



O que realmente está sendo avaliado quando avaliamos?

Rafael Filipe Novôa **Vaz**

Instituto Federal de Educação, Ciência e Tecnologia do Rio de Janeiro
Brasil

rafael.vaz@ifrj.edu.br

Daniel de Oliveira **Lima**

GPAM, PEMAT, Universidade Federal do Rio de Janeiro
Brasil

danielprof2006@gmail.com

Paula Monteiro **Baptista**

GPAM, PEMAT, Universidade Federal do Rio de Janeiro
Brasil

paulamonteirob@yahoo.com.br

Carlos Augusto **Aguilar Júnior**

Universidade Federal Fluminense
Brasil

carlosaugustobolivar@hotmail.com

Resumo

Esta comunicação traz um ensaio sobre avaliação, apresentando as conformidades e divergências ponderadas em uma multicorreção realizada com professores que ensinam matemática e pedagogos. Esta proposta se justifica tendo como objetivo o estudo sobre fatores que influenciam as discrepâncias de notas atribuídas por avaliadores. Para embasar a pesquisa, apresentam-se reflexões sobre confiabilidade e validade – conceitos fomentados pela docimologia – além da análise de dados de uma investigação. Tais dados foram produzidos durante um minicurso, a partir da proposta de correção de cinco soluções diferentes de uma única questão, feita por 11 avaliadores em dois momentos diferentes. A análise dos dados possibilitou a observação de que no segundo momento as correções estavam mais refinadas quanto à habilidade trabalhada e ao erro cometido, justificando a importância do estudo pelos avaliadores sobre o ato de avaliar.

Palavras-chave: Avaliação em Matemática; Multicorreção; Validade dos exames.

A pesquisa em multitorreção

A prova é o principal instrumento avaliativo utilizado por professores que ensinam matemática. As principais características das provas são: individuais, escritas, sem consulta, com tempo delimitado, esporádicas, intermitentes e breves, possibilitam a ausência de convivência como exigência para avaliar, oferecem tratamento genérico a todos os alunos com a pretensão de serem neutras. Tal instrumento, que está enraizado em nossa cultura educacional, foi concebido a partir de uma filosofia positivista que atesta um caráter imparcial, neutro e justo à avaliação escolar (Morgan, 2000; Fernandes, 2009, Guba & Lincoln, 2011). As concepções que estão presentes nos três pressupostos básicos que sustentam o pensamento geral na avaliação são:

primeiro, presume-se que os indivíduos possuem atributos (como conhecimento, compreensão, habilidade etc.) que são detectáveis e mensuráveis. Em segundo lugar, o objetivo principal da avaliação é descobrir e medir esses atributos. Finalmente, o processo de avaliação e seus resultados são considerados fundamentalmente benignos ou mesmo benéficos (embora efeitos colaterais infelizes e tentativas de melhorá-los possam ser reconhecidos) (Morgan, 2000, p. 225-226)

Precursos dos estudos psicométricos na avaliação escolar, Noizet e Caverni (1985) desenvolveram um vasto estudo sobre multitorreção envolvendo diversas disciplinas, que culminou na publicação do livro *Psychologie de l'évaluation scolaire* em 1978, traduzido em 1985 para a língua portuguesa. Neste livro, os autores apresentam uma análise do primeiro estudo sistemático de multitorreção, nomeado de inquérito internacional sobre os exames e provas de acesso, realizado em 1936 por Laugier e Weinberg.

O estudo de Laugier e Weinberg foi realizado a partir de dados do *baccalauréat*, exame que estudantes franceses fazem ao final do ensino médio para entrar na universidade. Neste estudo, seis avaliadores corrigiram 100 exercícios de seis disciplinas: Francês, Latim, Inglês, Matemática, Filosofia e Física. Laugier e Weinberg propuseram, nessa pesquisa, que as divergências observadas com a experiência de multitorreção eram frutos do acaso e que os erros cometidos pelos professores nas correções eram similares aos erros das ciências físicas, ou seja, variações aleatórias. De modo análogo à Física, o importante seria encontrar um modo de reduzir os erros cometidos pelos corretores para se obter a 'verdadeira nota' dos trabalhos e provas (Noizet & Caverni, 1985).

Segundo Merle (2018), as experiências de Laugier e Weinberg, em 1936, e de Piéron, em 1963, mostraram que as notas dos estudantes nas avaliações escolares são distribuídas de acordo com a curva de Gauss ou distribuição normal, mais ou menos centrada em torno da média. Um professor tende a ajustar o nível de ensino e avaliação do desempenho do aluno, a fim de manter, de ano para ano, aproximadamente a mesma distribuição (gaussiana) das notas. Se a prova era um instrumento de medida e o desempenho dos estudantes o objeto medido, a distribuição normal das notas e/ou pontuações seria a esperada. O que não ocorreu.

O objetivo dessas pesquisas era a construção de instrumentos de aferição das aprendizagens e de técnicas de correção que reduzissem discrepâncias entre os avaliadores. Entretanto, Noizet e Caverni (1985) identificam uma contradição na conclusão de Laugier e Weinberg, justamente no que se refere ao suposto comportamento normal dos resultados obtidos na multitorreção. Ao analisarem as tabelas com as notas dadas pelos corretores, identificaram,

além da grande dispersão, que a distribuição delas não estava de acordo com características gaussianas, constando que, em média, dentro de um intervalo que deveria abranger 95% das notas dadas, estavam apenas 45%, incompatível com a noção de divergências aleatórias (Noizet & Caverni, 1985). Esses resultados estão relacionados a um importante conceito da docimologia¹: a confiabilidade.

A confiabilidade de um exame

A confiabilidade refere-se à consistência das avaliações, ou seja, “para analisarmos se um exame é confiável temos que quantificar em que medida o desempenho dos examinandos se mantém sensivelmente o mesmo, se resolverem o exame em tempos ou ocasiões diferentes” (Fernandes, 2009, p. 134). Parece razoável supor que uma pequena variação de desempenho de um estudante em testes aplicados em momentos distintos de fato venha a ocorrer. Seja por fatores externos à escola – questões emocionais e fisiológicas dos estudantes – como pela própria variação das questões contidas nesses testes (Vaz & Nasser, 2019).

As correções dos exames podem variar muito de corretor para corretor, principalmente em questões não objetivas, de resposta aberta. Para Dolin e colaboradores (2018), a confiabilidade de uma avaliação está relacionada à precisão dos resultados em determinado contexto e para um determinado fim.

Existem muitos fatores que podem reduzir a confiabilidade de uma avaliação. Por exemplo, a confiabilidade é reduzida se os resultados dependerem de quem conduz a avaliação (qual professor ou examinador), de quem classifica os desempenhos de avaliação dos alunos (que apontam ou observador ou examinador externo em exames orais) ou sobre as questões específicas usadas em um teste escrito quando eles podem testar apenas uma amostra de todos os diferentes tópicos e níveis de aprendizado incluídos no currículo avaliado. (Dolin et al, 2018, p. 64, *tradução nossa*)

Diversos estudos de multicorreção, realizados em diferentes países, comprovam a existência de interferências do avaliador na correção e na pontuação dos itens, colocando “em xeque” a confiabilidade dos exames (Noizet & Caverni, 1985; Romagnano, 2001; Suchaut, 2008, Vaz & Nasser, 2019). Em experimentos de multicorreção, a amplitude das notas atribuídas a mesma prova é um indício relevante do forte grau de subjetividade na correção de testes escolares.

Mesmo considerando a hipótese que o conhecimento de alguém possa ser mensurado, que os professores sejam absolutamente neutros em sua atuação profissional e que todos os testes fossem construídos com embasamento científico de neutralidade e objetividade, a crença na possibilidade de usar o teste para “medir” de alguma forma o conhecimento também é questionada a partir desses resultados (Vaz & Nasser, 2019, p. 7)

Podemos, nesse sentido, admitir que a ideia de medir a aprendizagem de um estudante a partir de um exame seja realmente um mito, como afirmam Buriasco e colaboradores (2009). A objetividade, neste caso, seria “como o mítico pote de ouro no final do arco-íris, seria

¹ Esta palavra foi cunhada por Henri Piéron em 1920. Trata-se do estudo sistemático dos exames, em particular do sistema de atribuição de notas e dos comportamentos dos examinadores e examinados.

maravilhoso se pudéssemos tê-lo, mas ele não existe. Todas as avaliações da compreensão matemática dos alunos são subjetivas.” (Romagnano, 2001, p. 31, *tradução nossa*).

Certamente, compreendemos que as provas de resposta objetiva podem fornecer às avaliações uma maior confiabilidade, na medida em que as pontuações dadas em um exame de múltipla-escolha independem do corretor. No entanto, é imprescindível reconhecer que a prova ou qualquer outro instrumento avaliativo permite que o docente realize uma leitura da aprendizagem dos estudantes, pois se trata de algo humano e interpretativo.

Se por um lado, os experimentos mais recentes em multicorreção não fornecem nada de inédito à Educação Matemática, apenas corroboram com os estudos desenvolvidos desde o início do século passado na França. Por outro, tais estudos ainda podem oferecer contribuições distintas. Neste texto apresentamos uma investigação de multicorreção realizada com professores que ensinam matemática e pedagogos sobre a relação entre a correção de itens e as habilidades avaliadas nesses itens. Essa perspectiva remete a outro conceito da docimologia: a validade.

A validade de um exame

O conceito de validade está associado à capacidade de um instrumento de avaliar aquilo que ele foi projetado para avaliar. “A validade de uma avaliação se refere ao grau pelo qual as notas de um teste permitem tirar conclusões adequadas, significativas e úteis em relação ao(s) objetivo(s) do teste” (Fidalgo, 2006, p. 20). Dois tipos de validade serão importantes para este estudo: (1) a validade de conteúdo – ou o que se quer avaliar e (2) a validade de construto – se a avaliação mede exatamente a habilidade que deve medir. O primeiro se refere ao tópico matemático e o segundo a habilidade deste tópico que está sendo avaliada.

A validade de conteúdo refere-se a quão adequadamente a avaliação abrange o domínio do assunto que está sendo ensinado. No entanto, a cobertura de conteúdo não é suficiente para descrever o alcance total de um teste ou outra ferramenta de avaliação. Problemas matemáticos que pretendem medir habilidades e competências na aprendizagem de frações, por exemplo, mas que exigem boas habilidades de leitura e interpretação, podem oferecer um resultado enviesado, a menos que o construto avaliado inclua, por exemplo, a leitura como parte das demandas de aprendizagem. “É por isso que a validade de construto é um conceito de validade cada vez mais prevalente, abrangendo muitas das outras medidas de validade” (Dolin et al, 2018)

A Investigação realizada

Durante o Encontro Nacional de Professores que Ensinam Matemática (ENOPEM), realizado de forma online e organizado por instituições brasileiras, o primeiro autor deste trabalho ministrou um minicurso de 3h de Avaliação Escolar. Participaram deste minicurso, de forma contínua, 11 educadores, sendo 6 professores formados em matemática, 2 professores de Matemática com formação em Pedagogia, 2 licenciados em matemática e 1 pedagogo que não atua ensinando matemática.

O objetivo deste estudo é averiguar como a análise das habilidades investigadas em um item podem influenciar sua correção. Para este estudo, a habilidade é a capacidade de realizar

uma ação (Lima, 2018), ou seja, ela funciona como âncora para referir um saber mais amplo, que não se resume a apenas uma resolução de exercícios.

Portanto, para isso, foi solicitada, aos 11 participantes, a tarefa de corrigir cinco soluções distintas de um mesmo problema em dois momentos: um, antes das discussões realizadas no minicurso, acerca da importância de refletirmos sobre ‘o que realmente estamos avaliando quando avaliamos?’; outro, depois, na metade do minicurso.

A Figura 1 ilustra o enunciado da questão oferecida e a habilidade que deveria ser investigada.

Questão discursiva de uma prova

- Habilidade: Resolução de Problemas envolvendo o Teorema de Pitágoras
- 9º ano
- Enunciado

Em uma folha de papel retangular fez-se um corte, retirando um triângulo retângulo, como ilustra a figura abaixo:

Calcule o perímetro do triângulo retirado

Figura 1. Enunciado e habilidade da questão oferecida.

Cada participante acessou um formulário digital para realizar a correção. A primeira correção consiste apenas na atribuição de pontuações para cada uma das cinco soluções (Figura 2). As pontuações dadas em cada item poderiam ser 0; 0,5; 1; 1,5 ou 2.

Solução 1
 Diagram: Right triangle with legs 4 cm and 3 cm, hypotenuse 7 cm.
 Calculation: $3 + 4 + 7 = 14 \text{ cm}$

Solução 2
 Diagram: Right triangle with legs 4 cm and 3 cm, hypotenuse 5 cm.
 Calculations: $x^2 = 3^2 + 4^2$, $x^2 = 25$, $x = 5$, $3 + 4 + 5 = 13 \text{ cm}$

Solução 3
 Diagram: Right triangle with legs 4 cm and 3 cm, hypotenuse 12,5 cm.
 Calculations: $x^2 = 3^2 + 4^2$, $x^2 = 25$, $x = 12,5$, $\text{Perímetro} = 3 + 4 + 12,5 = 19,5 \text{ cm}$

Solução 4
 Diagram: Right triangle with legs 4 cm and 3 cm, hypotenuse 5 cm.
 Calculations: $x^2 = 3^2 + 4^2$, $x^2 = 25$, $x = 5$, $\text{Perímetro} = 3 \times 4 \times 5 = 60 \text{ cm}$

Solução 5
 Diagram: A pentagon with top edge 4 cm, right vertical edge 4 cm, bottom edge 8 cm, left vertical edge 7 cm, and a slanted edge of 5 cm.
 Calculations: $x^2 = 3^2 + 4^2$, $x^2 = 25$, $x = 5$, $\text{Perímetro} = 8 + 7 + 4 + 5 + 4 = 28 \text{ cm}$

Figura 2. As cinco soluções da questão oferecida.

As soluções apresentadas nos formulários foram construídas pelos autores do trabalho a partir de suas vivências nos diversos cotidianos escolares em que atuam. Na solução 1, o

estudante não utiliza a fórmula do teorema de Pitágoras, mas calcula o perímetro usando 7 como medida da hipotenusa. Na solução 2, o estudante aplica e desenvolve corretamente o teorema de Pitágoras, construindo a expressão correta para o cálculo do perímetro, mas erra a adição. O erro cometido na solução 3 é no cálculo da raiz quadrada. Nela o estudante aplica corretamente o teorema de Pitágoras, demonstrando também compreender o conceito de perímetro. O erro cometido na solução 4 é, exclusivamente, de perímetro. Nessa solução, o estudante realiza corretamente a construção e os cálculos do teorema de Pitágoras. Na solução 5, o estudante também desenvolve corretamente o cálculo da hipotenusa, porém calcula o perímetro do pentágono, não o do triângulo.

Na segunda correção, os respondentes acessam um formulário alternativo que continha, além das mesmas soluções do formulário anterior, orientações para a correção (Figura 3). Essas orientações tinham como objetivo provocar a reflexão sobre as possíveis naturezas dos erros cometidos e sobre a habilidade avaliada.

Atividade de Correção de Questões *
Realize nova correção, mas desta vez analise cada resposta utilizando os seguintes critérios:
1) O que você está avaliando exatamente? Que habilidade/competência/saberes?
2) A solução apresenta uma estratégia correta?
3) Que tipo de erro foi cometido? Conceitual ou operacional?
4) O quanto este erro demonstra um não saber? Pode ter sido um lapsos do estudante?

Figura 3. Orientações para a correção.

Ademais, os respondentes eram solicitados a justificar a (nova) correção, explicando a pontuação atribuída, sendo essa diferente, ou não, daquela dada no primeiro formulário. O objetivo era compreender se as orientações oferecidas e/ou a necessidade de realizar uma maior reflexão sobre a correção iria promover mudanças na pontuação e no modo de analisar as respostas dos estudantes.

Resultados

A tabela 1 apresenta as médias, modas e medianas das primeiras e segundas correções, nesta ordem, realizadas das cinco soluções por todos os participantes da pesquisa.

Tabela 1
Resultados das correções realizadas pelos avaliadores.

Solução	Média 1 ^a /2 ^a	Moda 1 ^a /2 ^a	Mediana 1 ^a /2 ^a
1	1,0 / 0,73	0,5 / 0,5	1,0 / 0,5
2	1,5 / 1,6	1,5 / 1,5	1,5 / 1,5
3	0,91 / 1,14	1,0 / 1,0	1,0 / 1,0
4	0,91 / 1,32	1,0 / 1,0	1,0 / 1,0
5	1,2 / 1,36	1,0 / 1,5	1,0 / 1,5

Fonte: elaborado pelos autores.

Os resultados apresentados no quadro acima indicam que a correção realizada no segundo momento foi mais apurada em relação à habilidade investigada e ao erro cometido. Foi nesse ínterim, entre as duas correções, que o palestrante do minicurso promoveu reflexões teóricas sobre a necessidade de analisar as habilidades que se pretendem avaliar, tanto na elaboração, quanto na correção de itens.

A solução 1 apresentou uma redução na pontuação média e mediana. Notem que se trata da única solução em que o estudante não domina a habilidade *resolução de problemas envolvendo o teorema de Pitágoras*, habilidade informada no enunciado da questão (vide fig.1). Segundo o corretor 2,

Pensando como objetivo central da avaliação o teorema de Pitágoras, o aluno não alcançou este objetivo, entretanto não podemos desvalidar a interpretação da figura. (corretor 2)

As soluções 2, 3, 4 e 5 tiveram suas pontuações médias aumentadas na segunda correção. Ambas as soluções apresentam a equação correta que relaciona o teorema de Pitágoras. Isto revela que esses quatro estudantes dominam o conhecimento geométrico avaliado no item, ou seja, conseguiram compreender que se tratava de um problema que deveria ser resolvido pelo teorema de Pitágoras, demonstrando, inclusive, saber como relacioná-lo algebricamente.

Os erros cometidos nas soluções 2, 4 e 5 estão relacionados ao cálculo do perímetro, que é uma habilidade secundária, ou seja, não é a habilidade avaliada no item. Em ambas as soluções, o estudante obtém a medida da hipotenusa. A solução 3 apresenta um erro no cômputo da raiz quadrada, no decorrer do cálculo da hipotenusa. Vale ressaltar alguns comentários expostos pelos professores, por exemplo: o corretor 1 considerou que a solução 2 apresenta um erro insignificante, caracterizando-o como erro mínimo:

Assinalei a opção 1,5, mas atribuiria uma nota muito próxima a 2,0, exemplo: para este aluno eu daria 1,9, para ele voltar a observar qual foi o mínimo erro, entretanto o pensamento dele foi excelente. (corretor 1)

A corretora 3 apresentou uma reflexão interessante sobre a própria dinâmica da pesquisa em multcorreção e sobre a avaliação formativa.

Aqui fiquei confusa, é difícil avaliar algo sem conhecer quem respondeu. A avaliação formativa depende disso, penso. Mas, considerando um erro conceitual de perímetro e a habilidade se referindo ao teorema, talvez valesse 1,9 ou 2 nesse caso. O importante é apontar o equívoco para não se repetir.

Um toque já basta para esse aluno, talvez nem seja um erro, mas um esquecimento conceitual.
(corretora 3)

A avaliação não é impessoal. É uma relação humana, construída na interação entre professor e aluno ao longo das aulas. Refletir sobre os erros cometidos e sobre as habilidades avaliadas pode promover uma correção mais coerente, com maior validade.

Considerações Finais

A avaliação é socialmente localizada, a concepção docente sobre a avaliação depende de alguns fatores, como a concepção do professor em relação à Matemática, a concepção em relação ao ensino, trajetórias acadêmicas e profissionais do docente, o contexto escolar e as famílias que são atendidas pela escola (Lima & Nasser 2022). Portanto, desenvolver um olhar sobre as multiorreções cria possibilidades de compreensão sobre os vieses comuns na correção de tarefas.

Ocorreu uma mudança na forma de se corrigir, fruto das reflexões teóricas que foram realizadas, ao longo do encontro, sobre a necessidade de analisar as habilidades que se pretende avaliar, tanto na elaboração, quanto na correção de itens. Outrossim, a validade de conteúdo mostrou-se como algo difícil de ser medido, confirmando que a cobertura de conteúdo não é suficiente para descrever o alcance total de um teste ou outra ferramenta de avaliação. Os corretores, chegaram a indicar a habilidade de calcular perímetro como um elemento importante para correção. Junto a isso, a necessidade de conhecer o estudante também foi outro elemento que surgiu no estudo, mostrando como o contexto escolar possui influência na hora da correção.

Compreendemos que cada contexto escolar é único, assim como a unicidade do sujeito que é o docente, por isso, a validade do construto que ele elabora está fortemente associado ao que ele compreende o que é importante de se avaliar.

Referências e bibliografia

- Black, P. & Wiliam, D. (1998) Inside the Black Box: Raising Standards through Classroom Assessment. *The Phi Delta Kappan*, Bloomington 80 (2), 139-148.
- Buriasco, R. L. C., Ferreira, P. E. A., & Ciani, A. B. (2009). Avaliação como prática de investigação (alguns apontamentos). *Boletim de Educação Matemática*, 22 (33), 69-96.
- Dolin, J. et al. (2018) Exploring relations between formative and summative assessment. In: Dolin, J.; Evans, R. (org.). *Transforming assessment*. Cham: Springer, 53-80.
- Fernandes, D. (2009). *Avaliar para aprender: fundamentos, práticas e políticas*. São Paulo, BR: Unesp.
- Fidalgo, S. S. (2006) A avaliação na escola: um histórico de exclusão social-escolar ou uma proposta sociocultural para a inclusão? *Revista Brasileira de Linguística Aplicada*, 6 (2), 15-31.
- Guba, E. G. & Lincoln, Y. S. (2011) *Avaliação de quarta geração*. Campinas, BR: Editora Unicamp.
- Hadji, C. (2001). *Avaliação desmistificada*. Porto Alegre, BR: Artmed Editora.
- Lima, D. (2018) A avaliação por habilidades e competências. Dissertação apresentada ao PROFMAT - IM, UFRJ.

- Lima, D. & Nasser, L. (2022) “Concepções docentes sobre a avaliação em Matemática – valores e instrumentos que compõem a prática docente”, *Revemop*, 4, p. e202206. doi: 10.33532/revemop.e202206
- Merle, P. (2018) *Les pratiques d'évaluation scolaire: historique, difficultés, perspectives*. Paris: Presses Universitaires de France/Humensis
- Morgan, C. (2000). Better assessment in mathematics education? A social perspective. In: Boaler, J. (Org.). *Multiple Perspectives on Mathematics Teaching and Learning*. Westport, AblexPublishing, 225-242.
- Noizet, G.& Caverni, J-P. (1985) *Psicologia da avaliação escolar*. Coimbra: Coimbra Editora.
- Romagnano, L. (2001) The myth of objectivity in mathematics assessment. *Mathematics Teacher*, 94 (1), 31-37.
- Suchaut; B. La loterie des notes au bac: un réexamen de l'arbitraire de la notation des élèves. *Les Documents de Travail de l'IREDU*. 2008.
- Vaz, R. F. N; Nasser, L. (2019). Um estudo de multicorreção com professores de matemática. In: Conferência Interamericana de Educação Matemática, 15. *Anais...* Medellín: CIAEM, 2019