Comparação da classificação na prova da OBMEP por meio da Teoria de Resposta ao Item (TRI) e da Teoria Clássica de Testes (TCT)

ALEX MOREIRA[I]
CRISTINA HENRIQUES NOGUEIRA[II]
http://dx.doi.org/10.22347/2175-2753v12i34.2151

Resumo

Este estudo tem por objetivo investigar se a Teoria de Resposta ao Item (TRI) pode ser aplicada nas provas da Olimpíada Brasileira de Matemática (OBMEP) e, assim, comparar com o modelo tradicional de Teoria Clássica de Testes (TCT). Para isso, utilizou-se uma amostra composta por 350 estudantes que cursavam o Ensino Médio no Instituto Federal de Educação, Ciência e Tecnologia do Sudeste de Minas Gerais (IFSudesteMG), *campus* Rio Pomba. Avaliou-se as questões contidas na prova da primeira da etapa da OBMEP de 2017, referente ao nível três. A análise dessas questões foi realizada utilizando o modelo de TRI, considerando o modelo com três parâmetros, cujas estimativas foram obtidas com o auxílio do programa computacional ICL (*IRT Command Language*). Com isso, conclui-se que o modelo de TRI se mostrou aplicável às provas da OBMEP, apresentando divergências nas classificações obtidas pela TCT, além de proporcionar maior capacidade de discernimento nesta classificação.
Palavras-chave: IF Sudeste MG. Avaliação. Modelos Logísticos. Matemática.

Submetido em: 24/01/2019
Aprovado em: 06/03/2020

[I] Instituto Federal de Educação Ciência e Tecnologia Sudeste de Minas Gerais (IFSudeste), Rio Pomba (MG), Brasil; https://orcid.org/0000-0001-6441-151X; e-mail: alex.moreira@educacao.mg.gov.br.
[II] Instituto Federal de Educação Ciência e Tecnologia Sudeste de Minas Gerais (IFSudeste), Rio Pomba (MG), Brasil; https://orcid.org/0000-0001-6401-1532; e-mail: cristina.nogueira@ifsudestemg.edu.br.

Comparison of the Classification in the OBMEP Test through Item Response Theory (IRT) and Classical Test Theory (TCT)

Abstract

This study aims to investigate whether the Item Response Theory (IRT) can be applied to the tests of the Brazilian Mathematical Olympiad (OBMEP) and, thus, compare with the traditional model of Classical Test Theory (TCT). For this, a sample composed of 350 students who attended high school at the Federal Institute of Education, Science and Technology of Southeastern Minas Gerais (IFSudesteMG), Rio Pomba *campus* was used. The questions contained in the first stage of the 2017 OBMEP were evaluated, referring to level three. The analysis of these questions was performed using the IRT model, considering the model with three parameters, whose estimates were obtained with the aid of the computer program ICL (IRT Command Language). Thus, it is concluded that the IRT model proved to be applicable to the OBMEP tests, showing divergences in the classifications obtained by TCT, in addition to providing greater capacity for discernment in this classification.
Keywords: IF Sudeste MG. Evaluation. Logistic Models. Mathematics.

Comparación de la clasificación en la prueba OBMEP usando la Teoría de Respuesta al Ítem (TRI) y de la Teoría Clásica de Pruebas (TCP)

Resumen

Este estudio tiene como objetivo investigar si la Teoría de Respuesta al Ítem (TRI) se puede aplicar a las pruebas de la Olimpiada Brasileña de Matemática (OBMEP) y, por lo tanto, comparar con el modelo tradicional de Teoría Clásica de Pruebas (TCP). Para esto, se utilizó una muestra compuesta por 350 estudiantes que asistían a la Escuela Secundaria en el Instituto Federal de Educación, Ciencia y Tecnología del Sureste de Minas Gerais (IF Sudeste MG), *campus* de Rio Pomba. Se evaluaron las preguntas contenidas en la prueba de la primera etapa de la OBMEP 2017, con respecto al nivel tres. El análisis de estas preguntas se realizó utilizando el modelo de TRI, considerando el modelo con tres parámetros, cuyas estimaciones se obtuvieron con la ayuda del programa informático ICL (*IRT Command Language*). Por lo tanto, se concluye que el modelo TRI demostró ser aplicable a las pruebas de OBMEP, mostrando divergencias en las clasificaciones obtenidas por TCT, además de proporcionar una mayor capacidad de discernimiento en esta clasificación.
Palabras clave: IF Sudeste MG. Evaluación. Modelos logísticos. Matemáticas.

Introdução

De modo geral, as avaliações são instrumentos frequentemente utilizados para mensuração da aprendizagem. Entretanto, Coelho (2014) ressalta que assinalar uma opção em uma prova pode ser um ato que revele plena consciência do assunto indicado no enunciado, mas, também, pode ser apenas mais um item assinalado na expectativa de não se deixar escapar alguma possível chance de acerto e, neste caso, configura o que se denomina acerto casual.

Em específico, a avaliação é utilizada na Olimpíada Brasileira de Matemática das Escolas Públicas (OBMEP), que se trata de uma competição criada em 2005 pelo Instituto de Matemática Pura e Aplicada (IMPA), cujo objetivo é estimular o estudo de matemática e revelar talentos na área. Dentre os seus objetivos específicos, a competição almeja contribuir para a melhoria da qualidade da educação básica, possibilitando que um maior número de alunos brasileiros possa ter acesso a material didático de qualidade, com videoaulas e aplicativos (INSTITUTO DE MATEMÁTICA PURA E APLICADA, 2017).

Os métodos de avaliação adotados nesta competição fundamentam-se na Teoria Clássica de Testes (TCT), logo, a classificação dos candidatos é baseada, apenas, na quantidade de acertos. Por este método, para o critério de pontuação, não há distinção entre as questões que compõem a avaliação, sendo desconsideradas informações como nível de dificuldade de cada questão.

Além disso, outro fato que merece destaque é que, ao utilizar a TCT, podem ocorrer empates entre os candidatos, ou seja, diversos candidatos podem obter a mesma pontuação na avaliação, o que ocasionaria dificuldades na classificação desses candidatos. Devido a isso, surge a necessidade de métodos mais refinados, capazes de diferenciar e selecionar os candidatos que possuem maior habilidade em uma determinada área do conhecimento.

Nessa perspectiva, tem-se observado o surgimento e aplicação de outros métodos avaliativos em processos seletivos, cujos itens são compostos por alternativas de múltiplas escolhas. Dentre esses, destaca-se a Teoria de Resposta ao Item (TRI), cuja metodologia computa a nota da prova de um participante por sua coerência em responder as questões, entendendo que a aquisição do conhecimento é feita de forma acumulativa. Nesse sentido, considera-se improvável que um candidato tenha habilidade para responder corretamente uma questão difícil, uma vez que o mesmo não ocorra com uma questão fácil. Assim, quando o candidato acerta uma questão

mais complexa, mas erra uma questão mais simples, a TRI considera que houve uma incoerência nas respostas, concluindo em um provável acerto casual e, com isso, na falta de habilidade do candidato para responder tal questão.

Para que seja feito esse discernimento, uma importante ferramenta na TRI é a Curva Característica do Item (CCI) que, de acordo com Couto e Prime (2011), trata-se de uma curva obtida para cada item referente a sua probabilidade de acerto em função da habilidade do indivíduo, de modo que, quanto maior a habilidade do indivíduo, maior será a probabilidade de que ele responda corretamente o item.

Para a caracterização de cada item podem ser utilizados três parâmetros, sendo eles: o parâmetro de discriminação, de dificuldade e de acerto casual. De maneira prática, o parâmetro dificuldade representa a habilidade mínima para que o indivíduo, com conhecimento necessário, possa responder de forma correta o item avaliado (PASQUALI; PRIMI, 2003). Já como parâmetro de discriminação, entende-se o quanto o item pode apresentar uma característica discriminativa, isto é, a clareza do nível de conhecimento apropriado que o indivíduo necessita para responder corretamente ao item (ANDRADE; LAROS; GOUVEIA, 2010). Por sua vez, o acerto casual expressa a probabilidade de que um indivíduo com baixa habilidade respondera corretamente ao item considerado.

Com isso, Klein (2009) afirma que a utilização desses parâmetros no modelo faz com que as questões deixem de ser avaliadas separadamente (individualmente), passando a ser avaliadas conjuntamente.

Entretanto, vale ressaltar que, de acordo com Bechger, Maris, Verstralen e Béguin (2003), embora a TRI apresente vantagens em relação à TCT, ela pode ser utilizada, ainda, em complementação à TCT, a fim de oferecer informações adicionais sobre a qualidade do teste.

Diante do exposto, o objetivo deste trabalho é investigar a aplicabilidade da metodologia de TRI aos resultados da primeira fase da Olimpíada Brasileira de Matemática das Escolas Públicas, ano de 2017, considerando os candidatos do Instituto Federal de Educação, Ciência e Tecnologia do Sudeste de Minas Gerais, *campus* Rio Pomba, além de fazer um comparativo da classificação dos estudantes mediante escores obtidos pelos modelos TRI e TCT.

A Teoria Clássica de Testes

Segundo Arantes (2016), a TCT originou-se pelo modelo de escore verdadeiro e de erro exposto por Charles Edward Spearman, em 1904, no qual argumentou que os "resultados dos testes são imperfeitos por natureza e, portanto, a correlação

observada entre duas medidas falhas seria com certeza inferior à correlação realmente existente".

Nesse sentido, a TCT avalia as aptidões pela soma dos itens, podendo apresentar um erro contido nesta soma por ser uma operação empírica, conforme explicita Pasquali e Primi (2003). Sendo assim, a TCT tem como objetivo a interpretação da resposta final, ou seja, o que a soma dos itens corretos descreve sobre as habilidades dos estudantes.

Ainda de acordo com Pasquali e Primi (2003), com relação à dificuldade do item, a TCT a define em termos de porcentagens de acertos, de forma que, quanto mais próximo de 100% a taxa de acertos, mais fácil é considerado o item.

Segundo Arantes (2016), a TCT é falha como teoria estatística, uma vez que não aborda a significância do item na composição do teste, além da escassez de regras sobre as quais possa ser baseada a decisão de incluir ou excluir um determinado item no teste.

A Teoria de Resposta ao Item

De acordo Soares (2014), a TRI surgiu na década de 1950, mas só foi consolidada a partir de 1970, com o desenvolvimento de computadores e programas capazes de realizar a análise de dados mais complexos.

Simão (2016) afirma que somente em 1995, após a retomada do Sistema de Avaliação do Rendimento Escolar do Estado de São Paulo (Saresp), do Governo de São Paulo e do Sistema de Avaliação da Educação Básica (Saeb) do Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira/Ministério da Educação (Inep/MEC), do Governo Federal, a TRI começou a ter maior desenvolvimento em pesquisas e aplicações no Brasil.

As primeiras aplicações da TRI no Brasil, são referenciadas ao Programa de Avaliação da Secretaria de Estado de Educação do Estado de São Paulo em 1993, mas com limitações na sua aplicação (SIMÃO, 2016).

Segundo Arantes (2016), na TRI, o conceito a ser avaliado ocupa um lugar central. Em muitas aplicações ele é concebido como uma variável que não é diretamente observável, por isso usa-se o termo variável latente. As variáveis observáveis, ou seja, as respostas dadas aos itens, são consideradas os indicadores da variável latente.

Basicamente, a TRI é formada por um conjunto de métodos estatísticos que, por meio de processos de estimação da característica de cada item aplicado e da aptidão do indivíduo avaliado em um teste, descreve o desempenho do indivíduo. Assim, esse método é capaz de quantificar, em uma escala métrica, a habilidade desenvolvida do indivíduo e sua competência, sendo essas também conhecidas como o seu escore.

De acordo com a TRI, uma resposta correta a um item do teste é considerada um sinal ou um indicador de maior habilidade do que uma incorreta. Nesse sentido, Verhelst (2010 apud ARANTES, 2016) afirma que:

> Os modelos da TRI se apoiam em um conjunto de pressupostos e formulam a "função do indicador" de uma resposta ao item de maneira probabilística: eles expressam a probabilidade de uma resposta correta para um item como uma função matemática da habilidade básica. Esta função é específica para cada item, e a dependência do item é expressa através da utilização de um ou mais parâmetros por item. Essas funções são chamadas de funções de Resposta ao Item (FRI) e, geralmente, pertencem à mesma família (p. 154).

Cordeiro (2014) complementa este conceito expondo que a TRI possui a propriedade de invariância de parâmetros, ou seja, compara os "resultados obtidos em provas distintas, aplicadas em grupos distintos de alunos".

Ao comparar essa metodologia com a TCT, Pasquali e Primi (2003) esclarecem que esta última tem como objetivo a interpretação da resposta final, ou seja, o que a soma dos itens diz sobre o aluno, enquanto a TRI tem o propósito de medir a habilidade do sujeito de acordo com as respostas dadas a cada item. Assim sendo, enquanto a TCT analisa o resultado final, a TRI analisa partes e probabilidades que geram o resultado final.

Para a obtenção das Funções de Resposta ao Item (FRI), as quais geram as CCI, são considerados modelos logísticos, compostos por um, dois ou três parâmetros. De acordo com Birnbaum (1968), o modelo logístico de três parâmetros é dado por:

$$P\left(U_{ij} = 1 | a, b, c, \theta_j\right) = c_i + (1 - c_i)\frac{1}{1 + e^{-Da_i(\theta_j - b_i)}} \ ,$$

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

I a quantidade total de itens;

n o número total de indivíduos analisados;

U_{ij} é uma variável dicotômica que assume os valores 1, quando o indivíduo j responde corretamente o item i, e 0 caso contrário;

θ_j representa a habilidade (traço latente) do *j*-ésimo indivíduo, o qual pode assumir qualquer valor real;

a_i é o parâmetro do item que representa a discriminação, de modo que esse assume apenas valores positivos da reta real;

b_i é o parâmetro do item que representa a dificuldade, podendo assumir qualquer valor real, porém, frequentemente assume valores compreendidos no intervalo entre -4 e + 4;

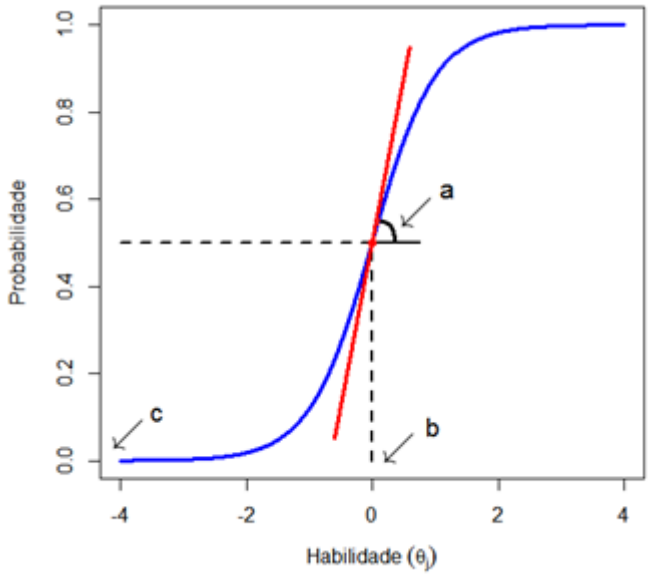
c_i refere-se o parâmetro do item que representa de acerto casual, ou seja, a probabilidade de que um indivíduo com baixa habilidade responder corretamente ao item i. Por se tratar de uma probabilidade, esse parâmetro deve assumir valores reais pertencentes ao intervalo de 0 a 1;

D é um fator de escala. De acordo com Pasquali (2007), o valor de 1,7 deve ser utilizado quando se deseja que a função logística forneça resultados semelhantes ao da ogiva normal;

$P(U_{ij} = 1|\theta_j)$ representa a probabilidade de um indivíduo *j*, com habilidade θ_j, responder corretamente ao item i.

A Figura 1 apresenta o comportamento dessa curva com a interpretação dos parâmetros: discriminação (a), dificuldade (b) e acerto casual (c).

Figura 1 – Curva característica do Item com a especificação dos três parâmetros



Fonte: O autor (2020).

Materiais e métodos

Este estudo trata-se de uma pesquisa quantitativa do tipo exploratória, na qual avaliou-se os itens (questões) da prova referente à primeira fase da OBMEP, considerando-se o ano de 2017. Utilizou-se como fonte de coleta de dados as provas dos estudantes matriculados no Instituto Federal de Educação, Ciência e Tecnologia do Sudeste de Minas Gerais (IFSudesteMG), *campus* Rio Pomba, localizado na cidade de Rio Pomba, Minas Gerais. Por ser ofertado apenas o Ensino Médio na rede de ensino avaliada, foram considerados apenas alunos que realizaram o nível três da OBMEP.

A amostra englobou um total de 350 alunos, distribuídos conforme especificado na Tabela 1. A primeira etapa do nível 3 da prova da OBMEP de 2017, contemplou 20 questões de múltipla escolha, as quais podem ser encontradas no site da competição (http://www.obmep.org.br/provas.htm) (INSTITUTO DE MATEMÁTICA PURA E APLICADA, 2017). Essas questões abordavam temas pertencentes às áreas de geometria, álgebra e estatística. A realização da prova ocorreu no dia 6 de junho de 2017, no próprio *campus* da instituição de ensino.

Tabela 1 – Distribuição dos alunos do Ensino Médio do IF Sudeste-MG, *campus* Rio Pomba, quanto à série cursada, que realizaram a prova da OBMEP em 2017

Série	Turma	Número de alunos
1° Ano	A	21
	B	23
	D	19
	E	30
	G	13
	Z	14
2° Ano	A	22
	B	13
	C	13
	D	18
	E	26
	F	19
	M	20
	Z	18
3° Ano	A	19
	B	08
	D	19
	C/E	35
Total	18	350

Fonte: O autor (2020).

A análise das questões foi realizada pelo ajuste de modelos de TRI, sendo os ajustes desses parâmetros realizados com o auxílio do programa computacional ICL (*IRT Command Language*) (HANSON; BEGUIN, 2002), com interface programada pelo

software R (THE R CORE TEAM, 2016), por meio dos pacotes *ltm* e *irtoys*, para estimação e calibração dos itens e do traço latente a partir da máxima verossimilhança marginal. Conforme Mendonça (2012), o programa ICL tem capacidade de estimar os modelos logísticos para itens dicotômicos, considerando um, dois ou três parâmetros.

Conforme Hanson (1998), para a estimação dos parâmetros foram utilizados estimadores de máxima verossimilhanças, os quais fundamentaram-se no algoritmo EM (*Expectation-Maximization*). Esse estimador possui propriedades assintóticas ótimas, contribuindo para estimação de parâmetros modelos através da escolha de parâmetros desconhecidos, as médias e variâncias, e por fim, considerando o vetor de observações pelas probabilidades iniciais. Para computar os parâmetros do modelo, cujo processo fundamenta-se na esperança e maximização da função, a qual representa a proporção de pessoas em cada classe de habilidade, a partir da discretização da variável latente.

O comportamento do ICL foi analisado frente aos seguintes fatores: o número de itens avaliados, o número de indivíduos examinados, a variação dos valores do parâmetro de discriminação dos itens, além da variação dos valores de dificuldade dos itens. Ao considerar a metodologia de TRI, foram abordados apenas modelos logísticos dicotômicos, considerando uma população apenas um traço latente, que se caracterizam como os modelos unidimensionais.

Os escores obtidos pelo ajuste dos modelos representaram as habilidades dos indivíduos, sendo calculados com base em uma distribuição normal de média 0 e variância 1. Assim, nessa escala, a habilidade de um aluno estaria num escore padrão dos resultados em torno de -4 e +4, pois os pacotes do programa R consideram 100% para todos os sujeitos da população (PEDROZA; CAVALCANTE, 2014).

Entretanto, por dificuldade de interpretação prática, devido à associação com a escala de habilidade medida pela TCT, é comum utilizar outras escalas para a apresentação dos resultados. Dessa forma, buscando maior facilidade na interpretação, foi utilizado uma escala de modo que os valores estivessem compreendidos dentro do intervalo de 0 a 1.000. Para isso, foi adotado média $\mu = 500$ e um desvio-padrão $\sigma = 100$ conforme a escala de proficiência adotado pelo Enem, onde convencionou-se utilizar estes valores a partir do exame de 2009, o primeiro a utilizar este valor. Assim, a habilidade do *j*-ésimo indivíduo (y_j) foi calculada por:

$$y_j = 500 + 100\theta_j, \qquad \text{para } i = 1, 2, ..., 350.$$

Conforme metodologia utilizada por Simão (2016), criou-se uma escala de dificuldade, por meio das estimativas do parâmetro (bi). Desse modo, conforme apresentado na Tabela 2, os itens da prova foram classificados em quatro níveis, definidos a partir dos percentis de 25% (Q_1), 50% (Q_2) e 75% (Q_3) dos valores estimados para esse parâmetro.

Tabela 2 – Classificação dos itens avaliados de acordo com o parâmetro de dificuldade

Classificação	b_i
Muito fácil	Abaixo de 25%
Fácil	25% a 50%
Médio	50% a 75%
Difícil	Acima de 75%

Legenda: b_i = percentil de valores estimados para o grau de dificuldade.
Fonte: Simão (2016).

Já a análise das questões segundo o parâmetro de discriminação, estimado para todas as questões da prova, foi feita com base na classificação proposta por Silva (1992). De acordo com este autor, tal classificação pode ser realizada conforme apresentado na Tabela 3.

Tabela 3 – Classificação dos itens em função do índice de discriminação

Faixas do Índice de Discriminação	
0,40 e acima	Itens muito bons.
0,30 a 0,39	Razoavelmente bons, mas possivelmente sujeitos a melhoramento
0,20 a 0,29	Itens marginais, usualmente necessitando e estando sujeitos a melhoramento.
0,19 e abaixo	Itens deficientes. Devem ser rejeitados ou melhorados por revisão.

Fonte: Silva (1992).

Para a comparação da metodologia da TRI com a TCT, foi calculada a habilidade dos indivíduos por meio dessa última. Assim, essa habilidade (γ) foi obtida pela porcentagem de acertos do indivíduo *j*, ou seja:

$$\gamma_j = \frac{n_j}{N} \times 100$$

em que:

n_j representa o número de acertos do *j*-ésimo indivíduo;

N é número total de questões sendo, neste caso, N = 20;

γ_j é a habilidade estimada do indivíduo *j*, por meio da TCT, em uma escala de 0 a 1.000.

Finalmente, a classificação dos indivíduos foi dada em ordem decrescente com relação à sua habilidade por meio da TRI, comparando as divergências dessa classificação com a classificação obtida por meio da TCT.

Resultados e discussões

Ao analisar descritivamente os resultados obtidos pelos estudantes que realizaram a prova da OBMEP de 2017, de acordo com a TCT, observa-se que 74% dos estudantes acertaram, no máximo, seis questões. Além disso, não houve nenhum resultado superior a 13 acertos, enquanto apenas 6% dos candidatos acertaram entre oito e 13 questões. É possível notar ainda que menos de 1% dos estudantes acertaram mais da metade das questões contidas na prova.

Tabela 4 – Distribuição de frequências de estudantes do IFSudesteMG com relação à quantidade de acertos na prova da OBMEP de 2017

Classe de acertos	Quantidade de estudantes	Porcentagem de estudantes
0 ⊢ 2	20	5,71
2 ⊢ 4	100	28,57
4 ⊢ 6	139	39,71
6 ⊢ 8	70	20,00
8 ⊢ 10	18	5,14
10 ⊢ 12	2	0,57
12 ⊢ 14	1	0,29

Fonte: O autor (2020).

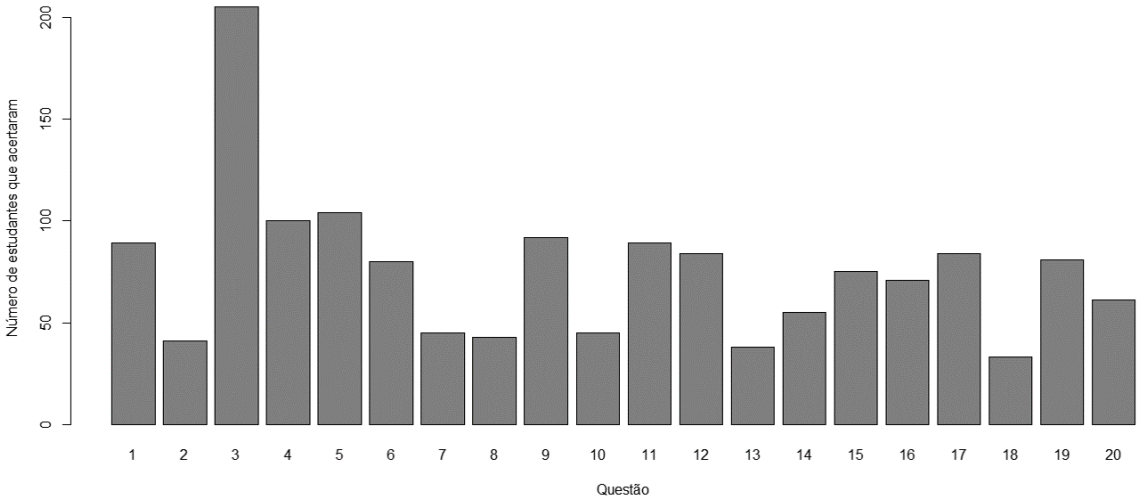
Esses percentuais revelam um cenário preocupante, tornando-se indispensável a identificação de suas causas, para que, em próximas edições, seja possível oferecer suporte para os conteúdos de habilidades específicas, buscando incentivar a participação discente e ressaltar a importância dessa avaliação no desenvolvimento acadêmico discente

Ao avaliar descritivamente os dados, encontra-se uma média de acertos 4,33 questões por aluno, enquanto a mediana foi quatro questões. Além disso, o número mínimo de acertos foi de 0 e o máximo foi de 13, sendo esse o maior número de acertos. Já a variância calculada foi de 3,91, de modo que, com isso, o desvio-padrão foi de 1,98, apontando que grande parte dos acertos estão dentro do intervalo 4,33 ± 1,98.

Analisando individualmente cada questão, conforme apresentado na Figura 2, nota-se que a questão número 3 apresentou o maior número acertos, pois aproximadamente 200 alunos responderam corretamente essa questão. Uma investigação sobre essa questão mostrou que o conteúdo abordado era relacionado à geometria plana, exigindo do estudante apenas conhecimentos sobre figuras geométricas para respondê-la corretamente.

Já as questões 2, 7, 8, 10, 13 e 18 foram as que apresentaram menor índice de acerto (Figura 1), sendo que aproximadamente 30 alunos acertaram cada uma das questões citadas. Uma análise sobre os conteúdos abordados nestas questões revelou que a questão número 2 exigia conhecimentos sobre produtos notáveis, a questão número 7 necessitava do domínio de funções do 2º grau, enquanto as questões 8 e 10 abrangiam conteúdos sobre circunferências e medida dos arcos. Já a questão 13 contemplava conceitos de relações trigonométricas e áreas de figuras planas, tópicos esses que frequentemente são apontados por diversos autores como geradores de grande dificuldade pelos discentes. Por fim, a questão 18 estava relacionada com conteúdo de estruturas algébricas que, por sua vez, trata-se de uma área pouco explorada pelas instituições de ensino básico.

Figura 2 – Frequência absoluta do número de acertos em cada questão da OBMEP de 2017



Fonte: O autor (2020).

Após a análise exploratória dos dados, procedeu-se a análise das questões por meio da TRI, que propõe o ajuste de modelos considerando três parâmetros para avaliar as

características de um candidato, que não podem ser observadas diretamente. Este modelo apresenta em sua composição os parâmetros discriminação, dificuldade e acerto casual, cujas estimativas para cada questão pertencente à prova da OBMEP de 2017, encontram-se na Tabela 5.

Ao analisar os resultados contidos na Tabela 5, referente ao parâmetro discriminação, o qual avalia a capacidade da questão em diferenciar indivíduos com diferentes habilidades, as questões 1, 2, 4 e 8 apresentaram maior poder discriminativo.

De acordo com a classificação de Silva (1992), todas as questões foram classificadas como itens muito bons, uma vez que as estimativas do parâmetro de discriminação foram superiores a 0,40 para todas as questões avaliadas. Esse resultado converge com as intenções das provas de OBMEP, a qual fundamenta-se em uma criteriosa busca por estudantes que possuem habilidades em conhecimentos matemáticos, sendo composta por questões capazes de discernir os estudantes com maiores habilidades.

Tabela 5 – Estimativas para o modelo de TRI com três parâmetros, referente a cada uma das questões da prova OBMEP de 2017

Item	Discriminação (a_i)	Dificuldade (b_i)	Acerto casual (c_i)
1	**1,6209**	1,3945	0,1041
2	**2,2429**	1,7925	**0,0448**
3	1,0006	**-0,0660**	0,1511
4	**1,3384**	1,3218	0,1024
5	0,7787	12,2330	**0,2900**
6	0,7472	4,1756	0,1837
7	0,8239	5,3136	0,1161
8	**1,1419**	3,1968	**0,0868**
9	0,7472	5,2654	0,2396
10	0,8318	6,2278	0,1240
11	0,6943	3,9458	0,1934
12	0,7788	**17,8659**	0,2360
13	0,8423	5,7438	0,1021
14	0,7931	5,0279	0,1375
15	0,7764	4,4212	0,1798
16	0,7789	**15,3474**	0,2008
17	0,7788	**16,0527**	0,2360
18	0,7789	**17,7099**	0,0981
19	0,7845	7,1614	0,2242
20	0,7788	**17,8816**	0,1738

Fonte: O autor (2020).

Ao considerar o parâmetro de acerto casual, uma vez que cada questão tem cinco alternativas sendo apenas uma correta, espera-se que a probabilidade de acerto casual esteja em torno de 0,20. Por meio da Tabela 5, observa-se que as questões 2 e 8

foram as que apresentaram as menores estimativas para esse parâmetro, o que reflete que elas possuem uma pequena possibilidade de acerto casual. É interessante notar que essas questões também apresentaram alto poder de discriminação, enfatizando a classificação desses itens como muito bons. Em contrapartida, a questão número 5 apresentou uma estimativa muito superior a 0,20, sugerindo que as opções de alternativas dessa questão pudessem contribuir para a escolha da alternativa correta e, por isso, ineficiente para mensurar a verdadeira habilidade do estudante.

Também na Tabela 5, estão expressas as estimativas para o parâmetro dificuldade, por meio das quais a questão pode ser classificada de acordo com seu nível de dificuldade. Esses resultados comparados à classificação proposta na Tabela 6 permitem inferir que as questões 12, 16, 17, 18, e 20 foram consideradas questões difíceis.

Em contrapartida, a questão 3 foi classificada como muito fácil, apresentando a menor estimativa para o parâmetro dificuldade, comungando com os resultados da TCT, que também apontou esta como a questão mais fácil devido ao maior número de acertos.

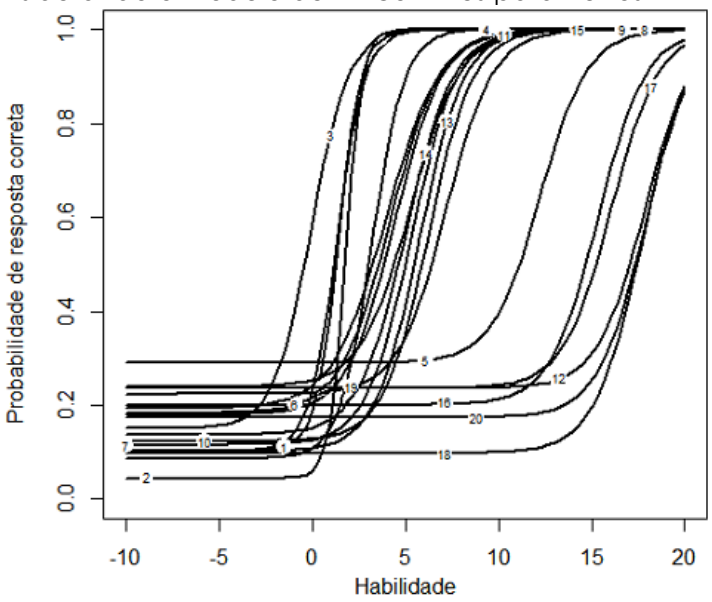
Tabela 6 – Classificação dos itens avaliados de acordo com o parâmetro de dificuldade, considerando o modelo de TRI com três parâmetros

Classificação	b_i
Muito fácil	$b < 3{,}7586$
Fácil	$3{,}7586 \leq b < 5{,}2896$
Médio	$5{,}2896 \leq b < 13{,}0116$
Difícil	$b \geq 13{,}0116$

Fonte: O autor (2020).

Esses resultados corroboram com o comportamento da CCI para cada uma das questões, conforme esboça a Figura 3. De acordo as CCI apresentadas neste gráfico, pode-se observar as questões 12, 16, 17, 18 e 20 como as que exigem maior habilidade do estudante, já que essas exigem maiores valores de habilidades para que a probabilidade de acerto se aproxime de 1. Em contrapartida, a curva da questão 3 continua localizada mais esquerda do gráfico, levando à conclusão de que esse é o item com o menor nível de dificuldade. Nota-se, também, que a questão 5 se destaca como o item como o item com a maior probabilidade de acerto casual. Já a curva da questão 2 destaca-se, além de apresentar a menor probabilidade de acerto casual, como a de maior poder discriminativo, uma vez que mostrou a maior inclinação da reta tangente à curva no seu ponto de inflexão.

Figura 3 – Curva característica do item para cada uma das questões da OBMEP de 2017, considerando o modelo de TRI com três parâmetros



Fonte: O autor (2020).

Ao utilizar as estimativas do modelo ajustado considerando três parâmetros, foram obtidas as estimativas da habilidade de cada estudante, as quais estão expostas na Tabela 7, juntamente com a ordem decrescente de classificação dos estudantes, de acordo com a mesma. Além disso, esta tabela contém, ainda, a habilidade calculada por meio da TCT e a classificação fundamentada nesta teoria, a qual torna possível a comparação entre os resultados obtidos pela TRI e pela TCT, observando divergências em suas classificações.

Tabela 7 – Estimativa da habilidade apresentada por cada estudante, segundo a TCT e TRI, considerando o modelo com três parâmetros

Estudante	Classificação TRI	Habilidade TRI	Classificação TCT	Habilidade TCT
204	1	902,4282	1	650
288	2	831,9180	22	350
95	3	823,1734	22	350
94	4	790,5166	22	350
211	5	784,1203	50	300
93	6	781,8693	4	450
115	7	774,0822	22	350
320	8	769,6986	10	400

Continua

Conclusão

92	9	763,0493	93	250
97	10	761,4681	50	300
96	11	754,4808	93	250

98	12	752,7547	50	300
100	13	752,1924	22	350
291	14	744,7463	4	450
91	15	741,8058	148	200
(...)	(...)	(...)	(...)	(...)
350	346	385,4537	232	150
290	347	385,4534	331	50
229	348	385,4532	348	0
230	348	385,4532	348	0
312	348	385,4532	348	0

Fonte: O autor (2020).

A partir desses resultados, verifica-se a diferença existente na classificação apresentada pelo modelo de TRI com três parâmetros e o modelo de TCT. Como exemplo desse fato, observa-se o estudante 93 e o 291: ambos acertaram exatamente nove questões e, com isso, com mesma classificação pela TCT, foram classificados, respectivamente, em 6ª e 14ª posição pela TRI. Além disso, os estudantes 211 e 93, por exemplo: enquanto estes estudantes ocupavam, respectivamente, a 50ª e a 4ª posição na classificação pela TCT, ao considerar a TRI, obtiveram como classificação, respectivamente, as posições 5 e 6. Tal situação pode ser justificada pelo fato de que estudantes tenham acertado questões com maior nível de dificuldade e errado as questões consideradas fáceis, cenário visto como incoerente pela TRI e, por isso, passível de acertos mediante casualidade, o que acarreta diminuição na mensuração da habilidade do estudante por meio do modelo de TRI.

Outro fato relevante refere-se aos estudantes 229, 230 e 312, os quais obtiveram a mesma classificação em ambos modelos. Deve-se salientar que, mesmo não acertando nenhuma questão, as notas desses estudantes pela TRI diferem de zero, cuja justificativa se dá pelo fato dos modelos de TRI adotarem uma metodologia comparativa entre as notas dos estudantes, de modo que essas foram distribuídas segundo uma distribuição normal com média 500 e desvio-padrão 100.

Conclusões

Diante do exposto, conclui-se que o modelo de TRI, o qual considera os parâmetros discriminação, dificuldade e acerto casual, mostrou-se um instrumento propício para mensurar a habilidade dos estudantes do IFSudesteMG, *campus* Rio Pomba, que realizaram as provas da OBMEP.

Ao comparar o modelo de TRI com o modelo tradicional de TCT, foi possível notar divergências em suas classificações. Nesse sentido, a TRI não considerou apenas o

número de acertos do estudante e a dificuldade inerente a cada questão, mas principalmente levou em consideração a coerência individual dos alunos. Isto significa que, ao contrário da TCT, um estudante que acertou um maior número de questões fáceis e que errou questões consideradas difíceis pode ter classificação superior a alguém que acertou o maior número de questões difíceis, porém errou questões fáceis.

A utilização da TRI possibilitou, ainda, realizar um estudo acerca da qualidade das questões que compunham a prova, destacando questões com maior poder discriminatório, questões com baixa eficiência devido à alta probabilidade de acerto casual e questões que apresentaram maior nível de dificuldade, o que torna possível diagnosticar os principais conteúdos com aprendizagem deficiente e, com isso, buscar alternativas para saná-la.

Por fim, a TRI mostrou-se superior à TCT com relação à capacidade de discernimento dos estudantes, uma vez que, enquanto vários estudantes foram classificados com mesma habilidade pela TCT, a TRI foi capaz diferenciá-los, evitando empates na classificação dos estudantes.

Referências

ANDRADE, J. M.; LAROS, J. A.; GOUVEIA, V. V. O uso da teoria de resposta ao item em avaliações educacionais: diretrizes para pesquisadores. *Aval. psicol.*, Porto Alegre, v. 9, n. 3, p. 421-435, 2010.

ARANTES, L. J. *Avaliando a aprendizagem do conceito de energia no ensino médio usando a TRI*. 2016. 156 f. Dissertação (Mestrado Profissional em Ensino de Física) – Faculdade de Física, Universidade Federal de Lavras, Lavras, 2016.

BECHGER, T. M.; MARIS, G.; VERSTRALEN, H. H. F. M.; BÉGUIN, A. A. Using classical test theory in combination with item response theory. *Applied Psychological Measurement*, [s. l.], v. 27, n. 5, p. 319-334, 2003.

BIRNBAUM, A. Some latent trait models and their use in inferring an examinee's ability. In: LORD, F. M.; NOVICK, M. R. (Ed.). *Statistical theories of mental test scores*. Oxford: Reading, 1968. p. 397-479.

COELHO, E. C *Teoria da resposta ao item*: desafios e perspectivas em exames multidisciplinares. 2014. 213 f. Tese (Doutorado em Ciências) - Programa de Pós-Graduação em Métodos Numéricos em Engenharia, Universidade Federal do Paraná, Paraná, 2014.

CORDEIRO, L. Sobre a inadequação da metodologia de cálculo das notas do SISU. *Educação & Sociedade*, [s. l.], v. 35, n. 126, p. 293-320, 2014.

COUTO, G.; PRIMI, R. Teoria de resposta ao item (TRI): conceitos elementares dos modelos para itens dicotômicos. *Boletim de Psicologia*, [s. l.] v. 23, n. 51, p. 1-15, 2011.

HANSON, B. A. Uniform DIF and DIF defined by differences in item response functions. *Journal of Educational and Behavioral Education*, [s. l.], v. 23, p. 244-253, 1998.

HANSON, B. A.; BEGUIN, A. A. Obtaining a common scale for item response theory item parameters using separate versus concurrent estimation in the common-item equating design. *Appl Psychol Meas*, [s. l.], v. 26, n. 1, p. 3-24, 2002.

INSTITUTO DE MATEMÁTICA PURA E APLICADA. 13ª Olimpíada Brasileira de Matemática das Escolas Públicas. *Portal*. Rio de Janeiro: OBMEP, 2017. Disponível em: http://www.obmep.org.br/. Acesso: 27 dez. 2018.

KLEIN, R. Utilização da teoria de resposta ao item do Sistema Nacional de Avaliação da Educação Básica (SAEB). *Revista Meta:* Avaliação, Rio de Janeiro, v. 1, n. 2, p. 125 – 140, maio/ago. 2009.

MENDONÇA, J. D. An*álise da eficiência de estimação de parâmetros da TRI pelo software ICL*. 2012. 125 p. Dissertação (Mestrado em Estatística e Experimentação Agropecuária) Universidade Federal de Lavras, Lavras, 2012.

PASQUALI, L.; PRIMI, R. Fundamentos da teoria da resposta ao item: TRI. *Avaliação Psicológica*, [s. l.], v. 2, n. 2, p. 99-110, 2003.

PASQUALI, L. *Teoria de resposta ao item*: teoria, procedimentos e aplicações. Brasília: LabPAM/UnB, 2007.

PEDROZA, J. K. B. R.; CAVALCANTE, P. R. N. IFRS para PMEs: uma Investigação quanto ao nível de compreensão de contadores amparada na Teoria de Resposta ao Item. *In*: CONGRESSO USP DE CONTROLADORIA E CONTABILIDADE, 14., 2014, São Paulo. *Anais*... São Paulo, 2014.

THE R CORE TEAM. *R*: a language and environment for statistical computing. Vienna: R Foundation for Statistical Computing, 2016. Disponível em: https://www.r-project.org/. Acesso em: 27 dez. 2018.

SILVA, C. S. *Medidas e avaliação em educação*. São Paulo: Editora Vozes, 1992.

SIMÃO, I. *TRI aplicada a análise de itens para o Ensino de Estatística*. 2016. 50 f. Dissertação (Mestrado em Estatística e Experimentação Agropecuária) - Universidade Federal de Lavras, Lavras, 2016.

SOARES, M. S. *Proposta de um software de banco de itens calibrados pela teoria da resposta ao item (TRI), para uso de professores de matemática da educação básica*. 2014. 127 f. Dissertação (Mestrado em Ciências) – Universidade Federal do Acre, Rio Branco, 2014.

VERHELST, N. Methodological advances in educational effectiveness research: using item response theory to measure outcomes and factors: an overview of item response theory models. In: CREEMERS, B P. M.; KYRIAKIDES, L.; SAMMONS, P. *Methodological advances in educational effectiveness research*. New York: Routledge, 2010. v. 1, p. 153-183.