 https://doi.org/10.1590/S0104-40362022003002362

ARTIGO

A questão da validade na avaliação educacional brasileira*

Girlene Ribeiro de Jesus [a]
Renata Manuelly de Lima Rêgo [b]
Victor Vasconcelos de Souza [c]

Resumo

O objetivo do presente estudo foi analisar como o atual conceito e as fontes de evidências de validade são utilizados no âmbito da avaliação educacional em larga escala no Brasil. Para tanto, foi realizada uma análise documental, por meio dos relatórios oficiais das maiores avaliações educacionais nacionais aplicadas no país. Os resultados obtidos indicaram que não são apresentadas fontes de evidências de validade suficientes para que se julgue que os testes educacionais analisados sejam válidos para o que se propõem. Ainda é necessário desenvolver o conhecimento técnico dos pesquisadores brasileiros que realizam estudos de validação de testes educacionais, especialmente no âmbito das informações que devem ser levantadas e apresentadas a fim de que se possam utilizar com confiança os resultados obtidos nas provas e nos testes educacionais aplicados no país.

Palavras-chave: Validade. Avaliação Educacional. Padrões para Testagem.

1 Introdução

A validade deve ser uma preocupação central quando se trata da construção de medidas educacionais. Por esse motivo, é preocupante a ausência de relatos de estudos de validação técnica e cientificamente rigorosos nos relatórios que tratam da qualidade dos testes, dos indicadores e dos questionários utilizados para a avaliação da Educação Básica e Superior no Brasil. De acordo com a literatura acerca da construção de medidas, a propriedade mais importante que uma medida educacional precisa ter é a validade (HALADYNA; RODRIGUES, 2013; PASQUALI, 2011;

* O presente trabalho é fruto de um intercâmbio internacional, realizado nos Estados Unidos, no *Educational Testing Service* (ETS). Contou com o apoio financeiro do Centro Brasileiro de Pesquisa em Avaliação e Seleção e de Promoção de Eventos (Cebraspe).

[a] Universidade de Brasília, Brasília, DF, Brasil.

[b] Centro Brasileiro de Pesquisa em Avaliação e Seleção e de Promoção de Eventos, Brasília, DF, Brasil.

[c] Universidade de Brasília, Brasília, DF, Brasil.

Recebido em: 02 jun. 2019
Aceito em: 28 out. 2021

SIRECI, 2013). Se os resultados obtidos em determinada medida educacional não apresentarem evidências suficientes para serem considerados válidos, todas as demais condutas e julgamentos que se baseiem naquela medida podem ser questionados do ponto de vista científico.

No Brasil, a discussão sobre validade como uma qualidade imprescindível em testes, em questionários e em indicadores já é presente, e, de certa forma, sistematizada, no campo da psicologia (veja, por exemplo a Resolução nº 002/2003 do Conselho Federal de Psicologia). Entretanto, a ausência de uma discussão mais compreensiva sobre as evidências de validade nos relatórios das avaliações educacionais brasileiras indica que a preocupação com esse tema na Educação ainda é incipiente, e não há resolução alguma sobre a qualidade desses testes que, muitas vezes, possuem enorme impacto na vida de indivíduos, nas escolas, nas universidades, e, em última análise, na sociedade.

Para compreender melhor o estado atual desse conceito, é necessário retomar a sua evolução histórica. O conceito de validade é central na psicometria e se expandiu gradualmente nos últimos 120 anos. O interesse nesse tema é tal que sua história é detalhada em artigos seminais da área (KANE, 2013; MESSICK, 1989; SIRECI; SUKIN, 2013). Para o cumprimento do objetivo desse artigo, revisaremos os trabalhos que tiveram maior impacto ao longo desse tempo.

1.1 Evolução histórica do conceito de validade

No início do século XX, quando os primeiros testes educacionais e psicológicos, baseados numa visão moderna de ciência, surgiram, duas definições de validade representavam duas grandes correntes de epistemologia da época (SIRECI, 2006). A primeira definia validade como o grau em que o teste mede o que se propõe (SMITH; WRIGHT, 1928) e a segunda definição afirmava que "um teste é válido para qualquer coisa com a qual se correlaciona" (GUILFORD, 1946, p. 429).

Essas duas definições iniciaram a discussão acadêmica sobre o que era validade e como ela deveria ser avaliada (SIRECI, 2016). A primeira definição levou a estudos baseados na análise fatorial, que buscou descobrir se os fatores provenientes da análise fatorial estavam de acordo com a teoria que fundamentou a construção da medida. A segunda definição promoveu estudos focados em estimar a relação entre os escores de um novo teste com outras medidas já testadas (ou outro critério, tais como desempenho ou idade).

Não obstante, Pressey (1920) observou que os esforços para validar testes com base em análises estatísticas estavam incompletos, pois a composição do teste

(seu conteúdo, por exemplo) estava sendo ignorada. Nessa época, em que a revolução promovida pela maturação da ciência moderna inibia o surgimento de métodos mais qualitativos, a volta da ideia de que a validade se referia mais do que à análise estatística foi um importante passo no desenvolvimento de uma teoria da validade mais complexa e nuançada. Esse autor apontou as limitações de uma abordagem puramente estatística e enfatizou a necessidade de emprego de uma conceituação mais ampla, que focasse em como os escores dos testes eram usados. Rulon (1946, p. 290) trouxe essa mesma preocupação, afirmando que "não podemos rotular um teste válido ou não válido, exceto para algum propósito".

Em comparação com a definição de Smith e Wright (1928), as noções de validade defendidas por Pressey (1920) e Rulon (1946) ajustam-se melhor com a definição contemporânea. Atualmente, a validade é definida como "o grau com que a evidência e a teoria apoiam as interpretações dos escores do teste para um propósito específico" (AERA; APA; NCME, 2014; MESSICK, 1989). Esse foco no propósito, no entanto, não é contrário à importância de estudos acerca do construto, baseados em análise fatorial, ou a estudos que investigam a correlação do teste com outras variáveis. Sireci (2016) enfatiza que essas definições anteriores permanecem importantes e necessárias; elas não foram anuladas ou substituídas pela teoria moderna; elas foram expandidas.

Em relação à velha definição de que validade é o grau em que o teste mede o que se propõe, a grande diferença da teoria contemporânea seria o reconhecimento de que os escores dos testes são utilizados para fins específicos, e é a relação do teste com esses usos que precisam ser validados, não o teste em si (SIRECI, 2016). Essa distinção revela uma preocupação com um aspecto mais prático da avaliação, e com as consequências do uso dos resultados dos testes. Além disso, reserva a possibilidade de se julgar um determinado escore inválido por motivos outros que não ao teste em si, como imprevistos no momento da aplicação, ou inadequação do teste para um determinado indivíduo ou amostra.

O documento utilizado como referência em todo o mundo na definição dos princípios constituintes da validade são os *Standards for Educational and Psychological Testing* (AERA; APA; NCME, 2014). Este livro é uma publicação elaborada por importantes associações americanas, em colaboração com os principais pesquisadores da área de Educação e de psicologia em todo o mundo: a *American Educational Research Association* (AERA – Associação Americana de Pesquisa Educacional), a *American Psychological Association* (APA – Associação Americana de Psicologia) e o *National Council on Measurement in Educational* (NCME – Conselho Nacional sobre Medida Educacional). O modelo apresentado

nos *Standards* é amplamente reconhecido como uma declaração do consenso profissional sobre padrões para testes devido à forma como os padrões foram desenvolvidos e aprovados pela comunidade acadêmica (LINN, 2006).

Segundo Sireci (2013), os *Standards* não representam as ideias de um único teórico da validade, mas apresentam o consenso de três organizações que promovem diretrizes sobre o desenvolvimento e uso de testes há mais de 50 anos. Para psicometristas e outros profissionais envolvidos no processo de desenvolvimento de testes, os *Standards* "são autoritativos; eles são as leis que governam nossa percepção coletiva de prática adequada; em certo sentido, eles representam as leis da nossa profissão" (SIRECI; PARKER, 2006, p. 1). Isso não significa que não haja discordâncias acerca de afirmações realizadas nos *Standards*, no entanto, essas discordâncias vão mais no sentido de ampliar o sentido de validade do que de se opor a essa perspectiva (por exemplo: NEWTON, 2016).

Atualmente, os *Standards* já foram revisados cinco vezes e a preocupação desde a primeira versão tem sido a mesma: a promoção de práticas adequadas para o desenvolvimento de testes e de interpretações dos seus escores com o objetivo de utilizar medidas mais justas e de responsabilizar as instituições responsáveis (LINN, 2006). Desde a primeira versão, com o nome de *Technical recommendations for psychological tests and diagnostic techniques* (APA, 1954), o conceito de validade é abordado, e, com o passar dos anos, o diálogo com a literatura de vanguarda possibilitou a evolução da visão adotada. O Quadro 1 apresenta a forma como a validade foi compreendida e como se constituíam as recomendações para o estudo dela em cada uma das versões.

Quadro 1 – A visão de validade nas diferentes versões dos *Standards*

Versão	Visão de validade	Estudo da validade
1. APA (1954)	Conceituação fragmentada	Quatro tipos de validade: 1. Conteúdo 2. Preditiva 3. Concorrente 4. Construto
2. APA (1966) 3. APA; AERA; NCME, (1974)	Conceituação fragmentada	Três tipos de validade: 1. Conteúdo 2. Relacionada ao critério 3. Construto Substituiu o termo "tipos" para "aspectos" da validade; e validade preditiva e concorrente foram combinadas e chamadas de validade relacionada ao critério.

Continua

Continuação

Versão	Visão de validade	Estudo da validade
4. AERA; APA; NCME (1985)	Afastou-se da visão fragmentada e afirmou que validade é um conceito unitário e está relacionada à adequação, significado e utilização das inferências feitas a partir dos resultados de um teste.	Três categorias de validade: 1. Validade de conteúdo 2. Validade de critério 3. Validade de construto
5. AERA; APA; NCME (1999) 6. AERA; APA; NCME (2014)	Adota a visão da validade como um conceito unitário e que a interpretação do escore e o uso do teste são inseparáveis.	Cinco fontes de evidências de validade: 1. Evidências baseadas no conteúdo 2. Evidências baseadas nas relações com variáveis externas 3. Evidências baseadas na estrutura interna 4. Evidências baseadas no processo de resposta 5. Evidências baseadas nas consequências da testagem A partir da versão de 1999, não se trata mais de "tipos" de validade, mas de evidências de validade.

Fonte: Elaborado pelos autores (2018)

Em consonância com a literatura da área, a primeira versão dos *Standards* especificou quatro tipos de validade: conteúdo, preditiva, concorrente e de construto (APA, 1954). É possível observar que a definição tradicional que prevalecia na literatura acadêmica da época era de que a validade era "o grau em que o teste mede aquilo que se propõe a medir" e formava a base dos estudos de validade de construto; já a segunda definição, "um teste é válido para qualquer coisa com a qual se correlaciona", era a fundação dos estudos de validade preditiva e concorrente.

A segunda (APA, 1966) e a terceira versões (APA; AERA; NCME, 1974) dos *Standards* adotaram ainda uma visão fragmentada da validade, e os tipos de validade estavam relacionados aos objetivos específicos dos testes: determinar como o indivíduo atua em um universo de situações (conteúdo), prever o futuro do indivíduo quanto a um critério (relacionado ao critério) e inferir a quantidade de um traço que o indivíduo possui (construto) (MESSICK, 1989). Apesar de adotar ainda uma visão fragmentada da validade, a edição de 1966 deu o primeiro passo para abandonar essa concepção, por meio da sugestão de que a validade seria mais bem interpretada como uma qualidade uníssona. A respeito disso, os autores tiveram a dizer: "estes três aspectos da validade são apenas conceitualmente independentes e raramente um deles é mais importante em uma

situação particular; um estudo completo de um teste normalmente envolveria todos os tipos de validade" (APA, 1966, p. 14).

Messick (1980, 1981) explica que a visão fragmentada da validade pode levar à confusão no estudo e no entendimento do conceito, de modo que um tipo de validade poderia ser entendido como um fim em si mesmo, e não apenas uma evidência de validade, que, na verdade, era uma qualidade única. De fato, Messick (1980, 1981) previu uma realidade que vigora ainda hoje.

Kane (2013) aponta que, com o passar dos anos, uma série de "validades" específicas foram desenvolvidas com a finalidade de avaliar interpretações e usos dos escores em situações específicas. Esses desenvolvimentos enriqueceram muito a concepção de validade adotada hoje, mas também tenderam a fragmentar o conceito, adicionando tipos e mais tipos de validade. Um exemplo dessa realidade pode ser constatado no levantamento realizado por Newton e Shaw (2013). Os autores analisaram 22 periódicos científicos que publicaram artigos entre 2005 e 2010 acerca das evidências de validade de medidas educacionais e psicológicas. Eles analisaram 208 artigos e identificaram que a visão da validade como conceito unitário ainda não foi absorvida completamente e muitos artigos ainda utilizam os "tipos" de validade: 29,3% dos trabalhos nomearam a análise como validade de construto, 13% utilizaram o termo validade incremental, 10,6% utilizaram validade preditiva, e assim por diante.

A grande contribuição da versão de 1974, no entanto, não foi acerca do conceito de validade, mas sobre uma sistematização de padrões para o uso de testes, explicando explicitamente a preocupação com o viés, com os abusos e com as consequências sociais advindas do mau uso de testes educacionais e psicológicos (MESSICK, 1989). A versão de 1985, em relação à concepção de validade, continuou na marcha em direção à visão unificada: afirma categoricamente que a validade é um conceito uníssono e substitui a expressão "tipos de validade" por "categorias de evidências de validade relacionadas ao conteúdo, ao critério e ao construto" (MESSICK, 1989). Assim, a versão de 1985 sustenta que "uma validação ideal inclui as três categorias tradicionais, com ênfase em obter uma combinação de evidências que, de forma otimizada, reflete o valor de um teste para um propósito pretendido" (AERA; APA; NCME 1985, p. 9).

Por fim, as duas últimas revisões abandonaram a visão fragmentada de validade e declararam que a interpretação do escore do teste e o uso dele são inseparáveis, pois a interpretação do escore é sempre seguida por uma ação (AERA; APA; NCME 2014). Os testes tendem a ser desenvolvidos para um uso e os escores dos testes não são utilizados para relatar simplesmente como o sujeito realizou certas tarefas em

uma ocasião específica sob certas condições. Em vez disso, os escores são usados para embasar afirmações que um examinador tem acerca dos atributos do sujeito, fundamentando, assim, uma decisão (seleção ou diagnóstico, por exemplo). Dessa forma, Kane (2013) afirma que o primeiro passo para o estudo de evidências de validade requer uma declaração clara do propósito da medida e das interpretações que serão realizadas a partir dos resultados do teste. A ideia central é declarar explicitamente o uso proposto da medida e avaliar a plausibilidade dessa proposta, a partir dos estudos de evidências de validade que serão realizados pelos responsáveis.

Atualmente, seguindo a versão mais atual dos *Standards* (AERA; APA; NCME 2014), são apontadas cinco fontes de evidências de validade: baseadas no conteúdo, no processo de resposta, na relação com variáveis externas, na estrutura interna e nas consequências da testagem. O Quadro 2 sintetiza as cinco fontes e traz os respectivos dados que são categorizados dentro dessas fontes.

Quadro 2 – As fontes de evidências de validade apresentadas na quarta versão dos *Standards* e os tipos de dados que são coletados para se levantar cada uma das fontes

Fonte	Tipos de Dados Coletados
Evidências com base no conteúdo	Acerca da representatividade da matriz e dos itens da medida, investigando se esses consistem em amostras representativas do domínio que se pretende avaliar.
Evidências com base no processo de resposta	Acerca dos processos mentais envolvidos na realização das tarefas propostas pela matriz ou pela teoria.
Evidências com base na estrutura interna	Acerca da representação do construto, com base nas dimensões avaliadas, na qualidade dos itens e na confirmação de hipóteses derivadas da teoria.
Evidências com base nas relações com variáveis externas	Acerca dos padrões de correlação entre os escores da medida e de outras variáveis que medem o mesmo construto ou construtos relacionados (convergência) e variáveis que medem construtos diferentes (divergência). Levantar também dados sobre a capacidade preditiva da medida com relação a outros fatos de interesse direto (critérios externos) que têm importância por si só e associam-se ao propósito direto do uso da medida.
Evidências com base nas consequências da testagem	Acerca das consequências sociais intencionais e não intencionais do uso da medida, com o fim de verificar se sua utilização está surtindo os efeitos desejados na sociedade, de acordo com o propósito para o qual foi criada.

Fonte: HUTZ (2009)

As evidências necessárias para a obtenção da propriedade de validade são determinadas pelas afirmações que serão realizadas; "afirmações mais ambiciosas exigem mais evidências do que afirmações menos ambiciosas" (KANE, 2013).

1.2 Estudos recentes sobre a validade no contexto brasileiro

No Brasil, no âmbito da Educação, o conceito atualizado de validade apresentado nos Standards (AERA; APA; NCME 2014) ainda é pouco conhecido e utilizado, poucos são os estudos que fazem uso. GOMES *et al*. (2020), por exemplo, realizaram um estudo cujo objetivo foi validar um instrumento de avaliação de efetividade da formação profissional. Esses autores realizaram um estudo com foco em uma categoria de validade, a de conteúdo, utilizando ainda a visão fragmentada do conceito.

Costa e Dias (2020), por sua vez, em um estudo sobre a proposta de um instrumento para avaliação da formação superior pelo discente, trataram o conceito de validade como a capacidade do teste em mensurar, de fato, o que pretende mensurar, ou seja, ainda fazendo uso do conceito antigo, e sem trabalhar com evidências de validade.

Outros estudos de validação de instrumentos no âmbito da Educação (CUNHA; SASAKI, 2020; DAVOGLIO; SANTOS; LETTNIN, 2016; FERREIRA, 2019; LIMA; SILVA, 2020; SILVA *et al*., 2017; SOUSA *et al*., 2018), também seguiram na mesma direção, ora tratando o conceito de validade de forma fragmentada, predominantemente considerando as características psicométricas, ora tratando o conceito como a capacidade de medir o que se propõe.

Por outro lado, os estudos realizados por Toffoli *et al*. (2016), assim como por Jesus, Rêgo e Souza (2018), ao tratar sobre o conceito de validade, utilizaram a visão mais atualizada, convergente com a última edição dos *Standards* (AERA; APA; NCME 2014).

Nesse cenário, considerando o conceito atual de validade, adotado nos *Standards*, que contempla cinco fontes de evidências de validade, o objetivo do presente estudo é analisar como o atual conceito e as fontes de evidências de validade são utilizados no âmbito da avaliação educacional em larga escala no Brasil. Para tanto, foi realizada análise documental dos relatórios das avaliações nacionais publicados pelo Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (Inep), a qual é apresentada a seguir.

2 Análise de evidências de validade em medidas de avaliação educacional em larga escala aplicadas no Brasil

Para realizar a análise foram selecionados os testes usados nas principais avaliações e/ou exames aplicados no Brasil, que abrangem tanto a Educação Básica quanto a Superior. Na Educação Básica, a principal avaliação aplicada no país, desde

2005 até 2017, foi a Prova Brasil, usando testes de língua portuguesa e de matemática. Essa avaliação era aplicada nos anos ímpares, a cada dois anos, de forma censitária na rede pública, e amostral na rede privada, nos 5º e 9º anos do Ensino Fundamental. Nas 3ª e 4ª séries do Ensino Médio, tanto na rede pública quanto na privada, a Prova Brasil era aplicada de forma amostral.

Em 2019, a Prova Brasil deixou de existir e foi incorporada pelo novo Saeb (Sistema de Avaliação da Educação Básica), que passou a aplicar testes de ciências da natureza e de ciências humanas (apenas no 9º ano do Ensino Fundamental, e somente de forma amostral, tanto na rede pública quanto na rede privada) e iniciou a avaliação do 2º ano do Ensino Fundamental, de forma amostral, tanto na rede pública quanto na privada, aplicando testes de língua portuguesa e de matemática. O novo Saeb também aplica testes de língua portuguesa e de matemática de forma censitária na rede pública, nos 5º e 9º anos do Ensino Fundamental, e nas 3ª e 4ª séries do Ensino Médio, a cada dois anos. A rede privada continua participando da avaliação, em todas as séries pesquisadas, de forma amostral. Devido à sua relevância como principal instrumento de acompanhamento da qualidade da Educação Básica no Brasil, essa avaliação foi escolhida para compor o rol de testes a serem analisados nesse estudo.

Além do Saeb, na Educação Básica também é aplicado um exame no Ensino Médio, o Exame Nacional do Ensino Médio (Enem), que, desde 2009, foi aperfeiçoado e passou a ser a principal porta de acesso ao Ensino Superior. O público do Enem são os estudantes concluintes e egressos do Ensino Médio. O Enem é aplicado anualmente a milhões de estudantes e é visto não como uma avaliação, mas como um processo seletivo para preenchimento de vagas no Ensino Superior. No Enem são aplicados testes de linguagens, códigos e suas tecnologias; de matemática e suas tecnologias; de ciências humanas e suas tecnologias; de ciências da natureza e suas tecnologias, além de uma prova de redação. Em virtude da sua relevância no contexto nacional, e por ser o principal meio utilizado como seleção para os cursos superiores no Brasil, o Enem foi selecionado para análise nesse estudo.

Por fim, na Educação Superior no Brasil, há apenas um teste que é aplicado aos estudantes, o Exame Nacional de Desempenho dos Estudantes (Enade). O Enade é aplicado aos concluintes dos cursos superiores, em ciclos avaliativos divididos por área do conhecimento, sendo cada área avaliada a cada três anos. O teste do Enade é composto por 30 questões do componente específico de cada curso e 10 questões de formação geral (comum a todos os cursos). Esse exame faz parte do Sistema Nacional de Avaliação da Educação Superior (Sinaes), que faz uso dos resultados do Enade para o cálculo de indicadores de avaliação. Por ser o

único exame nacional aplicado no Ensino Superior no Brasil, o Enade também foi selecionado para análise de evidências de validade no presente estudo.

Buscaram-se no sítio do Inep os relatórios desses testes, utilizando o seguinte critério: o relatório mais recente publicado até o ano de 2018. Ao analisar os relatórios da Prova Brasil, do Enem e do Enade, selecionados de acordo com o critério referido anteriormente, verificou-se que esses relatórios ainda adotam a visão fragmentada da validade, que já foi superada no contexto internacional. No Quadro 3, é possível comparar o tipo de evidência de validade atualmente considerada na literatura e o tipo de evidência apresentada nesses relatórios oficiais.

Quadro 3 – Fontes de evidências de validade nos relatórios das avaliações nacionais

Fonte	Relatório		
	Prova Brasil (a)	**Pedagógico: Enem (b)**	**Síntese de Área: Enade (c)***
Evidências com base no conteúdo	Parcialmente. O relatório apresenta as matrizes de referência para as diferentes séries, entre os anos de 2005 e 2015. Mas não descreve os procedimentos adotados para garantir a representatividade das matrizes nos testes aplicados. Ou seja, o documento não explicita as especificações dos testes. Por exemplo, não são apresentadas informações quanto a características básicas, como demanda cognitiva, dificuldade e número de itens por tópico da matriz.	Não. Embora o relatório apresente alguns itens como exemplo e os relacione à matriz, isso não é feito para a prova como um todo. No relatório, faz-se referência ao Mapa de Itens do Enem, que estaria disponibilizado no site, mas o *link* está desatualizado e, na época dessa pesquisa, encontrava-se fora do ar.	Não. A matriz de referência é apenas apresentada, mas não é feita qualquer relação dessa com os itens da prova. Mesmo na seção dedicada à análise de conteúdo das questões discursivas, em nenhum momento a matriz é retomada. Os itens são analisados independentemente da matriz.
Evidências com base no processo de resposta	Não. Nenhum estudo sobre o processo de resposta é relatado.	Não. Nenhum estudo sobre o processo de resposta é relatado.	Não. Nenhum estudo sobre o processo de resposta é relatado.

Continua

Continuação

Fonte	Relatório		
	Prova Brasil (a)	**Pedagógico: Enem (b)**	**Síntese de Área: Enade (c)***
Evidências com base na estrutura interna	Parcialmente. São apresentadas apenas análises dos itens, não é apresentada análise da prova como um todo. Não são relatadas análises de dimensionalidade e de fidedignidade das provas.	Parcialmente. São apresentadas apenas análises dos itens, não é apresentada análise da prova como um todo. Não são relatadas análises de dimensionalidade e de fidedignidade das provas.	Parcialmente. São apresentadas apenas análises dos itens, não é apresentada análise da prova como um todo. Não são relatadas análises de dimensionalidade e de fidedignidade das provas.
Evidências com base nas relações com variáveis externas	Não. Nenhum estudo de fatores associados é relatado.	Não. Nenhum estudo que relacione os escores das provas com variáveis externas é relatado.	Não. Nenhum estudo que relacione os escores das provas com variáveis externas é relatado.
Evidências com base nas consequências da testagem	Não. Não é feita referência às consequências e usos possíveis da medida.	Não. Não é feita referência às consequências e usos possíveis da medida.	Não. Mesmo na seção referente à distribuição dos conceitos, não é feita referência aos usos possíveis da medida.

(a) INEP, 2018. (b) INEP, 2015. (c) INEP, 2016.
*o mesmo modelo é utilizado para todas as áreas.
Fonte: Elaborada pelos autores (2018)

Nesses relatórios são apresentadas, de forma majoritária, diferentes análises estatísticas dos itens que compõem as provas, assim como estatísticas descritivas dos resultados e dos participantes. Entretanto, Messick (1989) destaca que não é suficiente avaliar as respostas dos itens isoladamente, mas é necessário avaliar o conjunto de respostas. Ou seja, é importante e fundamental avaliar o teste como um todo para emitir um julgamento sobre o desempenho. Além disso, pouca atenção é dada à fonte primária de validade, que se refere ao conteúdo da medida. Sireci (2013) destaca que é improvável que apenas uma fonte de evidência seja capaz de validar o uso de um teste para um propósito específico. Ademais, esse mesmo autor afirma que os dados nunca substituem um bom julgamento e que os testes não podem ser defendidos puramente por motivos estatísticos. De forma análoga, Borsboom, Mellenbergh e Van Heerden (2004) argumentam que uma grande parte da validade do teste deve ser colocada dentro do processo de construção.

Analisando os relatórios oficiais publicados, não é possível fazer um julgamento sobre a validade dos testes, a fim de justificar o uso da medida para o propósito

ao qual ela foi planejada. Kane (2013) diz que "afirmações públicas exigem justificativas públicas". Ou seja, a validação exige evidências suficientes para cada afirmação que é feita a partir dos escores do teste. Sem essas evidências, o julgamento emitido pelo Inep, com base nos escores desses testes, não tem transparência, consistindo em um prejuízo para a sociedade.

3 Exemplo de uso do conceito atual de validade

Com o objetivo de instruir a construção de medidas educacionais, Sireci (2013) indica três passos que devem ser realizados durante o desenvolvimento dos testes: definição clara de seu propósito, considerações acerca do uso indevido do teste e cruzamento do propósito do teste e uso indevido com as cinco fontes de evidências listadas nos *Standards*. Esses três passos podem ser usados para desenvolver um plano de validação e auxiliar na construção do argumento de validade.

O primeiro passo, então, é a definição do propósito do teste. "Como podemos desenvolver um bom teste se o seu propósito não for claramente enunciado?" (SIRECI, 2013, p. 101). O autor ressalta que esse passo tem se tornado cada vez mais importante, pois muitos programas de testagem contemporâneos se esforçam para alcançar múltiplos objetivos, mas raramente definem claramente esses objetivos (SIRECI, 2013).

O segundo passo trata de considerações acerca do mau uso do teste. Nessa etapa, confronta-se a forma como os resultados dos exames podem ser mal interpretados ou como o programa de testagem pode levar a resultados não desejados. Sireci (2013) destaca que nem sempre esse passo pode ser separado, porque muitas declarações de propósitos de testes já podem incluir precauções contra o mau uso da medida.

Sireci (2013) ainda aponta que uma maneira de identificar possíveis usos indevidos dos testes é ouvir as críticas que, comumente, são feitas ao instrumento, pois as críticas podem levar a questões que precisam ser investigadas e respondidas.

O último passo é o cruzamento do propósito do teste e o uso indevido com as cinco fontes de evidências de validade listadas nos *Standards*. Para ilustrar como esse passo poderia ser realizado, apresentamos o exemplo de um teste fictício de matemática criado por Sireci (2013). O Quadro 4 apresenta quais fontes de evidências de validade deveriam ser usadas para fundamentar o uso do teste para o propósito estabelecido – verificar a aprendizagem de matemática no país a partir da matriz curricular em vigor.

Quadro 4 – Teste fictício de matemática criado por Sireci (2013)

Propósito do teste (uso indevido*)	Fontes de evidências de validade				
	Conteúdo	Estrutura interna	Relação com variáveis externas	Processo de resposta	Consequências da testagem
Avaliar a proficiência em matemática dos estudantes tendo como referência a matriz curricular	x			x	
Determinar a proficiência em matemática dos alunos em três níveis: básico, proficiente ou avançado	x	x	x	x	
Fornecer informações acerca da proficiência em matemática dos estudantes que podem ser usadas pelo Estado, pela escola, pelo professor e como prestação de contas.	x		x		x
Fornecer informações que podem ser usadas para melhorar o Ensino na sala de aula, na escola e no estado.	x	x		x	x
(Os professores ensinam para aprovar os estudantes no teste, ao invés de seguir a matriz curricular)				x	x
(Os estudantes abandonam a escola para não responder os testes)					x

*O uso indevido está entre parênteses
Fonte: Sireci (2013)

O Quadro 4 mostra uma visão geral de como os 3 passos apontados por Sireci (2013) podem ser usados para desenvolver um argumento de validade. Observa-se também que o quadro ajuda os desenvolvedores de medidas a esboçar os tipos de estudos que podem ser realizados, afastando-os da visão fragmentada de validade. Caso uma avaliação intencione contemplar todos os propósitos elencados no

exemplo apresentado no Quadro 4, pode-se verificar que todas as evidências de validade deveriam ser levantadas.Os resultados de testes educacionais têm impacto direto na vida dos estudantes, dos professores, das escolas e das instituições de Ensino Superior. Portanto, é necessário investigar se as ações tomadas, a partir da mensuração, apresentam um argumento lógico e fundamentado em evidências.

4 Conclusão

O conceito de validade, e consequentemente, a forma como essa propriedade pode ser verificada, nos estudos de validação, sofreram diversas alterações ao longo dos anos. Este desenvolvimento, no entanto, tratou-se de uma evolução relativamente linear, caracterizado pela ampliação do entendimento do construto e da aceitação de novos dados como fontes de evidências de validade (JESUS; RÊGO; SOUZA, 2018). O conteúdo, a estrutura interna e as correlações com outras variáveis, por exemplo, estão presentes desde os primórdios na análise desse tema. Embora o entendimento do conceito de validade tenha-se tornado mais complexo e mais repleto de nuances, essas fontes de evidência de validade tiveram papel central na argumentação pela validade de um teste desde o início das discussões acerca do tema.

No âmbito das avaliações educacionais em larga escala brasileiras, o levantamento das evidências de validade ou não é feito ou não é relatado para a sociedade, impedindo o cidadão de auditar a qualidade e a justiça da utilização dos escores provenientes desses testes na classificação de candidatos, escolas ou instituições de Ensino Superior, por exemplo. Com base nos relatórios oficiais, não é possível fazer um julgamento acerca da validade dos testes aplicados atualmente no Brasil. Tal fato está na contramão do que propõem os *Standards* (AERA; APA; NCME, 2014) e os principais pesquisadores da área (KANE, 2013; MESSICK, 1989; PASQUALI, 2009; SIRECI, 2013), que utilizam o modelo para, entre outras coisas, a promoção da transparência.

Vale ressaltar que não apenas no âmbito da avaliação educacional em larga escala conduzida por entes públicos, mas outros estudos brasileiros que envolvem o desenvolvimento de medidas educacionais (BOLLELA; BORGES; TRONCON, 2018; FREDERICO-FERREIRA *et al*., 2017; SOUSA *et al*., 2018) também têm utilizado o conceito antigo de validade, considerando-a como a propriedade que o teste tem de medir o que se propõe, sendo composta por três tipos: validade de conteúdo, de critério e de construto.

Concluímos que, no que tange às avaliações educacionais em larga escala aplicadas no Brasil, ainda é necessário avançar tanto no entendimento do que é validade

quanto nas informações que são necessárias levantar e apresentar, a fim de que se possa utilizar com confiança os resultados obtidos nos diferentes processos avaliativos aplicados em âmbito nacional. Para que se melhore a qualidade dos testes no Brasil, cabem às instituições governamentais e à sociedade civil buscar formas para resguardar a qualidade dos testes, e aos que constroem os testes construir um texto técnico-científico para cada prova, argumentando, por meio de evidências em um modelo bem estabelecido, os motivos pelos quais os usos pretendidos pelos criadores dos testes estão justificados.

The question of validity in Brazilian educational assessment

Abstract

This study aimed to analyze how the current concept of validity and its sources of evidence have been utilized on large-scale assessment in Brazil. We made a documentary analysis using the official reports of the most important national educational assessments. The results indicated that the reports do not present adequate information about sources of evidence validity, which makes unfeasible to judge the validity of the educational tests analyzed. In this context, it is difficult to say if these tests evaluate what they propose. Therefore, we conclude that, in Brazil, it is still necessary to improve both understanding about what validity is, and what information is necessary to be able to use with confidence the results of different assessments applied in Brazil.

Keywords: *Validity. Educational Assessment. Standards for Testing.*

La cuestión de validez en la evaluación educativa brasileña

Resumen

El objetivo del presente estudio fue analizar cómo el actual concepto y las fuentes de evidencia de validez se utilizan en el marco de la evaluación educativa a gran escala en Brasil. Para ello, se realizó un análisis documental, a través de los informes oficiales de las mayores evaluaciones educativas nacionales aplicadas en el país. Los resultados obtenidos indicaron que no se presentan suficientes fuentes de evidencia de validez para juzgar que las pruebas educativas analizadas son válidas para lo que proponen. Aún es necesario desarrollar el conocimiento técnico de los investigadores brasileños que realizan estudios de validación de pruebas educativas, especialmente en el ámbito de la información que debe ser recolectada y presentada a fin de que se puedan utilizar con confianza los resultados obtenidos en los testes y pruebas educativas se apliquen en el país.

Palabras clave: *Validez. Evaluación Educativa. Estándares de Prueba.*

Referências

AMERICAN EDUCATIONAL RESEARCH ASSOCIATION – AERA; AMERICAN PSYCHOLOGICAL ASSOCIATION – APA; NATIONAL COUNCIL ON MEASUREMENT IN EDUCATION – NCME. *Standards for educational and psychological testing.* Washington, DC: American Psychological Association, 1985.

AMERICAN EDUCATIONAL RESEARCH ASSOCIATION – AERA; AMERICAN PSYCHOLOGICAL ASSOCIATION – APA; NATIONAL COUNCIL ON MEASUREMENT IN EDUCATION – NCME. *Standards for educational and psychological testing.* Washington, DC: American Educational Research Association, 1999.

AMERICAN EDUCATIONAL RESEARCH ASSOCIATION – AERA; AMERICAN PSYCHOLOGICAL ASSOCIATION – APA; NATIONAL COUNCIL ON MEASUREMENT IN EDUCATION – NCME. *Standards for educational and psychological testing.* Washington, DC: American Educational Research Association, 2014.

AMERICAN PSYCHOLOGICAL ASSOCIATION – APA. *Standards for educational and psychological tests and manuals.* Washington, DC: American Psychological Association, 1966.

AMERICAN PSYCHOLOGICAL ASSOCIATION – APA. Technical recommendations for psychological tests and diagnostic techniques. *Psychological Bulletin,* Washington, DC, v. 51*,* n. 2 pt. 2, p. 461-475, 1954. https://doi.org/10.1037/h0053479

AMERICAN PSYCHOLOGICAL ASSOCIATION – APA; AMERICAN EDUCATIONAL RESEARCH ASSOCIATION – AERA; NATIONAL COUNCIL ON MEASUREMENT IN EDUCATION – NCME. *Standards for educational and psychological tests*. Washington, DC: American Psychological Association, 1974.

BOLLELA, V. R.; BORGES, M. C.; TRONCON, L. E. A. Avaliação somativa de habilidades cognitivas: experiência envolvendo boas práticas para a elaboração de testes de múltipla escolha e a composição de exames. *Revista Brasileira de Educação Médica*, Brasília, DF, v. 42 n. 4, p. 74-85, out./dez. 2018.

BORSBOOM, D.; MELLENBERGH, G. J.; VAN HEERDEN, J. The concept of validity. *Psychological Review*, v. 111, n. 4, p. 1061-1071, 2004. https://doi.org/10.1037/0033-295X.111.4.1061

CONSELHO FEDERAL DE PSICOLOGIA. *Resolução CFP nº 002/2003.* Define e regulamenta o uso, a elaboração e a comercialização de testes psicológicos e revoga a Resolução CFP nº 025/2001. Brasília, DF, 2003.

COSTA, F. J.; DIAS, J. J. L. Avaliação da formação superior pelo discente: proposta de um instrumento. *Avaliação*, Campinas, v. 25, n. 2, p. 275-296, maio/ago. 2020. https://doi.org/10.1590/S1414-4077/S1414-40772020000200003

CUNHA, R. F. F.; SASAKI, D. G. G. Validação da nova versão do Test of Understanding Graphs in Kinematics (TUG-K) com estudantes de ensino médio. *Revista Brasileira de Ensino de Física*, São Paulo, v. 42, e20190149, 2020. https://doi.org/10.1590/1806-9126-RBEF-2019-0149

DAVOGLIO, T. R.; SANTOS, B. S.; LETTNIN, C. C. Validação da Escala de Motivação Acadêmica em universitários brasileiros. *Ensaio: Avaliação e Políticas Públicas em Educação*, Rio de Janeiro, v. 24, n. 92, p. 522-545, jul./set. 2016. https://doi.org/10.1590/S0104-40362016000300000

FERREIRA, L. M. L. Um estudo sobre a dimensionalidade das escalas de avaliação da proficiência oral do Certificado de Proficiência em Língua Portuguesa para Estrangeiros. *Educação e Pesquisa*, São Paulo, v. 45, e202512, 2019. https://doi.org/10.1590/S1678-4634201945202512

FREDERICO-FERREIRA, M. M., *et al*. Tradução e adaptação do questionário de validade das avaliações dos estudantes ao ensino e aos professores. *Avaliação*, Campinas, v. 22, n. 2, p. 458-468, jull./nov. 2017. https://doi.org/10.1590/S1414-40772017000200011

GOMES, D. E., *et al*. Efetividade da formação profissional ofertada na educação a distância: validação teórica de um instrumento. *Ensaio: Avaliação e Políticas Públicas em Educação*, Rio de Janeiro, v. 28, n. 108, p. 762-783, 2020. https://doi.org/10.1590/S0104-40362019002701667

GUILFORD, J. P. New standards for test evaluation. *Educational and Psychological Measurement,* Thousand Oaks, v. 6*,* p. 427-439, Dec.1946. https://doi.org/10.1177/001316444600600401

HALADYNA, T. M.; RODRIGUEZ, M. C. *Developing and validating test items*. New York: Routledge, 2013.

HUTZ, C. S. *Avanços e polêmicas em avaliação psicológica*. Itatiba: Casa do Psicólogo, 2009.

INSTITUTO NACIONAL DE ESTUDOS E PESQUISAS EDUCACIONAIS ANÍSIO TEIXEIRA – INEP. Enade 2014 – Exame Nacional de Desempenho dos Estudantes: relatório de área: Arquitetura e Urbanismo. Brasília, DF, 2016.

INSTITUTO NACIONAL DE ESTUDOS E PESQUISAS EDUCACIONAIS ANÍSIO TEIXEIRA – INEP. Relatório pedagógico Enem 2011-2012. Brasília, DF, 2015.

INSTITUTO NACIONAL DE ESTUDOS E PESQUISAS EDUCACIONAIS ANÍSIO TEIXEIRA - INEP. Relatório Saeb (Aneb e Anresc) 2005-2015: Panorama da década. Brasília, DF, 2018.

JESUS, G. R.; RÊGO, R. M. L.; SOUZA, V. V. Evidências de validade de conteúdo da prova de psicologia do Enade. *Estudos em Avaliação Educacional*, São Paulo, v. 29, n. 72, p. 858-884, set./dez.2018. https://doi.org/10.18222/eae.v29i72.4897

KANE, M. Validating the interpretations and uses of test scores. *Journal of Educational Measurement,* Malden, v. 50, n. 1, p. 1-73, Mar. 2013. https://doi.org/10.1111/jedm.12000

LIMA, D. B. M.; SILVA, G. O. L. Inventário de práticas docentes que favorecem a criatividade: estudo de adaptação e evidências de validade. *Educação em Revista,* Belo Horizonte, v. 36, e231110, 2020. https://doi.org/10.1590/0102-4698231110

LINN, R. L. The standards for educational and psychological testing: Guidance in test development. *In*: DOWNING, S. M.; HALADYNA, T. M. (org.). *Handbook of test development.* New Jersey: Lawrence Erlbaum Associates, 2006. p. 27-38.

MESSICK, S. Test validity and the ethics of assessment. A*merican Psychologist,* Washington, DC, v. 35, n. 11, p. 1012-1027, 1980. https://doi.org/10.1037/0003-066X.35.11.1012

MESSICK, S. Evidence and ethics in the evaluation of tests. *Educational Researcher,* v. 10, n. 9, p. 9-20, Nov. 1981. https://doi.org/10.3102/0013189X010009009

MESSIK, S. Validity. *In*: Linn, R. (ed.). *Educational measurement*. 3rd. ed. Washington, DC: American Council on Education, 1989. p. 13-103.

NEWTON, P. E. Macro- and micro-validation: beyond the 'Five Sources'Framework for Classifying Validation Evidence and Analysis. *Practical Assessment, Research & Evaluation*, v. 21, n. 12, p. 1-13, 2016. https://doi.org/10.7275/f75k-1y75

NEWTON, P. E.; SHAW, S. D. Standards for talking and thinking about validity. *Psychological Methods,* Washington, DC, v. 18, n. 3, p. 301-319, 2013.

PASQUALI, L. Psicometria. *Revista da Escola de Enfermagem da USP*, v. 43, n. esp., p. 992-999, 2009. https://doi.org/10.1590/S0080-62342009000500002

PRESSEY, S. L. Suggestions looking toward a fundamental revision of current statistical procedure, as applied to tests. *Psychological Review,* [s. l.], v. 27, n. 6, p. 466-472, 1920. https://doi.org/10.1037/h0075018

RULON, P. J. On the validity of educational tests. *Harvard Educational Review,* Cambridge, v. 16, p. 290-296, 1946.

SILVA, M. A, *et al*. Construção e estudo de evidências de validade da Escala de Avaliação Docente. *Revista Brasileira de Educação,* Rio de Janeiro, v. 22, n. 70, p. 690-707, jul./set. 2017. https://doi.org/10.1590/S1413-24782017227035

SIRECI, S. G. Agreeing on validity arguments. *Journal of Educational Measurement,* Malden, v. 50, n. 1 spe. iss., p. 99-104, Spring 2013.

SIRECI, S. G. On the validity of useless tests. *Assessment in Education: Principles, Policy & Practice,* Abingdon, v. 23, n. 2, p. 226-235, 2016. https://doi.org/10.1080/0969594X.2015.1072084.

SIRECI, S. G.; PARKER, P. Validity on trial: psychometric and legal conceptualizations of validity. *Educational Measurement: Issues and Practice,* [s. l.], v. 25, n. 3, p. 27-34, 2006. https://doi.org/10.1111/j.1745-3992.2006.00065.x

SIRECI, S. G.; SUKIN, T. Test validity. *In*: GEISINGER, K. F. (ed.). *APA handbook of testing and assessment in psychology*. Washington, DC: Burdett, 2013. v. 1, p. 61-84.

SMITH, H. L.; WRIGHT, W. W. *Tests and measurements.* New York: Silver, Burdett, 1928.

SOUSA, T. F., *et al*. Validade de constructo da escala Condições do Ambiente e Características de Aprendizagem na Universidade (CACAU). *Avaliação*, Campinas, v. 23, n. 3, p. 665-678, set./dez. 2018. https://doi.org/10.1590/S1414-40772018000300006

TOFFOLI, S. F. L., *et al*. Avaliação com itens abertos: validade, confiabilidade, comparabilidade e justiça. *Educação e Pesqui*sa, São Paulo, v. 42, n. 2, p. 343-358, abr./jun. 2016. https://doi.org/10.1590/S1517-9702201606135887

Informações sobre os autores

Girlene Ribeiro de Jesus: Doutora em Psicologia pela Universidade de Brasília, na área de Fundamentos e Medidas. Professora da Faculdade de Educação, professora e Pesquisadora no Programa de Pós-Graduação em Educação da mesma universidade. Contato: girlene@unb.br
https://orcid.org/0000-0002-1782-1089

Renata Manuelly de Lima Rêgo: Doutora em Psicologia pela Universidade de Brasília. Coordenadora de Ensino, Pesquisa e Avaliação do Centro Brasileiro de Pesquisa em Avaliação e Seleção e de Promoção de Eventos. Contato: renatamanuelly@gmail.com
https://orcid.org/0000-0002-8366-0981

Victor Vasconcelos de Souza: Doutorando em Psicologia pela Universidade de Brasília. Contato: victor.vscn@gmail.com
https://orcid.org/0000-0001-9786-8217