Estratégias para a coleta de evidências de validade de avaliações de sala de aula

MARCELO COPPI[I]

http://dx.doi.org/10.22347/2175-2753v14i45.3825

Resumo

Até recentemente, o parâmetro da validade dos instrumentos de avaliação era interpretado como a capacidade deste em mensurar o que foi planejado para medir. Nessa perspectiva, a validade era determinada por padrões psicométricos, originários da avaliação dos testes de larga escala. Contudo, devido às condições particulares do contexto de sala de aula, torna-se necessária a adequação da validade das avaliações nesta perspectiva. Assim, este estudo, mediante uma breve revisão do estado da arte, apresenta o atual conceito de validade, aponta a incompatibilidade da aplicação desse conceito entre os testes de larga escala e as avaliações de sala de aula e apresenta estratégias de coleta de evidências de validade para a interpretação propícia e para o uso adequado dos resultados das avaliações de sala de aula.

Palavras-chave: evidências de validade; validade de instrumentos de avaliação; estratégias de coleta de evidências; avaliação de sala de aula.

Submetido em: 08/02/2022
Aprovado em: 14/11/2022

[I] Centro de Investigação em Educação e Psicologia da Universidade de Évora (CIEP-UE), Évora, Portugal; http://orcid.org/0000-0001-6734-7592; e-mail: mcoppi@uevora.pt.

Strategies for collecting validity evidence of classroom assessments

Abstract
Until recently, the parameter of validity of assessment instruments was interpreted as their ability to measure what they were designed to measure. From this perspective, validity was determined by psychometric standards, originating from the evaluation of large-scale tests. However, due to the particular conditions of the classroom context, it is necessary to adjust the validity of assessments from this perspective. Thus, this study, through a brief review of the state of the art, presents the current concept of validity, points out the incompatibility of applying the concept of validity between large-scale tests and classroom assessments, and presents strategies for collecting validity evidence for the supportive interpretation and appropriate use of classroom assessment results.
Keywords: validity evidence; validity of assessment instruments; strategies for collecting evidence; classroom assessment.

Estrategias para recoger pruebas de validez de las evaluaciones en el aula

Resumen
Hasta hace poco, el parámetro de validez de un instrumento de evaluación se interpretaba como su capacidad para medir lo que se había diseñado para medir. Desde esta perspectiva, la validez se determinaba por medio de normas psicométricas, procedentes de la evaluación de pruebas a gran escala. Sin embargo, debido a las condiciones particulares del contexto del aula, es necesario ajustar la validez de las evaluaciones desde esta perspectiva. Así, este estudio, a través de una breve revisión del estado del arte, presenta el concepto actual de validez, señala la incompatibilidad de la aplicación del concepto de validez entre las pruebas a gran escala y las evaluaciones en el aula, y presenta estrategias para la recogida de pruebas de validez para la interpretación de apoyo y el uso adecuado de los resultados de la evaluación en el aula.
Palabras claves: pruebas de validez; validez de los instrumentos de evaluación; estrategias de recogida de pruebas; evaluación en el aula.

Introdução

A avaliação em sala de aula pode ser definida como "a coleta de informações de diversas fontes, com a intenção de promover um ensino e uma aprendizagem eficazes" (KANE; WOOLS, 2020, p. 11). Para isso, esse tipo de avaliação pode assumir uma grande variedade de formatos e de funções educacionais, entre elas: o planejamento e a modificação das atividades instrucionais, a identificação das dificuldades e das potencialidades dos alunos, o *feedback* e a comunicação do progresso para os alunos, a seleção e a atribuição de notas (KANE; WOOLS, 2020; MILLER; LINN; GRONLUND, 2009).

Em sala de aula, os professores avaliam seus alunos com o objetivo de tomar boas decisões educacionais a seu respeito (POPHAM, 2017). Para o autor "a principal contribuição da avaliação de sala de aula para a tomada de decisão de um professor é fornecer evidências razoavelmente precisas sobre a situação dos alunos" (POPHAM, 2017, p. 99).

A fim de cumprir este propósito, é necessário que as avaliações de sala de aula evidenciem uma alta qualidade técnica (PHAKITI; ISAACS, 2021). Russel e Airasian (2008, p. 18) corroboram a ideia, alegando que "a característica mais importante de uma boa avaliação é a sua capacidade de ajudar o professor a tomar as decisões adequadas", capacidade esta que os autores denominam como validade. De acordo com Popham (2017), a validade é a característica mais significante para a qualidade de uma avaliação, principalmente no contexto de sala de aula.

Contudo, até recentemente, a validade das avaliações de sala de aula era determinada por padrões psicométricos, os quais eram originários dos testes padronizados de larga escala (BROOKHART, 2003; KANE; WOOLS, 2020; MCMILLAN, 1999). Apesar de ser conceitualmente importante para a maioria dos tipos de avaliação, a validade na perspectiva psicométrica tem pouca relevância no contexto de sala de aula, uma vez que os objetivos das avaliações de sala de aula são diferentes daqueles dos testes de larga escala (MCMILLAN, 1999). De acordo com o autor, o foco dos professores referente à validade das avaliações de sala de aula está no uso e nas consequências dos resultados e não em uma inspeção detalhada do próprio instrumento de avaliação (MCMILLAN, 1999).

Alam e Aktar (2020) corroboram a ideia, afirmando que a concepção de validade, originada a partir do enquadramento dos testes de larga escala,

geralmente, ignora o contexto, que é "essencialmente uma parte inseparável e importante na avaliação de sala de aula" (ALAM; AKTAR, 2020, p. 162). Segundo os autores, sob a ótica dos testes de larga escala, o contexto é construto-irrelevante, já na perspectiva da sala de aula, este é relevante para a avaliação dos alunos. Desta forma, argumentam Alam e Aktar (2020), devido às suas implicações contextuais, os dois tipos de avaliação são fundamentalmente diferentes e nem sempre é viável aplicar os princípios de validade destas na conjuntura de sala de aula.

Vale ressaltar que, em 1988, Cronbach descreveu duas perspectivas a respeito da validade das avaliações, uma de medição e uma funcional. Enquanto a perspectiva da medição concentra-se no rigor e na precisão da pontuação como medida de algum construto – incluindo características como, por exemplo, a padronização, a consistência e a imparcialidade –, a perspectiva funcional está relacionada como o quão bem a avaliação serve aos propósitos pretendidos (CRONBACH, 1988). Para Kane e Wools (2020), embora as perspectivas se concentrem em diferentes critérios de avaliação, ambas são relevantes para a validação de todos os tidos de avaliação. No entanto, no contexto de sala de aula, "a perspectiva funcional é de preocupação central e a perspectiva de medição tem um papel de apoio" (KANE; WOOLS, 2020, p. 11).

Brookhart (2003) ressalta a importância da discussão da validade no contexto de sala de aula. De acordo com a autora, é preciso

> apontar o desajuste entre o contexto da psicometria em larga escala e da avaliação em sala de aula e sugerir maior adequação e potencialmente mais vias produtivas de desenvolvimento teórico para o campo da avaliação em sala de aula (BROOKHART, 2003, p. 6).

Phakiti e Isaacs (2021) também enfatizam a necessidade de os professores compreenderem o papel da validade no contexto da avaliação de sala de aula. Os autores alegam que muitos são os desafios, dentre eles:

> [...] avaliações de baixa qualidade, falta de clareza de instrução e de tarefas, número insuficiente de perguntas ou tarefas, presença de parcialidade, métodos de pontuação pouco claros e transparentes, baixa entrega de *feedback*, pouco envolvimento dos alunos nos procedimentos de avaliação, um estreitamento do currículo como resultado do ensino para o teste e/ou potenciais usos indevidos dos testes para fins diferentes daqueles para os quais foram destinados (PHAKITI; ISAACS, 2021, p. 4).

Foi nesse enquadramento que o estudo apresentado neste artigo pretendeu: a) apresentar a atual definição do conceito de validade; b) apontar a incompatibilidade da aplicação do conceito de validade entre os testes de larga escala e as avaliações de sala de aula; e c) apresentar estratégias de coleta de evidências de validade para a interpretação propícia e para o uso adequado dos resultados das avaliações de sala de aula, a fim de chancelar o processo avaliativo dos docentes em sala de aula.

Mediante uma breve revisão do estado da arte sobre o conceito de validade e sobre a sua aplicação nos testes de larga escala e nos instrumentos de avaliação de sala de aula, aspira-se responder às seguintes perguntas de investigação: 1) qual é a definição de validade mais aceita atualmente?; 2) o conceito de validade utilizado nos testes de larga escala pode ser aplicado às avaliações de sala de aula?; 3) quais estratégias podem ser utilizadas pelos professores para que as suas avaliações apresentem evidências de validade?

Nesse sentido, o artigo está estruturado em três seções: definição de validade, na qual é apresentado um sucinto histórico da evolução do conceito de validade, revelando os conceitos tradicional e atual de validade; incompatibilidade na aplicação do conceito de validade entre avaliações de sala de aula e testes de larga escala, que identifica os motivos de tal assimetria; e estratégias para a coleta de evidências de validade adequadas para avaliações de sala de aula, a qual designa indicações técnicas para a coleta de evidências de validade baseadas no conteúdo, nos processos de resposta e nas consequências das avaliações.

Definição de validade

Tradicionalmente, o parâmetro da validade de um instrumento de avaliação é interpretado como a capacidade deste medir o que foi planejado para medir (GIPPS, 2003). De acordo com Phakiti e Isaacs (2021), essa definição antiga, com registros do início do século 21, foi amplamente incontestada até a década de 1950, quando Cronbach e Meehl (1955) a questionaram, alegando, por exemplo, que se uma avaliação for mal administrada os resultados não podem ser válidos.

Em 1989, Messick realizou mudanças radicais no âmbito da validade (PHAKITI; ISAACS, 2021). Messick (1989) declarou que a validade não é uma propriedade do teste, mas do significado dos seus resultados, os quais podem sofrer interferência do contexto da avaliação e dos próprios respondentes. O autor também enfatizou as

consequências sociais, intencionais e não intencionais, da utilização dos resultados das avaliações (MESSICK, 1989).

Contudo, Phakiti e Isaacs (2021) alegam que, na prática, a noção de validade proposta por Messick (1989) tem sido considerada pouco funcional. De acordo com os autores, embora Messick tenha apresentado uma concepção ampla de validade, incluindo as repercussões do seu uso para as tomadas de decisões, o autor não descreveu suficientemente o "como implementar esta noção expandida de validade de forma que suas recomendações para validar avaliações pudessem ser implementadas na prática de acordo com seus críticos" (PHAKITI; ISAACS, 2021, p. 6).

Outros pesquisadores, como, por exemplo, Brookhart (2003) e Kane (2006), trouxeram contribuições para a discussão a respeito da validade das avaliações. Ainda assim, conforme alegam Phakiti e Isaacs (2021), a interpretação tradicional do parâmetro da validade ainda é comumente utilizada.

Nessa perspectiva, um instrumento de avaliação pode apresentar três tipos de validade: conteúdo, construto e critério. Basicamente, a validade de conteúdo está relacionada com relevância e a representatividade dos conteúdos que serão avaliados; a validade de construto refere-se à capacidade do teste ou do instrumento de avaliação em medir a habilidade que se propôs a avaliar, ou seja, o construto; e a validade de critério, que pode ser concorrente ou preditiva, e está relacionada com a previsão da performance dos indivíduos em comparação com algum critério ou instrumento externo (GIPPS, 2003; PHAKITI; ISAACS, 2021).

Todavia, atualmente, a literatura define validade como um conceito único (AMERICAN EDUCATIONAL RESEARCH ASSOCIATION; AMERICAN PSYCHOLOGICAL ASSOCIATION; NATIONAL COUNCIL ON MEASUREMENT IN EDUCATION, 2014; MILLER; LINN; GRONLUND, 2009; POPHAM, 2017), o qual corresponde ao "grau em que as evidências e a teoria apoiam as interpretações dos resultados dos testes para os usos propostos" (AMERICAN EDUCATIONAL RESEARCH ASSOCIATION; AMERICAN PSYCHOLOGICAL ASSOCIATION; NATIONAL COUNCIL ON MEASUREMENT IN EDUCATION, 2014, p. 11). Dentro dessa concepção, os tipos de validade dão lugar às evidências de validade, representadas por cinco categorias: evidências de validade baseadas no conteúdo, evidências de validade baseadas nos processos de respostas, evidências de validade baseadas na estrutura interna, evidências de validade baseadas em outras variáveis e evidências de validade baseadas nas consequências das avaliações (AMERICAN EDUCATIONAL RESEARCH ASSOCIATION;

AMERICAN PSYCHOLOGICAL ASSOCIATION; NATIONAL COUNCIL ON MEASUREMENT IN EDUCATION, 2014).

As evidências baseadas no conteúdo incluem a escolha do próprio conteúdo, o texto do enunciado, o formato dos itens e o processo de administração e de pontuação dos instrumentos de avaliação. Conforme consta nos *Standards for Educational and Psychological Testing*, este tipo de evidência inclui "análises lógicas ou empíricas da adequação com a qual o conteúdo do teste representa o domínio do conteúdo e a relevância do domínio do conteúdo para a interpretação da pontuação do teste proposto" (AMERICAN EDUCATIONAL RESEARCH ASSOCIATION; AMERICAN PSYCHOLOGICAL ASSOCIATION; NATIONAL COUNCIL ON MEASUREMENT IN EDUCATION, 2014, p. 14). Este tipo de evidência também pode ser obtida a partir do processo de apreciação dos itens e do instrumento como um todo por parte de um painel de especialistas (AMERICAN EDUCATIONAL RESEARCH ASSOCIATION; AMERICAN PSYCHOLOGICAL ASSOCIATION; NATIONAL COUNCIL ON MEASUREMENT IN EDUCATION, 2014), que pode analisar, por exemplo, a relevância e a correlação entre os itens e os conteúdos e as habilidades, a adequação da linguagem e do vocabulário e a presença de ambiguidades.

No caso das evidências baseadas nos processos de respostas, estas relacionam-se com os domínios cognitivos estipulados e exigidos pelos instrumentos de avaliação. De acordo com Popham (2017, p. 101), este tipo de evidência é útil para analisar "até que ponto a organização interna de um teste confirma uma avaliação precisa do construto que está supostamente sendo medido", pois "algumas interpretações do construto envolvem suposições mais ou menos explícitas sobre os processos cognitivos aplicados pelos participantes do teste" (AMERICAN EDUCATIONAL RESEARCH ASSOCIATION; AMERICAN PSYCHOLOGICAL ASSOCIATION; NATIONAL COUNCIL ON MEASUREMENT IN EDUCATION, 2014, p. 15).

Nesse sentido, a coleta deste tipo de evidência de validade se dá por meio de análises teóricas e empíricas da adequação entre o construto e a natureza das respostas dos avaliados. Além disso, as evidências baseadas nos processos de respostas também envolvem os procedimentos de correção dos instrumentos, podendo ser recolhidas mediante a análise da consistência dos critérios de avaliação e de interpretação dos avaliadores e do uso pretendido dos resultados (AMERICAN EDUCATIONAL RESEARCH ASSOCIATION; AMERICAN PSYCHOLOGICAL ASSOCIATION; NATIONAL COUNCIL ON MEASUREMENT IN EDUCATION, 2014).

As evidências baseadas na estrutura interna estão relacionadas com o nível de correspondência entre os itens do instrumento de avaliação e os construtos para os quais este instrumento foi construído para medir e sobre os quais as interpretações dos resultados serão baseadas (AMERICAN EDUCATIONAL RESEARCH ASSOCIATION; AMERICAN PSYCHOLOGICAL ASSOCIATION; NATIONAL COUNCIL ON MEASUREMENT IN EDUCATION, 2014). A coleta deste tipo de evidências de validade permite identificar itens que avaliam construtos distintos, os quais, dependendo do objetivo da avaliação, podem interferir na validade do uso destas informações (POPHAM, 2017).

Já as evidências de validade baseadas em outras variáveis compreendem a relação entre os resultados provenientes de um instrumento de avaliação com outro, normalmente, um instrumento de avaliação externa (AMERICAN EDUCATIONAL RESEARCH ASSOCIATION; AMERICAN PSYCHOLOGICAL ASSOCIATION; NATIONAL COUNCIL ON MEASUREMENT IN EDUCATION, 2014). De acordo com os *Standards*, este tipo de evidência fornece "indicadores sobre o grau em que essas relações são consistentes com o construto subjacente às interpretações da pontuação do teste propostas" (AMERICAN EDUCATIONAL RESEARCH ASSOCIATION; AMERICAN PSYCHOLOGICAL ASSOCIATION; NATIONAL COUNCIL ON MEASUREMENT IN EDUCATION, 2014, p. 16). Desta forma, a coleta de evidências baseadas em outras variáveis é útil para a obtenção de informações sobre o quanto um determinado construto foi precisamente medido pelo instrumento de avaliação em questão, a fim de que essas evidências sejam válidas para o uso pretendido (POPHAM, 2017).

Finalmente, as evidências de validade baseadas nas consequências da avaliação correspondem, fundamentalmente, à interpretação dos dados recolhidos e ao uso dos resultados. De acordo com os *Standards*, este tipo de evidência envolve a coleta de informações que auxiliam o avaliador a determinar a adequação da intepretação dos resultados para os usos propostos e desejados (AMERICAN EDUCATIONAL RESEARCH ASSOCIATION; AMERICAN PSYCHOLOGICAL ASSOCIATION; NATIONAL COUNCIL ON MEASUREMENT IN EDUCATION, 2014). A coleta deste tipo de evidências é fundamental para evitar repercussões não intencionais, as quais, geralmente, são negativas (AMERICAN EDUCATIONAL RESEARCH ASSOCIATION; AMERICAN PSYCHOLOGICAL ASSOCIATION; NATIONAL COUNCIL ON MEASUREMENT IN EDUCATION, 2014).

À vista disso, observa-se que, entendida como um conceito unitário, a validade corresponde ao acúmulo de diversos tipos de evidências necessários para amparar a interpretação e o uso dos resultados procedentes da aplicação dos instrumentos de avaliação (AMERICAN EDUCATIONAL RESEARCH ASSOCIATION; AMERICAN PSYCHOLOGICAL ASSOCIATION; NATIONAL COUNCIL ON MEASUREMENT IN EDUCATION, 2014; MESSICK, 1989; MILLER; LINN; GRONLUND, 2009; POPHAM, 2017). Além disso, a atual concepção de validade evidencia, ainda mais, a inadequação do debate a respeito da validade de um instrumento de avaliação, posto que esta é uma característica originada a partir da coleta de dados e de informações que serão atribuídos a um uso específico e não a característica do instrumento de avaliação em si (BROOKHART, 1999; POPHAM, 2017).

Incompatibilidade na aplicação do conceito de validade entre avaliações de sala de aula e testes de larga escala

Embora tenham ocorrido mudanças significativas no conceito de validade, seu desenvolvimento se deu principalmente na perspectiva dos testes padronizados (KANE; WOOLS, 2020). Phakiti e Isaacs (2021, p. 5) corroboram a ideia, alegando que "a psicometria é o campo da medição psicológica que influenciou o desenvolvimento da validade" e que, no século 21, o ato de demonstrar a validade envolvia dominar discussões em psicometria e medição educacional.

De acordo com Brookhart (2003), as discussões realizadas sobre a validade dos instrumentos de avaliação de sala de aula utilizam conceitos baseados na perspectiva psicométrica e nas noções de qualidade desenvolvidas para os testes de larga escala e adaptadas para o contexto de sala de aula. No entanto, as circunstâncias e as condições das avaliações de sala de aula distinguem-se daquelas dos testes em larga escala, tornando-se necessária a discussão da validade no contexto de sala de aula (BROOKHART, 2003; KANE; WOOLS, 2020).

Em 1963, Glaser (1963) já indicava a necessidade da distinção entre as análises das avaliações educacionais e dos testes psicométricos e psicológicos. Em seu artigo, o autor definiu e distinguiu duas categorias de medidas, as medidas referenciadas a critério e as medidas referenciadas a norma: "o que chamarei de medidas referenciadas a critério dependem de um padrão de qualidade absoluto, enquanto o que chamo de medidas referenciadas a norma dependem de um padrão relativo" (GLASER, 1963, p. 519).

Nas avaliações referenciadas à critério, o desempenho dos alunos é comparado com padrões pré-definidos de bom ou ruim, enquanto naquelas referenciadas à norma o desempenho de um aluno é comparado com o desempenho de outros alunos (RUSSEL; AIRASIAN, 2008). Em sua maioria, os testes de larga escala são referenciados à norma, enquanto as avaliações de sala de aula são referenciadas à critério (BROOKHART, 2004). Sendo assim, as avaliações de sala de aula "avaliam o desempenho do aluno em termos de um critério padrão" (GLASER, 1963, p. 520) e não com relação à pontuação dos outros alunos (BROOKHART, 2004), fornecendo informações quanto ao "grau de competência obtido por um aluno em particular, o qual é independente da referência ao desempenho de terceiros" (GLASER, 1963, p. 520).

Para Gipps (2003), diferentes conceitos e interpretações são aplicados nas avaliações referenciadas a critério quando comparadas àquelas referenciadas a norma. Por exemplo, itens de avaliações referenciadas à norma devem discriminar os alunos com maior e menor competência, a fim de dispô-los em uma distribuição normal (GIPPS, 2003). Sendo assim, os itens com um grande número de acertos ou de erros são descartados da análise, já que não contribuem para o valor de discriminação do teste (GIPPS, 2003; SCHUWIRTH; VLEUTEN, 2006).

No entanto, como as avaliações referenciadas a critério não intencionam discriminar os alunos, mas identificar quais as tarefas que eles podem ou não realizar, estes itens não discriminativos devem ser incluídos na análise caso eles sejam importantes para a tomada de decisão do professor (GIPPS, 2003). De acordo com a autora, "o fator importante nos testes referenciados a critério não é a alta discriminação, mas a representação de um *continuum* de tarefas relevantes" (GIPPS, 2003, p. 83).

Popham (1987) alega que o principal ganho das avaliações referenciadas a critério é o fato delas produzirem uma descrição explícita do que está sendo medido, possibilitando o estabelecimento de inferências mais precisas a respeito do significado das suas pontuações. Desta forma, quando referenciadas a critério, as avaliações de sala de aula permitem aos professores redirecionarem suas aulas para as áreas nas quais os alunos não se saíram muito bem (POPHAM, 1987), contribuindo para a melhoria da qualidade do processo de ensino e aprendizagem.

Outra dissonância entre a concepção de validade para os testes de larga escala e para as avaliações de sala de aula está relacionada com a forma como

ocorrem as análises das evidências (BONNER, 2013; KANE; WOOLS, 2020; WHITTINGTON, 1999). Em testes de larga escala, as análises se dão pela coleta de evidências psicométricas (KANE; WOOLS, 2020), as quais envolvem análises quantitativas e modelos estatísticos que frequentemente "aparecem por meio de coeficientes de validade ou tabelas de expectativas e são explicados aos profissionais de educação por meio de gráficos de dispersão, probabilidades, análises fatoriais e fórmulas" (WHITTINGTON, 1999, p. 15).

No entanto, essas análises requerem um conjunto de dados cujo tamanho não é acessível ao nível da avaliação de sala de aula (PHAKITI; ISAACS, 2021; WHITTINGTON, 1999). Bonner (2013) reforça a ideia, acrescentando que a maioria dos métodos quantitativos tradicionalmente utilizados no processo de validação da interpretação das avaliações de larga escala são inviáveis na perspectiva de sala de aula, não só devido ao tamanho das amostras e do número de itens das avaliações, mas também pela falta de qualificação técnica por parte dos professores para realizar tais análises.

Além disso, muitos autores argumentam que, na maioria das vezes, as avaliações de sala de aula destinam-se a produzir julgamentos qualitativos sobre os pontos fortes e fracos dos alunos e sobre a capacidade destes em desempenharem ou não uma tarefa, ao invés de ou além de, gerar pontuações, classificações numéricas, dimensionamentos e escalas (BONNER, 2013; DEPRESBITERIS; TAVARES, 2009; KANE; WOOLS, 2020; RUSSEL; AIRASIAN, 2008).

A padronização também é um fator divergente entre as duas perspectivas de avaliação. Messick (1989) alega que, enquanto nos testes em larga escala preza-se pela padronização, as avaliações de sala de aula valorizam a individualização. Whittington (1999, p. 21) corrobora a ideia, alegando que para os professores "é preferível adaptar os critérios às necessidades e às características de cada criança do que manter um padrão comum". De acordo com o autor, os professores que prezam pela individualização não adotarão as noções de validade em suas práticas a não ser que essas sejam utilizadas para promover o aprendizado e o bem-estar dos seus alunos (WHITTINGTON, 1999).

Por fim, o contexto apresenta-se como um aspecto divergente e, talvez, o mais significativo entre os dois formatos de avaliação. De acordo com Brookhart (2003), o contexto modifica a natureza da medida, inviabilizando a utilização das noções de validade dos testes de larga escala para as avaliações de sala de aula. No

enquadramento psicométrico, o contexto de medida é construto-irrelevante. Assim, as especificações do conteúdo descrevem um domínio, a administração pode ser padronizada e as pontuações podem ser equiparadas entre contextos e formatos de avaliação (BROOKHART, 2003).

Já no âmbito da sala de aula, o contexto de medida é construto-relevante. Deste modo, as especificações dos conteúdos refletem não apenas o domínio, mas também os objetivos de aprendizagem e, principalmente, o processo instrucional (BROOKHART, 2003), tornando-se um contexto específico, o qual não é passível de ser comparado com outros.

Kane e Wools (2020) respaldam essa ideia, afirmando que no contexto de testes de larga escala as características técnicas ou psicométricas, como a padronização e a consistência, desempenham um papel central. Isso porque seus resultados são utilizados para a tomada de decisões de grande impacto no sistema educativo, como, por exemplo, a elaboração e a implementação de políticas públicas ou a seleção de candidatos em um concurso ou em um vestibular. Já no contexto de sala de aula, as decisões envolvem inferências de uso local, cujos "resultados precisam ser práticos e úteis no cumprimento do principal objetivo da avaliação de sala de aula: promover a eficácia do ensino e da aprendizagem" (KANE; WOOLS, 2020, p. 12).

Além disso, no contexto psicométrico o processo de medida é externo às inferências realizadas e às ações tomadas, enquanto no contexto de sala de aula o processo de medida é interno (BROOKHART, 2003). Na conjuntura dos testes de larga escala, os alunos são considerados como sujeitos passivos e a validade é assumida como uma série de inferências significativas sobre os seus desempenhos para usos específicos que, normalmente, incluem a seleção, o dimensionamento e a elaboração de relatórios de prestação de contas para as instituições de ensino (BROOKHART, 2003). Em contrapartida, na perspectiva de sala de aula, as inferências e as ações são internas ao processo de medida. Sendo assim, os alunos são sujeitos ativos que interpretam as informações e tomam decisões junto com os professores, com o objetivo de melhorar a qualidade do ensino e do aprendizado (BROOKHART, 2003).

Nesse sentido, o contexto define os métodos de avaliação, os critérios de seleção e de correção, a forma de *feedback*, a instrução e a percepção dos professores com relação aos seus alunos (BROOKHART, 2003). Em sala de aula, as

práticas de instrução e o entendimento dos professores sobre o assunto e sobre os alunos são relevantes para a validade do uso e da interpretação dos resultados das avaliações, já que "as experiências instrucionais afetam como a linguagem e as tarefas são interpretadas e como as expectativas são definidas" (BROOKHART, 2003, p. 6).

Estratégias para a coleta de evidências de validade adequadas para avaliações de sala de aula

Diante do exposto, percebe-se que a concepção tradicional de validade falha ao atender as necessidades dos professores em sala de aula (BONNER, 2013). Sendo assim, emerge a necessidade de estabelecer estratégias que possibilitem a coleta de evidências de validade adequadas para as interpretações e para os usos dos resultados das avaliações de sala de aula e que levem em consideração o contexto, sem a perda das características técnicas de qualidade (BROOKHART, 2003).

Como visto anteriormente, atualmente existem cinco tipos de evidências de validade e dificilmente a coleta de apenas um tipo permitirá ao professor tomar uma decisão precisa, sendo "necessário recolher uma variedade de tipos de evidências de validade" (POPHAM, 2017, p. 100). No contexto de sala de aula, três tipos de evidências apresentam maior importância, a saber: as evidências baseadas no conteúdo, as evidências baseadas no processo de resposta e as evidências baseadas nas consequências da avaliação (POPHAM, 2017).

As evidências baseadas na estrutura interna e as evidências baseadas em outras variáveis raramente são utilizadas como fontes de evidência de validade das interpretações das avaliações de sala de aula. As primeiras porque partem do princípio de que os construtos devem ser unidimensionais e, muitas vezes, é difícil isolar construtos das avaliações de sala de aula (POPHAM, 2017), já que, normalmente, estas apresentam mais de um propósito e "cada abordagem do conteúdo do teste pode ser válido para alguns propósitos, mas não para outros" (BROOKHART, 2003, p. 10). Os *Standards* respaldam a ideia, alegando que, "raramente, se é que há, existe um único significado possível que possa ser associado à pontuação ou a um padrão de respostas de um teste" (AMERICAN EDUCATIONAL RESEARCH ASSOCIATION; AMERICAN PSYCHOLOGICAL ASSOCIATION; NATIONAL COUNCIL ON MEASUREMENT IN EDUCATION, 2014, p. 11).

Já as evidências baseadas em outras variáveis apresentam uma grande dificuldade de coleta no contexto de sala de aula: raramente são encontrados instrumentos semelhantes, capazes de avaliar exatamente os mesmos construtos da avaliação proposta pelo professor (POPHAM, 2017). Além disso, com relação ao desempenho dos alunos, a maioria das avaliações de sala de aula tem como objetivo avaliá-lo naquele determinado momento e não predizer os resultados para momentos futuros (MILLER; LINN; GRONLUND, 2009; POPHAM, 2017).

Dentre as três evidências de maior influência para o contexto de sala de aula, aquelas baseadas no conteúdo são as mais relevantes e estão no centro do processo educacional (BROOKHART, 2003; DEPRESBITERIS; TAVARES, 2009; MCMILLAN, 1999; POPHAM, 2017). Popham (2017) afirma que, geralmente, existem duas formas para garantir que as evidências derivadas dos conteúdos das avaliações sejam válidas para os usos propostos: os cuidados na elaboração dos instrumentos de avaliação e a revisão externa.

No que diz respeito aos cuidados na elaboração dos instrumentos, o autor reconhece a necessidade de aplicar um conjunto de procedimentos que garantam que os objetivos propostos estejam propriamente expressos no instrumento de avaliação (POPHAM, 2017). De acordo com Brookhart (1999), os resultados de uma avaliação correspondem aos objetivos de aprendizagem quando o conteúdo e os domínios cognitivos representam aqueles dos objetivos e da instrução, quando o número de itens ou tarefas é suficiente e representativo para os objetivos de aprendizado determinados e para os propósitos de sua utilização e quando esses itens ou tarefas são claros para os alunos.

Diversos autores sugerem a elaboração de uma tabela de especificação como uma estratégia para garantir a correspondência entre os objetivos e o instrumento de avaliação, assim como, para assegurar a representatividade do conteúdo (BROOKHART, 1999; DEPRESBITERIS; TAVARES, 2009; MILLER; LINN; GRONLUND, 2009; POPHAM, 2017; RUSSEL; AIRASIAN, 2008). A tabela de especificação é uma ferramenta simples e eficaz, amplamente utilizada na construção de instrumentos de avaliação (MILLER; LINN; GRONLUND, 2009), o qual auxilia os professores a evidenciar a representação do conteúdo e a selecionar os itens de forma adequada (BROOKHART, 1999).

Além disso, a tabela de especificação "melhora e documenta a validade, mapeando a maneira como o instrumento de avaliação, como um todo, representa

o domínio avaliado" (BONNER, 2013, p. 93). Popham (2017) reforça a importância da documentação, afirmando que é ela que, de fato, constitui uma importante forma de evidência de validade para usos específicos. De acordo com o autor, "quanto mais importante a avaliação – isto é, quanto mais importantes são os seus usos –, maior é a necessidade de que as atividades de desenvolvimento da avaliação sejam documentadas" (POPHAM, 2017, p. 106).

Em sua forma mais simples, uma tabela de especificação deve apresentar as competências ou os conteúdos que devem ser considerados, as habilidades ou os domínios cognitivos e a quantidade de itens a serem elaborados para verificá-los (DEPRESBITERIS; TAVARES, 2009), conforme exemplificado na Tabela 1.

Tabela 1 - Exemplo de organização básica de uma tabela de especificação

Dimensão do conteúdo	**Domínio cognitivo**			**Total**
	Compreender	**Analisar**	**Avaliar**	
Natureza da ciência (NC)	4	1	1	6
Impacto da ciência e da tecnologia na sociedade (ICTS)	2	1	4	7
Conteúdo da ciência (CC)	15	15	21	51
Total	21	17	26	64

Fonte: (COPPI; FIALHO; CID, 2022, p. 109).

Já na etapa de construção do instrumento, cuidados devem ser tomados para que este corresponda às especificações determinadas e para que os itens estejam claros para os alunos (MILLER; LINN; GRONLUND, 2009). Para isso, é necessário que os professores possuam o mínimo de conhecimento e de competências referentes aos tipos de avaliação e de itens e dos seus diferentes usos (STIGGINS, 1999). Brookhart (2004) corrobora a ideia, alegando que, a fim de avaliar seus alunos, os professores precisam saber quais as opções de avaliação estão disponíveis, como elaborar uma avaliação e como interpretar e utilizar as informações recolhidas.

Vale ressaltar, contudo, que diversas pesquisas revelam o fato de que os professores apresentam baixos índices de letramento em avaliação (PASTORE; ANDRADE, 2019; POPHAM, 2017; STIGGINS, 2001), definido por Popham (2011, p. 267) como "o entendimento individual dos conceitos e dos procedimentos fundamentais de avaliação, considerados passíveis de influenciar as decisões educacionais". De acordo com Pastore e Andrade (2019), muitos dos atuais professores pouco sabem sobre avaliação educacional. Os autores alegam, também, que esta lacuna nas

competências em avaliação é muito compreensível, declarando que uma das razões está nos programas de formação inicial dos professores, nos quais não há muita exigência em relação à avaliação pedagógica e, em muitos casos, o único contato com os conceitos e com as práticas de avaliação educacional se dá em algumas poucas aulas de psicologia ou em alguma unidade na disciplina de metodologia (PASTORE; ANDRADE, 2019).

Uma vez elaborado, o instrumento deve passar pelo processo de revisão externa (DEPRESBITERIS; TAVARES, 2009; POPHAM, 2017). No contexto de sala de aula, esse processo pode ser realizado por um ou mais professores da mesma disciplina, que podem se encarregar de revisar e verificar a adequação do instrumento de avaliação quanto ao vocabulário, às instruções, à dificuldade dos itens, à presença de ambiguidades, ao construto e à cobertura representativa dos conteúdos (POPHAM, 2017).

Os cuidados nas etapas de elaboração e de revisão dos instrumentos de avaliação são fundamentais para a validade das interpretações e dos usos pretendidos. Caso o instrumento contenha itens com vocabulário inapropriado, instruções pouco claras ou ambiguidades, a tomada de decisão e a utilização dos resultados será comprometida (MILLER; LINN; GRONLUND, 2009). Bons instrumentos de avaliação fornecem benefícios significativos para todos os utilizadores dos seus resultados (AMERICAN EDUCATIONAL RESEARCH ASSOCIATION; AMERICAN PSYCHOLOGICAL ASSOCIATION; NATIONAL COUNCIL ON MEASUREMENT IN EDUCATION, 2014), que, no contexto de sala de aula, são os alunos, os professores, os coordenadores pedagógicos e os pais ou responsáveis (BONNER, 2013).

Quanto às evidências baseadas nos processos de resposta, suas análises não se limitam às respostas dos alunos, mas, também, aos critérios de correção utilizados pelos avaliadores, que podem gerar vieses na interpretação e no uso dos resultados (AMERICAN EDUCATIONAL RESEARCH ASSOCIATION; AMERICAN PSYCHOLOGICAL ASSOCIATION; NATIONAL COUNCIL ON MEASUREMENT IN EDUCATION, 2014). Levando em consideração a inviabilidade de realizar análises psicométricas, os professores podem utilizar diferentes estratégias, como por exemplo, o emprego da rubrica (BONNER, 2013), cuja utilização "constitui um procedimento bastante simples para apoiar a avaliação de uma grande diversidade de produções e desempenhos dos alunos" (FERNANDES, 2021, p. 4).

A rubrica é uma escala de classificação descritiva que contém um conjunto de critérios de desempenho, os quais devem ser utilizados para auxiliar alunos e professores a focar no construto que será valorizado e considerado no instrumento de avaliação (DEPRESBITERIS; TAVARES, 2009; RUSSEL; AIRASIAN, 2008). Fernandes (2021) acrescenta que este conjunto de critérios traduz o que é desejável que os alunos aprendam, afirmando ser necessário estabelecer, para cada critério, um número de descrições de níveis de desempenho. De acordo com Russel e Airasian (2008, p. 223), as descrições

> [...] ajudam os professores a focar a instrução e a atribuição de escores ao trabalho dos alunos nos aspectos importantes inclusos na rubrica. A descrição também ajuda os alunos a melhor compreender o que os professores esperam deles para determinado o desempenho o produto.

Neste sentido, uma rubrica é, normalmente, composta por quatro elementos: a descrição geral da tarefa, indicando o objeto de avaliação; os critérios; os níveis de descrição do desempenho relativamente a cada critério, representados pelos indicadores ou descritores de desempenho; e a definição de uma escala, em que a cada ponto – numeral, letra do alfabeto ou percentagem – corresponda um determinado indicador ou descritor de desempenho (FERNANDES, 2021), conforme apresentado nos exemplos dos Quadros 1 e 2.

Quadro 1 - Exemplo da organização geral de uma rubrica de avaliação

Descrição Geral da Tarefa (Objeto de Avaliação)			
Critérios	**Níveis de Desempenho**		
	1	2	3
Critério 1	Descritor ou Indicador do Desempenho	Descritor ou Indicador do Desempenho	Descritor ou Indicador do Desempenho

Fonte: (FERNANDES, 2021, p. 9).

Quadro 2 - Exemplo de rubrica para a avaliação de mapas conceituais

Desempenho no âmbito dos Mapas Conceptuais			
Critérios	**Níveis de Desempenho**		
	1	2	3
Relações entre Conceitos	Relações entre os conceitos não são claras. Desorganização das componentes e subcomponentes.	Relações entre os conceitos são evidentes. Componentes e subcomponentes nem sempre organizadas.	Relações claras entre os conceitos. Componentes e subcomponentes hierarquicamente organizadas

Fonte: (FERNANDES, 2021, p. 10).

Quando elaboradas e utilizadas adequadamente, as rubricas valorizam a consistência entre as respostas dos alunos e os critérios de avaliação dos professores (FERNANDES, 2021; RUSSEL; AIRASIAN, 2008). Para Blass e Irala (2021, p. 205), as rubricas "surgiram como estratégia para integrar os estudantes, padronizar os critérios, imprimir maior transparência ao processo e efetivo acompanhamento de todas as etapas previstas, até a sua finalização". Além disso, a utilização da rubrica é um bom método para reduzir o viés na correção e no *feedback* dos resultados dos instrumentos de avaliação, auxiliando no processo de interpretação e do uso das evidências de validade baseadas nos processos de resposta (BONNER, 2013).

Vale ressaltar que, conforme afirma Brookhart (2013) acerca da natureza das rubricas, embora estas permitam avaliar, as rubricas são descritivas e não avaliativas. Isto, para Fernandes (2021, p. 4), deve-se ao fato de que "em vez de julgar o desempenho, professores e alunos verificam qual a descrição que melhor o pode representar. Assim, antes do mais, as rubricas permitem desenvolver uma avaliação de referência criterial" e não de referência normativa, como aquelas em que que o desempenho dos alunos é comparado com uma média ou com um grupo.

Blass e Irala (2021) afirmam que utilizar as rubricas na avaliação pode trazer elementos que evidenciam, por exemplo, a sua aplicabilidade em contextos reais, a possibilidade de ser efetuada de forma contínua e a capacidade de obter informações sobre o processo. Nesse sentido, as rubricas podem ser utilizadas tanto no contexto de avaliação formativa, ou avaliação para as aprendizagens, quando para a avaliação somativa, ou avaliação das aprendizagens, inserindo-se "no contexto da avaliação pedagógica, pois são utilizadas nas salas de aula e podem contribuir para apoiar as aprendizagens dos alunos e o ensino dos professores através daquelas duas modalidades de avaliação" (FERNANDES, 2021, p. 5).

Contudo, para que o uso da rubrica seja eficiente, esta deve ser partilhada com os alunos, para que estes possam se preparar para as avaliações (FERNANDES, 2021; RUSSEL; AIRASIAN, 2008). Além disso, Fernandes (2021) afirma que, sempre que possível, os alunos devem participar da identificação e dos critérios e da descrição dos desempenhos que serão considerados relevantes para as aprendizagens que estão sendo desenvolvidas.

Já com relação às evidências de validade baseadas nas consequências das avaliações, dois aspectos são de interesse dos professores: a sub-representação e a irrelevância do construto (AMERICAN EDUCATIONAL RESEARCH ASSOCIATION;

AMERICAN PSYCHOLOGICAL ASSOCIATION; NATIONAL COUNCIL ON MEASUREMENT IN EDUCATION, 2014; MILLER; LINN; GRONLUND, 2009). A sub-representação do construto "descreve uma deficiência de uma avaliação que falha ao recolher aspectos importantes do construto que está sendo avaliado", enquanto a irrelevância refere-se a "uma fraqueza da avaliação na qual as pontuações dos examinados são influenciadas por fatores alheios ao construto que está sendo avaliado" (POPHAM, 2017, p. 97).

De acordo com Kane e Wools (2020, p. 12), os dois aspectos "são relevantes principalmente em termos de seu impacto na eficácia da avaliação no apoio ao ensino e à aprendizagem". Além disso, os autores afirmam que "uma conclusão imprecisa sobre a habilidade de um aluno pode não ser catastrófica, mas provavelmente não será útil para o planejamento de instruções futuras e, portanto, no apoio à aprendizagem" (KANE; WOOLS, 2020, p. 12).

Esse tipo de evidência de validade depende da plausibilidade da interpretação e da adequação dos seus usos (AMERICAN EDUCATIONAL RESEARCH ASSOCIATION; AMERICAN PSYCHOLOGICAL ASSOCIATION; NATIONAL COUNCIL ON MEASUREMENT IN EDUCATION, 2014). Para se atingir a validade pretendida, é necessário que os professores determinem e especifiquem os objetivos da intepretação e do uso dos resultados antes de elaborar o instrumento de avaliação (RUSSEL; AIRASIAN, 2008). Desta forma, eles serão capazes de desenvolver instrumentos que atendam às especificações planejadas e poderão avaliar o quanto as interpretações e usos são justificados pelos resultados das avaliações (AMERICAN EDUCATIONAL RESEARCH ASSOCIATION; AMERICAN PSYCHOLOGICAL ASSOCIATION; NATIONAL COUNCIL ON MEASUREMENT IN EDUCATION, 2014; KANE; WOOLS, 2020).

A própria tabela de especificação também pode ser utilizada como ferramenta para auxiliar os professores a evitar a sub-representação e a irrelevância do construto no instrumento de avaliação. Sua elaboração "oferecerá maior segurança ao docente porque indicará uma amostra representativa das habilidades, competências e conteúdos desenvolvidos no processo de ensino, significativos para a aprendizagem do aluno" (DEPRESBITERIS; TAVARES, 2009, p. 76).

Considerações finais

A principal característica referente à qualidade da avaliação de sala de aula é a validade. No entanto, a noção de validade até então utilizada para este

contexto baseia-se no contexto dos testes de larga escala, uma perspectiva psicométrica.

Com o intuito de definir o conceito de validade, apontar a incompatibilidade da aplicação deste conceito entre os testes de larga escala e as avaliações de sala de aula e de apresentar estratégias de coleta de evidências de validade para a interpretação propícia e para o uso adequado dos resultados das avaliações de sala de aula, este artigo, atendendo a uma breve revisão do estado da arte sobre o tema, esclareceu o atual conceito de validade, apresentou argumentos que inviabilizam a aplicação da noção de validade utilizada em testes de larga escala nas avaliações de sala de aula e indicou estratégias que podem auxiliar os professores a recolher os três tipos de evidência de validade mais importantes para o contexto de sala de aula.

Conforme exposto no decorrer do artigo, a validade é um atributo relacionado com o acúmulo de diversos tipos de evidências que asseguram a interpretação e o uso dos resultados de um instrumento de avaliação. Por esse motivo, nenhuma avaliação possui ou não validade, são as interpretações e, principalmente, os usos dos resultados que podem ou não ser válidos, de acordo com as evidências recolhidas.

Apesar dos tipos de evidência de validade serem os mesmos para os testes de larga escala e para as avaliações de sala de aula, as estratégias de coleta das evidências de validade são divergentes. Isso porque, enquanto os testes de larga escala, geralmente, são referenciados a norma, carecem de uma análise quantitativa, psicométrica e estatística, visam a padronização e o processo de medida é externo às decisões tomadas; as avaliações de sala de aula, na sua maioria, são referenciadas a critério e passíveis de análises qualitativas, visam a individualização e o processo de medida é interno às decisões que serão tomadas.

Em síntese, no contexto de sala de aula, três tipos de evidência de validade estão mais suscetíveis à coleta pelos professores: as evidências baseadas no conteúdo, as evidências baseadas nos processos de resposta e as evidências baseadas nas consequências das avaliações. De modo geral, determinar, explicitar e compartilhar os objetivos da avaliação; seguir procedimentos adequados para a elaboração do instrumento de avaliação – o conteúdo e os domínios cognitivos que serão avaliados devem representar àqueles dos objetivos e da instrução, o avaliador deve elaborar a tabela de especificação e o instrumento deve ser revisado pelos pares –; e adotar mecanismos para a análise das respostas dos alunos – o emprego

da rubrica e a análise de sub-representação e irrelevância do construto, por exemplo – são estratégias eficazes para a coleta de evidências de validade para a interpretação e o uso apropriado dos resultados provenientes dos instrumentos de avaliação de sala de aula.

Por fim, sugere-se que estudos posteriores desenvolvam e validem diferentes metodologias que possibilitem aos professores recolherem evidências de validade compatíveis com o contexto das avaliações de sala de aula, ou seja, com o objetivo de fornecer evidências válidas e precisas sobre a situação dos alunos, a fim de tomar boas decisões educacionais a seu respeito. Propõe-se, também, a implementação de disciplinas específicas de avaliação nos cursos de formação de professores – licenciaturas – e a elaboração de cursos de formação continuada com a temática do letramento em avaliação para os professores em serviço.

Agradecimentos

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

Referências

ALAM, M. J.; AKTAR, T. Large-scale educational assessment against classroom assessment: pedagogy and measurement paradigm in EFL classroom. *Journal of Education & Social Policy*, [*S. l.*], v. 7, n. 3, p. 160-167, 2020.

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

BLASS, L.; IRALA, V. B. Usar ou não usar rubricas? Um olhar para as práticas avaliativas a partir dos desempenhos discentes. *Revista Insignare Scientia*, [*S. l.*], v. 4, n. 4, p. 203-226, 2021.

BONNER, S. M. Validity in classroom assessment: purposes, properties, and principles. *In*: MCMILLAN, J. H. (ed.). *SAGE handbook of research on classroom assessment*. Thousand Oaks, CA: SAGE Publications, Inc., 2013. p. 87-106.

BROOKHART, S. M. *The art and science of classroom assessment*: the missing part of pedagogy. Washington, DC: The George Washington University, Graduate School of Education and Human Development, 1999.

BROOKHART, S. M. Developing measurement theory for classroom assessment purposes and uses. *Educational Measurement*: Issues and Practice, v. 22, n. 4, p. 5-12, 2003.

BROOKHART, S. M. Assessment theory for college classrooms. *New Directions for Teaching and Learning*, San Francisco, n. 100, p. 5-14, 2004.

BROOKHART, S. M. *How to create and use rubrics for formative assessment and grading*. Alexandria: ASCD, 2013.

COPPI, M. A.; FIALHO, I.; CID, M. Evidências de validade baseadas no conteúdo de um instrumento de avaliação da literacia científica. *Cadernos de Pesquisa*, [Rio de Janeiro], v. 29, n. 2, p. 99-127, 2022.

CRONBACH, L. J. Five perspectives on validation argument. *In*: WAINER, H.; BRAUN, H. (ed.). *Test validit*. Hillsdale, NJ: Erlbaum, 1988. p. 3-17.

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

DEPRESBITERIS, L.; TAVARES, M. R. *Diversificar é preciso...*: instrumentos e técnicas de avaliação de aprendizagem. São Paulo: Senac São Paulo, 2009.

FERNANDES, D. *Rubricas de avaliação*. Lisboa: Ministério da Educação/Direção-Geral da Educação, 2021.

GIPPS, C. V. *Beyond testing*: towards a theory of educational assessment. Washington, DC: The Falmer Press, 2003.

GLASER, R. instructional technology and the measurement of learning outcomes: some questions. *American Psychologist*, Washington, DC, n. 18, p. 519-521, 1963.

KANE, M. T. Validation. *In*: BRENNAN, R. L. (ed.). *Educational measurement*. 4. ed. Washington, DC: Rowman and Littlefield Publishers Inc, 2006. p. 17-63.

KANE, M. T.; WOOLS, S. Perspectives on the validity of classroom assessments. *In*: BROOKHART, S. M.; MCMILLAN, J. H. (ed.). *Classroom assessment and educational measurement*. New York: Routledge, 2020. p. 11-26.

MCMILLAN, J. H. Establishing high quality classroom assessments. *ERIC Document Reproduction Service*, [New York], v. 429, n. 146, p. 1-15, 1999.

MESSICK, S. Validity. *In*: LINN, R. L. (ed.). *Educational measurement*. 3. ed. New York: Macmillan, 1989. p. 3-209.

MILLER, M. D.; LINN, R. L.; GRONLUND, N. E. *Measurement and assessment in teaching*. 10. ed. New Jersey: Pearson Education, 2009.

PASTORE, S.; ANDRADE, H. L. Teacher assessment literacy: a three-dimensional model. *Teaching and Teacher Education*, New York, v. 84, p. 128-138, 2019.

PHAKITI, A.; ISAACS, T. Classroom assessment and validity: psychometric and edumetric approaches. *The European Journal of Applied Linguistics and TEFL*, [*S. l.*], v. 10, n. 1, p. 3-24, 2021.

POPHAM, W. J. Two-plus decades of educational objectives. *International Journal of Educational Research*, Oxford, v. 11, n. 1, p. 31-41, 1987.

POPHAM, W. J. *Transformative assessment in action*: an anside look at applying the process. Alexandria: ASCD, 2011.

POPHAM, W. J. *Classroom assessment*: what teachers need to know. 8. ed. Los Angeles: Pearson, 2017.

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

SCHUWIRTH, L. W. T.; VLEUTEN, C. P. M. van der. A plea for new psychometric models in educational assessment. *Medical Education*, [*S. l.*], v. 40, p. 296-300, 2006.

STIGGINS, R. J. Evaluating classroom assessment training in teacher education programs. *Educational Measurement*: Issues and Practice, [*S. l.*], v. 18, n. 1, p. 23-27, 1999.

STIGGINS, R. J. The unfulfilled promise of classroom assessment. *Educational Measurement*: Issues and Practice, [*S. l.*], v. 20, n. 3, p. 6-10, 2001.

WHITTINGTON, D. Making room for values and fairness: teaching reliability and validity

in the classroom context. *Educational Measurement*: Issues and Practice, [*S. l.*], n. 18, p. 14-22, 1999.