5

VALIDADE

Juliana Cerentini Pacico
Claudio Simon Hutz

Algumas exigências acerca dos testes devem ser satisfeitas a fim de que eles possam ser considerados adequados para uso. Uma delas se refere ao quanto o teste é legítimo com relação àquilo que mede. Quando se utilizam variáveis como peso, por exemplo, sabemos que uma balança mede essa variável. O instrumento, a balança, é legítimo, pois mede peso. Essa ideia está relacionada ao conceito de validade.

A utilização do conceito de validade acompanha a história do desenvolvimento dos testes. Embora vários autores se refiram ao termo, ele foi utilizado com diferentes significados. Houve algumas tentativas de uniformizar e formalizar o conceito de validade e de procedimentos para validação. Entretanto, algumas não tiveram sucesso. Apenas em 1921, nos Estados Unidos, com o trabalho de um comitê, a National Association of Directors of Educational, foi possível concluir essa tarefa (Newton & Shaw, 2014). A definição clássica de validade foi apresentada no relatório resultante: validade se refere ao grau em que um teste mede aquilo que se propõe a medir (Buckingham, 1921; Markus & Borsboom, 2013). Isso significa que um teste é válido quando os itens medem os comportamentos que são a expressão do traço latente que se deseja mensurar. Os itens (idealmente) devem refletir o traço latente como se fosse um espelho, no que se refere ao conceito, ao conteúdo e às relações com outras variáveis. Em função de a validade, bem como a fidedignidade (que será discutida em outro capítulo deste livro), serem questões centrais tão fundamentais para a avaliação, a American Psychological Association (APA) produziu um relatório que trata desse assunto com detalhes.

Em 1954, a APA, juntamente com a American Educational Research Association (AERA) e o National Council on Measurement in Education (NCME), publicou a primeira versão dos padrões norte-americanos para testes: *Technical*

Recommendations for Psychological Tests and Diagnostic Techniques. Nesse documento, a validade foi classificada em três tipos: Validade de Conteúdo, Validade de Critério (preditiva ou concorrente) e Validade de Construto (Zumbo & Chan, 2014).

Em 1955, Cronbach e Meehl publicaram um artigo dando ênfase à validade de construto. Eles sugeriram, com esse trabalho, que uma forma diferente de abordar o fenômeno de interesse deveria ser utilizada. Enquanto a validade de conteúdo e de critério eram dadas pela proximidade entre o domínio examinado e o domínio que se pretendia examinar, na validade de construto, a lógica era outra. Ela estava relacionada ao quanto os construtos hipotéticos poderiam explicar os escores de um teste. Assim, para os dois primeiros tipos de validade, parte-se da teoria para o teste, os itens do teste devem cobrir um determinado conteúdo (validade de conteúdo) e relacionam-se de maneira definida com um critério (validade de critério). Na validade de construto, parte-se de uma hipótese, e os itens podem ou não confirmá-la. Assim, o problema não está em descobrir o construto a partir da representação comportamental (teste), mas em verificar se ele é uma representação legítima do construto. Cronbach e Meehl (1955), além de argumentarem que o foco deve ser a validade do construto, enfatizaram a importância de uma rede nomológica como forma de construção de teorias sobre o fenômeno psicológico de interesse.

A visão de validade classificada em três categorias foi a base para os *Standards for Educational and Psychological Testing* (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 1985). Essas normas influenciaram o desenvolvimento de outros *guidelines* e legislação sobre testagem, e muitos livros foram escritos tomando-as como referência. Além disso, elas estimularam a criação de outras categorias de validade, já que diferentes tipos de validade eram considerados adequados a diferentes testes (Markus & Borsboom, 2013). Foi nesse contexto que Campbell e Fiske (1959) introduziram a Validade Convergente e a Validade Discriminante.

Enquanto alguns autores propunham a noção de validade como composta por diferentes categorias, havia, ao mesmo tempo, um movimento que buscava a unificação dessas categorias. Loevinger (1957) e Cronbach (1971) sugeriram unificar as diferentes categorias como subtipos da validade de construto. Essa ideia não teve repercussão muito significativa por algum tempo, mas recebeu destaque com a publicação de Messick (1989), que defendeu uma visão unificada de validade, similar àquela apresentada por Cronbach (1971) e diferente da visão anterior, que propunha três categorias. Messick (1989) posicionou-se a favor de uma definição de validade que envolvesse um julgamento avaliativo integrado do grau em que as evidências empíricas e bases teóricas apoiam a adequação e o significado das inferências e ações baseadas nos escores dos

testes. O entendimento de Messick influenciou o campo teórico da avaliação e refletiu-se nos *Standards for Educational and Psychological Testing* (AERA, APA & NCME, 1999). Segundo essas normas, validade pode ser entendida como o grau em que as evidências e a teoria corroboram a interpretação dos escores de um teste obtidos pelo seu uso proposto. Ou seja, a validade é dada pelo grau em que todas as evidências de validade obtidas corroboram para interpretação dos escores de um teste. Essa perspectiva vem ganhando força entre os pesquisadores. De acordo com esse pensamento, não há uma fonte única de evidência de validade que seja suficiente para dar conta de todos os aspectos que precisam ser considerados para se admitir que a validade foi alcançada. Os diferentes tipos de evidências de validade cobrem aspectos distintos que devem ser considerados para que a validade possa ser alcançada.

Há vários pontos em que os autores concordam sobre validade e práticas de validação. Validade e validação são os tópicos mais fundamentais quando se fala em instrumentos de mensuração. Validade refere-se à qualidade das inferências, conclusões e decisões tomadas com base nos escores obtidos pelo uso de um instrumento. Validação é o processo em que se busca dar evidências de validade que apoiem a adequação, o significado e a utilidade das decisões tomadas com base nas inferências feitas a partir dos escores obtidos do teste (Zumbo & Chan, 2014). Embora exista essa nova perspectiva em validade, neste capítulo, serão apresentadas as três categorias clássicas de validade. Ao final, a definição mais moderna será discutida, mas é muito importante ressaltar que essas três categorias são muito importantes e fundamentais para determinar se um teste pode (ou deve) ser utilizado para um determinado fim, com um grupo específico de pessoas.

VALIDADE DE CONTEÚDO

A validade de conteúdo se refere ao quanto o teste pode ser uma amostra representativa dos comportamentos que são a expressão do traço latente em questão, ou, em outras palavras, se os itens do teste se constituem em uma amostra representativa do universo de itens do construto. Alguns testes são planejados para coletar amostras de comportamentos que se relacionam a inferências que se deseja fazer a partir dos escores obtidos, como no caso dos testes de desempenho. Esse tipo de validade somente é aplicável quando se pode definir *a priori* uma amostra de comportamentos que são capazes de representar o universo por meio do qual o traço latente se expressa (Urbina, 2007). O teste será válido, do ponto de vista do conteúdo, se a amostra de comportamentos selecionada para representar o universo de comportamentos por meio dos quais o traço latente se expressa for representativa. Por exemplo, um professor de Avaliação Psicológica dá um curso composto por cinco aulas com os seguintes conteúdos,

cada um trabalhado em uma aula: histórico da avaliação, construção de testes, adaptação de testes, validade e ética. Assim, há um conjunto finito de conteúdos que foram estudados. Esse professor irá avaliar o quanto os alunos apreenderam os conteúdos por meio de uma prova. Como é possível avaliar se essa prova terá validade de conteúdo? Bem, para isso, a prova deve conter um conjunto de questões que seja uma amostra representativa dos conteúdos dados em aula. A prova deverá ter questões sobre assuntos tratados em todas as cinco aulas, caso contrário, uma das aulas poderia ficar sub ou sobrerrepresentada com relação às demais. Por exemplo, se a prova tiver 10 questões, e 5 forem sobre história da avaliação, 3 sobre validade e 2 sobre ética, a prova não terá validade de conteúdo adequada. Entretanto, não seria necessário ter duas questões sobre cada tópico. É possível ter uma única questão mais complexa sobre um tópico e três questões mais simples sobre outro. O que garante a validade de conteúdo é a cobertura adequada de todos os tópicos.

As técnicas de validação são essenciais para que a validade de conteúdo seja atingida. A preocupação com ela começa antes mesmo que se construam ou se adaptem os itens para testar o construto. É preciso que seja feito um exame sistemático do construto que se deseja avaliar, a fim de que, ao determinar o conteúdo que se deseja testar, ele esteja corretamente definido. Tendo em mente essa definição, todos os aspectos que compõem o construto devem estar representados. Vejamos um exemplo: se um pesquisador decidir desenvolver um instrumento para avaliar esperança, por exemplo, ele pesquisará a literatura relacionada ao assunto e consultará os pesquisadores e peritos que trabalham nessa área, buscando definir todos os aspectos do construto. Possivelmente, ele concluirá que esperança é um estado emocional que emerge da interação entre desejo e expectativa. Staats (1989) trabalhou com esse conceito e desenvolveu a The Hope Index para avaliar o quão esperançoso o sujeito poderia ser. Assim, no instrumento, todos os aspectos da esperança devem ser avaliados (desejos e expectativas). O pesquisador deve evitar super-representação de um aspecto ou sub-representação de outro, especialmente quando é difícil desenvolver itens para cobrir um deles. A autora do instrumento original apontou, ainda, para a existência de dois outros aspectos da esperança: esperança autocentrada e esperança altruísta. A autocentrada refere-se a objetivos e desejos relacionados ao próprio sujeito. A esperança altruísta é composta por desejos relacionados a outras pessoas (como amigos, familiares, pessoas em geral) e a circunstâncias globais (paz universal, prosperidade global). Esses aspectos devem ser representados por itens dentro do instrumento. É preciso, então, que se tenha itens para desejos, para expectativas, para esperança autocentrada e para esperança altruísta (incluindo seus dois aspectos: outras pessoas e circunstâncias globais).

Algumas falhas nos procedimentos de construção ou adaptação podem levar à perda da validade de conteúdo. Um artigo sobre a adaptação de uma escala de esperança, originalmente construída para norte-americanos, alerta para esse risco (Pacico, Zanon, Bastianello, Reppold, & Hutz, 2013). Os autores, ao realizarem o procedimento de adaptação, decidiram certificar-se de que os itens do instrumento adaptado eram representativos do universo de comportamentos por meio do qual o traço latente se expressava. Ao realizarem entrevistas com sujeitos que fariam parte da amostra final, descobriram que brasileiros, além de desejarem atingir objetivos já representados na escala, desejavam outros, que não estavam no instrumento original. Assim, o mesmo traço latente, quando testado em norte-americanos e brasileiros, tinha expressão diferente, apontando para a necessidade de incluir mais cinco itens na escala original.

Quando o pesquisador concluir que todos os aspectos do construto foram considerados, deve certificar-se de que o conjunto de itens que elegeu para compor o teste é efetivamente uma amostra representativa do universo de comportamentos do qual foi retirado e de que representa a expressão do traço latente. Por exemplo, é possível avaliar o quanto o sujeito deseja e quais suas expectativas com relação ao item "ter um bom relacionamento amoroso". Esse item representa a esperança autocentrada de um sujeito associada às suas relações pessoais. O item sozinho, porém, não é representativo da esperança autocentrada. Para que esse aspecto do construto seja adequadamente representado, existe a necessidade de se utilizar outros itens, para avaliar outros aspectos da vida em que ele pode demonstrar esperança, como trabalho e escola. O sujeito poderia não ter esperança de um bom relacionamento amoroso porque teve uma experiência frustrante, o que não significa que ele não tenha esperança com relação a outras coisas. Por isso, deve-se utilizar um conjunto de itens que seja representativo dos comportamentos por meio dos quais a esperança pode se manifestar.

Lawshe (1975) desenvolveu um método para avaliar a validade de conteúdo utilizando como base a concordância entre avaliadores sobre a importância de um item no teste. De acordo com Cohen, Swerdlik e Sturman (2014), o teste era submetido à avaliação de juízes, que deveriam indicar se o item era essencial ao teste, útil, mas não essencial, ou não necessário. Se o item fosse considerado essencial por mais da metade dos avaliadores, ele teria validade de conteúdo. Quanto mais o item fosse indicado como essencial, mais validade de conteúdo teria. Lawshe (1975) descreveu isso por meio da fórmula de razão de validade de conteúdo (RVC):

$$RVC = (n_e - N/2)/ (N/2)$$

n_e = número de avaliadores que indicou o item como essencial
N = número de avaliadores

Entretanto, a concordância entre os juízes poderia se dar ao acaso. Se a chance disso fosse maior que 5%, o item deveria ser eliminado do teste. Para evitar que os valores de concordância entre os juízes fossem obtidos ao acaso, Lawshe (1975) apresentou uma tabela (Tab. 5.1) com valores mínimos de RVC. Se a RVC atingisse esse valor, conforme o número de avaliadores, seria improvável que a concordância entre eles tivesse ocorrido ao acaso.

Outro tipo de validade, que tem recebido pouca atenção nos dias de hoje e que frequentemente se confunde com validade de conteúdo, é validade de face, ou validade aparente (*face validity*). É importante estar atento a esse tipo de validade, pois ela pode trazer implicações para os resultados das avaliações que fazemos.

Validade de face, ou validade aparente, refere-se ao julgamento subjetivo que as pessoas fazem sobre o teste. Quando um teste é aplicado, o respondente forma uma opinião sobre ele. Pode achar, com base em sua percepção sobre os itens ou sobre as tarefas, que se trata de um teste interessante, que mede algo importante, ou que não mede nada relevante. Essa percepção pode afetar

TABELA 5.1
Valores mínimos para RVC não serem obtidas ao acaso

Número de juízes	Valor mínimo de RVC
5	0,99
6	0,99
7	0,99
8	0,75
9	0,78
10	0,62
11	0,59
12	0,56
13	0,54
14	0,51
15	0,49
20	0,42
25	0,37
30	0,33
35	0,31
40	0,29

as respostas que o respondente dará ao completar os itens, sua motivação para responder ao instrumento e, consequentemente, pode prejudicar seu desempenho. Há vários estudos mostrando que é importante determinar a validade de face dos instrumentos e que esse tipo de validade pode interferir em outras formas de validade (p. ex., Bornstein, 1996; Nevo, 1985).

VALIDADE DE CRITÉRIO

A validade de critério está relacionada ao quanto o teste pode predizer o desempenho do sujeito em tarefas especificadas (Anastasi & Urbina, 2000). O desempenho nessa tarefa especificada torna-se o critério por meio do qual a validade do teste será avaliada. A validade é dada pela avaliação da relação dos escores obtidos no teste em questão com os escores obtidos no teste que servirá de critério (Cohen et al., 2014). O critério deve preencher alguns requisitos para que possa ser utilizado: deve ser relevante, válido e não contaminado. Por relevância, pode-se entender que deve ter alguma relação com o assunto em questão. Para um teste de inteligência, por exemplo, um critério pertinente poderia ser o desempenho escolar. Além de ser relevante, o critério utilizado deve ser válido, ou seja, se um teste X é utilizado como critério para o teste Y, deve existir evidência de validade do teste X. Por fim, o critério não deve ser contaminado. Por exemplo: suponha que está sendo desenvolvido um teste de habilidade matemática, composto por questões semelhantes às provas escolares. O critério para esse teste é o desempenho escolar na disciplina de matemática. O critério será contaminado se na prova de matemática houver questões que foram também utilizadas no teste.

A validade de critério pode ser classificada como preditiva ou concorrente. A validade preditiva ocorre quando os escores do teste são obtidos em um momento, e as medidas de critério, em um momento futuro. Geralmente se obtêm as medidas de critério após um evento interveniente, como treinamento, capacitação, terapia, uso de medicação, etc. A medida da relação entre as provas de vestibular e as médias das notas dos alunos ingressantes fornece evidência de validade preditiva para as provas de vestibular (Cohen et al., 2014). É extremamente importante que se possa obter medidas que predizem resultados. Um teste que possa predizer o desempenho do sujeito seria de interesse para seleção de pessoas. Na área clínica, testes que têm validade preditiva relacionada ao desenvolvimento de determinados transtornos podem ser utilizados para tomar medidas preventivas.

Há, ainda, a questão da validade incremental. Pesquisadores e profissionais podem estar interessados no uso de múltiplos preditores para prever um critério. Entretanto, o preditor deve acrescentar algumas vantagens ao ser

incluído nas análises. Cohen e colaboradores (2014) ressaltaram a importância de que cada preditor utilizado tenha validade preditiva com relação ao critério em questão. Os preditores adicionais devem ter validade incremental com relação a ele, isto é, eles devem explicar algo sobre a medida de critério que os outros preditores não explicam. Ou seja, validade incremental é o grau em que o preditor adicional pode explicar algo sobre o critério que outros preditores não explicam. Por exemplo, sabe-se que inteligência é um importante preditor de desempenho acadêmico. Rand (2009) demostrou que esperança é um dos mais importantes preditores de desempenho, mesmo quando inteligência é considerada. Assim, esperança pode ser utilizada como preditor da medida de critério, já que pode fornecer informações que inteligência não forneceria. É sempre necessário cuidado quando mais de um preditor é utilizado. Quando se organiza uma avaliação psicológica, deve-se usar os instrumentos necessários, mas não mais do que o estritamente necessário. O uso excessivo de testes gera cansaço no testando e pode levá-lo a não responder todos os itens ou a começar a responder de forma aleatória a partir de certo ponto. Manuais de testes podem indicar se um instrumento é preditor de alguma medida de critério, mas não se ele agregaria (ou quanto agregaria) se for usado em conjunto com outros instrumentos. Essa resposta está na literatura da área, que deve sempre ser consultada.

A validade concorrente ocorre quando as duas medidas, o teste e o critério, são obtidas quase simultaneamente (uma logo após a outra). Um exemplo seria a aplicação de dois testes na mesma sessão. Por exemplo, um pesquisador quer desenvolver um teste para avaliar personalidade no Brasil. Uma vez realizados todos os procedimentos para a construção de um novo teste, para a obtenção de evidência de validade concorrente, esse pesquisador poderia aplicar seu teste e também um teste já existente e reconhecidamente válido para uso com a população brasileira, como, por exemplo, a Bateria Fatorial de Personalidade (BFP) (Nunes, Hutz, & Nunes, 2010).

O coeficiente de validade é um tipo de evidência estatística utilizado para inferir a validade de critério, seja preditiva, seja concorrente. O coeficiente de validade é calculado pela correlação entre a medida e o critério. Frequentemente, a correlação de Pearson é utilizada. Outros coeficientes de correlação podem ser utilizados, dependendo das características dos dados coletados.

Não há um parâmetro estabelecido para o tamanho que o coeficiente de validade deva assumir para que o pesquisador o adote como válido. Entretanto, os pesquisadores (Cohen et al., 2014; Cronbach & Glesser, 1965) afirmaram que ele deve ser grande o suficiente para permitir que as decisões corretas sejam tomadas pelo pesquisador no contexto em que o teste for utilizado. Assim, é possível entender que o coeficiente de validade deve ser grande o suficiente

para permitir que se identifique e se diferencie o sujeito com relação à variável critério. Outras evidências estatísticas, como os dados de expectativa, também podem ser utilizadas para inferir se há ou não validade de critério.

VALIDADE DE CONSTRUTO

A validade de construto foi reportada pela primeira vez em 1954, em um relatório técnico publicado pela American Psychological Association. No ano seguinte, foi publicado o artigo *Construct Validity in Psychological Tests* (Cronbach & Meehl, 1955). Nele, os autores chamaram atenção para um novo método de pesquisa. Em geral, os testes eram construídos com base em teorias, o que permitia que se examinasse sua validade por meio de técnicas de validação de conteúdo (pois o universo de comportamentos por meio do qual o traço latente poderia se expressar já estava bem definido). Também se utilizava muito a validade de critério, pois a teoria já referenciava como se dariam as relações com outras variáveis. Entretanto, havia construtos cujo universo de comportamentos por meio dos quais se expressavam não estava totalmente definido, levando a falhas no processo de validação por conteúdo. A relação com variáveis que poderiam servir de critério também não estava clara, o que poderia levar a falhas no procedimento de validação de critério. Quando o construto se encontrava nessa situação, Cronbach e Meels (1955) identificaram que ele ainda não havia sido "operacionalmente definido" e que a rede nomológica (as relações entre os construtos e as variáveis observáveis deles decorrentes) ainda estaria em construção. Por isso, outras técnicas de validação são necessárias para verificar sua validade. Assim, os autores sugeriram que, em lugar de partir da teoria para a obtenção do teste, pode-se partir do teste para a teoria e, assim, "clarear" a rede nomológica na qual o construto está inserido. O pesquisador elabora hipóteses teóricas acerca do construto e busca outra forma de verificar a validade (validade de construto), já que, nesse caso, as técnicas de validação de conteúdo e de critério não são suficientes para determinar a validade do construto envolvido. Mediante a validação de construto, ele verifica se as hipóteses formuladas a respeito do construto são verdadeiras ou não.

A validade de construto de um teste é a extensão em que se pode dizer que ele mede um construto teórico ou um traço (Anastasi & Urbina, 2000), como personalidade, ansiedade, autoeficácia, etc. A validação de construto ocorre pela acumulação gradual de informações que provêm de diversas fontes (Anastasi & Urbina, 2000; Cohen et al., 2014). Alguns procedimentos podem ser utilizados para conferir evidências de validade de construto, como:

a) Mensurar a homogeneidade do teste: os itens que compõem o teste são homogêneos, isto é, referem-se todos à mesma variável?
b) Examinar se os escores no teste variam conforme o esperado. Os escores no teste variam conforme previsto nas hipóteses? Por exemplo, variam com idade, com uma manipulação experimental (os escores diferem do pré para o pós-teste), grupos (os escores variam do grupo-controle para os grupos experimentais), conforme previsto nas hipóteses?
c) Mensurar a correlação do construto com outras variáveis. As relações com outros construtos ocorrem conforme previsto? Ou seja, as evidências de validade convergente e discriminante confirmam as hipóteses acerca do construto?

De acordo com esses procedimentos, várias técnicas de validação podem ser utilizadas. Entre elas, duas são frequentemente relatadas em artigos científicos: análise fatorial e análise da consistência interna.

A análise fatorial permite identificar fatores ou variáveis específicas (que são os atributos ou as dimensões nas quais os escores variam de um sujeito para o outro). Mediante a análise das intercorrelações dos dados comportamentais, é possível reduzir as categorias que descrevem o comportamento a um pequeno número de fatores. Dizendo de outra forma, a análise fatorial pode ser empregada como um método que reduz os dados provenientes de um conjunto de escores a um número menor de fatores, empregando a inter-relação entre os itens para atingir esse objetivo. Assim, um instrumento como a Bateria Fatorial de Personalidade, que utiliza 126 itens para descrever o comportamento, ao ser submetido à análise fatorial para mensurar as inter-relações entre eles, gera cinco grandes fatores que explicam os traços que o instrumento se propõe a medir. O fator é formado por um conjunto de itens que apresenta correlação entre si. Os itens desse fator podem apresentar correlações com os outros fatores, mas elas são mais baixas, e isso é suficiente para permitir que fiquem em fatores diferentes. É como se todos os dados comportamentais (os escores em cada item) do instrumento fossem como estrelas no céu. Aqueles que estão mais correlacionados apresentam-se como estrelas mais próximas, formando constelações (que poderiam ser comparadas aos fatores). As estrelas de uma constelação podem estar próximas de outras estrelas quaisquer, mas não o suficiente para caracterizar uma constelação (os itens podem estar correlacionados com outros, mas essa correlação não é alta o suficiente para caracterizar um fator). Assim, após os escores dos itens serem submetidos à análise fatorial, é possível identificar os fatores responsáveis pela expressão comportamental. Quando os fatores identificados correspondem àqueles descritos pela teoria ou hipóteses teóricas, pode-se dizer que existe evidência de validade de construto.

A análise fatorial pode ser exploratória (AFE) ou confirmatória (AFC). A primeira permite a extração de fatores e foi discutida no parágrafo anterior. A AFC testa o quanto os dados reais se ajustam a um modelo hipotético criado para descrever os dados. O pesquisador constrói um modelo teórico utilizando a variável em questão e outras que sejam relevantes para explicá-la. A análise apresenta um conjunto de índices de ajuste que informam o quanto o modelo reflete o conjunto de dados observados.

Outros procedimentos também contribuem para a obtenção da validade de construto. A validade convergente, a discriminante e a relação dos escores com instrumentos semelhantes são um exemplo deles. Quando se realiza de modo cuidadoso a validação de construto, é importante saber se o instrumento construído se relaciona com outras variáveis conforme foi previsto. Assim, se um instrumento foi construído para medir esperança, e ele se correlaciona com os escores obtidos em um teste de otimismo, esse resultado fornece evidência de validade convergente. Isso ocorre porque essa correlação é esperada teoricamente, e há estudos mostrando que ela efetivamente acontece. Da mesma maneira, é esperada uma correlação negativa de esperança com depressão. Se isso for verificado, também confere evidência de validade convergente ao instrumento.

Entretanto, não basta que o pesquisador saiba que os escores obtidos a partir do instrumento se correlacionam da maneira esperada com os escores de outro. É preciso que ele verifique se os escores não se correlacionam com os escores de testes com os quais não devem se relacionar. Por exemplo, não é previsto que esperança se correlacione com velocidade de processamento. Assim, ao verificar que os escores desses dois testes não apresentam correlação, o pesquisador terá encontrado evidência de validade discriminante.

A visão de validade de conteúdo, critério e construto apresentou alguns problemas. Talvez o principal deles seja o que associava a validade ao teste. Com a visão unificada de validade proposta por Messick (1989), os pesquisadores passaram a perceber a validade como um conceito único e integrado que se refere às ações decorrentes do uso dos testes (interpretações, inferências, conclusões, etc.). Talvez o mais importante disso é que Messick trouxe a noção de que a validação é algo em constante construção, sempre em busca de evidências que permitam que as conclusões acerca dos escores dos testes sejam progressivamente mais válidas. Isso levou à criação e à proliferação de diversos tipos de validade. Entretanto, provavelmente Messick queria evitar esse resultado, já que ele concebia a validade como única, sendo possível ao processo de validação contar com diferentes fontes de evidências de validade que seriam acumuladas e sintetizadas para conferir validade de construto e dar suporte às interpretações, inferências e conclusões feitas sobre os escores obtidos pelo uso

apropriado dos testes. Dessa forma, o pesquisador deve preocupar-se em coletar evidências de validade de diferentes fontes, sejam elas de conteúdo, sejam elas de critério, convergente ou divergente, todas em busca de validar o construto.

QUESTÕES

1. Qual a diferença entre validade e validação?
2. Ao que se refere a validade de construto?
3. Como pode ser obtida?
4. Como se avaliam a validade convergente e a discriminante?
5. Que cuidados deve-se ter durante a validação de conteúdo?
6. Qual a diferença entre validade de face e validade de conteúdo?
7. Como a visão de validade única proposta por Messick transcende a visão composta por três categorias?

REFERÊNCIAS

American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME). (1985). *Standards for education and psychological testing*. Washington: AERA, APA, NCME.

American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME). (1999). *Standards for educational and psychological testing*. Washington: AERA, APA, NCME.

American Psychological Association (APA). (1954). Technical recommendations for psychological tests and diagnostic techniques. *Psychological Bulletin, 51*(2 pt. 2), 1-38.

Anastasi, A., & Urbina, S. (2000). *Testagem psicológica*. Porto Alegre: Artmed.

Bornstein, R. F. (1996). Face validity in psychological assessment: Implications for a unified model of validity. *American Psychologist, 51*(9), 983-984.

Buckingham, B. R. (1921). Intelligence and its measurement: A symposium. *Journal of Educational Psychology, 12*(3), 123-147.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by multitrait-multimethod matrix. *Psychological Bulletin, 56*(2), 81-105.

Cohen, R. J., Swerdlick, M. E., & Sturman, E. D. (2014). *Testagem e avaliação psicológica: Introdução a testes e medidas* (8. ed.). Porto Alegre: AMGH.

Cronbach, L. J. (1971). Test validation. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 443-507). Washington: American Council on Education.

Cronbach, L. J., & Glesser, G. C. (1965). *Psychological tests and personnel decisions* (2nd ed.). Urbana: University of Illinois.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*(4), 281-302.

Lawshe, C. H. (1975). A quantitative approach to content validity. *Personnel Psychology, 28*(4), 563-575.

Loevinger, J. (1957). Objective tests as instruments of psychological theory. *Psychological Reports, 3*, 635-694.

Markus, K. A., & Borsboom, D. (2013). *Frontiers of test validity theory: Measurement, causation, and meaning* (Multivariate Applications Series). New York: Routledge.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13-103). New York: American Council on Education, MacMillan.

Nevo, B. (1985). Face validity revisited. Journal of Educational Measurement, 22(4), 287-293.

Newton, P. E., & Shaw, S. D. (2014). *Validity in educational & psychological assessment*. London: Sage.

Nunes, C. H. S. S., Hutz, C. S., & Nunes, M. F. O. (2009). *Bateria Fatorial de Personalidade (BFP: Manual técnico)*. São Paulo: Casa do Psicólogo.

Pacico, J. C., Zanon, C., Bastianello, M. R., Reppold, C. T., & Hutz, C. S. (2013). Adaptation and validation of the Brazilian version of the Hope Index. *International Journal of Testing, 13*(3), 193-200.

Rand, K. L. (2009). Hope and optimism: Latent structures and influences on grade expectancy and academic performance. *Journal of Personality, 77*(1), 231-260.

Staats, S. (1989). Hope: A comparison of two self-report measures for adults. *Journal of Personality Assessment, 53*(2), 366-375.

Urbina, S. (2007). *Fundamentos da testagem psicológica*. Porto Alegre: Artmed.

Zumbo, B. D., & Chan, E. K. H. (2014). *Validity and validation in social, behavioral, and health sciences* (Social Indicators Research Series). New York: Springer.

LEITURAS SUGERIDAS

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*(3), 297-334.

Hair, J. F., Jr, Black, W. C., Babin, B. J., Anderson, R. E., & Tathan, R. L. (2009). *Análise multivariada de dados* (6. ed.). Porto Alegre: Bookman.