ACTIO: Docência em Ciências
ISSN: 2525-8923

http://periodicos.utfpr.edu.br/actio

Análise das validades de conteúdo e de construto de instrumentos de avaliação de biologia

Marcelo Alves Coppi
mcoppi@uevora.pt
orcid.org/0000-0001-67343-7592
Centro de Investigação em Educação e Psicologia da Universidade de Évora (CIEP-UE), Évora, Évora, Portugal

RESUMO

O objetivo deste trabalho consistiu em analisar a validade de três instrumentos de avaliação da disciplina de biologia de uma turma da terceira série do ensino médio de um colégio particular de São Paulo, mediante a análise das validades de conteúdo e de construto. A validade de conteúdo compreendeu a verificação psicométrica dos parâmetros de dificuldade, de discriminação e de acerto ao acaso dos itens, os quais foram concebidos e tratados pela Teoria de Resposta ao Item (TRI). A validade de construto se deu por meio da averiguação da consistência interna, através do coeficiente KR (α_{20}), e da função de informação da TRI. Os resultados da validação de conteúdo evidenciaram que os itens dos três instrumentos de avaliação apresentaram níveis de discriminação situados entre os níveis moderado e alto, considerados como adequados, e níveis de dificuldade próximos a 50%, classificados como apropriados para as habilidades dos alunos da turma analisada. Todos os itens dos três instrumentos de avaliação se enquadraram dentro dos valores padrão do índice de acerto ao acaso. No caso da validação de construto, as três avaliações se situaram no nível satisfatório da validade de construto e, portanto, no geral, todas as avaliações foram consideradas válidas.

PALAVRAS-CHAVE: Validade de testes. Avaliação de testes. Instrumentos de avaliação. Validade de conteúdo. Validade de construto.

INTRODUÇÃO

Bons instrumentos de avaliação e de medidas educacionais representam uma motivação para o aluno, desafiando e estimulando o seu interesse e a sua curiosidade intelectual. Quando bem construídos, esses instrumentos orientam os alunos sobre o estudo, guiando-os quanto à o que estudar e, principalmente, ao como estudar (VIANNA, 2014a). No entanto, a análise de instrumentos de avaliação do rendimento escolar releva que, "ao lado de alguns poucos que realmente demonstram medir aquilo que se propõem, existe, infelizmente, um número elevado de instrumentos que apresentam completa carência de requisitos técnicos" (VIANNA, 2014a, p. 106).

O autor argumenta que o problema da qualidade técnica dos instrumentos é grave, principalmente devido à influência que exercem no processo de aprendizagem. Independente do seu aspecto formal, observa-se que muitos dos instrumentos de avaliação "não orientam, mas sim conduzem o estudante a adotar comportamentos sem grande relevância educacional, ou seja, estimulam a aprendizagem do efêmero e do factual, e, assim, transformam-se num elemento de frustação para o estudante" (VIANNA, 2014a, p. 106–107).

A baixa qualidade técnica dos instrumentos de mensuração educacional está relacionada com a falta de formação técnica dos elaboradores dos itens e dos testes (VIANNA, 2014a). Segundo o autor, os problemas mais significativos são: a carência de confiabilidade e da subjetividade da maioria dos julgamentos sobre o rendimento educacional; a elaboração dos instrumentos, que quase sempre, é realizada às pressas e sobre pressão; ausência de validade de conteúdo; a ênfase na trivialidade; a estrutura formal inadequada dos itens, a qual impossibilita avaliar capacidades complexas; a ausência de diretrizes orientadoras no processo de elaboração dos itens; a limitação do uso de muitos instrumentos apenas para gerar uma nota final; o desconhecimento do erro de amostragem nos escores de um teste; e a ausência da verificação da análise estatística.

Nesse sentido, este estudo teve por objetivo analisar a validade dos instrumentos de avaliação da disciplina de biologia do terceiro trimestre do ano letivo de uma turma da terceira série do ensino médio de um Colégio particular de São Paulo. A escolha pelo terceiro trimestre justifica-se pelo fato de que este era um trimestre especialmente orientado para a revisão do conteúdo de biologia do ensino médio para os vestibulares. Além disso, foi o único trimestre no qual foi possível trabalhar com um sistema de correção de provas que permitiu a tabulação e posterior análise dos itens de maneira individual e conjunta.

FUNDAMENTAÇÃO TEÓRICA

A validade é um procedimento que verifica a capacidade de precisão da medida do fenômeno que está a ser estudado (ALEXANDRE; COLUCI, 2011). Um instrumento é considerado válido quando ele mede efetivamente o que ele foi construído para medir (PASQUALI, 2009a).

A validade de um instrumento de coleta de dados pode ser demonstrada por diversas técnicas. Essas técnicas podem ser agrupadas em três categorias, a saber:

validade de conteúdo, validade de construto e validade de critério (PASQUALI, 2009a).

A primeira categoria apresenta diversas definições (CUNHA; NETO; STACKFLETH, 2016). No entanto, de modo geral, a validade de conteúdo pode ser definida como a análise da capacidade do instrumento em reproduzir uma amostra representativa do universo de conteúdos ou comportamentos que estão sendo investigados (CRESWELL, 2015; CUNHA; NETO; STACKFLETH, 2016; PASQUALI, 2009a).

A validade de conteúdo tem fundamental importância no processo de elaboração do instrumento de avaliação. Segundo Haladyna (2004), é durante esse processo que os elaboradores de itens e de instrumentos definem o domínio e o conteúdo que será avaliado, garantindo que cada item que compõe o instrumento esteja associado diretamente à essa definição. Nesse sentido, é por meio da validação de conteúdo que um elaborador de testes é capaz de averiguar o quão bem os conteúdos foram amostrados pelos itens e, consequentemente, pelo instrumento (RUBIO et al., 2003).

Esse fato pode ser evidenciado em avaliações de desempenho escolar. De acordo com Pasquali (2009, p. 189), a validade de conteúdo "é aplicável quando se pode delimitar *a priori* e com clareza um universo de comportamentos, como é o caso em testes de desempenho, que pretendem cobrir um conteúdo delimitado por um curso programático específico".

Alguns autores caracterizam a validade de conteúdo apenas como uma avaliação realizada por um painel de especialistas. No entanto, outros a caracterizam como um procedimento que se inicia no desenvolvimento do instrumento e é finalizado com a análise teórica e empírica dos itens (ALEXANDRE; COLUCI, 2011).

Pasquali (2009) pertence ao segundo grupo e argumenta que para viabilizar um teste com tal validade, é necessário definir e planejar algumas especificações antes da elaboração dos itens. Essas especificações estão relacionadas com a definição do conteúdo, dos processos psicológicos a serem avaliados e pela determinação da representação de cada um dos conteúdos no instrumento (PASQUALI, 2009a).

Esse planejamento influencia diretamente a validade de conteúdo. Segundo Alexandre e Coluci (2011), essa etapa consiste na identificação dos domínios e na elaboração dos itens, na qual, "se organiza uma amostra representativa de conhecimentos, de processos cognitivos e de comportamentos" Raymundo (2009, p. 87).

Com relação à segunda categoria, a validade de construto, primeiramente é necessário definir o que se entende por construto. Vianna (2014b, p. 35) define construtos como

> traços, aptidões ou características supostamente existentes e abstraídos de uma variedade de comportamentos que tenham significado educacional (ou psicológico). Assim, fluência verbal, rendimento escolar, aptidão mecânica, inteligência, motivação, agressividade, entre outros, são construtos.

Nesse sentido, a validade de construto pode ser considerada como a verificação da qualidade de conversão de um conteúdo em uma realidade

operacional (TAHERDOOST, 2016), como por exemplo, um item ou um teste. De acordo com Cronbach e Meehl (1955, p. 282), ao analisar a validade de construto, "o problema enfrentado pelo pesquisador é: qual construto é responsável pela variação na performance do teste?".

Além disso, os autores alegam que

> validação de construto ocorre quando um investigador acredita que seu instrumento reflete um particular construto, ao qual estão ligados certos significados. A interpretação proposta gera hipóteses testáveis específicas, que são um meio de confirmar ou não confirmar a reivindicação (CRONBACH; MEEHL, 1955, p. 290).

Assim como a validade de conteúdo, a validade de construto é de grande relevância para a área da avaliação educacional. Segundo Vianna (2014b, p. 35), tal importância se deve ao fato de que "a avaliação se vale, frequentemente, de construtos, que, após sua definição operacional, são medidos por intermédio de testes".

O autor afirma também que, por meio da validade de construto, é possível esclarecer as diferenças encontradas nas pontuações do instrumento de avaliação (VIANNA, 2014b). Para isso, é necessário levar em consideração que este tipo de validade

> resulta do acúmulo, por diferentes meios, de várias provas, que precisam ser analisadas em todos os seus detalhes, a fim de constatar, entre outros aspectos, quais as variáveis com as quais os escores do teste se correlacionam, quais os tipos de itens que integram o teste, o grau de estabilidade dos escores sob condições as mais variadas e o grau de homogeneidade do teste, com vistas a ter elementos que possam esclarecer o significado do instrumento (VIANNA, 2014b, p. 36).

A validade de construto deve, portanto, ser determinada "pela concordância de medidas obtidas por métodos tão diferentes quanto possível" (VIANNA, 2014b, p. 39). Pasquali (2009, p. 185) corrobora a ideia, alegando que, "dado que a convergência de resultados das várias técnicas constitui garantia para a validade do instrumento", o ideal é que se utilize técnicas distintas para estimar a validade de construto.

Já no que diz respeito à terceira categoria, a validade de critério, esta consiste na correlação entre as pontuações do instrumento em questão e com aquelas de algum critério externo (SOUZA; ALEXANDRE; GUIRARDELLO, 2017). Entende-se pelo critério externo um instrumento ou medida largamente aceita e que apresente as mesmas características do instrumento de avaliação que pretende-se validar (SOUZA; ALEXANDRE; GUIRARDELLO, 2017).

Para que um instrumento seja considerado como válido nessa categoria, seus resultados devem corresponder aos resultados do critério. Independente do construto que está a ser analisado, essa validade será verdadeira apenas quando as pontuações do instrumento a ser validado corresponderem aos escores do critério escolhido (SOUZA; ALEXANDRE; GUIRARDELLO, 2017). Tal correspondência é realizada por meio do coeficiente de correlação (FERREIRA; MARQUES, 1998).

A validade de critério pode ser dividida em dois tipos de validade, a saber: validade concorrente e validade preditiva. Estas se distinguem, basicamente, pelo

tempo entre a coleta da informação pelo instrumento a ser avaliado e a coleta da informação sobre o critério a ser comparado (PASQUALI, 2009a). A validade concorrente é testada quando ambos os instrumentos de medida são aplicados dentro de um período curto de tempo, enquanto a validade preditiva é analisada pela correlação entre um instrumento aplicado no presente e um critério aplicado no futuro (FERREIRA; MARQUES, 1998; PASQUALI, 2009a).

A validade de critério evidencia, portanto, "até que ponto os valores obtidos pelo instrumento estão relacionados com uma medida de critério" (FERREIRA; MARQUES, 1998, p. 14). Contudo, esse procedimento enfrenta alguns obstáculos e, muitas vezes, não é passível de ser analisado. O primeiro motivo é o fato de que nem sempre existe um critério correlacionado à mesma área do conhecimento do instrumento a ser validado (FERREIRA; MARQUES, 1998; PASQUALI, 2009a; SOUZA; ALEXANDRE; GUIRARDELLO, 2017).

A segunda razão é a dificuldade em superar a eficiência do critério, pois, ao elaborar um novo instrumento, o pesquisador espera que este apresente vantagens quando comparado ao critério, seja pelo menor tempo de administração, menor custo ou maior facilidade de aplicação e utilização (SOUZA; ALEXANDRE; GUIRARDELLO, 2017). Quanto a esse fator, Pasquali (2009, p. 188) defende que a validade de critério "só faz sentido se existirem testes comprovadamente válidos que possam servir de critério contra o qual se quer validar um novo teste e que este novo teste tenha algumas vantagens sobre o antigo (como por exemplo, economia de tempo, etc.)".

METODOLOGIA

O estudo foi realizado durante o terceiro trimestre do ano letivo, no qual participaram 34 alunos de uma turma da terceira série do ensino médio de um Colégio particular de São Paulo. Desses, 19 (56%) pertenciam ao sexo feminino e 15 (44%) ao sexo masculino, com idades entre 16 e 18 anos.

Foram utilizados três instrumentos de avaliação para a coleta de dados, a saber: a avaliação mensal 1 (M1), a avaliação mensal 2 (M2) e a avaliação global (GB). Esses instrumentos continham 20, 20 e 40 itens, respectivamente. As duas primeiras avaliações foram aplicadas em um tempo disponível para 1 aula, já a terceira foi aplicada durante duas aulas. As avaliações mensais foram elaboradas com itens provenientes de testes de anos anteriores de vestibulares públicos e do ENEM, enquanto a avaliação global contou apenas com itens do ENEM.

O processo de verificação da validade dos instrumentos de avaliação foi realizado seguindo o proposto por Pasquali (2009). Tendo em vista que para a análise da validade de critério é necessário um critério exterior e que esse não estivesse presente neste estudo, foram analisadas apenas as validades de conteúdo e de construto.

De acordo com Pasquali (2009), a validade de conteúdo é composta por sete etapas: definição do domínio cognitivo; definição do universo do conteúdo; definição da representatividade do conteúdo; elaboração da tabela de especificação; construção do instrumento; análise teórica dos itens; e análise empírica dos itens.

Definição do domínio cognitivo

A definição do domínio cognitivo está relacionada com a definição dos processos psicológicos que serão avaliados (PASQUALI, 2009a). Segundo o autor, é importante adotar algumas das taxonomias de objetivos educacionais existentes para estabelecer os objetivos que se pretende medir.

Dentre as taxonomias existentes, a de Bloom é amplamente utilizada por professores para descrever e indicar objetivos cognitivos (RUSSEL; AIRASIAN, 2008). De acordo com Costa et al.(2018, pp. 3-4), a Taxonomia de Bloom possibilita

> estabelecer uma linguagem comum acerca dos objetivos de aprendizado, servir de base para o desenvolvimento de instrumentos de avaliação, estimular o desempenho dos alunos, incentivar educadores a auxiliarem aos seus alunos de forma estruturada e consciente a adquirirem competências específicas e determinar coerências entre os objetivos educacionais, as atividades e as avaliações nos currículos.

Para a construção dos três instrumentos em questão, foi utilizada a Taxonomia atualizada de Bloom, elaborada por Anderson et al. (2001). A seleção dos processos psicológicos levou em consideração aqueles que correspondem a níveis superiores, como: a aplicação dos conhecimentos para a resolução de problemas; a análise de situações cotidianas; e a avaliação de procedimentos realizados. Estes, segundo Russel e Airasian, (2008), representam domínios cognitivos de níveis superiores, já que exigem mais do que a simples recordação de informações.

Além disso, a fim de evitar itens que avaliassem apenas a memorização de fatos, optou-se por itens interpretativos, ou seja, itens que apresentam em seu enunciado um problema, uma situação ou um procedimento em forma de textos, imagens, tabelas ou gráficos e que devem ser analisados para que os alunos possam chegar à resposta do item (RUSSEL; AIRASIAN, 2008).

Definição do universo do conteúdo

A necessidade do estabelecimento do universo do conteúdo justifica-se pelo fato de que os itens que compõem o instrumento devem configurar uma amostra significativa do conteúdo programático (PASQUALI, 2009a). De acordo com o autor, este processo envolve dividir o conteúdo em unidades e subunidades de ensino.

Levando em consideração que o terceiro trimestre teve por objetivo revisar os conteúdos de biologia do ensino médio e preparar os alunos para o vestibular, a definição do universo do conteúdo dos três instrumentos de avaliação utilizados durante esse trimestre levaram em consideração os conteúdos que mais são cobrados nos vestibulares.

No caso das avaliações M1 e M2, por se tratar de avaliações pequenas e de duração de apenas uma aula, os conteúdos escolhidos compreenderam uma ou duas áreas do conhecimento dentro da biologia. Na avaliação M1, os conteúdos avaliados foram botânica e parasitologia. Em botânica, privilegiou-se a anatomia e a fisiologia vegetal, enquanto na parasitologia deu-se ênfase nas verminoses causadas por platelmintos e nematelmintos. Com relação à avaliação M2, a área

avaliada foi a fisiologia humana. Dentro desse universo, a ênfase foi dada nos sistemas circulatório, respiratório, nervoso e hormonal.

Já a avaliação GB, por se tratar de uma avaliação mais longa do que as demais e com um caráter de simulado final, permitiu a utilização de um universo de conteúdos maior do que as anteriores. Como ela foi baseada no ENEM, foram utilizados itens com os conteúdos que mais vezes aparecem nesse exame nacional, a saber: citologia, genética, ecologia, fisiologia, evolução, parasitologia, botânica e alterações ambientais. No caso dos conteúdos de fisiologia, parasitologia e botânica, pelo fato de já terem sido cobradas nas avaliações mensais, foram utilizados itens que avaliassem aspectos não representados pelos itens das avaliações anteriores.

Definição da representatividade do conteúdo

A definição da representatividade do conteúdo, definida pela proporção com que cada conteúdo deve ser representado no instrumento (PASQUALI, 2009a), foi estabelecida de acordo com a representatividade desses conteúdos nos testes dos vestibulares públicos dos quais os itens foram retirados.

Sendo assim, os 20 itens da avaliação M1 foram compostos por 6 itens de anatomia vegetal, 7 itens de fisiologia vegetal e 3 itens de verminoses causadas por platelmintos e 4 itens de verminoses causadas por nematelmintos. No caso dos 20 itens da avaliação M2, estes foram divididos em 4 itens para cada um dos sistemas fisiológicos, circulatório, respiratório, nervoso e hormonal. Já os 40 itens da avaliação GB foram distribuídos da seguinte maneira: 2 itens de citologia, 3 de genética, 7 de ecologia, 9 de fisiologia, 5 de evolução, 4 de parasitologia, 3 de botânica e 7 de alterações ambientais.

Elaboração da tabela de especificação

A tabela de especificação correlaciona as duas primeiras etapas do processo de validação de conteúdo, a da definição dos domínios cognitivos e do universo do conteúdo (PASQUALI, 2009a). Nesse sentido, esta tabela foi elaborada mediante a atribuição de correspondências entre estas duas etapas.

Construção do instrumento

A construção do instrumento de avaliação envolve, principalmente, o processo de elaboração dos itens que o constituirão. Normalmente, esta etapa compreende decisões relativas ao objetivo dos itens, ao seu formato, às técnicas de construção e à configuração do item (PASQUALI, 2009a).

Como os itens utilizados nas três avaliações foram retirados de vestibulares e exames anteriores, as técnicas de construção e sua configuração já estavam estabelecidas. No caso dos objetivos, foram escolhidos aqueles que se adequavam aos domínios cognitivos estabelecidos, ou seja, aplicação, análise e avaliação. E com relação ao formato, optou-se por itens objetivos compostos por 4 ou 5 opções de resposta.

Análise teórica dos itens

No que diz respeito à análise teórica dos itens, a qual consiste em uma verificação por parte de um painel de especialistas quanto à "representatividade dos itens em relação às áreas de conteúdo e à relevância dos objetivos a medir" (RAYMUNDO, 2009, p. 87), esta não foi realizada por este estudo. Os itens utilizados nos três instrumentos de avaliação foram obtidos dos testes de vestibulares e de exames de anos anteriores e, por isso, presume-se que já haviam passado por esta etapa.

Contudo, as três avaliações foram verificadas por outro professor de biologia, coordenador da área de Ciências Naturais, o qual analisou a pertinência dos itens com relação aos conteúdos e domínios cognitivos propostos pelo plano de ensino da disciplina.

Análise empírica dos itens

A análise empírica dos itens compreende a verificação de um conjunto de características dos itens, a qual é capaz de revelar se os itens avaliam de forma adequada o que eles se propõem a medir (PASQUALI, 2009a). Normalmente, nesta etapa são analisadas os seguintes parâmetros dos itens: dificuldade, discriminação e o acerto ao acaso (PASQUALI, 2009a).

Esses parâmetros foram concebidos e tratados pelo modelo logístico de 3 parâmetros da Teoria de Resposta ao Item (TRI), a qual pode ser utilizada para identificar "qual é a probabilidade e quais são os fatores que afetam esta probabilidade de cada item individualmente ser acertado ou errado (em testes de aptidão)"(PASQUALI, 2009b, p. 993).

Na TRI, para os modelos logísticos de 1 ou 2 parâmetros, o índice de dificuldade representa o nível de habilidade necessário para que a probabilidade de resposta correta a um item seja de 50% (ARAUJO; ANDRADE; BORTOLOTTI, 2009). Este parâmetro é nomeado como b e é definido pelo ponto, na escala de habilidade, onde a probabilidade de acertar e de errar o item é de 50% (PASQUALI, 2009a).

No entanto, dentro do modelo logístico de 3 parâmetros, "o parâmetro de dificuldade do item é o ponto na escala de habilidade onde: $P(\theta) = c + (1 - c)(.5) = (1 + c)/2$" (BAKER, 2001, p. 29). Ou seja, ao invés de utilizar uma escala na qual o 0 é a base e o 1 é o topo, este modelo utiliza uma escala onde a base é o valor de c, o valor do acerto ao acaso. Nesse sentido, "o parâmetro de dificuldade define o ponto na escala de habilidade onde a probabilidade de resposta correta é a metade do caminho entre essa base e 1,0" (BAKER, 2001, p. 30).

Quanto ao índice de discriminação, este é o parâmetro que descreve o quanto um item é capaz de diferenciar os respondentes que dominam dos que não dominam a habilidade estipulada pelo item em questão (PASQUALI, 2009a). Este índice é representado pela letra a e indica o ângulo de incidência da CCI no momento em que ela intercepta o ponto b, ou seja quando a probabilidade de responder corretamente o item é de 50% (PASQUALI, 2009a), já considerando o parâmetro c. Quanto maior for a inclinação do ângulo, maior será o poder de discriminação do item (BAKER, 2001).

Como dito anteriormente, o modelo logístico de 3 parâmetros inclui o acerto ao acaso. Este parâmetro, identificado como c, representa a "probabilidade de responder o item corretamente somente devido à adivinhação" (BAKER, 2001, p. 28). Vale ressaltar que o valor de c não varia em função do nível de habilidade e, por isso, alunos com habilidades mais altas e mais baixas possuem as mesmas probabilidades de acertar o item através da adivinhação (BAKER, 2001).

Para a verificação da validade de construto, foram utilizadas as análises da consistência interna e da função de informação da TRI, conforme proposto por Pasquali (2009).

Consistência interna

A consistência interna "avalia a consistência com que um determinado conjunto de itens de medida estima um determinado constructo ou dimensão latente" (MAROCO; GARCÍA-MARQUES, 2006, p. 70). Esta técnica permite dizer se os itens de um teste produzem pontuações consistentes (TANG; CUI; BABENKO, 2014) e auxilia "na definição do construto, especialmente ao indicar se o teste mede um único traço ou se, ao contrário, mede diversos traços" (VIANNA, 2014b, p. 40).

Além disso, a consistência interna é uma das técnicas empregadas para a análise da homogeneidade e da confiabilidade, ou fidedignidade, de um teste (CRONBACH, 1951; PASQUALI, 2009a; TANG; CUI; BABENKO, 2014). Sua análise compreende a verificação da correspondência entre todos os itens de um mesmo teste (PASQUALI, 2009a). Segundo o autor, a análise da consistência interna consiste no cálculo da correlação entre os itens do teste entre si ou com o escore total do teste.

Para a análise da consistência interna dos instrumentos de avaliação utilizou-se o coeficiente alfa de Kuder-Richardson 20 (α_{20}), o mais indicado para a análise de itens dicotômicos (PASQUALI, 2009a).

Função de informação da TRI

A função de informação da TRI identifica a qualidade do item para o construto, fornecendo dados sobre a precisão dos níveis de habilidade que estão sendo estimados pelo instrumento de medida (BAKER, 2001). Os dados resultantes dessa técnica revelam o quanto de informação um item é capaz de gerar para determinado valor de habilidade e, também, em qual valor de habilidade o item gera maior quantidade de informação (COUTO; PRIMI, 2011).

Utilizada como ferramenta para o processo de validação de construto, a função de informação da TRI assume-se como excelente método para descrever itens e testes, permitindo a seleção dos itens já que ela analisa a quantidade de informação para a medida da aptidão (PASQUALI, 2009a).

A função de informação da TRI pode ser trabalhada a nível do item e do teste, por meio das suas representações gráficas, a curva de informação do item e do teste, respectivamente. Neste trabalho, optou-se por utilizar a função de informação do teste, já que ela representa a soma das informações dos itens em

um determinado nível de habilidade, onde gera muito mais informação que a dos itens individuais (BAKER, 2001) e mostra para quais níveis de habilidade teta o teste é válido (PASQUALI, 2009a).

Tanto a fase empírica da validação de conteúdo como a validação de construto foram calculados utilizando o programa Microsoft Excel 2019, por meio do assistente EIRT.

ANÁLISE E DISCUSSÃO

Os resultados da análise da pesquisa serão apresentados em duas partes. Na primeira, constarão os dados provenientes da verificação da validade de conteúdo, mais precisamente da sua última etapa, a análise empírica dos itens e, na segunda parte, serão expostos os dados referentes à validade de construto.

Análise empírica dos itens

Os dados resultantes da análise empírica dos itens dos três instrumentos de avaliação estudados revelaram uma certa correspondência nos parâmetros de discriminação, dificuldade e acerto ao acaso. No caso da avaliação M1, os resultados evidenciaram uma média de 1,25, -0,29 e 0,17 para os parâmetros de discriminação, dificuldade e acerto ao acaso, respectivamente, assim como consta na Tabela 1.

Tabela 1 – Índices de discriminação, dificuldade e acerto ao acaso dos itens do instrumento de avaliação M1

Item	Discriminação (a)	Dificuldade (b)	Acerto ao acaso (c)
1	1,41	-0,49	0,16
2	0,47	-0,96	0,17
3	0,52	1,47	0,17
4	1,69	0,43	0,18
5	0,99	-3,01	0,17
6	0,27	4,24	0,2
7	0,8	0,89	0,18
8	1,22	2,65	0,15
9	0,77	-0,89	0,17
10	0,93	1,51	0,18
11	0,7	0,53	0,18
12	0,79	-0,46	0,17
13	1,16	-0,87	0,16
14	1,14	1,48	0,18
15	3,07	1,19	0,12
16	1,2	-2,63	0,17
17	1,14	-2,73	0,17
18	0,47	-7,31	0,17
19	3,1	0,4	0,18
20	3,15	-1,28	0,16
Média	**1,25**	**-0,29**	**0,17**

Fonte: Autoria própria (2021).

Os resultados da avaliação M2 expressaram, respectivamente, os valores médios de 1,14, 0,14 e 0,17 para os parâmetros de discriminação, dificuldade de acerto ao acaso, como mostra a Tabela 2 abaixo.

Tabela 2 – Índices de discriminação, dificuldade e acerto ao acaso dos itens do instrumento de avaliação M2

Item	Discriminação (a)	Dificuldade (b)	Acerto ao acaso (c)
1	2,39	-0,81	0,15
2	0,88	-0,36	0,18
3	1,22	-0,19	0,16
4	1,01	2,19	0,13
5	0,76	1,67	0,15
6	2,69	-0,08	0,18
7	0,44	1,95	0,18
8	2,91	-0,15	0,13
9	0,36	-0,83	0,17
10	0,36	0,05	0,17
11	1,06	-1,31	0,17
12	4,05	-1,3	0,15
13	0,74	5,29	0,2
14	1,19	1,25	0,13
15	0,34	-6,37	0,17
16	1,42	0,17	0,18
17	1,45	0,04	0,18
18	2,19	-0,91	0,17
19	3,16	-0,85	0,16
20	0,79	3,29	0,18
Média	**1,47**	**0,14**	**0,17**

Fonte: Autoria própria (2021).

Já no que diz respeito à avaliação GB, seus itens apresentaram valores médios de discriminação, dificuldade e acerto ao acaso de 1,17, -0,88 e 0,16, respectivamente, assim como está exposto na Tabela 3.

Tabela 3 – Índices de discriminação, dificuldade e acerto ao acaso dos itens do instrumento de avaliação GB

Item	Discriminação (a)	Dificuldade (b)	Acerto ao acaso (c)
1	0,62	1,9	0,17
2	0,79	0,69	0,17
3	0,85	-2,77	0,17
4	0,94	-0,31	0,17
5	1	-1,75	0,17
6	1,47	-0,14	0,16
7	0,65	-2,5	0,17
8	4,7	-0,16	0,16
9	2,07	0,1	0,15
10	1,84	-0,76	0,16
11	0,38	-0,32	0,17
12	0,7	-0,19	0,17
13	0,7	-1,44	0,16

Item	Discriminação (a)	Dificuldade (b)	Acerto ao acaso (c)
14	1,74	-0,17	0,14
15	0,82	1,06	0,16
16	0,32	1,56	0,18
17	0,66	-1,74	0,17
18	1,99	1,21	0,11
19	1,3	0,37	0,17
20	1,91	-1,23	0,16
21	0,53	-1,14	0,17
22	1,04	0,18	0,16
23	1,05	-2,37	0,17
24	2,19	-1,59	0,16
25	0,79	-2,5	0,17
26	1,11	-1,01	0,16
27	0,43	-3,11	0,17
28	0,8	-0,36	0,17
29	2,46	0,27	0,16
30	1,1	-2,26	0,17
31	0,5	-5,29	0,17
32	0,86	0,59	0,16
33	0,86	0,58	0,16
34	0,7	0,43	0,16
35	0,2	-8,87	0,17
36	0,86	0,26	0,17
37	3,32	0,36	0,16
38	1,32	0,87	0,15
39	0,81	-4,39	0,17
40	0,57	0,61	0,17
Média	**1,17**	**-0,86**	**0,16**

Fonte: Autoria própria (2021).

No que diz respeito aos valores numéricos do índice de discriminação, Baker (2001) argumenta que, teoricamente, esses valores podem variar de -∞ a +∞. O autor, apresenta o seguinte referencial para os intervalos de discriminação: 0 - não discriminativo; de 0,1 a 0,34 – muito baixo; de 0,35 a 0,64 – baixo; de 0,65 a 1,34, moderado; de 1,35 a 1,69 – alto; e acima de 1,7 – muito alto. De acordo com esses valores referenciais, observa-se que o poder de discriminação médio dos itens da avaliação M1 é moderado (1,25), da M2 é alto (1,47) e da GB é moderado (1,17).

Constata-se, portanto, que o índice médio de discriminação dos três instrumentos varia de moderado a alto. Além disso, os dados revelam que as avaliações M1, M2 e GB apresentaram apenas 20% – 4 (20%), 4 (20%) e 8 (20%), respectivamente – de itens com poder de discriminação baixo, ou seja, valores abaixo de 0,64.

Tendo em vista que o índice de discriminação é parâmetro responsável por identificar sujeitos com habilidades distintas, infere-se que, em geral, os três instrumentos foram capazes de realizar tal função, diferenciando os alunos que dominavam daqueles que não dominavam a habilidade requerida pelos itens das avaliações.

Com relação ao parâmetro da dificuldade dos itens dos instrumentos, ou seja, o ponto na escala de habilidades onde a probabilidade de resposta correta é de (1+*c*)/2 para o modelo de três parâmetros, "a maneira correta de interpretar um valor numérico do parâmetro de dificuldade do item é em termos de onde o item se posiciona na escala de habilidades" (BAKER, 2001, p. 34).

A análise dos dados revelou que a posição média dos valores dos índices de dificuldade dos itens, em uma escala de -3 a +3, das avaliações M1, M2 e GB foi de -0,29, 0,14 e -0,88, respectivamente. De acordo com Ferreira (2018), a dificuldade dos itens pode ser categorizada em cinco níveis: muito fácil (< -1.28), fácil (-1.27 a -0.52), médio (-0.51 a 0.51), difícil (0.52 a 1.27) e muito difícil (> 1.28). Nesta perspectiva, o índice de dificuldade médio dos itens das avaliações M1 e M2 é classificado na categoria médio, enquanto o da GB na categoria fácil.

Comparando esses valores com os valores médios das habilidades dos alunos nessas mesmas avaliações, medidos pela TRI, os quais se aproximaram de 0 – M1 = 0,001; M2 = 0,0001 e GB = 0,07 –, percebe-se que, todas as outras avaliações apresentaram, em média, itens com valores de dificuldade condizentes com a habilidade dos alunos que realizaram a avaliação.

Ademais, Pasquali (2009) afirma que, devido ao fato de 50% dos alunos acertarem e 50% errarem, itens com índices de dificuldade em torno de 50% são os itens que mais geram informação para o elaborador do teste. Tendo em vista que na escala da TRI os 50% encontram-se no valor 0, verifica-se que a média do parâmetro de dificuldade das três avaliações se enquadra dentro dos padrões ideais para a dificuldade do item.

Relativamente ao que concerne o índice de acerto ao acaso, o qual revela a probabilidade um aluno com baixa habilidade responder corretamente a um item, sua escala vai de 0 a 1, ou seja, 0 < = c < = 1.0. Valores acima de 0,35 não são considerados aceitáveis, nesse sentido, a escala utilizada como base de interpretação para esse parâmetro foi de 0 a 0,35, no caso 0 < = c < = 0,35 (BAKER, 2001).

Levando em consideração esse critério, nota-se que a média de c nas três avaliações apresenta-se abaixo de 0,35 e não superiores a 0,20. Isso mostra que, mesmo se fossem utilizados os critérios de valores máximos iguais a 0,25 e 0,20, referentes à probabilidade de acertar um item com quatro ou cinco opções de resposta, respectivamente, todos os itens estariam enquadrados dentro dos valores padrão.

Análise da validade de construto

A primeira técnica utilizada para a análise da validade de construto foi a consistência interna, mediante a determinação do coeficiente KR (α_{20}). Os resultados revelaram valores de consistência interna de 0,53 para a avaliação M1, 0,70 para a M2 e 0,80 para a GB.

Souza et al. (2017) argumentam que, embora o coeficiente alfa seja o mais utilizado para a avaliação da consistência interna, não existe um consenso quanto à sua interpretação. Cronbach (1951) alega que quando os valores se encontram acima de 0,70, são considerados aceitáveis e, quando acima de 0,80, são bons. Por

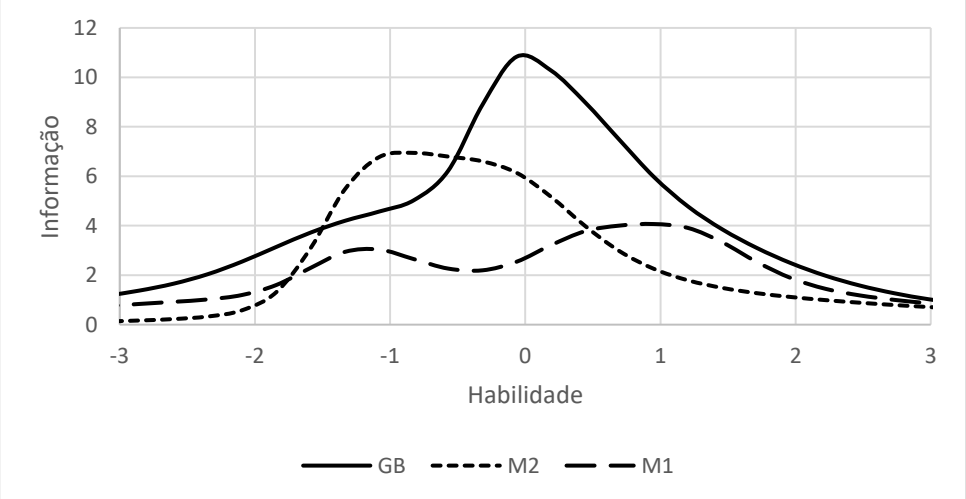
outro lado, a maioria dos estudos determinam como ideais valores de α acima de 0,70 e alguns pesquisadores consideram valores abaixo de 0,70, mas próximos de 0,60, como satisfatórios (SOUZA; ALEXANDRE; GUIRARDELLO, 2017).

De acordo com essas classificações, observa-se que os valores de alfa das avaliações M2 e GB encontram-se acima do padrão aceitável, enquanto a da avaliação M1 pode ser considerada como satisfatório, ou próximo desse enquadramento, apesar de expressar um alfa inferior a 0,70.

Segundo Souza et al. (2017), os valores de alfa são altamente influenciados pelo número de itens do instrumento de avaliação. Sendo assim, esses valores encontrados justificam-se, possivelmente, pelo número de itens e, também, pela quantidade de construtos medidos em cada avaliação. O fato da avaliação M1 ter sido constituída por 13 itens do construto de botânica e 7 itens do construto de parasitologia, diferente dos 20 itens de fisiologia da M2 e dos 40 itens de biologia geral da GB que, embora apresente diversos construtos, possui o dobro de itens das outras avaliações, pode ter sido o motivo para o baixo valor de consistência interna da avaliação M1 e altos valores das avaliações M2 e GB.

Por fim, a análise das curvas de informação dos testes revelou que a avaliação M1 gera maior informação para os intervalos de habilidades em torno de -1,08 e 0,95, sendo o segundo intervalo aquele que produz em maior quantidade. Com relação às avaliações M2 e GB, as curvas de informação evidenciaram que a maior quantidade de informação é proveniente dos intervalos dos níveis de habilidade em torno de -0,8 e -0,06, respectivamente. As curvas de informação estão expostas no Gráfico 1.

Gráfico 1 - Função de informação total dos instrumentos de avaliação



Fonte: Autoria própria (2021).

Levando em consideração a média das habilidades dos alunos nas avaliações M1, M2 e GB foram, respectivamente, 0,001, 0,0001 e 0,07, observa-se que a avaliação GB foi a que melhor se adequou nas habilidades dos alunos, pois a maior quantidade de informação (10,8) se localizou na habilidade -0,06. Com relação à avaliação M2, apesar do pico da curva se localizar na habilidade -0,8, ela apresenta uma variação de informação de apenas -0,8 com relação à habilidade 0. Já no caso da avaliação M1, verifica-se que a curva possui dois ápices (-1,8 e 0,95), os quais

encontram-se um pouco mais afastados da média de habilidade 0,001 apresentada pelos alunos, revelando que essa avaliação gera maiores informações para outros níveis de habilidade.

Além disso, é notória a diferença entre a quantidade de informação gerada pelas avaliações. A avaliação GB é a que mais gera informação, seguida pelas avaliações M2 e M1. Assim como na consistência interna, esses resultados podem estar relacionados com a quantidade de itens e com os construtos que compuseram as avaliações.

CONSIDERAÇÕES FINAIS

A fim de analisar a validade dos instrumentos de avaliação aplicados a uma turma da terceira série do ensino médio de um colégio particular de São Paulo no último semestre do ano letivo, as avaliações foram submetidas às análises da validação de conteúdo e de construto.

A validação de conteúdo, mais especificamente da análise empírica dos itens realizada por meio da TRI, revelou que o poder de discriminação dos itens das três avaliações se enquadrou dentro do intervalo de discriminação entre moderado e alto, sendo considerados como adequados. O índice de dificuldade também se revelou adequado para os três instrumentos de avaliação, todos apresentaram valores próximos a 50% e, portanto, geram mais informação e se adequaram às habilidades dos alunos. O mesmo ocorreu para o parâmetro acerto ao acaso, cujos índices dos itens das três avaliações se mostraram abaixo dos 0,20 e, portanto, dentro dos padrões ideais.

Relativamente à validação de construto, as análises da consistência interna, verificadas pelo coeficiente alfa de Kuder-Richardson 20 (α_{20}) e da função de informação da TRI evidenciaram que as avaliações M2 e Global são confiáveis e adequadas. No que diz respeito à M1, o valor da consistência interna apresentou-se abaixo do ideal, mas ainda próximo do índice considerado satisfatório e, na análise da função de informação da TRI não se mostrou tão eficiente quanto às outras duas na finalidade de gerar mais informações para a habilidade dos alunos avaliados.

Entende-se que, no geral, os três instrumentos de avaliação foram considerados válidos. A quantidade de itens e de construtos por avaliação foram os fatores que, provavelmente, geraram as diferenças notadas nos dados psicométricos resultantes do processo de validação de construto. Sugere-se, portanto, que para avaliações futuras, o número de itens seja maior ou que as avaliações envolvam apenas um construto.

Recomenda-se, também, que estudos futuros utilizem dados de todos os instrumentos de avaliação aplicados no decorrer do ano letivo por determinada disciplina. Dessa forma, será possível coletar dados mais robustos, realizar análises mais amplas e gerar resultados sobre a validade de todos os instrumentos utilizados na avaliação dos alunos.

Content and construct validities analysis of biology assessments instruments

ABSTRACT

The aim of this work was to analyze the validity of three classroom assessment instruments of the subject of biology a third grade of high school class in a private school in São Paulo, through an analysis of the content and construct validities. The content validity included the psychometric analysis of the parameters of difficulty, discrimination and random matching of the items, which were conceived and processed by the Item Response Theory (IRT). The construct validity was achieved through the verification of internal consistency, through the KR coefficient (α20), and the information function of the IRT. The results of the content validation showed that the items of the three classroom assessment instruments showed levels of discrimination between the moderate and high levels, considered as adequate, and levels of difficulty close to 50%, classified as appropriate for the latent skills of the students in the class analyzed. All items of the three assessment instruments were within the standard values of the random matching index. In the case of the construct validation, the three assessments were at the satisfactory level of construct validity and, therefore, in general, all assessments were considered valid.

KEYWORDS: Test validity. Test evaluation. Assessment tools. Content Validity. Construct validity.

AGRADECIMENTOS

Este trabalho é financiado por fundos nacionais através da FCT – Fundação para a Ciência e a Tecnologia, I.P., no âmbito da Bolsa de Investigação com referência UI/BD/151034/2021 e do Projeto UIDB/4312/2020.

REFERÊNCIAS

ALEXANDRE, N. M. C.; COLUCI, M. Z. O. Validade de conteúdo nos processos de construção e adaptação de instrumentos de medidas. **Ciência & Saúde Coletiva**, v. 16, n. 7, p. 3061–3068, 2011.

ANDERSON, L. W. et al. **A taxonomy for learning, teaching and assessing: a revison of Bloom's taxonomy of educational objectives**. New York: Addison Wesley Longman, 2001.

ARAUJO, E. A. C. DE; ANDRADE, D. F. DE; BORTOLOTTI, S. L. V. Teoria da resposta ao item. **Revista da Escola de Enfermagem USP**, v. 3, n. especial, p. 1000–1008, 2009.

BAKER, F. B. **The basics of item response theory**. Washington, DC: ERIC, 2001.

COSTA, G. O. F. et al. Taxonomy of educational objectives and learning theories in the training of laparoscopic surgical techniques in a simulation environment. **Revista do Colégio Brasileiro de Cirurgiões**, v. 45, n. 5, p. 1954, 2018.

COUTO, G.; PRIMI, R. Teoria de resposta ao item (TRI): Conceitos elementares dos modelos para itens dicotômicos. **Boletim de Psicologia**, v. 62, n. 134, p. 1–15, 2011.

CRESWELL, J. **Educational research**: planning, conducting, and evaluating quantitative and qualitative research. New York: Pearson, 2015.

CRONBACH, L. J. Coefficient alpha and the internal structure of tests. **Psychometrika**, v. 16, p. 297–334, 1951.

CRONBACH, L. J.; MEEHL, P. E. Construct validity in psychological tests. **Psychology Bulletin**, v. 52, n. 4, p. 281–302, 1955.

CUNHA, C. M.; NETO, O. P. DE A.; STACKFLETH, R. Principais métodos de avaliação psicométrica da validade de instrumentos de medida. **Rev. Aten. Saúde**, v. 14, n. 47, p. 75–83, 2016.

FERREIRA, E. A. **Teoria de resposta ao item** – TRI: análise de algumas questões do ENEM - habilidades 24 a 30. Dissertação de Mestrado, Universidade Federal da Grande Dourados, Mato Grosso do Sul, 2018.

FERREIRA, P. L.; MARQUES, F. B. Avaliação psicométrica e adaptação cultural e linguística de instrumentos de medição em saúde: princípios metodológicos

gerais. **Centro de Estudos e Investigação em Saúde da Universidade de Coimbra**, p. 0–24, 1998.

HALADYNA, T. M. **Developing and validating multiple-choice test items**. 3 ed. London: Lawrence Erlbaum Associates, 2004.

KRASILCHIK, M. **Práticas do ensino de biologia**. 4 ed. São Paulo: EDUSP, 2016.

MAROCO, J.; GARCÍA-MARQUES, T. Qual a fiabilidade do alfa de Cronbach? Questões antigas e soluções modernas? **Laboratório de Psicologia**, v. 4, n. 1, p. 65–90, 2006.

OLIVEIRA, C. B. C. DE; VALLE, M. G. DO; AVELAR, B. Y. S. Concepções de Professores de Biologia sobre Avaliação: um estudo de Caso. **Revista Meta: Avaliação**, v. 10, n. 28, p. 29, 2018.

PASQUALI, L. **Psicometria teoria dos testes na psicologia e na educação**. 4 ed. Petrópolis: Vozes, 2009a.

PASQUALI, L. Psicometria. **Revista da Escola de Enfermagem da USP**, v. 43, n. spe, p. 992–999, 2009b.

RAYMUNDO, V. P. Construção e validação de instrumentos: um desafio para a psicolinguística. **Letras de Hoje**, v. 44, n. 3, p. 86–93, 2009.

RUBIO, D. M. G. et al. Objectifyng content validity: Conducting a content validity study in social work research. **Social Work Research**, v. 27, n. 2, p. 94–104, 2003.

RUSSEL, M. K.; AIRASIAN, P. W. **Classroom assessment**: concepts and applications. New York: McGrall-Hill, 2008.

SILVA, L. M.; BEZERRA, M. L. DE M. B. Instrumentos de avaliação na disciplina de biologia: identificação, reflexão e ações do PIBID. I Congresso de Inovação Pedagógica em Arapiraca. **Anais...** Arapiraca: 2015. Disponível em: www.seer.ufal.br/index.php/cipar/article/download/1949/1449.

SOUZA, A. C. DE; ALEXANDRE, N. M. C.; GUIRARDELLO, E. DE B. Propriedades psicométricas na avaliação de instrumentos: avaliação da confiabilidade e da validade. Epidemiologia e Serviços de Saúde: **Revista do Sistema Único de Saúde do Brasil**, v. 26, n. 3, p. 649–659, 2017.

TAHERDOOST, H. Validity and reliability of the research instrument; How to test the validation of a questionnaire/survey in a research. **International Journal of Academic Research in Management**, v. 5, n. 3, p. 28–36, 2016.

TANG, W.; CUI, Y.; BABENKO, O. Internal consistency: do we really know what it is and how to assess it? **Journal of Psychology and Behavioral Science**, v. 2, n. 2, p. 205–220, 2014.

VIANNA, H. M. Qualificação técnica e construção de instrumentos de medida educacional. **Estudos em Avaliação Educacional**, v. 25, n. 60, p. 106–117, 2014a.

VIANNA, H. M. Validade de construto em testes educacionais. **Estudos em Avaliação Educacional**, v. 25, n. 60, p. 136–152, 2014b.

VIEIRA, N. N. **As provas das quatro áreas do ENEM vista como prova única na ótica de modelos da teoria da resposta ao item uni e multidimensional**. Dissertação de Mestrado, Universidade Federal de Santa Catarina, Florianópolis, 2016.

Recebido: 14 jan. 2021
Aprovado: 06 mai. 2021
DOI: 10.3895/actio.v6n2.13715
Como citar:
COPPI, M. A. Análise das validades de conteúdo e de construto de instrumentos de avaliação de biologia. **ACTIO**, Curitiba, v. 6, n. 2, p. 1-19, mai./ago. 2021. Disponível em: <https://periodicos.utfpr.edu.br/actio>. Acesso em: XXX.
Correspondência:
Marcelo Alves Coppi
Rua da Barba Rala, n. 1, Edifício B, Évora, Évora, Portugal.