Psicometria

PSYCHOMETRICS

PSICOMETRÍA

Luiz Pasquali[1]

RESUMO
A psicometria fundamenta-se na teoria da medida em ciências para explicar o sentido que têm as respostas dadas pelos sujeitos a uma série de tarefas e propor técnicas de medida dos processos mentais. Neste artigo são apresentados os conceitos e modelos da psicometria moderna e discutidos os parâmetros de validade e precisão dos testes.

DESCRITORES
Psicometria.
Reprodutibilidade dos testes.
Validade dos testes.
Estudos de validação.

ABSTRACT
Psychometrics has foundations on the theory of measurement in Sciences and is aimed at explaining the meaning of responses provided by subjects submitted to a series of tasks, and proposing techniques for the measurement of mental processes. This article presents concepts and models of modern psychometrics and discusses the validity and reliability parameters of the applied tests.

KEY WORDS
Psychometrics.
Reproducibility of results.
Validity of tests.
Validation studies.

RESUMEN
La Psicometría se fundamenta en la teoría de la medida en las ciencias buscando explicar el sentido en las respuestas de los que fueron sujetos a una serie de tareas, además de proponerse técnicas de medida de sus procesos mentales. En este artículo son presentados los conceptos y modelos de psicometría moderna, así como son discutidos los parámetros de validez y precisión de los testes.

DESCRIPTORES
Psicometría.
Reproducibilidad de resultados.
Validez de las pruebas.
Estudios de validación.

[1] Professor Pesquisador Associado do Departamento de Psicologia Social e do Trabalho do Instituto de Psicologia da Universidade de Brasília. Brasília, DF, Brasil. luiz.pasquali@pq.cnpq.br

Recebido: 15/06/2008
Aprovado: 15/12/2008

INTRODUÇÃO

A medida em ciências psicossociais

Etimologicamente, psicometria representa a teoria e a técnica de medida dos processos mentais, especialmente aplicada na área da Psicologia e da Educação. Ela se fundamenta na teoria da medida em ciências em geral, ou seja, do método quantitativo que tem, como principal característica e vantagem, o fato de representar o conhecimento da natureza com maior precisão do que a utilização da linguagem comum para descrever a observação dos fenômenos naturais.

Historicamente, a psicometria tem suas origens na psicofísica dos psicólogos alemães Ernst Heinrich Weber e Gustav Fechner. O inglês Francis Galton também contribuiu para o desenvolvimento da psicometria, criando testes para medir processos mentais; inclusive, ele é considerado o criador da psicometria. Foi, contudo, Leon Louis Thurstone, o criador da análise fatorial múltipla, que deu o tom à psicometria, diferenciando-a da psicofísica. Esta foi definida como a medida de processos diretamente observáveis, ou seja, o estímulo e a resposta do organismo, enquanto a psicometria consistia na medida do comportamento do organismo por meio de processos mentais (lei do julgamento comparativo).

A medida em ciências tem provocado diatribes entre os pesquisadores, particularmente na área das ciências sociais. Contudo, a definição mais aceita de medida foi dada por Stanley Smith Stevens em 1946, quando dizia que: medir consiste em *assinalar números a objetos e eventos de acordo com alguma regra*[1]. As regras de assinalar tais números são definidas na proposta do mesmo autor sobre os quatro níveis de medida ou escalas de medida: nominal, ordinal, intervalar e de razão. A medida nominal sendo aquela que aplica os números aos fenômenos da natureza, salvando somente os axiomas de identidade do número, ou seja, o número é utilizado somente como numeral ou símbolo gráfico. Ao utilizar o número, a escala ordinal já salva os axiomas de ordem, ou seja, a característica mais marcante do número, isto é, a magnitude - um número é por definição maior ou menor que outro, não somente diferente, ou melhor, um número é diferente do outro precisamente porque é maior ou menor que outro. As outras escalas salvam também axiomas de aditividade. Essa história dos axiomas foi detalhada por Whitehead e Russell em 1910 a 1913 e 1965, no livro *Principia Mathematica*, onde descrevem os famosos 27 axiomas do número matemático[2].

PSICOMETRIA: CONCEITUAÇÃO E MODELOS

A psicometria moderna tem duas vertentes: a teoria clássica dos testes (TCT) e a teoria de resposta ao item (TRI). A TCT foi axiomatizada por Gulliksen[3] e a TRI foi inicialmente elaborada por Lord[4] e por Rasch[5] e, finalmente, axiomatizada por Birnbaum[6] e por Lord[7].

De um modo geral, a psicometria procura explicar o sentido que têm as respostas dadas pelos sujeitos a uma série de tarefas, tipicamente chamadas de itens. A TCT se preocupa em explicar o resultado final total, isto é, a soma das respostas dadas a uma série de itens, expressa no chamado escore total (T). Por exemplo, o T em um teste de 30 itens de aptidão seria a soma dos itens corretamente acertados. Se for dado 1 para um item acertado e 0 para um errado, e o sujeito acertou 20 itens e errou 10, seu escore T seria de 20. A TCT, então, se pergunta o que significa este 20 para o sujeito? A TRI, por outro lado, não está interessada no escore total em um teste; ela se interessa especificamente por cada um dos 30 itens e quer saber qual é a probabilidade e quais são os fatores que afetam esta probabilidade de cada item individualmente ser acertado ou errado (em testes de aptidão) ou de ser aceito ou rejeitado (em testes de preferência: personalidade, interesses, atitudes). Dessa forma, a TCT tem interesse em produzir *testes* de qualidade, enquanto a TRI se interessa por produzir *tarefas* (itens) de qualidade. No final, então, temos ou testes válidos (TCT) ou itens válidos (TRI), itens com os quais se poderão construir tantos testes válidos quantos se quiser ou o número de itens permitir. Assim, a riqueza na avaliação psicológica ou educacional, dentro do enfoque da TRI, consiste em se conseguir construir armazéns de itens válidos para avaliar os traços latentes, armazéns estes chamados de bancos de itens para a elaboração de um número sem fim de testes.

A psicometria procura explicar o sentido que têm as respostas dadas pelos sujeitos a uma série de tarefas, tipicamente chamadas de itens.

O modelo da TCT foi elaborado por Spearman e detalhado por Gulliksen[3], o modelo é o seguinte:

$$\mathbf{T = V + E}$$

Onde,

T = escore bruto ou empírico do sujeito, que é a soma dos pontos obtidos no teste;

V = escore verdadeiro, que seria a magnitude real daquilo que o teste quer medir no sujeito e que seria o próprio T se não houvesse o erro de medida;

E = o erro cometido nesta medida.

Dessa forma, *o escore empírico é a soma do escore verdadeiro e do erro* e, conseqüentemente, E = T - V, bem como, V = T - E.

A Figura 1 mostra a relação entre estes vários elementos do escore empírico, onde se vê que este é a união do escore verdadeiro (V) e do erro (E), ou seja, o escore empírico ou bruto do sujeito (T – resultado no teste, conhecido como o escore tau – τ) é constituído de dois com-

ponentes: o escore real ou verdadeiro (V) do sujeito naquilo que o teste pretende medir e o erro (E) de medida, este sempre presente em qualquer operação empírica. Em outras palavras, estamos aqui assumindo que, diante do fato de que o escore bruto do sujeito difere do seu escore verdadeiro, esta diferença é devida ao erro ou, melhor, esta diferença é o próprio conceito de erro.

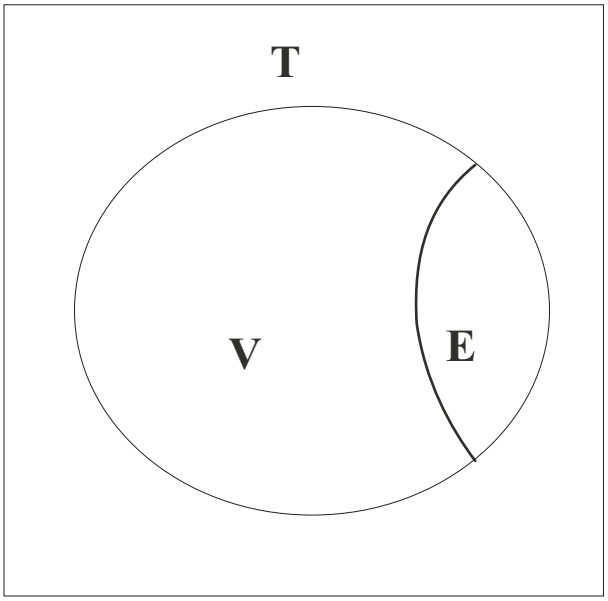

Figura 1 – Os componentes do escore T

Assim, a grande tarefa da TCT consiste em elaborar estratégias (estatísticas) para controlar ou avaliar a magnitude do E. Os erros são devidos a toda uma gama de fatores estranhos, detalhados por Campbell e Stanley[8], tais como defeitos do próprio teste, estereótipos e vieses do sujeito, fatores históricos e ambientais aleatórios.

Por outro lado, o *modelo da TRI* trabalha com traços latentes e adota dois axiomas fundamentais:

1) O desempenho do sujeito numa tarefa (item do teste) se explica em função de um conjunto de fatores ou traços latentes (aptidões, habilidades etc.). O desempenho é o efeito e os traços latentes são a causa;

2) A relação entre o desempenho na tarefa e o conjunto dos traços latentes pode ser descrita por uma equação monotônica crescente, chamada de CCI (Função Característica do Item ou Curva Característica do Item) e exemplificada na Figura 2, onde se observa que sujeitos com aptidão maior terão maior probabilidade de responder corretamente ao item e vice-versa (θ_i é a aptidão e $P_i(\theta)$ a probabilidade de resposta correta dada ao item).

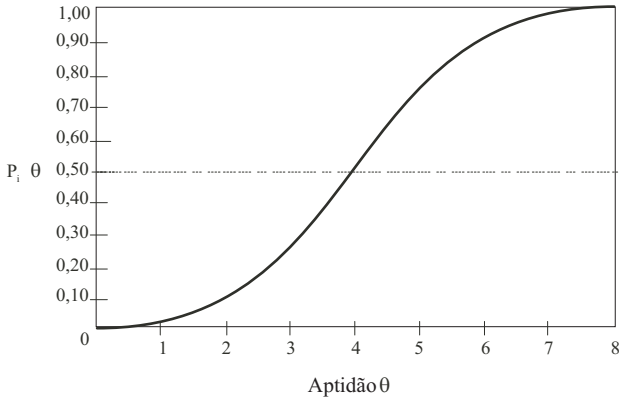

Figura 2 – A curva característica do item

Concretamente, a TRI está dizendo o seguinte: você apresenta ao sujeito um estímulo ou uma série de estímulos (tais como, itens de um teste) e ele responde aos mesmos. A partir das respostas dadas pelo sujeito, isto é, analisando as suas respostas aos itens especificados, pode-se inferir sobre o traço latente do sujeito, hipotetizando relações entre as respostas observadas deste sujeito com o nível do seu traço latente. Estas relações podem ser expressas por meio de uma equação matemática que descreve a forma de função que estas relações assumem.

De fato, pode-se imaginar um número ilimitado de modelos matemáticos que podem expressar esta relação, dependendo do tipo de função matemática utilizada e/ou do número de parâmetros que se quer descobrir para o item. Uma preciosa vantagem sobre a teoria clássica que a TRI tem quanto aos modelos que usa consiste em que os modelos utilizados pela TRI permitem desconfirmação. Na verdade, a demonstração da adequação do modelo aos dados (*model-data goodness-of-fit*) é um passo necessário nos procedimentos desta teoria. Para trabalhar com a TRI são necessários pacotes estatísticos especializados, que já existem em abundância no mercado[a].

OS PARÂMETROS DOS TESTES: VALIDADE E PRECISÃO

Tanto na TCT quanto na TRI, os dois parâmetros mais importantes de legitimidade de uma medida ou teste são a validade e a precisão.

A validade dos testes

A validade constitui um parâmetro da medida tipicamente discutido no contexto das ciências psicossociais. Ela não é corrente em ciências físicas, por exemplo, embora haja nessas ciências ocasiões em que tal parâmetro se

(a) Dois muito utilizados são o BILOG para testes de aptidão e o PARSCALE para testes de personalidade.

aplicaria. Nestas últimas ciências, a preocupação principal na medida se centra na questão da precisão, a dita calibração dos instrumentos. Esta é importante também na medida em ciências psicossociais, mas ela não tem nada a ver, conceitualmente, com a questão da validade. A razão disto está no fato de que a validade diz respeito ao aspecto da medida ser *congruente* com a propriedade medida dos objetos e não com a exatidão com que a mensuração, que descreve esta propriedade do objeto, é feita. Em Física, o instrumento é um objeto físico que mede propriedades físicas; então parece fácil se 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 esta propriedade (comprimento), isto é, o metro, usa a sua propriedade de comprimento para medir a comprimento de outro objeto; então estamos medindo 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, onde 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 conseqüências, em hipóteses dela derivadas ou deriváveis e verificáveis. Neste caso específico, o problema da precisão da medida diz respeito à quão exata pode ser feita a mensuração das distâncias entre as linhas espectrais no osciloscópio, ao passo que o de validade diz respeito a se esta medida das distâncias das linhas espectrais, por mais exata e perfeita que ela possa ser, tem 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 adequação (legitimidade) da representação ou da modelagem da velocidade galáctica via distâncias das linhas espectrais.

Este caso da astronomia ilustra o que tipicamente acontece com a medida em ciências psicossociais e, conseqüentemente, torna a prova da validade dos instrumentos nestas ciências algo fundamental e crucial, ou seja, é uma condição *sine qua non* demonstrar a validade dos instrumentos nestas ciências. Isto é particularmente o caso nos enfoques que, em Psicologia, trabalham com o conceito de traço latente, onde 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, no seu parâmetro fundamental e indispensável.

Nos manuais de Psicometria, costuma-se definir a validade de um teste dizendo que ele é válido se de fato mede o que supostamente deve medir. Embora esta definição pareça uma tautologia, na verdade ela não é, considerada a teoria psicométrica que admite o traço latente. O que se quer dizer com esta definição é que, ao se medirem os comportamentos (itens), que são a representação física do traço latente, está-se medindo o próprio traço latente. Tal suposição é justificada se a representação comportamental for legítima. Esta legitimação somente é possível se existir uma teoria prévia do traço que fundamente que a tal representação comportamental constitui uma hipótese dedutível desta teoria. A validade do teste (este constituindo a hipótese), então, será 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 esta ênfase felizmente está voltando ou deverá voltar ao seu devido lugar na Psicometria.

A validade diz respeito ao aspecto da medida ser congruente com a propriedade medida dos objetos e não com a exatidão com que a mensuração, que descreve esta propriedade do objeto, é feita.

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

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.é., válidos) da amostra inicial de itens[9].

A validação da representação comportamental do traço, isto é, do teste, embora constitua o ponto nevrálgico da Psicometria, apresenta dificuldades importantes que se situam em três níveis ou momentos do processo de elaboração do instrumento, a saber, ao nível 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 encon-

tra ainda em estado embrionário, destituída quase que totalmente de qualquer nível de axiomatização, resultando disto uma pletora de teorias, muitas vezes até contraditórias. Basta lembrar de teorias como behaviorismo, psicanálise, psicologia existencialista, psicologia dialética e outras, que, existindo simultaneamente, postulam 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. Havendo esta confusão no campo teórico dos construtos, torna-se extremamente difícil para o psicometrista operacionalizar estes mesmos construtos, isto é, formular hipóteses claras e precisas para testar ou, então, formular hipóteses psicologicamente úteis. Ainda quando a operacionalização for um sucesso, a coleta da informação empírica não será isenta de dificuldades, como, por exemplo, a definição inequívoca de grupos critérios onde estes construtos possam ser idealmente estudados. Mesmo ao nível das análises estatísticas encontramos problemas. Pela lógica da elaboração do instrumento, a verificação da hipótese da legitimidade da representação dos construtos se faz por análises do tipo da análise fatorial (confirmatória), que 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, a análise fatorial 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[10].

Diante destas dificuldades, os psicometristas recorrem a uma série de técnicas para viabilizar a demonstração da validade dos seus instrumentos. Fundamentalmente, estas técnicas podem ser reduzidas a três grandes classes (o modelo trinitário): técnicas que visam a validade de construto, validade de conteúdo e validade de critério[11-12].

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 conceito de 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[12].

O conceito de validade de construto foi elaborado com o já clássico artigo de Cronbach e Meehl[13] *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*). Estas 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 qualidade, o qual não tenha sido *definido operacionalmente*[13]. Reconhecem, entretanto, que a validade de construto reclamava por um novo enfoque científico. De fato, definir esta 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 está em 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 psicométrica 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) constitui uma representação legítima, adequada, do construto. Este enfoque exige uma colaboração, bem mais estreita do que existe, entre psicometristas e Psicologia Cognitiva[14]. 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[15-16].

A *validade de critério* de um teste consiste no 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 por meio de técnicas que são independentes do próprio teste que se quer validar.

Costuma-se distinguir dois tipos de validade de critério: (1) validade preditiva e (2) validade concorrente. A diferença fundamental entre os dois tipos é basicamente uma questão do tempo que ocorre entre a coleta da informação pelo teste a ser validado e a coleta da informação sobre o critério. Se estas 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 deste tipo de validação dos testes, a saber: (1) definir um critério adequado e (2) medir, válida e independentemente do próprio teste, este critério.

Quanto à adequação dos critérios, pode-se afirmar que há uma série destes que são normalmente utilizados quais sejam:

1) *Desempenho acadêmico*. Talvez seja ou foi o critério mais utilizado na validação de testes de inteligência. Consiste na obtenção do nível de desempenho escolar dos alunos, seja através das notas dadas pelos professores, seja pela média acadêmica geral do aluno, seja pelas honrarias acadêmicas que o aluno recebeu ou seja, mesmo, pela avaliação puramente subjetiva dos alunos em termos de *inteligente* por parte dos professores ou colegas. Embora seja amplamente utilizado, este 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. Esta 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 esta tarefa é dispendiosa, o professor tipicamente não se dá ao trabalho de validar (validade de conteúdo) suas provas acadêmicas.

Neste contexto, é também utilizado como critério de desempenho acadêmico o *nível escolar* do sujeito: sujeitos mais avançados, repetentes e evadidos. A suposição sendo de que quem continua regularmente ou está avançado academicamente em relação à sua idade possui mais habilidade. Evidentemente, nesta história não entra somente a questão da habilidade, mas muitos outros fatores sociais, de personalidade, etc., tornando este 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 músicos, pilotos, atividades mecânicas ou eletrônicas especializadas, etc. No final deste 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, neste 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 deste 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 vs. neuróticos, psicopatas vs. 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 esta técnica, sobretudo, em testes de personalidade, onde é difícil encontrar avaliações mais objetivas. Assim, os sujeitos avaliam seus colegas em categorias ou dão escores em traços de personalidade (agressividade, cooperação, etc.), baseados 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 juizes poderá diminuir os vieses subjetivos nestas 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 deste 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. Entretanto, quando não existem testes aceitos como definitivamente validados para avaliar algum traço latente, a utilização desta validação concorrente é extremamente precária. Esta situação infelizmente é a mais comum. De fato, nós temos 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. Neste caso, pode-se utilizar estes testes como critérios de validação, mas o risco é demasiadamente grande, porque se está utilizando como critério testes cuja validade é pelo menos duvidosa.

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 este novo teste tenha algumas vantagens sobre o antigo (como, por exemplo, economia de tempo etc.). 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 esta tarefa-teste contra medidas inferiores, representadas pela medida dos vários critérios aqui apresentados. Justifica-se validar medidas supostamente superiores por medidas inferiores?[17]. Com as críticas de Thurstone em 1952 e sobretudo de Cronbach e Meehl em 1955[13,18], a validade de critério deixou de ser a técnica panacéia de validação dos testes psicológicos em favor da validade de construto. Contudo, estes critérios 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 dos mesmos é precária, deixando, por isso, muita dúvida quanto ao processo de validação do teste. Entretanto, há exemplos famosos de testes validados através deste método, como é o caso do MMPI.

A *validade de conteúdo* de um teste consiste em verificar se o teste 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 em testes de desempenho, que pretendem cobrir um conteúdo delimitado por um curso programático específico[11].

A precisão dos testes

O parâmetro da precisão ou da fidedignidade dos testes vem referenciado sob uma série elevada e heterogênea de nomes. Alguns destes nomes resultam do próprio conceito deste parâmetro, isto é, eles procuram expressar o que ele de fato representa para o teste. Estes nomes são, principalmente, precisão, fidedignidade e confiabilidade. Outros nomes deste parâmetro resultam mais diretamente do tipo de técnica utilizada na coleta empírica da informação ou da técnica estatística utilizada para a análise dos dados empíricos coletados. Entre estes nomes, podemos relacionar os seguintes: estabilidade, constância, equivalência, consistência interna.

A fidedignidade ou a precisão de um teste diz respeito à característica que ele deve possuir, a saber, a de medir sem erros, donde os nomes precisão, confiabilidade ou fidedignidade. Medir sem erros significa que o mesmo teste, medindo os mesmos sujeitos em ocasiões diferentes, ou testes equivalentes, medindo os mesmos sujeitos na mesma ocasião, produzem resultados idênticos, isto é, a correlação entre estas duas medidas deve ser de 1. Entretanto, como o erro está sempre presente em qualquer medida, esta correlação se afasta tanto do 1 quanto maior for o erro cometido na medida. A análise da precisão de um instrumento psicológico quer mostrar precisamente o quanto ele se afasta do ideal da correlação 1, determinando um coeficiente que, quanto mais próximo de 1, menos erro o teste comete ao ser utilizado.

O problema da fidedignidade dos testes era tema preferido da psicometria clássica, onde a parafernália estatística de estimação deste parâmetro mais se desenvolveu, mas ele perdeu muito em importância dentro da psicometria moderna em favor do parâmetro de validade.

De qualquer forma, dentro da TCT o *coeficiente de fidedignidade*, r_{tt}, é definido estatisticamente como a correlação entre os escores dos mesmos sujeitos em duas formas paralelas de um teste, T_1 e T_2. Assim o coeficiente de fidedignidade se define como função da covariância [$Cov(T_1,T_2)$] entre as formas do teste pelas variâncias ($S^2_{T_1}$ e $S^2_{T_2}$) das mesmas, isto é, $r_{tt} = \frac{S^2_V}{S^2_T}$

onde,

r_{tt}: coeficiente de fidedignidade

S^2_V: Variância verdadeira do teste

S^2_T: Variância total do teste.

Praticamente, existem duas grandes técnicas estatísticas para decidir a precisão de um teste, ou seja, a correlação e a análise da consistência interna.

A técnica da correlação é utilizada no caso do teste – reteste e das formas paralelas de um teste. Nestes casos temos os resultados dos mesmos sujeitos submetidos ao mesmo teste em duas ocasiões diferentes ou respondendo a duas formas paralelas do mesmo teste. O índice de precisão, neste caso, consiste simplesmente na correlação bivariada entre os dois escores dos mesmos sujeitos.

Para o caso da análise da consistência interna existe uma parafernália complexa de técnicas estatísticas, que finalmente se reduzem a duas situações: a divisão do teste em parcelas - mais comumente em duas metades - com a subseqüente correção pela fórmula de predição de Spearman-Brown, e as várias técnicas do coeficiente alfa, sendo o mais conhecido o alfa de Cronbach. Nesses casos, existe a aplicação de somente um teste numa única ocasião; as análises consistem em verificar a consistência interna dos itens que compõem o teste. Trata-se, portanto, de uma estimativa da precisão, cuja lógica é a seguinte: se os itens se *entendem*, isto é, covariam, numa dada ocasião, então irão se entender em qualquer ocasião de uso do teste.

CONCLUSÃO

Para assegurar que os testes apresentem os parâmetros de qualidade cientificamente exigidos, a American Psychological Association (APA) estabeleceu os *Standards for Educational and Psychological Testing*, tendo várias edições a partir de 1985.

REFERENCES

1. Stevens SS. On the Theory of Scales of Measurement. Science. 1946;103(2684):677-80.

2. Whitehead AN, Russell B. Principia mathematica. Cambridge: Cambridge University Press; 1910-1913, 1965. 3 v.

3. Gulliksen H. Theory of mental tests. New York: Wiley; 1950.

4. Lord FM. A theory of test scores. Iowa (IA): Psychometric Society; 1952. (Psychometric Monograph, n. 7).

5. Rasch G. Probabilistic models for some intelligence and attainment tests. Copenhagen: Danish Institute for Educational Research and St. Paul; 1960.

6. Birnbaum A. Some latent trait models and their use in inferring and examinee's ability. In: Loed FM, Lord MR. Novick, statistical theories of mental test scores. Reading: Addison Wesley; 1968. p.17-20.

7. Lord FM. Applications of item response theory to practical testing problems. Hillsdale: Erlbaum; 1980.

8. Campbell DT, Stanley J. Experimental and quasi-experimental designs for research. Skokie: Rand McNally; 1973.

9. Anastasi A. Evolving concepts of test validation. Ann Rev Psychol. 1986;37(1):1-15.

10. Pasquali L, organizador. Instrumentos psicológicos: manual prático de elaboração. Brasília: LabPAM/IBAPP; 1999.

11. Pasquali L. Análise fatorial para pesquisadores. Porto Alegre: Artmed; 2005.

12. American Psychological Association (APA).Technical recommendations for psychological tests and diagnostic techniques. Washington; 1954.

13. Cronbach LJ, Meehl PE. Construct validity in psychological tests. Psychol Bull. 1955;52(4):281-302.

14. Pasquali L.Validade dos testes psicológicos: será possível reencontrar o caminho? Psicol Teor Pesq. 2007; 23 (n.esp):99-107.

15. Pasquali L. Psicometria: teoria dos testes na psicologia e na educação. Petrópolis: Vozes; 2004.

16. Pasquali L. TRI - Teoria de Resposta ao Item: teoria, procedimentos e aplicações. Brasília: LabPAM/UnB; 2007.

17. Ebel RL. Must all tests be valid? Am Psychol. 1961;16 (10):640-7.

18. Thurstone LL.The criterion problem in personality research. Chicago: University of ChicagoPress; 1952.

Correspondência: Luiz Pasquali
Campus Darci Ribeiro, ICC Sul
LabPAM, sala AI-096
Plano Piloto - Asa Norte
CEP 70910-900- Brasília, DF, Brasil