UNIVERSIDADE FEDERAL DO OESTE DO PARÁ
INSTITUTO DE CIÊNCIAS DA EDUCAÇÃO
PROGRAMA DE CIÊNCIAS EXATAS

ANDREY CAMURÇA DA SILVA

ANÁLISE PSICOMÉTRICA E PEDAGÓGICA DA PROVA DA PRIMEIRA FASE DA OBMEP RESPONDIDA POR UM GRUPO DE ALUNOS DO 6° E 7° ANO DO ENSINO FUNDAMENTAL

SANTARÉM
2019

ANDREY CAMURÇA DA SILVA

ANÁLISE PSICOMÉTRICA E PEDAGÓGICA DA PROVA DA PRIMEIRA FASE DA OBMEP RESPONDIDA POR UM GRUPO DE ALUNOS DO 6° E 7° ANO DO ENSINO FUNDAMENTAL

Trabalho de Conclusão de Curso apresentado à Universidade Federal do Oeste do Pará como parte das exigências para a obtenção do título de licenciado em Matemática e Física

Orientador: Prof. Dr. Mario Tanaka Filho

Coorientador: Prof. Dr. Claudir Oliveira

SANTARÉM

2019

Dados Internacionais de Catalogação-na-Publicação (CIP)
Sistema Integrado de Bibliotecas – SIBI/UFOPA

S586a Silva, Andrey Camurça da
Análise psicométrica e pedagógica da prova da primeira fase da OBMEP respondida por um grupo de alunos do 6º e 7º ano do ensino fundamental. / Andrey Camurça Silva . – Santarém, Pará, 2019.
85fls.:il.
Inclui bibliografias.

Orientador: Mário Tanaka Filho
Coorientador: Claudir Oliveira
Trabalho de Conclusão de Curso (Graduação) – Universidade Federal do Oeste do Pará, Instituto de Ciências da Educação, Licenciatura em Biologia.

1. Avaliação educacional. 2. OBMEP. 3. Teoria clássica de Teses I. Oliveira, Claudir. II. Tanaka Filho, Mário, *orient*. III. Título.

CDD: 23 ed. 530

Bibliotecária - Documentalista: Selma Souza – CRB/2 1.096

ANDREY CAMURÇA DA SILVA

ANÁLISE PSICOMÉTRICA E PEDAGÓGICA DA PROVA DA PRIMEIRA FASE DA OBMEP RESPONDIDA POR UM GRUPO DE ALUNOS DO 6 E 7 ANO DO ENSINO FUNDAMENTAL

Trabalho de Conclusão de Curso apresentado à Universidade Federal do Oeste do Pará como parte das exigências para a obtenção do título de licenciado em Matemática e Física

Data da Aprovação: 11 de dezembro de 2019

Prof. Dr. Mario Tanaka Filho
Orientador
UFOPA-ICED

Prof. Dr. Claudir Oliveira
Coorientador
UFOPA-ICED

Prof. Dr. Edilan de Sant'Ana Quaresma
UFOPA - ICED

Prof. Me. Tarcísio da Costa Lobato
UFOPA - ICS

Dedico este trabalho aos meus pais
Amarildo Lemos da Silva e Eurides Camurça Lemos.

Agradecimentos

Agradeço minha família, em especial minha irmã Eurice Camurça e meus pais, que sempre me deram apoio para que eu pudesse continuar firme nos estudos. Agradeço a Erlene Sousa por ter me dado apoio financeiro nos momentos mais difíceis. Ao mesmo tempo, agradeço meus colegas de aula, especialmente ao Tamilson, Marcos Olivetto e Juliane. E claro, agradeço aos professores Sebastian Mancuso, Glauco, Aroldo, Mario Tanaka e todos os demais que se dedicaram e se dedicam à missão de oferecer um curso de licenciatura de qualidade.

Resumo

A Olimpíadas de Matemática das Escolas Públicas (OBMEP) é uma política educacional importante para a difusão dos conhecimentos matemáticos nas escolas, sendo ela composta por provas e programas de treinamento específico presentes em todo o Brasil. As provas da OBMEP da primeira fase são constituídas por itens de múltipla escolha, cuja finalidade é meramente classificatória. Não se realiza, assim, uma avaliação da aprendizagem nos moldes do Sistema de Avaliação da Educação Básica (Saeb), em que se posiciona as escolas e/ou indivíduo em níveis de aprendizagem. Tendo em vista a pouca utilização do referido certame para o diagnóstico da aprendizagem dos alunos em matemática, o presente trabalho considera para a composição do banco de dados as respostas de um grupo de aluno da rede pública de Santarém (PA) que realizaram a prova de nível 1 da 13ª edição da competição. O objetivo foi apresentar uma análise psicométrica e pedagógica dos itens da prova da primeira fase OBMEP, nível 1, 13ª edição, com base na Teoria Clássica dos Testes e Teoria de Resposta ao Item (TRI). Os resultados expressos por meio de índices da TCT e TRI mostraram que o certame possui baixa qualidade para a amostra considerada, sobretudo no que se refere a consistência interna do instrumento. Pelo menos 7 itens não se adequaram aos pressupostos da TRI e que foram analisados individualmente. O estudo mostrou que a extração de informações pedagógicas da prova da OBMEP pode fornecer *feedback* sobre dificuldades e erros que os alunos comentem por não terem adquirido a proficiência exigida para a resolução dos problemas matemáticos propostos – informações que são relevantes tanto para a formação continuada de professores quanto para os cursos de Licenciatura em Matemática.

Palavras-chaves: Avaliação Educacional, Olimpíadas Brasileira de Matemática das Escolas Públicas, Teoria Clássica de Testes, Teoria de Resposta ao Item.

Abstract

The Public Schools Mathematics Olympics (OBMEP) is an important educational policy for the dissemination of mathematical knowledge in schools. It consists of tests and specific training programs present throughout Brazil. The OBMEP tests of the first phase consist of multiple choice items, the purpose of which is merely classificatory. Thus, there is no assessment of learning along the lines of the Basic Education Assessment System (Saeb), in which schools and/or individuals are positioned at learning levels. Given the low use of this event for the diagnosis of students' learning in mathematics, the present work considers for the composition of the database the answers of a group of students from the public schools of Santarém (PA) who took the test. level 1 of the 13th edition of the competition. The objective was to present a psychometric and pedagogical analysis of the first phase OBMEP test items, level 1, 13th edition, based on the Classical Test Theory (CTT) and Item Response Theory (IRT). The results expressed by CTT and IRT indices showed that the event has low quality for the sample considered, especially regarding the internal consistency of the instrument. At least 7 items did not fit the IRT assumptions and were analyzed individually. The study showed that extracting pedagogical information from the OBMEP test can provide feedback on difficulties and errors that students comment for not having acquired the proficiency required to solve the proposed mathematical problems – information that is relevant to both. the continuing education of teachers as for undergraduate mathematics courses.

Key-words:ducational Assessment, Brazilian Olympics for Public Schools Mathematics, Classical Test Theory, Item Response Theory.

Lista de ilustrações

Lista de tabelas

Sumário

1 INTRODUÇÃO

Mais do que uma competição matemática, a Olimpíadas Brasileira de Matemática das Escolas Públicas e Privadas (OBMEP) é uma política pública relevante para a educação brasileira. Ela é realizada pelo Instituto Nacional de Matemática Pura e Aplicada (IMPA), juntamente com Ministério da Educação e Ministério da Ciência, Tecnologia, Inovação e Comunicações (MCTI), visando estimular o estudo da matemática e revelar talentos na área, contribuindo assim para o aperfeiçoamento dos processos de ensino e aprendizagem em matemática (MARANHÃO, 2011).

Na forma mais sintética, a OBMEP tem sua culminância por meio de provas aplicadas em duas fases. No primeiro momento aplicam-se testes com 20 itens (questões) de múltipla escolha, cuja finalidade é selecionar os alunos com maior número de acertos. Até aí nota-se o caráter classificatório da avaliação, suja correção não utiliza métodos sofisticados e tampouco são difundidos os seus resultados em âmbito nacional para a verificação dos níveis de habilidades atingidos pelos alunos na área de matemática. De todo modo, as provas da competição mobilizam professores e até escolas inteiras na busca por bons desempenhos, ganhando reforço com programas de treinamento específico da OBMEP destinados a alunos premiados. Tudo isso apresenta impactos positivos na qualidade da educação brasileira, conforme mostra (MARANHÃO, 2011).

A segunda fase, por outro lado, é composta por itens de perguntas abertas, que avaliam a capacidade do aluno propor soluções para uma certa variedade de problemas matemáticos. Nessa etapa participam os alunos com melhores desempenhos, por escola, da primeira fase. As provas avaliam habilidades previstas no Parâmetro Curricular Nacional de Matemática, embora, talvez, o tipo de tarefas/itens cobradas nas provas não sejam, com frequência, tratados no contexto da sala de aula.

Embora se concretize por meio de provas, a OBMEP não é uma política de avaliação da educação básica. No estudo de caso realizado por Costa (2015), coloca-se em questão a possibilidade de incorporar aos objetivos da OBMEP dimensões da avaliação da aprendizagem, visando o melhor aperfeiçoamento e aproveitamento das provas. É importante ressaltar que a OBMEP possui alcance de cerca de 99% dos municípios brasileiros (MARANHÃO, 2011).

Paralelo a isso, viu-se nas últimas três décadas a realização de importantes avaliações externas conduzidas pelo sistema de Avaliação da Educação Básica do Brasil (Saeb). Diferente da OBMEP, essas avaliações utilizam testes bem elaborados do ponto de vista psicométrico e pedagógico, os quais assumem o papel de diagnosticar sucessos e fracassos do sistema educacional, tendo como contribuições avanços científicos, metodológicos e tecnológicos agregados ao campo da avaliação educacional (FONTANIVE, 2005).

Segundo Fontanive (2005, p.139), “a medida em que aumenta o interesse pela ava-

liação no cenário educacional cresce também, em nossos dias, a discussão sobre o uso dos testes como medida do processo educativo". Desde suas origens, os testes foram empregados na mensuração de atributos psicológicos e educacionais, cujas metodologias começaram a ser desenvolvidas no campo da psicometria no final do século XIX (PASQUALI, 2017).

O uso dos testes padronizados de múltipla escolha tornaram-se presentes a partir dos anos 70 e foram ligeiramente incorporados nos concursos e vestibulares. As avaliações externas de sistemas escolares aderiram a essa metodologia, dando ênfase aos aspectos mensuráveis do currículo. Esse, talvez, seja o motivo pelo qual esse tipo de avaliação é submetida a importantes críticas por parte de alguns especialistas em avaliação educacional (FONTANIVE, 2005).

As críticas atraídas pelas avaliações externas e seus usos equivocados, chama atenção para a necessidade de se pensar numa avaliação que promova de fato mudanças na qualidade do ensino, capazes de orientar a prática docente. Além do que hoje se entende como avaliação formativa, cujo papel é fundamental nas relações de ensino e aprendizagem, Klein (2005), Costa (2015) e Fontanive (2005) apontam que a análise pedagógica das questões (itens), com apoio de análises estatísticas/psicométricas dos testes, são fontes de informações importante para o diagnostico de dificuldades enfrentadas pelos alunos. É possível com essas análises observar erros que os alunos cometem com frequência e habilidades que eles ainda não conseguiram atingir.

Atualmente são poucas as pesquisas que ligam a atividade docente ao avanço do desempenho dos alunos, por meio de avaliações externas. "Essa não é uma área de fácil investigação, pois exige o uso de instrumentos variados que requerem um grande investimento de tempo para desenvolvê-los e validá-los, treinar pessoas para aplicá-los, analisar e consistir os bancos de dados" (FONTANIVE, 2013, p.15).

Apesar de possuir limitações, estudos exploratórios nesse sentido podem produzir resultados customizados sobre a situação da aprendizagem dos alunos. Klein (2005) mostra exemplos de como os índices produzidos no âmbito da Teoria Clássica de Testes (TCT) e Teoria de Resposta ao Item (TRI) indicam erros e falhas de aprendizagem dos alunos em alguns itens que versavam sobre número racional.

Estudos mais recentes realizados por Costa (2015), Vilarinho (2015) e Silva (2019) discutiram resultados de aplicações da TCT e TRI aos dados da OBMEP, como meio de fornecer *feedback* aos professores sobre dificuldades, desempenho e erros que os alunos comentem na prova da primeira fase da OBMEP, visando o aperfeiçoamento da prática docente.

O estudo exploratório desenvolvido por Silva (2019) mostrou que a prova da 1ª fase da competição, 13ª edição, respondida por um grupo de alunos da rede municipal de ensino da cidade de Santarém, no estado do Pará, teve baixa qualidade global dos

itens, traduzido por meio de elevadas medidas de dificuldade dos itens, baixa consistência e poder discriminativo do teste.

Na investigação conduzida por Vilarinho (2015), também com dados da 1ª fase da OBMEP, mas desta vez oriundos de cinco escolas públicas da cidade de Brasília, mostraram o mesmo problema. Nos dois estudos os alunos participantes demonstraram falta de habilidade em grande parte dos itens.

Costa (2015) chama atenção para a necessidade de se utilizar a prova da primeira fase da OBMEP para além de um mecanismo de classificação baseado apenas no paradigma da TCT, mas como uma ferramenta de diagnóstico da aprendizagem, capaz de orientar a prática docente e aperfeiçoar o ensino da Matemática.

Nesse sentido, o presente trabalho contribui com as pesquisas conduzidas nesse tema, considerando como objeto de estudo as respostas de um grupo de aluno da rede pública de Santarém (PA) que realizaram a prova de nível 1 da 13ª edição da OBMEP.

O objetivo deste trabalho é apresentar uma análise psicométrica e pedagógica dos itens da prova da primeira fase OBMEP, nível 1, 13ª edição, com base na Teoria Clássica dos Testes e Teoria de Resposta ao Item (TRI). Para tanto, pretense-se: apresentar as características e relevância da OBMEP no contexto do ensino e da avaliação em matemática; discutir os principais impactos das teorias psicométricas no âmbito da avaliação educacional e seu desenvolvimento ao longo da história; aplicar os pressupostos da TCT e TRI na análise da qualidade dos itens e da prova, respondida por grupo de alunos do 6° e 7° ano da rede pública de ensino de Santarém, no estado do Pará; analisar os itens que se destacam no estudo psicométrico, numa perspectiva pedagógica, buscando identificar temas e habilidades nos quais os alunos apresentam deficiência de aprendizagem.

O trabalho está organizado em mais cinco capítulos, sendo que no Capítulo 2 é apresentado um apanhado histórico das competições matemáticas até a culminância da OBMEP em 2005, a qual tem suas características brevemente apresentada. O Capítulo 3 discute resultados teóricos, percurso histórico e aplicações da Psicometria no contexto da avaliação. Uma vez apresentado todo arcabouço teórico, o Capítulo 4 trata dos procedimentos metodológicos. No capítulo seguinte apresenta-se e discute-se os resultados do estudo exploratório. Por fim, faz-se as conclusões no Capítulo 6.

2 OLIMPÍADAS BRASILEIRA DE MATEMÁTICA DAS ESCOLAS PÚBLICAS

Este capítulo apresenta um apanhado histórico das competições na área da matemática, dando enfoque, sobretudo, na Olimpíadas Brasileira de Matemática das Escolas Públicas (OBMEP), hoje considerada uma das mais abrangentes e importantes políticas de difusão da matemática e seu ensino no Brasil. Tratamos especialmente das características da competição, tais como abrangência, níveis e elaboração.

Realizadas na maioria das vezes com intuito de mobilizar estudantes e matemáticos para tratar de uma variedade de problemas, as competições de matemática são organizadas desde longo tempo. Segundo Maciel e Basso (2009), os desafios nos quais importantes matemáticos submetiam seus talentos em busca de soluções universalmente aceitas vem desde o século XVI, e eram promovidas por instituições europeias ligadas ao campo da ciência. Como recompensa, tais instituições, principalmente universidades, ofereciam quantias em dinheiro e até mesmo posições de destaque nas suas cátedras.

Por volta do século XIX, matemáticos húngaros passaram a organizar competições evolvendo desafios matemáticos chamadas *Eotvos*. Tais eventos são entendidos como a gênese do que hoje se entende por Olimpíadas de Matemática (OM). Em 1934, foi a vez dos soviéticos organizarem um evento com características das atuais OM, tendo sua culminância na cidade de Leningrado (URSS), atual São Petersburgo (Rússia). Mais tarde, em 1959, a primeira Olimpíada Internacional de Matemática (IMO) foi organizada na cidade de Bucareste (Romênia) (MACIEL; BASSO, 2009).

A primeira edição da IMO foi organizada no escopo das competições modernas, mas com alcance ainda restrito. Possibilitou o ingresso de vários países que podiam enviar até 8 participantes, mas no total, obteve participação de 7 países, dentre eles Bulgária, Hungria, Alemanha e União Soviética. Os anfitriões foram os campões, obtendo a primeira colocação, seguido da Hungria, que ficou em segundo lugar (IMO, 2015).

As edições seguintes da IMO continuaram tendo pequena abrangência, geralmente atraindo equipes europeias. A primeira equipe das Américas a enviar participantes foi Cuba, em 1971, seguido do Brasil que, apenas em 1979, teve sua primeira participação no evento, obtendo a penúltima colocação. Dois anos depois, todos os continentes estavam representados, incluindo Austrália e países africanos (IMO, 2015).

No Brasil, o inicio das competições foram introduzidas bem mais tarde. Somente em 1977 a Academia Paulista de Ciências criou a Olimpíada Paulista de Matemática, cujo objetivo era:

> 1) Incentivar o ensino de Matemática. 2) Proporcionar o entrosamento dos professores de Matemática de uma mesma escola e entre os de di-

> ferentes escolas de uma mesma região. 3) Favorecer a participação da comunidade local em problemas e atividades educacionais de jovens em idade escolar. 4) Avaliação e rendimento do ensino de Matemática no Estado (DUARTE; GALVÃO, 2014).

Dois anos mais tarde a Sociedade Brasileira de Matemática (SBM) reuniu esforço para criar uma competição a nível nacional, intitulada de Olimpíada Brasileira de Matemática (OBM) (OBM, 2010).

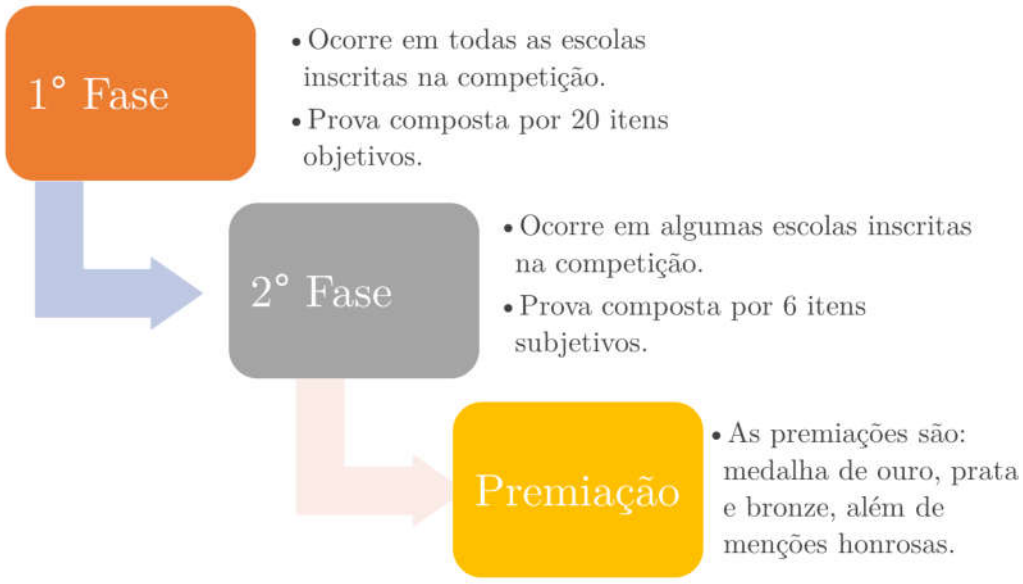
A consolidação da OBM fortaleceu a disseminação da matemática no país e abriu precedentes para a criação de uma OM de grande proporção, capaz de atingir o alunado da maioria das escolas públicas. Tal feito culminou em 2005 com criação da OBMEP.

A OBMEP é realizada pelo Instituto Nacional de Matemática Pura e Aplicada (IMPA), juntamente com Ministério da Educação e Ministério da Ciência, Tecnologia, Inovação e Comunicações (MCTI), visando estimular o estudo da matemática e revelar talentos na área.

A competição tem como público alvo alunos do 6° ao 9° ano do ensino fundamental e alunos do Ensino Médio. Para atender a esse público, a competição foi dividida em três níveis de provas. A prova do nível 1 é respondida por alunos do 6° e 7° ano do Ensino Fundamental, enquanto a prova do nível 2 tem como alvo os alunos do 8° e 9° ano. E por fim, participam do nível 3 os candidatos matriculados em qualquer ano do Ensino Médio.

O evento inclui três etapas, sendo duas fases de prova e um evento de premiação, conforme apresenta a Figura 1.

Figura 1 – Fases da OBMEP.



Fonte: OBMEP (2018)

Para se ter uma noção da grandeza da OBMEP, a primeira edição da competição

teve mais de 10 milhões de inscritos, oriundos de mais de 30 mil escolas públicas do País. De lá até aqui o número de participantes aumentou consideravelmente. Na edição de 2017 esse número atingiu mais de 18 milhões de inscritos, englobando, pela primeira vez, candidatos das escolas privadas.

A Tabela 1 mostra, com mais detalhe, o número de escolas inscritas desde a primeira edição da competição (OBMEP, 2018).

Tabela 1 – Número de escolas, alunos inscritos e percentual de municípios participantes da OBMEP, 1 fase, desde a primeira edição.

Ano	Escolas	Alunos	Municípios (%)
2005	31 031	10 520 831	93,50
2006	32 655	14 181 705	94,50
2007	38 450	17 341 732	98,10
2008	40 397	18 326 029	98,70
2009	43 854	19 198 710	99,10
2010	44 717	19 665 928	99,16
2011	44 691	18 720 068	98,90
2012	46 728	19 166 371	99,42
2013	47 144	18 762 859	99,35
2014	46 711	18 192 526	99,41
2015	47 580	17 972 333	99,48
2016	47 474	17 839 424	99,59
2017	53 231	18 240 497	99,57

Fonte: OBMEP (2018)

Em relação a sua primeira edição ocorrida no ano de 2005, a competição realizada em 2017 teve um aumento de 73,4% no número de participantes inscritos. A quantidade de escolas inscritas também teve um aumento de 71,5% no período citado, revelando que a OBMEP tem aumentado seu alcance.

Na percepção de Maranhão (2011), a OBMEP é entendida como

> (...) uma política pública mundialmente reconhecida, uma das maiores iniciativas governamentais voltadas ao processo de ensino-aprendizagem em matemática, visando melhorar a motivação, o interesse e o desempenho dos alunos nas escolas públicas brasileiras (MARANHÃO, 2010, p.13).

A autora se refere à OBMEP não apenas como uma competição, considerando-a uma política relevante para o processo de ensino e aprendizagem em matemática no país, a qual tem como objetivos principais:

- Estimular e promover o estudo da matemática entre alunos das escolas públicas;
- Contribuir para a melhoria da qualidade da educação básica;

- Identificar os jovens talentos e incentivar seu ingresso nas áreas científicas e tecnológicas;
- Incentivar o aperfeiçoamento dos professores das escolas públicas, contribuindo para a sua valorização profissional;
- Contribuir para a integração das escolas públicas com as universidades públicas, os institutos de pesquisa e sociedades científicas;
- Promover a inclusão social por meio da difusão do conhecimento (OBMEP, 2018).

Para promover o estudo da matemática, a OBMEP possui programas de treinamentos específicos. O programa de Iniciação Científica Jr. (PIC), por exemplo, fornece material e tutoria para alunos premiados na competição, que conta ainda com bolsas do Conselho Nacional de Desenvolvimento Científico e Tecnológico (Cnpq), no valor de R$ 100,00. Um programa mais avançado, intitulado Programa de Mentoria, insere os alunos no contexto das universidades federais para serem tutorados por professores universitários. Nesse ambiente, os alunos estudam tópicos avançados de matemática e resolvem problemas voltados a Matemática Olímpica (OBMEP, 2018). Ao longo de suas edições, os programas citados já encaminharam cerca de 47 mil alunos da rede pública de ensino a aulas de matemática com tutores presenciais e professores universitários, promovendo a inclusão social e o aperfeiçoamento acadêmicos (BRASIL, 2018).

Com investimentos em divulgação e elevado alcance, a OBMEP tem influenciado escolas públicas em vários pontos do Brasil a criar uma cultura de estudos voltados para problemas matemáticos que estimulam a criatividade a autonomia dos alunos. Essas mudanças foram potencializadas com o programa OBMEP nas Escolas, criado em 2015. Voltados para professores de Matemática das escolas públicas, o programa estimula atividades extraclasse, compostas por aulas presenciais, estudo do bando de questões OBMEP, apostilas, videoaulas e livros produzidos ou divulgados pela OBMEP. Para tanto, o programa oferece capacitação aos professores para desenvolver atividades em sua escola ou em outras escolas da mesma cidade (CRUZEIRO, 2018; COSTA, 2015; SILVA, 2019; OBMEP, 2018; VILARINHO, 2015).

Cabe ressaltar que as contribuições da OBMEP para a melhoria da qualidade da educação brasileira vão além de seus programas de ensino. Suas provas podem funcionar como gatilho de buscas por novas concepções e abordagens matemáticas capazes de estimular o posicionamento crítico por parte de professores e alunos, em contraposição ao paradigma da *monumentalização do saber* – metáfora utilizada pelo pesquisador francês Yves Chevallard para se referir ao ensino que funciona como uma visita às obras (conteúdos), sem levar em conta sua utilidade e razão de ser (OTERO et al., 2013).

Nesse sentido, acredita-se que as ações conduzidas pela OBMEP produzem resultados positivos nos campos do ensino, aprendizagem e da avaliação na área da matemática,

uma vez que cidades e escolas que possuem um número expressivo de alunos premiados, chama atenção das autoridades e em alguns casos é tomado como indicador de qualidade da educação matemática empreendida, como no caso de Cocal dos Alves, município do interior do Piauí de pouco mais de 6 mil habitantes que barganhou 172 premiações até o ano de 2015 (OBMEP, 2019). Apesar de possuir um Índice de Desenvolvimento Humano (IDH) posicionado entre os 50 piores do Brasil, o município apresentou, em 2017, um Ideb (escrever por extenso) nos anos finais do Ensino Fundamental de 6,4 – índice superior à meta estabelecida (4,9) (QEDU, 2019).

Tais afirmações podem ser reforçadas por meio do estudo de impacto da OBMEP na qualidade da educação e econômica, conduzido por Biondi, Vasconcellos e MENEZES-FILHO (2009). A pesquisa utilizou dados do Sistema de Avaliação da Educação Básica (Saeb) e envolveu alunos da 8° série (atual 9° ano do Ensino Fundamental), constatando-se impacto positivo e significativo da OBMEP nas notas médias de Matemática obtidas pelos alunos na Prova Brasil de 2007. O mesmo estudo revelou que a competição "proporciona benefícios para a qualidade da educação pública do país, com impacto direto nas avaliações educacionais e ganhos futuros em termos de rendimento no mercado de trabalho dos participantes" (BIONDI; VASCONCELLOS; MENEZES-FILHO, 2009, p. 11).

Por outro lado, Silva (2019), Costa (2015) questionam os métodos de como são avaliadas as respostas dos alunos na primeira fase da competição – hoje caracterizada pela simples soma do número de acertos, baseado na paradigma da TCT. Os autores evidenciam que o certame poderia também ser usado como ferramenta avaliativa, visando um melhor aproveitamento e aperfeiçoamento do ensino e das práticas docentes. Nesse sentido, torna-se pertinente o uso das respostas dos alunos à prova da OBMEP para fins de análises sob diversos enfoques metodológicos, tendo em vista que os resultados inerentes à primeira fase da competição não têm sido amplamente utilizados para fornecer *feedback* ao professor da educação básica.

No capítulo seguinte apresenta-se um rol de teorias e modelos úteis para avaliar tanto o desempenho dos alunos quanto dificuldades que eles possuem no certame. As teorias apresentadas produzem, assim, resultados mais detalhados sobre proficiências e habilidades dos alunos que se baseiam em suas respostas a testes.

3 MEDIDA DE PROFICIÊNCIA

Parcela dos avanços teóricos em avaliação de habilidades, dentre elas a medida de proficiência escolar, obtidas por meio de métodos quantitativos, nasceu com o advento da psicometria, cujas formulações podem ser aplicadas tanto às avaliações psicológicas quanto ao marketing, agronomia, nutrição, qualidade de vida e nas avaliações educacionais conduzidas em larga escala (SARTES; SOUZA-FORMIGONI, 2013; ANDRADE; TAVARES; VALLE, 2000).

Neste capítulo discute-se duas abordagens da psicometria que são importantes para a consecução de avaliações e testes escolares, a saber, a Teoria Clássica de Testes (TCT), formulada no âmbito da psicometria clássica, e a Teoria de Resposta ao Item (TRI) – por muitos concebida como a "psicometria moderna". Antes, apresenta-se um breve histórico da psicometria, uma vez que os fundamentos teóricos e metodológicos usados para a realização de avaliações escolares tiveram origem nesse campo da psicologia.

3.1 ASPECTOS HISTÓRICOS DA PSICOMETRIA

O desenvolvimento da psicometria trouxe avanços expoentes no estudo de respostas fornecidas por um indivíduo à testes compostos por estímulos (perguntas) e respostas. Na sua maioria, os testes tinham finalidades específicas, sendo usados no estudo do comportamento, inteligência e aptidão (PASQUALI, 2017). De acordo com Sartes e Souza-Formigoni (2013, p. 241), "o desenvolvimento de instrumentos de avaliação psicológica se iniciou no século XIX, paralelamente ao avanço da ciência positivista e da necessidade de medidas objetivas e válidas para o desenvolvimento de pesquisas clínicas".

No início do século XX, os primeiros esforços foram direcionados ao desenvolvimento de métodos que avaliassem as propriedades psicométricas dos instrumentos. Tais propriedades resumem-se, precisamente, à validade e à confiabilidade dos testes, tendo em vista a crescente demanda por testagens objetivas, no âmbito de estudos psicológicos.

Na mesma época, as vertentes da psicologia de orientação empiricista e a psicologia mentalista de *Binet* vinham se desenvolvendo, mas sem contribuições expoentes para ser considerada a base do campo da psicometria (PASQUALI, 2017). De acordo com Pasquali (2017), as origens da psicometria estão presentes nos trabalhos fisicalistas de Galton (1883). Adiante, por volta de 1905, as contribuições do inglês Charles Spearman deram os primeiros passos para a formalização estatística da Psicometria (NASCIMENTO; RUEDA, 2014; PASQUALI, 2017). Segundo Nascimento e Rueda (2014, p.308), "(...) Spearman investigou se as habilidades intelectuais se correlacionavam entre si e se eram dependentes ou independentes de um fator geral, comum a todas elas".

> Não se deve estranhar que a Psicometria surgisse no campo das aptidões

> humanas (mentais, físicas, psicofísicas), pois, além de ser a temática psicológica da época, se coadunava melhor a um estudo quantitativo, pois se pode ali contabilizar o comportamento em termos de acertos e erros(PASQUALI, 2017, p. 14).

Ainda segundo Pasquali (2017), o desenvolvimento da psicometria se dava segundo dois enfoques, sendo um mais prático (experimental) e o outro voltado às questões teóricas da própria psicometria. No primeiro enfoque, psicólogos lidavam com questões psicopedagógicas e clínicas, visando detectar fenômenos psicológicos, dentre eles os então chamados retardos mentais. O segundo enfoque, ainda que desenvolvido por psicólogos, tinha uma tendência fundamentalmente estatística, e por isso a psicometria clássica se desenvolveu com rigor teórico e metodológico nas primeiras décadas do século XX. A Figura 2 mostra algumas eras importantes até o surgimento da Psicometria Moderna.

Figura 2 – Linha do tempo da Psicometria Clássica

Fonte: Adaptado de (PASQUALI, 2017).

Com o avanço gradativo da psicometria entre 1880 e 1940, vários testes foram ganhando popularidade no estudo das aptidões, personalidade e inteligência. Por volta de 1940, Louis Leon Thurstone deu impulso inovador com o uso da análise fatorial ao estudo das aptidões. Entre 1940 e 1980, a sistematização da teoria clássica dos testes psicológicos e da análise fatorial proporcionou avanços significativos na psicometria. Paralelo aos avanços da teoria clássica, Lord e Novick (1968) deram início à teoria do traço latente, em parte apresentada na obra *Statistical Theory of Mental Tests Scores*, e que mais tarde culminou com a teoria moderna da Psicometria, a Teoria de Resposta ao Item (TRI), proposta por (LORD, 1980 apud PASQUALI, 2017).

A partir de 1980, a TRI teve protagonismo no entendimento de características psicológicas não observáveis, que na literatura especializada ganha o nome de traço latente. No âmbito da educação, também possibilitou avaliações que permitem comparação entre indivíduos que não foram submetidos ao mesmo teste (ANDRADE; TAVARES; VALLE,

2000). Ainda assim, a TRI não substitui toda a Psicometria clássica, mas apenas partes dela (PASQUALI, 2017).

Ainda de acordo com Pasquali (2017), a TRI define a qualidade dos testes (variáveis observáveis) em função das variáveis latentes (θ), o que rompe com a Psicometria tradicional, concebida nestes trabalho como Teoria Clássica dos Testes (TCT), que define a qualidade dos testes em função de comportamentos presentes ou futuros. A próxima subseção tratará com mais detalhes o modelo da TCT, aplicação no âmbito da educação, seus pressupostos e limitações.

3.2 TEORIA CLÁSSICA DE TESTES (TCT)

Inicialmente é estabelecido os principais conceitos para o entendimento do modelo da TCT dentro do contexto da avaliação em larga escala. Considera-se N indivíduos, de índice j, $j = 1, 2, 3 \cdots, n$, submetidos a um conjunto de tarefas (item) i, $i = 1, 2, 3, \cdots, I$. As respostas dadas serão denotadas pela variável aleatória U_{ij} dicotômica que assume os valores 1, quando o indivíduo j responde corretamente o item i, ou 0 no caso em que o indivíduo erra o item i. A soma das respostas corretas de cada indivíduo é o chamamos de escore bruto T_i. Genericamente, a escala de uma prova ou teste pode ser definida pelo escore T_i, conforme representa a Equação 3.1,

$$T_i = \sum_{j=1}^{n} \xi_j U_{ij} \tag{3.1}$$

em que ξ representa o peso de cada item. Na TCT, temos interesse por ξ constante para todos os j (FLETCHER, 2010).

O foco da TCT é responder os significados dos escores T_i para a população analisada, não cabendo a ela avaliações individualizadas dos itens. Seu pressuposto básico é de que o escore não corresponde, necessariamente, à habilidade do indivíduo, dado que, como todo processo de medição empírica, a magnitude do escore empírico deve conter uma porção de erros (PASQUALI, 2017).

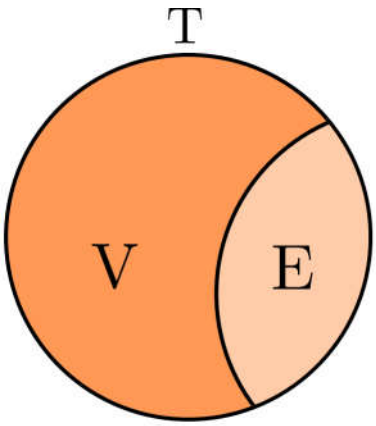
Mais recentemente o objetivo da TCT é estimar os erros associados aos testes, sobretudo em suas aplicações na avaliação educacional. De acordo com Pasquali (2017), o modelo clássico possui limitações na avaliação da proficiência do aluno, mas pode apontar a capacidade preditiva da ferramenta de avaliação. Em outras palavras, tem como foco questões relacionadas à qualidade do teste e é isso que está em discussão nesta seção.

Os postulados básicos da psicometria clássica se desdobra a partir do modelo

$$T = V + E \tag{3.2}$$

em que T corresponde ao escore bruto, V o escore verdadeiro e E o erro. De forma mais detalhada:

Figura 3 – Componentes do modelo fundamental da psicometria clássica.



Fonte: Pasquali (2017)

- T = escore bruto ou empírico do sujeito, que é a soma dos pontos obtidos no teste;
- V = escore verdadeiro, que seria a magnitude real daquilo que o teste quer medir no sujeito e que seria o próprio T se não houvesse o erro de medida;
- E = o erro cometido nesta medida.

Esse modelo foi sintetizado por Gulliksen (1950), no qual assume-se que a diferença entre o escore bruto e verdadeiro é o próprio erro, conforme mostra a Figura 3.

3.2.1 Pressupostos e derivações do modelo da TCT

Uma característica notável do modelo clássico é que, a princípio, não se conhece o erro E e consequentemente o escore verdadeiro V, gerando uma equação de duas incógnitas, o que inviabiliza a determinação do escore verdadeiro. No entanto, quando aplicado a um número grande de indivíduos, as distribuições de frequências em T, V e E, podem trazer resultados razoáveis sobre as características das mesmas (PASQUALI, 2017).

A aplicação a um número elevado de sujeito produz três distribuições de frequências, a saber, de T_i, V_i e E_i. Então, analisando estas três distribuições, pode-se fazer estimativas importantes a respeito dos parâmetros envolvidos no teste.

As consequências estatísticas do modelo clássico, expresso por meio da Equação 3.2, são listados como propriedades numerada de 1 a 3:

1. $\mu(E|T) = 0$. A esperança do erro de escore é zero;
2. $\rho(T, E) = 0$. O escore verdadeiro e o erro de escore são não correlacionados;
3. $\rho(E_1, E_2) = 0$. Os erros de escores de aplicações distintas não estão correlacionados.

Pasquali (2017) explica que essas suposições são válidas para um número suficientemente grande de aplicações do teste (τ). A propriedade (1) decorre do fato dos erros serem considerados assistemáticos. A segunda propriedade é tomada como verdadeira porque não há qualquer indicativo de que escores verdadeiros maiores terão erros positivos. A propriedade 3 resulta da aleatoriedade dos erros, uma vez que os erros cometidos no teste τ_1 são independentes dos erros cometidos no teste τ_2.

As derivações estatísticas desses pressupostos dizem respeito à média dos escores, variância e correlação entre escore e o erro. A primeira propriedade listada tem como imediata consequência o fato de que a esperança dos escore bruto é igual a esperança matemática do escore verdadeiro. Uma simples demonstração pode ser feita. Considere a Equação 3.2.

$$\sum_{i=1}^{N} \frac{T_i}{N} = \sum_{i=1}^{N} \frac{V_i}{N} + \sum_{i=1}^{N} \frac{E_i}{N}. \tag{3.3}$$

De acordo com a propriedade 1 do modelo elementar da TCT, a esperança matemática do erro é nula, e portanto a Equação 3.3 se reduz à

$$\sum_{i=1}^{N} \frac{T_i}{N} = \sum_{i=1}^{N} \frac{V_i}{N}. \tag{3.4}$$

Assim, conclui-se que o escore esperado é igual ao escore verdadeiro. Por outro lado, a propriedade 2 tem importantes consequências para o entendimento da variância dos escores. Considera-se o pressuposto básico (Equação 3.2) e a variância do escore bruto nas equações que seguem

$$\begin{aligned} \sigma_T^2 &= \sum_{i=1}^{N} \frac{(T_i - \bar{T})^2}{N} && (3.5) \\ &= \sum_{i=1}^{N} \left[\frac{(V_i - \bar{V})}{N} + \frac{(E_i - \bar{E})}{N} \right]^2. && (3.6) \end{aligned}$$

Substituindo os desvios, $(T_i - \bar{T})$ por t_i, $(V_i - \bar{V})$ por v_i e $(E_i - \bar{E})$ por e_i, se obtém

$$\begin{aligned} \sigma_T^2 &= \sum_{i=1}^{N} \frac{t_i^2}{N} && (3.7) \\ &= \sum_{i=1}^{N} \frac{(v_i + e_i)^2}{N} && (3.8) \\ &= \bar{v^2} + 2\bar{v}\bar{e} + \bar{e^2} && (3.9) \end{aligned}$$

Considerando que a covariância $\bar{v}\bar{e}$ é nula, a Equação 3.9 pode ser rescrita na forma

$$\sigma_T^2 = \bar{v^2} + \bar{e^2} \tag{3.10}$$

$$= \sum_{i=1}^{N} \frac{(V_i - \bar{V})^2}{N} + \sum_{i=1}^{N} \frac{(E_i - \bar{E})^2}{N} \tag{3.11}$$

Logo, conclui-se que a variância do escore bruto é igual a soma das variância do escore e do erro, conforme expresso na Equação 3.12

$$\sigma_T^2 = \sigma_V^2 + \sigma_E^2. \tag{3.12}$$

A correlação entre escores é outra derivação importante do modelo. Por definição, a correlação entre os escore empírico T e verdadeiro V é

$$\rho_{TV} = \frac{cov(T, V)}{\sigma_T \sigma_V}. \tag{3.13}$$

Usando a mesma notação para os desvios, tem-se

$$\rho_{TV} = \sum_{i=1}^{N} \frac{t_i v_i}{N \sigma_T \sigma_V} \tag{3.14}$$

$$= \sum_{i=1}^{N} \frac{(v_i + e_i) v_i}{N \sigma_T \sigma_V} \tag{3.15}$$

$$= \frac{\sum_{i=1}^{N} v_i^2 + \sum_{i=1}^{N} v_i e_i}{N \sigma_T \sigma_V}. \tag{3.16}$$

Como a covariância do erro e o escore verdadeiro é nula, sobra da Equação 3.16,

$$\rho_{TV} = \frac{\sigma_V^2}{\sigma_T \sigma_V} = \frac{\sigma_V}{\sigma_T}. \tag{3.17}$$

Conclui-se então que a correlação entre o escore bruto e o escore verdadeiro é igual ao quociente entre o desvio padrão do escore verdadeiro e o desvio padrão do escore empírico. Os mesmos argumentos usados para mostrar a correlação existentes entre os escores pode ser usado para mostrar a correlação entre o escore empírico e o erro, que é expresso pela Equação 3.18.

$$\rho_{TE} = \frac{\sigma_E}{\sigma_T}. \tag{3.18}$$

Outras consequências do modelo da TCT são tautologias, e não serão discutidas nesse estudo. Para uma maior cobertura do assunto, sugere-se Pasquali (2017), Lord e Novick (1968).

Além dos pressupostos e derivações básicas do modelo, a TCT é composta por uma série de índices que ajudam na análise empírica dos itens, respondendo a questões como

dificuldade e descriminação do item, fidedignidade do teste, tendenciosidade de resposta e precisão. Nesse sentido, a TCT é uma abordagem importante para a consecução de testes educacionais, particularmente quando aplicados em uma testagem de larga escala.

As subseções a seguir apresentam, de forma resumida, os parâmetros da TCT que serão usados na análise empírica dos itens da prova da OBMEP, 13ª edição, aplicada a um grupo de alunos das escolas públicas do município de Santarém, Pará.

3.2.2 Dificuldade do Item

O primeiro parâmetro analisado no âmbito da TCT baseia-se numa medida de dificuldade do item. Classicamente, em um teste de múltipla escolha, o índice de dificuldade ID de um item é definido como a proporção de respostas dada à alternativa correta de um item, ou seja, a porcentagem de respondentes acertam o itens (CONDé, 2008; RODRIGUES, 2006).

Pode-se expressar este índice por meio da Equação 3.19

$$ID = \frac{A}{N} \tag{3.19}$$

em que A é o número de indivíduos que responderam corretamente ao item e N o número total de respondentes. Nota-se que a dificuldade do item não leva em conta as respostas corretas obtida por chutes. Na tentativa de corrigir essa limitação, Pasquali (2017) sugere a Equação 3.20

$$ID = \frac{A - \left(\frac{E}{K-1}\right)}{N} \tag{3.20}$$

para estimar a dificuldade do item. Nela, E representa o número de sujeitos que erraram o item e K o número de distratores do item. Porém esse modelo assume que cada $K-1$ resposta incorreta, há uma resposta correta conseguida por acaso, o que não é regra geral.

Condé (2001) apresenta classificações dos itens em termo de seu índice de dificuldade, no qual considera-se:

- item fácil: $ID > 0,70$;
- item de média dificuldade: $0,30 < ID \leq 0,70$;
- item difícil $ID \leq 0,30$.

É importante frisar que a classificação proposta se apoia no fato de que quanto mais indivíduos acertam um determinado item, mais fácil ele é considerado. Segundo Rodrigues (2006), é importante que provas realizadas em larga escala tenham itens de dificuldade variada, que alcancem o *continuum* da escala. Em testes classificatórios, como supostamente é o caso da OBMEP, o nível de dificuldade dos itens tendem a ser mais elevados. Espera-se então que uns e outros itens não avaliam adequadamente a proficiência

dos participantes de baixo rendimento, uma vez que melhor discriminara os participantes de maior aptidão.

3.2.3 Discriminação do Item

A discriminação do item é outro parâmetro importante, que na TCT é definido como a capacidade do item diferenciar indivíduos com bom rendimento no teste, daqueles que possuem baixo rendimento. Dois modelos estatísticos podem ser usados para estimar a discriminação do item, a saber, o dos grupos-critério e a correlação do item com o escore total dos itens (PASQUALI, 2017; RODRIGUES, 2006).

Na discriminação via grupo-critério, se utiliza o escore bruto do próprio teste na obtenção do critério discriminativo – tipicamente grupos extremos de sujeitos: grupo superior e grupo inferior. Na proposta de Kelley (1939) apud Pasquali (2017), considera-se na aplicação de um teste a muitos indivíduos os 27% superiores e os 27% inferiores para compor os grupos de referência (PASQUALI, 2017).

A forma mais simples de obter a discriminação D por meio dessa metodologia, é fazendo a diferença entre a proporção de acertos do grupo superior (27% dos alunos com maior desempenho) a do inferior e inferior (27% com o menor despenho). Assim, o procedimento consiste em calcular a proporção de acertos em cada item desses dois grupos e subtraí-los.

O teste t pode ser usado para comparas as médias do grupo superior e inferior, requisitando apenas parâmetros da estatística descritiva de fácil obtenção. A equação para o cálculo do teste t é

$$t = \frac{\bar{X}_s - \bar{X}_i}{\sqrt{\frac{S_s^2}{n_s} + \frac{s_i^2}{n_i}}}, \tag{3.21}$$

cujo grau de liberdade $GL = n_s + n_i - 2$. $\bar{X}_s$ e $\bar{X}_i$ são as médias do grupo superior e inferior, S_s^2 e S_i^2 são as variâncias dos grupos, n_s e n_i representam o número de sujeito em cada grupo. É importante mencionar que a consecução desse método pressupõe que os escores sigam uma distribuição normal de probabilidade e que nenhuma dessas variâncias sejam nulas (SOARES, 2018).

Como mencionado, a correlação item total também estabelece o índice de discriminação. Entre outras correlações, a correlação ponto bisserial r_{pb} é bastante utilizada, pois consiste em uma correlação de Pearson para variável dicotômica. A equação decorrente dessa correlação é da forma

$$r_{pb} = \frac{\bar{X}_A - \bar{X}_T}{S_T}\sqrt{\frac{p}{q}} \tag{3.22}$$

em que: $\bar{X}_A$ é a média no teste dos sujeitos que acertaram o item; $\bar{X}_T$ a média total do teste; S_T é o desvio padrão de escore; p a proporção de sujeitos que aceitaram o item e $q = 1 - p$.

Outra media discriminativa do item resulta da correlação bisserial r_b. Para sua definição, postula-se a existência de uma variável contínua não observável associada à habilidade do indivíduo testado. Para cada resposta U_{ij} (dicotomizada), supõe-se a existência de um valor da variável para cada item, de forma que: i) o aluno acerta o item se a habilidade estimada pela variável é maior ou igual ao ponto; ii) o aluno erra o item se a habilidade não alcança o ponto (KLEIN, 2005).

Ainda segundo Klein (2005), a correlação bisserial é conceitualmente definida como a correlação de Pearson da variável latente com a medida de desempenho, ou seja, consiste numa medida de associação do acerto do item com o desempenho do sujeito avaliado. Por isso pressupõe-se que a variável contínua tenha uma distribuição normal de probabilidade.

O coeficiente r_b está relacionado com a correlação ponto bisserial por meio da Equação 3.23

$$r_b = \frac{r_{pb}\sqrt{p(1-p)}}{h(p)} \tag{3.23}$$

em que: $h(p)$ é o valor da função de densidade da distribuição normal com média nula e variância unitária no ponto cuja área da curva à esquerda desse ponto é igual a proporção de acerto p no item (KLEIN, 2005).

As vantagens da correlação bisserial com relação ao coeficiente ponto bisserial é ser menos dependente do grupo testado (KLEIN, 2005). Por outro lado, a expressão da correlação bisseral pode resultar em valores maiores que 1 caso a condição de normalidade da variável contínua não seja satisfeita. Neste caso, Pasquali (2017) recomenda o uso da correlação ponto bisserial.

De acordo com Quaresma (2014, p. 71), "altas correlações entre o item e escore configuram alta contribuição do teste para aumentar a variância dos escores, ajudando na discriminação dos sujeitos".

> O item é mais discriminativo quanto maior for o seu valor. O índice de discriminação pode assumir qualquer valor entre -1 e +1, correspondendo à diferença entre o índice de dificuldade dos indivíduos que obtiveram uma pontuação elevada no escore total do teste e o índice de dificuldade dos indivíduos que obtiveram uma pontuação baixa no escore total do teste (SARTES; SOUZA-FORMIGONI, 2013, p.243).

Segundo Rabelo (2013), é desejável que a discriminação do item em testes educacionais seja superior a 40%, conforme mostra a Tabela 2.

3.2.4 Precisão do teste

Quanto ao parâmetro de precisão dos testes, Fletcher (2010) diz que, embora não seja possível obter o escore verdadeiro, conhecer o grau de proximidade entre os escores observados e os verdadeiros não só é factível como substancial. É talvez por esse motivo que o grau de precisão tem papel crucial na análise da fidedignidade do teste.

Tabela 2 – Classificação dos itens de acordo com o índice discriminativo

Valores	Classificação
Discriminação $< 0{,}20$	Item deficiente, deve ser rejeitado
$0{,}20 \leq$ Discriminação $< 0{,}30$	Item marginal, sujeito a reelaboração
$0{,}30 \leq$ Discriminação $< 0{,}40$	Item bom, mas sujeito a aprimoramento
Discriminação $\geq 0{,}40$	Item bom

Fonte: Rabelo (2013)

Ainda segundo Fletcher (2010), o índice de fidedignidade natural seria o coeficiente de correlação entre os escores observados e verdadeiros (ρ_{TV}), porém, como não é possível observar diretamente os escores verdadeiros, esta formulação tem efeito apenas sobre a teoria em si. Felizmente é possível estabelecer uma identidade matemática entre ρ_{TV} e a correlação dos escores observados em dois testes paralelos[1], t_1 e t_2.

O resultado da correlação dos escores de t_1 e t_2, paralelos, resultam no índice de precisão ρ_{vv} (*index of reliabilit*) expresso em termos dos escores T e V

$$\rho_{vv} = \left(\frac{\sigma_V}{\sigma_T}\right)^2 = (\rho_{TV})^2 \tag{3.24}$$

Essa expressão também assume a forma

$$\rho_{vv} = \frac{\sigma_T^2 - \sigma_E^2}{\sigma_T^2} = 1 - \frac{\sigma_E^2}{\sigma_T^2} \tag{3.25}$$

dado que $\sigma_T^2 = \sigma_V^2 + \sigma_E^2$.

> O coeficiente de fidedignidade serve para estimar o efeito dos erros de medida sobre as correlações e corrigir sua atenuação, o problema que inspirou a obra de Spearman no início do século. A teoria clássica demonstra que a variância dos escores verdadeiros aumenta junto com o quadrado do número de itens no teste, ao mesmo tempo que a variância dos erros aumenta apenas linearmente com o número de itens. Essa relação explica o aumento progressivo da fidedignidade do teste junto com o aumento do número de itens (FLETCHER, 2010, p.11).

Outro importante indicador de consistência interna conhecido é o coeficiente *Alfa de Cronbach*, proposto por Lee J. Cronbach em 1951, que é expresso pela Equação

$$\alpha = \frac{n}{n-1}\left(1 - \frac{\sum_{i=1}^{N}\sigma_i^2}{\sigma_T^2}\right) \tag{3.26}$$

em que: n é o número de itens; $\sum_{i=1}^{N}\sigma_i^2$ a soma das variâncias dos n itens e σ_T^2 a variância total dos escores do teste (PASQUALI, 2017).

A Equação 3.26 mostra que, quanto menor são as variâncias dos itens individuais σ_i^2 e maior a variância dos escores do teste σ_T^2 (variância que todos os item possui em comum),

[1] Segundo Anjos (2013), denomina-se testes paralelos aqueles cuja distribuição de de erro é a mesma em cada um dos testes ou quando os escores verdadeiros são os mesmos em ambos os testes para o mesmo indivíduo.

o índice α tende a valores mais elevados. Nesse sentido, o *Alfa de Cronbach* proporciona de fato uma medida de consistência interna, cujos valores pertencem ao intervalo [0,1]. Para valores de α próximos de 1, diz-se que a consistência interna do item é elevada e, de modo análogo, para α próximos de 0, diz-se que o item possui baixa consistência interna (PASQUALI, 2017).

3.2.5 Análise Gráfica do Item (AGI)

Batenburg e Laros (2002) descrevem uma representação visual para a análise da discriminação de itens de múltipla escolha produzidos no âmbito da avaliação de habilidades, sobretudo no meio educacional. O método conhecido como Análise Gráfica de Itens (AGI) exibe a relação entre os escores totais atingidos pelos alunos em um teste e as proporções de resposta dada a cada uma das alternativas. O método da AGI teve lugar nas análises pedagógicas e psicométrica dos itens da prova do Sistema de Avaliação do Ensino Básico (SAEB), a fim de se discutir a qualidade do instrumento através de uma representação visual de fácil entendimento, a partir de ideias básicas da TCT.

Segundo Pasquali (2017) a AGI se baseia em duas ideias centrais. A primeira delas é de que indivíduos que fornecem resposta correta ao item costumam saber mais do que os indivíduos que erram o item. O segundo pressupõe, por exemplo, que se o sujeito S_1 obteve escore maior do que o sujeito S_2 em um teste que visa avaliar o domínio de conteúdos em um determinado tema, espera-se, então, que o sujeito S_1 saiba mais tais conteúdos do que o sujeito S_2 .

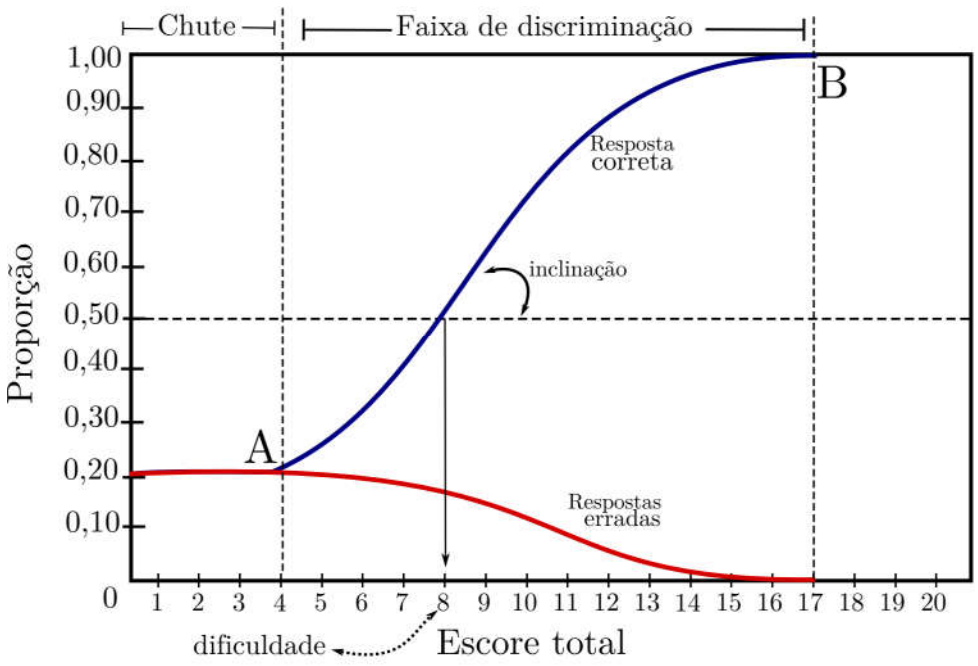
As consequências desses pressupostos é de que sujeitos que possuem elevados escores sabem mais do que os sujeitos com escores menores, o que torna pertinente a seguinte afirmação: "sujeitos com escores maiores tenderão a acertar mais um certo item que sujeitos com escores menores"(PASQUALI, 2017) .

Assim, espera-se que a proporção de estudantes que escolhem a resposta correta de um item aumente em função do crescimento da pontuação total, ao passo que a proporção de estudantes que escolhem os distratores errados diminua com o aumento do escore. “Quanto mais rápido a proporção de alunos com a resposta correta aumenta com aumento da pontuação total, melhor o poder de discriminação e maior a qualidade de um item” (BATENBURG; LAROS, 2002, p. 320). Por outro lado, consideram-se itens de baixa qualidade aqueles que, com o aumento dos escores dos estudantes, mantém proporções elevadas de escolhas aos distratores. Não se espera, portanto, que grupos de alunos com elevada pontuação optem pelas respostas erradas.

Na Figura 4 apresenta-se os elementos dessa análise a partir de um item teoricamente de boa qualidade. O caso apresentado é de um teste de 20 itens de múltipla escolha com 5 distratores, no qual o escore máximo foi de 17 acertos. No eixo das ordenadas apresenta-se a proporção de respostas dos indivíduos e no eixo das abscissas os escores

atingidos pelos participantes.

Figura 4 – AGI de um item de boa qualidade



Nota-se na Figura 4 que a proporção de indivíduos que respondem o item corretamente, conforme indicado pela curva azul, aumentam sistematicamente com o crescimento dos escores. Os distratores incorretos, representado pela curva vermelha, diminui ao passo que os escores aumentam. Nessa representação, o intervalo [0, 4] do eixo das abscissas consiste em um intervalo de acerto ao acaso, pois um indivíduo que escolhesse responder o teste com cinco alternativas e de forma aleatória contaria com 20% de probabilidade de escolher a resposta correta em cada um dos itens. Respondendo de forma aleatória, o indivíduo poderia obter escore igual a 4 no teste. Neste caso, a partir do escore 4 a proporção de resposta correta deve aumentar em função do escore e as proporções de respostas aos distratores incorretos devem diminuir.

Outro elemento que aparece na AGI da Figura 4 é o de inclinação da curva das respostas corretas quando a ocorrência de resposta é equivalente a 50%, isto é, ângulo em relação ao eixo das abscissas quando o eixo das ordenadas equivale a 0,5. Essa inclinação é uma estimativa da discriminação do item, pois quanto maior o ângulo mais rapidamente a ocorrência de respostas corretas cresce com o aumento do escore. O valor associado à ocorrência de respostas igual a 50% no eixo das abscissas indica uma medida de dificuldade do item. No exemplo dado, a dificuldade seria 8 (PASQUALI, 2017).

Em síntese, a análise gráfica apresenta estimativas visuais e numéricas que ajudam a responder sobre a capacidade discriminativa do item, dificuldade e até mesmo propõe intervalos onde os escores podem ser obtidos por meio de escolhas aleatórias.

3.2.6 Algumas limitações da TCT

A TCT apresenta em seu escopo parâmetros e resultados importante na análise da qualidade dos testes. Contudo, seus resultados são baseados no estudo do escore, e depende do conjunto de itens e da população testada. Essa dependência do teste em relação ao grupo impõe problemas práticos como o de determinação de uma amostra significativa que contenha todas as característica da população (PASQUALI, 2017). Além disso, os resultados apresentam características restritas de um grupo que, no caso educacional, inviabiliza comparações entre grupos e testes distintos (KLEIN, 2005).

Segundo Klein (2005, p. 120) as limitações da TCT são:

- Os escores do teste que avaliam os alunos dependem do conjunto de itens apresentados.
- A TCT só pode ser utilizado em caso de todos os alunos realizarem o mesmo teste (ou formas "paralelas" de teste).
- A teoria não prevê um modelos para o desempenho de um aluno em um item, apenas a qualidade do instrumento.
- Nas aplicações da Teoria Clássica dos Testes deve-se ter o cuidado de que os erros de medida nem sempre possuem a mesma variabilidade para todos os alunos.

3.3 TEORIA DE RESPOSTA AO ITEM (TRI)

A TRI surgiu para suprir algumas limitações da TCT, em particular a dependência íntima dos resultados produzidos com o grupo testado e o conjunto de itens, uma vez que leva em conta apenas as estatísticas dos escores, obtidos de um grupo particular, que inviabilizam comparações dos resultados de diferentes testes e grupos de indivíduos testados (KLEIN, 2005).

Em avaliações de larga escala[2], busca-se avaliar a proficiência dos alunos com respeito ao currículo escolar, os quais, na sua maioria, envolvem um número muito grande de conteúdos. E dessa forma, sem o uso de uma abordagem que permita avaliar provas com diferentes itens e grupo de pessoas, os testes tenderiam a ser extensos para garantir que os grupos de alunos testados recebam testes equivalentes, que cubram todos os tópicos da matriz curricular (KLEIN, 2013).

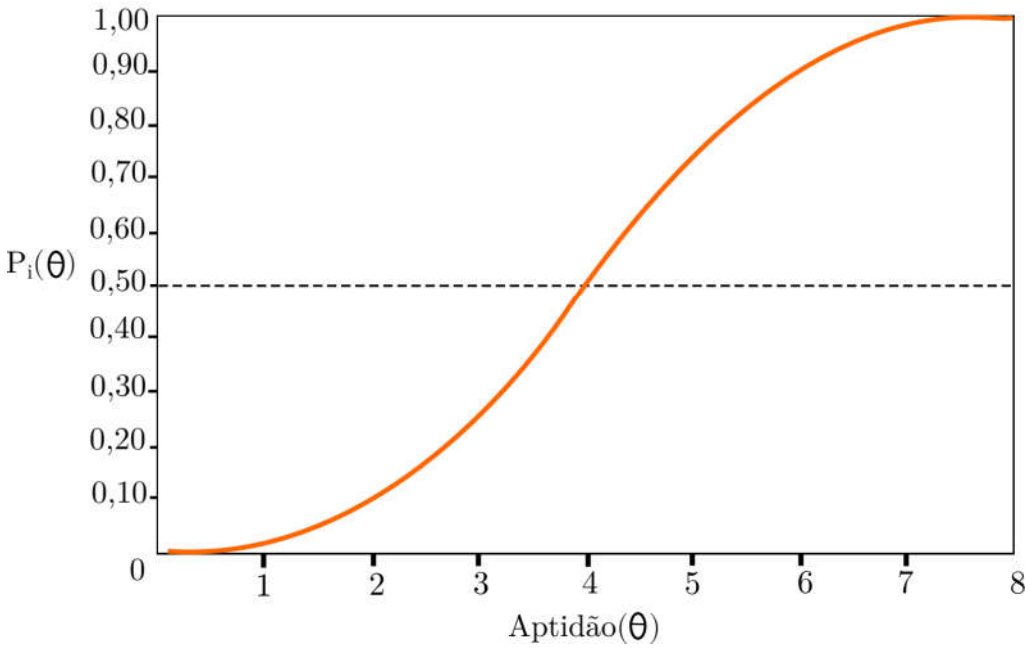
[2] Avaliações em larga escala fogem à dinâmica professor-aluno, sendo conduzida por entidades externas e que avaliam de forma independente constructos (proficiência, satisfação etc.) de uma grande quantidade de sujeitos, permitindo construir estatísticas que avaliam redes inteiras de ensino quanto sua qualidade. De acordo com (LEIRIãO, 2017), a "aplicação em larga escala tira do foco da avaliação o aluno e coloca a escola/ rede/ município ou estado".

Mas com o avanço da TRI, em parte viabilizado pelo avanço da computação e pela consolidação de modelos matemáticos consistentes, a análise dos testes passaram a levar em conta as características individuais do item, tendo como foco o estudo da variável latente (variável não observável) θ. Assim, o desempenho do sujeito em um item do teste se explica em função dos traços latentes que, no caso de avaliações educacionais, são aptidões e habilidades (PASQUALI, 2017).

De acordo com Klein (2005), tanto parâmetros dos itens quanto das proficiências dos indivíduos são invariantes. Essa propriedade da TRI conduz ao fato de que parâmetros dos itens respondidos por diferentes grupos são invariantes. O mesmo vale para o caso da proficiência. Ou seja, grupos diferentes de itens produzem resultados independentes, mudando somente origem e escala.

Outra característica elementar da TRI, é que o desempenho do indivíduo em um item particular do teste pode ser expresso por uma função monótona crescente, de variável θ e imagem no intervalo [0,1], cujo gráfico genérico está ilustrado na Figura 5. Essa função determina que sujeitos com maior aptidão terão maior probabilidade de responder o item corretamente(PASQUALI, 2017).

Figura 5 – Função de probabilidade de resposta ao item $P_i(\theta)$



Fonte: Pasquali (2017)

A abordagem da TRI unidimensional pressupõe duas hipóteses fundamentais, a saber, a de unidimensionalidade da proficiência e de independência condicional das respostas de um sujeito as itens que compõe um teste, dado sua proficiência. Essa última hipótese também é denominada de "independência local"(KLEIN, 2005). Cabe ressaltar que já exitem resultados robustos sobre a TRI multidimensional, recomendada para instrumentos

que possuem mais um traço latente predominante, mas isso é uma outra história. Este estudo se concentra apenas nos modelos unidimensionais.

A hipótese da unidimensionalidade postula que as respostas aos itens, de uma certa prova, dependem de apenas uma aptidão. Teoricamente, outras variáveis latentes podem influenciar as respostas fornecidas pelo indivíduo, mas para satisfazer o postulado, considera-se a aptidão dominante que influencia o conjunto de respostas (PASQUALI, 2017). Caso o instrumento não possa ser reduzido a uma única dimensão (aptidão), os resultados da TRI não serão satisfatórios. Nesse sentido, uma verificação do primeiro postulado pode ser feito via Análise Fatorial (AF).

Ainda sobre a hipótese da independência local, Pasquali (2017) discute que, mantidas invariantes os traços latentes que afetam o teste, ou seja, a de unidimensionalidade, as respostas fornecidas a cada item são estatisticamente independentes. Assim, a probabilidade de resposta a um conjunto de itens (i), dada uma aptidão dominante, é igual aos produtos das probabilidades das respostas do sujeito j dada a cada item. Matematicamente, pode-se denotar esse resultado na forma

$$Prob(U_{ij}|\theta_j) = P(U_{1j}|\theta_j) \cdot P(U_{2j}|\theta_j) \cdot \cdots \cdot P(U_{In}|\theta_j) \tag{3.27}$$

$$= \prod_{i=1}^{I} P(U_{ij}|\theta_j) \tag{3.28}$$

em que:

- θ_j é variável contínua que representa a aptidão que afeta o conjunto de itens do j-ésimo indivíduo;
- U_{ij} é a variável dicotômica que representa a resposta do sujeito j ao item i;
- $P(U_{ij}|\theta_j)$ a probabilidade de resposta ao item ou Função de Resposta ao Item (FFI). Em particular, $P(U_i = 1|\theta_j)$ significa a probabilidade de resposta correta ao item, dado a aptidão θ. No caso em que $P(U_i = 0|\theta_j)$, o interesse passa a ser na probabilidade de errar o item, dado θ_j.

Os pressupostos da unidimensionalidade e independência local são equivalentes, uma vez que a consequência de supor a unidimensionalidade é de que as respostas dadas à cada item i pelo individuo j, sejam independentes. O contrário também é válido, pois a independência local exige que as respostas dependam apenas da habilidade dominante, e não dos demais itens ou de outros traços latentes (ANDRADE; TAVARES; VALLE, 2000; KLEIN, 2005).

Os modelos matemáticos da TRI são definidos de acordo com o tipo de função e com o número de parâmetros. Um modelo bastante usado para avaliar sujeitos submetidos à testes de múltipla escolha, geralmente com mais de duas alternativas, é intitulado de modelo logístico de 3 parâmetros (ML3). Nessa revisão, será dado enfoque apenas no ML3,

pois ele é um dos mais robustos para análise de respostas dicotomizadas. Os modelos de 2 parâmetros e 1 parâmetro logístico são tratados como caso particular do ML3.

3.3.1 Modelo logístico de 3 parâmetros - ML3

Como dito anteriormente, o ML3 unidimensional é um modelo bastante usual em avaliação educacional. Como exemplo, o Sistema de Avaliação da Educação Básica (Saeb) e o Exame Nacional do Ensino Médio (Enem) utilizam esse modelo para estimação da proficiência dos alunos.

O ML3 foi desenvolvido por Lord (1980) e é expresso matematicamente por:

$$P(U_{ij} = 1|\theta_j) = c_i + (1 - c_i)\frac{1}{1 + e^{-Da_i(\theta_j - b_i)}} \tag{3.29}$$

- O parâmetro a_i é uma medida da discriminação do item, cujo valor é proporcional à inclinação da Curva Característica do Item (CCI) no ponto de inflexão onde $P(U_{ij}|\theta_j) = 0,5$;
- b_i é o parâmetro localizado na escala de habilidade para qual $P(U_{ij}|\theta_j) = 0,5$, e indica uma medida de dificuldade do item;
- c_i consiste em uma medida de probabilidade de acerto casual;
- D é uma constante real, conhecido como fator de escala, sendo que $D = 1,7$ faz com que a função logística forneça resultados parecidos aos da função ogiva normal (ANDRADE; TAVARES; VALLE, 2000).

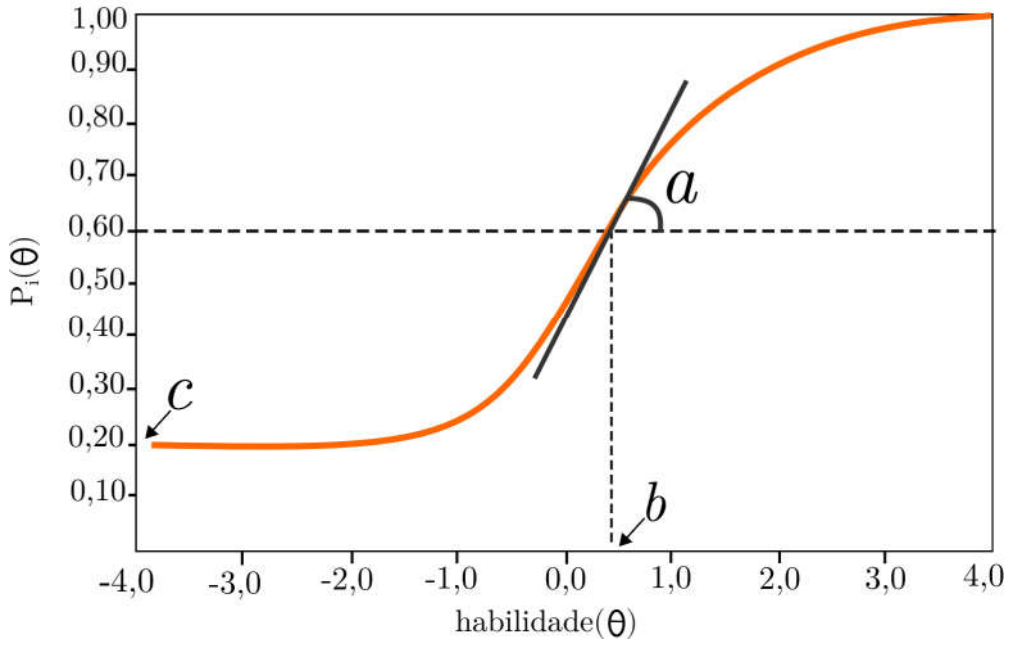
É importante notar que o modelo concorda com a ideia básica da TRI, a de que indivíduos com maior habilidade tem maior probabilidade de responder corretamente o item, supondo ainda que essa relação não é linear. E de fato, os modelos lineares propostos apresentaram dificuldades práticas e baixa capacidade de predição do desempenho dos alunos nos testes (PASQUALI, 2017).

Os parâmetros a_i, b_i e c_i modelam a Curva Característica do Item (CCI), tal como origem, deslocamento na escala de habilidade e inclinação. A Figura 6 mostra as contribuições de cada parâmetro para a formação da CCI.

Segundo Andrade, Tavares e Valle (2000, p. 10-11), "o parâmetro b representa a habilidade necessária para uma probabilidade de acerto igual a $(1 + c)/2$". Nesse sentido, o aumento de b implica que a probabilidade do sujeito j responder corretamente ao item i diminui, o que o caracteriza como um item difícil.

No que diz respeito à avaliação educacional, Rabelo (2013) sugere que a prova tenha valores de b variado e propõe uma classificação de níveis de dificuldade para os

Figura 6 – Exemplo de uma Curva Característica do Item – CCI



itens a partir de intervalos em que se encontra o parâmetro b, conforme apresentado na Tabela 3.

Tabela 3 – Distribuição esperada e classicação do item em relação ao seu nível de dicul-dade

Distribuição esperada	Classificação	Dificuldade do item
10%	Muito fáceis	Até -1,28
20%	Fáceis	De -1,27 a -0,52
40%	Medianos	De -0,51 a 0,51
20%	Difíceis	De 0,52 a 1,27
10%	Muito difíceis	1,28 ou mais

Fonte: (RABELO, 2013)

Ao tomar $c = 0$, isto é, que a probabilidade de acerto ao acaso é descartada, o parâmetro b passa a representar o ponto na escala de habilidade onde a probabilidade de responder ao item corretamente é 0,5. De fato, $(1 + 0)/2 = 0,5$, sendo esse um ponto importante na análise de questionários, cujas respostas são sim ou não (KLEIN, 2005). Mas em testes educacionais, onde o indivíduo sempre tem uma chance de acertar o item ao acaso, o parâmetro c é importante e por isso geralmente é considerado.

O parâmetro a indica a discriminação do item. Ao supor-se a constante para todos os itens do teste, chegamos ao modelo logístico de 1 parâmetro, conhecido como modelo de *Rasch*, o qual considera apenas o parâmetro de dificuldade do item. Segundo Klein (2005), o modelo de *Rasch* foi obtido a partir de algumas hipóteses independente do desenvolvimento da TRI e tem a desvantagem de apresentar poucas informações do item.

Uma proposta de classificação baseada no poder discriminativo do item é proposta por Rabelo (2013). A Tabela 4 exibe faixa de discriminação que vai desde a ausência de discriminação, quando $c = 0$, até itens com índices de discriminação muito alto ($a \geq 1,70$).

Tabela 4 – Classificação do item de acordo com seu potencial discriminativo

Discriminação do item	Valor do parâmetro a
Nenhuma	a = 0
Muito baixa	$0 < a \leq 0,35$
Baixa	$0,35 < a \leq 0,65$
Moderada	$0,65 < a \leq 1,35$
Alta	$1,35 < a \leq 1,70$
Muito alta	$a \geq 1,70$

Fonte: (RABELO, 2013)

Assim, a análise empírica dos itens depende essencialmente dos valores dos parâmetros do modelo utilizado. Além das medidas de discriminação, dificuldade e de acerto ao acaso, a TRI produz resultados sobre a qualidade de cada item e do teste. Uma desses resultados é obtido pela curva de informações do item, a qual recebe o nome de função de informação do item (FII) – dependente apenas da variável latente θ. A FII é expressa matematicamente pela Equação 3.30.

$$I_i(\theta) = \frac{\left[\frac{d}{d\theta}P_i(\theta)\right]^2}{P_i(\theta)Q_i(\theta)} \tag{3.30}$$

- $I_i(\theta)$ é a "informação" fornecida pelo item no nível de proficiência θ;
- $P_i(\theta) = P(U_{ij} = 1|\theta)$, ou seja, a probabilidade de o sujeito j de habilidade θ responder o item i corretamente ;
- $Q_i(\theta) = 1 - P_i(\theta)$ (ANDRADE; TAVARES; VALLE, 2000).

Derivando a função de resposta ao item do modelo logístico de 3 parâmetros em relação a θ e fazendo algumas manipulações, obtêm-se a FFI do ML3, dada por:

$$I_i(\theta) = D^2 a_i^2 \frac{Q_i(\theta)}{P_i(\theta)} \left[\frac{P_i(\theta) - c_i}{1 - c_i}\right]^2 \tag{3.31}$$

Outra relação importante na abordagem da TRI é a função de informação do teste, por sua vez definida como a soma das informações fornecidas pelos itens individualmente. Assim,

$$I(\theta) = \sum_{i=1}^{I} I_i(\theta) \tag{3.32}$$

O erro-padrão da medida – no contexto da TRI conhecido como erro-padrão de estimação, consiste em outra maneira de representar a função $I(\theta)$, sendo expresso por:

$$EP(\theta) = \frac{1}{\sqrt{I(\theta)}} \tag{3.33}$$

em que $EP(\theta)$ é erro-padrão da habilidade θ.

3.3.2 Estimação de Parâmetros

Conforme discutido na subseção anterior, a probabilidade do sujeito responder a um item corretamente depende apenas dos parâmetros dos itens e de sua habilidade (ANDRADE; TAVARES; VALLE, 2000). Mas em geral nenhuma dessas informações são conhecidas a *priori*. É aí que os métodos de estimação de parâmetros surgem como etapa fundamental na produção dos resultados da TRI, pois é a partir dela que se calibra os itens e se estima as habilidades dos alunos. A informação utilizada na estimação são as respostas fornecidas por um grupo de sujeitos a um teste τ. Assim, a calibração dos itens tende a ser influenciada pelo número de itens e de respondentes.

A etapa de calibração é complexa, pois envolve métodos iterativos que nem sempre produzem bons ajustes. Em termos práticos, Andrade, Tavares e Valle (2000) divide o processo de estimação nas três situações seguintes:

- quando os parâmetros dos itens são conhecidos, isto é, calibrados, e portanto resta somente estimar as habilidades dos sujeitos submetido ao teste;
- quando se conhece a habilidade dos sujeitos testados e resta estimar os parâmetros dos itens;
- e por fim, e de maneira mais recorrente, quando não se conhece nem os parâmetros dos itens e tampouco as habilidades dos sujeitos.

No caso em que se deseja estimar os parâmetros e a habilidades dos sujeitos, simultaneamente, Andrade, Tavares e Valle (2000) evidencia duas abordagens. A primeira delas é a estimação conjunta (parâmetros dos itens e habilidades) e a segunda é um processo de estimação por etapas, começando pela estimação dos parâmetros e, em seguida, das habilidades.

O detalhamento matemático da estimação para essas suas abordagens e diferentes métodos pode ser conferido em Andrade, Tavares e Valle (2000). Os Aspectos computacionais utilizado no processo de estimação é tratado nos métodos deste trabalho.

3.3.3 Dimensionalidade da proficiência

A unidimensionalidade é um axioma assumido na abordagem unidimensional da TRI. Conforme já mencionado, se o pressuposto da unidimensionalidade não é satisfeito, então o modelo da TRI produz ajustes débeis. Nesse sentido, é de suma importância que o instrumento de fato avalie uma única habilidade dominante. A verificação da dimensionalidade do teste é feita, em geral, com auxílio da Análise Fatorial (AF) e de Análise de Componentes principais.

A AF consiste numa série de métodos estatísticos que trabalham com análises multivariadas e matrizes de intercorrelações entre variáveis ou itens. Apresenta-se como uma maneira de verificar se o construto mensurado pelos itens de um teste podem ser reduzidos a uma única dimensão dominante, com que todos os itens possuem alta correlação, sem perda de informações. Boas medidas de correlação dos itens com o fator dominante, comumente chamado de carga fatorial, indicam que os itens do teste medem uma única habilidade. Segundo Pasquali (2017), um conjunto de itens que têm alta carga no fator constituem um instrumento unidimensional. Por outro lado, itens com cargas menores que 0,3 não medem a mesma habilidade.

Hair et al. (2009) apresenta uma tabela de valores para identificar itens com carga no fator para um nível de significância igual a 0,05, levando em consideração o tamanho da amostra. A Tabela 5 mostra o valor de carga fatorial minimamente aceitável de acordo com o tamanho da população.

Tabela 5 – Diretrizes para identificação de cargas fatoriais significantes com base no tamanho da amostra

Carga fatorial	Tamanho necessário da amostra para significância de 5%
0,30	350
0,35	250
0,40	200
0,45	150
0,50	120
0,55	100
0,60	85
0,65	70
0,70	60
0,75	50

Fonte: (HAIR et al., 2009)

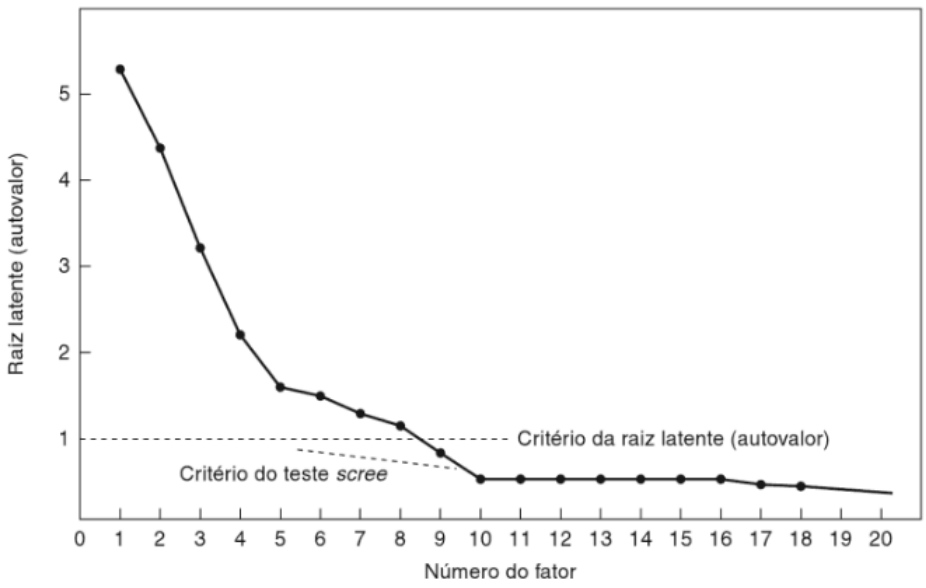
Na análise fatorial, dois tipos de abordagem podem ser utilizadas. A primeira delas é a análise fatorial exploratória, cujo emprego ocorre na tentativa de determinar o número de fatores dominantes que o teste possui. A análise fatorial confirmatória, por sua

vez, é comumente aplicadas em testes nos quais seus itens já foram calibrados e deseja-se produzir uma confirmação da qualidade do ajuste para uma dimensão considerada (QUARESMA, 2014).

A primeira etapa da análise fatorial exploratória consiste na identificação do número de fatores dominantes. Uma técnica visual usada para analisar a presença de um fator dominantes é o *scree plot* (PASQUALI, 2017). Segundo Hair et al. (2009, p. 114) "o teste de scree é usado para identificar o número ótimo de fatores que podem ser extraídos antes que a quantia de variância única comece a dominar a estrutura da variância comum". O método baseia-se no ponto de corte do gráfico dos autovalores de fatores em função do número de fatores ou componentes. É esperado que o ângulo de inclinação do primeiro fator em relação ao segundo fator comece a decrescer rapidamente, até que chegue a um ponto em que os ângulos de inclinação entre cada fator e o seu sucessor começam a diminuir lentamente, fazendo com que os fatores fiquem aproximadamente alinhados. É nesse ponto que o valor da abscissa indica o número de componentes ou fator a ser considerado no estudo (HAIR et al., 2009).

A Figura 7 é um *scree plot* baseado no autovalor dos fatores e o número de fatores para 18 itens. Nota-se que os autovalores vão decrescendo rapidamente e, a partir do 10 autovalor, passa a decrescer quase que linearmente, indicando a presença de aproximadamente 10 fatores. Outro critério importante é considerar o autovalor de fatores de ordem maior que 1 (HAIR et al., 2009). No gráfico, há 8 autovalores que respeitam a regra, indicando mais precisamente a presença de 8 fatores dominante no instrumento em questão. Esse seria um caso em que o pressuposto da unidimensionalidade não pode ser assumido.

Figura 7 – Gráfico de autovalor para o critério de teste *scree.*



Também se determina o número de fatores ou componentes em uma matriz de dados por meio do gráfico *scree plot* dos autovalores sucessivos das componentes ou fatores,

baseado no método de análise fatorial paralela. A análise paralela é que compara os *scree* de fatores dos dados observados com os de uma matriz de dados aleatórios do mesmo tamanho que o original (REVELLE, 2011; PASQUALI, 2017).

Segundo Pasquali (2017), se o pressuposto da unidimensionalidade for satisfeito, então os dois *scree plot* serão parecidos, mas com o primeiro autovalor da matriz obtida pelo método randômico menor que os da matriz de correlação dos dados reais.

Uma vez reduzida a dimensão dos dados, a análise fatorial confirmatória (AFC) é utilizada para testar se de fato os dados poderão ser reduzidos a uma determinada dimensão, conforme estabelecido nas hipóteses iniciais (VICINI; SOUZA, 2005). A partir daí calcula-se os valores das cargas fatoriais, comunalidade e outros índices que expressam a qualidade do ajuste.

Por outro lado, a aplicação da AF nem sempre é recomendada. O teste de Kaiser-Meyer-Olkin (KMO) compara as correlações simples com as correlações parciais e que pode ser usado para analisar a adequação da análise fatorial para o conjunto de dados. O modelo é expresso por

$$KMO = \frac{\sum_{i \neq j} \sum r_{ij}^2}{\sum_{i \neq j} \sum r_{ij}^2 + \sum_{i \neq j} \sum a_{ij}^2} \tag{3.34}$$

em que r_{ij} é a matriz de correlação entre variáveis e a_{ij} a matriz de do covariância parcial. A estatística KMO é baseada na medida da proporção de variação entre variáveis. Quanto menor a proporção, mais adequados são os dados para a Análise fatorial (GLEN, 2016).

As medidas de KMO variam entre 0 e 1 e avaliam a adequação da matriz de dados quanto ao grau de correlação parcial entre os valores. Valores próximos de zero indicam correlação fraca entre as variáveis do banco de dados, o que sugere uma inadequação do emprego da AF. Já valores de KMO próximos de 1, o emprego da AF é adequada (GLEN, 2016). A Tabela 6 mostra alguns intervalos de medidas de KMO usados na literatura como indicadores de adequação da AF para um conjunto de dados.

Tabela 6 – Emprego da estatística Kaiser-Meyer-Olkin (KMO)

KMO	**Análise Fatorial**
1 – 0,9	Muito boa
0,8 – 0,9	Boa
0,7 – 0,8	Média
0,6 – 0,7	Razoável
0,5 – 0,6	Má
<0,5	Inaceitável

Outro indicador de adequação é o teste de Esfericidade de *Bartlett*. O teste verifica se matriz de correlações pode ser ser a matriz identidade com determinante unitário. Assim, a hipótese nula (H0) é de que a matriz de correlações é uma matriz identidade. Havendo significância no teste (valor-p $< \alpha$), considera-se que a matriz de correlações não apresenta perdas significativa no processo de de extração de fatores (HAIR et al., 2009).

4 MATERIAIS E MÉTODOS

4.1 MATERIAL

Os dados usados nesta pesquisa são provenientes da prova da primeira fase da OBMEP, nível 1, aplicada a um grupo de 3875 alunos da escolas municipais da cidade Santarém, estado do Pará. A referida prova (ver no Apêndice A) foi realizada no ano de 2017 em âmbito nacional, incluindo, pela primeira vez, escolas particulares. Conforme discutido no Capítulo 2, a prova de nível 1 tem como público alvo os alunos do 6º e 7º ano do Ensino Fundamental.

A restrição do grupo de respondentes serem oriundos apenas da cidade de Santarém, deve-se ao fato dos dados não estarem disponíveis para análise na forma de microdados, como no caso do Exame Nacional do Ensino Médio (ENEM) e das avaliações conduzidas pelo Saeb, dificultando assim, o acesso às respostas de todos os alunos participantes da competição, seja no âmbito municipal, estadual ou nacional. Desta maneira, é factível coletar localmente os cartões respostas, indo de escola em escola solicitar dos coordenadores e gestores o material com as respostas dos alunos. Esse restrição populacional produz análises restritas e por isso não se tem pretensão de produzir análises conclusivas sobre qualidade dos itens da prova da OBMEP do referido nível e ano de aplicação.

Qualquer avaliação externa está estreitamente vinculada a uma matriz de referência proveniente de documentos oficiais que indiquem as habilidades e competências que os alunos devem atingir ao final de cada etapa da educação escolar (FONTANIVE, 2005; RODRIGUES, 2006; KLEIN, 2005). No caso da OBMEP não há esse estreitamento dos conteúdos cobrados com uma matriz de referência curricular. De acordo com um levantamento feito por Costa (2015), as provas da OBMEP não possuem uma matriz curricular específica, mas respeitam as habilidades e competências por ano/série apresentadas nos Parâmetros Curriculares Nacionais (PCN). Faz sentido, assim, classificar e analisar esses itens de acordo com os temas e habilidades previstos nos PCN.

A Tabela 7 contém a classificação dos itens da referida prova de acordo com os temas do PCN de matemática, para que se possa ter uma visão geral do conteúdos tratados na prova.

Tabela 7 – Classificação dos itens por temas do PCN

Tema	Itens
Números e operações	3, 4, 9, 10, 11, 12, 14, 15, 17, 18, 19, 20
Espaço e forma	8, 13, 16
Grandezas e Medidas	1, 2, 6, 7
Tratamento da Informação	5

Fonte: OBMEP, 2017

No que se refere as características da prova, a OBMEP elabora cadernos com 20 itens para serem aplicados na primeira fase, sendo que cada um desses itens possui 4 distratores e o gabarito (resposta correta), indicados pelas letras A, B, C, D e E. Inicialmente, as provas possuem objetivo meramente classificatório. As provas da primeira fase são corrigidas pelos professores da própria escola (lócus de aplicação da prova), que somam a pontuação total dos alunos. Os alunos com melhores desempenho passam para a segunda fase.

Os cartões respostas ficam disponíveis na escola por um tempo determinado. Após não serem mais úteis para fim de esclarecimentos aos participantes e organizadores, os cartões são descartados. Com finalidade de obter os cartões respostas, enviou-se aos responsáveis pela aplicação da prova em cada escola, um Termo de Consentimento de Livre e Esclarecido (Apêndice A), apresentado objetivos, necessidade do uso das respostas dos estudantes e cláusula que assegura o sigilo do nome da escola e dos alunos participantes.

O pesquisador coletou os cartões respostas da primeira fase de todas as escolas municipais inscritas na primeira fase da OBMEP da rede municipal de Santarém. As escolas participantes estão apresentadas na Tabela 8.

Tabela 8 – Escolas municipais inscritas na OBMEP, número de inscritos e de alunos que compareceram para fazer a prova nível 1 na cidade de Santarém

Código MEC	Nome	Alunos N1 inscritos	Alunos N1 participantes
15011810	E M E F ADERBAL TAPAJOS CAETANO CORREA	35	18
15012050	E M E F BRIGADEIRO EDUARDO GOMES	110	96
15518620	E M E F DEPUTADO UBALDO CORRÊA	653	492
15012310	E M E F DOM ANSELMO PIETRULLA	60	57
15162354	EMEF DOM LINO VOMBOMMEL	254	229
15012417	E M E F DRA MARIA AMALIA QUEIROS DE SOUSA	176	108
15572560	E M E F ELOINA COLARES E SILVA	98	82
15012522	E M E F FLUMINENSE	291	257
15162370	EMEF FREI MIGUEL KELLET	164	144
15012662	E M E F HAROLDO VELOSO	195	144
15012670	E M E F HELENA LISBOA DE MATOS	167	166
15012727	E M E F JOAO BATISTA MILEO	119	98
15162362	EMEF JOAO BIANOR MOTA FREITAS	140	118
15156753	E M E F MAESTRO WILDE DIAS DA FONSECA	269	253
15012255	E M E F MAGALHAES BARATA	130	101
15012883	E M E F MARIA DE LOURDES ALMEIDA	279	216
15162389	EMEF PADRE JOAO FELIPE BETTENDORF	300	205
15013472	EMEF PADRE MANUEL ALBUQUERQUE	285	220
15013430	E M E F PAULO RODRIGUES DOS SANTOS	163	109
15013553	E M E F PRINCESA IZABEL	318	191
15013677	E M E F PROF EILAH GENTIL	264	160
15013618	E M E F PROFA NAZARE DEMETRIO MUSSI	151	140
15013685	E M E F PROFA HILDA MOTA	120	77
15013758	E M E F ROTARY	160	107
15518566	E M E F S FRANCISCO DE ASSIS	167	87
TOTAIS		**5068**	**3875**

Fonte: Coordenação regional OBMEP, 2017

Após o processo de coleta do material nas escolas, os cartões foram escaneados e organizados em pastas e subpastas com o nome de cada escola e nível.

4.2 MÉTODOS

Uma pesquisa deve ser classificada de acordo com natureza (básica ou aplicada), finalidade (exploratórias, descritivas e explicativas) e procedimentos metodológicos (estudo de caso, bibliográfica, experimental, etc.) (GIL, 2002). Assim, esta é uma pesquisa aplicada que visa explorar as respostas de uma amostras de aluno em um teste para avaliar dificuldades e subsidiar discussões pedagógicas sobre o ensino de matemática. Para esta finalidade, faz-se um levantamento bibliográfico sobre psicometria no âmbito da educação e um estudo de caso baseado nos dados coletados da amostra. Engloba, para sua finalidade maior, as abordagens qualitativa e quantitativa.

O presente estudo perpassa por diferentes etapas, sendo que no primeiro momento consultou-se os os autores Quaresma (2014), Fontanive (2005), Rodrigues (2006), Costa (2015), Vilarinho (2015), Klein (2005), que tratam da avaliação educacional, testes educacionais e medida de proficiência, em consonância com a possibilidades de produzir alguma mudança positiva no contexto educacional, a partir da análise das respostas fornecidas pelos alunos nos testes.

A análise dessas respostas numa perspectiva quantitativa, requer revisão de pressupostos e modelos consolidados no campo da psicometria clássica e moderna, visto em Pasquali (2017), Lord e Novick (1968), Andrade, Tavares e Valle (2000), Klein (2005).

Posteriormente, verificou-se por meio de dados fornecidos pela coordenação local (Tabela 4.1) que a zona urbana de Santarém teve 5068 inscritos na primeira fase, correspondente ao nível 1. Desse total, 3875 realizaram a prova – número considerado o tamanho da população neste estudo. Devido à dificuldade de se tabular os 3875 gabaritos, optou-se por desenvolver o estudo amostral.

O tamanho da amostra foi determinado com um nível de confiança de 95% e seu cálculo foi realizado por meio da expressão

$$n = \frac{p(1-p)Z^2N}{\varepsilon^2(N-1) + Z^2p(1-p)} \tag{4.1}$$

a qual determina o tamanho de uma amostra com finalidade de estimar a proporção populacional com determinadas característica. Na equação 4.1, n é o tamanho da amostra; p a proporção esperada (0,5); Z o valor da distribuição normal para o nível de confiança de 95%; N o tamanho da população (3875) e ε o erro amostral. Com essas informações, o tamanho da amostra obtido foi de 349 respondentes.

Por conseguinte, a amostragem estratificada usada para selecionar esta amostra foi do tipo aleatória sistemática. A técnica consiste em escolher um dos gabaritos de forma aleatória entre a população, e, por conseguinte, selecionar os gabaritos subsequentes usando os intervalos definido por K, por usa vez determinado pela razão entre a população e o tamanho da amostra.

Também foi levado em conta a proporção de alunos por escola. Assim, o sorteio de um gabarito G, usando a função $= ALEATÓRIO()$ no Excel, foi repetido para cada escola, cuja lista de gabaritos leva em conta a proporção de alunos matriculados na escola e o tamanho da amostra considerada. A escolha dos $N-1$ indivíduos restantes ocorreu mediante uma sucessão aritmética, da forma

$$G, G+K, G+2K, G+3K, \cdots, G+(n-1)K.$$

Concomitante ao processo de amostragem, criou-se uma planilha no *software Excel* com finalidade de organizar as respostas e gabarito da prova, sendo que este último inserido na segunda linha da planilha, conforme mostra a Figura 8. Em seguida, os dados da planilha foram salvos no formato *txt*. A escolha desse formato ocorreu devido a necessidade de efetuar a leitura pelo *software R*.

Figura 8 – Modelo: organização das respostas referentes a amostra numa planilha do Excel

	Item1	Item2	Item3	Item4	Item5	Item6	Item7	...
Gabarito	A	B	C	D	E	C	D	...
A1	A	C	A	B	A	A	C	...
A2	A	A	E	E	A	A	B	...
A3	A	C	D	D	D	E	B	...
A4	A	B	B	E	E	E	A	...
A5	A	C	B	E	E	D	B	...
A6	B	B	C	A	C	C	C	...
A7	D	C	A	E	E	A	D	...
A8	A	A	D	A	D	D	E	...
A9	B	C	D	B	E	A	C	...
A10	A	B	A	D	A	C	D	...
A11	B	D	B	B	E	A	C	...
A12	A	B	C	E	B	B	A	...
A13	A	E	D	D	D	A	B	...
A14	A	C	C	A	D	C	B	...
A15	B	D	C	E	D	C	D	...
A16	A	A	D	E	B	B	D	...
A17	B	C	B	D	E	F	B	...
A18	A	B	B	A	C	C	C	...
A19	B	C	B	E	A	A	B	...
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋱

4.2.1 Softwares e Pacotes

Para ler os dados e realizar as análises segundo os pressupostos da TCT e TRI, utilizou-se o *software R* – linguagem e ambiente para computação estatística e gráficos do projeto GNU (TEAM, 2018). Optou-se pelo ambiente de desenvolvimento RStudio - software gratuito amplamente utilizado na comunidade científica, acadêmica e corporativa. O RStudio é um ambiente de desenvolvimento integrado (IDE) para *R*. Ele inclui um console, editor de realce de sintaxe que suporta execução direta de código, bem como ferramentas para plotagem, histórico, depuração e gerenciamento de espaço de trabalho (TEAM, 2018).

Quatro pacotes desenvolvidos em linguagem *R* foram fundamentais na análise dos dados, a saber: *A package for personality, psychometric, and psychological research Des-*

cription (Psych), *Classical Test Theory Functions* (CTT), *Latent Trait Models for Item Response Theory Analyses* (Ltm), *Full information maximum likelihood estimation of IRT models* (Mirt) e *Análisis y calificación de pruebas objetivas* (Itan). Cada um desses pacotes desenvolve funções relativos ao campo da Psicometria, tendo como foco executar análises de dados dicotômicos ou de múltiplas categorias.

- **Psych** – possui funções mais úteis para pesquisas psicométricas, que abrange a análise fatorial e diversos índices confirmatório e exploratório da pesquisa psicológica (REVELLE, 2011).
- **CTT** – o pacote é usado para executar uma variedade de tarefas e análises associadas à TCT, bem como pontuar respostas de múltipla escolha, realizar análises de confiabilidade, conduzir análises de itens e transformar pontuações em diferentes escalas (WILLSE; SHU, 2014).
- **Mirt** – esse é um dos mais completos pacotes usados em *R* na análise de dados de respostas dicotômicas e politômicas, baseado em modelos de traços latentes unidimensionais e multidimensionais sob o paradigma Teoria da Resposta ao Item (CHALMERS et al., 2012).
- **Ltm** – além de fornecer índices relativos à TCT, bem como coeficiente de consistência interna, medida de correlação ponto bisserial e índice de dificuldade clássica, o pacote possui funções que permite converter, com bastante praticidade, o *data.frame* composto por caracteres ("A","B","C", "D", "E", etc.) em um novo *data.frame* dicotômico, conveniente para a analise posterior dos dados. "O pacote ltm foi desenvolvido para a análise de dados multivariados dicotômicos e politômicos usando modelos de variáveis latentes, sob o enfoque da Teoria da Resposta ao Item (RIZOPOULOS, 2006)".
- **Itan** – é um pacote bastante útil para a análise pedagógica dos itens, pois permite gerar as AGI, corrigir o teste e verificar a qualidade do item com base no índice de correlação bisserial relativo a cada uma das alternativas (QUIROZ, 2017).

4.2.2 Análise empírica do teste

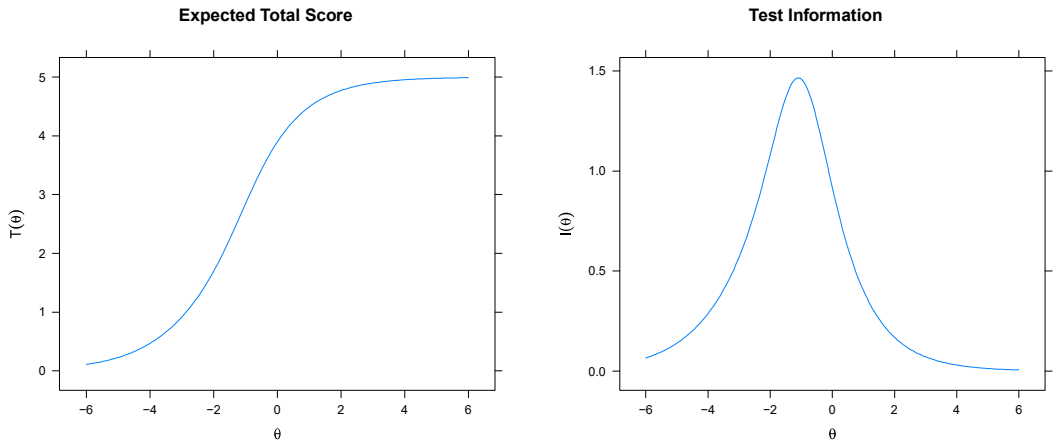
A primeira análise realizada para inferir a qualidade do teste foi baseada na medida *Alfa de Cronbach*. Para essa finalidade utilizou-se a função *descript()* do pacote *ltm*, a qual tem como uma de suas saídas as medidas individuais α_i e do teste α_T. Cabe salientar que a análise clássica é baseada no escore total T, e que este já pressupõe que o teste é a priori homogêneo e válido. Isto quer dizer que o escore total é a crença de que os itens são somáveis e representam adequadamente o traço de um mesmo traço latente (PASQUALI,

2017). Assim, α_T não é a prova de que o teste é válido e homogêneo, mas um indicador de consistência da prova.

Observando a necessidade de se verificar a unidimensionalidade do instrumento, fez-se um teste de unidimensionalidade baseado no *scree plot* e na análise fatorial paralela (*fa.parallel()*). A adequação da AF para o conjunto de dados foi verificado por meio das funções *KMO ()* e *cortest.bartlett()*.

Como medida de validade do instrumento utilizou-se a Curva de Informação do Teste (CIT), baseada na teoria do traço latente θ. Ajustado o modelo e estimado os parâmetros, a função *plot()* foi utilizada para plotar a CIT e a Curva Característica do Teste (CCT). A Figura 9 exibe os gráficos fornecido pelo *Mirt* para um teste de uma dimensão. A Figura 9(a) mostra a curva característica de 5 itens da base de dados LISART7. A curva de informação de um item dessa mesma base de dados é exibida na Figura 9(b).

Figura 9 – Curvas características do Teste e de Curva de Informação do Teste



(a) Curva Característica do Teste (b) Curva de informação do item

Fonte: Chalmers et al. (2012)

4.2.3 Análise empírica dos itens

A análise empírica dos itens envolve a execução de diversos procedimentos estatísticos, os quais, na prática, são a tentativa de verificar se os itens avaliam adequadamente a habilidade que o teste propõe mensurar. De acordo com Pasquali (2017), os parâmetros que apresentam tal análise são, em geral, os de unidimensionalidade, dificuldade, vieses (entre estes, particularmente o chute e a função diferencial, isto é, o DIF), tendenciosidade de resposta, discriminação, validade e precisão.

O primeiro passo dado na análise dos itens, segundo os pressupostos da psicometria clássica, foi o cálculo do índice de dificuldade clássico. O cálculo do índice ocorreu

com base na proporção de respostas dada ao gabarito (resposta correta) de cada item, permitindo, em seguida, classificá-los com os rótulos: "fácil", "média dificuldade"ou "difícil", conforme proposto por (CONDÉ, 2001). Essa tarefa foi executada com auxílio de um *script* elaborado pelo pesquisador no *RStudio* e teve seus valores confirmados na saída dos computadas por meio da função *descript*() do pacote *Ltm*. Os indicadores de dificuldade calculado são importantes porque permitem obter uma visão geral de itens em que os alunos apresentaram baixo domínio e igualmente os itens nos quais tiveram bons desempenhos.

Além da proporção de acertos e erros, a função *descript*() retorna uma medida que pode ser usada como índice de discriminação dos itens, a saber, a medida de correlação Ponto Bisserial (r_{pb}) do item i com o escore total T, expresso por meio da Equação 3.22. Essa medida é frequentemente utilizada na análise empírica do conjunto de tarefas ou itens que compõe um teste, pelo fato de avaliar a contribuição de um item na diferenciação de sujeitos. Em outras palavras, indica a correlação que têm as respostas dadas pelos sujeitos a cada um dos itens com o escore total.

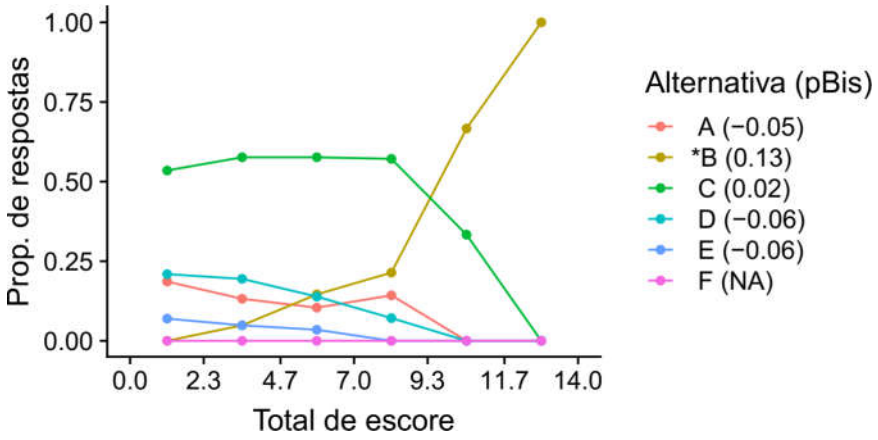
De acordo com o que foi discutido no Capítulo 3, valores de r_{pb} próximos de 1 indicam que as respostas estão altamente correlacionadas com o escore total, e isso é o que se espera de um item, que o mesmo diferencie sujeitos com desempenhos (escores) distinto adequadamente. As implicações dessas medidas na análise empírica dos itens é que itens com baixo poder discriminativo podem ter algum problema de formulação ou que avaliam atributos que os sujeitos submetidos ao teste não possuem (PASQUALI, 2017).

Outra saída da função *descript*() é o *Alfa de Cronbach*, que pode ser usado como medida de precisão do instrumento avaliativo e/ou de consistência interna dos itens. O parâmetro indica a fidedignidade do item e é determinado em termos da variância dos itens e de seus índices de discriminação de acordo como discutido na Subseção 3.2.4. Na análise feita, observou-se itens com consistência interna negativa como indicativo de baixa consistência interna, acarretado pela má formulação do item ou pela falta de habilidade coletiva dos participantes na resolução da tarefa solicitada.

Outra maneira conveniente de verificar a qualidade dos itens é por meio da análise gráfica (AGI). Para essa tarefa utilizou-se o *Itan*, que tem como saída medidas clássicas de coeficientes de correlação Ponto Bisserial e as AGIs individuais dos itens. A Figura 10 mostra que a proporção de respostas fornecidas à respostas correta (alternativa B) cresce em grupos de alunos com escores maiores. É visível que o traço amarelo, correspondente à letra B, tem sua proporção de resposta aumentada conforme o escore alunos aumenta, mostrando de maneira visual que o item em questão é discriminativo.

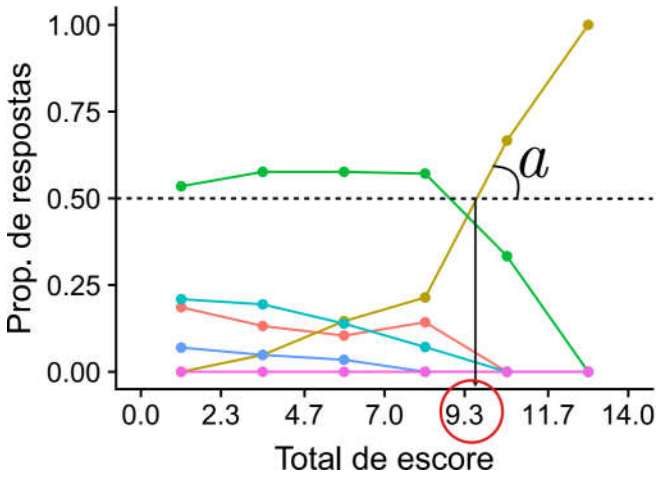
Do gráfico é possível inferir ainda que o item em questão é bastante difícil para esse grupo particular de respondentes, pois o valor do eixo das abscissas que corresponde

Figura 10 – Análise gráfica de um item plotado com auxílio do pacote Itan



à 0,5 no eixo das ordenadas é de aproximadamente 9, quase a metade da maior escore (20). Esses parâmetros são explicitados na Figura 11. A inclinação da linha amarela, neste mesmo ponto, também é bastante considerável, mostrando que esse pode ser um item com boa discriminação. O que pode comprovar essa afirmação visual são as medidas de correlação Ponto Bisserial de cada alternativa. Espera-se minimamente que o gabarito tenha correlação positiva e os demais tenham correlação negativa ou nula.

Figura 11 – Análise gráfica de um item: mostrando os índices de dificuldade e a inclinação da reta



Fonte: Elaboração própria

De posse dos parâmetros baseado na TCT, a etapa seguinte foi calcular os parâmetros a, b e c dos itens com auxílio do pacote *Mirt* e explorar a CCI e CFI. O algorítimo de estimação do *Mirt* decorre de iterações de quadratura fixa de Gauss-Hermite, e exibe boa precisão na estimação de parâmetros nos casos em que o instrumento não possui múltiplas dimensões (CHALMERS et al., 2012). A estimação dos parâmetros logísticos foram obtidos com a função *mirt()*, que utiliza o método de média a *posteriori*. A função

mirt() calibra os itens e realiza o cálculo das cargas fatoriais pelo método *Full Information* (FIFA). O método de carga fatorial costuma ser usual na validação dos itens, pois estabelece a correlação de cada item com o fator predominante, no caso em que o constructo é unidimensional. Os valores de cargas fatoriais e comunalidade foram extraídos pela função *summary()*.

Para a verificação do modelo que melhor se ajusta aos dados, fez-se uso da função *anova()*, do pacote *Mirt*. A função *anova()* compara os modelos logísticos usando estatísticas de razão de verossimilhança, bem como critérios de informação como o AIC e o BIC, cuja hipótese de nulidade postula que há diferença entre os modelos analisados. As duas hipóteses do teste são, portanto,

- H_0 : há diferença entre os modelos analisados.
- H_1 : não há diferença entre os modelos analisados.

Para um nível de significância $\alpha = 0,05$, se $p_{valor} < 0,05$ então rejeita-se a hipótese nula H_0. Os modelos são calculados pelas equações

$$AIC_p = 2\ln(Lp) + 2[(p+1)+1]$$

$$BIC_p = 2\ln(Lp) + [(p+1)+1]\ln n,)$$

em que n é o tamanho da amostra, Lp é a função de verossimilhança do modelo e p é o número de variáveis explicativas. Sua única diferença na prática é o tamanho da penalidade; O BIC penaliza a complexidade do modelo com mais intensidade (SCHWARZ et al., 1978; AKAIKE, 1974).

O *Mirt* possui a função *itemplot*(), usado neste trabalho para obter as CCI e CFI dos itens. Um painel com as CCIs dos itens foi usado para avaliar quais itens seguem os pressupostos da TRI. As CFIs foi o meio de verificar a qualidade do item no que se refere ao seu potencial de informação no processo de mensuração das habilidade dos participantes.

5 RESULTADOS E DISCUSSÕES

Os resultados do estudo exploratório estão organizado em três Seções, os quais apresentam análises exploratórias dos itens do nível 1 da primeira fase da OBMEP, 13ª edição, aplicada a um grupo de alunos da rede pública de ensino da cidade de Santarém, no estado do Pará. A análise ocorre por meio da TCT e TRI. Os índices extraídos no âmbito dessas teorias, subsidiam a análise pedagógica dos itens, visando apresentar dificuldades e até erros cometido pelos alunos em matemática.

5.1 ANÁLISE EXPLORATÓRIA DOS ITENS POR MEIO DA TCT

Os resultados obtidos por meio da TCT, expressos neste trabalho em termos dos índices de dificuldade, correlação Ponto Bisserial e de consistência interna (*Alpha de Cronbach*), mostram que a prova possui baixa qualidade para o grupo testado.

A medida de consistência interna obtido pelo método *Alpha de Cronbach*, também usado como indicador de precisão do teste, foi $\alpha_T = 0,18$, mostrando que o conjunto de itens possui baixa consistência. A análise via *Alpha de Cronbach* busca "verificar a consistência interna do teste através da análise da consistência interna dos itens, isto é, verificando a congruência que cada item do teste tem com o restante dos itens do mesmo teste"(PASQUALI, 2017). Ainda de acordo com Pasquali (2017), a baixa consistência sugere que os itens produzem resultados com elevada variância para a mesma amostra, implicando no aumento do erro, e daí o indicativo de que o teste teve baixa precisão para o grupo de alunos considerado neste estudo.

As medidas de consistência dos itens, juntamente com as de discriminação, obtida pela correlação Ponto Bisserial, estão apresentados na Tabela 9. Junto a esses dois parâmetros, a tabela contém os índices de dificuldade D_i, proporção de erro ERR e a proporção com que cada alternativa foi marcada pelos alunos. Observa-se na Tabela 9 que todas as medidas de α estão abaixo dos valores aceitáveis na literatura especializada. O ideal seria que as medidas estivessem próximas de 1 (PASQUALI, 2017; KLEIN, 2005). Os resultados das correlações Ponto Bisserial também apresentadas na Tabela 9, indicam um baixo poder discriminativo do conjunto de item e, portanto, baixa fidedignidade do teste.

Uma proposta de tomada de decisão a partir dos índices de discriminação é sugerido em Rabelo (2013), conforme apresentado na Tabela 2. Segundo o autor, itens com medida de discriminação abaixo de 0,20 devem ser rejeitados, pois são incapazes de discriminar grupos de alunos com alta proficiência dos que possuem baixa proficiência. Usando os critério proposto por (RABELO, 2013), verificou-se que os itens 5, 10, 11 e 20, com valores ρ_{pb} menores que 0,20, são os que possuem menor poder discriminativo para o grupo de

Tabela 9 – Resultados da análise TCT e estatística dos resultados.

43 height Item	GAB	Índices clássicos				Proporção de resposta por alternativa				
		ERR	D_i	ρ_{pb}	α	A	B	C	D	E
1	A	0,359	0,641	0,357	0,137	0,641	0,135	0,049	0,086	0,089
2	B	0,903	0,097	0,284	0,143	0,126	0,097	0,567	0,166	0,043
3	C	0,911	0,089	0,220	0,164	0,190	0,346	0,089	0,222	0,153
4	D	0,846	0,155	0,320	0,120	0,233	0,204	0,140	0,155	0,268
5	E	0,689	0,311	0,165	0,208	0,254	0,092	0,170	0,173	0,311
6	C	0,740	0,260	0,273	0,162	0,246	0,223	0,260	0,171	0,101
7	D	0,758	0,242	0,240	0,170	0,153	0,305	0,161	0,242	0,138
8	B	0,738	0,262	0,359	0,138	0,444	0,262	0,084	0,150	0,061
9	D	0,885	0,115	0,302	0,134	0,250	0,359	0,141	0,115	0,135
10	B	0,899	0,101	0,146	0,180	0,394	0,101	0,171	0,142	0,191
11	E	0,901	0,099	0,086	0,193	0,134	0,299	0,154	0,314	0,099
12	B	0,747	0,253	0,334	0,151	0,353	0,253	0,261	0,046	0,086
13	A	0,661	0,339	0,295	0,162	0,339	0,148	0,151	0,119	0,243
14	D	0,886	0,114	0,212	0,169	0,484	0,178	0,155	0,114	0,070
15	C	0,609	0,391	0,210	0,204	0,046	0,124	0,391	0,322	0,118
16	A	0,732	0,268	0,223	0,185	0,268	0,193	0,075	0,098	0,366
17	D	0,865	0,135	0,284	0,153	0,225	0,280	0,210	0,135	0,150
18	A	0,633	0,367	0,212	0,208	0,367	0,166	0,201	0,149	0,117
19	C	0,865	0,135	0,203	0,187	0,282	0,343	0,135	0,144	0,095
20	E	0,839	0,161	0,130	0,211	0,259	0,199	0,205	0,176	0,161

Fonte: OBMEP-STM:2017.

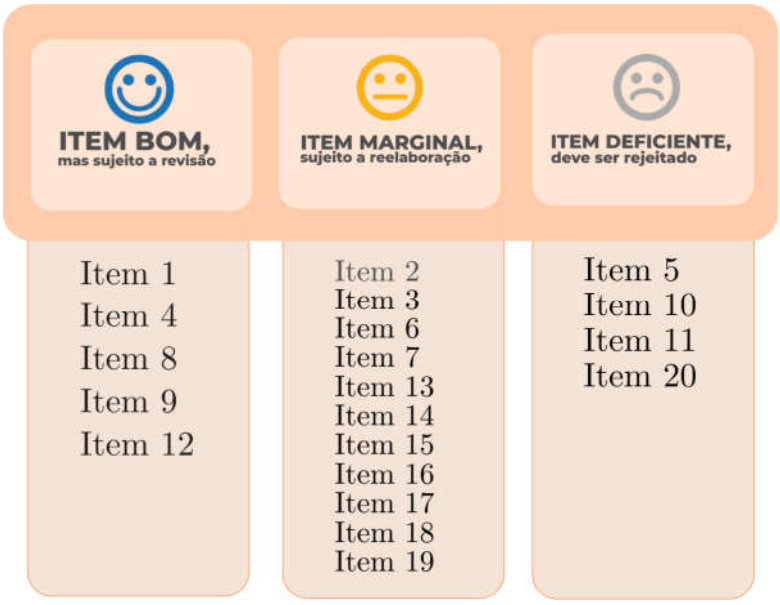
alunos considerado no estudo. Esses itens são deficientes e deveriam, desta forma, não fazer parte do teste, caso o objetivo da prova fosse unicamente avaliar a habilidade desse grupo de alunos.

Os itens 1, 4, 8, 9 e 12, com valores entre 0,30 e 0,40, são considerados itens bons, mas sujeito a aprimoramento. O restante dos itens possuem valores de correlação Ponto Bisserial compreendido entre 0,20 e 0,30, sendo que estes são considerados "itens marginais", sujeitos à reelaboração. Percebe-se que nenhum dos itens do teste poderia ser considerado "bom"segundo o critério proposto por Rabelo (2013). O esquema da Figura 12 mostra a distribuição dos itens de acordo com o seu poder discriminativo.

O índice D_i apresentado na Tabela 9 é considerada uma medida de dificuldade dos itens, o qual numericamente representa a proporção de respostas corretas fornecida a cada item pelo grupo de aluno. De acordo com a classificação proposta por Condé (2001), são de média dificuldade os itens 1, 5, 13, 15 e 18. Os demais itens são difíceis, pois possuem proporção de acerto menor que 0,3. Curiosamente, não há nenhum item "fácil"para o grupo de alunos em análise, sugerindo que o teste tem um grau de dificuldade muito elevado, na visão dos respondentes.

Segundo Rodrigues (2006), é importante que um teste possua itens de todos os níveis de dificuldade, alcançando todo o contínuo da escala. Esse resultado reforça o que foi dito anteriormente: a prova possui fragilidade técnica caso o objetivo fosse avaliar a

Figura 12 – Análise dos itens pelo poder discriminativo representado pelas medidas de correlação Ponto Bisserial (ρ_{pb})



aprendizagem dos alunos nos moldes das avaliações de larga escala.

É interessante notar que a proporção de respostas erradas é consideravelmente elevada para praticamente todos os itens, em particular nos itens 2, 3, 10 e 11 com proporção de respostas erradas superior a 90%. Essa simples estatística chama atenção para uma aparente dificuldade que os respondentes possuem em um número expressivo de situações/problemas cobrados na prova da OBMEP.

5.2 ANÁLISE EXPLORATÓRIA DOS ITENS POR MEIO DA TRI

5.2.1 Unidimensionalidade da prova

A análise da dimensionalidade de um instrumento é feita, em geral, com auxílio da AF. Não obstante, os dados nem sempre se ajustam adequadamente aos pressupostos da AF. Tal verificação é uma etapa importante nesse tipo de estudo, e pode ser feita com as estatística de KMO e Teste de Esferecidade de *Bartlett*.

A estatística de *KMO=0,53* mostra que os dados do estudo possuem má adequação aos métodos da AF. O teste de Esferecidade de *Bartlett* retornou um *p-valor=0,004*, mostrando, por outro lado, que existem correlações suficientes entre variáveis para se continuar à AF.

A extração de fatores para uma única dimensão baseada nos resíduos mínimos, mostra que a maioria dos dos itens da prova possui baixa correlação no fator, conforme a Tabela 10. Observa-se nos dados apresentado valores das cargas fatoriais (MR1), comunalidade (h2) e especificidade (u2), sendo que a comunalidade diz respeito à porção

de variância de cada variável explicada pelo fator e a especificidade (erro) a porção de variância das variáveis que não é explicada nesse fator (HAIR et al., 2009). Na literatura é esperado medidas de comunalidade maior que 0,5 e de especificidade inferior a 0,5.

Tabela 10 – Cargas fatoriais para um único fator

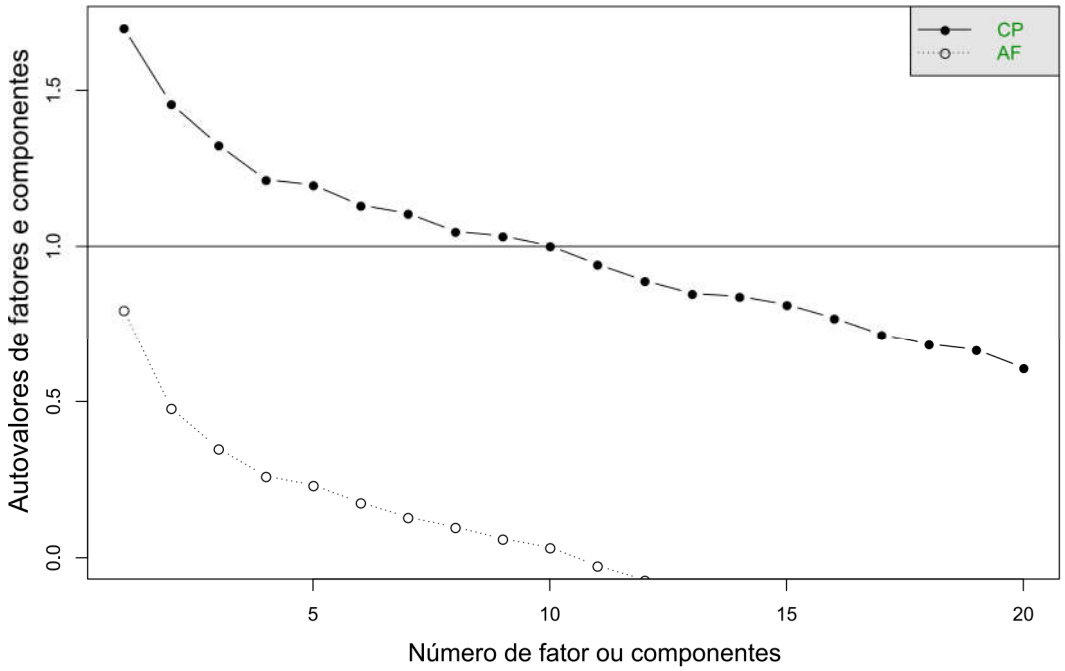
Itens	MR1	h2	u2
1	0,430	0,189	0,810
2	0,840	0,706	0,290
3	0,250	0,064	0,940
4	0,080	0,007	0,990
5	-0,280	0,081	0,920
6	0,200	0,042	0,960
7	-0,100	0,009	0,990
8	0,280	0,076	0,920
9	0,410	0,165	0,840
10	0,360	0,129	0,870
11	-0,300	0,091	0,910
12	0,220	0,050	0,950
13	0,140	0,019	0,980
14	0,160	0,027	0,970
15	0,090	0,008	0,990
16	-0,030	0,001	1,000
17	-0,050	0,003	1,000
18	0,170	0,029	0,970
19	-0,310	0,095	0,910
20	-0,350	0,121	0,880

Observa-se valores negativos de carga fatorial nos itens 5, 7, 11, 16, 17, 19 e 20. Essa medidas não são esperadas para um instrumento unidimensional. Na verdade, deseja-se que os valores da carga fatorial sejam maiores que 0,3 para uma população maior ou igual a 350 indivíduos. Nessas condições, somente os itens 1, 2, 9 e 10 possuem cargas fatoriais aceitáveis para um único fator. Esse resultado chama atenção porque a obtenção de cargas fatoriais para um único fator traz perdas significativas na análise da estrutura do conjunto de dados deste estudo.

O *scree plot* apresentado na Figura 13 mostra que os os autovalores devolvidos pela AF estão abaixo de 1. Esse resultado sugere que o método não é conclusivo para a verificação da dimensionalidade do teste. Apesar disso, os autovalores das componentes principais (PC) indicam a presença de 10 componentes superiores a 1. Na análise fatorial paralela foi indicado a presença de 10 fatores e 9 componentes para o conjunto de dados, reforçando a suspeita de que o instrumento não pode ser reduzido a um único fator.

De acordo com o que foi comentado em seções anteriores, a TRI não verifica a unidimensionalidade do traço latente, apenas a assume como um postulado. Assim, é esperado que para o particular conjunto de dados a TRI não produza bons resultados em todos os itens, uma vez que o pressuposto da unidimensionalidade não é satisfeito. Ainda

Figura 13 – Scree Plot dos dados



assim, nesse trabalho é apresentado os resultados obtidos no âmbito da TRI para fins meramente pedagógicos.

5.2.2 Escolha do modelo e qualidade do teste

Uma das vantagem da TRI é que ela fornece as principais informações relativas à qualidade dos itens após o processo de calibração, que consiste na estimação dos parâmetros característicos dos itens. Antes de estimar os parâmetro pelo método da média a *posteriori* (EAP), verificou-se os modelos nos quais os dados tem melhor ajuste. Para isso foi utilizado a função *anova()*.

A comparação entre o modelo de *Rasch*, de um parâmetro, e o modelo logístico de dois parâmetros (ML2), para um nível de significância $\alpha' = 0,05$, mostra que não há diferenças apreciáveis entre os dois modelos para esse conjunto de dados (Tabela 11). Apesar do valor de AIC ser menor no ML2, a diferença não é significante, mostrando que não há ganhos consideráveis com o modelo de 2 parâmetros. O valor de BIC, por outro lado, é menor no modelo de Rasch. Cabe salientar que o teste BIC penaliza a complexidade do modelo com mais intensidade.

Usando a mesma função para comparar o ML2 ao modelo logístico de 3 parâmetros (ML3), tem-se agora uma diferença considerável entre os dois modelos para o mesmo nível de significância (Tabela 12) . Tanto o valor de AIC quanto de BIC possuem valores elevados para o ML3, mostrando que o ML2 é o mais adequado para o particular conjunto

Tabela 11 – Comparativo entre os ajustes realizados por meio do modelo Rasch e ML2

Modelo	**AIC**	**BIC**	**logLik**	**X^2**	**df**	**p**
Rasch	6789,086	6870,042	-3373,543	NaN	NaN	NaN
ML2	6762,729	6916,932	-3341,365	64,357	19	0

Tabela 12 – Comparativo entre os ajustes realizados por meio do modelo ML2 e ML3

Modelo	**AIC**	**BIC**	**logLik**	**X^2**	**df**	**p**
ML2	6762,729	6916,932	-3341,365	NaN	NaN	NaN
ML3	6779,479	7010,784	-3329,740	23,25	20	0,277

de dados.

O ML2 não considera o parâmetro de acerto casual, oferecendo informações sobre a dificuldade e a discriminação do item. Esperava-se que o modelo mais adequado aos dados fosse o ML3, pois, em geral, testes de múltipla escolha que não contem uma alternativa do tipo "não sei a resposta", como forma de diminuir o chute, o parâmetro c de acerto ao acaso possui importância, pois agrega informações aos itens e consequentemente à análise do teste.

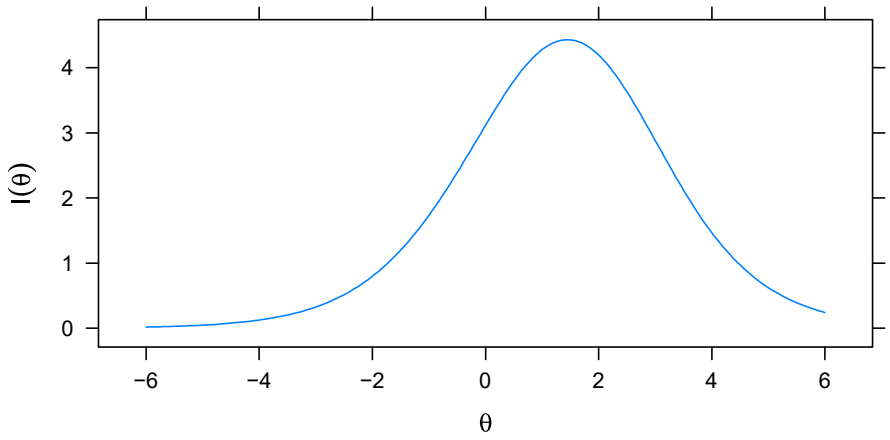
Dentre outras maneiras de avaliar a qualidade do teste com auxílio da TRI, a curva de informação e característica são métodos práticos e visuais. A Figura 14 mostra a curva de informação da prova para os modelo *Rasch*, ML2 e ML3. Observa-se que o modelo de *Rasch* apresenta mais informação sobre a prova em questão que o ML2, o qual contém metade da informação $I(\theta)$ da oferecida no modelo de *Rasch* (Figura 14(a) e na Figura 14(b)).

No caso do ML3, o máximo de informação chega próximos de 80, porém limitado a um pequeno intervalo de θ. Em medidas de habilidade entre -6 e 0, o teste oferece pouca informação. Por esse motivo a forma da curva é fechada e deslocada à direita. Espera-se que o teste ofereça mais informação sobre o traço latente, porém levando em consideração um amplo intervalo da escala de habilidade, e isso não ocorre para o ML3. A curva de informação do ML3 mostra que o teste oferece informações somente sobre candidatos com alto nível de habilidade.

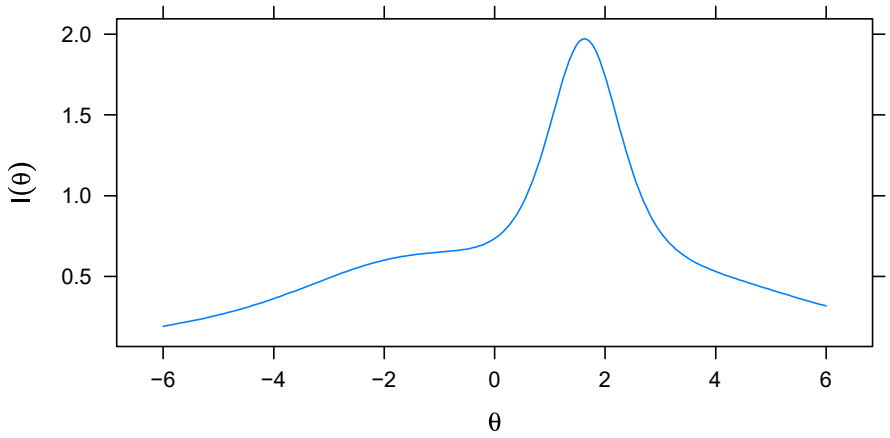
Usando a curva característica do teste (CCT) (Figura 15), observa-se comportamentos estranhos tanto no resultados obtidos para o ML2 quanto para o ML3. A ideia básica da TRI é de que quanto maior é o nível de habilidade do sujeito submetido a um teste, maior é a probabilidade que ele tem de responder corretamente um item ou conjunto de itens. Para valores de θ entre -6 e 0, a CCT obtida no ML3 mostra um decrescimento com o aumento do nível de habilidade, o que é incoerente com o paradigma da TRI. O mesmo ocorre no ML2.

Os resultados gráficos obtidos na TRI mostram que o ML3 possui os piores ajustes para os dados do estudo. Assim, o ML2, apesar de apresentar um comportamento estranho

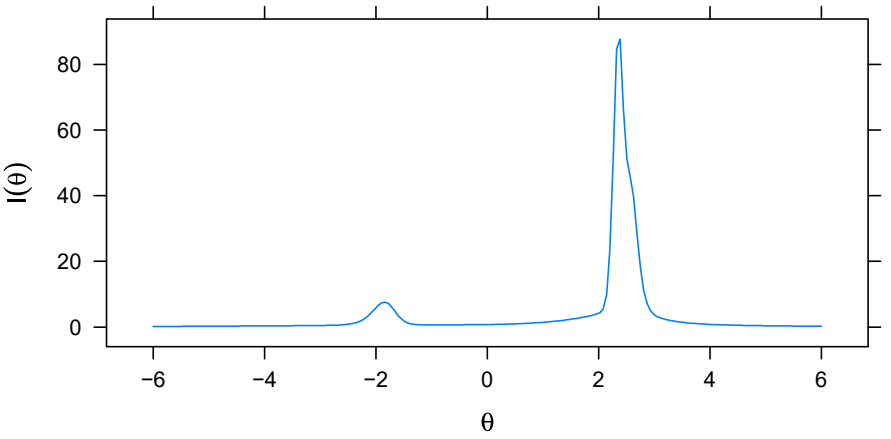
Figura 14 – Curvas de informação do teste para três modelos da TRI



(a) Informação do teste: Rasch



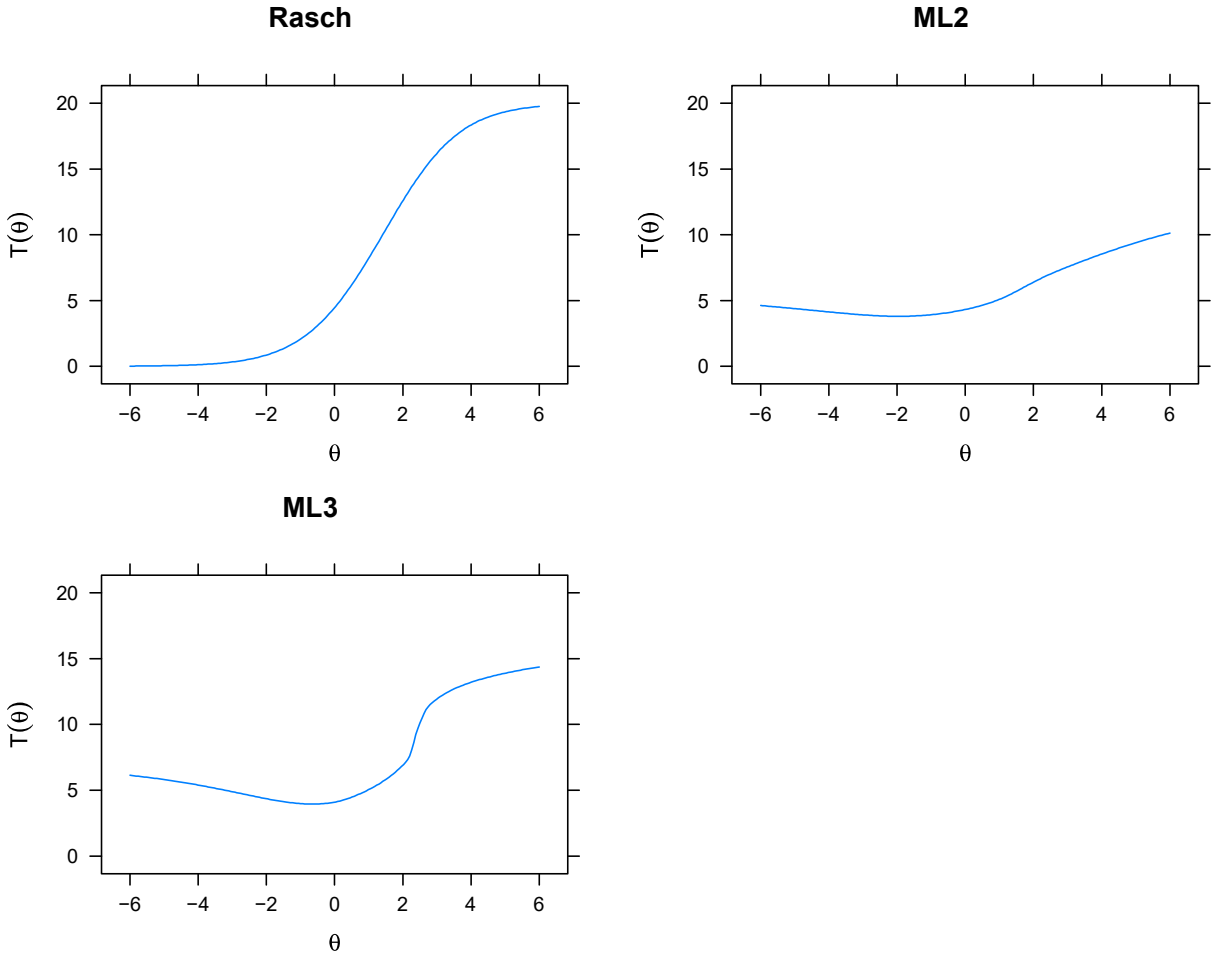
(b) Informação do teste: ML2



(c) Informação do teste: ML3

na sua CCT, pode ser usado para fornecer informações da dificuldade e discriminação dos itens.

Figura 15 – Curva característica do teste, modelos Rasch, ML2 e ML3



5.2.3 Análise individuais dos itens

A Tabela 13 contém parâmetros de discriminação dos itens (a_i) e dificuldade (b_i), calculados por meio do ML2. A calibração revela itens com ajustes inadequados aos pressupostos da TRI.

Os parâmetros a_i assumem valores negativos nos itens 5, 7, 11, 16, 17, 19 e 20, evidenciando que eles não estão de acordo com o pressuposto fundamental da TRI; o de que maiores habilidades resultam em maiores probabilidades de acerto ao item. Em uma pré-testagem, esses itens seriam eliminados do instrumento avaliativo. Segundo Rodrigues (2006), resultados ruins à luz das teorias psicométricas não necessariamente indicam problemas de formulação do item, como podem sugerir a falta de habilidade coletiva, exigida para fornecer as respostas corretas. Tal constatação faz sentido principalmente no caso

Tabela 13 – Parâmetros de dificuldade e discriminação dos itens, obtidos via TRI

Item	a_i	b_i
1	0,84	-0,79
2	2,35	1,63
3	0,51	4,77
4	0,15	11,43
5	-0,49	-1,71
6	0,37	2,93
7	-0,16	-7,34
8	0,63	1,80
9	0,76	2,96
10	0,59	3,94
11	-0,48	-4,80
12	0,42	2,71
13	0,24	2,89
14	0,26	7,90
15	0,20	2,20
16	-0,09	-11,77
17	-0,06	-30,28
18	0,30	1,88
19	-0,40	-4,73
20	-0,87	-2,16

Fonte: OBMEP-STM:2017.

aqui investigado, visto que as provas da OBMEP têm perfil classificatório, sendo composta por itens cuja maioria dos alunos não estão habituados a responder em sala de aula.

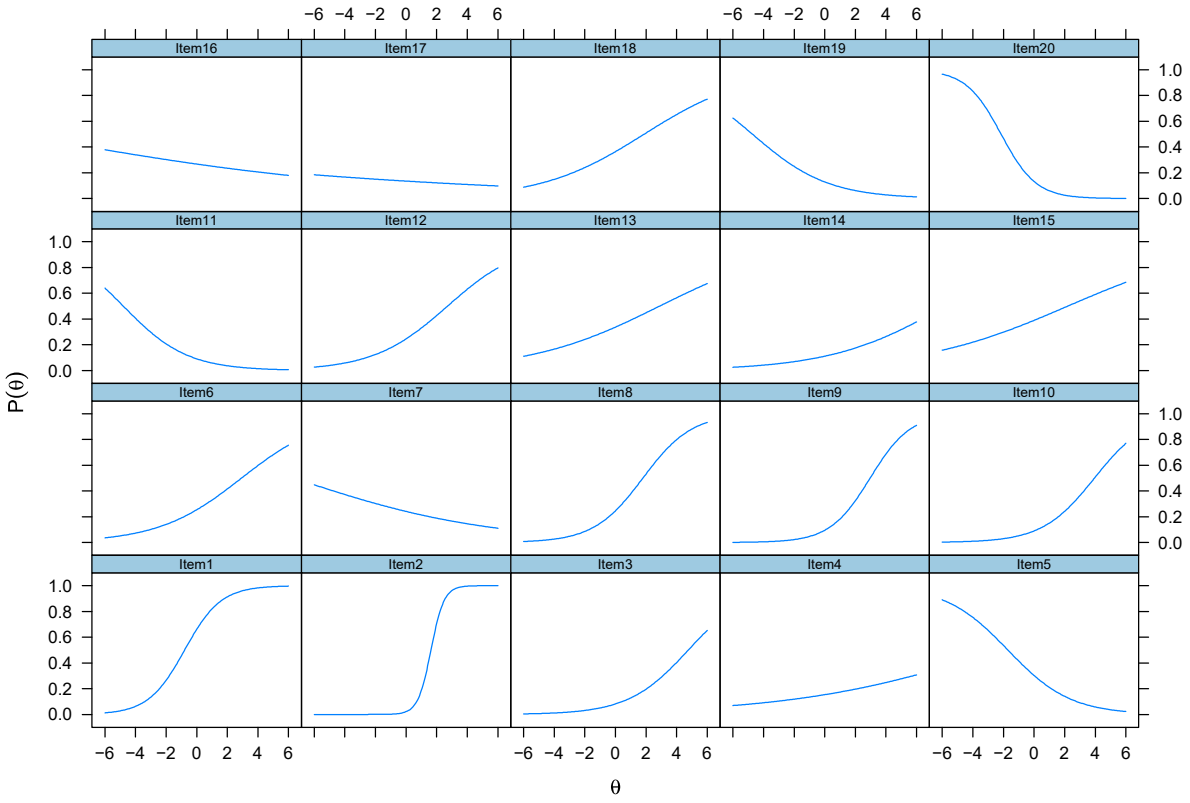
É interessante notar que os itens 5, 7, 11, 16, 17, 19 e 20 já haviam sido detectados como itens enviesados na AF exploratória. Felizmente a TRI detectou que eles não estão se referindo à mestra proficiência. Em outras palavras, os referidos itens deveriam ser analisados separadamente se o objetivo do trabalho fosse estimar a proficiência dos alunos. Neste caso, a estimação da proficiência para o conjunto de dados com 9 fatores predominantes, exigiria a produção de 9 escalas de habilidade – tarefa que não está no escopo desse trabalho. Outra maneira de lidar com a muldimensionalidade do conjunto de dados é excluindo itens com baixa carga no fator ou itens com valores de discriminação negativo.

De acordo com a metodologia de classificação dos itens, fundamentado no nível de dificuldade, apenas o item 1 pode ser considerado fácil (RABELO, 2013). Excluindo dessa análise os itens com parâmetro a_i negativo, todos os demais itens são considerados muito difíceis para o publico testado. Nesse caso, a TRI apenas confirmou que a prova possui alto grau de dificuldade, segundo os respondentes.

As CCI apresentadas na Figura 16, fornece uma representação visual da probabilidade de resposta ao item em função do traço latente. O Traço latente θ é representado no eixo das abscissas e a probabilidade de resposta ao item é dado no eixo das ordenadas.

Observa-se que itens anteriormente mencionados (5, 7, 11, 16, 17, 19 e 20) possuem curva decrescentes, enquanto os demais crescem com o aumento do valor de θ. Esse método visual permite identificar, sem esforço, os itens que não seguem os pressupostos da TRI. Visualmente, o item 4 tem baixo poder disciminativo, uma vez que exibe um pequeno ângulo no ponto em que a probabilidade de resposta ao item é igual a 0,5.

Figura 16 – Curva Característica dos Itens obtidas via ML2



Os parâmetros também foram estimados via ML3, tendo em vista que nesse modelo apresentou mais informações aos grupos com nível de habilidade elevada. O ML3 considera a probabilidade de acerto ao acaso e isso pode trazer informações relevantes. A Tabela **??** contém os parâmetros de dificuldade, discriminação e de acerto ao acaso. Chama atenção as medidas de acerto ao acaso superiores a 20% nos itens 7, 16 e 18. Supostamente, esses três itens são os que os sujeitos com baixa habilidade possuem maiores chances respondê-los corretamente, sem o domínio das habilidades exigidas.

Os itens 7, 14 e 17 são itens com índice de discriminação elevado, pois seus valores de a_i são atípicos entre os itens do estudo. Nas medidas de discriminação a_i, observa-se valores negativos nos itens 5, 11, 16, 19 e 20, mostrando que esses itens não se adéquam aos axiomas da TRI. De acordo com Andrade, Tavares e Valle (2000), parâmetros de discriminação negativo afetam os parâmetros de dificuldade, o que afeta a análise de seus índices de dificuldade.

Usando a classificação proposta por Rabelo (2013) para os níveis de dificuldade baseado no parâmetro b_i, considera-se fácil apenas o item 1. O item 15 pode ser considerado difícil e os demais, todos muito difíceis. É importante ressaltar que os itens 5, 11, 16, 19 e 20 não estão sendo levados em consideração nessa classificação.

Tabela 14 – Resultados da análise TRI via ML3:parâmetro de discriminação a_i, dificuldade b_i e de acerto ao acaso c_i.

Item	a_i	b_i	c_i
1	1,071	-0,666	0,001
2	2,176	1,665	0,000
3	2,988	2,319	0,067
4	2,846	2,498	0,140
5	-0,559	-2,103	0,083
6	0,385	2,811	0,001
7	15,907	2,349	0,235
8	0,631	1,790	0,002
9	1,290	2,502	0,049
10	0,605	3,993	0,009
11	-4,381	-3,748	0,000
12	0,999	2,500	0,165
13	0,389	3,377	0,154
14	14,879	2,323	0,105
15	0,371	1,241	0,000
16	-6,653	-1,805	0,237
17	12,752	2,580	0,131
18	0,654	3,069	0,268
19	-0,390	-6,065	0,049
20	-0,775	-2,372	0,000

Fonte: OBMEP-STM:2017.

5.3 ANÁLISE PEDAGÓGICA DE ALGUNS ITENS

Os resultados da TRI e AF exploratória elucidaram que no particular conjunto de dados, alguns itens não são adequados para o grupo de respondentes, supostamente pelo elevado nível de dificuldade ou pela falta de habilidade coletiva dos alunos. A metodologia empregada mostrou, mais precisamente, que os itens 5, 7, 11, 16, 17, 19 e 20 não avaliam adequadamente os estudantes da amostra, mas tal constatação isolada não traz informações apreciáveis para uma discussão pedagógica. Deve-se conhecer mais a fundo os assuntos e as habilidades avaliadas em cada um desses itens, bem como as tendências com que os alunos responderam a prova e consequentemente equívocos cometidos em função da falta de domínio dos conteúdos, conceitos e procedimentos matemáticos.

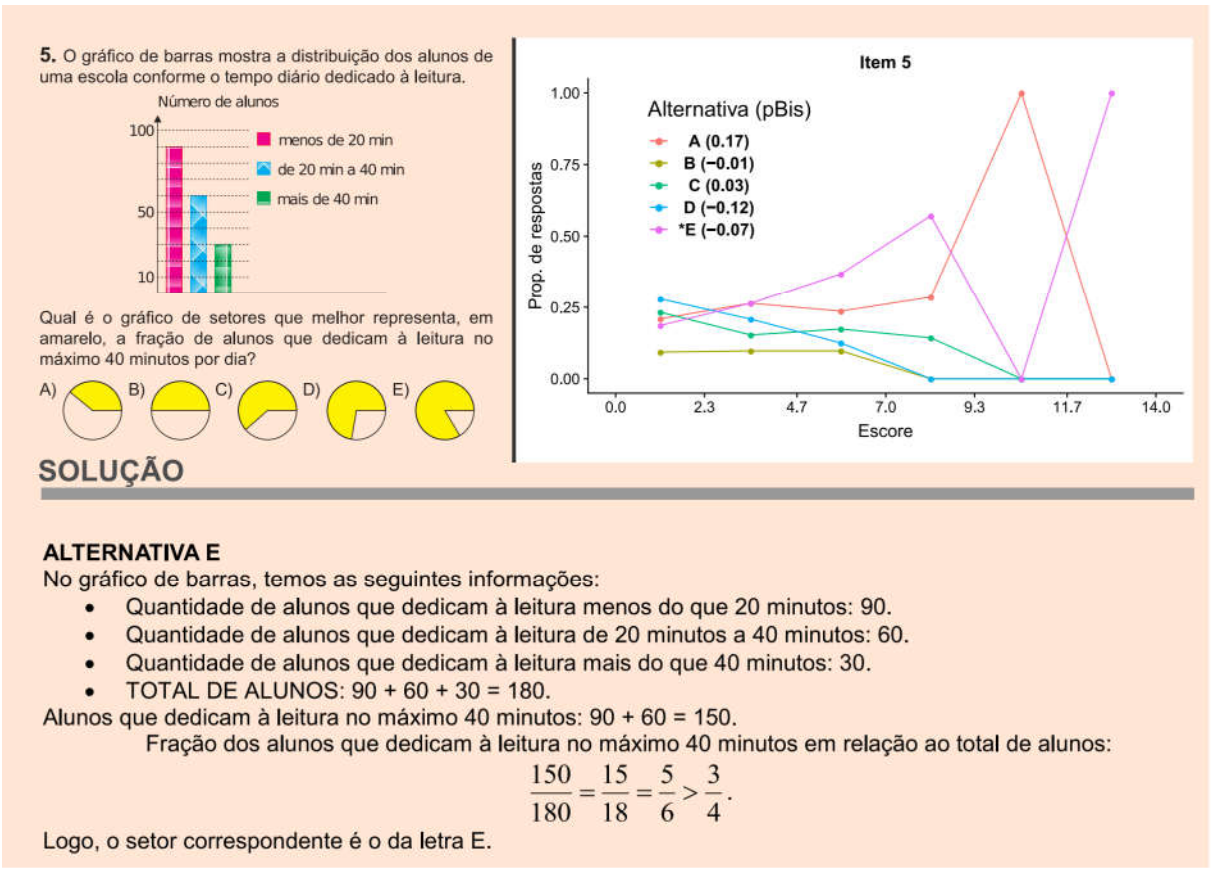
Em geral, a análise pedagógica é feita com o apoio de documentos oficiais que determinam conteúdos e objetivo de aprendizagem. O PCN – documento que define habi-

lidade e competências por ano/série escolar desde 1997/1998 –, é utilizado como diretriz deste estudo, uma vez que os itens não são baseados em uma matriz de referência específica, de elaboração da comissão organizadora da OBMEP. A recente Base Nacional Comum Curricular (BNCC) também não foi considerada na elaboração do certame.

Nesta análise, apresenta-se temas e habilidades cobradas nos itens 5, 7, 11, 16, 17, 19 e 20, além de possíveis equívocos cometidos em função da falta de domínio dos conceitos e procedimentos matemáticos previsto no PCN. Cada item é acompanhado da resolução proposta pelos organizadores do certame, AGI e coeficientes Ponto Bisserial do gabarito e distratores. O Gabarito de cada item é diferenciado dos distratores pelo símbolo "*"(asterisco), como por exemplo "*C".

O item 5, apresentado na Figura 17, tem como tema do PCN *Tratamento da informação* e diz respeito ao conceito de "Coletar, organizar os dados e utilizar-se de recursos visuais adequados (fluxogramas, tabelas e gráficos) para sintetizá-los, comunica-los e permitir a elaboração de conclusões"(BRASIL, 1998, p.). Além desses conceitos e procedimento, o item engloba *Conhecimentos Numéricos*, pois exige o "reconhecimento de números racionais em diferentes contextos cotidianos e históricos e exploração de situações-problema em que indicam relação parte/todo, quociente, razão ou funcionam como operador"(BRASIL, 1998, p. 71).

Figura 17 – Item 5: Análise gráfica e resposta.



O índice de dificuldade clássico elucida que apenas 31% dos alunos responderam o item corretamente, o que caracteriza o item 5 como difícil, isso de acordo com as respostas dos 349 alunos participantes da pesquisa. Aproximadamente 25% dos alunos optaram pela letra A.

Para responder o item corretamente, o aluno deve identificar no gráfico, a quantidade total de alunos que se dedicam à leitura e a quantidade de alunos que se dedicam a leitura por no máximo 40 minutos diário. Depois disso, deve expressar o número de alunos que leem no máximo 40 minutos diários e a quantidade total de leitores por meio de uma fração e reduzi-la. O aluno teria ainda que identificar uma representação gráfica dessa fração.

A AGI mostra correlação Ponto Bisserial positiva para o distrator **A** e **C**. Isso quer dizer que esse distrator tem alta correlação com o escore total, ou seja, conforme aumentam o escore dos alunos aumenta também a proporção de sujeitos que optam por esse distrator. Para marcar a alternativa **A**, que tem maior medida de correlação, o respondente deveria considerar a fração de leitores que leem de 20 a 40 minutos e o total de leitores. Essa fração é $\frac{60}{180} = \frac{1}{3}$, que corresponde aproximadamente ao gráfico de setores do distrator **A**. Supostamente grande parte dos alunos, mesmo os com bons desempenhos, não souberam fazer a leitura do gráfico e/ou interpretar o que o item pede como resposta.

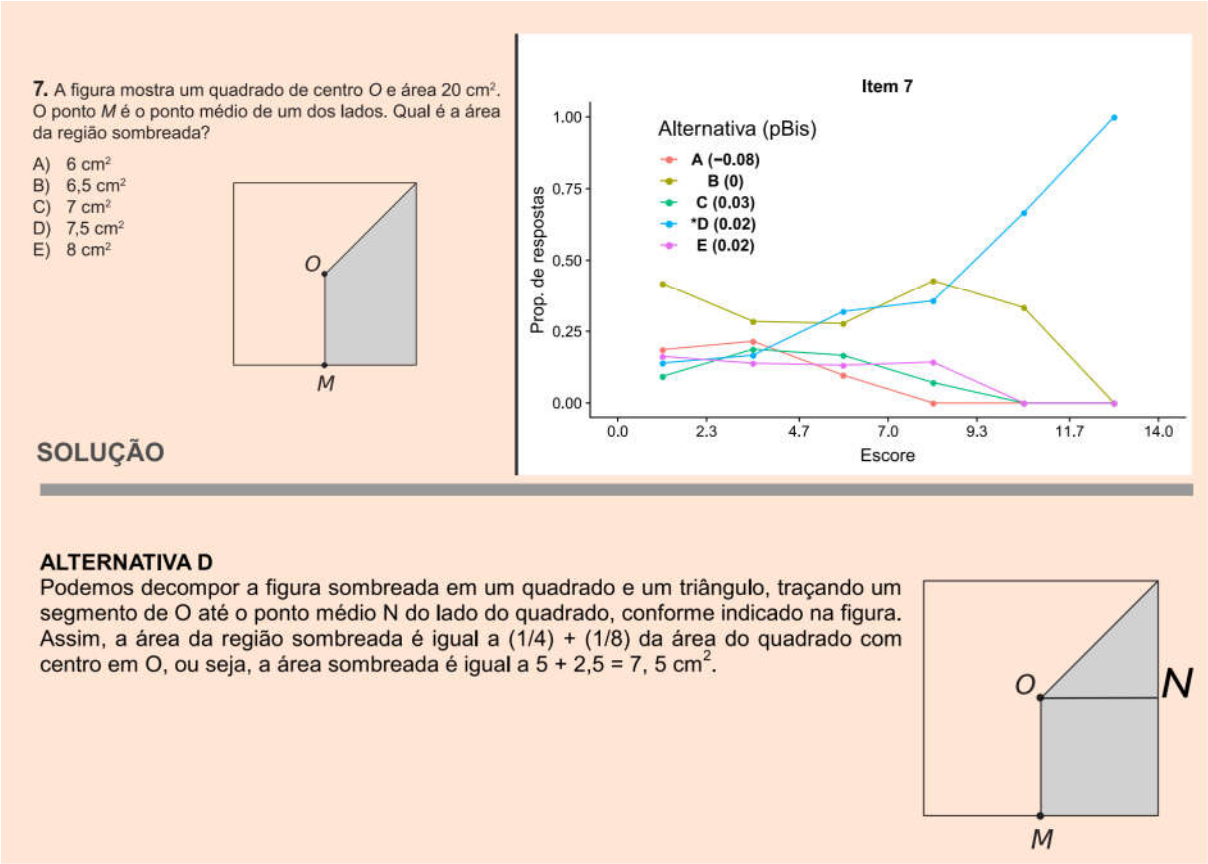
Apresentado na Figura 18, o item 7 refere-se aos temas *Grandezas e Medidas* e *Espaço e Forma*. Os conceitos e procedimentos matemáticos abaliza-se em "calcular área de figuras planas pela decomposição e/ou composição em figuras de áreas conhecidas, ou por meio de estimativas"(BRASIL, 1998, p. 74).

Um percentual de 24% dos alunos responderam o item corretamente, o que o caracteriza como um item difícil. A AGI exibida na Figura 18 demonstra que a alternativa correta se destaca a medida que o escore aumenta. O valor da correlação Ponto Bisserial é positivo no gabarito, mas é no distrator **C** que possui a mais alta medida de correlação com o escore total ($\rho_{pBis} = 0,03$). Apesar de não possuir bom ajuste no ML2, visualmente o item é discriminativo e, por suposto, tem melhor ajuste no ML3, cujo poder discriminativo é evidenciado com mais clareza.

A Figura 19, apresenta o item 11, cujo tema é *números e operações*. Os conceitos e procedimentos cobrados no item está relacionado com "reconhecer os números racionais em diferentes contextos – cotidiano e históricos – e explorar situações-problema em que indicam relação parte/todo, quociente, razão ou funcionam como operador"(BRASIL, 1998). Portanto, mais um item que avaliar a capacidade do sujeito utilizar números racionais para resolver situações-problema.

Os índices clássicos mostram que o item em questão é difícil para o grupo testado, pois somente 9,9% dos alunos responderam o item corretamente. O item é pouco discriminativo, tendo um valor de correlação $\rho_{pb} = 0,08$.

Figura 18 – Item 7: Análise gráfica e resposta.



Na situação-problema vista no item 11, os alunos deveriam usar o conceito de parte por um todo para obter a quantidade de litros de cada cor primária necessária para formar cada cor secundária. O problema poderia ser solucionado sem o uso de frações, bastando aplicar o método pictórico de barras, conhecido na literatura como método chinês (MALTA; LOPES, 2018). A solução da questão usando o método pictórico é apresentado na Figura 20.

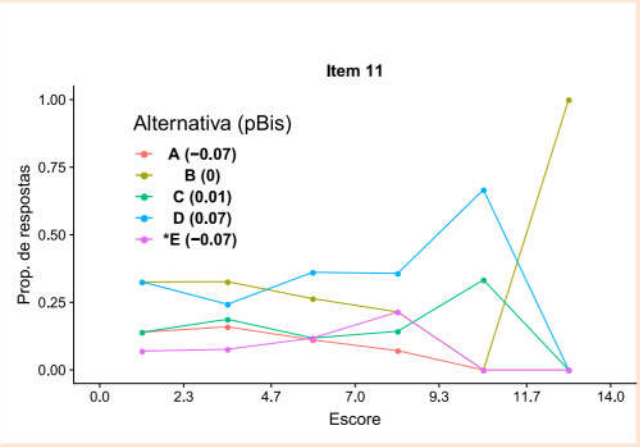
No item em análise, o coeficiente Ponto Bisserial é positivo nos distratores **C** e **B**. Tal fato chama atenção porque os alunos tiveram tendência a marcar esses dois distratores, incluindo aqueles com escore bruto elevado.

Os alunos estariam chutando o item ou cometendo algum erro devido à falta de domínio dos conceitos e procedimento necessário à resolução do problema? No que se refere ao chute, os parâmetros da TRI não oferece informação. Já na AGI, que considera a maneira como estão dispostas as proporções com que os alunos escolhem cada alternativa em diferentes níveis de escore, pode-se pressupor que os candidatos não estão preparados para lidar com esse tipo de situação-problema. Apesar dos esforços, não foi detectado passos lógicos pelos quais os alunos pudessem chegar nos distratores **C** e **B**. Como sugestão aos professores, recomenda-se o uso desta categoria de problema em sala de aula, explorando diferentes formas de resolução, inclusive pelo método pictórico.

Figura 19 – Item 11: Análise gráfica e resposta.

11. Para obter tinta de cor laranja, devem-se misturar 3 partes de tinta vermelha com 2 partes de tinta amarela. Para obter tinta de cor verde, devem-se misturar 2 partes de tinta azul com 1 parte de tinta amarela. Para obter tinta de cor marrom, deve-se misturar a mesma quantidade de tintas laranja e verde.
Quantos litros de tinta amarela são necessários para obter 30 litros de tinta marrom?

A) 7
B) 8
C) 9
D) 10
E) 11

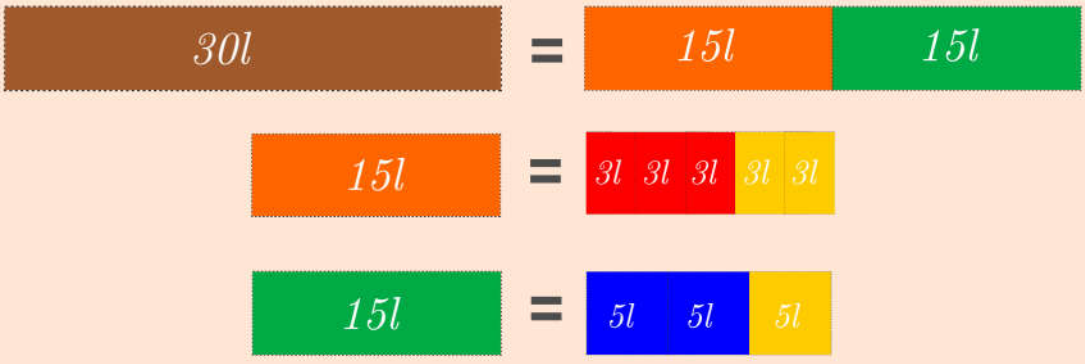


SOLUÇÃO

ALTERNATIVA E
Para obter 30 litros de tinta marrom, precisaremos de 15 litros de cada uma das cores laranja e verde. A primeira condição nos diz que, para obter essa quantidade de tinta laranja, precisaremos da fração $\frac{2}{3+2} = \frac{2}{5}$ da quantidade total exclusivamente de tinta amarela, ou seja, $\frac{2}{5} \cdot 15 = 6$ litros de tinta amarela. Da mesma forma, a segunda condição nos diz que, para obter 15 litros de tinta verde, precisaremos da fração $\frac{1}{2+1} = \frac{1}{3}$ da quantidade total exclusivamente de tinta amarela, ou seja, $\frac{1}{3} \cdot 15 = 5$ litros de tinta amarela. Portanto, a quantidade total de tinta amarela necessária é $5 + 6 = 11$ litros.

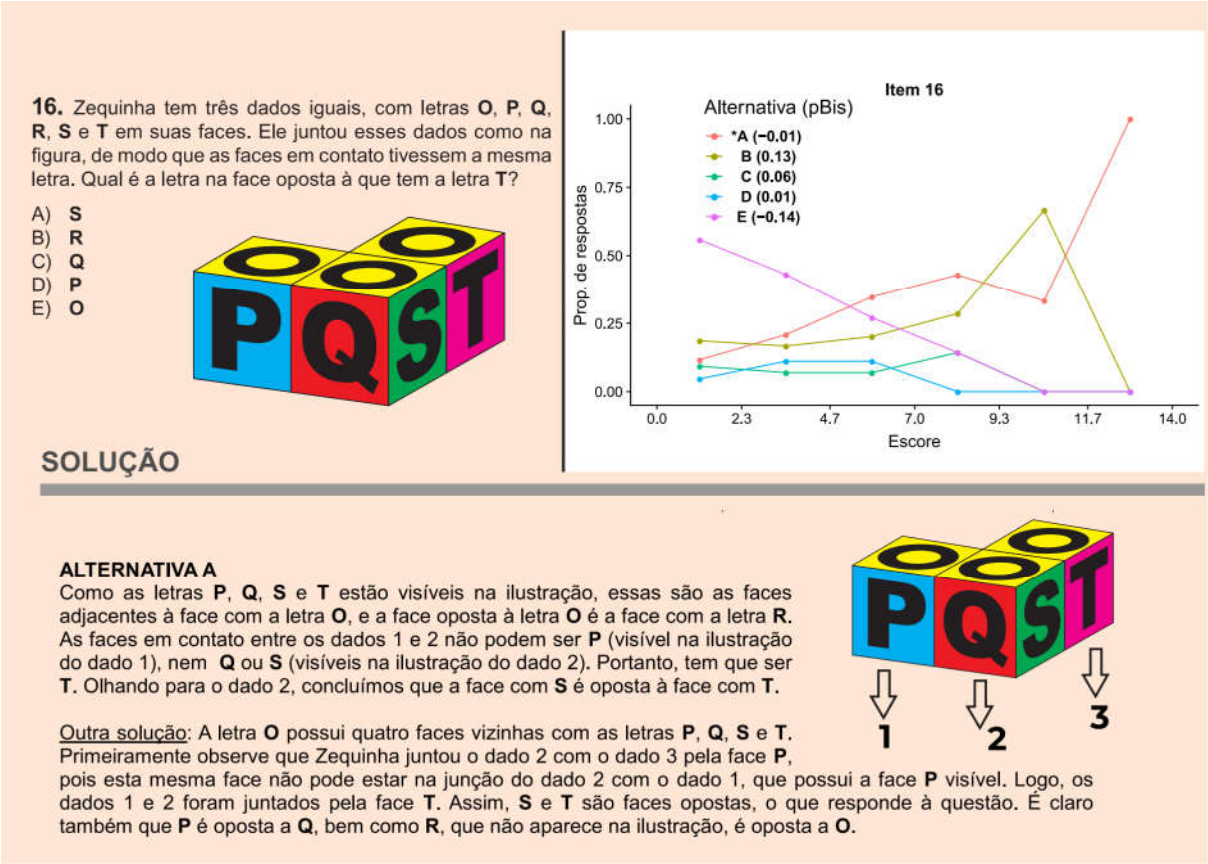
Figura 20 – Solução pictórica do item 11

SOLUÇÃO PICTÓRICA



Logo, a quantidade de tinta amarela é
$3l + 3l + 5l = 11l$.

Figura 21 – Item 16: Análise gráfica e resposta.



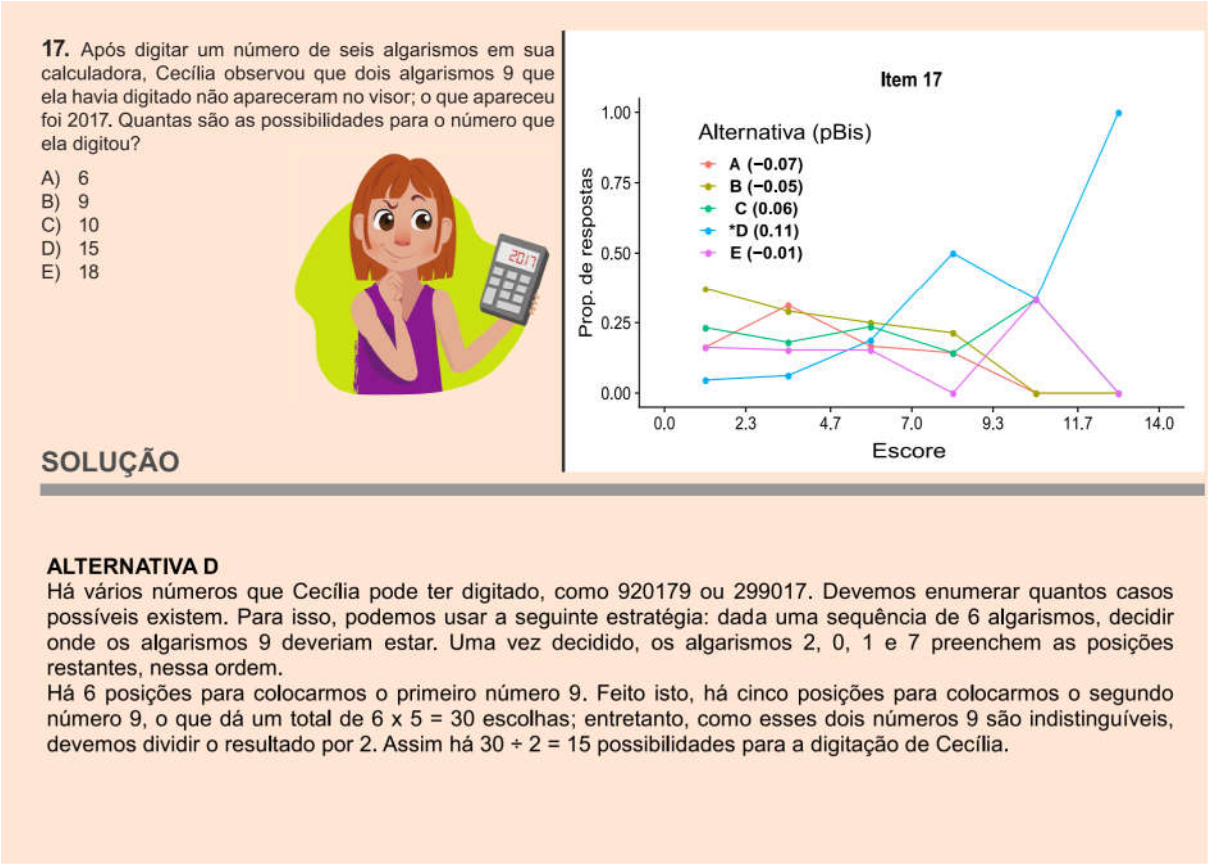
Aproximadamente 26,8% dos alunos responderam o item 16 corretamente. O item em questão não se encaixa explicitamente em nenhum dos conceitos e procedimentos definidos no PCN, mas atende ao seguinte objetivo de aprendizagem: "estabelecer relações entre figuras espaciais e suas representações planas, envolvendo a observação das figuras sob diferentes pontos de vista, construindo e interpretando suas representações" (BRASIL, 1998). Versando sobre o tema *espaço e forma*, o item exige que o aluno saiba utilizar noções de visualização geométrica e raciocínio lógico para resolver a situação problema.

Na análise gráfica relativa ao item 16, observa-se que o distrator **B** possui alta medida de correlação Ponto Bisserial. O gabarito, por sua vez, tem valor de correlação negativo, o que não é típico em avaliação educacional.

O item 17 tem como tema números e operações, e propõe uma situação no qual os alunos devem dominar conceitos e procedimentos de "Resolução de problemas de contagem, incluindo os que envolvem o princípio multiplicativo, por meio de estratégias variadas, como a construção de esquemas e tabelas" (BRASIL, 1998, p. 72)

Para resolver o problema proposto os alunos deveriam dominar o princípio fundamental da contagem ou mesmo contar cada caso usando um diagrama de árvore e/ou tabela.

Figura 22 – Item 17: Análise gráfica e resposta.



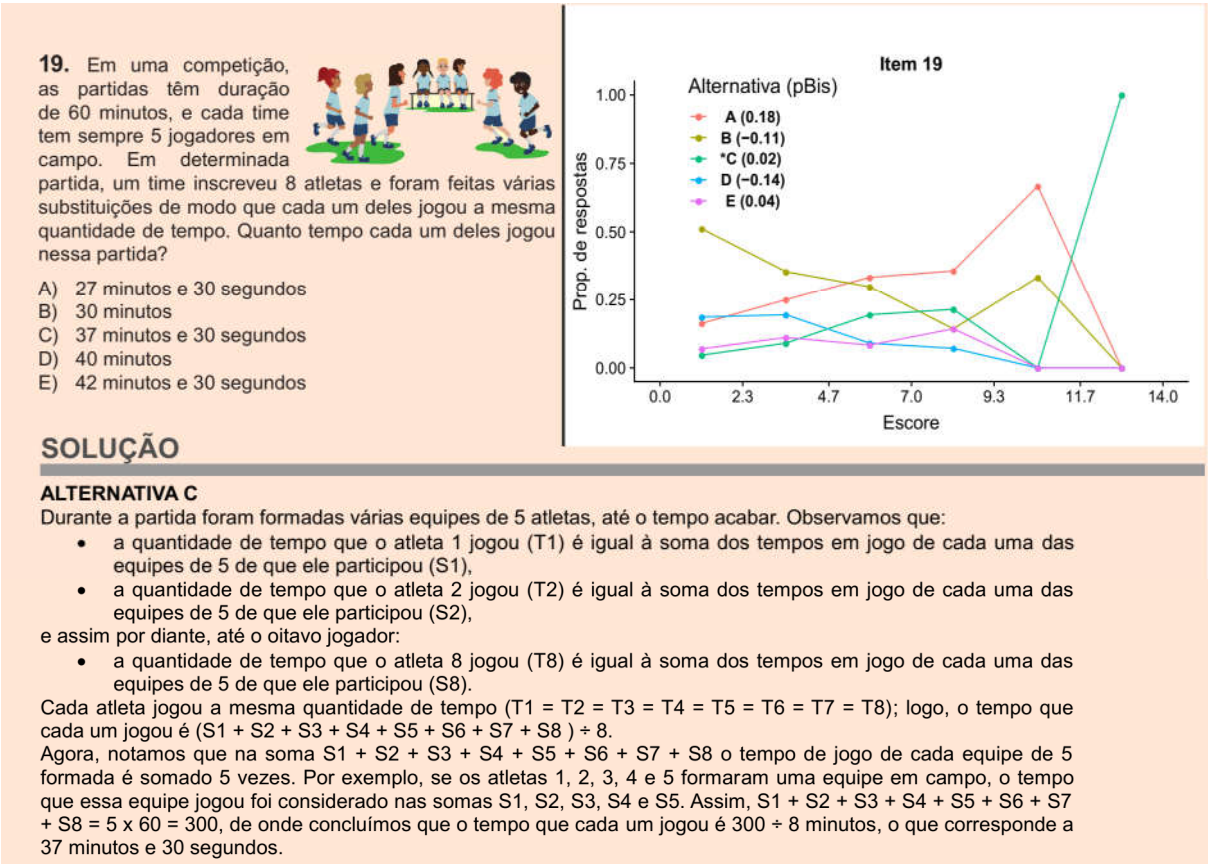
Os índices da TCT advertem que o item em discussão possui alto grau de dificuldade, uma vez que apenas 13,5% dos alunos o respondem corretamente. Na AGI exibida na Figura 22 observa-se que a resposta correta, apresentada na letra **D**, tem correlação Ponto Bisserial positiva, o que é esperado em avaliações educacionais. O distrator **C** também possui correlação positiva, indicando uma tendência dos alunos a optarem pela alternativa, ou por algum tipo de erro ou "chute".

Os parâmetros característicos do item 17 alcançado via ML3, mostra que ele tem elevado poder discriminativo ($a_{17} = 12,75$) e dificuldade ($b_{17} = 2,58$). O parâmetro de acerto ao acaso foi $c_{17} = 0,13$, reforçando a hipótese de que parcela dos alunos tenham respondido o item de forma aleatória, sem o domínio dos conceitos e procedimentos necessários para a resolução do problema.

Com índice de dificuldade igual a 0,135, o item 19 foi difícil para o particular grupo de alunos. A situação-problema se refere aos temas números e operações e grandezas e medidas, cujos conceitos e procedimentos são:

> Análise, interpretação, formulação e resolução de situações problema, compreendendo diferentes significados das operações, envolvendo números naturais, inteiros e racionais, reconhecendo que diferentes situações-problema podem ser resolvidas por uma única operação e que eventual-

Figura 23 – Item 19: Análise gráfica e resposta.



> mente diferentes operações podem resolver um mesmo problema (BRASIL, 1998, p. 71).
>
> Reconhecimento de grandezas como comprimento, massa, capacidade, superfície, volume, ângulo, tempo, temperatura, velocidade e identificação de unidades adequadas (padronizadas ou não) para medi-las, fazendo uso de terminologia própria (BRASIL, 1998, p. 73).

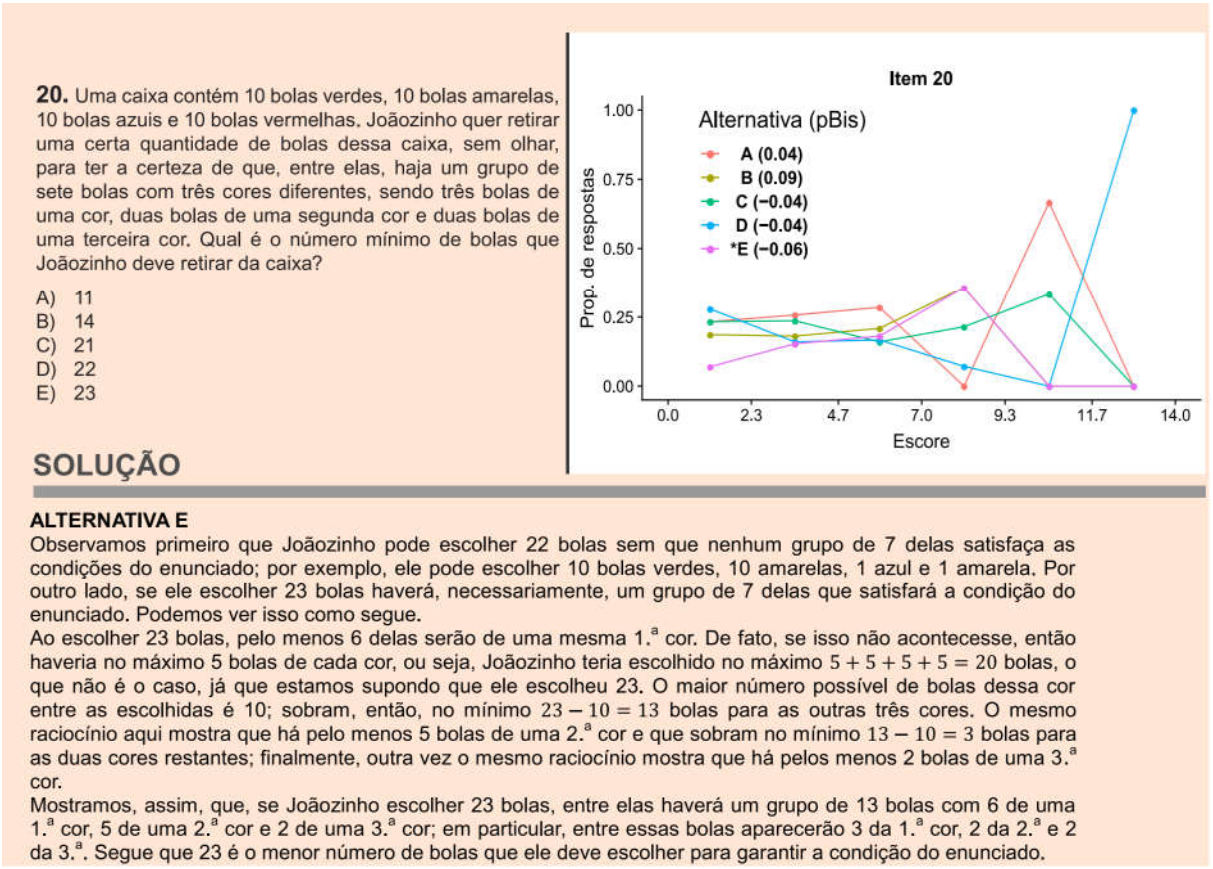
Para resolver a situação-problema deve organizar o pensamento na forma de operações com números inteiros e reconhecer a grandeza tempo, na escala minutos e segundos. Os alunos deveriam perceber que independentemente da quantidade de jogadores na equipe, o tempo total jogado só depende da quantidade de jogadores em campo. Se a partida demora 60 minutos, então o tempo total jogado pela equipe, incluindo as substituições, é 5×60. Por conseguinte, eles deveriam saber quanto cada jogador iria jogar, bastando dividir por 8. Por fim deveriam apresentar o resultado em minutos e segundos, exigindo o conhecimento de 1 minuto possui 60 segundos.

A AGI e as medidas de correlação de cada alternativa, apontam que o item 19 é inadequado para fins avaliativos, tendo em vista determinados critérios. O primeiro deles adverte que o distrator A possui a mais alta correlação ($pBis = 0,18$) com o escore do teste, ao passo que a resposta correta, presente na letra C, exibe correlação na ordem de 0,02, havendo uma tendência dos alunos a optarem por essa alternativa, mesmo aqueles

com alto escore. Outro critério baseado na AGI é o de que a proporção de alunos que optam pela resposta correta diminui conforme aumenta seus escores brutos, o que não é esperado de um item cuja finalidade é avaliar níveis de aprendizagem do sujeito.

Há, assim, um indicativo de que os alunos não sabem lidar com situações como a apresentada no item, faltando habilidade para interpretar e propor soluções baseadas em operações básicas.

Figura 24 – Item 20: Análise gráfica e resposta.



O último item da prova versa sobre números e operações. Na situação proposta, os alunos devem dominar conceitos e procedimentos pertinente ao item 19.

O item é considerado difícil na visão dos respondentes, exibindo índice de dificuldade clássico na ordem de 0,161. A AGI mostra que não há um destaque da alternativa correta, apontando que os alunos podem ter respondido este item de maneira aleatória. Ademais, observa-se medidas de correlação $pBis$ positivo nos distratores A e B, enquanto no gabarito, valor negativo.

6 CONSIDERAÇÕES FINAIS

A OBMEP foi apresentada neste trabalho como política educacional relevante no contexto do ensino da matemática no Brasil, uma vez que desponta de ações integradas que visam a difusão de conhecimentos matemáticos. Apesar de ser uma política que envolve atividades de ensino, a OBMEP é mais conhecida pelas suas provas, cuja finalidade é descobrir talentos para a matemática em todo país. E para tanto, as provas ocorrem em duas fases, sendo que na primeira os alunos são submetidos à diversas situações-problemas presentes em itens de múltipla escolha, cuja finalidade é meramente classificatória. Na segunda fase, os alunos respondem às situações-problemas de maneira discursiva.

Neste estudo extraiu-se informações pedagógicas dos itens da primeira fase da OBMEP, que pudessem indicar erros frequentes e até dificuldades que o particular grupo de alunos participantes da pesquisa possuem. A prova da primeira fase da OBMEP aplicada em 2017 a uma amostra de 349 alunos da rede municipal de ensino da zona urbana de Santarém, no estado do Pará, se mostrou inapropriada para fins de avaliação das proficiências dos sujeitos considerados, conforme evidenciam os índices da TCT e os parâmetros da TRI. A medida de consistência interna clássica do teste $\alpha_T = 0,18$, aponta a baixa fidedignidade do teste, cujo pressuposto da unidimensionalidade não é satisfeito na análise fatorial paralela. Cabe ressaltar que esses resultados obtidos nesta pesquisas concordam com os resultados alcançados por Silva (2019), Vilarinho (2015), Costa (2015), pois em todos os casos foi verificado baixa consistência interna dos instrumentos.

Por conseguinte, dois modelos da TRI não apresentaram bons ajustes aos dados, pois observou-se parâmetros de discriminação negativos em sete itens oriundos da estimação via ML2 e cinco itens não discriminativos no ML3, sendo que este último produz ajustes mais débeis que o primeiro. Tais resultados chamam atenção para dois supostos problemas, a saber, o da má formulação dos itens e o da falta de habilidade coletiva dos alunos, sobretudo no que diz respeito aos conceitos e procedimentos exigidos em cada um dos itens.

A análise pedagógica realizada apontou tendências, erros e dificuldades que os alunos possuem na resolução de problemas matemáticos acerca de diversos assuntos. Observou-se uma maior dificuldade nas resoluções de problemas que envolvem a aplicação de números racionais. Erros de leitura e interpretação gráfica também foi marcante precisamente no item 5.

No decorrer deste estudo mostrou-se que a extração de informações pedagógicas da prova da OBMEP podem fornecer *feedback* sobre dificuldades e erros que os alunos comentem por não terem adquirido a proficiência exigida para a resolução dos problemas matemáticos propostos.

Referências

AKAIKE, H. A new look at the statistical model identification. In: **Selected Papers of Hirotugu Akaike**. Springer, 1974. p. 215–222. Disponível em: <https://link.springer.com/chapter/10.1007/978-1-4612-1694-0_16>. Acesso em: 25 de junho de 2019.

ANDRADE, D. F.; TAVARES, H. R.; VALLE, R. d. C. **Teoria da Resposta ao Item: conceitos e aplicações**. São Paulo: ABE, 2000.

ANJOS, A. **Teorias de Avalia¸c~ao**. 2013. Disponível em: <https://docs.ufpr.br/~aanjos/CE095/slides2014/TCT.pdf>. Acesso em: 16 de dezembro de 2019.

BATENBURG, T. A. van; LAROS, J. A. Graphical analysis of test items. **Educational Research and Evaluation**, Taylor & Francis, v. 8, n. 3, p. 319–333, 2002.

BIONDI, R. L.; VASCONCELLOS, L.; MENEZES-FILHO, N. A. d. Avaliando o impacto da olimpíada brasileira de matemática das escolas públicas (obmep) no desempenho de matemática nas avaliações educacionais. **São Paulo: Fundação Getúlio Vargas, Escola de Economia de São Paulo**, 2009.

BRASIL. **Cadernos de Estudos Desenvolvimento Social em Debate n 30**. Brasília, DF: Ministério do Desenvolvimento Social; Secretaria de Avaliação e Gestão da Informação., 2018.

BRASIL, P. C. N. matemática. **Brasília: Ministério da Educação/Secretaria de Educação Fundamental**, 1998.

CHALMERS, R. P. et al. mirt: A multidimensional item response theory package for the r environment. **Journal of Statistical Software**, v. 48, n. 6, p. 1–29, 2012.

CONDÉ, F. N. **Análise empírica de itens**. Brasília: INEP, 2001.

CONDé, F. N. **Relação entre características do teste educacional e estimativa de habilidade do estudante. 2008. 151f**. Tese (Doutorado) — Universidade de Brasília, 2008.

COSTA, R. Q. G. **Análise da prova da primeira fase da OBMEP como subsídio para orientar a prática docente**. 2015. Dissertação de Mestrado – Universidade de Brasília.

CRUZEIRO, H. G. C. **Comparação de desempenhos na escola e na OBMEP de estudantes do ensino médio de uma rede de escolas privadas do Distrito Federal**. 2018. Dissertação de Mestrado – Universidade de Brasília.

DUARTE, A. R. S.; GALVÃO, M. E. E. L. Olimpíada paulista de matemática: quase quatro décadas de incentivo ao estudo da matemática. **Revista Brasileira de História da Matemática**, Publicação Oficial da Sociedade Brasileira de História da Matemática, v. 14, n. 29, p. 129–143, 2014.

FLETCHER, P. R. Da teoria clássica dos testes para os modelos de resposta ao item. **Rio de Janeiro: Escola Nacional de Ciências Estatísticas**, 2010.

FONTANIVE, N. S. O uso pedagógico dos testes. In: SOUZA, ALBERTO DE MELLO. **Dimensões da Avaliação Educacional**. Petrópoles, RJ: Vozes, 2005.

GIL, A. C. Como classificar as pesquisas. **Como elaborar projetos de pesquisa**, v. 4, p. 44–45, 2002. Disponível em: <http://www.madani.adv.br/aula/Frederico/GIL.pdf>. Acesso em: 25 de junho de 2019.

GLEN, S. **Kaiser-Meyer-Olkin (KMO) Test for Sampling Adequacy**. 2016. Disponível em: <https://www.statisticshowto.datasciencecentral.com/kaiser-meyer-olkin/>. Acesso em: 01 de outubro de 2019.

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

HAIR, J. F. et al. **Análise multivariada de dados**. Porto Alegre: Bookman Editora, 2009.

IMO. **Olimpíadas Internacional de Matemática. História**. 2015. Disponível em: <https://imof.co/about-imo/history/>. Acesso em: 07 de maio de 2019.

KLEIN, R. Teste de rendimento escolar. In: SOUZA, ALBERTO DE MELLO. **Dimensões da Avaliação Educacional**. Petrópoles, RJ: Vozes, 2005.

KLEIN, R. Alguns aspectos da teoria de resposta ao item relativos à estimação das proficiências. **Ensaio: avaliação e políticas públicas em educação**, SciELO Brasil, v. 21, n. 78, p. 35–56, 2013.

LEIRIãO, L. **Avaliação externa de larga escala, o que é?** 2017. Disponível em: <https://www.tuneduc.com.br/avaliacao-externa-de-larga-escala/>. Acesso em: 16 de dezembro de 2019.

LORD, F. M. Some how and which for practical tailored testing. **Psychometrics for educational debates**, Wiley New York, p. 189–205, 1980.

LORD, F. M.; NOVICK, M. R. **Statistical theories of mental test scores**. New York: IAP, 1968.

MACIEL, M. V. M.; BASSO, M. V. d. A. Olimpíada brasileira de matemática das escolas públicas (obmep): as origens de um projeto de qualificação do ensino de matemática na educação básica. **In: X Encontro Gaúcho de Educação Matemática**, Ijuí, 2009.

MALTA, G. H.; LOPES, S. A. **Resolução de Problemas pelo Métodos Pictórico**. Sociedade Brasileira de Matemática, 2018. Disponível em: <http://twixar.me/BRq1>. Acesso em: 1 de outubro de 2019.

MARANHÃO, T. d. P. Avaliação de impacto da olimpíada brasileira de matemáticas nas escolas públicas (obmep–2005/2009). **Avaliação do impacto da Olimpíada Brasileira de Matemática nas escolas públicas, Série Documentos Técnicos**, v. 11, 2011.

NASCIMENTO, M. M.; RUEDA, F. J. M. Estudo da estrutura interna do teste de inteligência–ti. **Psico-USF**, Universidade São Francisco, v. 19, n. 2, 2014.

OBM. **Olimpíadas Brasileira de matemática. Histórico da OBM.** 2010. Disponível em: <http://www.obm.org.br/quem-somos/historico/>. Acesso em: 23 de fevereiro de 2018.

OBMEP. **Apresentação: Olimpíadas Brasileira de Matemática das Escolas Públicas.** 2018. Disponível em: <http://www.obm.org.br/quem-somos/historico/>. Acesso em: 23 de fevereiro de 2018.

OBMEP. **8ª reportagem da série OBMEP 10 anos homenageia um professor da capital piauiense da matemática**. 2019. Disponível em: <http://www.obmep.org.br/noticias.DO?id=320>. Acesso em: 10 de dezembro de 2019.

OTERO, M. et al. La teoría antropológica de lo didáctico en el aula de matemática. **Buenos Aires: Editorial Dunken**, 2013.

PASQUALI, L. **Psicometria: teoria dos testes na psicologia e na educação.** 3ª. ed. Petrópolis: Vozes, 2017.

QEDU. **Ideb do Município Cocal dos Alves**. 2019. Disponível em: <https://www.qedu.org.br/cidade/4971-cocal-dos-alves/ideb>. Acesso em: 10 de dezembro de 2019.

QUARESMA, E. d. S. **Modelagem para construção de escalas avaliativas e classificatórias em exames seletivos utilizando teoria da resposta ao item uni e multidimensional**. Tese (Doutorado) — Universidade de São Paulo, 2014.

QUIROZ, A. A. Item analysis for multiple choice tests. **R package version**, v. 1, 2017. Disponível em: <https://cran.r-project.org/web/packages/itan/itan.pdf>. Acesso em: 25 de junho de 2019.

RABELO, M. **Avaliação Educacional: Fundamentos, Metodologia e Aplicações no Contexto Brasileiro**. Rio de Janeiro: SBM, 2013.

REVELLE, W. An overview of the psych package. **Department of Psychology Northwestern University**, Citeseer, v. 3, n. 2012, p. 1–25, 2011.

RIZOPOULOS, D. 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.

RODRIGUES, M. M. Proposta de análise de itens das provas do saeb sob a perspectiva pedagógica e a psicométrica. **Estudos em Avaliação Educacional**, v. 17, n. 34, p. 43–78, 2006.

SARTES, L. M. A.; SOUZA-FORMIGONI, M. L. O. d. Avanços na psicometria: da teoria clássica dos testes à teoria de resposta ao item. **Psicologia: Reflexão e Crítica**, Curso de Pós-Graduação em Psicologia da Universidade Federal do Rio Grande do Sul, 2013.

SCHWARZ, G. et al. Estimating the dimension of a model. **The annals of statistics**, Institute of Mathematical Statistics, v. 6, n. 2, p. 461–464, 1978. Disponível em: <https://projecteuclid.org/download/pdf_1/euclid.aos/1176344136>. Acesso em: 25 de junho de 2019.

SILVA, W. L. G. **AVALIAÇÃO EM LARGA ESCALA COMO POLÍTICA DO ESTADO: um estudo comparativo entre a Teoria Clássica dos Testes e a Teoria da Resposta ao Item na Olimpíada Brasileira de Matemática das Escolas Públicas (OBMEP)**. 2019. Dissertação de Mestrado – Universidade Federal do Oeste do Pará.

SOARES, D. J. M. **Teoria clássica dos testes e teoria de resposta ao item aplicadas em uma avaliação de matemática básica**. 2018. Dissertação (Mestrado em Estatística)—Universidade Federal de Viçosa.

TEAM, R. R. C. **RStudio: integrated development for R**. 2018. Disponível em: <http://www.rstudio.com>. Acesso em: 23 de fevereiro de 2018.

VICINI, L.; SOUZA, A. M. Análise multivariada da teoria à prática. **Santa Maria: UFSM, CCNE**, p. 32, 2005.

VILARINHO, A. P. L. **Uma proposta de análise de desempenho dos estudantes e de valorização da primeira fase da OBMEP**. 2015. Dissertação (Mestrado Profissional em Matemática)—Universidade de Brasília.

WILLSE, J. T.; SHU, Z. Ctt: Classical test theory functions. **R package version**, v. 2, 2014.

7 APÊNDICES

APÊNDICE A - TERMO DE CONSENTIMENTO LIVRE E ESCLARECIMENTO

TERMO DE CONSENTIMENTO LIVRE E ESCLARECIDO

Prezado(a) participante:

Sou estudante do curso de licenciatura em Matemática e Física da Universidade Federal do Oeste do Pará. Estou realizando uma pesquisa sob supervisão do professor Dr. Mario Tanaka Filho, cujo objetivo é: apresentar uma análise psicométrica e pedagógica dos itens da prova da primeira fase OBMEP, nível 1, 13ª edição, com base na Teoria Clássica dos Testes e Teoria de Resposta ao Item (TRI).

Sua participação envolve ceder o uso dos gabaritos respondidos pelos estudantes do 6° e 7° ano do ensino fundamental às provas da OBMEP de 2017 e possibilitar a divulgação do resultado da pesquisa e as propostas de ação ao grupo de docentes.

A participação nesse estudo é voluntária e se você decidir não participar ou quiser desistir de continuar em qualquer momento, tem absoluta liberdade de fazê-lo.

Na publicação dos resultados desta pesquisa, sua identidade será mantida no mais rigoroso sigilo. Serão omitidas todas as informações que permitam identificá-lo, identificar a escola e/ou identificar os estudantes em estudo.

Mesmo não tendo benefícios diretos em participar, indiretamente você contribuirá para a compreensão do fenômeno estudado e para a produção de conhecimento científico.

Quaisquer dúvidas relativas à pesquisa poderão ser esclarecidas pelo pesquisador Andrey Camurça da Silva por telefone (93)99157-7256 ou por e-mail andreycamurca@gmail.com.

Atenciosamente
Santarém, Pará, Brasil

_____
Nome e assinatura do(a) estudante

_____
Nome e assinatura do(a) do professor orientador

[ ] Consinto em participar deste estudo e declaro ter recebido uma cópia deste termo de consentimento.

Santarém, Pará, Brasil

_____
Nome e assinatura do(a) participante

APÊNDICE B - *SCRIPT* USADO NO SOFTWARE R

```
install.packages("psych")
install.packages("mirt")
install.packages("CTT")
install.packages("itan")
install.packages("ggplot2")
install.packages("reshape")
install.packages("ggplot")
install.packages("cowplot")
install.packages("lavaan")
library(CTT)
library(ltm)
library(psych)
library(mirt)
library(ggplot2)
library(reshape)
library(itan)
library(ggplot)
library(cowplot)
library(lavaan)

#-----------------------------------------------------------------------
1) Leitura da base de dados
#-----------------------------------------------------------------------
dados <- read.table("D:/OneDrive/TCC/R/dados_na.txt", header=T)
head(dados)

#-----------------------------------------------------------------------
2) Dicotomizando a base de dados
#-----------------------------------------------------------------------
dados1 = as.matrix(dados[-1,])
dados1

gabarito=as.character(as.matrix(dados[1,]))
gabarito

certame <- mult.choice(dados1, gabarito)
```

```
#2.1 omitindo dados com "NA"
dados_sna<- na.omit(certame)
dados_sna

#-------------------------------------------------------------------------
2) AGI teste
#-------------------------------------------------------------------------
plots <- agi(dados1, gabarito, nGrupos = 6, nOpciones = 5)
plots[[1]][[1]]
plots[[1]][[2]]
plots[[1]][[3]]
plots[[1]][[4]]
plots[[1]][[5]]
plots[[1]][[6]]
plots[[1]][[7]]
plots[[1]][[8]]
plots[[1]][[9]]
plots[[1]][[10]]
plots[[1]][[11]]
plots[[1]][[12]]
plots[[1]][[13]]
plots[[1]][[14]]
plots[[1]][[15]]
plots[[1]][[16]]
plots[[1]][[17]]
plots[[1]][[18]]
plots[[1]][[19]]
plots[[1]][[20]]

#-------------------------------------------------------------------------
3) Calculando índices da TCT
#-------------------------------------------------------------------------
certame.desc <- descript(certame)
certame.desc

#-------------------------------------------------------------------------
#3.1) Classificando itens quanto ao poder discriminativo

cpbs<- c()
```

```
for (j in 20:1) {
bicor<-  biserial.cor(rowSums(dados_sna[-1,]), dados_sna
[-1,j],level=2)
cpbs<- rbind(bicor, cpbs)
}
correlacao.pbs <- data.frame(item=1:20, cpbs)
correlacao.pbs

clas.pbs <- data.frame(correlacao.pbs, classificacao = 0)

for (i in 1:20) {
    ifelse(clas.pbs[i,2]<0.2,
        clas.pbs[i,3]<-"Item deficiente, deve ser rejeitado",
            ifelse(clas.pbs[i,2]< 0.3,
            clas.pbs[i,3]<-"Item marginal, sujeito a reelaboração",
                ifelse(clas.pbs[i,2]< 0.4,
                clas.pbs[i,3]<-"Item bom, mas sujeito a aprimoramento",
                    clas.pbs[i,3]<-"Item bom" )))}
clas.pbs

#-------------------------------------------------------------------------
#3.2) Classificando o item de acordo com o índice de dificuldade

D <- c()
for(i in 20:1){
  index <- sum(dados_sna[,i])/nrow(dados_sna)
  D <- rbind(index, D)}

dif <- data.frame(D, Classificação=0)
for (i in 1:20) {
  ifelse(dif[i,1]<0.3,
         dif[i,2]<-"Difícil",
            ifelse(dif[i,1]<= 0.7,
                   dif[i,2]<-"Média dificuldade",
                   dif[i,2]<-"Fácil"))}
dif


#-------------------------------------------------------------------------
#3.3 Calculando a proporção de respostas por alternativa
```

```
prop <- c()
for(i in 20:1) {
  it <- prop.table(table(dados_sna[-1,i]))
  prop<- rbind(it, prop)
}
prop
#-------------------------------------------------------------------------
#3.4 Calculando o escore empírico dos sujeitos
notas <- rowSums(dados_sna)
notas

mednot<-mean(notas)
mednot
desv <- sqrt(var(notas))
cv <- 100*desv/mednot
cv
#-------------------------------------------------------------------------
4) Realizando a análise fatorial exploratória
#-------------------------------------------------------------------------
cortest.bartlett(certame, n=NULL, diag=TRUE)
KMO(certame)
fa.parallel(certame, cor="tet")
factor.plot(fa(certame, cor="tet"), cut=0.3)
fa.diagram(fa(certame,  cor="tet"), cut = 0.3)
irt.fa(certame, plot=T)
fa(certame, cor="tet")

#-------------------------------------------------------------------------
5) Analisando os dados à luz da TRI
#-------------------------------------------------------------------------

mod1<-mirt (certame, 1 , itemtype= "Rasch")
mod2<-mirt (certame, 1 , itemtype ="2PL")
mod3<-mirt (certame, 1 , itemtype ="3PL")

#-------------------------------------------------------------------------
#5.1 Plotando curvas características e de informação dos itens

plot(mod2, type="trace", main=" ")
```

```
itemplot(mod2, 1, col="black", lty=1 , lwd =3, main=" "))

#-------------------------------------------------------------------------
#5.2 Comparando modelos da TRI

anova(mod1, mod2)
anova(mod2, mod3)
#-------------------------------------------------------------------------
6) Plotando outros gráficos
#-------------------------------------------------------------------------

df <- prop.table(table(dados[-1,1]))
barplot(df , col = c(4,8,8,8,8),
        ylim=c(0,1), space=.2,width=c(.1,.1),
        main=c("Item 1"),
        xlab="Anternativas", ylab="Proporção")

histograma<- hist(notas, col="08", border="black", freq = TRUE, xlab="Acertos",
                  ylab="Frequência", main="Histograma")
```

8 ANEXOS

ANEXO A - PROVA DA PRIMEIRA FASE DA OBMEP, NÍVEL 1, 13ª EDIÇÃO

13ª OLIMPÍADA BRASILEIRA DE MATEMÁTICA DAS ESCOLAS PÚBLICAS
OBMEP 2017
Somando novos talentos para o Brasil

Nível 1

6.º e 7.º anos do Ensino Fundamental

1.ª FASE – 6 de junho de 2017

Nome completo do(a) aluno(a): _____

INSTRUÇÕES

1. Preencha o cartão-resposta com seu nome completo, sexo, telefone, endereço eletrônico, data de nascimento, ano e turno em que estuda, e lembre-se de assiná-lo.
2. A duração da prova é de 2 horas e 30 minutos.
3. Cada questão tem cinco alternativas de resposta: A), B), C), D) e E) e **apenas uma** delas é correta.
4. Para cada questão marque a alternativa escolhida no cartão-resposta, preenchendo todo o espaço dentro do círculo correspondente, a lápis ou a caneta esferográfica azul ou preta (é preferível a caneta).

Ⓐ ● Ⓒ Ⓓ Ⓔ

5. Marque apenas uma alternativa para cada questão. **Atenção:** se você marcar mais de uma alternativa, perderá os pontos da questão, mesmo que uma das alternativas marcadas seja correta.
6. Não é permitido o uso de instrumentos de desenho, calculadoras ou quaisquer fontes de consulta.
7. Não é permitido o uso de celulares, *tablets* ou quaisquer outros equipamentos eletrônicos.
8. Os espaços em branco na prova podem ser usados para rascunho.
9. Ao final da prova, entregue-a ao professor junto com o cartão-resposta.

Visite nossas páginas na Internet: www.obmep.org.br www.facebook.com/obmep

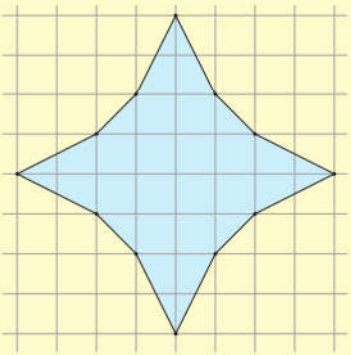
SBM impa MINISTÉRIO DA CIÊNCIA, TECNOLOGIA, INOVAÇÕES E COMUNICAÇÕES MINISTÉRIO DA EDUCAÇÃO GOVERNO FEDERAL

1. Nas balanças da figura, objetos iguais têm pesos iguais. Qual dos objetos é o mais pesado?

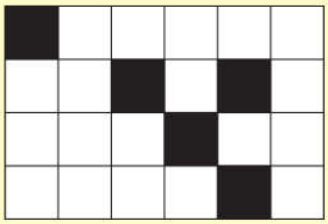
A) 
B) 
C) 
D) 
E) 

2. A área da figura azul é igual à soma das áreas de quantos quadradinhos do quadriculado?

A) 12
B) 22
C) 32
D) 64
E) 100



3. Na figura, quantos quadradinhos brancos ainda devem ser pintados de preto para que o número total de quadradinhos pretos passe a ser o dobro do número de quadradinhos brancos?



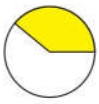
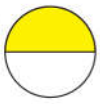
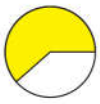
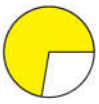
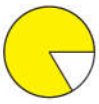
A) 9
B) 10
C) 11
D) 12
E) 13

4. Vânia preencheu os quadradinhos da conta abaixo com os algarismos 1, 2, 3, 4, 5, 6, 7 e 8. Ela usou todos os algarismos e obteve o maior resultado possível. Qual foi esse resultado?

A) 402
B) 609
C) 618
D) 816
E) 876

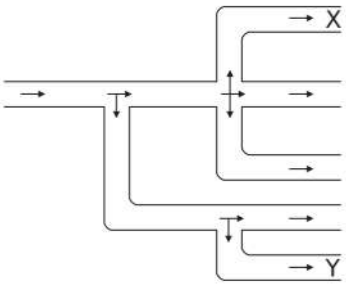
5. O gráfico de barras mostra a distribuição dos alunos de uma escola conforme o tempo diário dedicado à leitura.

Qual é o gráfico de setores que melhor representa, em amarelo, a fração de alunos que dedicam à leitura no máximo 40 minutos por dia?

A) 
B) 
C) 
D) 
E) 

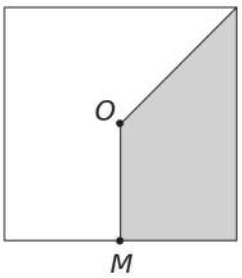
6. Na rede de distribuição de água representada abaixo, a água passa pelos canos como indicado pelas setas e se distribui igualmente em cada ramificação. Em uma hora passaram 200 mil litros de água pela saída X. Quantos litros de água passaram pela saída Y nessa mesma hora?

A) 100 mil litros
B) 130 mil litros
C) 300 mil litros
D) 450 mil litros
E) 600 mil litros

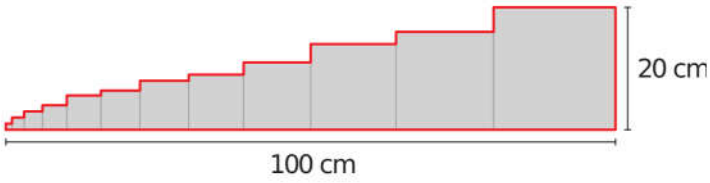


7. A figura mostra um quadrado de centro O e área 20 cm^2. O ponto M é o ponto médio de um dos lados. Qual é a área da região sombreada?

A) 6 cm^2
B) 6,5 cm^2
C) 7 cm^2
D) 7,5 cm^2
E) 8 cm^2



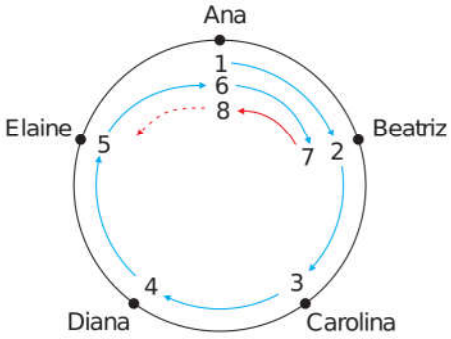
8. Vários quadrados foram dispostos um ao lado do outro, em ordem crescente de tamanho, formando uma figura com 100 cm de base. O lado do maior quadrado mede 20 cm. Qual é o perímetro (medida do contorno em vermelho) da figura formada por esses quadrados?



A) 220 cm
B) 240 cm
C) 260 cm
D) 300 cm
E) 400 cm

9. Ana, Beatriz, Carolina, Diana e Elaine, em roda, brincam de falar números consecutivos. Ana começa falando 1, depois Beatriz fala 2 e assim por diante, conforme ilustrado na figura. Elas iniciam a brincadeira no sentido horário e mudam o sentido toda vez que o número falado for múltiplo de 7. Qual delas vai falar o número 32?

A) Ana
B) Beatriz
C) Carolina
D) Diana
E) Elaine



10. Em uma mesa há nove cartões numerados de 1 a 9. Ana e Beto pegaram três cartões cada um. A soma dos números dos cartões de Ana é 7 e a soma dos números dos cartões de Beto é 23. Qual é a diferença entre o maior e o menor dos números dos três cartões deixados sobre a mesa?

A) 3
B) 4
C) 5
D) 6
E) 7

11. Para obter tinta de cor laranja, devem-se misturar 3 partes de tinta vermelha com 2 partes de tinta amarela. Para obter tinta de cor verde, devem-se misturar 2 partes de tinta azul com 1 parte de tinta amarela. Para obter tinta de cor marrom, deve-se misturar a mesma quantidade de tintas laranja e verde.

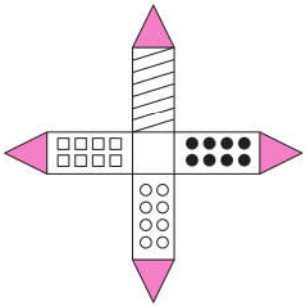
Quantos litros de tinta amarela são necessários para obter 30 litros de tinta marrom?

A) 7
B) 8
C) 9
D) 10
E) 11

12. Uma roda-gigante está parada com o banco 8 na posição mais baixa e o banco 3 na posição mais alta. Seus bancos estão igualmente espaçados e numerados em ordem a partir do número 1. Quantos bancos tem essa roda-gigante?

A) 8
B) 10
C) 12
D) 14
E) 16

13. Em um dos lados de uma folha de papel grosso, Pedro desenhou a figura ao lado. Depois, recortou-a e montou uma torre em miniatura. Das cinco imagens abaixo, quais podem representar a torre montada por Pedro?



A) Imagens 1, 3 e 5
B) Imagens 1, 4 e 5
C) Imagens 1, 2 e 3
D) Imagens 2, 3 e 4
E) Imagens 3, 4 e 5

14. Mônica e seu namorado foram assistir a uma peça de teatro. O auditório era organizado em fileiras paralelas ao palco, todas com o mesmo número de cadeiras dispostas lado a lado. Eles se sentaram um ao lado do outro nos dois últimos lugares vagos. Mônica percebeu que havia, no total, 14 pessoas nas fileiras à sua frente e 21 pessoas nas fileiras atrás da sua. Quantas cadeiras havia no auditório?

A) 37
B) 38
C) 40
D) 42
E) 49

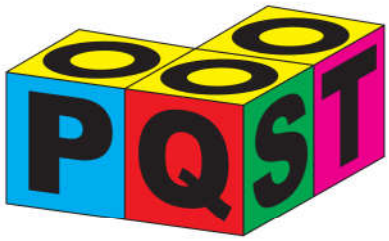
15. Na conta armada, cada letra representa um algarismo, e letras diferentes representam algarismos diferentes. Qual é o algarismo que a letra T representa?

A) 0
B) 1
C) 3
D) 5
E) 7

$$\begin{array}{r} GOTA \\ GOTA \\ GOTA \\ GOTA \\ +GOTA \\ \hline AGUA \end{array}$$

16. Zequinha tem três dados iguais, com letras **O**, **P**, **Q**, **R**, **S** e **T** em suas faces. Ele juntou esses dados como na figura, de modo que as faces em contato tivessem a mesma letra. Qual é a letra na face oposta à que tem a letra **T**?

A) **S**
B) **R**
C) **Q**
D) **P**
E) **O**



17. Após digitar um número de seis algarismos em sua calculadora, Cecília observou que dois algarismos 9 que ela havia digitado não apareceram no visor; o que apareceu foi 2017. Quantas são as possibilidades para o número que ela digitou?

A) 6
B) 9
C) 10
D) 15
E) 18

18. Uma escola fez uma pesquisa com todos os alunos do sexto ano para verificar se eles gostavam de banana, maçã ou laranja. Cada aluno assinalou pelo menos uma dessas três frutas. A tabela abaixo apresenta os resultados da pesquisa.

	6º A	6º B	6º C
Banana	20	15	14
Maçã	12	20	12
Laranja	18	5	10

Por exemplo, 20 alunos do 6º A assinalaram que gostam de banana. Quantos alunos há, no mínimo e no máximo, no sexto ano dessa escola?

A) No mínimo 54 e no máximo 126 alunos.
B) No mínimo 54 e no máximo 58 alunos.
C) No mínimo 27 e no máximo 54 alunos.
D) No mínimo 27 e no máximo 126 alunos.
E) No mínimo 31 e no máximo 58 alunos.

19. Em uma competição, as partidas têm duração de 60 minutos, e cada time tem sempre 5 jogadores em campo. Em determinada partida, um time inscreveu 8 atletas e foram feitas várias substituições de modo que cada um deles jogou a mesma quantidade de tempo. Quanto tempo cada um deles jogou nessa partida?

A) 27 minutos e 30 segundos
B) 30 minutos
C) 37 minutos e 30 segundos
D) 40 minutos
E) 42 minutos e 30 segundos

20. Uma caixa contém 10 bolas verdes, 10 bolas amarelas, 10 bolas azuis e 10 bolas vermelhas. Joãozinho quer retirar uma certa quantidade de bolas dessa caixa, sem olhar, para ter a certeza de que, entre elas, haja um grupo de sete bolas com três cores diferentes, sendo três bolas de uma cor, duas bolas de uma segunda cor e duas bolas de uma terceira cor. Qual é o número mínimo de bolas que Joãozinho deve retirar da caixa?

A) 11
B) 14
C) 21
D) 22
E) 23