Análise Clássica de Testes: Uma proposta de análise de desempenho dos estudantes na primeira fase da OBMEP

Test Classical Analysis: A performance analysis of the proposal of the students in the first phase of OBMEP

Raíra Elberhardt Nogueira Knüpfer [*]
Aruana do Amaral[*]
Elisa Henning[*]

Resumo

Neste artigo buscamos comparar o desempenho de duas turmas do sexto ano do Ensino Fundamental II na primeira fase da Olimpíada Brasileira de Matemática das Escolas Públicas (OBMEP), em uma escola municipal na região de Joinville, Santa Catarina, por meio da quantidade de acertos na prova da olimpíada. A prova aplicada foi a mesma para esse nível, portanto o objetivo era verificar se os alunos diferenciavam, por turma, na quantidade de acertos. Ainda, efetuamos uma análise da prova por meio da Teoria Clássica dos Testes (TCT) que visa avaliar as questões da olimpíada, seguindo os níveis em grau de dificuldade (muito fácil, fácil, moderado, difícil e muito difícil) e, também, busca avaliar os itens no quesito qualidade de acordo com os acertos e erros dos alunos em cada questão.

Palavras-chave: Teoria Clássica dos Testes. OBMEP. Educação Matemática.

Linha Temática: Educação Matemática

1 Introdução

De acordo com Vilarinho (2015), a utilização das avaliações nas últimas décadas levou a um aumento na preocupação com a metodologia utilizada para obtenção dos resultados. Além disso, a autora ressalta a importância quanto à qualidade e validade dos instrumentos usados para elaboração de itens que

[*] Mestranda do Programa de Pós-Graduação em Ensino de Ciências, Matemática e Tecnologias (PPGECMT), rairaelb@gmail.com
[*] Mestranda do Programa de Pós-Graduação em Ensino de Ciências, Matemática e Tecnologias (PPGECMT), aruana.amaral@gmail.com
[*] Professora Doutora do Programa de Pós-Graduação em Ensino de Ciências, Matemática e Tecnologias (PPGECMT), elisa.henning@udesc.br

cumpram com a finalidade educacional avaliativa, no sentido de buscar informações acerca de métodos e estatísticas adotadas, para que seja possível analisar e interpretar, a partir do desempenho dos estudantes, ações de cunho pedagógico, visando melhorar e aprimorar o ensino aprendizado dos mesmos.

Diante disso, Borgatto e Andrade (2012), apontam que a avaliação do desempenho dos estudantes, depende, fundamentalmente, da qualidade dos itens da prova. Sendo assim, em uma avaliação educacional, a análise pela Teoria Clássica dos Testes (TCT), chamada também de análise clássica, constituem métodos estatísticos que podem contribuir para esta finalidade.

A TCT baseia-se em seus parâmetros descritivos, auxiliando na interpretação da distribuição das respostas de cada alternativa, isto é, considera a prova como um todo e seus resultados são expressos no número total ou no percentual dos itens respondidos corretamente. Sendo assim, as propriedades psicométricas dos itens de uma prova relacionam-se aos seguintes parâmetros: índice de dificuldade; índice de discriminação e correlação bisserial (BORGATTO e ANDRADE, 2012).

No índice de dificuldade, calcula-se a proporção de acertos, ou seja, a razão entre o número de estudantes que responderam o item corretamente e o número total de estudantes submetidos ao item. O índice varia de 0 a 1, em que o extremo inferir indica que ninguém acertou e o extremo superior que todos acertaram. Quanto menor a porcentagem de acerto, maior será o grau de dificuldade (VILARINHO, 2015).

Vilarinho (2015), com base em Pasquali (2003), destaca que, para que uma avaliação educacional tenha um nível de dificuldade ideal, é indicada uma distribuição de níveis de dificuldade de itens no teste dentro de uma curva normal. A autora apresenta a seguinte tabela para classificação e percentual esperado para os índices de dificuldade na TCT.

Tabela 01: Classificação e percentual esperado para os índeces de dificuldade na TCT

Quantitativo ideal de itens na avaliação (% esperado)	Índice de dificuldade do item	Classificação do item em relação ao índice de dificuldade
10%	Superior a 0,9	Muito fáceis
20%	De 0,7 a 0,9	Fáceis
40%	De 0,3 a 0,7	Medianos
20%	De 0,1 a 0,3	Difíceis
10%	Até 0,1	Muito difíceis

Fonte: Vilarinho, 2015, p. 27.

Já o índice de discriminação, analisa, para determinado item, as porcentagens de acertos dos grupos de estudantes com melhor e pior desempenho. Para o cálculo desse, os participantes são divididos em três grupos: o grupo superior (27% dos participantes com maiores pontuações), o grupo inferior (27% dos participantes com menores pontuações) e os demais 46% dos participantes, que compõem o grupo intermediário. Esse parâmetro corresponde à diferença entre o percentual de acerto do primeiro e do segundo grupo. Almeja-se que, a porcentagem de acerto seja maior para o grupo com melhor desempenho e, quanto maior for a diferença entre as porcentagens de acertos dos dois grupos, maior será a discriminação do item (VILARINHO, 2015).

De acordo com a autora, por meio de Rabelo (2013), é esperado que em uma avaliação educacional, o poder de discriminação do item seja superior a 40, conforme a tabela:

Tabela 02: Poder de discriminação

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*

Fonte: Vilarinho, 2015, p. 28.

Quanto ao coeficiente bisserial, é uma medida de associação entre o desempenho do indivíduo no item e o desempenho na prova como um todo, ou seja, com seu escore bruto. Ele estima a correlação entre a variável de desempenho no teste e uma variável latente (não observável) com distribuição normal que, por hipótese, representa a habilidade que determina o acerto ou erro do item (BORGATTO e ANDRADE, 2012).

Para o cálculo do coeficiente bisserial, de acordo com Borgatto e Andrade (2012), utiliza-se a seguinte fórmula:

$$r_{bis} = \frac{M+ - M-}{S} \cdot \frac{p(1-p)}{h(p)}.$$

Sendo que, de acordo com Borgatto e Andrade (2012), classifica-se:

- M+: média da medida de desempenho para os alunos que acertaram o item;
- M^{-}: média da medida de desempenho no teste para os alunos que erraram o item;
- S*:* desvio-padrão da medida de desempenho no teste para todos os alunos;

- p*:* percentual de respostas e;
- h(p)*:* valor da densidade da distribuição normal com média 0 e variância 1 no ponto em que a área da curva à esquerda deste ponto é igual a p .

Espera-se que a opção correta do item tenha correlação positiva, e que as opções erradas, chamadas de distratores, tenham correlação negativa. Desta forma, é possível dizer que alunos com melhores desempenhos no teste como um todo estão acertando o item (VILARINHO, 2015).

Quanto a Olimpíada Brasileira de Matemática das Escolas Públicas (OBMEP), é realizada pelo Instituto Nacional de Matemática Pura e Aplicada (IMPA), que de acordo com o portal OBMEP, tem como meta, incentivar o estudo da matemática, visando descobrir talentos na área. Cabe ressaltar que as provas são realizadas anualmente em dia específico designado pela organização e, em geral, as escolas públicas têm grande interesse em participar.

Existem três níveis de dificuldade nas provas, os quais se intensificam gradativamente. O Nível 1 é aplicado para sextos e sétimos anos do Ensino Fundamental II, Nível 2 para oitavos e nonos anos do Ensino Fundamental II e Nível 3 para o Ensino Médio. Ainda, a olimpíada é composta de duas fases, sendo a primeira uma prova objetiva com vinte questões e a segunda uma prova discursiva com seis questões.

Para a escola, o resultado das provas pode servir como indicador do conhecimento matemático de seus alunos, podendo também ajudar a criar estratégias para o ensino da disciplina, buscando melhores resultados.

Para os professores de matemática, as provas contribuem para o estímulo dos estudantes que sentem afinidade com a disciplina, uma vez que as questões da prova (OBMEP) são tidas como desafios para os alunos. Ainda, o professor consegue trabalhar com essas questões em sala, podendo também conquistar aqueles alunos que ainda não se sentem confiantes com a Matemática.

Desse modo, a presente pesquisa busca fazer uma análise qualitativa e quantitativa das respostas de estudantes na prova da primeira fase da OBMEP, aplicada em duas turmas do sexto ano de uma escola do Ensino Fundamental II, localizada na cidade de Joinville, Santa Catarina. Vale ressaltar que a prova aplicada foi a mesma para ambas as turmas e, pretendemos verificar se os alunos da primeira turma obtiveram um melhor desempenho em comparação aos alunos da segunda turma. O que se deseja é analisar o desempenho desses estudantes, utilizando os fundamentos da TCT para tentar compreender em que medida a prova pode servir como base de instrumento para que se possa mensurar o ensino aprendizado na disciplina de Matemática aos estudantes e docentes nas escolas.

2 Método

O enquadramento metodológico utilizado neste trabalho baseia-se em Henning, Ramos e Konrath (2013) e Vilarinho (2015). Nesta pesquisa foram utilizados os dados referentes às avaliações da OBMEP de 46 alunos de duas turmas do sexto ano (6º C e 6º D) do Ensino Fundamental II de uma escola pública na cidade de Joinville. A prova foi composta por 20 (vinte) questões, abrangendo conteúdos da disciplina de Matemática para o nível de ensino em questão.

Para a análise da TCT foram consideradas as respostas dos alunos como sendo, 1 (um) o acerto, e 0 (zero) o erro. Para as análises foi utilizado o R (R Core Team, 2016), com auxílio dos pacotes ltm (Latent Traint Model) (Rizopoulos, 2006) e psychometric.

3 Resultados e discussões

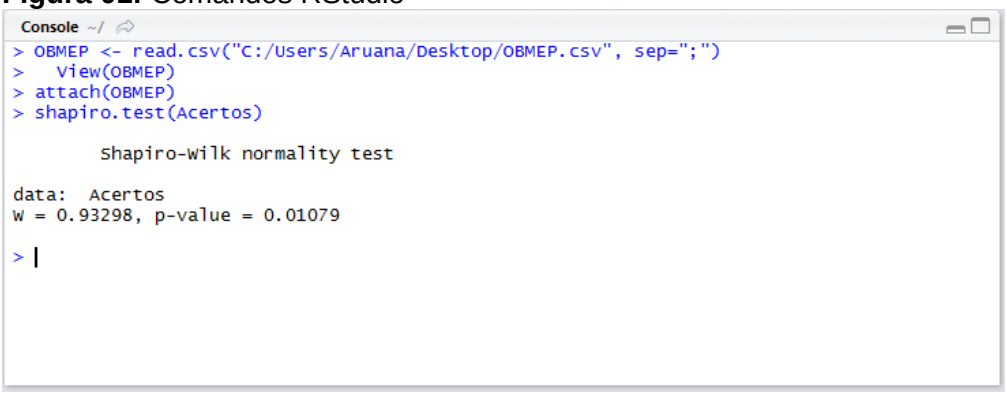
Inicialmente, aplicamos o teste de normalidade Shapiro-Wilk para verificar que tipo de teste seria utilizado, paramétrico ou não-paramétrico. Sendo as hipóteses nula e alternativa:

H_0: As amostras possuem uma distribuição normal;

H_1: As amostras não possuem uma distribuição normal.

Logo:

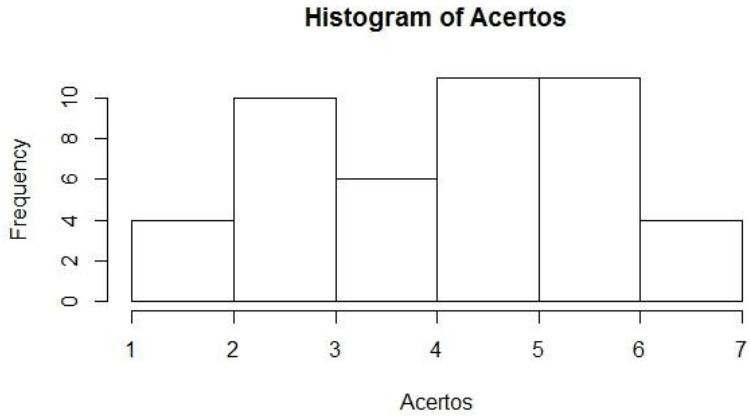
Figura 01: Comandos RStudio



Fonte: Dados das autoras, 2016.

Após, verificamos que a distribuição dos dados não era normal, pois para um nível de significância de 5%, tivemos p-valor = 0,01079 < α. Logo, devemos rejeitar a hipótese inicial de que as amostras possuem distribuição normal. Analisando o Histograma também foi possível verificar a não normalidade de distribuição dos dados, como é possível observar a seguir:

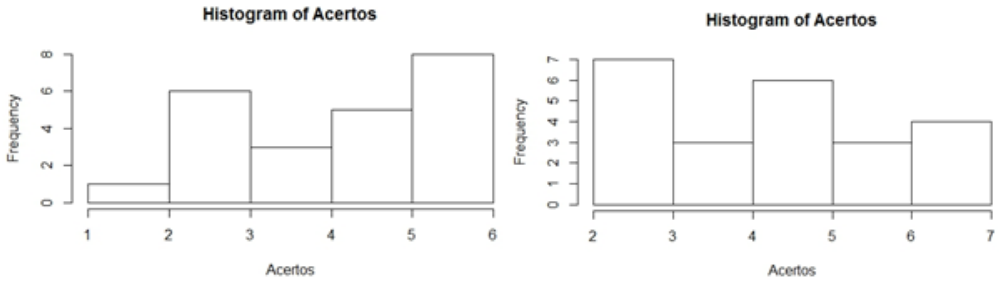
Figura 02: Histograma



Fonte: Dados das autoras, 2016.

Ainda, analisando separadamente por turma, também pudemos perceber a não normalidade dos dados, como é possível observar nos histogramas a seguir:

Figura 03: Histograma por turma



Fonte: Dados das autoras, 2016.

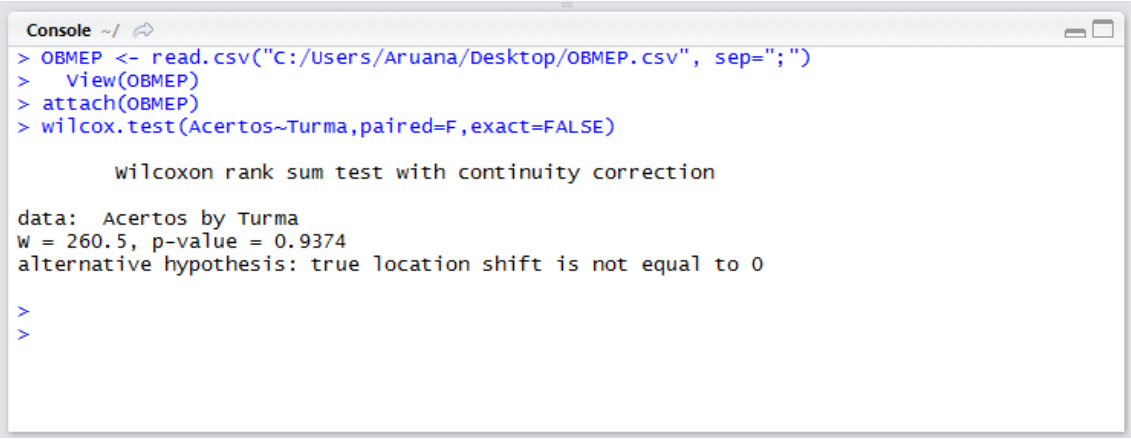
Portanto, concluímos que o teste deveria ser não-paramétrico. Ainda, precisávamos analisar se as amostras eram pareadas ou independentes para então decidir o teste ideal. Como a pesquisa consiste em comparar duas turmas diferentes, concordamos que seriam amostras independentes. Dessa forma, o teste utilizado deveria ser Mann-Whitney e foi o que aplicamos. Sendo as hipóteses nula e alternativa:

H_0: As turmas não diferenciam a quantidade de acertos;

H_1: As turmas diferenciam a quantidade de acertos.

Assim:

Figura 04: Comandos RStudio



Fonte: Dados das autoras, 2016.

Por intermédio da análise dos resultados, temos que, para um nível de significância de 5%, ou seja, para α = 0,05, devemos aceitar a hipótese nula (H_0), uma vez que p-valor = 0,9374 > α. Portanto, concluímos que as turmas não diferenciam a quantidade de acertos, ou seja, não há diferença entre o número de acertos de alunos do 6º ano C e do 6º ano D.

Posteriormente, foi calculada a proporção de acertos, que corresponde à proporção de participantes que responderam ao item corretamente; a proporção de erros; o índice de discriminação por item e a correlação bisserial. Os resultados da análise clássica podem ser visualizados na Tabela 03 e os escores totais na Tabela 04.

Tabela 03: Proporção de acertos, erros e correlação bisserial

Questões	Proporção de acertos	Proporção de erros	Índice de Discriminação	Correlação Bisserial
01	0.1304	0.8696	0.06666667	0.1090
02	0.5652	0.4348	0.13333333	0.1222
03	0.5435	0.4565	0.46666667	0.4485
04	0.2174	0.7826	0.00000000	-0.0564
05	0.0870	0.9130	0.26666667	0.4366
06	0.1739	0.8261	0.20000000	0.2035
07	0.0652	0.9348	0.06666667	0.1314
08	0.0870	0.9130	0.13333333	0.1369
09	0.1957	0.8043	0.33333333	0.3163
10	0.0435	0.9565	0.13333333	0.2671
11	0.2391	0.7609	-0.06666667	-0.0072
12	0.4783	0.5217	0.26666667	0.2696
13	0.0870	0.9130	0.06666667	0.0869
14	0.2609	0.7391	0.13333333	0.1352
15	0.2391	0.7609	0.46666667	0.3559
16	0.3913	0.6087	0.20000000	0.2257
17	0.1957	0.8043	0.33333333	0.3163
18	0.1957	0.8043	0.00000000	-0.0031
19	0.3043	0.6957	0.20000000	0.1862
20	0.0652	0.9348	0.13333333	0.3024

Fonte: Dados das autoras, 2016.

Tabela 04: Escores brutos

Escores totais	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Frequência	0	1	3	10	6	11	11	4	0	0	0	0	0	0	0	0	0	0	0	0	0

Fonte: Dados das autoras, 2016.

Analisando os dados da Tabela 03, a questão identificada como mais difícil foi a questão 10, com apenas 4,35% de acertos, e a mais fácil foi a questão 02, com 56,52% de acertos.

Também foi possível verificar que, a prova não apresentou questões com todas as variações de graus de dificuldade, indo de encontro com a proposta de Vilarinho (2015) com base em Vianna (1989), que sugere que as provas contenham itens que alcancem todo o *continuum* da escala, ou seja, devem conter questões fáceis, medianas e difíceis.

Percebemos também que apenas as questões 02; 03; 12; 16 e 19 apresentaram-se com dificuldade moderada, entretanto, a maioria das questões pode ser classificada como difícil ou muito difícil, conforme pode ser analisado pela Tabela 05 a seguir.

Tabela 05: Distribuição das questões em relação ao parâmetro dificuldade

Classificação	**Valores**	**Questões**	**Percentual de questões na prova**
Muito fácil	0,9 ou mais	Nenhum	0%
Fácil	0,7 ≤ valor ≤ 0,9	Nenhum	0%
Moderado	Entre 0,3 e 0,7	02; 03; 12; 16; 19	25%
Difícil	Entre 0,1 e 0,3	01; 04; 06; 09; 11; 14; 15; 17; 18	45%
Muito difícil	Até 0,1	05; 07; 08; 10; 13; 20	30%

Fonte: Dados das autoras, 2016.

Pela Tabela 04, podemos visualizar que, para as vinte questões o escore mínimo foi um, com um aluno acertando apenas uma questão, e o máximo foi sete, com quatro alunos acertando sete questões.

Na TCT, a partir da correlação bisserial, podemos verificar que o item que aparece como o mais discriminante é a questão 03. Este item apresenta a maior diferença entre a proporção de acertos dos 27% de alunos com notas mais altas e

a proporção de acertos dos 27% de alunos com notas mais baixas, com valor de 0.46666667.

Ainda em relação à discriminação dos itens, a prova apresentou onze questões consideradas deficientes, cinco questões consideradas marginais, sujeitas a reelaboração, duas questões boas, mas sujeitas a aprimoramento e duas questões consideradas boas. Os itens que menos discriminaram, em ordem crescente são os itens 11; 04; 18; 01; 07; 13; 02; 08; 10; 14 e 20. Já os itens que tiveram o maior poder de discriminação foram os itens 03 e 15, conforme pode ser verificado na Tabela 06.

Tabela 06: Distribuição das questões em relação à discriminação

Classificação	Valores	Questões	Percentual de questões na prova
Item deficiente	Até 0,2	01; 02; 04; 07; 08; 10; 11; 13; 14; 18; 20	55%
Item marginal	0,2 ≤ Disc < 0,3	05; 06; 12; 16; 19	25%
Item bom, sujeito a aprimoramento	0,3 ≤ Disc < 0,4	09; 17	10%
Item bom	Disc ≥ 0,4	03; 15	10%

Fonte: Dados das autoras, 2016.

4 Conclusão e Considerações finais

De acordo com as nossas análises, auxiliadas pelo software RStudio, concluímos que não houve diferença na quantidade de acertos, por turma. Portanto, pudemos concluir que os dois sextos anos avaliados possuem o mesmo nível de desempenho na OBMEP. Ainda, observamos que a prova não apresentou questões nos graus de dificuldade muito fácil e fácil, uma vez que os escores propostos não foram atingidos. Por fim, constatamos que a prova possuía

onze questões deficientes, fato esse que surpreende, visto que a prova da olimpíada possuía vinte questões.

Referências

BORGATTO, Adriano Ferreti e ANDRADE, Dalton Francisco de. **Análise Clássica de Testes com diferentes graus de dificuldade**. Estudos em Avaliação Educacional, São Paulo, 23 (52), 146-156. 2012. Disponível em < www.fcc.org.br/pesquisa/publicacoes/eae/arquivos/1733/1733.pdf>. Acesso em: 21 jun. 2016.

HENNING, Elisa. RAMOS, Marcelo Sávio e KONRATH, Andréa Cristina. **Análise de itens de uma prova de raciocínio probabilístico**. In: Actas del VII CIBEM. Montevideo, Uruguai. p. 2025-2032. 2013. Disponível em < www.cibem7.semur.edu.uy/7/actas/pdfs/1180.pdf>. Acesso em: 01 jul. 2016.

PORTAL OBMEP. Disponível em <http://www.obmep.org.br>. Acesso em: 21 jun. 2016.

VILARINHO, Ana Paula Lima. **Uma proposta de análise de desempenho dos estudantes e de valorização da primeira fase da OBMEP**. Dissertação (Mestrado Profissional em Matemática). Universidade de Brasília, Brasília, 2015.

TEAM, R Core. **A Language and Environment for Statistical Computing.** R Foundation for Statistical Computing. Vienna, Austria. 2016. Disponível em: https://www.R-project.org/.