGE, possui 170.835 habitantes. Estes dados são referentes à ultima publicação do IBGE em 30 de agosto de 2017. Segundo o próprio IBGE o município de Jaraguá do Sul possui uma taxa de escolarização de 98,3% entre as pessoas de 6 a 14 anos de idade. Além disso, no último IDEB, referente ao ano de 2015, o município atingiu a nota 6,9 para os anos iniciais e 5,6 para os anos finais, sendo superiores aos índices nacionais de 5,3 e 3,8 para os anos iniciais e finais, respectivamente (INEP, 2018). A Figura 5.1 mostra a localização do município em relação ao mapa do Brasil.

Figura 5.1 - Localização do município de Jaraguá do Sul

Assim como existe uma preocupação individual de cada professor, no município de Jaraguá do Sul, sempre houve uma preocupação muito grande com a qualidade do ensino. A feira municipal de matemática e a feira municipal de ciências e tecnologia (Figura 5.2) são exemplos de ações em prol de uma Educação melhor. Porém, para elaborar ações mais pontuais nas escolas necessita-se, primeiramente, saber qual a real situação de cada Unidade de Ensino e seus respectivos estudantes. Para isso, a rede municipal de ensino de Jaraguá do Sul iniciou em 2014 um programa de simulados

bienais com o intuito de criar, futuramente, um índice de cada uma das disciplinas curriculares para o município. É claro que a ideia é identificar os problemas e apontar possíveis soluções para estes problemas, de modo geral ou para determinado local específico.

Figura 5.2 - VI Feira Municipal de Educação Matemática de Jaraguá do Sul, e Feira Municipal Científica e Tecnológica (FECITEC) 2017

Fonte: https://jdv.com.br

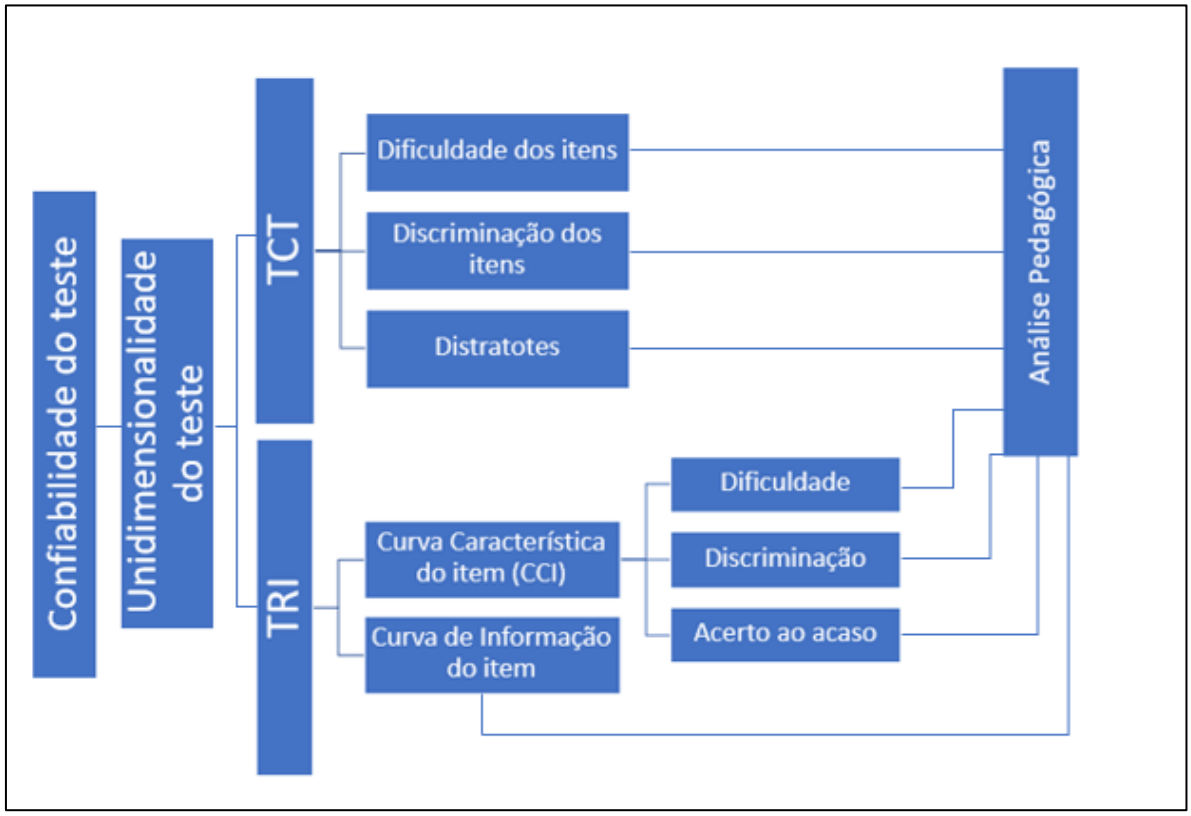
Aplicada pela primeira vez em 2015, a prova Jaraguá abrangeu todas as 27 escolas do município de Jaraguá do Sul, para as turmas de 6º ao 9º ano da Educação Básica. A 2ª edição da prova está programada para este ano (2018). Fazem a prova aproximadamente 10.000 estudantes da rede pública municipal, compreendidos em uma faixa etária de 10 a 15 anos. A prova é formada pelas 9 disciplinas do currículo escolar: Português, Matemática, Ciências, História, Geografia, Ensino Religioso, Artes, Inglês e Educação Física. As provas são formadas por 10 questões de cada disciplina. Os estudantes realizam estas provas em três dias consecutivos, sendo três disciplinas por dia. É disponibilizado aos estudantes o período de 3 horas/aula (135 minutos) para que concluam as 30 questões de cada dia. A elaboração desta prova envolveu uma equipe de professores da Secretaria Municipal de Educação que recebeu um treinamento na Universidade Federal de Santa Catarina com o professor Dalton Francisco de Andrade.

Para a realização deste trabalho foram levadas em consideração as respostas de 300 estudantes das turmas do 8º ano que responderam um teste com 10 questões referente à disciplina de matemática no ano de 2015. Por ser um número muito grande de estudantes, a Secretaria Municipal

da Educação (SEMED) tomou esta amostra de 300 alunos para as devidas análises, tais como índice por escola, índice por turma, índice do município, etc. Assim, com a disponibilização destes dados pela SEMED, realizamos todo o trabalho.

A Figura 5.3 mostra um fluxograma discriminando passo a passo as análises que serão feitas neste trabalho.

Figura 5.3 - Análises realizadas



Estas análises foram feitas utilizando a TCT e a TRI aplicando o modelo logístico de três parâmetros. Como o INEP (2014) desconsidera a inclusão de questões que possuem uma correlação ponto-bisserial inferior a 0,20, adotamos este critério para a exclusão dos ites deficientes da prova Jaraguá, tanto das análises pela TCT quanto pela TRI. Em relação a análise pedagógica dos itens tentamos apontar alguns fatores que possam ter influenciado o baixo nível de correlação. A obtenção dos resultados se deu por meio do software versão 3.4.2 (R CORE TEAM, 2017), com os pacotes mirt versão 1.28 (CHALMERS, 2012), CTT versão 2.3.2 (WILLSE, 2017), psych 1.8.4 (REVELLE, 2017) e ltm versão 1.1-1 (RIZOPOULOS, 2006). O conjunto de dados utilizado para as análises e a rotina estão nos anexos II e III, respectivamente.

6 Resultados e Discussão

Neste capítulo vamos apresentar os resultados referentes às duas teorias de estudo deste trabalho aplicadas na prova Jaraguá. Estes resultados vão servir de base para a análise estatística e pedagógica.

6.1 Algumas estatísticas descritivas

Nas tabelas e figuras a seguir apresentam-se os resultados obtidos e comentados através das teorias TCT e TRI.

Iniciamos apresentando a Tabela 6.1. Esta tabela contém os valores referentes a quantidade de acertos das questões, ou seja, quantos estudantes acertaram apenas uma questão, quantos estudantes acertaram duas questões, e assim sucessivamente.

Tabela 6.1 - Frequência das pontuações totais

Quantidade de questões	0	1	2	3	4	5	6	7	8	9	10
Frequência	1	7	21	55	65	57	43	28	17	5	1

Fonte: Do autor.

Apenas um dos alunos respondeu corretamente todas as questões. Da mesma maneira, apenas um dos alunos não acertou nenhuma das questões. Além disso, a maior parte dos alunos acertou cinco questões ou menos.

Na Tabela 6.2 podemos ver uma representação dos itens no que diz respeito as quantidades de erros e acertos.

Tabela 6.2 - Proporções para cada nível de resposta.

Item	Erros	Acertos
1	0,1967	0,7967
2	0,4933	0,5033
3	0,3967	0,6033
4	0,6667	0,3300
5	0,7800	0,2200
6	0,7267	0,2700
7	0,2133	0,7800
8	0,3533	0,6400
9	0,6000	0,3967
10	0,8600	0,1400

Fonte: do autor.

Na Tabela 6.3 temos os valores de outras estatísticas descritivas da prova.

Tabela 6.3 - Outras estatísticas descritivas

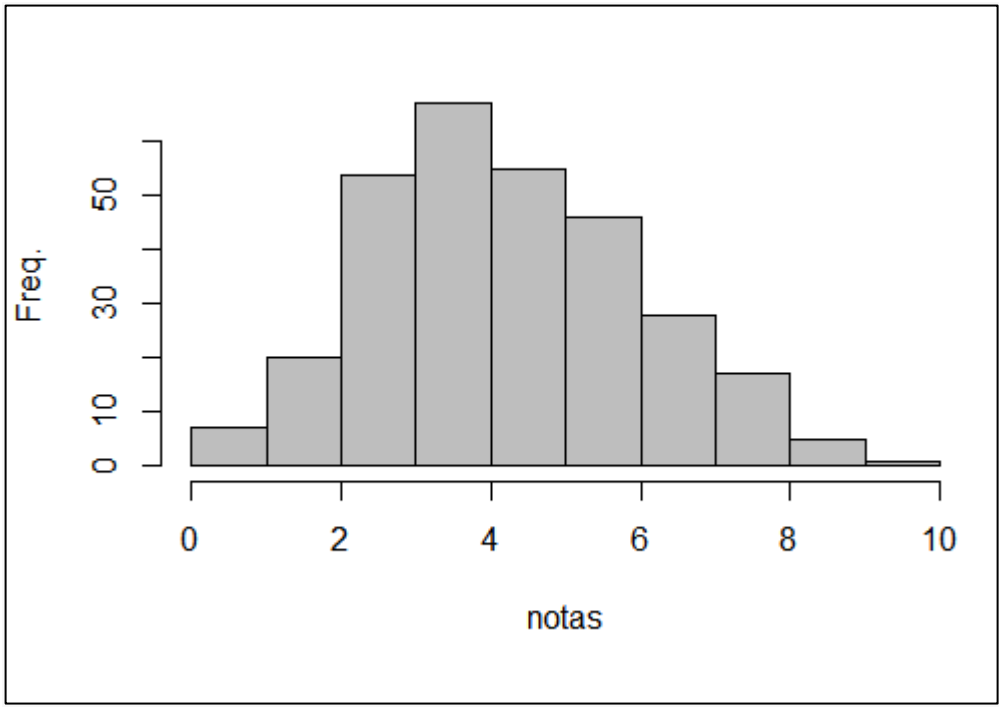
Função	Valor
Média	4,71
Moda	4,00
Mediana	5,00
Q1	3,00
Q3	6,00
Nota Máxima	10,00
Nota Mínima	0,00
Variância	3,26
Desvio Padrão	1,80

Fonte: Do autor.

Como as notas variam de 0 a 10, o valor igual a 6 para o terceiro quartil (Q3) monstrou que a maior parte das notas altas estão mais próximas de 5 do que de 10, o que já era de se esperar após analisarmos os resultados da tabela de frequência (Tabela 6.1).

Na Figura 6.1 apresentamos um histograma das notas dos alunos.

Figura 6.1 - Histograma das notas dos alunos



Fonte: do autor.

De acordo com o histograma (Figura 6.1) podemos ver que concentração de notas baixas é maior do que a concentração de notas altas. Podemos ver também que grande parte dos estudantes atingiu notas entre 3 e 6.

6.2 Confiabilidade do teste

Uma das formas de avaliar a confiabilidade do teste é através do valor do alfa de Cronbach, que foi igual a 0,429. Na Tabela 6.4 temos a alteração no valor do alfa de Cronbach mediante a exclusão de um determinado item.

Tabela 6.4 - Consequência no valor do alfa de Cronbach com a exclusão de cada item do teste.

Excluindo	Valor
Item 1	0,4040
Item 2	0,4515
Item 3	0,3582
Item 4	0,3918
Item 5	0,4353
Item 6	0,4210
Item 7	0,3906
Item 8	0,3871
Item 9	0,3845
Item 10	0,3984

Fonte: Do autor

Constatamos que a retirada dos itens 2 e 5 implicaria em uma pequena melhora do valor do alfa de Cronbach. Apesar de melhorar o coeficiente α, o mesmo ainda não atingiria o nível aceitável de 0,70.

Na Tabela 6.5 estão os resultados inerentes a confiabilidade do teste utilizado como base deste estudo, levando em consideração correlação ponto-bisserial.

Tabela 6.5 - Correlação ponto-bisserial.

Item	Correlação ponto-bisserial
1	0,23925079
2	0,06158151
3	0,35052307
4	0,25559223
5	0,10074721
6	0,15913489
7	0,28967397
8	0,26803129
9	0,27147925
10	0,29806435

Fonte: Do autor.

Os itens 1, 4, 7, 8 e 9 necessitam de revisão, pois possuem correlação ponto-bisserial inferior a 0,29, índice mínimo conforme orienta Pasquali (2003). Além disso, os itens que atingiram o valor mínimo da correlação ponto-bisserial estão muito próximos do limite, o que mostra uma baixa consistência dos itens.

Na Tabela 6.6 estão os dados referentes a correlação Ponto Bisserial com a pontuação total dos escores, o que nos permite fazer uma análise sobre quais itens podem ser excluídos para obter um teste com maior poder de discriminação.

Tabela 6.6 - Correlação Ponto Bisserial com Escores Total

Item	Incluído	Excluído
1	0,3930	0,184
2	0,3304	0,0593
3	0,5156	0,2773
4	0,4421	0,2012
5	0,2963	0,0721
6	0,3581	0,1217
7	0,4233	0,2115
8	0,4665	0,2239
9	0,4526	0,2024
10	0,3765	0,1971

Fonte: Do autor

A exclusão dos ites não implica em melhoras significativas dos coeficientes de correlação ponto-bisserial, o que pode ser explicado pela pequena quantidade de itens no teste.

6.3 Unidimensionalidade do teste

Vamos utilizar dois métodos para determinar a dimensionalidade do teste: o método das cargas fatoriais e o método scree-plot. Iniciamos com a Tabela 6.7 que apresenta as cargas fatoriais dos itens.

Tabela 6.7 - Cargas fatoriais

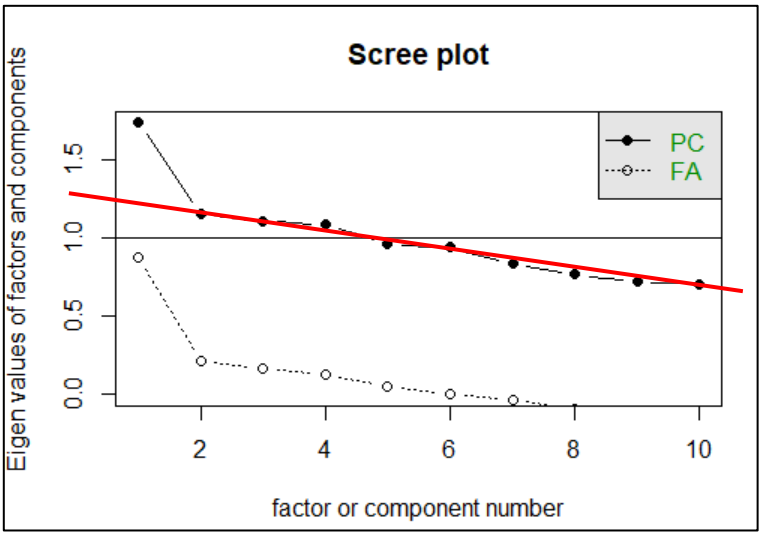
Item	F1
1	0,800
2	0,364
3	0,930
4	0,973
5	0,964
6	0,659
7	0,466
8	0,703
9	0,707
10	0,589

Fonte: Do autor

Conforme apresentado na Tabela 6.7, percebemos que a carga fatorial dos itens 2 e 7 são as mais baixas, porém ultrapassam o valor mínimo de 0,30, critério mínimo de unidimensionalidade proposto por Pasquali (2003).

Na Figura 6.2 apresentamos o gráfico scree-plot da prova Jaraguá.

Figura 6.2 - Gráfico scree-plot da prova Jaraguá



Fonte: do autor.

Aplicando a “regra do cotovelo” não foi possível determinar uma multidimensionalidade no teste, pois apenas um dos fatores apresentou autovalor acima da reta, o que indica um teste unidimensional. Assim, o teste é constituído por poucas questões e, aparentemente, possui apenas uma dimensão.

6.4 Análise no âmbito da TCT

Como o INEP (2014) desconsidera a inclusão de questões que possuem uma correlação ponto-bisserial inferior a 0,20, os itens 2, 5 e 6 foram excluídos, tanto das análises pela TCT quanto pela TRI. Em relação a análise pedagógica dos itens tentamos apontar alguns fatores que possam ter influenciado o baixo nível de correlação.

Iniciamos as análises da TCT com a Tabela 6.8 em que consta a classificação dos itens da prova Jaraguá de acordo com o nível de dificuldade.

Tabela 6.8 - Itens da prova Jaraguá na escala de dificuldade da TCT

Classificação	Itens	%	% esperado
Muito fáceis	1	14,1%	10%
Fáceis	3, 7 e 8	**43,6%**	20%
Medianos	9	**14,1%**	40%
Difíceis	4	14,1%	20%
Muito difíceis	10	14,1%	10%

Fonte: do autor.

De acordo com a Teoria Clássica dos Testes o item mais difícil da prova Jaraguá foi o item 10, enquanto o item mais fácil foi o item 1. Classificando os itens de acordo com a Tabela 2.2, criada pelo autor, os itens 1, 4 e 10, considerados muito fácil, difícil e muito difícil, respectivamente, atingiram uma porcentagem próxima da desejada. Porém os itens considerados fáceis compreendem uma porcentagem acima do esperado e, consequentemente, os itens considerados medianos compreendem uma porcentagem abaixo do esperado (negrito). Isso mostra que, de modo geral, o teste apresentou um nível considerado fácil.

Listamos na Tabela 6.9, em ordem decrescente, os itens da prova Jaraguá de acordo com o seu parâmetro de discriminação na Teoria Clássica dos Testes.

Tabela 6.9 - Parâmetro de discriminação dos itens da prova Jaraguá

Item	Parâmetro de discriminação
3	0,35052307
10	0,29806435
7	0,28967397
9	0,27147925
8	0,26803129
4	0,25559223
1	0,23925079

Fonte: do autor.

Considerando os valores que constam na tabela 8 e a classificação dos itens conforme a sua discriminação, temos um item com discriminação aceitável que não precisa de revisão (item 3), e seis itens que necessitam de revisão (itens 10, 7, 9, 8, 4 e 1).

Por meio da análise psicométrica clássica, de modo geral, verificamos que os itens da prova Jaraguá não apresentam muito poder discriminativo. Segundo Oliveira (2017) isso pode ter ocorrido por conta de uma quantidade insuficiente de questões. Uma quantidade pequena de questões não fornece uma medida confiável sobre o que se pretende medir, mesmo que esses itens sejam bastante correlacionados entre si (OLIVEIRA, 2017).

Para concluir as análises utilizando a Teoria Clássica dos Testes, analisamos os distratores por meio dos resultados que estão apresentados no Quadro 6.1. Os itens 2, 5 e 6 foram mantidos na análise dos distratores para que pudéssemos apontar alguns possíveis motivos pelos quais estes itens apresentaram resultados ruins:

Quadro 6.1 - Análise de distratores

Item 1			
Resposta	**Proficiência**		
	Baixa	**Média**	**Alta**
A	7	1	0
B	6	0	0
C	35	7	3
*D	100	47	95

Item 2			
Resposta	**Proficiência**		
	Baixa	**Média**	**Alta**
*A	55	26	72
B	52	**15**	**16**
C	15	7	6
D	26	7	4

Item 3			
Resposta	**Proficiência**		
	Baixa	**Média**	**Alta**
A	24	9	1
*B	58	35	89
C	39	8	4
D	27	3	4

Item 4			
Resposta	**Proficiência**		
	Baixa	**Média**	**Alta**
*A	22	20	59
B	48	**14**	**15**
C	65	**18**	**21**
D	13	3	3

Item 5			
Resposta	**Proficiência**		
	Baixa	**Média**	**Alta**
A	27	4	7
B	25	10	14
*C	20	10	37
D	76	**31**	**40**

Item 6			
Resposta	**Proficiência**		
	Baixa	**Média**	**Alta**
*A	24	11	48
B	39	**12**	**22**
C	50	**17**	**23**
D	35	15	5

Item 7			
Resposta	**Proficiência**		
	Baixa	**Média**	**Alta**
A	23	0	1
*B	93	52	92
C	20	1	3
D	12	2	2

Item 8			
Resposta	**Proficiência**		
	Baixa	**Média**	**Alta**
A	13	1	1
B	29	12	2
*C	71	36	88
D	35	6	7

Item 9			
Resposta	**Proficiência**		
	Baixa	**Média**	**Alta**
A	69	14	21
*B	27	30	64
C	15	2	3
D	37	9	10

Item 10			
Resposta	**Proficiência**		
	Baixa	**Média**	**Alta**
A	57	**25**	**26**
B	53	**16**	**26**
C	32	6	17
*D	6	8	29

Fonte: Do autor

Em todos os itens que podem permanecer no teste, um maior número de alunos com alta proficiência obteve êxito, ou seja, os alunos que se mostraram preparados para o teste, acertaram as questões esperadas. Em destaque no Quadro 6.1 (negrito) temos os distratores que atraíram respondentes de média e alta proficiência, o que aponta possíveis falhas na elaboração do item.

Vários fatores importantes para os resultados da prova e para futuros trabalhos similares podem ser observados com a análise que foi feita utilizando a Teoria Clássica dos Testes como: a exclusão de itens que não satisfazem requisitos mínimos, ênfase nas alternativas que atraíram muitos respondentes com alta proficiência (distratores), além dos índices de discriminação e dificuldade que podem ser comparados com os resultados obtidos com a TRI.

6.5 Análise no âmbito da TRI

De maneira semelhante à nossa análise utilizando a TCT, iniciamos a análise na TRI apresentando a classificação dos itens levando em consideração o seu nível de dificuldade (parâmetro *b*).

Tabela 6.10 - Parâmetro de dificuldade - TRI

Item	Dificuldade
1	0,625
3	0,269
4	1,258
7	-1,675
8	-0,162
9	1,029
10	2,438

Fonte: do autor.

Na sequência classificamos os itens em relação ao seu nível de dificuldade. Na Tabela 6.11 verifica-se o que cada índice de dificuldade apresentado na Tabela 6.10 representa na escala que vai de muito fácil a muito difícil.

Tabela 6.11 - Classificação dos itens em relação ao nível de dificuldade – TRI

Classificação	Itens	%	% esperado
Muito fáceis	7	14,1%	10%
Fáceis	8	14,1%	20%
Medianos	3	**14,1%**	40%
Difíceis	1, 4 e 9	**43,6%**	20%
Muito difíceis	10	14,1%	10%

Fonte: Do autor

Destacamos na Tabela 6.11 (negrito) as porcentagens que destoaram muito do que se espera para um teste equilibrado no que diz respeito a dificuldade dos itens. O item 10, considerado o mais difícil, e o item 7, considerado o mais fácil, foram resultados esperados. A surpresa maior ficou para o item 1, o qual será descrito na subseção de comparação entre TCT e TRI.

Na Tabela 6.12 temos os coeficientes de discriminação (parâmetro *a*) dos itens da prova Jaraguá.

Tabela 6.12 - Discriminação dos itens - TRI

Item	Discriminação
1	2,266
3	4,304
4	7,211
7	0,897
8	1,682
9	1,699
10	1,241

Fonte: Do autor.

Um dos itens com alto índice de discriminação que não se traduz em sua curva característica é o item 1. Isso ocorreu devido ao alto índice de acerto ao acaso (Tabela 6.14) deste item que, conforme vimos no capítulo 3, interfere diretamente no cálculo do índice de discriminação. Lembramos que o nível de dificuldade deste apresentou surpresa, o que também pode ser explicado pelo alto índice de acerto ao acaso, pois o mesmo interfere diretamente no cálculo do parâmetro *b*.

Na Tabela 6.13 são apresentados os valores de acerto ao acaso de todos os itens da prova Jaraguá.

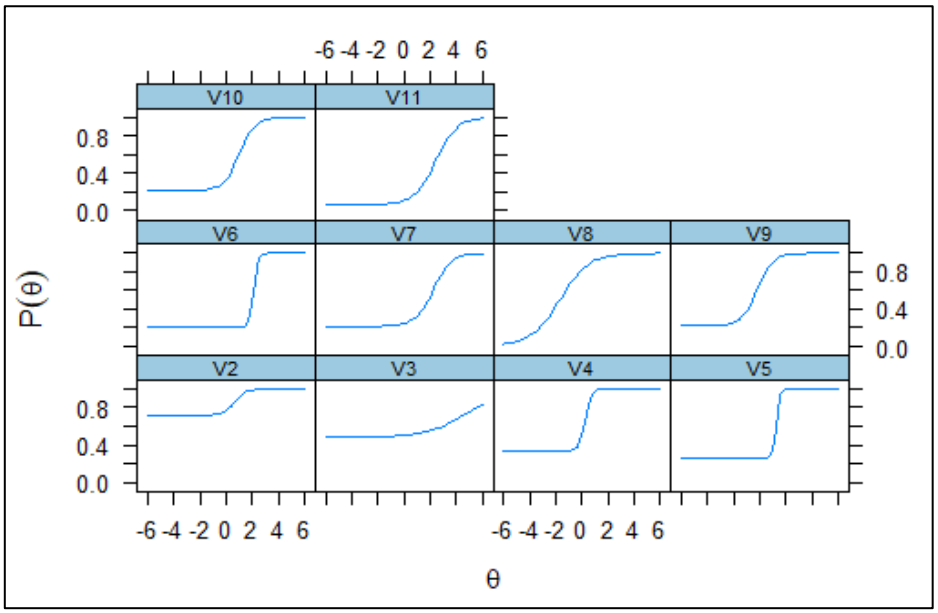
Tabela 6.13 - Acerto ao acaso

Item	Acerto ao acaso
1	0,715
3	0,337
4	0,250
7	0,006
8	0,223
9	0,216
10	0,065

Fonte: Do autor.

A seguir temos a Figura 6.3 que contém a curva características de cada um dos itens da prova Jaraguá. De modo geral, o comportamento da curva apresenta informações sobre o item no que diz respeito a dificuldade, discriminação e acerto ao acaso.

Figura 6.3 - CCIs considerando V2 = item 1 e assim sucessivamente até V11 = item 10.

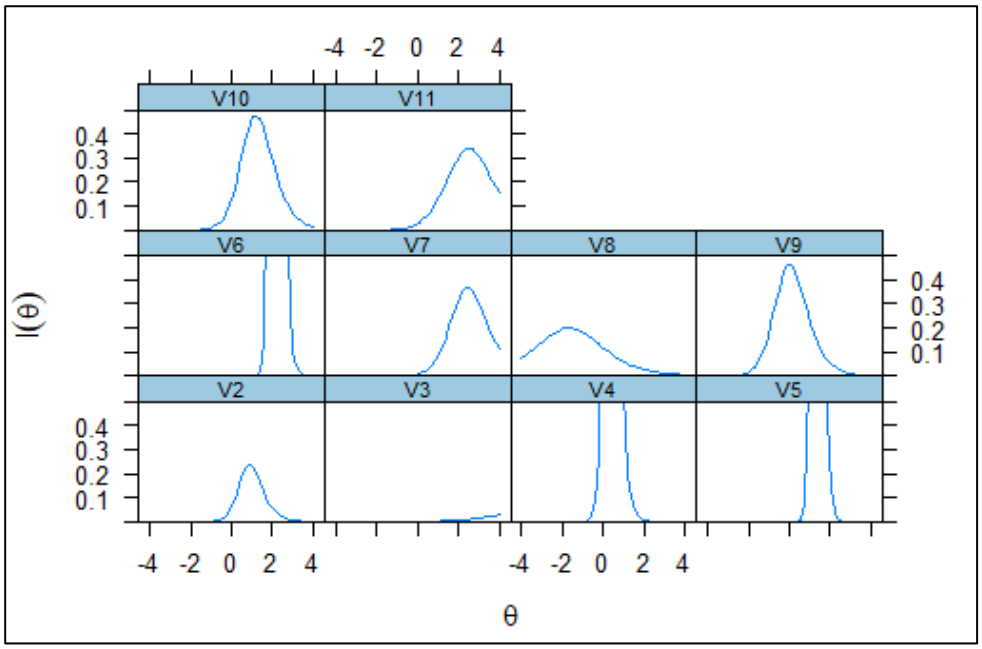


Fonte: Do autor.

Novamente mencionamos o item 1, representado na Figura 6.3 por V2. Analisando a CCI podemos perceber um grande achatamento da mesma, o que indica pouco poder de discriminação, ou seja, o item não é capaz de separar os respondentes que possuem altas habilidades dos que possuem baixas habilidades. O item 10 (V11), considerado o mais difícil do teste, também teve um bom índice de discriminação. Vale mencionar também a CCI do item 2 (V3). O grande achatamento desta curva mostra que o item fornece pouca informação sobre os respondentes.

Na Figura 6.4 apresentamos a função de informação de cada item do teste. A observação da configuração da curva de informação do item representa a maneira como o item se desempenhou para cada nível de habilidade no que diz respeito ao índice de discriminação.

Figura 6.4 - Função de informação de cada item.

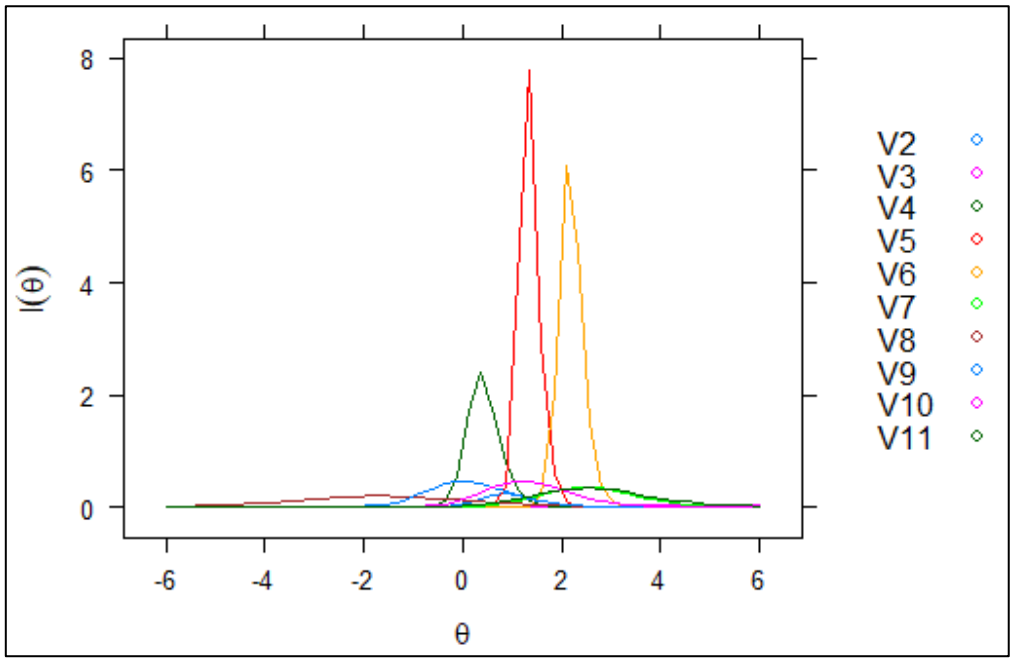


Fonte: do autor.

A curva de informação do item 2 (V3) reforça o que falamos anteriormente, o item realmente não apresenta informações sobre os respondentes. Os itens 3 e 4 do teste (V4 e V5, respectivamente) são os que mais discriminam os respondentes. Estes itens serão discutidos na comparação entre TCT e TRI. Observamos também a curva de informação do item 10 (V11) que apresenta informações sobre os respondentes com altas habilidades.

A opção pelo valor da escala se deu pelo alto valor de discriminação dos itens 4, 5 e 6 (V5, V6 e V7) o que acabou interferindo na visualização dos outros itens. A Figura 6.5, na qual podemos observar todas as curvas de informação do teste (referete aos 10 utens) complementa a Figura 6.4 e serve também para justificar a mudança na escala.

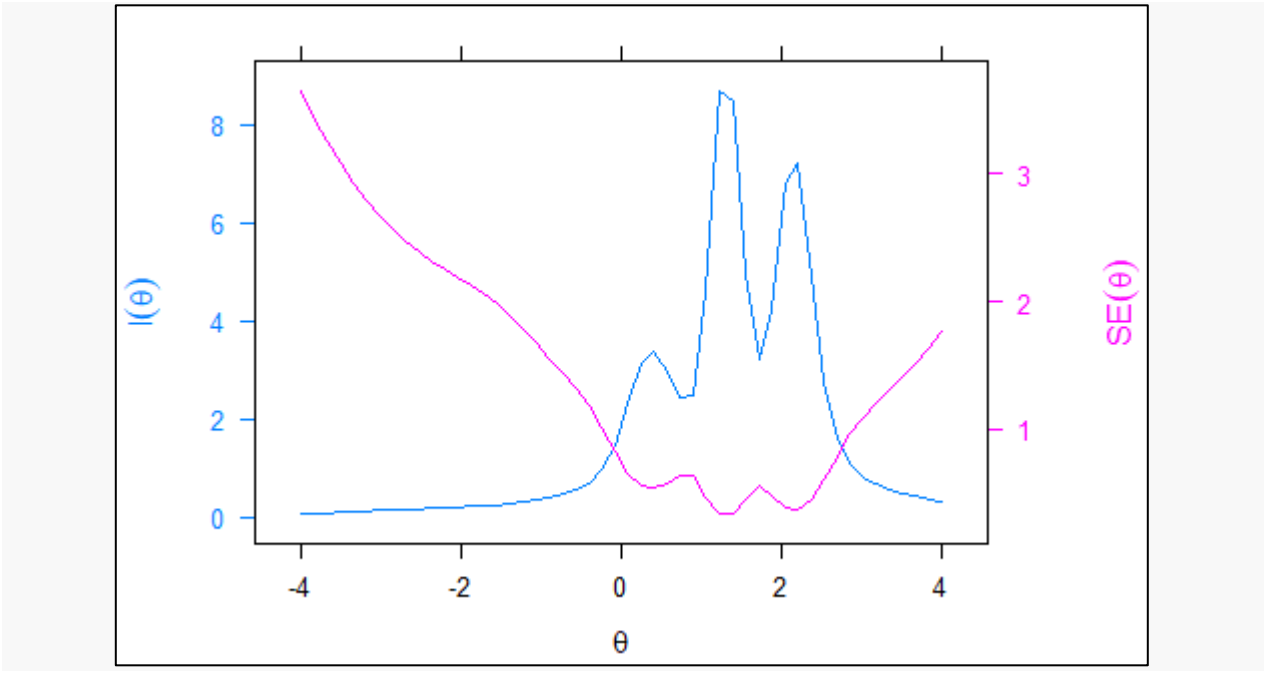
Figura 6.5 - Função de informação dos 10 itens



Fonte: do autor.

Em seguida temos a curva de informação do teste, juntamente com o erro padrão.

Figura 6.6 - Curvas de Informação do Teste e Erro Padrão



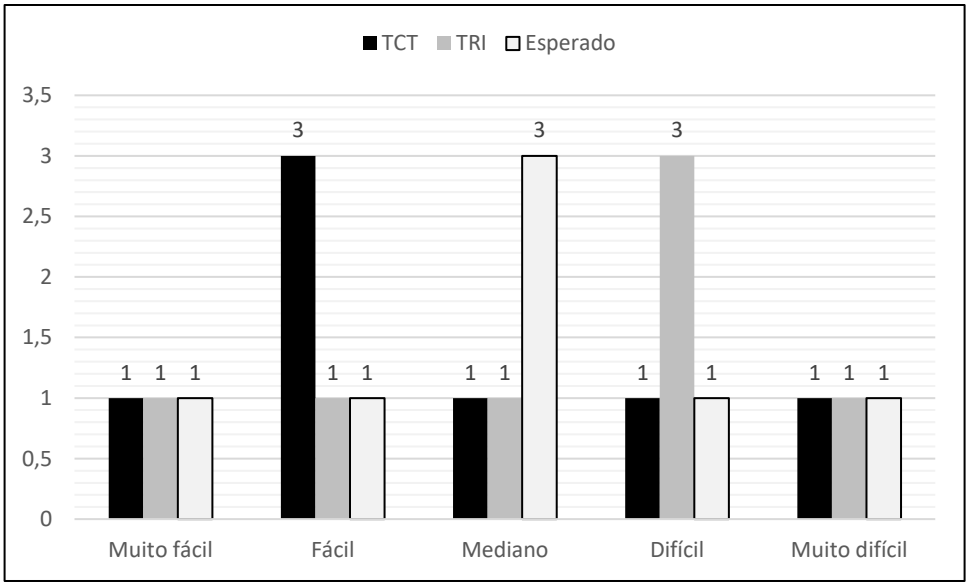
Fonte: Do autor.

Analisando a curva de informação do teste e erro padrão, pode-se perceber que o mesmo apresenta maior informação para os respondentes de proficiência em torno de 0,5 a 3,0, ou seja, nesse intervalo há uma maior precisão na estimativa das proficiências. Isso significa que a capacidade de discriminar os respondentes que possuem as habilidades necessárias para resolver o teste daqueles que não possuem, é maior para os respondentes que possuem um nível θ de habilidade no intervalo (0,5;3,0). Não podemos deixar de observar uma menor informação em duas regiões, próximo a (+1) e próximo a (+2) unidades da habilidade medida. Isso significa que a inclusão de itens que possuem estes níveis de dificuldade resultaria em uma maior confiabilidade no instrumento de medida. Percebe-se também que o erro de estimativa é maior próximo a estes dois níveis de habilidade, o que implica uma dificuldade de diferenciar de forma precisa os respondentes com níveis relativamente próximos a esses.

6.6 Comparação entre os resultados na TCT e na TRI

Nesta subseção vamos apresentar uma comparação entre as duas teorias no que diz respeito a dificuldade dos itens e discriminação dos itens, analisando os itens da prova Jaraguá com base nas escalas de dificuldade e discriminação deste estudo. Na Figura 6.7 comparamos a classificação dos itens por níveis de dificuldade.

Figura 6.7 - Classificação dos itens por nível de dificuldade



Fonte: do autor

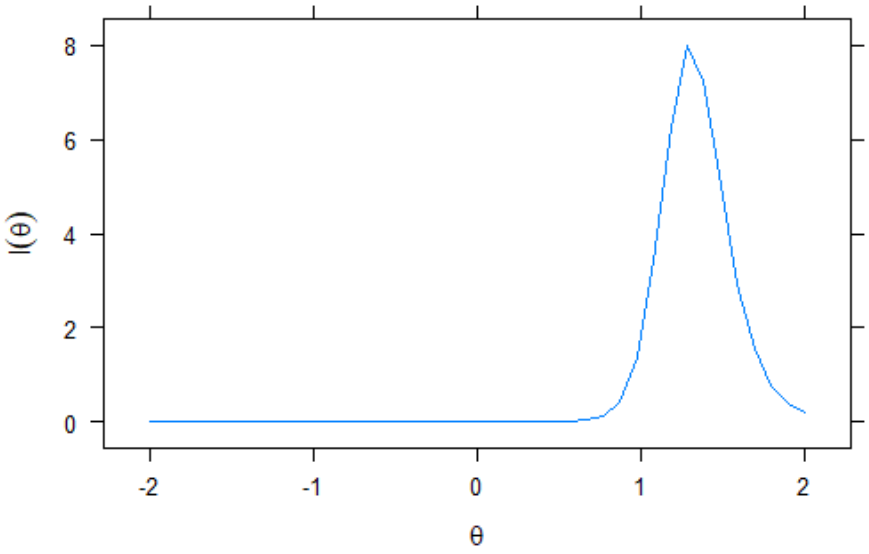
Verifica-se que ambas as teorias não atingiram uma quantidade desejável de itens medianos, ficando abaixo do esperado. Além disso, percebe-se que, de acordo com a TCT, o teste se mostrou mais fácil do que com a TRI. Nas faixas extremas os resultados estão dentro do esperado.

O item mais difícil do teste considerado por ambas as teorias foi o item 10. O item considerado mais fácil pela TCT foi o item 1, enquanto o item mais fácil de acordo com a TRI foi o item 7. O item 1 apresentou um alto índice de acerto ao acaso (0,715), o que interfere diretamente no cálculo do parâmetro de dificuldade da TRI (parâmetro *b*). Apesar do item 7 não ter sido considerado o mais fácil pela TCT, o mesmo foi considerado um item fácil, o que mostra que a análise das duas teorias apresentou similaridades.

Ressaltamos a importância da análise dos itens com a Teoria Clássica, o que nos fez olhar com mais atenção para o item 1, percebendo que o seu alto índice de acerto ao acaso interferiu na classificação dos itens de acordo com o parâmetro *b*.

Apesar de ter um alto índice de discriminação (7,211), analisando a função de informação do item 4 (Figura 6.8) podemos ver que esta discriminação é particularmente válida para os respondentes com aptidões dentro de um pequeno intervalo próximo a 0,8.

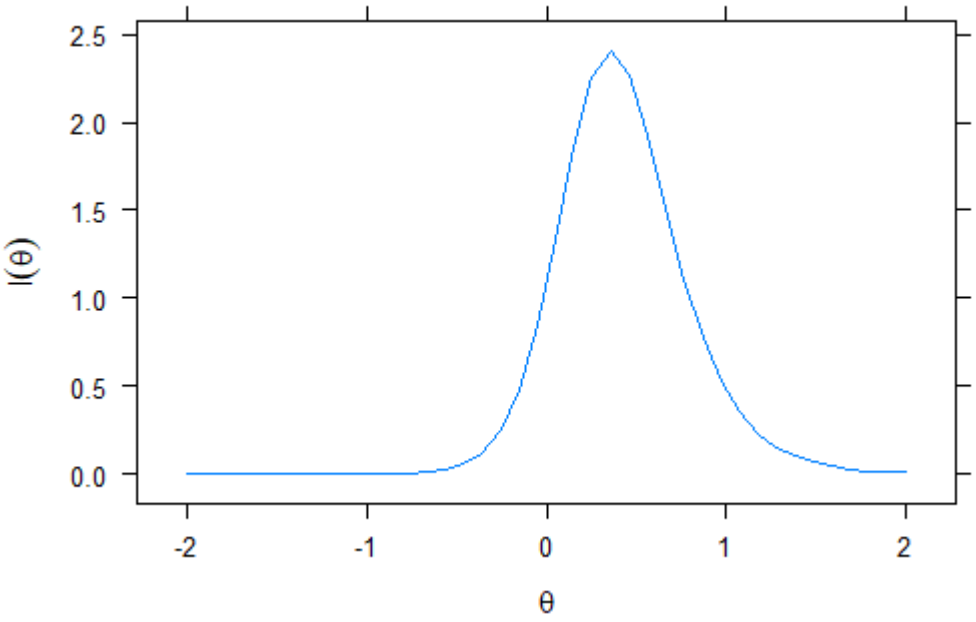
Figura 6.8 - Função de informação do item 4



Fonte: do autor.

Na Figura 6.9 temos a função de informação do item 3.

Figura 6.9 - Função de informação do item 3



Fonte: Do autor.

O item 3, o mais discriminativo na TCT (0,35), também teve um bom índice de discriminação na TRI (4,304), porém, da mesma forma que aconteceu com o item 4, através da sua função de informação temos que esta discriminação é particularmente válida para os respondentes com nível de aptidão em torno de 0,5.

7 Análise Pedagógica

Neste capítulo faremos uma análise pedagógica de sete itens da prova Jaraguá: o item 1, considerado o mais fácil pela TCT com aproximadamente 80% de acertos, os itens 2, 5 e 6 que, por não atingirem um índice mínimo de 0,20 da correlação ponto-bisserial devem ser excluídos do teste, o item 4 que apresentou distratores ruins, o item 10, considerado o mais difícil da prova em ambas as teorias e o item 7, considerado um item fácil em ambas as teorias.

Em relação a análise das questões pelo item 1 da prova Jaraguá. Dois fatos curiosos em relação a análise deste item estão em seus índices de dificuldade e discriminação. Se observarmos a CCI (Figura 6.3), podemos perceber que não é necessário possuir grandes habilidades para resolver o item, pois b = 0,625. Mesmo assim, de acordo com o estudo apresentado neste trabalho, o nível de dificuldade deste item foi dado como mediano.

Em relação ao poder de discriminação do item 1, a = 2,266, o grande achatamento da curva indica que o item não discrimina os respondentes de diferentes níveis de habilidade, o que novamente vai de encontro ao seu índice mediano de discriminação.

Estes fatos são explicados pelo alto índice de acerto ao acaso c = 0,715, ou seja, a alternativa mais procurada quando se desejou "chutar" a resposta foi justamente a alternativa correta, letra (D), o que dificulta interpretar se o respondente realmente possuía as habilidades necessárias para resolver o item, ou se ele simplesmente "chutou" e acertou a questão.

Questão 1

1) Renato levou 2 horas para digitar um texto de 8 páginas. Se ele trabalhar durante 4 horas, no mesmo ritmo, é possível que ele digite um texto de:

(A) 4 páginas
(B) 8 páginas
(C) 12 páginas
(D) 16 páginas

Esta é uma questão em que se deseja que o estudante aplique regra de três. Porém, como a razão é igual a 2, o estudante nem precisaria aplicar a regra, pois basta dobrar o número de páginas, assim como aconteceu com as horas. Uma possível alteração no enunciado que provavelmente faria com que o estudante aplicasse de fato a regra de três seria:

1) Renato levou 4 horas para digitar um texto de 22 páginas. Se ele trabalhar durante 10 horas, no mesmo ritmo, é possível que ele digite um texto de:

(A) 50 páginas
(B) 60 páginas
(C) 44 páginas
(D) 55 páginas

Com essa mudança no enunciado a proporção entre os valores dados não é tão direta, o que faz com que o respondente tente aplicar a regra de três.

Conforme foi relatado no capítulo 6, a retirada dos itens 2, 5 e 6 ocasionariam em um teste mais eficiente. Agora, de forma mais detalhada, analisaremos alguns dos motivos que podem ter influenciado este resultado.

Questão 2

2) Ao resolver corretamente a expressão -1-(-5).(-3)+(-4).3:(-4), o resultado é:

(A) -13
(B) -2
(C) 0
(D) 30

Um detalhe que podemos perceber na elaboração desta questão está relacionado à sua formatação. Os sinais operacionais não são totalmente claros, podendo haver confusão na hora de realizar os cálculos. A seguir mostramos como deveria estar apresentada a expressão: $-1-(-5)\cdot(-3)+(-4)\cdot 3:(-4)$.

Apesar de ter sido uma prova aplicada para o 8º ano, o conteúdo é do 7º ano. A ideia foi aplicar um simulado no início do ano letivo com os conteúdos referentes ao ano anterior para diagnosticar as deficiências dos alunos e trabalhar estas deficiências antes do simulado seguinte. Eu, por exemplo, passei a trabalhar um pouco mais de álgebra com os estudantes do 8º ano, pois foi dignosticado uma certa dificuldade neste conteúdo com as análises da prova Jaraguá.

Outro detalhe a ser observado é que essa questão envolve conteúdos do 7º ano (operações entre números inteiros) e do 6º ano (expressões numéricas). A questão 2 foi elaborada para medir o conhecimento dos estudantes sobre as operações que envolvem números inteiros, e os principais erros da questão foram na resolução da expressão numérica, e não das operações em si.

Em seguida apresentamos a questão 4, na qual as alternativas (B) e (C) se apresentaram como distratores ruins por atraírem muitos respondentes com média e alta habilidade.

Questão 4

4) Um carro percorre 5 km, enquanto no mesmo intervalo de tempo um homem caminha 40 m. Observando a razão entre os espaços percorridos pelo carro e pelo homem, concluímos que:

(A) o carro percorre 125 m enquanto o homem percorre 1 m.
(B) o carro percorre 5 000 m enquanto o homem percorre 4 m.
(C) o carro percorre 500 m enquanto o homem percorre 40 m.
(D) o carro percorre 1 250 m enquanto o homem percorre 1 m.

Uma explicação para os respondentes terem assinalado as alternativas (B) e (C) está no fato de que ambas possuem um número de quilometragem que é múltiplo da quilometragem do enunciado da questão envolvendo potências de base 10 ($5 \cdot 10^3$ e $5 \cdot 10^2$). Uma mudança no enunciado que minimizaria o erro cometido por estes respondentes seria a não utilização dos múltiplos de 5 com potências de base 10. Esta mudança faria com que os respondentes cumprissem as etapas necessárias para a resolução da questão: transformação de unidades (5 km em 5000 m) e proporção entre os números do problema e os números da resposta.

Questão 5

5) (Saresp) Marcos fez um empréstimo de R$ 120.000,00 que deverá pagar com juros simples de 1% sobre o valor emprestado a cada mês. Sabendo que ele pagou R$ 6.000,00 de juros, quantos meses levou para pagar o empréstimo?

(A) 3 meses
(B) 4 meses
(C) 5 meses
(D) 6 meses

A questão 5 também foi designada para ser retirada do teste. A análise dos distratores mostra que a alternativa (D) atraiu mais respondentes com altas habilidades do que a alternativa correta (C). Uma maneira de resolver esta questão é:

$6000 = \left(\frac{1}{100} \cdot 120000\right) \cdot t \ \rightarrow t = 5$, em que t representa o tempo que Marcos levou para pagar o empréstimo. A maioria dos estudantes que assinalou a alternativa (D), provavelmente não executou os cálculos, tentando utilizar um raciocínio lógico ao ver relação entre os 6 meses da resposta e os R$6.000,00 de juros. Dessa forma a alternativa (D) se apresentou como um distrator ruim, pois foi assinalada por aproximadamente 40% dos respondentes com nível alto de habilidade e aproximadamente 56% dos respondentes com um nível médio de habilidade.

Este conteúdo, além de porcentagem, envolve matemática financeira, conteúdo trabalhado por muitos professores somente no 9º ano. Portanto, o fato de muitos alunos não terem estudado o conteúdo necessário para a resolução deste problema é o principal motivo para que este item seja excluído do teste.

Outro fator que pode ter influenciado o desempenho ruim de alguns estudantes neste item foi a falta de autonomia durante o processo de aprendizagem. Essa falta de autonomia, juntamente com a falta de persistência na resolução de um problema, podem ser detectadas pelos erros nos simulados. A alternativa (D) da questão 5, que atraiu mais respondentes do que a alternativa correta em todos os três níveis de habilidade, demonstra uma resposta assinalada sem muito esforço, apresentando uma busca rápida por uma solução que faça algum sentido, mesmo que este sentido seja mínimo. Estudantes que tenham autonomia nos estudos, que tenham o costume de persistir na busca pela solução do problema, de modo geral, não cometem este tipo de erro.

Questão 6

6) Uma bicicleta, cujo preço é R$ 1200,00, pode ser comprada da seguinte maneira:

a) a vista, com 15% de desconto.

b) pagamento para 90 dias, com acréscimo de 25% sobre o preço inicial.

Assinale a alternativa que corresponde a diferença, em reais, entre as duas opções de compra?

(A) R$ 480,00
(B) R$ 180,00
(C) R$ 300,00
(D) R$ 150,00

A questão 6 foi classificada como uma questão difícil (b = 2,263). De fato, a resolução desta questão envolve algumas etapas:

15% de R$1.200,00 = R$180,00

25% de R$1.200,00 = R$300,00

R$180,00 + R$300,00 = R$480,00

A resposta correta é a letra A (R$480,00). Como as alternativas B e C apresentam valores que aparecem durante a resolução da questão (R$180,00 e R$300,00), estas duas alternativas atraíram vários respondentes. Por outro lado, grande parte dos respondentes com altas habilidades escolheram a opção correta, o que mostra que as alternativas B e C não foram distratores tão ruins.

Esta questão, assim como a questão 5, envolve porcentagem e foi indicada para ser retirada do teste. Uma sugestão para os professores é que trabalhem o conteúdo de porcentagem com um pouco mais de ênfase no 7º ano.

Questão 7

7) (Saresp) Uma pesquisa foi respondida por 200 pessoas, que indicaram o local que mais frequentam nos finais de semana. A distribuição das respostas está registrada na tabela seguinte:

	Shopping	Clube	Restaurante	Praia
Número de respostas	100	50	30	20

O gráfico de setores que representa o resultado dessa pesquisa pode ser:

(A) 

(B) 

(C) 

(D) 

O item 7 foi considerado o mais fácil pela TRI e considerado fácil pela TCT. Um aluno que tenha lido o enunciado viu que 200 pessoas participaram da pesquisa e, dessas 200, 100 responderam shopping, ou seja, a metade. A única alternativa que apresenta uma divisão que contém um setor que corresponde à metade é a alternativa (B).

Questão 10

10) CMB) Um feirante comprou 15 quilos (kg) de alho para vender em pacotes de 150 gramas (g). Ao final do dia, ele tinha vendido a metade dos pacotes. Dentre as opções abaixo, a única que apresenta a sequência de operações que determina a quantidade de pacotes que restaram ao final do dia é:

(A) [(15.100) : 150] : 2
(B) [(15:100) : 150] . 2
(C) [(15:1000) : 150] : 2
(D) [(15.1000) : 150] : 2

As três alternativas incorretas (A), (B) e (C) atraíram uma quantidade considerável de respondentes com alta habilidade. As alternativas (A) e (B) também foram escolhidas por vários respondentes com média habilidade. Como este foi considerado um item muito difícil (b = 2,438), é esperado que todas as alternativas pareçam corretas. Além disso, a alternativa que atriu mais respondentes com altas habilidades foi a alternativa correta, o que mostra que os distratores, de modo geral, cumpriram bem a sua função.

Este item atingiu bons níveis de discriminação. De fato, o item foi bem elaborado. Diferentemente do que acontece com a maioria das questões que pedem a resposta do problema, esta questão avalia se o estudante entendeu o processo, algo difícil de se avaliar em questões objetivas. Além de interpretar a situação do problema o estudante deveria lembrar que 1 kg corresponde a 1.000 g, ou seja, deve-se multiplicar 15 por 1.000.

Podemos dizer que a Teoria Clássica dos Testes e a Teoria de Resposta ao Item foram muito úteis no processo de avaliação da prova Jaraguá. Fomos capazes de apontar falhas de elaboração, salientar acertos, discutir os problemas e as suas possíveis soluções, fatores que interferem diretamente na aprendizagem dos estudantes.

De modo geral não foi um bom teste. A pequena quantidade de questões pode ter interferido nos resultados, além da amostra de estudantes que também poderia ser maior. Uma das nossas sugestões para melhorar este teste seria o aumento do número de questões e a sua aplicação a um número maior de estudantes.

Nas novas edições deste teste alguns cuidados devem ser tomados em relação a elaboração dos itens como verificar se o item realmente está avaliando aquilo que se deseja, se o respondente necessita realmente da habilidade a ser medida para resolver o item, ou se o item pode ser resolvido de uma forma mais simples.

8 Percepções e inquietações do autor

No contexto geral, o foco principal deste trabalho está nas avaliações, mas no meu ponto de vista outras questões pertinentes merecem ser discutidas, em minha visão. Desde que eu comecei a lecionar matemática no município de Jaraguá do Sul em 2014, percebo nos estudantes, de modo geral, pouco tempo de dedicação aos estudos. É claro que a atuação da família e dos professores tem grande influência nisso. Em relação aos professores tenho percebido alguns problemas que exigem uma atenção especial. Diferentemente do que ocorre em algumas instuições de ensino da rede privada, na rede municipal de Ensino de Jaraguá do Sul os estudantes só têm aulas com um professor com formação em matemática a partir do 6º ano. Antes disso, as aulas são dadas por um professor formado em pedagogia que leciona a maior parte das disciplinas curriculares. A grande maioria dos professores das séries iniciais (pedagogos) com os quais eu discuti sobre matemática disseram ter uma certa dificuldade em ensinar alguns conteúdos específicos de matemática, como frações e contagem.

Com base em minhas vivências de sala de aula, o maior obstáculo que existe entre os estudantes e o entendimento da matemática é a falta de autonomia por parte desses estudantes. Os estudantes que apresentam um bom desempenho, de modo geral, quando se deparam com um desafio, persistem na busca pela resposta, mesmo que tenham fracassado nas primeiras tentativas de resolver o desafio. Em contrapartida, os estudantes que geralmente não apresentam bom desempenho em matemática costumam desistir logo após as primeiras tentativas, buscando logo a resposta sem ter feito um grande esforço. Pelo que conversamos com os professores da UDESC durante as disciplinas deste mestrado, essa falta de autonomia acompanha os estudantes até os cursos de graduação, juntamente com a falta de tempo destinado aos estudos.

Tentando mudar este cenário do Ensino Fundamental, pelo menos das turmas em que eu leciono, algumas estratégias fazem parte do meu trabalho. Uma dessas estratégias é entregar para os estudantes uma lista de exercícios (algo em torno de 50 exercícios) para que eles resolvam em casa e me entreguem no dia da prova. A entrega desta lista consta, para a nota da prova, um acréscimo que varia de 0,1 a 1,0 ponto, dependendo da quantidade de exercícios reolvidos, organização e, é claro, exatidão. A qualquer momento os estudantes podem pedir auxílio para resolver os exercícios mais difíceis, pedir dicas, ou até que eu mesmo resolva, quando estiver muito próximo da prova. O que eu sempre digo é que: - vocês não têm a obrigação de conseguir resolver

todos os exercícios, mas tem a obrigação de tentar. O resultado está sendo bem produtivo. Se eu demoro alguns dias para entregar a lista eles mesmos me cobram, pois geralmente uso alguns exercícios parecidos na prova, então eles entendem que a resolução da lista é uma boa forma de preparação.

Outra estratégia que eu utilizo, desta vez nas provas, é fazer uma questão extra, valendo 1,0 ponto. Esta questão possui um nível bem maior do que as outras questões da prova, e tem como objetivo motivar e desafiar os estudantes, o que geralmente acontece com os de bom desempenho. Entre as questões utilizadas estão problemas da segunda fase da OBMEP, problemas da OBM, questões de vestibulares, ou questões mais aprofundadas do conteúdo que não foram trabalhadas em sala. A figura 16 representa uma questão que eu já utilizei em provas.

Figura 8.1 - Questão aplicada em uma prova de matemática

(1,0 ponto) (OBMEP 2015: FASE 2) Comece uma sequência escrevendo dois números inteiros não negativos, sendo o primeiro maior do que o segundo. Depois, para encontrar os próximos termos da sequência, repita o seguinte procedimento:

- se o último termo escrito for maior do que o penúltimo, a sequência termina;
- caso contrário, o próximo termo a ser escrito será o penúltimo menos o último. Um exemplo é a sequência 120, 71, 49, 22, 27; ela começa com 120 e 71 e possui cinco termos.

a) Escreva a sequência que começa com 30 e 16.

b) Escreva a sequência que possui exatos cinco termos, sendo o quarto termo igual a 1 e o quinto termo igual a 2.

c) Uma sequência que começa com 25 tem exatamente três termos. Quais são os valores possíveis para o segundo termo?

d) Uma sequência que começa com 60 tem o maior número possível de termos. Qual é o valor do segundo termo dessa sequência?

Fonte: OBMEP

Além de incentivar os estudantes com bom desempenho, esta questão bônus, mesmo que de uma forma singela, serve de preparação para que eles tenham um melhor desempenho em outras avaliações, além da OBMEP.

Quando incentivamos o raciocínio, a persistência na busca pela solução, e a perseverança, independente da quantidade de fracassos, os estudantes estão sendo preparados não apenas para as provas de matemática, mas para qualquer situação de dificuldade que encontrar na vida.

9 Considerações Finais

O objetivo do presente trabalho é analisar a qualidade de um simulado de matemática aplicado para estudantes do 8º ano da prefeitura municipal de Jaraguá do Sul, além de discutir os itens do simulado pedagogicamente, contribuindo na elaboração de futuros testes. Para cumprir este objetivo, foram aplicadas na pesquisa a metodologia da Teoria Clássica dos Testes e da Teoria de Resposta ao item. Estas análises nos levaram a discutir pedagogicamente a construção de um item, enriquecendo as discussões sobre a TCT e a TRI e contribuindo para futuros trabalhos similares. Além da análise de forma individual utilizando as duas teorias, realizamos também uma comparação entre os resultados apresentados pela TCT e pela TRI, apontando algumas limitações, semelhanças e diferenças nos resultados.

A análise feita com o auxílio da TCT e da TRI, nos proporcionou a clareza e a confiabilidade dos resultados obtidos, identificando os itens que são capazes de discriminar os estudantes que possuem as habilidades necessárias para resolver o teste daqueles que não possuem (itens 3, 4 e 10). Além disso, também fomos capazes de classificar os itens de acordo com o nível de dificuldade, o que pode ser muito importante na hora de elaborar um teste.

Os resultados e discussões deste trabalho podem ser de grande valor para a aplicação de testes similares, tanto para a Secretaria Municipal de Educação de Jaraguá do Sul quanto para outras instituições que tem por objetivo avaliar a qualidade de suas provas, corrigindo e aprimorando o que se mostrar necessário

Concluímos que os itens 2, 5 e 6 deveriam ser retirados do teste, porém não refizemos os procedimentos de análise e discussão. Isso porque o maior objetivo deste trabalho está voltado para a análise pedagógica dos itens, auxiliando na elaboração de testes futuros.

Como sugestão para trabalhos futuros, a partir do estudo realizado, recomendamos a aplicação a TCT e da TRI em simulados de outras disciplinas, além de matemática, com a presença de uma análise pedagógica dos itens, o que auxilia o processo de minimização dos erros na elaboração do teste. Sugerimos também que sejam realizados os mesmos procedimentos de análise deste trabalho excluindo os itens 2, 5 e 6, além da utilização dos modelos de 2 e 4 parâmetros (ML2 e ML4), discutindo os resultados e fazendo uma comparação paralela aos resultados discutidos no presente trabalho. Sugerimos também analisar se os itens da prova Jaraguá satisfazem os requisitos necessários para serem classificados como âncora ou quase âncora, o que agregaria mais valor aos resultados do trabalho.

10 Referências

ANDRADE, D. F. D.; TAVARES, H. R.; VALLE, R. D. C. **Teoria da Resposta ao Item:** conceitos e aplicações. Ceará: Sinape, 2000. Disponível em <http://egov.ufsc.br/portal/sites/default/files/livrotri.pdf>

ANDRIOLA, W. B. Psicometria Moderna: características e tendências. **Estudos em Avaliação Educacional**, São Paulo, v. 20, n. 43, mai/ago, 319-340, 2009.

ANJOS, A.; ANDRADE, D. F. **Teoria de Resposta ao Item com o uso do R.** In: Simpósio Nacional de Probabilidade e Estatística, 2012. Disponível em <https://docs.ufpr.br/~aanjos/CE095/RTRIsinape.pdf>. Acesso em 12 de agosto de 2018.

ARIAS, M. R. M.; LLOREDA, M. J. H.; E LLOREDA, M.V. H. **Psicometria.** Madrid. Alianza Editorial, S. A., 2006.

BARNARD-BRAK, Lucy; LAN, William Y.; YANG, Zhanxia. Differences in mathematics achievement according to opportunity to learn: A 4pL item response theory examination. **Studies in Educational Evaluation**, v. 56, p. 1-7, 2018.

CHALMERS, R. P. mirt: A multidimensional Item Response Theory Package for the R Environment**. Journal of Statistical Software**, v. 48, n. 6, p. 1-29, 2012. Acesso em 02 de setembro de 2018.

CORTESÃO, L. **Formas de Ensinar, formas de avaliar**. Lisboa: Ministério da Educação, Departamento do Ensino Básico: [s.n.], 2002.

FERREIRA, Francisco Fialho G. **Escala de Proficiência para o ENEM: utilizando teoria da resposta ao item**. Dissertação de Mestrado apresentada ao Programa de Pós-graduação em Matemática e Estatística, UFPA, 2009. Disponível em: <http://www.ppgme.ufpa.br/doc/diss/fialhoguedes.pdf>. Acesso em 16 de julho 2018.

HAIR, J. F.; BLACK, W. C.; BABIN, J. B.; ANDERSON, R. E.; TATHAN, R. L. **Análise Multivariada de Dados.** 6ª edição. Artmed S. A. Porto Alegre. 2005.

INEP. Exame Nacional de Desempenho de Estudantes (ENADE 2014). **Relatório Síntese:** Matemática. 2014. Disponível em: <http://download.inep.gov.br/educacao_superior/enade/relatorio_sintese/2014/2014_rel_matem matem.pdf> Acesso em 20 de julho de 2018.

IBGE. **Censo Demográfico.** 2017. Disponível em <https://www.ibge.gov.br/estatisticas-novoportal/por-cidade-estado-estatisticas.html?t=destaques&c=4208906>. Acesso em 12 de agosto de 2018.

INEP. **IDEB – Resultados e Metas.** Disponível em <http://ideb.inep.gov.br/resultado/home.seam?cid=5721>. Acesso em 20 de julho de 2018.

JUNKER, B. W. **Some aspects of classical reliability theory & classical test theory**. Carnegie Mellon University. Pittsburgh. 2012.

KLEIN, R. Alguns aspectos da Teoria de Resposta ao Item relativos à estimação das proficiências. **Ensaio:** Avaliação e Políticas Públicas em Educação, v. 21, n. 78. 2013

KLEIN, R.; FONTANIVE, N. S. Avaliação em larga escala: tendências e desafios**. Em aberto**, vol. 15, n. 66, p. 29-34, abr./jun. 1995.

KOLEN, Michael J.; BRENNAN, Robert L. **Test equating: Methods and practices**. Springer Science & Business Media, 2013.

LUCKESI, C. C. **Avaliação da Aprendizagem escolar**. 19. ed. São Paulo: Cortez, 2008.

LUCKESI, C. C. **Verificação ou Avaliação:** O Que Pratica a Escola. Governo do Estado do Ceará: CAED, 2007.

MACHADO, N. **Epistemologia e didática**. São Paulo. 1996.

MAIA, J. L. **O uso da Teoria Clássica dos Testes - TCT e da Teoria de Resposta ao Item - TRI na avaliação da qualidade métrica de testes de seleção.** Tese de doutorado apresentada ao programa de pós-graduação em educação brasileira, UFC, 2009. Disponível em: <http://repositorio.ufc.br/bitstream/riufc/3235/1/2009_Tese_JLMaia.pdf>. Acesso em 16 de julho de 2018.

MENDONÇA, J. **Análise da eficiência de estimação de parâmetros da TRI pelo software ICL.** Dissertação de Mestrado. Universidade Federal de Lavras. Lavras – MG, 2012.

MUÑIZ, J. **Teoría de Respuesta a los ítems:** un nuevo enfoque en la evolución psicológica y educativa. Madri: Pirámide, 1990.

MUÑIZ, J. **Introducción a la teoria de respuesta a los ítems**. Madri: Pirámide, 1997.

MUÑIZ, J. **Teoria Clássica dos Testes**. Madrid: Pirámide, S. A, 2003.

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

RABELO, M. **Avaliação Educacional:** Fundamentos, metodologia e aplicação no contexto brasileiro. Rio de Janeiro: SBM, 2013.

REVELLE, W. **The "New Psychometrics" – Item Response Theory.** Disponível em: <http://www.personality-project.org/r/book/Chapter8.pdf>. Acesso em 01 de setembro de 2018.

REVELLE, William R. psych: **Procedures for personality and psychological research**. 2017. Disponível em: <https://CRAN.R-project.org/package=psych. Acesso em 01 de setembro de 2018.

RIZOPOULOS, Dimitris. ltm: An R package for latent variable modeling and item response theory analyses. **Journal of statistical software**, v. 17, n. 5, p. 1-25, 2006. Disponível em <http://www.jstatsoft.org/v17/i05/>

SILVA, F. E. F. D. **Teoria de resposta ao Item (TRI) em Avaliações de Matemática na EEM Professor Gabriel Epifânio dos Reis**. Trabalho de conclusão de curso apresentado ao Corpo Docente do Mestrado Profissional em Matemática em Rede Nacional - PROFMAT - UFERSA. Mossoró. 2015. Disponível em < https://sca.profmat-sbm.org.br/sca_v2/get_tcc3.php?id=598>. Acesso em 01 de setembro de 2018.

TORRES, F. C. **Uma Aplicação da Teoria de Resposta ao Item em um Simulado de Matemática no Modelo ENEM.** Trabalho de conclusão de curso apresentado ao Departamento de Matemática da Universidade de Brasília, UNB, 2015. Disponível em: <https://sca.profmat-sbm.org.br/sca_v2/get_tcc3.php?id=79363>. Acesso em 16 de julho de 2018.

URBINA, S. **Fundamentos da Testagem Psicológica**. Porto Alegre: Artmed, 2007.

VIANNA, H. M. **Testes em Educação**. 4 Ed. ed. São Paulo : Ibrasa, 1982.

VIANNA, H. M. Avaliação Educacional: Uma perspectiva histórica. **Estudos em Avaliação Educacional**, São Paulo, v. 25, p. 7-24, Dez 1995.

VIANNA, H. M. **Estudos em Avaliação Educacional**. São Paulo: [s.n.], v. 25, 2014.

WILLSE, J. T. (2017). **CTT: Classical Test Theory Functions**. R package version 2.3. <https://CRAN.R-project.org/package=CTT>

Anexo I

Nas páginas seguintes encontram-se os 10 itens do simulado de Matemática aplicado para as turmas do 8º ano da rede municipal de educação de Jaraguá do Sul em 2016.

MATEMÁTICA – 8º ANO

1) Renato levou 2 horas para digitar um texto de 8 páginas. Se ele trabalhar durante 4 horas, no mesmo ritmo, é possível que ele digite um texto de:

(A) 4 páginas
(B) 8 páginas
(C) 12 páginas
(D) 16 páginas

2) Ao resolver corretamente a expressão -1-(-5).(-3)+(-4).3:(-4), o resultado é:

(A) -13
(B) -2
(C) 0
(D) 30

3) O consumo de determinadas frutas é benéfico à saúde. Um exemplo é a pêra, cujo consumo auxilia na circulação do sangue, no controle da pressão arterial e facilita a digestão. Cada 100g dessa fruta equivale a 56 calorias. Uma pessoa que ingere 450g dessa fruta, fornece ao organismo:

(A) 156 calorias
(B) 252 calorias
(C) 468 calorias
(D) 504 calorias

4) Um carro percorre 5 km, enquanto no mesmo intervalo de tempo um homem caminha 40 m. Observando a razão entre os espaços percorridos pelo carro e pelo homem, concluímos que:

(A) o carro percorre 125 m enquanto o homem percorre 1 m.
(B) o carro percorre 5 000 m enquanto o homem percorre 4 m.
(C) o carro percorre 500 m enquanto o homem percorre 40 m.
(D) o carro percorre 1 250 m enquanto o homem percorre 1 m.

5) (Saresp) Marcos fez um empréstimo de R$ 120.000,00 que deverá pagar com juros simples de 1% sobre o valor emprestado a cada mês. Sabendo que ele pagou R$ 6.000,00 de juros, quantos meses levou para pagar o empréstimo?

(A) 3 meses
(B) 4 meses
(C) 5 meses
(D) 6 meses

6) Uma bicicleta, cujo preço é R$ 1200,00, pode ser comprada da seguinte maneira:

a) a vista, com 15% de desconto.

b) pagamento para 90 dias, com acréscimo de 25% sobre o preço inicial.

Assinale a alternativa que corresponde a diferença, em reais, entre as duas opções de compra?

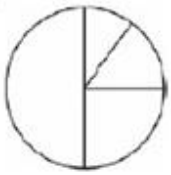
(A) R$ 480,00
(B) R$ 180,00
(C) R$ 300,00
(D) R$ 150,00

7) (Saresp) Uma pesquisa foi respondida por 200 pessoas, que indicaram o local que mais frequentam nos finais de semana. A distribuição das respostas está registrada na tabela seguinte:

	Shopping	Clube	Restaurante	Praia
Número de respostas	100	50	30	20

O gráfico de setores que representa o resultado dessa pesquisa pode ser:

(A) 

(B) 

(C) 

(D) 

8) Tenho que comprar lápis e canetas. Se comprar 7 lápis e 3 canetas, gastarei R$ 16,50. Se comprar 5 lápis e 4 canetas, gastarei R$ 15,50. Qual o preço de cada lápis e cada caneta?

(A) Preço do lápis é R$ 5,00 e preço da caneta é R$ 2,00
(B) Preço do lápis é R$ 0,50 e preço da caneta é R$ 1,00
(C) Preço do lápis é R$ 1,50 e preço da caneta é R$ 2,00
(D) Preço do lápis é R$ 1,50 e preço da caneta é R$ 3,50

9) Uma casa, com 250 m² de área construída, tem 4 dormitórios do mesmo tamanho. Qual é a área de cada dormitório, se as outras dependências da casa ocupam uma área de 170 m²?

(A) 80 m²
(B) 20 m²
(C) 5 m²
(D) 60 m²

10) CMB) Um feirante comprou 15 quilos (kg) de alho para vender em pacotes de 150 gramas (g). Ao final do dia, ele tinha vendido a metade dos pacotes. Dentre as opções abaixo, a única que apresenta a sequência de operações que determina a quantidade de pacotes que restaram ao final do dia é:

(A) [(15.100) : 150] : 2
(B) [(15:100) : 150] . 2
(C) [(15:1000) : 150] : 2
(D) [(15.1000) : 150] : 2

Anexo II

Dados da planilha MAT8.dat

KEYDABACABCBD
001DABACABCBC
002CDBBDABCBC
003DDCBDBBBAA
004CCBBDDBCAB
005DACCDBBBAC
006DACBDBBCAA
007DCBBBDBCDC
008AABCDBCCDA
009DABBAABBCA
010CADCBADCDA
011DBCACCBBAC
012DBBDDABDAC
013DDBBDBBCAB
014DAACDBBCBB
015DABBDCDCAD
016DADACCBCCC
017CADCAAADDA
018DBBADABCBA
019DBDCBABCAB
020CCBCDDBCBB
021DAACBDAADC
022DCDBBDBCBA
023CCBCDBBCAB
024CBBCDCBCBC
025DCBBDCBCAA
026CBCBDCBCBA
027DAACDCCDAA
028DBCBDCBCAB
029DADBDBCDBA
030BBCCCABAAD
031CABDABBADC
032CBCCCCBCAB
033DABCDABCBC
034DDCDCCBCAB
035CBCBDCADAC
036CBDADDBCAA
037DBBDADCCDD
038DBBBDABCBA
039DABCDDBCAD
040CADBDCCBDA
041CDDBDABCCA
042DABCDABCBA
043DDBCDACACD
044DABCDABABB
045AAACDAACBB
046DACACABDAB
047DABCDBBCDD
048DACDCBBBBA
049DCDBBDDCBB
050DAAAADCBCBB
051DCBABBBCBA
052DCBCDDBCBC
053DBBACDBCBA
054DABBCBBCBD
055DBCBCBBDAB
056DACADCAABA
057DDBACCBDDB
058DABACABCBA
059DABCDCBCDB
060CDBBDBADBC
061DABCDCBDAA
062DABACCBDAB
063DABBCBBCBD
064DDBACABCBB
065DABACCBDAB
066DBBDDCBCBA
067CCBABACCBA
068DDBACCBCBD
069DBBBBDBDAB
070DACBCDBBAA
071DCBCACBCBA
072DDACBDBCCB
073DABDCBCBBD
074DADADABCAB
075DDBCACDAAC
076DABBCCADCB
077DBCCADBDAB
078DAACCBBCAB
079DABCDBBDBB
080BABCABBCCA
081CACBDCBADA
082DABABCBCAD
083BBADDCADBC
084DDCABCAABC
085DBBCDBBCBA
086DBCCABDCCA
087DBCACDBAAC
088DAACDACCDB
089DCCCDABCBA
090BBBCCDACAB
091DBBADBBADD
092DDBCDCBCCB
093DDBBDBBCDB
094CBBBAADCBA
095DBDBCBBCBD
096DABADBBCBC
097DABCDDBCDA
098DABABABDCB
099DABDBCBCBB
100DADBABBCAA
101DCBACABCBD
102DCBACCBCBD
103DCBBABCCAB
104CBBABBBCBD
105DABDBABCDD
106DABADCBCDD
107DABCDDBDDA
108DABADCBCBD
109DAADBAABAA
110DABCDCBBDB
111DACBCABCAD
112CDBCCBBCAB
113DCBADBBCBA
114DACCDABBBC
115DABACABCBA
116DDBDDBBDBB
117CBBCDBBCAA
118DCACACBBAA
119DABBDDBDBB
120DADABCBBAD
121DCBADDBCCB
122DCDABBBCBA
123DAABBBBCAA
124DABACACCCD
125DABCAABBBC
126DBDCDABBAC
127DAAABBBCAA
128DABADCBCAA
129DCBBACBCAC
130DBBBDBBCBA
131DBBCDDBCAA
132DABABCBCBB
133DBAADCDDAB
134DACBCABBAB
135DCBBADBBBA
136DBACBABBDA
137DBCCBCACDB
138ADBCAACBDA
139DDCDCCDBBB
140CABACBBCDA
141DDCDDCBBDA
142DBCBCCBDBC
143DAAABBCAAA
144DBDCDABBDC
145DABADCBBAA
146DBBACABCDA
147DABADABCBB
148DBBCDDBCDB
149DABCBABCAA
150DABADCBDBA
151DDBCCDBBBA
152DABADABCBA
153DBDCDCBCAB
154ABAAABBCCD
155DBBABBBCBB
156DBBABDBCAA
157DAACBBBCAA
158DBBADCBCAB
159DABCDAACBA
160CACBDDBCCB
161DAAADCBDAC
162DABACCCCDA
163DABADBBCBC
164CACCABBCAC
165DBBCDABCBB
166DABCBCBCAD
167DABCBCBBDC
168DACBDCACAB
169DABADCBCAB
170DABADBBCBD
171CDBADBBCBA
172CACCDBABAB
173DAABCCBDBB
174DDCABBBDAB
175BAAADBBBAB
176DBBADABCAC
177DBBABDBBAD
178DABADCBCBC
179CADDCBADBA
180ABCAABDBDA
181DBCADBCCAA
182DABBDCBCAD
183DABBCCBCAA
184DABACABCBD
185AAABCABBAD
186DCBCCBBCBA
187CDBCDDBACC
188DCBDADBCBB
189DACBCABCBC
190DADADABCDA
191DBBCACBCDC
192DBDCBCBCAA
193DABCDABCBB
194DACCBBBBCA
195DDBBDABCBB
196CBBADABCDC
197DDDCDABBBA
198DABCCCBCAC
199DBBCCABCAB
200DBBAACBCBB
201DACCBCBCDB
202DBCBDCBCBD
203DBDAACABBA
204DAACDABCAA
205DADCDDDCAB
206DABCDABCBB
207DADBDDBCBA
208DABBACBBAB
209DACCDCBCDC
210DADCDDBDAA
211DDBBCBCBAB
212DCCCDBBCAA
213CCBBCDBCAB
214CBABAADCCD
215DABBDABCBA
216DADBDCBDBA
217CAAADCBBBD
218DCAAACADDD
219DABCBABCBA
220DACADCBDAB
221DABBDBBCBC
222DBBCDCDDAC
223DABADABCBA
224DACCDCBCAB
225DACBDDBCDA
226DABCADBCAA
227ABBCACBCAB
228DABADABCBD
229CBADCABCDB
230DBCBDCBABC
231DDABBBAADB
232DABBDCBDAC
233DDBCCDBCAB
234DACBDCBDDA
235DBBCBCBCBA
236CABACCBBDB
237DABCCBBCAB
238DCCBDBBCBC
239DABBCABDBB

240DABABBBCBD
241AABCAABBDC
242DACBCBDCDD
243DABAAABCBB
244DACABABDAC
245DACABDBCBC
246DABCAABCBA
247DBBACBBCBC
248DCBAAABCBD
249CDBCCCBBBC
250DABADBBCBA
251DABBACCCDA
252DBDACBADBC
253DABBDCACAB
254DABCDBACDB
255DBBADCBCDA
256DABCCABCDB
257DDBADDBCDA
258CBBAADCDCC
259DBBADDBCBD
260DABACDBCAC
261CBDBDDBBAA
262DBBADDCCDA
263DBBABDCBDC
264DBBDAACDBB
265CBBADABDCA
266DBBABDBDBC
267CAACDABCBB
268BBDDCDDCBA
269DABBDCBCBB
270DBACBDBCAB
271DAABDCBCBA
272CBDCDDBCAA
273DDBCBDBBAA
274DBBBDBBCBA
275DBBBDDCCBA
276DABCDABCDC
277CAACDAACAB
278DBBABABDBB
279DBBADBACAB
280DDBBAADCBD
281DADBDDBDAA
282DABBCBBCDC
283DABCDDBBBB
284DABBDBCCDB
285DABACBBCBB
286DBBACABCAC
287CAACDCBDAB
288DAACBDCDCB
289DDDCDCBCDB
290CABCDABCBD
291DBBCBDBCBB
292DABADABCAD
293DABADABCAD
294DBCADCBCBA
295DABAABBCBA
296DDDCDABDAA
297CABCDABDAA
298DBBACCDCBA
299DDBBDCBCBA
300CDBBCBBCDA

Anexo III

Rotina no software R

```
#Carregamento e conferência da planilha denominada dados
dados = read.fwf('C:\\Users\\Marcos\\Desktop\\Dissertacao\\MAT8.dat',widths=c(3,rep(1,10)),header=F,dec=',')

head(dados)

tail(dados)

View(dados)

#Exclusão da primeira linha e da primeira coluna. Nova planilha dados2
dados2<-dados[-1,-1]

#Inclusão do gabarito
gabarito=c("D","A","B","A","C","A","B","C","B","D")

#Erros e acertos, frequência, ponto-bisserial item total, alfa de Cronbach com exclusões
descript(dad.binario)

#carregamento dos pacotes utilizados
library(psych)
library(CTT)
library(mirt)
library(ltm)

#Transformação da planilha dados2 em valores binários
dad.binario <- key2binary(dados[2:nrow(dados),2:11],t(dados[1,2:11]))

#Confiabilidade (Coeficiente Alfa)
reliability(as.matrix(dad.binario,itemal=TRUE))

#Aplicação do modelo logístico de três parâmetros
meu.modelo.1<-mirt(dad.binario,1,"3PL")

#Cargas Fatoriais
summary(meu.modelo.1)

#TRI - discriminação, dificuldade e acerto ao acaso
coef(meu.modelo.1,simplify=T,IRTpars=T)

#Erros e acertos, frequência, ponto-bisserial item total, alfa de Cronbach com exclusões
descript(dad.binario)

#moda#
statmod<-function(dad.binario){z<-table(as.vector(dad.binario));names(z)[z==max(z)]}
statmod(rowSums(dad.binario))

#mínimo, Q1, mediana, Q3, média, máximo
summary(rowSums(dad.binario))

#desvio padrão
sd(rowSums(dad.binario))

#variância
var(rowSums(dad.binario))

#Curva característica do item 7
itemplot(meu.modelo.1,7,zeros=T)
```

```
#Curva característica do item 10
itemplot(meu.modelo.1,10,zeros=T)

#Curva característica de cada item do teste
plot(meu.modelo.1,type="trace")

#Curva de informação de cada item do teste
plot(meu.modelo.1,type="infotrace", theta_lim=c(-4,4), ylim=c(0,0.5))

#Curva de informação do teste e curva do erro padrão
plot(meu.modelo.1,type="infoSE")

#Curva de informação do item 3 e curva do erro padrão
itemplot(meu.modelo.1,3,type="infoSE")

#Análise dos distratores
distractor.analysis(dados2,gabarito)

#Correlação ponto-bisserial
iA <- itemAnalysis(dad.binario)
```

A aplicação de simulados é uma prática comum quando se deseja avaliar um grupo de estudantes. Porém, para que os resultados apresentados sejam válidos, é necessário que os itens que compõem o simulado sejam capazes de discriminar os participantes que possuem as habilidades que estão sendo medidas dos estudantes que não possuem. Desta forma, o presente trabalho se propôs a analisar a qualidade dos itens de um simulado de Matemática aplicado na rede municipal de educação de Jaraguá do Sul. Esta análise foi feita utilizando e comparando os resultados da Teoria Clássica dos Testes (TCT) e da Teoria de Resposta ao Item (TRI). Concluiu-se que a maior parte dos itens necessita de revisão. Os resultados obtidos com as análises da TCT e da TRI foram confrontados, garantindo uma maior fidedignidade nas conclusões sobre os itens. Os resultados das duas teorias apresentaram divergências em relação aos itens com maior poder de discriminação. Por outro lado, o item considerado mais difícil do teste foi o mesmo nas duas teorias. Além disso, foi realizada uma análise pedagógica nos itens considerados deficientes, com o intuito de contribuir com futuros trabalhos similares. Algumas sugestões de alteração no enunciado da questão e das alternativas foram feitas neste trabalho.

Orientadora: Elisa Henning

Joinville, 2018