VALIDADE DOS TESTES

TEST VALIDITY

VALIDEZ DE LOS TESTES

Luiz Pasquali

RESUMO

A validade ocupa uma posição central na teoria da medida, constituindo-se um parâmetro fundamental e indispensável. Atualmente, é definida como a medida em que as evidências empíricas embasam as interpretações e os usos propostos para o teste. Neste estudo, o objetivo principal é apresentar o conceito de validade, a história desse parâmetro e as principais formas de medi-lo. A primeira parte do artigo explora as bases conceituais da história do parâmetro em três períodos. Em cada um deles, há a predominância de um dos tipos atualmente conhecidos de validade. Em seguida, são detalhados os procedimentos qualitativos e quantitativos recomendados para investigar validade na visão atual. Por fim, é apresentado o conceito de validade ecológica, que não constitui uma nova forma de coletar evidências de validade, mas sim de identificar como tais evidências devem ser buscadas.

Palavras-chave: validade; medidas educacionais; psicometria.

ABSTRACT

Validity occupies a key position in measurement theory, constituting a fundamental and indispensable parameter. Currently, it is defined as the degree to which empirical evidence supports interpretation and proposed test uses. The primary objective of this study is to present the concept of validity, the history behind this parameter, and the main ways of measuring it. The first part of the article explores the conceptual roots of the history of the validity parameter in three periods; in each one of them, there is the predominance of one of the recognized types of validity over the others. Subsequently, we detail the recommended qualitative and quantitative procedures to investigate validity under the current paradigm. Finally, we present the concept of ecological validity, which should not be understood as a new form of collecting evidence validity, but rather as a form of identifying how such evidence is to be sought.

Keywords: validity; educational measures; psychometrics.

RESUMEN

La validez ocupa una posición central en la teoría de la medida, constituyéndose como un parámetro fundamental e indispensable. Actualmente, es definida como la medida en que las evidencias empíricas embazan las interpretaciones y los usos propuestos para el test. En este estudio, el objetivo principal es presentar el concepto de validez, la historia de ese parámetro y las principales formas de medirlo. La primera parte del artículo explora las bases conceptuales de la historia del parámetro de validez en tres períodos. En cada uno de ellos, la predominancia de uno de los tipos actualmente conocidos de validez. En seguida, son detallados los procedimientos cualitativos y cuantitativos que son recomendados para investigar la validez en la visión actual. Por fin, es presentado el concepto de validez ecológica, que no se constituye como una nueva forma de recoger evidencias de validez, pero sí identificar como tales evidencias deben ser buscadas.

Palabras clave: validez; medidas educacionales; psicometría.

Introdução

A validade constitui um parâmetro da medida tipicamente discutido no contexto das ciências psicossociais, que trabalham com a modelagem latente. Ela não é corrente em ciências físicas, por exemplo, embora haja nessas ciências ocasiões em que tal parâmetro se aplicaria. Nestas, a preocupação principal na medida se centra na questão da precisão, na dita calibração dos instrumentos. Essa é importante também na medida em ciências psicossociais, mas ela não tem nada a ver, conceitualmente, com a questão da validade. A razão disso está no fato de que a validade diz respeito ao aspecto da medida de ser congruente com a propriedade medida dos objetos e não com a exatidão com que a mensuração, que descreve essa propriedade do objeto, é feita. Em Física, o instrumento é um objeto físico que mede propriedades físicas; então parece fácil ver que a propriedade do objeto mensurante é ou não congruente com a propriedade do objeto medido. Tome, por exemplo, o caso da propriedade "comprimento" do objeto. O instrumento que mede essa propriedade, isto é, o metro, usa a sua propriedade de comprimento para medir o comprimento de outro objeto; então, mensura-se comprimento

com comprimento, tomados estes termos univocamente. Não há necessidade de provar que a propriedade "comprimento" do metro seja congruente com a mesma propriedade no objeto medido; os termos são unívocos, eles são conceitualmente equivalentes, aliás, idênticos.

O caso já se torna menos claro quando, por exemplo, o astrônomo mede a propriedade "velocidade" galáctica de aproximação ou afastamento via efeito Doppler, no qual a aproximação/afastamento das linhas espectrais da luz da galáxia seria o instrumento da medida. Aqui já temos, na verdade, um problema de validade do instrumento de medida, a saber, é verdade ou não que as distâncias das linhas espectrais têm a ver com a velocidade das galáxias? Pode-se fazer tal suposição, mas ela tem que ser demonstrada empiricamente de alguma maneira, isto é, pelo menos em suas consequências, em hipóteses dela derivadas ou deriváveis e verificáveis. Nesse caso específico, o problema da precisão da medida diz respeito a quão exata pode ser feita a mensuração das distâncias entre as linhas espectrais, ao passo que o problema da validade diz respeito ao fato de essa medida, por mais exata e perfeita que ela possa ser, ter algo a ver ou não com a velocidade de afastamento da galáxia. Em outras palavras, a validade em tal caso diz respeito à demonstração da legitimidade da representação ou da modelagem da velocidade galáctica via distâncias das linhas espectrais.

Esse caso da astronomia ilustra o que tipicamente acontece com a medida em ciências psicossociais e, consequentemente, torna a prova da validade dos instrumentos nessas ciências algo fundamental e crucial, isto é, é uma condição *sine qua non* demonstrar a validade dos instrumentos nessas ciências. Isso é particularmente o caso nos enfoques que, em Psicologia, trabalham com o conceito de traço latente, pelos quais se deve demonstrar a correspondência (congruência) entre traço latente e sua representação física (o comportamento). Não causa estranheza, portanto, que o problema de validade tenha tido, na história da Psicologia, uma posição central na teoria da medida, constituindo-se, na verdade, o seu parâmetro fundamental e indispensável. Aliás, a história desse parâmetro é repleta de diatribes que espelham concepções teóricas antagônicas da própria teoria psicológica. À questão de "como legitimar ou justificar a pertinência da medida do comportamento humano?" foram dadas respostas diferentes na história da Psicometria. Essa diatribe pode ser ilustrada distinguindo várias etapas

de predominância de uma concepção do parâmetro validade sobre outras e que aparecem sempre atreladas a uma concepção mais geral da própria Psicologia, como já anotava Anastasi em 1986.

Desenvolvimento

Com efeito, poder-se-ia delinear, em traços bem gerais, a história do parâmetro da validade em três períodos. Em cada um deles, há a predominância de um dos tipos atualmente conhecidos de validade, desde o famoso trabalho de Cronbach e Meehl (1955), expressos sob o modelo trinitário, a saber, validade de conteúdo, de critério e de construto.

Predomínio da validade de conteúdo – 1º período (1900-1950)

Nessa época, estavam em voga as teorias da personalidade e com elas predominava o interesse pelos traços de personalidade (tipos, temperamentos, traços, aptidões etc.). Essas teorias (Psicanálise, Fenomenologia, Gestaltismo etc.) apresentavam em geral pouca fundamentação empírica, assumindo um caráter bastante nebuloso, quando não fantasioso. Nessa atmosfera, os testes dos traços eram considerados válidos na medida em que seu conteúdo correspondesse ao conteúdo dos traços teoricamente definidos pela teoria psicológica em questão.

Fora alguns poucos (teste de Binet-Simon, de Raven, de Thurstone e alguns testes projetivos ainda em voga), as dezenas de testes criados nessa época já fazem parte de uma história passada e podem ser ditos representantes da pré-história dos testes psicológicos.

Predomínio da validade de critério – 2º Período (1950-1970)

Prevalecia em Psicologia o enfoque do Behaviorismo Skinneriano, que influenciou também a Psicometria. Os testes eram concebidos como uma amostra de comportamentos e tinham como função predizer outros comportamentos ou comportamentos futuros. Um teste era, consequentemente, válido se predizia com precisão os comportamentos em uma futura ou em outra condição. Esse se tornava, assim, o critério de validade do teste. Não interessava saber por que o teste predizia algo, bastava mostrar que de fato ele o fazia e isso era o critério de sua validade. Esse modo de conceber os testes ainda persiste hoje em dia, mas parece que aos poucos sua

relevância vem se tornando secundária, tornando-se tão somente uma etapa, juntamente com a validade de conteúdo, no processo de elaboração dos testes psicológicos (ANASTASI, 1986).

Esse período se caracteriza por uma acentuada fuga do pensar teórico que definia a época anterior. O teste não era mais construído para representar traços de personalidade, e os itens (tarefas) eram selecionados a partir de um grande elenco (*pool of items*) que parecia se referir àquilo para o qual se queria uma medida, fazendo uso praticamente exclusivo e *a posteriori* de análises estatísticas, especialmente da correlação. Não era mais a teoria psicológica e sim a estatística que definia a qualidade do teste. Esse processo de empirismo cego se assemelha ao pescador que lança a rede não importa onde para ver o que pode colher e, em cima do colhido, decidir o que quer. No processo, tipicamente se perdem "toneladas" de itens puramente por não satisfazerem critérios estatísticos (KURTZ, 1948; CURETON, 1950; PRIMOFF, 1952). A atitude dos psicometristas dessa época é explicada por razões históricas, eles queriam se desfazer do que lhes parecia um teorizar gratuito e fantasioso do início do século XX em Psicologia. Contudo, já na década de 1970, os psicometristas procuravam voltar a um teorizar psicológico mais relevante e em cima dele elaborar seus testes, o que deu início ao terceiro período na concepção dos testes e de sua validade.

Predomínio da validade de construto – 3º Período (1970)

Esse período teve suas fontes históricas no artigo de Cronbach e Meehl (1955) sobre o modelo trinitário da validade (conteúdo, critério, construto). Eles próprios já diziam que a validade de construto exigia um novo tipo de teorizar em Psicometria. Entretanto, o impacto prático dessa visão dos autores só se faria sentir após os anos 1970. Na verdade, a volta à teoria psicológica em Psicometria se deve a vários fatores; entre eles, salientam-se os seguintes.

1) Preocupação com o desenvolvimento da teoria da personalidade e, em especial, da inteligência, com maior base empírica e valendo-se sobretudo das técnicas da análise fatorial (COMREY, 1970; GUILFORD, 1967; JACKSON, 1974; MILLON, 1983; CATTELL, 1965; CATTELL; STICE, 1957; CATTELL; WARBURTON, 1967).

2) Realização de estudos dos processos cognitivos (STERNBERG, 1977, 1984; STERNBERG; DETTERMAN, 1986; STERNBERG; RIFKIN, 1979).

3) Realização de estudos do processamento da informação (NEWELL; SHAW; SIMON, 1958; NEWELL; SIMON, 1958).

4) Insatisfação com os resultados do uso de testes na educação e no trabalho. Na clínica, ainda se utilizavam bastante os testes projetivos, em que predominava, aliás, o pensamento da primeira época dos testes, os quais se baseavam nas teorias dos traços de personalidade.

5) O impacto da Teoria de Resposta ao Item (TRI) por sua insistência no traço latente. A influência decisiva dessa teoria ocorre somente após os anos 1980, devido ao atraso na área da informática para fazer uso prático das análises estatísticas complexas que tal enfoque exige.

Na validação dos instrumentos psicológicos, a preocupação agora se concentra na validade de construto ou de traços latentes. Não está ainda finalizada a disputa entre a ênfase nos traços ou a ênfase nas situações (*construct-centered versus task-centered*) ou, como diz Messick (1994), entre a avaliação *task-driven versus construct-driven*. Parece, entretanto, que o conceito de validade dos testes psicológicos irá finalmente se reduzir à validade de construto, e o conteúdo e o critério serão apenas aspectos desta (ANASTASI, 1986; MESSICK, 1989, 1994; EMBRETSON, 1983; WIGGINS, 1989; CRONBACH, 1989; o qual, já em 1955, de algum modo, previa tal desenlace). Essa tendência é obviamente favorecida também pelos psicólogos da linha cognitivista (STERNBERG, 1985, 1990; GARDNER, 1983).

Retoma-se, agora, a validade dos testes. Nos manuais de Psicometria, costuma-se dizer que um teste é válido quando de fato mede o que supostamente deve medir. Embora essa definição pareça uma tautologia, na verdade ela não é, uma vez considerada a teoria psicométrica sobre o traço latente, exposta neste trabalho. O que se quer dizer com a definição é que, ao se medirem os comportamentos (itens), que são a representação do traço latente, está-se medindo o próprio traço latente. Tal suposição é justificada se a representação comportamental for legítima. Essa legitimação somente é possível se existir uma teoria prévia do traço que fundamente que a tal representação comportamental constitui uma hipótese dedutível da teoria. A validade do teste (este constituindo a hipótese) será, então,

estabelecida pela testagem empírica da verificação da hipótese. Pelo menos, esta é a metodologia científica. Assim, fica muito estranha a prática corrente na Psicometria de se agrupar intuitivamente uma série de itens e, *a posteriori*, verificar estatisticamente o que eles estão medindo. A ênfase na formulação da teoria sobre os traços foi muito fraca no passado. Com a influência da Psicologia Cognitiva, essa ênfase felizmente está voltando ou deverá voltar ao seu devido lugar na Psicometria.

Aliás, a Psicometria Clássica entende "aquilo que supostamente deve medir" como sendo o "critério", este representado por teste paralelo. Assim, "aquilo que" é o traço latente na concepção cognitivista da Psicometria e é o critério (escore no teste paralelo) na visão comportamentalista.

Diz Anastasi (1986, p. 3) que o processo de validação de um teste "inicia com a formulação de definições detalhadas do traço ou construto, derivadas da teoria psicológica, pesquisa anterior, ou observação sistemática e análises do domínio relevante do comportamento. Os itens do teste são então preparados para se adequarem às definições do construto. Análises empíricas dos itens seguem, selecionando-se finalmente os itens mais eficazes (i.e., válidos) da amostra inicial de itens".[1]

A validação da representação comportamental do traço, isto é, do teste, embora constitua o ponto nevrálgico da Psicometria, apresenta dificuldades importantes em três níveis ou momentos do processo de elaboração do instrumento; a saber, níveis da teoria, da coleta empírica da informação e da própria análise estatística da informação.

No nível da teoria, se concentram talvez as maiores dificuldades. Na verdade, a teoria psicológica se encontra ainda em estado embrionário, destituída quase que totalmente de qualquer nível de axiomatização, o que resulta em uma pletora de teorias, muitas vezes até contraditórias. Basta lembrar de teorias como Behaviorismo, Psicanálise, Psicologia Existencialista, Psicologia Dialética e outras que existiram simultaneamente e postularam princípios irredutíveis entre as várias teorias e pouco concatenados dentro de uma mesma teoria ou, então, em número insuficiente para se poder deduzir hipóteses úteis para o conhecimento psicológico. Com essa confusão no

1 A questão da elaboração de testes psicológicos é detalhadamente tratada em Pasquali, 2010.

campo teórico dos construtos, torna-se extremamente difícil para o psicometrista operacionalizá-los, isto é, formular hipóteses claras e precisas para testar ou, então, formular hipóteses psicologicamente úteis. Ainda quando a operacionalização for um sucesso, a coleta de informação empírica não será isenta de dificuldades, como, por exemplo, a definição inequívoca de critérios com os quais estes construtos possam ser idealmente estudados. Mesmo em nível das análises estatísticas, são encontrados problemas. Pela lógica da elaboração do instrumento, a verificação da hipótese da legitimidade da representação dos construtos se faz por análise do tipo fatorial (confirmatória), por meio da qual se procura identificar, nos dados empíricos, os construtos previamente operacionalizados no instrumento. Mas acontece que a análise fatorial faz algumas postulações fortes que nem sempre se coadunam com a realidade dos fatos. Por exemplo, essa análise assume que as respostas dos sujeitos aos itens do instrumento são determinadas por uma relação linear destes com os traços latentes. Há, ainda, o grave problema da rotação dos eixos, a qual permite a demonstração de um número sem fim de fatores para o mesmo instrumento.

Entretanto, infelizmente, a história da validade dos testes psicológicos é ainda uma área pelo menos obscura. Duas definições já demostram isso.

1) Validade pode ser mais bem definida como a extensão para a qual certas inferências podem ser feitas com base em escores de um teste ou em outras medidas (MEHRENS; LEHMANN, 1984).

2) Validade consiste na extensão para a qual um teste é verídico, preciso ou relevante ao medir um traço que pretende medir.

A primeira definição se coaduna com o pensar de Cronbach e Mehel (1955), da rede nomológica, depois sistematizada por Samuel Messick (1989) sob validade de construto. A segunda definição melhor se coadunaria com uma visão dualista do ser humano, na qual traço significa processos mentais. De qualquer forma, as duas definições se concentram no que veio a ser denominado de construto. Agora, o que é um construto?

Na visão de Messick e da grande maioria dos psicometristas atuais, o construto é uma ficção; ou melhor, são racionalizações, na expressão de Messick, pois para ele a validade de um teste consiste num julgamento

integrado e avaliativo do grau em que evidências empíricas suportam a adequação e a propriedade de inferências e ações fundamentadas nos escores de testes ou outras formas de avaliação. Na visão da Psicologia Cognitiva e das neurociências, construto é uma realidade mental.

As duas visões apresentam problemas gigantescos. Primeiramente, construto como ficção implica irracionalidade epistemológica: pois o construto é a causa dos comportamentos; mas estes são reais, então como é possível se conceber uma ficção (construto) que possa produzir uma realidade? Por outro lado, construto como realidade mental implica o conhecimento desta (sua estrutura e funcionamento); mas a Psicologia não conhece nem a estrutura nem o funcionamento de tais processos mentais. Essa concepção de construto salva a racionalidade epistemológica, mas nos coloca no reino do incógnito. Contudo, ela permite e justifica a procura e a pesquisa desses processos, tarefa que as neurociências vêm tentando realizar. Fica a sensação de que as duas versões nos deixam, no presente, "em um mato sem cachorro".

Diante de tamanhas dificuldades, os psicometristas recorrem a uma série de técnicas para viabilizar a demonstração da validade dos seus instrumentos. Fundamentalmente, essas técnicas foram as seguintes.

1) Na visão clássica – Técnicas que podem ser reduzidas a três grandes classes (modelo trinitário): as que visam a validade de construto, as que visam a validade de conteúdo e as que visam a validade de critério (APA, 1954).

2) Na visão atual – Técnicas que objetivam procurar evidências de validade com base em:

   - conteúdo;
   - processos de resposta;
   - estrutura interna;
   - relação com outras variáveis;
   - consequências da testagem (AERA; APA; NCME, 1999).

A visão atual é a predominante em Psicologia, embora ela tenha finalmente perdido totalmente o conceito de construto (COLLIVER; CONLEE; VERHULST,

2012), focalizando a validação dos testes psicológicos por meio do acúmulo de provas circunstanciais (ditas evidências de validade) para legitimar as decisões tomadas com base nos escores (às vezes chamada de validade consequencial porque foca nas consequências que se tiram a partir dos escores dos testes e não mais no construto). Mesmo assim, essas associações de Psicologia deixaram de fora um tipo de validade que sobretudo as neurociências vêm insistindo ser importante, a validade ecológica. A seguir será apresentado um pouco de tudo isso, tomando a opinião de Aera, Apa e NCME (1999) simplesmente como organograma.

Validade com base no conteúdo

Refere-se ao conceito tradicional de validade de conteúdo. Um teste tem validade de conteúdo se ele constitui uma amostra representativa de um universo finito de comportamentos (domínio). É aplicável quando se pode delimitar *a priori* e com clareza um universo de comportamentos, como é o caso dos testes de desempenho, que pretendem cobrir um conteúdo delimitado por um curso programático específico.

Para viabilizar um teste com validade de conteúdo, é preciso que se façam as especificações dele antes da construção dos itens. Essas especificações comportam a definição de três grandes temas: *i*) definição do conteúdo; *ii*) explicitação dos processos psicológicos (dos objetivos) a serem avaliados; e *iii*) determinação da proporção relativa de representação no teste de cada tópico do conteúdo.

Quanto ao conteúdo, trata-se de detalhá-lo em tópicos (unidades) e subtópicos e de explicitar a importância relativa de cada tópico dentro do teste. Tais procedimentos evitam as indevidas super-representação de alguns tópicos e sub-representação de outros por vieses e pendores pessoais do avaliador. Claro que será sempre o avaliador ou a equipe de avaliadores que vai definir esse conteúdo e a relativa importância de suas partes, mas essa definição deve ser estabelecida antes da construção dos itens, a fim de garantir certa objetividade pelo menos nas decisões.

Quanto aos objetivos, um teste não deve ser elaborado para avaliar exclusivamente um processo. Como na aprendizagem entram em ação vários

processos psicológicos, há interesse em todos, ou naqueles que se quer que sejam avaliados por um teste de conteúdo. Por exemplo, o teste deverá conter itens que avaliam a memória (reproduzir), a compreensão (conceituar, definir), a capacidade de comparação (relacionar) e a capacidade de aplicação dos princípios aprendidos (solução de problemas, transferência da aprendizagem).

A validade de conteúdo de um teste é praticamente garantida pela técnica de construção deste. Assim, é importante esboçar essa técnica. Ela comporta os seguintes passos.

1) Definição do domínio cognitivo
   Definir os objetivos ou os processos psicológicos que se quer avaliar. Para essa tarefa, é útil se inspirar em alguma taxonomia clássica de objetivos educacionais, como, por exemplo, a taxonomia de Bloom (1956). Com base em uma taxonomia, definem-se os objetivos gerais e específicos que se deseja medir no teste, como:

   - conhecer tais tópicos;
   - compreender tais tópicos;
   - aplicar tais tópicos;
   - analisar tais tópicos.

2) Definição do universo de conteúdo
   Como o teste constitui uma amostra representativa do conteúdo, é preciso definir e delimitar o universo do conteúdo programático em divisões e subdivisões (tópicos e subtópicos) ou quantas outras subclassificações forem necessárias. Isso implica delimitar o conteúdo em suas unidades e subunidades de ensino.

3) Definição da representatividade de conteúdo
   Definir a proporção com que cada tópico e subtópico deve ser representado no teste, decidindo, assim, a importância com que cada um deles aparece no conteúdo total do universo.

4) Elaboração da tabela de especificação
   Nela são relacionados os conteúdos com os processos cognitivos a serem avaliados, bem como a importância relativa a ser dada a cada unidade.

5) Construção do teste
Elaborar os itens que irão representar o teste seguindo as técnicas de construção de itens (MAGER, 1981; PASQUALI, 2010).

6) Análise teórica dos itens
Essa análise visa verificar a compreensão das tarefas propostas no teste por parte dos testandos (análise semântica) e a avaliação da pertinência do item à unidade correspondente, bem como o processo cognitivo envolvido (análise de juízes).

7) Análise empírica dos itens
Após a aplicação do teste, os dados obtidos podem ser utilizados para validação empírica deste, para seu uso futuro. Essa análise implica basicamente na determinação dos níveis de dificuldade e de discriminação dos itens. A técnica da teoria da resposta ao item (TRI) pode ser de grande valia nessa etapa.

Para facilitar a especificação do teste, pode-se utilizar uma tabela de dupla entrada, com o detalhamento dos objetivos (processos) à esquerda, o detalhamento dos tópicos no topo, e, no corpo da tabela, o número de itens.

Validade com base nos processos de resposta

> Alguns estudos mais recentes fazem uma análise teórica-empírica das relações entre os processos mentais ligados ao construto em causa e as respostas aos itens do instrumento. A partir de propostas explicativas dos processos mentais subjacentes às respostas aos itens, formulam-se modelos explicativos sobre como a pessoa processa as informações dos itens do teste. A partir disso tenta-se prever aspectos da resposta como acertos e tempo de reação a diferentes itens em razão das suas características e demandas consequentes aos processos cognitivos ou emocionais. Assim busca-se analisar a coerência entre as explicações teóricas e os dados empíricos (NUNES; PRIMI, 2010, p. 122).

Contudo, é difícil ver nisso demonstração de validade do teste; trata-se de uma relevante curiosidade de estudo da Psicologia Cognitiva, mas não de uma prova de validade.

Validade com base na estrutura interna

Entre as cinco fontes de validade dos testes nos padrões de Aera, Apa e NCME (1999), este tipo de validade seria o único que poderia salvar o conceito de construto. Numa visão cognitivista de construto, a validade de construto ou de conceito é considerada a forma mais fundamental de validade dos instrumentos psicológicos e com toda a razão, dado que ela constitui a maneira direta de verificar a hipótese da legitimidade da representação comportamental dos traços latentes e, portanto, se coaduna exatamente com a teoria psicométrica aqui defendida. Historicamente, o termo construto entrou na Psicometria por meio da American Psychological Association Committee on Psychological Tests, que trabalhou entre 1950 e 1954 e cujos resultados se tornaram as recomendações técnicas para os testes psicológicos (APA, 1954).

O conceito de validade de construto foi elaborado com o clássico artigo de Cronbach e Meehl (1955) *Construct validity in psychological tests*, embora o conceito já tivesse uma história sob outros nomes, tais como validade intrínseca, validade fatorial e até validade aparente (*face validity*). Essas várias terminologias demonstram a confusa noção que construto possuía. Embora tenham tentado clarear o conceito de validade de construto, Cronbach e Meehl ainda o definem como a característica de um teste enquanto mensuração de um atributo ou de uma qualidade, o qual não tenha sido " definido operacionalmente". Reconhecem, entretanto, que a validade de construto reclamava por um novo enfoque científico. De fato, definir essa validade do modo que eles a definiram parece um pouco estranho em ciência, dado que conceitos não definidos operacionalmente não são suscetíveis de conhecimento científico. Conceitos ou construtos são cientificamente pesquisáveis somente se forem, pelo menos, passíveis de representação comportamental adequada. Do contrário, serão conceitos metafísicos e não científicos. O problema é que, sintetizando a atitude geral dos psicometristas da época, para definir validade de construto, os autores partiram do teste, isto é, da representação comportamental, em vez de partir da teoria psicológica que se fundamenta na elaboração da teoria do construto (dos traços latentes). O problema não é descobrir o construto a partir de uma representação existente (teste), mas sim descobrir se a representação (teste) é legítima, adequada do construto. Esse enfoque exige uma colaboração, bem mais estreita do que existe, entre os psicometristas

e a Psicologia Cognitiva. A validade de construto de um teste pode ser trabalhada sob vários ângulos: a análise da representação comportamental do construto, a análise por hipótese, a curva de informação da TRI, além do falsete estatístico do erro de estimação da Teoria Clássica dos Testes.

Erro de estimação

Essa forma de avaliar a validade de um teste era típica da Psicometria Clássica. Esse é um modelo de psicometria que poderia ser chamado de positivista, uma vez que ele se fundamenta exclusivamente nos dados empíricos coletados de um conjunto de itens agrupados inicialmente mais ou menos de maneira intuitiva. Na verdade, o teste (conjunto de itens) é construído mediante seleção de uma amostra de itens coletados de um universo de itens que parecem medir um dado construto. Essa maneira de construir instrumentos psicométricos se fundamenta na ideia de que existe, para cada construto, um universo indefinido de itens (*pool of itens*), do qual uma amostra é extraída para constituir o teste. Como é que se sabe inicialmente que os itens incluídos na amostra se referem a um construto somente ou que estamos retirando itens de um universo unidimensional para compor o teste? Apela-se aqui à famosa ou malfadada validade aparente (*face validity*), isto é, os itens parecem estar se referindo à mesma coisa! Por mais estranho que isto pareça ser, honestamente, é o que se faz na tradição positivista da Psicometria. É que nessa tradição falta todo o teorizar prévio sobre o construto (traço latente) para o qual se quer construir o instrumento de medida. Sem os procedimentos teóricos sobre o traço latente, os itens não são construídos para representá-lo comportamentalmente, mas são coletados mais ou menos a esmo ("chutados"), com base na validade aparente, e verificados depois, por meio de análises estatísticas, para ver se de fato eles estão ou não se referindo a alguma coisa (construto) comum. Assim, a Psicometria se torna, no máximo, um ramo da Estatística, como, aliás, era normalmente definida, e não um ramo da Psicologia, como deve ser concebida. Para a Estatística, número é número, não interessa de onde ele vem; mas para a Psicologia (Psicometria) o número é uma representação de conteúdo psicológico, então interessa muito de onde ele vem. Na tradição clássica da Psicometria, apela-se demasiadamente à Estatística para salvar a teoria psicológica. Isso não se aplica. Não se pode abdicar da teoria psicológica em favor da Estatística. É preciso, primeiramente, desenvolver e avançar a teoria psicológica (dos traços latentes) e apelar, em seguida, à Estatística para auxiliar na tomada

de decisões mais objetivas sobre a demonstração de hipóteses psicologicamente significativas e relevantes, estas deduzidas da teoria psicológica e não levantadas intuitiva e aleatoriamente. A Psicometria Clássica, e também a moderna, necessita urgentemente da ajuda da Psicologia Cognitiva neste particular, a fim de que possa instrumentalizar-se com a teoria dos traços latentes, para os quais ela quer desenvolver instrumentos de observação quantitativa (medida).

De qualquer forma, também na TCT se procura demonstrar a validade dos testes. Como é que isso era feito?

Nesse contexto, a Psicometria Clássica procura legitimar a validade de um instrumento segundo o conceito de erro de estimação, isto é, quanto o escore obtido pelo sujeito no teste se afasta do escore verdadeiro.

A fórmula para o cálculo do erro de estimação (EE), na qual um critério é predito com base em um teste, é a seguinte:

$$EE = s_C \sqrt{1 - r_{TC}^2}$$

na qual, s_c é o desvio-padrão da medida do critério e r_{TC}^2 é o coeficiente de validade, isto é, a correlação entre o teste e o critério.

Essa fórmula está fundamentada na ideia de se computar o erro mínimo que se pode cometer ao se predizer o escore de um teste a partir do escore de um teste paralelo.

Para poder obter o erro de estimação, é necessário possuir a medida de um critério, este que supostamente é a medida da aptidão. Por mais precário que tal procedimento pareça ser, é um dos poucos de que dispõe a Psicometria Clássica para estabelecer o erro de estimação e, por consequência, a validade de um teste, entendida como a precisão com a qual o teste pode predizer o escore verdadeiro. A fórmula deixa claro que, se o coeficiente de validade r_{TC}^2 for zero, então o erro de estimação será igual ao desvio-padrão da medida, pois o fator sob a raiz equivaleria a 1. Tal ocorrência implicaria que o teste não é capaz de predizer o escore verdadeiro melhor do que uma simples adivinhação, isto é, ele é totalmente inútil para predizer qualquer coisa. Agora, se o coeficiente de validade for

diferente de 0, então o teste tem poder maior de predizer do que uma simples adivinhação. Quanto maior? Se o coeficiente de validade fosse igual a 1, o erro de estimação seria 0, pois ele seria o desvio-padrão multiplicado por 0. Suponha-se que um teste tenha coeficiente de validade de 0,80, o que constitui um coeficiente de grandeza extraordinária em termos práticos, nesse caso, qual seria a força de predição do teste com respeito ao critério que pretende medir? Calculando o erro de estimação do teste, tem-se

$$EE = 1 \times \sqrt{1 - 0{,}80^2} = \sqrt{1 - 0{,}64} = \sqrt{0{,}36} = 0{,}60.$$

Assim, a predição do teste é 40% (1,00 – 0,60) superior à predição feita ao acaso ou por adivinhação. Isso não parece grande coisa dado um coeficiente de validade tão elevado, mas sempre é melhor que a pura adivinhação. Também, felizmente, o erro de estimação e o coeficiente de validade não são os únicos nem os melhores procedimentos para estabelecer a validade de um teste, como se verá a seguir.

Na verdade, há nesse procedimento do erro de estimação um certo equívoco ao se supor que o escore verdadeiro seja a medida daquilo que o teste pretende medir. De fato, o escore verdadeiro constitui um agregado de medida daquilo que o teste pretende medir mais as características peculiares dos itens que compõem o teste, sem que estas tenham a ver com o que este pretende medir.

Análise da representação

São utilizadas duas técnicas para demonstrar a adequação da representação do construto pelo teste: a análise da consistência interna e a análise fatorial.

Análise da consistência interna do teste

A análise da consistência interna consiste em calcular a correlação que existe entre cada item do teste e o restante dos itens ou o total (escore total) dos itens. Dado que o item analisado contribui para o escore total, ele teoricamente não deve entrar nesse escore, já que é ele que está sendo escrutinado. Assim, a correlação legítima será a do item com o restante dos itens. Essa preocupação é importante quando o número de itens do teste for pequeno, pois nesse caso o próprio item em análise afeta substancialmente o escore total a seu favor. Por exemplo, em um teste com 10 itens, cada

um contribui e influencia o escore total em 10%. Quanto maior, contudo, o número de itens que compõem o teste, menos relevante a influência de cada um em particular no escore total. Em um teste com 100 itens, por exemplo, cada um afeta o escore total em apenas 1%. Consequentemente, no caso de um teste com grande número de itens (n ≥ 30), a correlação do item com o escore total ou com o restante dos itens não vai fazer diferença relevante.

A análise da consistência interna do teste implica o cálculo das correlações de cada item individualmente com o restante do teste. Essa análise apresenta um problema lógico que se situa no escore total. Na verdade, o escore total é o critério contra o qual cada item é avaliado; mas acontece que os itens são os que vão constituir o escore total, antes mesmo de se saber se eles são válidos e somáveis (unidimensionais, isto é, que medem um e o mesmo traço latente). O escore total constitui, assim, uma dificuldade, dado que ele somente faz sentido se o teste já é *a priori* homogêneo. A correlação de cada item com o escore total já pressupõe que os itens são somáveis, isto é, homogêneos e válidos; em outras palavras, se pressupõe que todos os itens constituam uma representação adequada do traço e de um mesmo traço latente (unidimensionalidade). Além disso, a consistência interna pressupõe que os itens estejam intercorrelacionados, isto é, que as correlações entre eles mesmos sejam elevadas. Entretanto, as intercorrelações entre os itens não são uma demonstração de que estes estejam medindo o mesmo construto. Suponha a situação de três itens saturados em três fatores, como apresentados a seguir.

Tabela 1 Itens saturados em fatores

Item	F1	F2	F3
1	0,80	0,30	0,30
2	0,30	0,80	0,30
3	0,30	0,30	0,80

As correlações entre os três itens são todas de 0,57, altas e significativas, mas nem por isso se pode dizer que eles estejam medindo uma e a mesma coisa. Na verdade, o item 1 mede especificamente o fator 1, pois está altamente saturado somente neste fator e não nos outros dois, e os outros itens medem outros fatores. Consequentemente, a análise da consistência

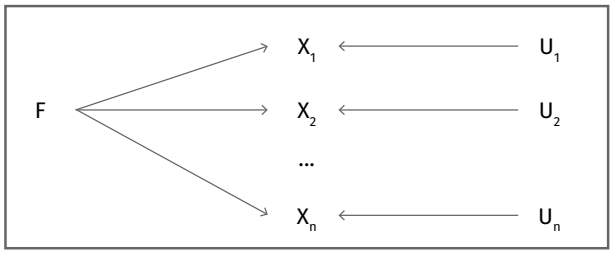
interna dos itens não parece garantir que eles sejam uma representação unidimensional de um construto.

A conclusão que se impõe dessas observações é a de que a análise da consistência interna não constitui prova cabal de validade de construto do teste.

Análise fatorial

Por outro lado, a análise fatorial tem como lógica precisamente verificar quantos construtos comuns são necessários para explicar as covariâncias (intercorrelações) dos itens. As correlações entre os itens são explicadas, pela análise fatorial, como resultantes de variáveis-fonte que seriam as causas dessas covariâncias. As variáveis-fonte são os construtos ou traços latentes de que fala a Psicometria. A análise fatorial também postula que um número menor de traços latentes (variáveis-fonte) é suficiente para explicar um número maior de variáveis observadas (itens), como se verifica na figura 1.

Figura 1 Representação do modelo fatorial



O modelo da figura 1 mostra que *n* variáveis (X) podem ser explicadas por um fator comum a todas as variáveis (F) e um fator específico para cada uma delas (U), de sorte que cada variável tem sua equação expressa em termos destes dois fatores.

Por exemplo:
$X_1 = a_1F + d_1U_1$.

O a1 é a saturação, a correlação, a covariância (dita carga fatorial) da variável X1 no fator F. Ela representa o percentual de relação que tem com o fator, isto é, quanto porcento ela se constitui em representação do fator

(traço latente); indica, em outras palavras, se ela é uma boa representação comportamental do traço latente. Além disso, as cargas fatoriais são as que determinam a correlação entre as próprias variáveis empíricas. Assim, a correlação entre X_1 e X_2 é definida por a_1a_2.

Dessa forma, a validade de construto de um teste é determinada pela grandeza das cargas fatoriais (que são correlações que vão de -1 a +1) das variáveis no fator, sendo aquelas a representação comportamental do fator, que, por sua vez, é o traço latente para o qual elas foram inicialmente elaboradas como representação empírica. Essas cargas fatoriais representam a parte fundamental do escore verdadeiro (V) da equação da Psicometria Clássica: T = V + E. Diz-se parte fundamental porque outra parte do V é constituída pela contribuição específica do item (contida no fator U do modelo fatorial) para o escore empírico T do teste. De fato, a variância total de um item ou variável pode ser decomposta em variância comum, variância específica e variância erro.

A variância comum representa o que as variáveis do teste têm em comum (expressa pelas intercorrelações entre elas) e que é recolhida nas cargas fatoriais no fator comum F. É esta que constitui a questão da validade do teste, isto é, quanto do traço latente (fator F) é representado empiricamente pelas variáveis (itens). O restante da variância dos itens é recolhido na chamada unicidade (U) de cada item que representa tanto o que é específico de cada um deles quanto os erros de medida. Estes dois últimos aspectos da variância (especificidade e erro) são agrupados em um conceito só, a saber, a unicidade, porque eles não contribuem para a validade do teste, pois é a porção do item que não constitui representação do traço latente.

Se não houvesse dificuldades com o modelo da análise fatorial, ele constituiria uma demonstração empírica cabal da validade de construto de um teste, pois forneceria a expressão exata de quanto o teste estaria representando o traço latente. Mas, infelizmente, a análise fatorial apresenta alguns problemas importantes. Duas razões são a preocupação principal neste particular. Primeiramente, o modelo fatorial se fundamenta em equações exclusivamente lineares entre variáveis e fatores. Embora seja rotineiro em Matemática tentar, em primeira aproximação, um modelo linear, parece difícil admitir que as intercorrelações empíricas entre os itens e a relação destes com os fatores (variáveis-fonte) possam ser todas reduzidas a

equações lineares. Isso é tanto mais plausível quando se observa que em quiçá nenhum campo da Psicologia e das ciências psicossociais em geral se encontram tais equações. Encontram-se, sim, equações logarítmicas, exponenciais e outras, isto é, equações não-lineares, como, por exemplo, nas leis da Psicofísica (leis de potência) e da análise experimental do comportamento (lei da igualação). Em segundo lugar, existe o grave problema da rotação dos eixos, para a qual não existe nenhum critério objetivo, a não ser a interpretabilidade psicológica (semântica) dos fatores. Essa ocorrência permite, em tese, a descoberta de qualquer fator que se queira, o que torna a solução extremamente arbitrária. Contudo, se o teste for construído via teoria psicológica de traços latentes e não a esmo (como a coleta de uma amostra de itens com base em um universo arbitrário deles, como é de praxe na construção de testes), tem-se um critério objetivo de rotação dos eixos em função dos traços latentes para os quais os itens foram inicialmente construídos como representação comportamental. Nesse caso, a análise fatorial será utilizada como teste de hipótese e não como pesca de hipóteses, assumindo, assim, a Estatística, como é legítimo, o papel de testagem de hipóteses psicológicas formuladas pela teoria psicológica e não o papel de criar ela (Estatística) as hipóteses psicológicas (*a posteriori*).

Análise por hipótese

Essa análise se fundamenta no poder de um teste psicológico ser capaz de discriminar ou predizer um critério externo a ele mesmo; por exemplo, discriminar grupos-critério que difiram especificamente no traço que o teste mede. Esse critério é procurado de várias formas. Há quatro entre as mais salientes e normalmente utilizadas, a saber, a validação convergente-discriminante, a idade, outros testes do mesmo construto e a experimentação.

A técnica da validação convergente-discriminante (CAMPBELL; FISKE, 1959) parte do princípio de que para demonstrar a validade de construto de um teste é preciso determinar duas coisas: *i*) o teste deve correlacionar significativamente com outras variáveis, com as quais o construto medido pelo teste deveria, pela teoria, estar relacionado (validade convergente); e *ii*) não se correlacionar com variáveis com as quais ele teoricamente deveria diferir (validade discriminante).

A idade é utilizada como critério para a validação de construto de um teste quando este mede traços que são intrinsecamente dependentes de

mudanças no desenvolvimento cognitivo/afetivo dos indivíduos, como é o caso, por exemplo, na teoria piagetiana do desenvolvimento dos processos cognitivos e da teoria de Spearman sobre a inteligência. A hipótese a ser testada nesse método é a de que o teste que mede o traço X, o qual muda claramente com a idade, é capaz de discriminar distintamente grupos de idades diferentes.

A prova que se faz nesse caso é a da diferença entre a média no teste de sujeitos mais jovens ($\overline{T}_j$) e a média de sujeitos mais adultos ($\overline{T}_a$), a saber

$$t = \frac{\overline{T}_a - \overline{T}_j}{\sqrt{\frac{s_a^2}{n_a} + \frac{s_j^2}{n_j}}}$$

na qual, $\overline{T}_j$ e $\overline{T}_a$ são as médias no teste do grupo jovem e do grupo adulto, s_j^2 e s_a^2 são as variâncias destas médias e n_j e n_a são os números de sujeitos nos dois respectivos grupos.

Os graus de liberdade para verificar a significância do teste de Student "t" são $n_j + n_a - 2$.

Na história dos testes psicológicos, esse procedimento de validação foi talvez o primeiro a ser utilizado quando Binet e Simon (1905) empregaram o critério de diferenciação por idade na seleção dos itens do seu famoso teste de inteligência. Embora a preocupação explícita dos autores fosse construir um teste que fosse capaz de predizer o desempenho acadêmico de alunos do primeiro grau, eles se basearam numa hipótese de caráter conceitual, isto é, de que as habilidades cognitivas aumentam sistematicamente com a idade cronológica (na infância) e, para medi-las, escolheram tarefas específicas, cuja execução correta correspondia a determinada faixa etária.

Esse método contém um problema, o qual consiste no fato de que a maturação psicológica pode assumir dimensões e conotações muito distintas em culturas diferentes, por um lado; por outro, outras variáveis que não o traço em questão podem ser dependentes dessa maturação, dificultando ou impossibilitando a definição dos grupos-critério somente em função

da idade. Assim, se outras variáveis se alteram com a idade, pode bem ser que estas sejam as responsáveis pelas mudanças no escore e não a idade especificamente. Isso não seria um grave problema se essas outras variáveis covariassem sistematicamente com o traço latente que o teste quer medir e, além disso, variassem do mesmo modo em qualquer contexto cultural ou sócio-econômico, o que obviamente é difícil de assumir. Dentro de uma mesma cultura, o método pode se apresentar como importante para a determinação da validade de construto.

A correlação com outros testes que meçam o mesmo traço é também utilizada como demonstração da validade de construto. O argumento é de que, se um teste X mede validamente o traço Z e o novo teste N se correlaciona altamente com o teste X, então o novo teste mede o mesmo traço medido por aquele teste.

Essa técnica também contém um problema, o qual consiste no fato de que normalmente um teste de um traço qualquer não se apresenta com tal pureza a se poder afirmar que ele mede exclusivamente o tal traço. De fato, ele mede o traço em termos de um certo nível de covariância: por exemplo, existe uma correlação de 0,70 entre o teste X e o traço, o que equivale a uma comunalidade de 49% entre os dois. Agora, o novo teste N correlaciona 0,78 com aquele teste X. Há, portanto, comunalidade de 61% entre os dois testes. Qual será, nesse caso, a comunalidade do novo teste com o traço em si? Por azar poderia acontecer que a comunalidade de 61% entre os dois testes ocorra precisamente com os 51% do primeiro teste que não covariam com o traço; nesse caso, a comunalidade do novo teste com o traço seria de apenas 10%, isto é, o novo teste seria uma representação quase totalmente equivocada do traço.

O uso da intervenção experimental aparece logicamente como uma das melhores técnicas para se decidir a validade de construto de um teste. Essa técnica consiste em verificar se o teste discrimina claramente grupos-critério "produzidos" experimentalmente em termos do traço objeto de medida do teste. Assim, um teste que mede ansiedade teria validade de construto (ansiedade) se discriminasse grupo não-ansioso de grupo ansioso, definidos estes grupos em termos de manipulações experimentais: o ansioso, por exemplo, criado assim por meio de experiências provocadoras

de ansiedade. Na medida em que se puder garantir que as manipulações feitas nos grupos-critério atingirem exclusivamente o traço em questão, a testagem da hipótese é válida. Como, normalmente, essas manipulações supostamente de uma variável de fato podem afetar uma série de outras variáveis, sobretudo se as variáveis interagirem, fica confusa a decisão sobre em que especificamente os grupos-critério diferem e, consequentemente, fica inconclusiva a decisão sobre a hipótese de que o teste discrimina os grupos-critério exclusivamente em termos do traço que ele pretende medir. Podendo-se garantir que não ocorre tal alastramento das manipulações, a hipótese fica corretamente colocada.

Em conclusão, a técnica da validação de construto via hipótese, que, de um ponto vista da metodologia científica, se apresenta como a mais direta e óbvia, esbarra na dificuldade que existe na definição inequívoca do critério a ser utilizado como representante da manifestação do traço.

Deve-se, na verdade, concluir que todas estas técnicas de validação apresentam dificuldades. Nem por isso se justifica o simples abandono delas. Primeiramente porque em ciência empírica nada existe de perfeito e isento de erro e, em segundo lugar, a consciência das dificuldades deve servir para melhorar e não abandonar as técnicas. Aliás, é recomendável o uso de mais de uma das técnicas analisadas para demonstrar a validade de construto do teste, dado que a convergência de resultados das várias técnicas constitui garantia para a validade do instrumento.

Validade com base na relação com outras variáveis

Esse tipo de validação dos testes praticamente se confunde com o conceito tradicional de validade de critério.

Concebe-se como validade de critério de um teste o grau de eficácia que ele tem em predizer um desempenho específico de um sujeito. O desempenho do sujeito torna-se, assim, o critério contra o qual a medida obtida pelo teste é avaliada. Evidentemente, o desempenho do sujeito deve ser medido/avaliado mediante técnicas que são independentes do próprio teste que se quer validar.

Costuma-se distinguir dois tipos de validade de critério: *i*) validade preditiva e *ii*) validade concorrente. A diferença fundamental entre os dois tipos é basicamente com relação ao tempo que ocorre entre a coleta da informação pelo teste a ser validado e a coleta da informação sobre o critério. Se essas coletas forem (mais ou menos) simultâneas, a validação será do tipo concorrente; caso os dados sobre o critério sejam coletados após a coleta da informação sobre o teste, fala-se em validade preditiva. O fato de a informação ser obtida simultaneamente ou posteriormente à do próprio teste não é um fator tecnicamente relevante à validade do teste. Relevante, sim, é a determinação de um critério válido. Aqui se situa precisamente a natureza central desse tipo de validação dos testes, a saber: *i.* definir um critério adequado e *ii.* medir, de forma válida e independentemente do próprio teste, esse critério.

Quanto à adequação dos critérios, pode-se afirmar que há uma série deles que são normalmente utilizados, encontram-se listados a seguir.

1) Desempenho acadêmico – Talvez seja ou tenha sido o critério mais utilizado na validação de testes de inteligência. Consiste na obtenção do nível de desempenho escolar dos alunos, seja por meio das notas dadas pelos professores, seja pela média acadêmica geral do aluno, seja pelas honrarias acadêmicas que o aluno recebeu ou seja, até mesmo, pela avaliação puramente subjetiva dos alunos por parte dos professores ou colegas. Embora seja amplamente utilizado, esse critério tem igualmente sido muito criticado, não em si mesmo mas pela deficiência que ocorre na sua avaliação. É sobejamente sabida a tendenciosidade por parte dos professores em atribuir as notas aos alunos, tendenciosidade nem sempre consciente, mas decorrente de suas atitudes e simpatias em relação a este ou aquele aluno. Essa dificuldade poderia ser sanada até com certa facilidade se os professores tivessem o costume de aplicar testes de rendimento que possuíssem validade de conteúdo, por exemplo. Como essa tarefa é dispendiosa, o professor tipicamente não se dá ao trabalho de validar (validade de conteúdo) suas provas acadêmicas.

   Nesse contexto, é também utilizado como critério de desempenho acadêmico o nível escolar do sujeito: sujeitos mais avançados,

repetentes e evadidos. A suposição é de que quem continua regularmente ou está avançado academicamente em relação a sua idade possui mais habilidade. Evidentemente, nessa história não entra somente a questão da habilidade, mas muitos outros fatores sociais, de personalidade etc., o que torna o critério bastante ambíguo e espúrio.

2) Desempenho em treinamento especializado – Trata-se do desempenho obtido em cursos de treinamento em situações específicas, como no caso de atividades ligadas à música, à pilotagem, atividades mecânicas ou eletrônicas especializadas etc. No final desse treinamento, há tipicamente uma avaliação, a qual produz dados úteis para servirem de critério de desempenho do aluno. As observações críticas feitas ao ponto 1 valem também neste parágrafo.

3) Desempenho profissional – Trata-se, nesse caso, de comparar os resultados do teste com o sucesso/fracasso ou o nível de qualidade do sucesso dos sujeitos na própria situação de trabalho. Assim, um teste de habilidade mecânica pode ser testado contra a qualidade de desempenho mecânico dos sujeitos na oficina de trabalho. Evidentemente continua a dificuldade de levantar adequadamente a qualidade do desempenho dos sujeitos em serviço.

4) Diagnóstico psiquiátrico – Muito utilizado para validar testes de personalidade/psiquiátricos. Os grupos-critério são aqui formados em termos da avaliação psiquiátrica que estabelece grupos clínicos: normais *versus* neuróticos, psicopatas *versus* depressivos etc. Novamente, a dificuldade continua sendo a adequação das avaliações psiquiátricas feitas pelos psiquiatras.

5) Diagnóstico subjetivo – Avaliações feitas por colegas e amigos podem servir de base para estabelecer grupos-critério. É utilizada essa técnica sobretudo em testes de personalidade, nos quais é difícil encontrar avaliações mais objetivas. Assim, os sujeitos avaliam seus colegas em categorias ou dão escores em traços de personalidade (agressividade, cooperação etc.), com base na convivência que eles têm com os colegas. Nem precisa mencionar as dificuldades enormes que tais avaliações apresentam em termos

de objetividade; contudo, a utilização de um grande número de juízes poderá diminuir os vieses subjetivos nessas avaliações.

6) Outros testes disponíveis – Os resultados obtidos por meio de outro teste válido, que prediga o mesmo desempenho que o teste a ser validado, servem de critério para determinar a validade do novo teste. Aqui fica a pergunta óbvia: para que criar outro teste se já existe um que mede validamente o que se quer medir? A resposta se baseia numa questão de economia, isto é, utilizar um teste que demanda muito tempo para ser respondido ou apurado como critério para validar um teste que gaste menos tempo.

No caso desse tipo de validade, é preciso atender a duas situações bastante distintas. Primeiramente, quando existem testes comprovadamente validados para a medida de algum traço, eles certamente constituem um critério contra o qual se pode com segurança validar um novo teste. Infelizmente essa situação ocorre quase exclusivamente no caso da medida da inteligência, em que dispomos de alguns testes cuja validade já tem sido comprovada repetidas vezes, como é o caso das escalas de Wechsler (1975), de Stanford-Binet (TERMAN; MERRILL, 1960) e quiçá os dois fatores de inteligência fluida e cristalizada de Cattell (1971) e o fator G de Spearman (1927). Nos outros campos, há muita confusão. Talvez em relação à personalidade já existam alguns instrumentos válidos, como, por exemplo, o Questionário de Personalidade de Eysenck (*Eysenck Personality Questionnaire* – EPQ em EYSENCK; EYSENCK, 1975), no qual ele se refere às variáveis extroversão e neuroticismo ou ansiedade. O que vale aqui é o princípio de que se houver um teste comprovadamente válido para a medida de algum traço latente, ele certamente pode servir de critério para a validação de um novo teste. Espera-se nesse caso que a correlação do novo teste seja elevada em pelo menos 0,75.

Entretanto, quando não existem testes aceitos como definitivamente validados para avaliar algum traço latente, a utilização dessa validação concorrente é extremamente precária. Essa situação infelizmente é a mais comum. De fato, existem testes para medir praticamente "não importa o quê", como atestam os Buro's Mental Measurement Yearbooks, que são publicados periodicamente com centenas e milhares de testes psicológicos existentes no mercado. Nesse caso, pode-se utilizar esses testes como critérios de validação, mas o risco é demasiadamente

grande, porque se está utilizando como critério testes cuja validade é pelo menos duvidosa.[2]

Pode-se concluir que a validade concorrente só faz sentido se existirem testes comprovadamente válidos que possam servir de critério contra o qual se quer validar um novo teste e que esse novo teste tenha algumas vantagens sobre o antigo (como, por exemplo, economia de tempo etc.).

Contudo, uma pergunta frustrante fica ao final desta exposição sobre validade de critério. Se o pesquisador empregou toda a sua habilidade para construir um teste sob as condições de maior controle possível, por que iria ele validar essa tarefa-teste contra medidas inferiores, representadas pela medida dos vários critérios aqui apresentados. "Justifica-se validar medidas supostamente superiores por medidas inferiores?" – pergunta Ebel (1961).

Com as críticas de Thurstone (1952) e sobretudo de Cronbach e Meehl (1955), a validade de critério deixou de ser a técnica panaceia de validação dos testes psicológicos em favor da validade de construto. Contudo, os critérios apresentados acima podem ser considerados bons e úteis para fins de validação de critério. A grande dificuldade em quase todos eles se situa na demonstração da adequação da medida deles; isto é, em geral, a medida deles é precária, e, por isso, deixa muita dúvida quanto ao processo de validação do teste. Entretanto, há exemplos conhecidos de testes validados mediante esse método.

Validade com base nas consequências da testagem

Embora as consequências ou o uso dos escores de um teste não pareçam ter a ver com a validade dele (GREEN, 1998; MEHRENS, 1997; COLLIVER; CONLEE; VERHULST, 2012; CIZEK; BOWEN; CHURCH, 2010), as interpretações e o uso que se fazem dos escores dos testes adquiriram grande importância e consenso entre os pesquisadores e usuários dos testes para uso legítimo destes (KANE, 2006; PERIE; MARION; GONG, 2009 NICHOLS; WILLIAMS, 2009). Trata-se mais de responsabilidade social dos testes do que prova de sua validade como

2 No Brasil existe uma saída pragmática para isso: se o teste está com avaliação favorável no Satepsi, então ele é um bom teste.

medida. Isso porque o uso dos escores de um teste para tomar decisões de intervenção precisa legitimar tal ato. Assim, o uso de escores inadequados para tomar tais decisões em uma dada situação torna a atitude do psicólogo até criminosa, o que no final das contas vai respingar sobre a qualidade do teste do qual se extraíram os escores que fundamentaram as decisões. Enfim, é uma visão pragmática dos testes psicológicos na medida em que eles são utilizados para o bem-estar do ser humano. Tal intento é legítimo e necessário. Mas será essa atitude diante dos testes psicológicos, de se tornar a preocupação central da avaliação psicológica, útil para desenvolver o conhecimento e a teoria psicológica, dado que esta se fundamenta em inferências com base nos escores para processos mentais (construtos)? Mehrens (1997, p. 17) afirma: "Pode-se investigar a validade da inferência de que um escore seja um indicador razoável do montante do construto que possui independentemente de qualquer uso específico do escore". Como consequência, não se pode utilizar análises dos efeitos do uso do teste como evidência de sua validade. Enfim, incorporar as consequências de uso dos escores de um teste na demonstração de validade do teste se apresenta ainda como uma diatribe do tipo quixotesco. Kane (2013) confessa que o usuário do teste pode confundir a invalidade do uso do escore do teste com a invalidade do significado do escore, e Mehrens (1997, p. 18), por outro lado, afirma que "se validade é tudo, então validade não é nada".[3] Enfim, Cizek, Bowen e Church (2010) afirmam que as consequências da testagem como fonte de evidência de validade simplesmente não existem na literatura profissional e na medida aplicada, bem como em trabalhos de política na área. Os autores concluem que esses achados implicam a necessidade de se buscar, pelo menos, refinamentos na teoria e práticas atuais de validação dos testes.

De qualquer forma, quais são, então, as precauções a serem tomadas nesse contexto para salvaguardar a validade de um teste considerando-se que as consequências de uso dos escores do teste impactam a validade dele?

Em primeiro lugar e sobretudo, embora um teste possa ser utilizado para várias situações ou atividades, nenhum teste é adequado para todas as situações e atividades do ser humano. Assim, o teste deveria ser elaborado para situações ou atividades específicas, do que resulta que, no final das contas,

3 Para perceber a confusão nessa área, veja Michael T. Kane (2013), que procura mostrar, em um emaranhado discurso, o enfoque fundamentado em argumentação, a validação de um teste mediante a validação das interpretações e o uso dos escores do teste.

se deveria elaborar um teste diferente para cada situação ou atividade. É isso razoável ou possível? Green (1998) e Reckase (1998) opinam que impor tal tarefa de coletar evidência para as consequências de uso dos escores do teste sobrecarrega o seu criador com uma tarefa impossível. De qualquer forma, fica como responsabilidade dele mostrar para que situações ou atividades o teste produz escores adequados para a tomada de decisões com base nele.

Validade ecológica

Finaliza-se esse texto com a validade ecológica. Esta realmente não constitui uma nova forma de coletar evidências de validade, mas sim a forma como tais evidências devem ser buscadas. Ou seja, validade ecológica significa que os métodos, os materiais e as situações de um estudo dessa natureza devem se aproximar ao máximo do mundo real que está sendo examinado (BREWER, 2000).

Um exemplo: testar alunos na sala de aula. Se eles são assim acostumados, então a validade ecológica é alta, porque o processo não irá afetar o comportamento deles. Se, ao contrário, os alunos forem testados individualmente em uma sala isolada, então a validade ecológica cai drasticamente, porque não é o ambiente em que eles estão acostumados a ser testados nem a forma como costumam ser testados.

Assim, afirmar que validade ecológica consiste em tornar a situação de pesquisa similar aos fenômenos do mundo real é correto, mas é somente um aspecto da questão. Mark A. Schmuckler (2001) apresenta três critérios ou dimensões para iniciar uma compreensão adequada do que é validade ecológica, quais sejam: natureza do ambiente, dos estímulos e da resposta.

1) Natureza do ambiente de pesquisa – Brunswik (1943) iniciou esse debate criticando a artificialidade e o isolamento das situações de pesquisa (laboratórios) com respeito à realidade de vida dos sujeitos, pois eles não são representativos dos padrões amplos da vida. Nesse sentido, Bronfenbrenner (1977, p. 516; 1979) deu uma definição clássica de validade ecológica: “validade ecológica se refere à extensão na qual o ambiente experienciado pelos sujeitos na investigação científica possui as propriedades

que são da experiência dos sujeitos do experimento". Assim a representatividade e a naturalidade do ambiente de pesquisa constitui elemento fundamental da validade ecológica, isto é, o comportamento do indivíduo deve ser aferido em um ambiente verdadeiro em que os atores se comportam costumeiramente.

2) Natureza dos estímulos – Como no caso do contexto, aqui também vale o princípio da representatividade e da naturalidade, ou seja, os estímulos ou questões devem consistir em ocorrências atuais e estáveis do mundo real (naturalidade) e que sejam relevantes ao sujeito com respeito ao objeto de interesse a ser investigado (representatividade, importância). Isto é, os estímulos (questões) não devem ser esdrúxulos e extravagantes.

3) Natureza da tarefa, do comportamento ou da resposta – Novamente, a tarefa e a resposta pedida ao sujeito deve fazer parte de sua vida, de seu dia a dia, e não ter acontecido uma vez em sua vida. Bronfenbrenner (1977, p. 513; 1979) se revolta contra o que acha ocorrer na pesquisa em Psicologia do Desenvolvimento ao afirmar que ela é "a ciência do comportamento estranho das crianças em situações estranhas com adultos estranhos durante um período curtíssimo de tempo".

A validade ecológica cria tensões entre os pesquisadores, porque uns acham que atender às demandas desta torna a pequisa menos precisa (deficiência no controle das variáveis em jogo), enquanto outros argumentam que a artificialidade da pesquisa imposta pelo controle das variáveis em jogo torna os resultados irrelevantes para as situações reais da vida. A solução desse problema provavelmente se encontra no equilíbrio entre as duas preocupações. Algo similar foi observado por Campbell e Stanley (1973) no que diz respeito à preocupação com a validade interna e à preocupação com a validade externa das pesquisas científicas: "pesquisa sem validade interna produz somente erros; pesquisa sem validade externa produz resultados inúteis".[4]

4 *"Magni passus, sed extra viam"*, diriam os romanos ("Grandes passos, mas fora do caminho!").

Referências

AMERICAN PSYCHOLOGICAL ASSOCIATION (1954). *Technical recommendations for psychological tests and diagnostic techniques*. Washington, D.C.: American Psychological Association, 1954.

AMERICAN EDUCATIONAL RESEARCH ASSOCIATION (AERA), AMERICAN PSYCHOLOGICAL ASSOCIATION (APA), NATIONAL COUNCIL ON MEASUREMENT IN EDUCATION (NCME). *Standards for psychological and educational testing*. Washington, D.C.: American Psychological Association, 1999.

ANASTASI, A. Evolving concepts of test validation. *Annual Review of Psychology*, v. 37, p. 1-15, 1986.

BINET, A.; SIMON, TH. Le développement de l'intelligence chez les efants. *Année Psychologique*, v. 14, p. 1-94, 1905.

BLOOM, B. S. *Taxonomy of educational objectives*: The classification of educational goals. Handbook I. Cognitive domain, New York: McKay, 1956. p. 201-207.

BREWER, M. B. Research design and issues of validity. In: REIS, H. T. (Ed.); JUDD, C. M. (Ed.). (2000). *Handbook of research methods in social and personality psychology*. New York, U.S.: Cambridge University Press, XII, 2000. p. 3-16.

BRONFENBRENNER, U. *The Ecology of Human Development*: Experiments by Nature and Design. Cambridge, MA: Harvard University Press, 1979.

BRONFENBRENNER, U. Toward an experimental ecology of human development. *American Psychologist*, v. 32, p. 515–531, 1977.

BRUNSWIK, E. Organismic achievement and environmental probability. *Psychological Review*, v. 50, p. 255-272, 1943.

CAMPBELL, D. T.; FISKE, D. W. Convergent and discriminant validation by the multitrait–multimethod matrix. *Psychological Bulletin*, v. 6, p. 81-105, 1959.

CAMPBELL, D. T.; STANLEY, J. *Experimental and quasi–experimental design for research*. Skokie, IL: Rand McNally, 1973.

CATTELL, R. B. *The scientific analysis of personality*. Baltimore, MD: Penguin Books, Inc., 1965.

CATTELL, R. B . *Abilities:* their structure, growth and action. New York: Houghton Mifflin, 1971.

CATTELL, R. B.; STICE, G. F. *The Sixteen Personality Factor Questinonnaire* ("The 16 P.F."). Champaign, IL: Institute for Personality and Ability Testing, 1957.

CATTELL, R. B.; WARBURTON, F. W. *Objective personality and motivation tests*. Urbana, IL: University of Illinois Press, 1967.

CIZEK, G. J.; BOWEN, D.; CHURCH, K. Sources of Validity Evidence for Educational and Psychological Tests: A Follow–Up Study. *Educational and Psychological Measurement*, v. 70, n. 5, p. 732-743, 2010.

COLLIVER, J. A.; CONLEE, M. J.; VERHULST, S. J. From test validity to construct validity... and back? *Medical Education in Review*, v. 46, p. 366-371, 2012.

COMREY. A. L. *The Comrey Personality Scales*. San Diego, CA: Educational and Industrial Testing Service, 1970.

CRONBACH, L. J. Construct validation after thirty years. In: LINN (Org.), *Intelligence: Measurement, theory and public policy* – Proceedings of a symposium in honor of Lloyd G. Hamphreys, Chicago, IL: University of Chicago Press, 1989.

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

CURETON, E. E.. Validity, reliability and baloney. *Educational and psychological measurement*, v. 10, p. 94-96, 1950.

EBEL, R. L. Must all tests be valid? *American Psychologist*, v. 16, n. 10, p. 640-647, 1961.

EMBRETSON, S. E. Construct validity: Construct representation versus nomothetic span. *Psychological Bulletin*, v. 93, p. 179-197, 1983.

EYSENCK, H. J.; EYSENCK, S. G. B. *The Eysenck Personality Questionnaire*. Sevenoaks: Hodder; Stoughton, 1975.

GARDNER, H. *Frames of Mind*. New York: Basic Book Inc., 1983.

GREEN, D. R. Consequential aspects of the validity of achievement tests: A publisher's point of view. *Educational Measurement*, v. 17, n. 2, 16-19, 1998.

GUILFORD, J. P. *The nature of human intelligence*, New York: McGraw–Hill, 1967.

JACKSON, D. N. *Personality Research Form*. New York: Research Psychologists Press, 1974.

KANE, M. T. Validation. In: BRENNAN, R. L. (Ed.), *Educational measurement*. 4th ed. Washington, D.C.: The National Council on Measurement in Education & the American Council on Education, 2006, p. 17-64.

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

KURTZ, A. K. A research test of the Rorschach test. *Personal Psychology*, v. 1, p. 41-51, 1948.

MAGER, R. F. *Medindo os objetivos de ensino ou "conseguiu um par adequado"*. Porto Alegre: Editora Globo, 1981.

MEHRENS, W. A. The Consequences of Consequential Validity. *Educational Measurement:* Issues and Practice, v. 16, p. 16-18, jun. 1997. DOI: 10.1111/j.1745–3992.1997.tb00588.x.

MEHRENS, W. A.; LEHMANN, I. J. *Measurement and evalution in education and psychology*. 3rd ed. New York: Holt, Rinehart, & Winston, 1984.

MESSICK, S. V. In: LINN, R. L. (Ed.), *Educational measurement.* 3rd ed. New York: Macmillan, p. 13-103, 1989.

MESSICK, S. V. The interplay of evidence nad consequences in the validation of performance assessments. *Educational Researcher*, v. 23, n. 2, p. 13-23, 1994.

MILLON, T. *Millon clinical Multiaxial Inventory Manual*. 2nd ed. Minneapolis, MN: National Computer Systems, 1983.

NEWELL, A., SHAW, J. C.; SIMON, H.A. Elements of a theory of human problem solving. *Psychological Review*, v. 65, p. 151-166, 1958.

NEWELL, A.; SIMON, H. A. Simulation of cognitive processes: A report on the summer research training institute. *Items*, v. 12, p. 37-40, 1958.

NICHOLS, P. D.; WILLIAMS, N. Consequences of test score use as validity evidence: Roles and responsibilities. *Educational Measurement:* Issues and Practice, v. 28, p. 3-9, 2009.

NUNES, C. H. S. S.; PRIMI, R. Aspectos técnicos e conceituais da ficha de avaliação dos testes psicológicos. In: CONSELHO FEDERAL DE PSICOLOGIA (Org.). *Avaliação psicológica: diretrizes na regulamentação da profissão.* Brasília: CFP, 2010. p. 101–128.

PASQUALI, L. *Instrumentação Psicológica*. Brasília, DF: Editora Vetor, 2010.

PERIE, M.; MARION, S.; GONG, B. Moving Toward a Comprehensive Assessment System: A Framework for Considering Interim Assessments. *Educational Measurement*: Issues and Practice, v. 28, p. 5-13, 2009. DOI:10.1111/j.1745–3992.2009.00149.x.

PRIMOFF, E. S. Job analysis attests to rescue trade testing from make–believe and shrinkage. *American Psychologist*, v. 7, p. 386, 1952.

RECKASE, M. D. Consequential Validity From the Test Developer's Perspective. *Educational Measurement*: Issues and Practice, v. 17, p. 13–16, 1998. DOI:10.1111/j.1745–3992.1998.tb00827.x.

SCHMUCKLER, M. A. What is ecological validity? A dimensional analysis. *Infancy*, v. 2, p. 419-436, 2001. DOI:10.1207/S15327078IN0204_02.

SPEARMAN, C. *The abilities of a man*. New York: MacMillan, 1927.

STERNBERG, R. J. Intelligence, information processing, and analogical reasoning: The componential analysis of human abilities. Hillsdale, NJ: Erlbaum, 1977.

STERNBERG, R. J. General intellectual ability. In STERNBERG, R. J. (ed.) *Human abilities*: An information–processing approach. New York: Freeman, 1984. p. 5-29.

STERNBERG, R. J.; DETTERMAN, D. K. *Human intelligence*: Perspectives on its theory and measurement. Norwood, NJ: Ablex, 1986.

STERNBERG, R. J.; RIFKIN, B. The development of analogical reasoning processes. *Journal of Experimental Child Psychology*, v. 27, p. 196-232, 1979.

STERNBERG, R. J. *Beyond IQ:* A Triarchic Theory of Intelligence. Cambridge: Cambridge University Press, 1985.

STERNBERG, R. J. *Metaphors of mind: Conceptions of the nature of intelligence*. New York: Cambridge University Press, 1990.

TERMAN, L. M.; MERRILL, M. A. *Stanford–Binet Intelligence Scale*: Manual for the third revision, Form L–M. Boston: Houghton Mifflin, 1960.

THURSTONE, L. L. The criterion problem in personality research. *Psychometric Lab. Rep.*, IL: University of Chicago, Chicago, n. 78, 1952.

WECHSLER, D. Intelligence defined and undefined: A realistic appraisal. *American Psychologist*, v. 30, p. 135-139, 1975.

WIGGINS, J. S. A true test: Toward more authentic and equitable assessment. *Phi Delta Kappa*, v. 79, p. 703-713, 1989.

Luiz Pasquali
Doutorado em Psicologia pela Université Catholique de Louvain, Bélgica
Professor da Universidade de Brasília
luiz.pasquali@gmail.com