UM ESTUDO SOBRE A CONFIABILIDADE DA AVALIAÇÃO EM MATEMÁTICA

Lilian Nasser
PEMAT - UFRJ
lnasser.mat@gmail.com

Rafael Filipe Novôa Vaz
Instituto Federal do Rio de Janeiro–IFRJ; PEMAT - UFRJ
rafael.vaz@ifrj.edu.br

Resumo:
Neste trabalho questionamos a ideia de associar a avaliação a um instrumento de medida e defendemos que a avaliação está mais próxima de uma leitura do estágio de aprendizagem de cada aluno. Nosso argumento está sustentado pelos vieses cometidos por professores na correção de testes discursivos de Matemática. Consideramos que a avaliação seja parte integrante do sistema educacional e que os testes, individuais, escritos e sem consulta, são procedimentos predominantes na avaliação em Matemática. Mas até que ponto esses instrumentos de avaliação são válidos e confiáveis? Isto é, até que ponto esses instrumentos são isentos de subjetividade, garantindo equidade de condições para todos? Os primeiros resultados de estudos já desenvolvidos mostram uma grande variação das notas atribuídas a uma mesma prova por diferentes avaliadores, dependendo, entre outros fatores, da ordem em que as questões resolvidas são apresentadas. Essa amplitude de notas se deve a vieses identificados na correção. Para garantir um resultado justo, é sugerida a dupla diversificação dos instrumentos avaliativos, tanto em número quanto na forma.

Palavras-chave: Avaliação em Matemática; Vieses na Correção; Confiabilidade; Multicorreção.

INTRODUÇÃO

Perrenoud (1999) escreveu, há cerca de 20 anos, que a avaliação era tradicionalmente associada, na escola, à criação de hierarquias de excelência, pois os alunos são comparados e depois classificados. Pouca coisa mudou na dinâmica das aulas: os quadros negros foram substituídos pelos quadros brancos e, posteriormente, pelos slides, mas a aula continua sendo essencialmente expositiva.

Apesar dos avanços nos métodos de ensino, da inserção de novas tecnologias e de uma preocupação crescente com práticas mais inclusivas, as avaliações da aprendizagem, com raras exceções, continuam as mesmas, centradas principalmente em exames individuais, escritos e sem consulta. São avaliações somativas, com o intuito de verificar e quantificara aprendizagem dos estudantes num ciclo de aprendizagem, amparadas em uma filosofia positivista, com a predominância de exames para 'medir' a aprendizagem dos estudantes. Tais procedimentos avaliativos proporcionam um tratamento genérico a todos os estudantes, não levando em consideração particularidades e necessidades do indivíduo.

Paulo Freire (1981), ao identificar o educador como um depositário de saberes cunhou o termo Educação Bancária.

> Em lugar de comunicar-se, o educador faz "comunicados" e depósitos que os educandos, meras incidências, recebem pacientemente, memorizam e repetem. Eis aí a concepção "bancária" da educação, em que a única margem de ação que se oferece aos educandos é a de receberem depósitos, guardá-los e arquivá-los. (FREIRE, 1981, apud ROMÃO, 1999, p. 58)

Nesse contexto os alunos, meros receptores, devem 'devolver' os saberes recebidos e devidamente memorizados através das avaliações, geralmente, testes e provas, constituindo o que Romão (1999) denominou de Avaliação Bancária.

> A avaliação bancária consistiria na capacidade do aluno de buscar nos seus "arquivos mentais" os depósitos ali deixados, exatamente como foram feitos, sem interpretação, sem acréscimos, sem qualquer tipo de juros ou deduções, e devolvê-los ao depositante, mediante requisição: a prova, o teste, o exame final (FIDALGO, 2006, p. 19).

Na educação e avaliação bancárias, "os alunos se transformam em meros arquivos especulares das 'verdades' descobertas previamente pelos professores na sua formação e na preparação de suas aulas" (ROMÃO, 1999, p.58). Ao reconhecer a educação bancária como uma prática ainda comum, algumas reflexões em torno da função da escola e do significado real da aprendizagem podem ser feitas. Qual é o papel da memória para o trabalhador na era do Google? Se Rubem Alves estiver correto em sua famosa frase, o aprendizado é o que fica depois que o esquecimento fez seu trabalho, os testes – individuais, escritos, sem consulta e com tempo delimitado – provavelmente não farão mais sentido.

Há dois pontos iniciais que devem ser analisados, o primeiro é se a aprendizagem de um indivíduo pode ser mensurada. O segundo ponto está relacionado à eficiência dos testes em realizar essa medição.

Nesta investigação, partiremos do pressuposto de que a avaliação é parte integrante e fundamental do sistema educacional. A partir daí, consideramos que haja uma necessidade de realizar exames e avaliações variadas nesse sistema para que professores, responsáveis e comunidade escolar realizem uma leitura da aprendizagem dos estudantes, ainda que essa leitura seja apenas "uma" leitura dentre outras possíveis.

Este estudo apresenta reflexões teóricas e resultados empíricos que permitem reafirmar que a avaliação está mais próxima de uma leitura do que de uma medida e alertam para a necessidade de mudanças na concepção avaliativa predominante no meio escolar.

OS EXAMES E A IDEIA DA MEDIÇÃO

Na ciência a ideia de medição pode ser interpretada como associar um fenômeno a um número. “Medir significa comparar grandezas de mesma espécie, tomando-se uma delas como unidade” (ROMÃO, 1999, p. 47).A objetividade da avaliação escolar, em que seria possível medir o conhecimento de alguém, está associada a uma filosofia positivista, na qual, a neutralidade e imparcialidade são seus pilares (MORGAN, 2000; VAZ; NASSER, 2019).

Fischer (2008) realizou uma pesquisa com professores universitários que atuam em cursos de licenciatura, a fim de investigar suas concepções em relação à avaliação. Neste trabalho, foi constatado que os professores formadores associam a objetividade na avaliação escolar à clareza, à uniformidade nos critérios de avaliação e à neutralidade no campo da matemática. A autora cita que

> as características apontadas como constituintes do *habitus* desse professor, como a busca pela objetividade, a concepção positivista de rigor no trato dessa ciência e de seu ensino, um certo descrédito do fazer pedagógico e a adoção de uma postura pouco flexível, têm fortes marcas desse paradigma de ciência. (Fischer, 2008, p.96)

Segundo Hadji (2001), o julgamento do avaliador é “sempre infiltrado por elementos provenientes do contexto escolar e social, desde a carga afetiva e a dimensão emocional devido à presença efetiva dos alunos” e, geralmente, “ignora que se baseia em parte em uma representação construída do aluno e em convicções íntimas que nada têm de científicas” (p.32). Para esse autor, a avaliação não é uma medida, porque “o avaliador não é um instrumento” e porque o “que é avaliado não é um objeto no sentido imediato” (p.34).

Romão (1999) afirma que as notas não fazem sentido matematicamente, pois elas

> se constituem em simples ordenações – e ainda assim de legitimidade duvidosa. Os intervalos entre as diversas notas não serão uniformes, porque não é possível estabelecer uma rígida regularidade entre os graus crescentes de dificuldade das situações-problema ou questões formuladas, nem estabelecer rígidos limites entre a qualidade das respostas. (ROMÃO, 1999, p. 48)

Buriasco, Ferreira e Ciani (2009) defendem que a avaliação escolar é composta por um rito e um mito. “O rito de avaliar – aplicar uma prova ou um teste escrito e converter as resoluções e respostas de cada estudante a um valor numérico” (p.70) está associado ao mito de “medir e classificar de maneira precisa os alunos” (p.71). Para Buriasco e colaboradores,

> via de regra, negligencia-se que o quantitativo advém do qualitativo, e, no caso da avaliação, a nota atribuída não emerge de maneira pura e unívoca dos instrumentos utilizados, mas é produzida pelo avaliador, que, para fazê-lo, pode se valer de instrumentos. Por fim, o rito de avaliar se constitui numa prática que confere uma validade ilusória ao mito da possibilidade do exercício da precisão e da justiça. (BURIASCO; FERREIRA; CIANI, 2009, p.72)

Na predominância desse modelo positivista de avaliação escolar configura-se o que Fernandes (2009) denomina de *paradigma psicométrico de avaliação*. Segundo este autor, há uma tendência de a avaliação centrar-se mais nos resultados ou nos produtos do que no processo de aprendizagem. As três características principais do paradigma são:

(1) é possível determinar exatamente o que os alunos sabem e são capazes de fazer;

(2) as aprendizagens dos alunos constituem uma realidade que pode ser avaliada de forma objetiva, neutra e sem quaisquer inferências valorativas;

(3) testes de naturezas diversas – cientificamente construídos e, como tal, objetivos e neutros – permitem a quantificação das aprendizagens dos alunos. (Fernandes, 2009, p.81-82).

Na realidade, isso nem sempre acontece, como mostram os estudos de multicorreção de avaliações escolares, remetendo à necessidade de investigar outras formas de avaliar.

VALIDADE E CONFIABILIDADE

A Validade e a Confiabilidade são consideradas pela literatura como as principais características psicométricas de uma avaliação (FERNANDES, 2009). Para que as avaliações sejam sólidas e atinjam o propósito para o qual foram elaboradas, elas devem estar livres de preconceitos e distorções. Confiabilidade e Validade são dois conceitos importantes para definir e medir o viés e a distorção de uma avaliação.

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 úteisem relação ao(s) objetivo(s) do teste” (FIDALGO, 2006, p. 20). A literatura apresenta diversos tipos de validade. Fidalgo (2006) destaca quatro tipos:

(1) a validade de conteúdo– ou o que se quer avaliar;

(2) a validade de construto– se a avaliação mede exatamente a habilidade que deve medir;

(3) a validade aparente – como as pessoas veem a avaliação, se acreditam nos instrumentos e nos processos envolvidos e

(4) o efeito de refluxo – que se refere ao efeito que os resultados obtidos em provas e testes têm sobre o ensino, sobre a prática educativa.

Sobre a validade de conteúdo, Fernandes (2009) acrescenta que tal tipo está relacionado à medida que um teste contém de uma amostra significativa do conteúdo relevante. Este autor destaca, além dos dois primeiros tipos indicados por Fidalgo, a validade de previsão (em que medida um teste é bom para indicar desempenhos futuros), validade concorrente (em que medida um teste se correlaciona com outros testes), a validade de critério (em que medida o teste prevê um desempenho sob determinado critério).

A validade de conteúdo refere-se a quão adequadamente a avaliação abrange o domínio do assunto que está sendo ensinado e é geralmente baseada no julgamento de especialistas no assunto. 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, p. 63).

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.

Para Kellaghan e Madaus (2003, apud FERNANDES, 2009, p. 135), "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 testado. (DOLIN et al., 2018, p. 64, *tradução nossa*)

OS EXAMES ESCOLARES SÃO CONFIÁVEIS?

Precursores dos estudos psicométricos na avaliação escolar, Noizet e Caverni (1985) desenvolveram um vasto estudo sobre multicorreção envolvendo diversas disciplinas, que culminou na publicação do livro Psychologie de l'évaliationscolaire em 1978, traduzido em 1985 para o português. Neste livro, os autores fazem uma análise do primeiro estudo sistemático de multicorreção, nomeado de inquérito internacional sobre os exames e provas de acesso, realizado em 1936 por Laugier e Weinberg.

O estudo de Laugier e Weinbergfoi realizado a partir de dados do *baccalauréat*, exame que estudantes franceses fazem ao final do ensino médio para entrar na universidade. Nesse 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 multicorreção 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" (p.118-119). De Landsheere (1992) definiu essa regra como lei ou efeito Posthumus, em homenagem a uma professora holandesa que associou a curva à distribuição das notas dos estudantes.

> A lei de Posthumus se deve ao fato de que existe um fenômeno de adaptação do professor ao nível escolar de seus alunos. Quando o nível médio de educação é baixo, o professor adapta logicamente o conteúdo do aprendizado ao nível de seus alunos, bem como a dificuldade dos exercícios e controles escritos. Um fenômeno oposto está em ação quando os alunos são de bom ou muito bom nível acadêmico. As avaliações escolares são, de forma mais ou menos consciente, destinadas a diferenciar os alunos, sejam eles muito bons ou muito fracos. (MERLE, 2018, p. 120, *tradução nossa*)

De acordo com Merle (2018), as notas obtidas por um aluno dependem do nível médio da turma. Uma avaliação mais objetiva só seria alcançada em avaliações externas,

padronizadas, já que utilizam uma amostra representativa e importante de alunos e não permitem a influência do efeito Posthumus.

Noizet e Caverni (1985) identificam uma contradição na conclusão de Laugier e Weinberg (1936), justamente no que se refere ao suposto comportamento normal dos resultados de multicorreção. Para exemplificar essa contradição, Noizet e Caverni (1985) utilizaram os resultados apresentados no inquérito internacional sobre os exames e provas de acesso, no qual as notas atribuídas por 76 avaliadores que corrigiram provas de francês não apresentaram características da aleatoriedade.

Tabela 1 – Distribuição de notas do Inquérito de Laugier e Weinberg (1936)

Nota	0 - 1	2 - 3	4 - 5	6 – 7	8 - 9	10 - 11	12 - 13
Número de avaliadores	1	6	20	34	10	3	2

Fonte: Noizet; Caverni (1985, p. 43)

Além da grande dispersão observada na tabela 1, as notas foram reguladas em classes de duas para fazer aparecer o caráter gaussiano da distribuição. No entanto, “o intervalo de confiança de 5% só compreende 45% das notas, ao invés das 95% esperadas e 55% dos avaliadores em vez dos 5% toleráveis estão, portanto, fora dos limites de confiança correspondentes”. Esse resultado é “totalmente incompatível com a noção de divergências aleatórias” (NOIZET; CAVERNI, 1985).

Vaz e Nasser (2018a, 2018b, 2019) desenvolveram diversos estudos inseridos em uma pesquisa de multicorreção de testes discursivos de Matemática. Inicialmente a pesquisa foi realizada com licenciandos em Matemática e, posteriormente, com professores de Matemática formados e atuantes. Nessa pesquisa, foi solicitado aos corretores que realizassem a correção de um mesmo teste resolvido por um estudante fictício, atribuindo a pontuação de cada questão e uma nota ao teste. Foi disponibilizado a cada um dos participantes do estudo o gabarito do teste, em que todas as questões tinham o mesmo valor.

Os resultados, em consonância com as pesquisas francesas, mostraram uma imensa amplitude entre as notas geradas pelos corretores. A figura 1 mostra a distribuição de notas dadas pelos 45 licenciandos.

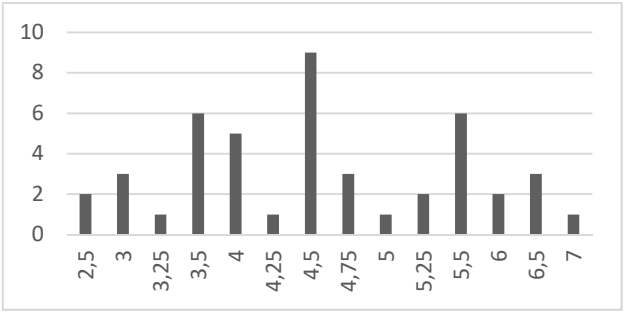

Figura 1 – Frequência de notas dadas pelos licenciandos.
Fonte: Vaz; Nasser (2018b)

Quanto aos 14 professores entrevistados, a amplitude das notas ainda foi grande, como mostra a figura 2.

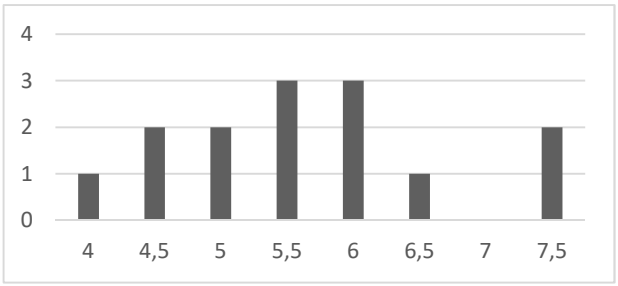

Figura 2 – Frequência de notas dadas professores.
Fonte: Vaz; Nasser (2019)

Mesmo não tendo uma variação tão grande, a amplitude das notas dos professores pode ser considerada significativa, pois trata-se de uma amostra menor. Vaz e Nasser também identificaram em seus estudos a existência de vieses cometidos pelos corretores, identificados anteriormente em pesquisas na França: o efeito halo e o efeito âncora (NOIZET; CAVERNI, 1985; MERLE, 2018).

O efeito *halo* consiste "no julgamento do todo a partir de características obtidas inicialmente, e se apresenta quando uma impressão é formada a partir de uma característica inicial influenciando múltiplos julgamentos ou classificações de fatores não relacionados" (VAZ; NASSER, 2018a).O efeito halo não é peculiar à avaliação escolar, é inerente a todos os processos de julgamento, é um produto de crenças e valores, um efeito de estereótipos e ideologias. Esse viés cognitivo é um erro sistemático que leva a uma avaliação tendenciosa (MERLE, 2018).

Vaz e Nasser (2018a; 2019), constataram que o ordenamento da correção das questões poderia influenciar o corretor e alterar a pontuação final de um teste discursivo em Matemática. A tabela 2 ilustra o resultado de medidas de tendência central realizados a partir das notas geradas por licenciandos na correção de um teste discursivo de Matemática contendo 4 questões. A diferença entre os testes era o ordenamento das questões e suas respectivas soluções. No teste A, a primeira questão apresentava uma solução resolvida

corretamente, e a quarta uma solução resolvida de modo incorreto. No teste B, as ordens dessas duas questões e suas soluções eram alteradas, mantendo as questões 2 e 3, parcialmente corretas, na sua posição de origem. Em ambos os testes, as questões intermediárias apresentavam resoluções parcialmente corretas.

Tabela 2 – Medidas de tendência central da Fase 1

	Teste A	Teste B
Média	6,43	7,17
Moda	6,25	7,25
Mediana	6,5	7,25

Fonte: Vaz; Nasser (2018a)

Uma explicação para os resultados de Vaz e Nasser (2018a) é que estes podem estar relacionados ao efeito *âncora*. Segundo Noizet e Caverni (1985), o termo âncora é oriundo da psicologia da percepção e pode ser definido como "uma correspondência privilegiada entre um objeto e uma categoria de resposta" (p.116). Na correção de uma prova, um objeto privilegiado pode ser, por exemplo, a resolução de uma das questões iniciais, como as questões correta e incorreta presentes nos testes A e B de Vaz e Nasser (2018a).

Os vieses de correção e a dispersão desses resultados sugerem ou, por que não, comprovam a existência de caráter subjetivo na correção de testes de Matemática.

> 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)

Romagnano (2001) apresenta um modo de caracterizar se os exames são confiáveis. Segundo esse autor, a confiabilidade de um exame pode ser diagnosticada em duas situações: quando professores diferentes usam um método consistente para avaliar o conhecimento de um determinado aluno e as avaliações desses professores geram resultados semelhantes, ouquando dois alunos com aproximadamente o mesmo nível de compreensão de um conjunto de ideias matemáticas obtêm resultados similares na mesma avaliação.

A partir daí, há duas interpretações possíveis das discrepâncias encontradas nos resultados de Vaz e Nasser (2008a, 2008b, 2019): ou há uma falência no método avaliativo utilizado (os testes) ou a ideia de realizar tal medição a partir de um teste é falaciosa.

(1) A primeira alternativa seria considerar que o método avaliativo utilizado na pesquisa não era consistente, ou seja, os dois testes elaborados pelos autores não eram

instrumentos precisos e confiáveis. Talvez por não atender a determinados critérios desconhecidos, inclusive, pelos autores. No entanto, foram elaborados e utilizados testes usuais, considerados comuns, até mesmo, pelos licenciandos e professores que participaram da investigação.

O teste utilizado na fase1, descrita em Vaz e Nasser (2018a), era mais simples, composto de uma questão do tipo resolva, com quatro itens (***a***, ***b***, ***c*** e ***d***). Nas fases seguintes, o teste foi mais elaborado e focava em outro conteúdo. Passou a conter 5 questões diferenciadas sobre áreas de figuras planas, sendo uma delas mais argumentativa. A figura 3 mostra uma das versões do teste, em que a primeira questão está respondida corretamente e a solução da quinta questão está errada.

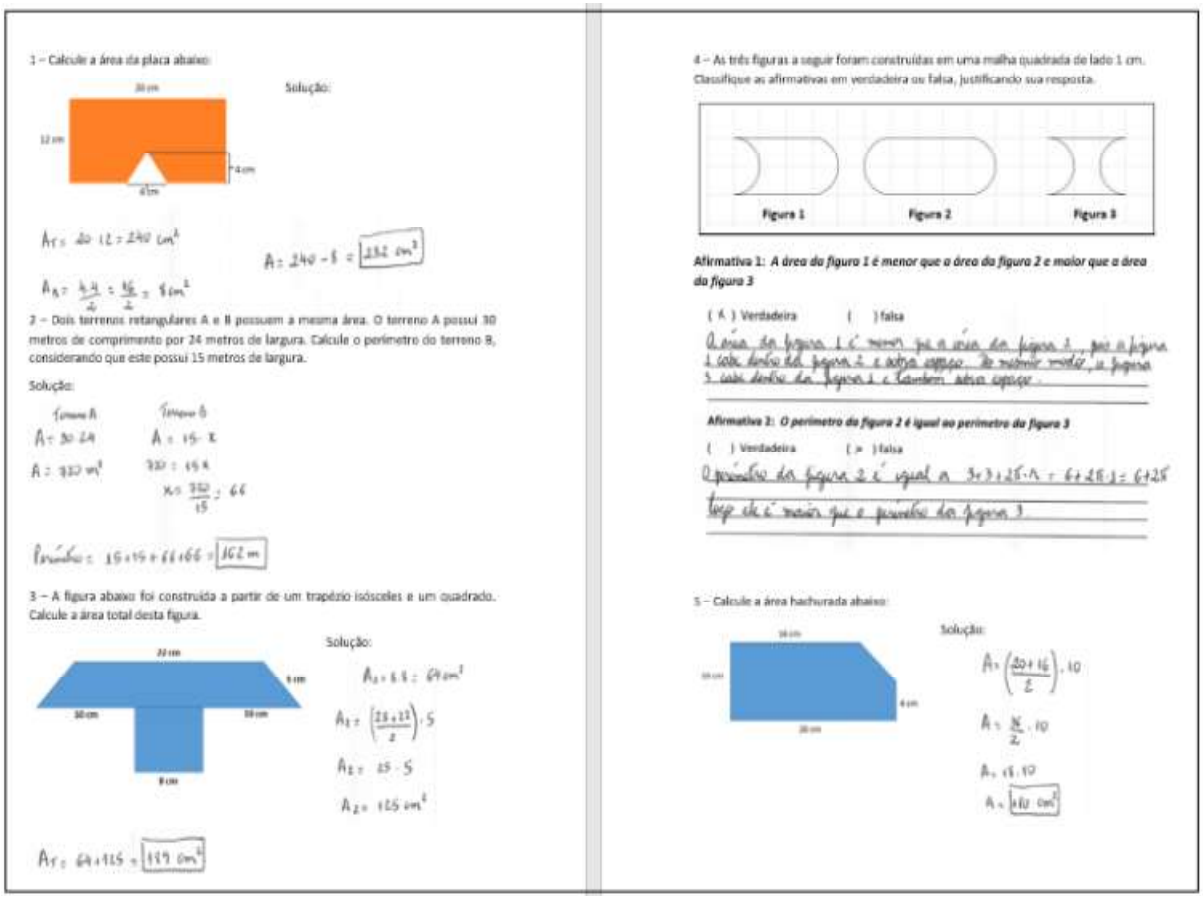

Figura 3 – Teste corrigido pelos professores
Fonte: Vaz; Nasser (2019)

As avaliações em larga escala, cientificamente construídas, podem fornecer resultados mais confiáveis. Dois alunos com conhecimentos similares tenderiam a obter resultados semelhantes e, com certeza, a partir de um mesmo gabarito, dois professores corrigiriam e pontuariam, igualmente, o mesmo exame. Então, por que não utilizamos testes de múltipla escolha? Provavelmente, por causa da Validade! Até que ponto o número de acertos em um

teste dessa natureza pode refletir o quanto o aluno adquiriu dos saberes e se desenvolveu as competências planejadas pelo professor?

Segundo Fernandes (2009, p. 135), para diminuir essas ameaças à confiabilidade do exame, o que em geral se faz é "padronizar as condições de administração" e estabelecer critérios de correção claros e detalhados. No entanto, "quanto mais rigorosas forem essas condições, mais limitações acabam por surgir quanto ao tipo de tarefas a serem incluídas no exame e, portanto, de conhecimentos ou domínios do currículo, que no fim, se pode avaliar".

(2) Outra interpretação possível é 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 maravilhoso se pudéssemos tê-lo, mas ele não existe. Todas avaliações da compreensão matemática dos alunos são subjetivas." (ROMAGNANO, 2001, p. 31, *tradução nossa*).

Nesta linha de pensamento, em consonância com as ideias de Romagnano (2001, p. 35) ao afirmar que "o conhecimento é multifacetado, complexo, construído individualmente e inseparavelmente ligado ao contexto no qual o aprendizado ocorre", não seria possível realizar a medição significativa da aprendizagem, ou do conhecimento, de um indivíduo independente do instrumento.

CONSIDERAÇÕES FINAIS

Para Romão (1998, p. 51) "a correção de uma questão, por mais fidedigna que seja, estará condicionada à subjetividade de quem vai corrigi-la". Segundo Buriasco (1999), a avaliação é sempre um processo inacabado carregado de subjetividade. Reconhecer a presença e o papel da subjetividade no processo avaliativo é um caminho árduo e de complexa transformação. Ainda hoje exercemos a educação e avaliação bancárias.

Os resultados apresentados neste trabalho apresentam razões propulsoras para o rompimento de tais perspectivas. Consideramos que seja "necessário passarmos de uma preocupação centrada no produto (que se pretendia medir, pesar...) para uma preocupação centrada no processo de produção" (BURIASCO, 1999, p. 218). Reconhecemos, talvez de forma preconceituosa, que tal transformação seja ainda mais difícil para aqueles professores das matérias ditas "exatas".

Fernandes (2009, p. 95) defende que "a diversidade de métodos de coleta de informações permite avaliar mais domínios do currículo, lidar melhor com a grande

diversidade de alunos que hoje estão nas salas de aula e, também, reduzir os erros inerentes à avaliação".Concordamos com a ideia de Fernandes, e a partir daí, propomos como um caminho para minimizar os vieses e tornar o processo avaliativo mais confiável seja o que denominamos de *dupla diversificação avaliativa*.

A dupla diversificação avaliativa consiste em diversificar as avaliações em relação aos momentos em que ocorrem, dando um caráter maior de continuidade, e aos instrumentos, permitindo leituras mais esclarecedoras. A avaliação está mais próxima de um filme do que de uma foto, deste modo, não deve ser restrita aos exames esporádicos em momentos específicos (normalmente nos finais de ciclos). A alternância de instrumentos, alguns breves outros longos, trabalhos, portfólios, seminários, testes discursivos e testes de múltipla escolha, podem proporcionar maior confiabilidade e validade nos resultados, além de oferecer um maior leque de "feedbacks" para os estudantes sobre aprendizagem.

Para Perrenoud (1999, p. 173), "mudar a avaliação significa possivelmente mudar a escola". Reconhecendo a ideia de medição contida nos exames como um mito (ROMAGNANO, 2001; BURIASCO; FERREIRA; CIANI, 2009; VAZ; NASSER, 2019), é preciso ponderar sob que condições a avaliação escolar pode ser de fato um momento de trocas, de aprendizagem e de investigação entre estudantes e professores.

Se para que a avaliação abandone seu papel classificatório e excludente (PERRENOUD, 1999; BURIASCO, 1999; FERNANDES, 2009) seja de fato necessário que mudemos a escola, então que comecemos a transformação em prol de novas possibilidades de avaliação, mais confiáveis, válidas e, principalmente, mais justas.

REFERÊNCIAS BIBLIOGRÁFICAS

BURIASCO, R. L. C. **Avaliação em Matemática:** um estudo das respostas de alunos e professores. 1999. 238 f. Tese (Doutorado em Educação) – Universidade Estadual Paulista. Marília, SP, 1999.

BURIASCO, R. L. C.; FERREIRA, P. E. A.; CIANI, A. B. Avaliação como prática de investigação (alguns apontamentos). **Boletim de EducaçãoMatemática**, Rio Claro, v. 22, n. 33, p.69-96, 2009.

DOLIN, J.; BLACK, P.; HARLEN, W.; TIBERGHIEN, A. Exploring relations between formative and summative assessment. In: **Transforming assessment**. Springer, Cham, p. 53-80, 2018.

FERNANDES, D. **Avaliar para aprender: fundamentos, práticas e políticas**. São Paulo: Editora Unesp, 2009.

FIDALGO, S. S. A avaliação na escola: um histórico de exclusão social-escolar ou uma proposta sociocultural para a inclusão?**Revista Brasileira de Linguística Aplicada**, v. 6, n. 2, p. 15-31, 2006.

FISCHER, M. C. B. Os formadores de professores de matemática e suas práticas avaliativas. In VALENTE, W. R. (Org.). **Avaliação em matemática: história e perspectivas atuais.** Campinas: Papirus, 2008. p. 75 -100.

HADJI, C. **Avaliação desmistificada**. Porto Alegre: Artmed Editora, 2001. 136 p.

MERLE, P. **Les pratiquesd'évaluationscolaire**:historique, difficultés, perspectives. Paris: Presses Universitaires de France/Humenisis, 2018.

MORGAN, C. Better assessment in mathematics education? A social perspective. In: BOALER,J.(Org.). **Multiple Perspectives on Mathematics Teaching and Learning**. Westport, AblexPublishing,2000.p. 225-242.

NOIZET, G.; CAVERNI, J-P. **Psicologia da avaliação escolar**. Coimbra: Coimbra Editora, 1985.

ROMAGNANO, L. The myth of objectivity in mathematics assessment. **MathematicsTeacher**, v. 94, n. 1, p. 31-37. 2001.

ROMÃO, J. E. Avaliação dialógica. **Desafios e perspectivas.** São Paulo, 1998.

VAZ, R. F. N; NASSER, L. Um estudo sobre o efeito halo na correção de provas. In: ENCONTRO ESTADUAL DE EDUCAÇÃO MATEMÁTICA DO RIO DE JANEIRO, 7., 2018a, Rio de Janeiro. **Anais...** Rio de Janeiro: SBEM, 2018a.

VAZ, R. F. N; NASSER, L. Avaliação em matemática: um estudo sobre multicorreção In: SEMINÁRIO INTERNACIONAL DE PESQUISA EM EDUCAÇÃO MATEMÁTICA, 7., 2018b, Foz do Iguaçu. **Anais...**Foz do Iguaçu: SBEM, 2018b.

VAZ, R. F. N; NASSER, L. Um estudo de multicorreção com professores de matemática. In: CONFERÊNCIA INTERAMERICANA DE EDUCAÇÃO MATEMÁTICA, 15., 2019, Medelín. **Anais...** Medelín: CIAEM, 2019.