http://dx.doi.org/10.18222/eae.v29i72.4897

EVIDÊNCIAS DE VALIDADE DE CONTEÚDO DA PROVA DE PSICOLOGIA DO ENADE*

GIRLENE RIBEIRO DE JESUS[I]
RENATA MANUELLY DE LIMA RÊGO[II]
VICTOR VASCONCELOS DE SOUZA[III]

I Universidade de Brasília (UnB) e Centro Brasileiro de Pesquisa em Avaliação e Seleção e de Promoção de Eventos (Cebraspe), Brasília-DF, Brasil; *girlene.ribeiro@gmail.com*

II Universidade de Brasília (UnB) e Centro Brasileiro de Pesquisa em Avaliação e Seleção e de Promoção de Eventos (Cebraspe), Brasília-DF, Brasil; *renatamanuelly@gmail.com*

III Universidade de Brasília (UnB) e Centro Brasileiro de Pesquisa em Avaliação e Seleção e de Promoção de Eventos (Cebraspe), Brasília-DF, Brasil; *victor.souza@cebraspe.org.br*

***** O presente trabalho é fruto de um intercâmbio internacional realizado nos Estados Unidos, no Educational Testing Service (ETS).

RESUMO

O presente estudo tem como objetivo analisar evidências de validade com base no conteúdo da prova de psicologia aplicada no Exame Nacional de Desempenho dos Estudantes (Enade) de 2015. Foi utilizada a blueprint, *ferramenta recomendada pela literatura internacional que embasa o planejamento dos testes. Houve divergência significativa entre as competências e habilidades apresentadas na matriz de referência e a demanda cognitiva solicitada na maior parte das questões objetivas. Há habilidades presentes na matriz que não foram contempladas na prova e outras foram contempladas com apenas um único item, o que compromete a confiabilidade da medida. As análises mostraram a necessidade do uso de técnicas que podem melhorar essa fonte primária de evidência.*

PALAVRAS-CHAVE VALIDADE • EXAME NACIONAL DE DESEMPENHO DOS ESTUDANTES (ENADE) • BLUEPRINT • DEMANDA COGNITIVA.

EVIDENCIAS DE VALIDEZ DE CONTENIDO DE LA PRUEBA DE PSICOLOGÍA DEL ENADE

RESUMEN

El presente estudio tiene el propósito de analizar evidencias de validez en base al contenido de la prueba de psicología aplicada en el Exame Nacional de Desempenho dos Estudantes (Enade) del 2015. Se utilizó la blueprint, herramienta recomendada por la literatura internacional que sirve como base para la planificación de las pruebas. Hubo divergencia significativa entre las competencias y habilidades presentadas en la matriz de referencia y la demanda cognitiva solicitada en la mayoría de las preguntas objetivas. Hay habilidades presentes en la matriz que no se tuvieron en cuenta en la prueba y otras fueron contempladas con tan solo un ítem, lo que compromete la confiabilidad de la medida. Los análisis mostraron la necesidad del uso de técnicas que pueden mejorar esa fuente primaria de evidencia.

PALABRAS CLAVE VALIDEZ • EXAME NACIONAL DE DESEMPENHO DOS ESTUDANTES (ENADE) • BLUEPRINT • DEMANDA COGNITIVA.

EVIDENCE OF CONTENT VALIDITY OF THE ENADE PSYCHOLOGY ASSESSMENT

ABSTRACT

This study aims to analyze evidence of validity based on the content of the psychology test applied in the Exame Nacional de Desempenho dos Estudantes (Enade) [National Exam of Student Proficiency] of 2015. We used Blueprint, a tool recommended in the international literature for test design. There were significant discrepancies between the skills and abilities presented in the test specifications, and in the cognitive demand demanded in almost all multiple-choice questions. Furthermore, there are skills and abilities presented in test specifications that were not in the test, and other skills and abilities that were represented by only one item, which compromise measurement reliability. We conclude that the evidence based on the content for this test shows the necessity to use tools to improve this primary source of evidence.

KEYWORDS VALIDITY • NATIONAL EXAM OF STUDENT PROFICIENCY (ENADE) • BLUEPRINT • COGNITIVE DEMAND.

INTRODUÇÃO

Testes educacionais são aplicados anualmente para milhões de estudantes da educação básica e milhares da educação superior no Brasil. Os resultados obtidos pelos estudantes são utilizados para ingresso no ensino superior, responsabilização dos sistemas de ensino, das unidades escolares e das instituições de ensino superior. Além disso, os escores obtidos nos testes servem para o cálculo de indicadores e para a indução de políticas públicas educacionais.

Diante do protagonismo que os testes educacionais ocupam no cenário nacional, é preocupante a falta de padrões para verificação da validade dos escores emitidos, pois a propriedade mais importante que um teste educacional deveria apresentar refere-se à validade (HALADYNA; RODRIGUEZ, 2013; PASQUALI, 2010). Se os escores obtidos não têm evidências suficientes de que são válidos, todas as demais atividades relacionadas ao uso do teste não podem se dizer baseadas em critérios científicos.

No contexto internacional, existem padrões muito bem estabelecidos para guiar o processo de testagem. Os *Standards*

for Educational and Psychological Testing, daqui em diante tratados apenas como *Standards*, foram publicados em 2014 e são uma publicação conjunta da American Educational Research Association (AERA), American Psychological Association (APA) e National Council on Measurement in Education (NCME). São amplamente reconhecidos como uma declaração autorizativa proveniente de um consenso profissional em relação aos padrões para a testagem. Os *Standards* são "uma força global para a testagem" e desempenham um papel pedagógico importante, tanto na comunidade americana como na internacional, pois abordam conceitos tão fundamentais (validade, fidedignidade, normas, equalização, etc.) para a construção de testes e avaliação que podem ser facilmente empregados em diferentes contextos (ZUMBO, 2014, p. 33). Os *Standards* gozam desse *status* tanto pela forma como foram desenvolvidos e aprovados, contando com a chancela das associações mais importantes da área de psicologia e educação americanas, quanto pela ampla história que têm (LINN, 2006).

De acordo com a mais recente edição dos *Standards* (AMERICAN EDUCATIONAL RESEARCH ASSOCIATION; AMERICAN PSYCHOLOGICAL ASSOCIATION; NATIONAL COUNCIL ON MEASUREMENT IN EDUCATION, 2014, p. 11), validade pode ser definida como o "grau com que a evidência e a teoria apoiam as interpretações dos escores para determinado uso do teste" (MESSICK, 1989). Dessa forma, a propriedade de validade está relacionada aos escores, não ao teste (KANE, 2013).

Segundo Cizek (2016), essa definição aborda duas questões:

1. O que os escores significam?
2. Os resultados dos testes podem ser usados para o propósito estabelecido (por exemplo, acesso ao ensino superior)?

Kane (2013) afirma que não é possível avaliar a validade dos escores dos resultados sem adotar, explicitamente ou implicitamente, alguma proposta de interpretação ou uso. O autor acrescenta que caso alguém lhe apresente um teste e um conjunto de orientações de administração do teste e solicite sua validade, provavelmente as primeiras perguntas

seriam acerca de como os escores do teste serão interpretados e usados, qual a população-alvo e qual o contexto de aplicação. Somente após uma proposta de interpretação e uso enunciada, as afirmações que serão feitas a partir do escore podem ser avaliadas.

Tipicamente estudada na psicologia, ciência na qual são investigados fenômenos que não são passíveis de observação direta, a validade constitui-se em um parâmetro fundamental e indispensável na avaliação (PASQUALI, 2017). Haladyna e Rodriguez (2013) entendem a validade como um tema crucial no desenvolvimento dos testes educacionais, indicando que os estudos sobre evidências de validade visam a garantir a definição clara dos construtos e a adequada representação desses.

Os *Standards* apresentam uma visão que incorpora o modelo com cinco fontes de evidências de validade como guia que deve orientar os desenvolvedores dos testes nos estudos sobre essa questão: 1) evidências baseadas no conteúdo do teste; 2) evidências baseadas no processo de resposta; 3) evidências baseadas na estrutura interna; 4) evidências baseadas na relação com outras variáveis; e 5) evidências baseadas nas consequências da testagem (AERA; APA; NCME, 2014).

QUADRO 1 – Fontes de evidências de validade

FONTE	PROCEDIMENTOS
Evidências com base no conteúdo	Levantar dados sobre a representatividade da matriz e dos itens do teste, investigando se esses consistem em amostras abrangentes do domínio que se pretende avaliar.
Evidências com base no processo de resposta	Levantar dados sobre os processos mentais envolvidos na realização das tarefas propostas pela matriz.
Evidências com base na estrutura interna	Levantar dados sobre a 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	Levantar dados sobre os padrões de correlação entre os escores do teste e 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 do teste com relação a outros fatos de interesse direto (critérios externos) que têm importância por si só e associam-se ao propósito direto do uso do teste (por exemplo, sucesso no trabalho).
Evidências com base nas consequências da testagem	Examinar as consequências sociais intencionais e não intencionais do uso do teste para verificar se sua utilização está surtindo os efeitos desejados, de acordo com o propósito para o qual foi criado.

Fonte: Hutz (2009, p. 251).

Embora a visão atual sobre validade contemple cinco fontes de evidência (AERA; APA; NCME, 2014), no âmbito dos estudos na área de avaliação educacional brasileira, tem sido dada grande ênfase ao levantamento de evidências de validade com base na estrutura interna do instrumento e pouca atenção ao estudo das evidências com base no conteúdo. 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. Além disso, o autor explica que os dados nunca substituem um bom julgamento e que os testes não podem ser defendidos puramente por motivos estatísticos. Comparativamente, Borsboom, Mellenbergh e Van Heerden (2004) argumentam que uma grande parte da validade do teste deve ser colocada dentro do processo de construção – um estágio do processo de testagem que tem recebido pequena atenção, comparada com a ênfase enorme que tem sido dada à análise estatística do teste.

Considerando o cenário internacional e a literatura da área (AERA; APA; NCME, 2014; HALADYNA; RODRIGUEZ, 2013; MESSICK, 1989) frente à escassez de estudos no Brasil, no que tange ao estudo da propriedade de validade das medidas educacionais, o presente estudo tem como objetivo realizar um julgamento de validade acerca das evidências com base no conteúdo da prova de psicologia aplicada no Exame Nacional de Desempenho dos Estudantes (Enade) de 2015.

No Brasil, especialmente no que tange à avaliação educacional conduzida nacionalmente em larga escala, são escassos os estudos e a apresentação das evidências de validade com base no conteúdo, que é a fonte primária de evidência de um teste, visto que todos os escores serão reflexos de sua composição (KANE, 2013). Embora todas as avaliações educacionais realizadas no país levem à responsabilização, seja dos sistemas de ensino, das instituições de ensino, ou mesmo dos estudantes, não existem padrões estabelecidos para se conduzir um processo avaliativo.

EVIDÊNCIAS DE VALIDADE COM BASE NO CONTEÚDO

A validade ocupa posição central na avaliação, perpassando todos os processos dessa, desde o estabelecimento do objetivo ou concepção do teste, desenvolvimento e aplicação da medida e interpretação dos resultados até o uso desses resultados para um objetivo específico (MISLEVY, 2007). Nesse sentido, qualquer esforço realizado durante a construção de um teste deve ter como objetivo último a garantia da validade.

A preocupação com as evidências de validade com base no conteúdo tem levado pesquisadores a desenvolver métodos que possibilitem analisar a representação do construto de forma mais confiável (DEVILLE, 1996; LYNN, 1986; POLIT; BECK; OWEN, 2007; SIRECI; GEISINGER, 1992). Entende-se construto como uma característica que não pode ser observada ou medida diretamente, visto que não há um único referente ou um conjunto de referentes que cubram o construto como um todo (CRONBACH; MEEHL, 1955). Dessa forma, seriam exemplos de construtos a ansiedade, a depressão, a proficiência em português ou matemática, entre outros.

Os estudos de evidências de validade com base no conteúdo dos testes visam a investigar se o teste constitui uma representação adequada do construto (PASQUALI, 2009). Sireci (1998) definiu os quatro componentes críticos da validade de conteúdo: definição do construto, representação do construto, relevância do construto e adequação dos procedimentos de desenvolvimento do teste. Esses quatro componentes reforçam a ideia apresentada por Kane (2006), que afirma que a evidência de validade relacionada ao conteúdo está intimamente ligada ao desenvolvimento do teste. Sireci (1998, p. 106) afirma que nunca se pode escapar do problema da validade de conteúdo: se o evitarmos durante a construção do teste, "ele levantará sua cabeça incômoda no momento da interpretação do escore".

A validade de conteúdo surgiu para evitar que as avaliações dos testes fossem estritamente numéricas, cometendo, assim, ameaças graves à validade das inferências derivadas do escore (SIRECI, 1998). Os estudos da área indicam que essa preocupação é frequente desde o início do desenvolvimento dos primeiros testes. Kelley (1927), por exemplo, expressou

preocupação com uma perspectiva puramente estatística sobre validação e sugeriu um julgamento mais amplo, envolvendo profissionais da área, para complementar as avaliações da validade do teste. Nessa linha de raciocínio, Rulon (1946) recomendou que a avaliação da validade deveria incluir uma avaliação do conteúdo do instrumento e a sua relação com o objetivo da mensuração. Esses pesquisadores, entre outros, sinalizaram a mudança de concepção e a prática de validação de testes. Essa mudança expandiu o conceito de validade para além da noção de testes de correlações e enfatizou que a validação deveria considerar a adequação do conteúdo do teste em relação ao propósito do teste (SIRECI, 1998). Na realidade, desde 1966, associações americanas reconhecem a necessidade do levantamento de evidências de validade de conteúdo como imperativo para testes educacionais (AMERICAN PSYCHOLOGICAL ASSOCIATION – APA, 1966).

Os procedimentos usados para avaliar a validade de conteúdo são geralmente classificados como de julgamento. Esses métodos referem-se aos estudos nos quais especialistas da área são consultados para avaliar se os itens do teste estão representados de forma adequada e se os tópicos mais importantes do conteúdo são avaliados na medida. No Brasil, há diferentes publicações, na área da psicologia, ressaltando a importância desse procedimento no desenvolvimento do teste (NASCIMENTO; SOUZA, 2017; PASQUALI, 1999, 2010).

Mislevy (2007) afirma que fortes evidências de validade podem ser apresentadas quando essa é pensada no desenvolvimento do teste. Dessa forma, podemos argumentar que o desenvolvimento de um teste requer um processo de documentação bem organizado para reunir evidências de validade suficientes para apoiar as inferências propostas em relação aos resultados. De fato, esse processo é utilizado em instituições internacionalmente reconhecidas, como o *Educational Testing Service* (ETS) (MISLEVY; ALMOND; LUKAS, 2003). Downing (2006) estabelece 12 passos para o desenvolvimento de um teste; aqui, porém, apresentamos somente os cinco primeiros, aqueles mais relacionados à validade de conteúdo:

1. *Plano geral*: orientação sistemática para todas as atividades de desenvolvimento dos testes: construto;

inferências desejadas; formato do teste; principais fontes de evidência de validade; propósito claro; modelo psicométrico; segurança; controle de qualidade;

2. *Definição do conteúdo*: plano amostral do domínio (ou construto) investigado; emprego de vários métodos para avaliar evidência de validade relacionada ao conteúdo do teste; delineamento do construto;
3. *Especificações do teste*: definição operacional do conteúdo; plano de estudos relacionados à validade, visando reunir evidências que o conteúdo do domínio selecionado tem relação com o construto investigado; características desejadas dos itens;
4. *Desenvolvimento do item*: formato do item; treinamento dos escritores e revisores dos itens; análise para verificar se os itens avaliam variância irrelevante do construto;
5. *Montagem do teste*: criação de formas paralelas do teste; seleção de itens para formatos específicos do teste; utilização da *blueprint*. A *blueprint* é uma forma tabelar da estrutura de conteúdo de um teste utilizado para manter consistência entre versões diferentes de um mesmo teste (ALDERMAN, 2015).

Em conformidade com a visão apresentada por Downing (2006), Haladyna e Rodriguez (2013) discorrem sobre a importância da construção de um conjunto de especificações do teste e dos itens, pois essas orientações constituem-se como um guia valioso para o desenvolvimento do teste e base importante para as evidências de validade com base no conteúdo. O termo especificações do teste se refere a um documento que deve conter, no mínimo:

1. os tipos de itens a serem usados e o fundamento para sua seleção;
2. as instruções sobre como criar os itens, incluindo informações sobre o estilo do item, a demanda cognitiva, a especificação se os itens do teste apresentarão estímulos visuais, como fotografias ou gráficos, o limite de tempo de resposta para cada item e os princípios a serem seguidos na elaboração;

3. a classificação dos itens por conteúdo e demanda cognitiva;
4. a tabela *blueprint*, que provê a base para o planejamento do teste; ela auxilia na visualização de quantos itens há disponíveis e quantos mais são necessários para os vários conteúdos e demandas cognitivas;
5. como os escores serão interpretados (norma ou critério).

O documento de especificações deve ser disponibilizado de tal forma que os interessados no processo tenham consciência dos altos padrões empregados para assegurar que o conteúdo do teste representa o construto avaliado. Além disso, as especificações do teste e dos itens são úteis para todos os profissionais que trabalham na construção e no desenvolvimento da medida. Apesar de o principal objetivo da *blueprint* ser auxiliar a construção dos itens e a montagem do teste, ela também traz transparência para a composição do teste, para que todos os envolvidos saibam o que é esperado de um testando.

A construção da *blueprint* é flexível e deve se adaptar às necessidades do teste. Comumente são descritos os conteúdos, as habilidades, a demanda cognitiva, a quantidade de itens para cada tópico, o tipo de item (questões de múltipla escolha ou resposta construída), o peso de cada item para a nota e o tempo de resposta por item (RUTKOWSKI; VON DAVIER; RUTKOWSKI, 2014). É considerada um elemento-chave no processo de montagem de diferentes versões de um mesmo teste, pois pode garantir que as mesmas habilidades estejam sendo mensuradas nas diferentes versões.

Por exemplo, o *Partnership for Assessment of Readiness for College and Careers* (PARCC) é um teste padronizado aplicado nos Estados Unidos. Suas diversas versões são equalizadas com base não só nas análises estatísticas, mas também na construção da prova seguindo uma *blueprint*. A *blueprint* do PARCC traz seis colunas: o tema; o conjunto de itens, que descreve o tipo de atividade exigida; o número de questões (N) associado a esse conjunto; as afirmações sobre as habi-

lidades dos candidatos; os números de pontos nas questões objetivas e nas questões discursivas (PARTNERSHIP FOR ASSESSMENT OF READINESS FOR COLLEGE AND CAREERS – PARCC, 2017).

O ETS, maior organização de desenvolvimento de testes do mundo, define, no seu manifesto acerca da qualidade dos testes lá produzidos, os *Standards for Quality and Fairness* (EDUCATIONAL TESTING SERVICE – ETS, 2014), que os desenvolvedores precisam ter uma *blueprint* detalhada para que possam montar um teste. Esse manifesto ainda especifica que o desenvolvimento de cada item está atrelado ao documento de especificações.

O *Test of English as Foreign Language* (TOEFL) é um exemplo de teste de renome internacional dessa organização que utiliza a *blueprint* com objetivo de informar evidências de validade e manter a consistência e a comparabilidade entre os escores das provas. A *blueprint* correspondente é publicada para acesso livre em seu *site*. Essa *blueprint*, além das informações essenciais, informa acerca da separação do teste em diferentes seções e estabelece o tempo necessário para resposta de cada seção (CHAPELLE; ENRIGHT; JAMIESON, 2011).

Outro exemplo de teste educacional americano que utiliza a *blueprint* é o *Scholastic Assessment Test* (SAT), que avalia conteúdos de matemática, inglês, história, línguas e ciências e que é respondido todos os anos por cerca de 1,6 milhão de estudantes (COLLEGE BOARD, 2015). O escore do SAT serve para compor um conjunto de notas utilizadas para admissão do aluno na universidade. A *blueprint* da prova do SAT apresenta: tempo de aplicação da prova; quantidade de palavras por passagens; quantidade de questões da prova; tipo de questão (múltipla escolha ou resposta construída); quantidade de questões de cada conteúdo selecionado; e dados acerca da complexidade textual, entre outros (COLLEGE BOARD, 2015).

Um último exemplo de teste que utiliza a *blueprint* como um plano que orienta o desenvolvimento da medida é o *National Assessment of Educational Progress* (NAEP). O NAEP é uma avaliação norte americana utilizada para avaliar e monitorar o desempenho dos estudantes do ensino fundamental e médio daquele país. As provas são respondidas

por estudantes do 4º, 8º e 12º anos e abordam as diferentes disciplinas estudadas na escola. Cada disciplina tem uma *blueprint* construída sob orientação de professores, especialistas da área, especialistas em avaliação, formuladores de políticas e membros do público em geral. A *blueprint* de leitura, por exemplo, especifica o tipo de texto que deve ser usado (literários e informativos); o tipo e a quantidade de textos para cada série (ficção, não ficção, poesia, exposição, argumentação, por exemplo); o tipo de material que será utilizado no comando como estímulo; a demanda cognitiva das questões segundo a Taxonomia Revisada de Bloom; e a quantidade de questões em cada nível (NATIONAL ASSESSMENT OF EDUCATIONAL PROGRESS – NAEP, 2015).

O ENADE

O Enade é um dos pilares da avaliação do Sistema Nacional de Avaliação da Educação Superior (Sinaes), criado em 2004. Estruturam o Enade dois componentes: o primeiro, denominado Componente de Formação Geral, configura a parte comum às provas das diferentes áreas, avaliando competências, habilidades e conhecimentos gerais de temas exteriores ao âmbito específico da profissão; o segundo, denominado Componente de Conhecimento Específico, contempla a especificidade de cada área no domínio dos conhecimentos e habilidades esperados para o perfil profissional.

As provas do Enade são compostas por 40 questões, sendo 35 de múltipla escolha e cinco discursivas, que abordam as habilidades e competências que o egresso do curso deve ter. Essas habilidades e competências, bem como as matrizes de referência de cada área, são publicadas em portarias específicas. Por exemplo, para os egressos do curso de psicologia, a Portaria Inep n. 243, de 10 de junho de 2015 (BRASIL, 2015b, p. 27) descreve as competências e as habilidades que o estudante deveria ter desenvolvido ao longo do curso de psicologia:

> I – avaliar, sistematizar e decidir as condutas profissionais, com base em evidências científicas;
>
> II – planejar, conduzir e relatar investigações científicas de

distintas naturezas, apoiado em análise crítica das diferentes estratégias de pesquisa;
III – identificar e analisar necessidades de natureza psicológica, elaborar projetos, planejar e agir de forma coerente com referenciais teóricos e características da população-alvo;
IV – elaborar relatos científicos, pareceres técnicos, laudos e outras comunicações profissionais, inclusive materiais de divulgação;
V – utilizar os recursos da matemática, da estatística e da informática para a análise e apresentação de dados e para a preparação das atividades profissionais em Psicologia;
VI – diagnosticar, planejar e intervir em processos educativos em diferentes contextos;
VII – diagnosticar, planejar e intervir em processos de gestão, em distintas organizações e instituições;
VIII – diagnosticar, planejar e intervir em processos de prevenção e promoção da saúde, em nível individual e coletivo;
IX – diagnosticar, planejar e intervir em processos de assistência e apoio psicossocial a grupos, segmentos e comunidades em situação de vulnerabilidade individual e social;
X – realizar psicodiagnóstico, psicoterapia e outras estratégias clínicas frente a questões e demandas individuais e coletivas;
XI – coordenar e manejar processos grupais, em diferentes contextos, considerando as diferenças individuais e socioculturais dos seus membros;
XII – avaliar os resultados e impactos das intervenções psicológicas conduzidas em diferentes contextos.

Tendo em vista a complexidade do processo de levantamento de evidências de validade com base no conteúdo, composto por diversas etapas, e considerando que a análise terá como foco uma prova já elaborada e aplicada, este artigo apresenta uma das etapas do processo de levantamento de evidência com base no conteúdo, a construção da *blueprint* – um dos componentes mais importantes do processo de documentação de um teste. O presente artigo discorre sobre as evidências baseadas no conteúdo do teste e apresenta uma versão reduzida da *blueprint*.

Originalmente, a *blueprint* foi criada como um plano para organização dos itens de teste conforme dados acerca da utilidade e da relevância de determinados temas para a avaliação (HALADYNA; RODRIGUEZ, 2013). Neste estudo, utilizamos a *blueprint* como uma ferramenta para diagnóstico da representatividade do conteúdo. Será usado, a título de exemplo, o componente específico da prova do Enade para o curso de psicologia aplicada em 2015 (BRASIL, 2015a).

A TAXONOMIA REVISADA DE BLOOM

Um dos elementos da *blueprint* é a informação acerca do processo cognitivo demandado pelos itens. Para identificar e realizar a classificação do processo cognitivo, é necessário fazer uso de alguma taxonomia.

A função principal das taxonomias é, certamente, prover um modelo para que os educadores possam elaborar objetivos educacionais (MARZANO; KENDALL, 2007), os quais, em geral, têm como foco o desenvolvimento de competências e habilidades.

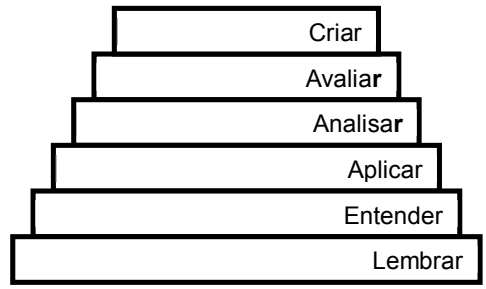
Em termos de estrutura, os objetivos educacionais iniciam com um verbo e finalizam com o substantivo. O **verbo** deve indicar o processo cognitivo (por exemplo, lembrar, aplicar, analisar) empregado no objetivo, e o **substantivo** descreve o conhecimento que se espera que os estudantes adquiram ou construam.

A separação entre verbo e substantivo, ou seja, entre o processo cognitivo e o conhecimento, foi fundamental no processo de revisão, pois conferiu à taxonomia original um caráter bidimensional: (a) Dimensão Conhecimento e (b) Dimensão do Processo Cognitivo (FERRAZ; BELHOT, 2010). Segundo Anderson *et al.* (2001), processo cognitivo é o meio pelo qual o conhecimento pode ser adquirido ou construído, bem como usado para resolver problemas.

No presente estudo, é trabalhada a dimensão do processo cognitivo, uma vez que essa é uma informação que precisa constar da *blueprint*. A dimensão do processo cognitivo é composta por seis categorias: *lembrar*, *entender*, *aplicar*, *analisar*, *avaliar* e *criar*. De acordo com Anderson *et al.* (2001),

o contínuo subjacente à dimensão do processo cognitivo é organizado por níveis de complexidade cognitiva, ou seja, *entender* é cognitivamente mais complexo do que *lembrar*, *aplicar* é cognitivamente mais complexo do que *entender*. Essa organização confere um caráter hierárquico à taxonomia, como pode ser visto na Figura 1.

FIGURA 1 – Categorias do domínio cognitivo



Fonte: Anderson *et al.* (2001) e Ferraz e Belhot (2010).

Como se pode verificar, as categorias vão dos processos cognitivos mais comumente encontrados nos objetivos educacionais, como *lembrar*, *entender* e *aplicar,* até aqueles menos frequentemente encontrados, como *analisar*, *avaliar* e *criar* (Anderson et al., 2001). A seguir há uma descrição das seis categorias da dimensão do processo cognitivo propostas por Anderson *et al.* (2001).

1. *Lembrar*: significa recuperar conhecimentos relevantes da memória de longo prazo. Os objetivos educacionais planejados nesse nível promovem um trabalho com foco na memorização do material apresentado, da forma mais similar possível à forma como foi visto. Os dois processos cognitivos associados a essa categoria são reconhecer e recordar.
2. *Entender*: tem como foco construir significados a partir de conteúdos orais, escritos e comunicações gráficas. Quando o estudante é capaz de reproduzir com suas próprias palavras a informação que viu, ouviu ou leu.
3. *Aplicar*: executar ou usar um procedimento em determinada situação. Envolve o uso de procedimentos

para realizar exercícios ou resolver problemas. Dessa forma, a categoria *aplicar* consiste de dois processos cognitivos: executar, quando a tarefa é um exercício (familiar), e implementar, quando a tarefa é um problema (não familiar).

4. *Analisar*: fracionar o material em suas partes constituintes e determinar como as partes se relacionam umas com as outras e com o todo. Os objetivos educacionais classificados nessa categoria incluem aprender a determinar as partes relevantes de uma mensagem (diferenciar), as formas por meio das quais as partes de uma mensagem são organizadas (organizar) e o propósito subjacente da mensagem (atribuir). *Analisar* pode ser considerada uma extensão de *entender* ou como um prelúdio para *avaliar* ou *criar*.
5. *Avaliar*: realizar julgamentos baseados em critérios e padrões. Os critérios mais frequentemente utilizados são qualidade, efetividade, eficiência e consistência. Por sua vez, os padrões podem ser tanto quantitativos quanto qualitativos. A categoria *avaliar* inclui os processos cognitivos de checar e criticar.
6. *Criar*: juntar elementos para formar um todo, reorganizar elementos formando uma nova estrutura. Os objetivos educacionais classificados como *criar* levam os estudantes a fazer algo novo, por meio da reorganização mental de elementos ou partes formando algo novo, não existente antes.

MÉTODO

Para analisar a propriedade da validade, levantaram-se evidências com base no conteúdo da prova de psicologia aplicada no Enade 2015. Foram utilizadas duas estratégias nesse processo: a *blueprint* e o cálculo da razão de validade de conteúdo (RVC).

Quando se utiliza a *blueprint*, o correto é relacionar cada questão às habilidades constantes da matriz de referência. Além disso, a demanda cognitiva de cada questão foi classificada levando em consideração a Taxonomia Revisada de Bloom (ANDERSON et al. 2001). Assim, realizou-se a

análise qualitativa das 30 questões referentes ao componente específico do Enade, sendo 27 de múltipla escolha e três discursivas. Buscou-se verificar a representatividade do domínio por meio dos itens do teste.

Para o cálculo do RVC, método proposto por Lawshe (1975), as 30 questões foram submetidas à avaliação de cinco especialistas que deveriam indicar se o item era: (a) essencial ao teste; (b) útil, mas não essencial; e (c) não necessário. Esse método produz uma medida da relação entre o número de avaliadores que classificou um determinado item numa categoria e o número total de avaliadores por meio da seguinte fórmula:

$$RVC = \frac{n_e - N/2}{N/2} \qquad (1),$$

sendo n_e o número de avaliadores que classificou o item na categoria proposta (essencial ao teste) e N o número total de avaliadores.

RESULTADOS E DISCUSSÃO

O Quadro 1 apresenta a *blueprint* reduzida, no qual é possível observar a classificação de cada questão quanto: (a) ao tipo de questão; (b) processo cognitivo; e (c) tópico dos conteúdos curriculares do componente específico. Destaca-se que não foi possível relacionar cada questão às habilidades propostas na matriz de referência devido à divergência entre o processo cognitivo declarado na matriz de referência e o efetivamente exigido nos itens. As competências e habilidades, descritas na Portaria Inep n. 243, de 10 de junho de 2015 (BRASIL, 2015b, p. 27), exigem processos cognitivos de alta complexidade, por exemplo: avaliar, sistematizar e decidir as condutas profissionais; planejar, conduzir e relatar investigações científicas de distintas naturezas; coordenar e manejar processos grupais; e avaliar os resultados e impactos das intervenções psicológicas. Essa demanda cognitiva alta não foi observada na maioria dos itens objetivos, o que tornou impossível relacionar os itens à matriz de referência. Dessa forma, optou-se por utilizar na *blueprint* os conteúdos curriculares

do componente específico. Além dessas informações, do Quadro 2 também consta o resultado da RVC.

QUADRO 2 – Componente específico: questões 9 a 35

CONTEÚDOS CURRICULARES DO COMPONENTE ESPECÍFICO	QUESTÃO	TIPO	PROCESSO COGNITIVO	RVC
1) Fundamentos epistemológicos e históricos				
a) Constituição da psicologia como campo de conhecimento	9	ME	L	-0,6
b) Constituição da psicologia como campo de atuação profissional no Brasil	10	ME	L	-0,6
c) Constituição, fundamentos e pressupostos epistemológicos dos principais sistemas psicológicos	11	ME	L	-0,6
2) Fundamentos, métodos e técnicas de coleta e análise de dados para investigações científicas				
a) Fundamentos das medidas em psicologia	12	ME	L	-0,2
b) Instrumentos e procedimentos de coleta de dados	14	ME	E	-0,2
c) A lógica da argumentação científica em psicologia	15	ME	An	-0,6
d) Concepção, planejamento, delineamento e comunicação de investigação científica	16	ME	A	-0,2
3) Fenômenos psicológicos				
a) Processos psicológicos de atenção, memória, percepção, linguagem, pensamento, consciência e inteligência	13, 27	ME	L, An	-1, -0,2
b) Emoção, afetos e motivação	19	ME	E	-0,2
c) Desenvolvimento humano	20	ME	L	-1
d) Personalidade e subjetividade	23	ME	E	-1
e) Processos psicopatológicos	18	ME	E	-0,2
f) Indivíduo, sociedade e cultura	21	ME	L	-1
g) Processos grupais, organizacionais e institucionais	22, 29	ME	L, L	-0,2, -0,6
h) Princípios e processos de aprendizagem	24	ME	A	-0,2
i) Psicofarmacologia e comportamento	25	ME	E	-0,2
j) Bases biológicas e evolutivas do comportamento	-	-	-	-
k) Neurociência das emoções, cognição e comportamento	26, 34	ME	E, L	-1, -0,6
4) Principais domínios de atuação do psicólogo				
a) Intervenções em processos educativos	28, 33	ME	E, E	-0,2, 0,2
b) Intervenções em processos organizacionais e de gestão de pessoas	-	-	-	-
c) Intervenções em processos de trabalho, saúde e bem-estar do trabalhador	30	ME	E	-0,6
d) Atenção e promoção da saúde (básica, secundária e terciária)	32	ME	E	-0,2
e) Avaliação psicológica/psicodiagnóstico	17, 31	ME	An, E	1, -0,6
f) Intervenções em grupos, instituições e comunidades	35	ME	E	-0,2
g) Psicoterapias	-	-	-	-

Fonte: Elaboração dos autores.

Nota: A: aplicar; An: analisar; C: criar; E: entender; L: lembrar; ME: múltipla escolha; RVC: razão de validade de conteúdo; S: sintetizar.

De forma geral, apenas com a utilização da *blueprint* já é possível observar muita divergência entre a matriz de competências e habilidades e as questões objetivas, tendo em vista o contraste entre a baixa demanda cognitiva solicitada nas questões e a alta demanda especificada na matriz. Com relação ao processo cognitivo demandado, prevaleceram itens classificados nos níveis mais baixos da taxonomia: *lembrar*, *entender* e *aplicar*. Dos 27 itens de múltipla escolha, dez estão no nível mais baixo de complexidade cognitiva (*lembrar*), 12 no segundo nível (entender), somente dois itens exigiam aplicação de conceitos (aplicar) e três exigiam análise (*analisar*). Todas as questões discursivas foram classificadas no nível mais alto da taxonomia (*criar*).

Quanto à RVC, se o item fosse considerado essencial por mais da metade dos avaliadores, ele teria validade de conteúdo; assim, quanto mais o item fosse indicado como essencial, mais validade de conteúdo teria (HUTZ; BANDEIRA; TRENTINI, 2015). Entretanto, para evitar que a concordância entre os juízes se desse ao acaso, Lawshe (1975) apresentou uma tabela com valores mínimos de RVC. Com cinco juízes avaliando a questão, o valor mínimo de RVC deveria ser 0,99. Caso a RVC atinja esse valor, é improvável que a concordância entre os juízes tenha ocorrido ao acaso.

Como se pode observar no Quadro 2, somente a questão objetiva 17 alcançou o valor mínimo estabelecido para a RVC; a questão 33, apesar de atingir o valor de RVC de 0,2, classificou-se como item essencial ao teste por três avaliadores. Além disso, todas as três questões discursivas foram classificadas como essenciais ao teste. Destaca-se que nove questões objetivas receberam a classificação "item não necessário" e seis questões foram classificadas como "item útil, mas não essencial". Esse resultado indica que a prova do Enade precisa sofrer grandes alterações para alcançar o objetivo proposto, que é convergir com a matriz de referência da prova, cuja proposta é a avaliação de competências por meio de um perfil profissional dos concluintes dos cursos superiores.

Tendo em vista que o objetivo da prova é avaliar competências adquiridas durante a formação, percebe-se a necessidade de apresentar, de forma predominante, itens com

situações que exijam do examinando níveis de raciocínio mais elevados. A quase ausência de níveis mais altos da taxonomia (*analisar*, *avaliar* e *criar*) nos itens objetivos da prova demonstra que boa parte dos itens solicita somente reconhecer, recordar, reproduzir informações ou comparar fatos no processo de resposta. Dessa forma, a demanda cognitiva não atende aos objetivos da avaliação definidos na matriz de referência, composta por competências como: avaliar condutas profissionais; planejar investigações; elaborar relatos científicos; intervir em processos científicos; e coordenar processos grupais, entre outros.

Em qualquer assunto, um estudante pode ter conhecimento e demonstrar capacidade de recordar o conteúdo. Entretanto, recordar um conteúdo estudado anteriormente não significa que o estudante compreende, de fato, o significado do que estudou. Além disso, os estudantes podem não ter a capacidade de aplicar o conhecimento em situações diferentes daquela em que foi aprendida ou combinar com um conhecimento adicional para criar novos *insights*, competências exigidas na matriz de referência da prova. É importante lembrar que a matriz da prova do Enade, documento oficial que guia a construção do teste, parece estar de acordo com o reconhecimento generalizado da importância de invocar processos de ordem superior (*higher-order thinking*) tanto no currículo quanto na avaliação (MOMSEN et al., 2010; SCULLY, 2017). Assim, o que precisa ser repensado é o processo de construção dos itens que compõem a prova.

Convém citar, ainda, que os itens de múltipla escolha podem ser usados para avaliar pensamentos de ordem superior (SCULLY, 2017) e que os itens que exigem tal pensamento melhoram a amplitude e a profundidade da cobertura de conteúdo em um teste (CIZEK; WEBB; KALOHN, 1995). Scully (2017) apresenta algumas sugestões para construir itens de múltipla escolha que avaliam pensamento crítico: (a) utilizar os verbos indicados para cada nível da taxonomia, por exemplo: conhecimento (identificar, definir, listar, nomear, etc.); (b) usar distratores de alta qualidade; e (c) construir itens que exijam do examinando o conhecimento de mais de um fato

ou conceito, para que ele precise combinar as informações para escolher a resposta correta.

Scully (2017) ainda afirma que construir itens de múltipla escolha avaliando pensamento de ordem superior é, sem dúvida, tarefa desafiadora e demorada, mas possível de ser feita. Além disso, algumas pesquisas indicam que estudantes que respondem avaliações que exigem raciocínio de ordem superior são posteriormente mais propensos a adotar metas significativas para seu estudo, evitando estratégias superficiais de aprendizagem, e que tais avaliações auxiliam os professores a dar *feedback* mais detalhado e específico que, por sua vez, pode promover e orientar a aprendizagem futura (JENSEN et al., 2014; LEUNG; MOK; WONG, 2008; MOMSEN et al., 2010).

Em trabalhos clássicos de coleta de evidências de validades para testes, são utilizadas análises após a aplicação do teste com o objetivo de garantir a validade de construto por meio da análise da dimensionalidade dos testes. Nessa categoria da análise, estão a análise de componentes principais e a Teoria da Resposta ao Item (TRI). No entanto, Huddleston (1956) já chamava a atenção para a preocupação de iniciar o processo de coleta de evidências de validade desde o início do processo de desenvolvimento dos testes. Um exemplo de abordagem nesse sentido, desenvolvido pelo ETS, é o *Evidence-Centered Design* (ECD) (ZIEKY, 2014). Trata-se de um método de criação e documentação do processo de testagem que se inicia com a definição de frases operacionais sobre o que se espera do examinando que esteja bem preparado para o teste. Esse método se adequa às recomendações dos *Standards*, visto que promove de forma clara as interpretações possíveis acerca dos escores dos estudantes.

Dessa forma, a coleta de evidências de validade de conteúdo já deve começar pela documentação do processo de decisão dos objetos de conhecimento, competências e habilidades que devem compor a matriz. A *blueprint* auxilia a montagem do teste, deixando claro, para os desenvolvedores do teste, os tópicos que estão sendo avaliados e aqueles que não serão abordados, o nível cognitivo de cada questão, o tipo do item, entre outras informações. Os testes padronizados nos Estados Unidos, por exemplo, têm documentos de especificação (*test*

specifications) que descrevem detalhes sobre o histórico e o objetivo do teste, uma *blueprint* mais ampliada que descreve os conteúdos a serem abordados, o número de questões de cada tipo e o tipo de contexto especificado para os itens.

Uma dificuldade que pode surgir no momento de montagem da *blueprint* é a alocação de itens que avaliam mais de uma habilidade. A técnica da *blueprint* parte do princípio de que são respeitadas as recomendações de Haladyna (2004), que sugere que os itens devem ser unidimensionais, uma vez que itens que avaliam diferentes combinações de dimensões num teste perdem a interpretabilidade e não são passíveis de avaliação por meio da TRI.

Por fim, enfatizando a validade como um aspecto central, Downing e Haladyna (2009) apresentam diversos procedimentos, qualitativos e quantitativos, que podem ser adotados pelos desenvolvedores de testes para analisar ou melhorar o grau de validade dos escores. Li e Sireci (2013) também apresentam um método que envolve análise qualitativa e quantitativa da análise de conteúdo de uma prova americana aplicada em larga escala. Todas essas ferramentas, inclusive a *blueprint*, são meios que facilitarão a construção adequada de instrumentos, fundamentando a interpretação dos escores e a tomada de decisão.

CONSIDERAÇÕES FINAIS

No contexto brasileiro, não são relatadas evidências de validade com base no conteúdo dos testes nos relatórios das avaliações educacionais. Comumente, são apresentadas somente as matrizes de referência e as análises estatísticas realizadas após a aplicação do teste. Newton (2016) afirma que se pensar em validade como algo estanque ou utilizando apenas um estudo para demonstrar a "validade do teste" é uma relíquia do passado. A visão atual de validade reitera que o estudo dessa propriedade deve ser programa contínuo de pesquisa, utilizando todos os tipos de evidências e análises indicadas na literatura (NETWON, 2016).

Outras questões importantes poderiam ser incluídas nos relatórios educacionais produzidos no Brasil, considerando

outros aspectos além das análises estatísticas; por exemplo, uma avaliação da adequação do item às especificações da matriz de referência, do enunciado e dos distratores (para questões de múltipla escolha), da complexidade textual e do contexto. No caso da avaliação de aptidão em matemática, por exemplo, se a questão traz um texto, esse deve ser adequado ao nível de escolaridade do examinando, visto que não deve ser tão complexo que elimine os candidatos que não consigam interpretá-lo por falhas no raciocínio verbal, e não por falhas na habilidade matemática que está sendo testada.

Um problema comum no estudo de evidências de validade com base no conteúdo é que esse tipo de avaliação é geralmente qualitativo e intuitivo, sem o uso de protocolos padronizados e, geralmente, não passa por nenhuma análise empírica. No entanto, existem recomendações para que outros tipos de análise sejam realizados, empregando técnicas mais rigorosas que quantifiquem, com certa precisão, a congruência entre as avaliações dos especialistas e a especificação do construto (DEVILLE; PROMETRIC, 1996; LYNN, 1986; POLIT; BECK; OWEN, 2007; SIRECI; GEISINGER, 1992).

Por fim, além da análise individual dos itens, o teste precisa ser avaliado como um todo, a fim de possibilitar a verificação de áreas do construto que não foram abordadas no teste e, assim, avaliar o grau de sub-representação. O ideal seria a completa representação do construto especificado, porém, na prática, isso não ocorre (HOGAN, 2006). Por esse motivo, é possível observar a importância de examinar se o conteúdo do teste é suficiente para fornecer a informação desejada a partir de uma amostra representativa do conteúdo previsto.

REFERÊNCIAS

ALDERMAN, J. *Test development process at ETS*. Princeton: ETS Global Institute Course, 2015.

AMERICAN EDUCATIONAL RESEARCH ASSOCIATION; AMERICAN PSYCHOLOGICAL ASSOCIATION; NATIONAL COUNCIL ON MEASUREMENT IN EDUCATION. *Standards for educational and psychological testing.* Washington, DC: APA, 2014.

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

ANDERSON, L. W. et al. (Ed.). *A taxonomy for learning, teaching, and assessing*: a revision of Bloom's Taxonomy of educational objectives. 2. ed. New York: Longman, 2001.

BORSBOOM, D.; MELLENBERGH, G. J.; VAN HEERDEN, J. The concept of validity. *Psychological Review*, Washington, v. 111, n. 4, p. 1061-1071, nov. 2004.

BRASIL. Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira. *Exame nacional do desempenho dos estudantes*: psicologia. Brasília, DF: Inep, 2015a. Disponível em: <http://download.inep.gov.br/educacao_superior/enade/provas/2015/09_psicologia.pdf>. Acesso em: 8 ago. 2017.

BRASIL. Portaria Inep n. 243, de 10 de junho de 2015. Estabelece as diretrizes da área de psicologia. *Diário Oficial da União*, Brasília, DF, 12 jun. 2015b. Seção 1, p. 27.

CHAPELLE, C. A.; ENRIGHT, M. K.; JAMIESON, J. M. *Building a validity argument for the test of English as a foreign language*. New York: Routledge, 2011.

CIZEK, G. J. Validating test score meaning and defending test score use: different aims, different methods. *Assessment in Education:* Principles, Policy & Practice, London, v. 23, n. 2, p. 212-225, Aug. 2016.

CIZEK, G. J.; WEBB, L. C.; KALOHN, J. C. The use of cognitive taxonomies in licensure and certification test development: reasonable or customary? *Evaluation & The Health Professions*, Thousand Oaks, v. 18, n. 1, p. 77-91, Mar. 1995.

COLLEGE BOARD. *Test specifications for the redesigned SAT.* New York: College Board, 2015.

CRONBACH, L. J.; MEEHL, P. E. Construct validity in psychological tests. *Psychological Bulletin*, Washington, DC, v. 52, p. 281-302, July 1955.

DEVILLE, C. W. An empirical link of content and construct validity evidence. *Applied Psychological Measurement*, Thousand Oaks, v. 20, n. 2, p. 127-139, June 1996.

DEVILLE, C. W.; PROMETRIC, S. An empirical link of content and construct validity evidence. *Applied Psychological Measurement*, Thousand Oaks, v. 20, n. 2, p. 127-139, 1996.

DOWNING, S. M. Twelve steps for effective test development. In: DOWNING, S. M.; HALADYNA, T. M. (Org.). *Handbook of test development.* New Jersey: Lawrence Erlbaum Associates, 2006. p. 3-25.

DOWNING, S. M.; HALADYNA, T. M. Test item development: validity evidence from quality assurance procedures. *Applied Measurement in Education*, Oxford, v. 10, n. 1, p. 61-82, Dec. 2009.

EDUCATIONAL TESTING SERVICE. *ETS standards for quality and fairness.* Princeton: ETS, 2014.

FERRAZ, A. P. C. M.; BELHOT, R. V. Taxonomia de Bloom: revisão teórica e apresentação das adequações do instrumento para definição de objetivos instrucionais. *Gestão & Produção*, São Carlos, v. 17, n. 2, p. 421-431, 2010.

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

HALADYNA, T. M.; RODRIGUEZ, M. C. *Developing and validating test items*. New York: Taylor & Francis Group, 2013.

HOGAN, T. P. Introdução à prática de testes psicológicos. São Paulo: LTC, 2006.

HUDDLESTON, E. M. Test development on the basis of content validity. *Educational and Psychological Measurement*, Thousand Oaks, v. 16, n. 3, p. 283-293, Oct. 1956.

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

HUTZ, C. S.; BANDEIRA, D. R; TRENTINI, C. M. *Psicometria*. Porto Alegre: Artmed, 2015.

JENSEN, J. L. et al. Teaching to the test… or testing to teach: exams requiring higher order thinking skills encourage greater conceptual understanding. *Educational Psychology Review*, New York, v. 26, n. 2, p. 307-329, Jan. 2014.

KANE, M. Content-related validity evidence in test development. In: DOWNING, S. M.; HALADYNA, T. M. (Org.). *Handbook of test development.* New Jersey: Lawrence Erlbaum Associates, 2006. p. 131-153.

KANE, M. Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, New Jersey, v. 50, n. 1, p. 1-73, mar. 2013.

KELLEY, T. L. *Interpretations of educational measurement*. Yonkers-on-Hudson: World Book, 1927.

LAWSHE, C. H. A quantitative approach to content validity. *Personnel Psychology*, Hoboken, v. 28, n. 4, p. 563-575, dez. 1975.

LEUNG, S. F.; MOK, E.; WONG, D. The impact of assessment methods on the learning of nursing students. *Nurse Education Today*, v. 28, n. 6, p. 711-719, Aug. 2008.

LI, X.; SIRECI, S. G. A new method for analyzing content validity data using multidimensional scaling. *Educational and Psychological Measurement*, Thousand Oaks, v. 73, n. 3, p. 365-385, Jan. 2013.

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.

LYNN, M. R. Determination and quantification of content validity. *Nursing Research*, London, v. 35, n. 6, p. 382-385, Nov./Dec. 1986.

MARZANO, R. J.; KENDALL, J. S. *The new taxonomy of educational objectives*. 2. ed. Thousand Oaks: Corwin, 2007.

MESSICK, S. Validity. In: LINN, R. L. (Ed.). *Educational measurement*. Washington, DC: American Council on Education; National Council on Measurement in Education, 1989. p. 13-103.

MISLEVY, R. J. Validity by design. *Educational Reseacher*, Washington, v. 36, n. 8, p. 463-469, Nov. 2007.

MISLEVY, R. J.; ALMOND, R. G.; LUKAS, J. F. A brief introduction to Evidence--Centered Design. *ETS Research Report Series*, Princeton, v. 03-16, n. 1, p. i-29, July 2003.

MOMSEN, J. L. et al. Just the facts? Introductory undergraduate biology courses focus on low-level cognitive skills. *CBE-Life Sciences Education*, Bethesda, v. 9, n. 4, p. 435-440, Dec. 2010.

NASCIMENTO, T. G.; SOUZA, E. C. L. Escala trifatorial da identidade social (ETIS): evidências de sua adequação psicométrica. *Psico-USF*, Bragança Paulista, v. 22, n. 2, p. 217-234, May/Aug. 2017.

NATIONAL ASSESSMENT OF EDUCATIONAL PROGRESS. *Reading Framework for the 2015 National Assessment of Educational Progress*. Washington: U.S. Government Printing Office, Jan. 2015.

NEWTON, P. E. Macro- and micro-validation: Beyond the "five sources" framework for classifying validation evidence and analysis. *Practical Assessment, Research & Evaluation*, College Park, v. 21, n. 12, p. 1-13, Dec. 2016.

PARTNERSHIP FOR ASSESSMENT OF READINESS FOR COLLEGE AND CAREERS. *PARCC Grades 6-11 High Level Blueprints*. EUA: 2017. Disponível em <http://www.parcconline.org/files/83/Spring%202016/388/Grades%206-11%20High%20Level%20Blueprint%20(Updated).pdf>. Acesso em: 26 jun. 2017.

PASQUALI, L. *Instrumentos psicológicos*: manual prático de elaboração. Brasília, DF: LabPAM/IBAPP, 1999.

PASQUALI, L. Psicometria. *Revista da Escola de Enfermagem da USP*, São Paulo, v. 43, p. 992-999, dez. 2009. Edição especial.

PASQUALI, L. *Instrumentação psicológica*: fundamentos e prática. Porto Alegre: Artmed, 2010.

PASQUALI, L. Validade dos testes. *Examen*: Pesquisa em Avaliação, Certificação e Seleção, Brasília, DF, v. 1, n. 1, p. 14-48, jul./dez. 2017.

POLIT, D. F.; BECK, C. T.; OWEN, S. T. Is the CVI an acceptable indicator of content validity? Appraisal and recommendations. *Research in Nursing & Health*, Thousand Oaks, v. 30, n. 4, p. 459-567, Aug. 2007.

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

RUTKOWSKI, L.; VON DAVIER, M.; RUTKOWSKI, D. *Handbook of international large-scale assessment*: background, technical issues, and methods of data analysis. Boca Raton: CRC/Taylor & Francis Group, 2014.

SCULLY, D. Constructing multiple-choice items to measure higher-order thinking. *Practical Assessment, Research & Evaluation*, College Park, v. 22, n. 4, p. 1-13, May 2017. Disponível em: <http://pareonline.net/getvn.asp?v=22&n=4>. Acesso em: 27 ago. 2017.

SIRECI, S. G. The construct of content validity. *Social Indicators Research*, New York, v. 45, n. 1-3, p. 83-117, Nov. 1998.

SIRECI, S. G. Agreeing on validity arguments. *Journal of Educational Measurement*, New Jersey, v. 50, n. 1, p. 99-104, Mar. 2013.

SIRECI, S. G.; GEISINGER, K. F. Analyzing test content using cluster analysis and multidimensional scaling. *Applied Psychological Measurement*, Thousand Oaks, v. 16, n. 1, p. 17-31, Mar. 1992.

ZIEKY, M. J. An introduction to the use of Evidence-Centered Design in test development. Psicología Educativa, Madrid, v. 20, n. 2, p. 79-87, dic. 2014.

ZUMBO, B. D. What role does, and should, the test standards play outside of the United States of America? *Educational Measurement*: Issues and Practice, Philadelphia, v. 33, n. 4, p. 31-33, Dec. 2014.

Recebido em: 14 AGOSTO 2017
Aprovado para publicação em: 20 JUNHO 2018