PSICOLOGIA,SAÚDE & DOENÇAS, 2013,*14 (1),* 245-263
EISSN - **2182-8407**
Sociedade Portuguesa de Psicologia da Saúde - SPPS - www.sp-ps.com

MEDIDA NA AVALIAÇÃO PSICOLÓGICA

José Luis Pais-Ribeiro

Faculdade de Psicologia e de Ciências da Educação, Universidade do Porto, Porto, Portugal

RESUMO - Neste estudo discute-se o papel da medição na avaliação psicológica criticando a excessiva valorização das explicações da qualidade dos instrumentos com recurso a números simples, assim como do abuso de estatísticas complexas com escalas que não são apropriadas para tais estatísticas. Defende-se que em vez de recorrer a modelos matemáticos para legitimar a avaliação psicológica se deveria recorrer a modelos psicológicos. Explica-se que desde as origens da avaliação psicológica, ou estes procedimentos estatísticos não eram utilizados, ou estavam limitados pelo tipo de escalas que essas avaliações utilizavam. Explica-se também que a validade, desde a década de 80 do século passado, quando se adotou uma perspetiva integrada de validade, passou a basear-se na combinação de ações complementares (rede nomológica), que podem integrar procedimentos estatísticos mas não se esgotam neles. Defende-se que numa época em que muita da instrumentação psicológica e de técnicas ou procedimentos que foram originalmente desenvolvidos pela psicologia são utilizados por muitos não psicólogos, os psicólogos devem ter especial cuidado com o uso da avaliação psicológica, com os vários passos desta, a começar na escolha das técnicas de avaliação, na sua aplicação e, principalmente, na interpretação dos resultados e do seu uso, todos eles subordinados a procedimentos técnicos e éticos próprios da psicologia. É esta sequência informada, que torna a avaliação psicológica um instrumento da psicologia e dos psicólogos. Muitos investigadores, muitos profissionais, em muitas áreas, utilizam os instrumentos da psicologia para os mais diversos fins, mas tal, defendemos, não é avaliação psicológica.

Palavras chave- avaliação psicológica; medição; validade

MEASUREMENT IN PSYCHOLOGICAL ASSESSMENT

ABSTRACT - The present study discusses the role of measurement in psychological assessment, criticising the abuse in the use of single numbers, as well as complex statistics with inadequate. At its inception, psychological assessment did not use complex statistics. The study defends that psychologists must base their procedures on psychological models, rather than mathematical models. We state that starting in the 1980's, when validity became a unified concept, a combination of different and integrated procedures including - eventually- statistics, started being utilized. We argue that at a time when much of the psychological techniques and procedures originally developed by psychology are used by non-psychologists, special care with the use of psychological assessment tools should be taken, including the choice of evaluation techniques, their application, and in particular regarding the interpretation and use of results, which must be guided by the rules and ethical principles of psychology. It is this informed sequence which makes psychological evaluation an instrument that needs to be applied by psychologists. Many researchers, professionals, and fields use techniques and instruments originally developed by, or following the principles of psychology, but that does not in itself constitute psychological assessment.

Key words- Psychological assessment; measurement; validity

Recebido em 19 de Feveriro de 2013/ Aceite em 25 de Março de 2013

A avaliação psicológica, desde os primórdios da psicologia, já no século XIX com a proposta de Fechner (1860) de métodos para avaliar a intensidade das sensações, ocupou uma posição central na afirmação da psicologia como ciência. Mas foi principalmente nos primeiros decénios do século XX, até à década de 30, que se constata o sucesso da avaliação psicológica, principalmente da avaliação da inteligência na sociedade norte americana, nomeadamente na seleção de candidatos para a Primeira Grande Guerra Mundial. A par da inteligência e da personalidade muitos outros conceitos ou construtos foram avaliados com outros instrumentos, muitos dos quais não tinham conversão numa pontuação *(score)*, como seja o caso dos testes projetivos de personalidade, de estrutura cognitiva de inspiração Piagetiana ou do raciocínio moral.

Hoje, em parte pelo desenvolvimento dos computadores e dos programas estatísticos, parece haver um abuso do enfoque na pontuação dada pelos instrumentos de avaliação. É fortemente questionável se a avaliação psicológica se esgota numa pontuação e, mais importante, se a avaliação da qualidade destes instrumentos são melhor apreciados por uma abordagem matemática.

Salienta-se que a validade de todo e qualquer instrumento de avaliação não se expressa por um número: ela requer uma análise complexa que relacione vários aspetos, nomeadamente, objetivos da avaliação, contexto, variáveis a avaliar, sujeito ou população avaliada e, essencialmente, os resultados, as consequências da avaliação. Cronbach e Meehl (1955) expressavam claramente que "*Construct validity cannot generally be expressed in the form of a single simple coefficient*" (p.300). Michell, (2013) defende que as diferenças na realização das pessoas são melhor explicadas em termos de diferenças qualitativas, entre recursos cognitivos relevantes, do que em termos de diferenças de magnitudes do tipo de estruturas quantitativas baseadas na psicometria. Também Ferris (2004), na sua análise do conceito de medição em geral, explica que esta não consiste em apresentar a realidade num número.

Para descrever a validade, principalmente a partir do momento em que a validade passou a adotar uma perspetiva unificada, é necessário conhecer as teorias, os conceitos/construtos, e avaliar e conhecer as teorias psicológicas que deram origem ao instrumento, e que explicam esses conceitos/construtos. Há mais de 50 anos, Guilford (1952) alertava para eventuais abusos do uso dos números como ameaça à psicologia. Dizia ele:

> The use of a complicated statistical procedure like factor analysis does not permit one to forget about the usual safeguards that should surround scientific observations. Statistical operations do not compensate for carelessness in making observations. Rather, they presuppose careful observations. They then serve as an important aid in seeing order in the observations and in making sense of that order. Under inappropriate conditions of observation, data may appear to have an order that is misleading if not fictitious. There is no statistical magic that will give a good ordered view of nature when the data do not permit (p.26-27).

Mais recentemente Hambleton (2001) recomendava que um investigador cauteloso deverá aplicar diversos procedimentos estatísticos e interpretá-los em combinação com a evidência.

Numa revisão sobre a teoria dos testes nos últimos 50 anos, Blinkhorn (1997) explicava que se verificava uma ênfase em modelos estatísticos em vez de em modelos psicológicos, o que tornava esses modelos inacessíveis para grande parte dos utilizadores. Barrett (2008) refere que modelos estatísticos sofisticados são utilizados para produzir resultados pouco relacionados com a prática diária e, sequer, com consequências científicas úteis. Como afirma Borsboom (2008), se construirmos uma base de dados de tal modo que contenha números, e esses números forem tratados com as análises estatísticas mais usuais – como seja, p.ex., a análise de variância ou análise em componentes principais – as conclusões relativas a esses números são de uma forma simples generalizados como atributos psicológicos em que o investigador está interessado. Ou seja, tende-se a assumir um isomorfismo entre atributos psicológicos e os números da base de dados, quando se devia ter em consideração que os atributos medidos não se conformam automaticamente aos números da base de dados. Diz o mesmo autor que tal se baseia num sistema de pensamento operacionalista o qual defende que os atributos teóricos são iguais aos atributos medidos, enquanto seria esperado que a maioria dos psicólogos subscrevesse a tese que os atributos teóricos e as suas medidas são aspetos distintos (Borsboom, 2006).

De facto a pontuação observada (*observed scores*) não substitui o atributo teórico. Borsboom, (2006) ilustra do seguinte modo:

> both in textbooks on psychological methods and in actual research, the dominant idea is that one has to find an "operationalization" (read: observed score) for a construct, after which one carries out all statistical analyses under the false pretense that this observed score is actually identical to the attribute itself. In this manner, it becomes defensible to construct a test for, say, self-efficacy, sum up the item scores on this test, subsequently submit these scores to analysis of variance and related techniques, and finally interpret the results as if they automatically applied to the *attribute* of self-efficacy because they apply to the *sumscore* that was constructed from the item responses (Borsboom, 2006, p.428).

Ou seja, neste caso da auto-eficácia, como seria noutro atributo, o instrumento, os itens são um elemento periférico desse atributo e não o próprio atributo. McGrath (2005), afirma que o conceito de validade supõe que os contrutos são independentes da sua medição.

A psicologia como ciência

No final do século XX, verificou-se uma reorientação da ciência em geral para uma ciência que Almeida (2009) designa por Realista, de inspiração Darwinista, agora ao lado das ciências exatas, mais clássicas ou duras para utilizar a linguagem de Becher (1994). As ciências sociais, como a psicologia, ficam deste modo reduzidas a pouco ou nada, diz, e "qualquer disciplina que queira assenhorear-se do epíteto "científico" não tem outro caminho à sua frente a não ser o de seguir o modelo das ciências naturais" (Almeida, 2009, p.34/5).

A ideologia atual, as próprias crises políticas e económicas que se manifestaram a partir da década de 80 do século passado, facilitaram a emergência de uma ciência "Realista", centrada nos interesses económicos, duras por isso. Nussbaum (2010), numa perspetiva filosófica, critica este movimento da ciência para as áreas duras, salientando a falta de uma perspetiva mais humanista que grassa na ciência atual. Curiosamente, o mesmo faz um editorial da Nature (2005) um dos jornais científicos de referência para a publicação de "ciência dura".

Na psicologia, nos últimos decénios, verifica-se um interesse pela utilização de variáveis e medidas mais duras, como sejam os constituintes químicos do metabolismo humano, imagens

do cérebro, ou registos gráficos das respostas elétricas do cérebro ou do coração, para legitimar a investigação psicológica como se, assim, a psicologia passasse a ser uma ciência dura e, por isso, mais séria. A análise estatística mais sofisticada facilitada por *software* e *hardware* cada vez mais poderosos são uma das vertentes desta orientação mais "dura" da psicologia.

Focando a psicologia, Michell (2008), explica que a adoção da perspetiva Realista se deveu a dois grupos de interesses: ideológicos e económicos. Os ideológicos estão relacionados com o que ele designa por Cientismo, para significar que, "Knowing something *scientifically* means *measuring* it" (p.10). Os interesses económicos têm, por um lado, a ver com a comercialização da instrumentação usada na psicologia, mas principalmente com a necessidade, após a segunda guerra mundial se desenvolver a Grande Ciência, e de os governos ocidentais terem feito grandes investimentos na investigação científica. As bolsas de investigação tornaram-se um instrumento fundamental para afirmação dos investigadores na sua carreira, para as suas disciplinas se afirmarem e, para sustentar as instituições científicas e académicas. Continua Michell (2008), que este imperativo levou a que disciplinas como a psicologia, nas margens das ciências estabelecidas, que tinham que se candidatar aos restos do que se despendia com a "boa ciência", tentassem desenvolver um rigorismo que a valorizasse aos olhos das boas ciências, e deste modo pudessem ser candidatas a bolsas de investigação disponibilizadas pelos organismos científicos oficiais ou outros.

Imperativo Quantitativo, Praticalismo, Operacionalismo, Realismo Empírico, são ideias (ideologias) modernamente associadas à ciência, que têm conduzido a psicologia, incluindo a sua vertente de avaliação psicológica, para campos cada vez mais estreitos e, por isso, provavelmente mais afastados das raízes da psicologia. Por definição a psicologia não se esgota na avaliação psicológica e muito menos, na medição.

Há a ideia naíve de que para qualquer coisa ser considerada científica tem que envolver medição. Designada por "Imperativo Quantitativo", consiste na ideia que a medição é uma característica necessária a toda a ciência (Michell, 1990). Desenvolveu-se nos últimos 26 séculos com origem na filosofia de Pitágoras, e foi o motor da filosofia da revolução científica no século XVII (Barrett, 2003). Assume que a natureza e a realidade, em geral, se revelam através de princípios matemáticos e numéricos, razão pela qual eles têm servido para explicar os fenómenos físicos e psicológicos de modo a permitir que sejam científicos. Na mesma linha encontra-se o "Praticalismo", ideia que a ciência deverá servir fins práticos (Michell, 1997). Explica este autor que, no entanto, a ciência enquanto tentativa para compreender e explicar o modo como a natureza funciona, ignora totalmente o Praticalismo: este não é necessário nem útil para o conhecimento científico em si. Deve-se juntar ainda o "Operacionalismo", a ideia que o significado de um conceito está sediado, se expressa no conjunto de operações utilizadas para o especificar ou identificar. Na psicologia, mais concretamente na avaliação psicológica, ele expressa-se, p.ex. na teoria clássica dos testes (*classic test theory*) cuja ideia central é que os atributos teóricos são iguais aos observados (Borsboom, 2006). Stevens (1935) foi um dos principais defensores do Operacionismo na psicologia.

Outra atitude científica que se propõe fazer uma melhor defesa da ciência é o "Realismo" (empírico, científico), explica Michell (1997), o qual assume que o mundo que a ciência descreve é o mundo real, ou seja, que é independente do que pensamos que ele é. Passado para a validade dos instrumentos de avaliação o Realismo assume que os construtos psicológicos existem enquanto realidade objetiva mesmo que a capacidade de os medir seja

fraca (McGrath, 2005). Esta é a ideia central do Positivismo e é também designado por Realismo Naïve (Guba & Lincoln 1998). O Positivismo defendia que o objetivo do conhecimento era descrever os fenómenos que se podem observar e medir. Conhecimento para além disso seria impossível. A emergência do Pós-Positivismo constituiu a total rejeição da perspetiva do Positivismo, assumindo o "Realismo Crítico", a saber, que toda a observação é enquadrada por uma teoria, e que é falível. É crítico sobre a possibilidade de conhecer a realidade com exatidão, com certeza. Nesta perspetiva pós positivista toda a observação é falível e contém erros, levando a que toda a teoria pode/deve revista: ou seja o Realismo Crítico critica a nossa capacidade para conhecer uma realidade sem incerteza (Robson, 2002).

Medição não é sinónima de avaliação psicológica

A medição é considerada um dos aspetos centrais no método científico, embora seja surpreendente a falta de uma discussão apurada sobre este assunto na literatura metrológica (Michell, 2005). No entanto a avaliação psicológica é muito mais do que, e é independente de, medição.

A avaliação psicológica tem aparecido estreitamente ligada à ideia de medição, embora esta ligação seja ambígua. Ferris (2004) discute, no inglês, inúmeras definições de medição e em resultado da análise e da crítica a essas definições, propõe a seguinte: "*Measurement is an empirical process, using an instrument, effecting a rigorous and objective mapping of an observable into a category in a model of the observable that meaningfully distinguishes the manifestation from other possible and distinguishable manifestations*"(p.107). Salienta que a medição descreve a relação observador-contexto-observado, e que o seu resultado expressa a compreensão do que observador observa sobre o observado. Esta compreensão tem o suporte de um modelo que é prévio à avaliação, e a técnica de avaliação é escolhida e utilizada no âmbito desse modelo, ambas (modelo e técnica) são enquadradas por uma teoria psicológica.

Na primeira metade do século XX Stevens (1946, p.677), definia medição, em sentido lato, como "*the assignment of numerals to objects or events according to some rule*". Pelo facto desta atribuição de números a objetos ou eventos ser feita segundo regras leva, dizia o autor, a diferentes tipos de escalas e a diferentes tipos de medição. Torna-se assim necessário, continua Stevens, tornar explícitas: a) as regras para atribuição de números, b) as propriedades matemáticas (ou estrutura de grupo) das escalas resultantes, c) as operações estatísticas que são aplicáveis às medições realizadas com cada tipo de escala. No mesmo artigo ele propõe os clássicos tipos de escalas que a psicologia utiliza, mais as correspondentes estatísticas que elas permitem nomeadamente, escalas nominais, ordinais, intervalares, de razão. A maioria das escalas utilizadas em psicologia são ordinais, continua, e "*in the strictest propriety the ordinary statistics involving means and standard deviations ought not to be used with these scales*" (Stevens, 1946, p.679).

Stevens desenvolveu uma teoria coerente de representações numéricas. A ideia básica em Stevens é que a medição envolve a modelação numérica de aspetos do mundo real (Realismo) (Stevens, 1951). Os aspetos modelados diferem em complexidade dando origem a diferentes tipos de escalas. Assim, modelar uma classificação dá origem a uma escala nominal; modelar uma ordem dá origem a uma escala ordinal; modelar diferenças no nível de um atributo a uma escala intervalar; modelar níveis de rácios de um atributo dá origem a uma escala de razão. A sua teoria de escalas de medição e a sua elaboração constituem um recurso inestimável para a psicologia (Michell, 2002)

A terminologia introduzida por Stevens em 1946, ainda é a referência utilizada na maioria dos manuais de avaliação psicológica, e parece estar correta, embora os psicólogos tendam a menosprezar as propriedades métricas das escalas (Michell, 2008) e a tratá-las como se fossem variáveis contínuas (para utilizar a linguagem da estatística), ou intervalares ou de razão (para utilizar a linguagem da avaliação psicológica introduzida por Stevens). Barrett (2003) e Michell (2008) afirmam que à primeira vista a organização proposta por Stevens parece razoável. As críticas atuais, no entanto, dizem que a diferenciação que Stevens propôs não chega.

Michell (1999) explica que, dado que a medição envolve a assunção da existência de atributos quantitativos, ela impõe uma questão prévia: o atributo é ou não quantitativo? Se sim, a medição pode prosseguir, se não o exercício está todo errado. Kline (1997) defende que uma ciência quantitativa se inicia com duas tarefas: primeiro confirmar a hipótese de que o atributo em estudo é quantitativo, seguida da tarefa prática, fundamental, de escolher os procedimentos para medir a magnitude dos atributos assumidos como quantitativos. Conclui dizendo que o mal é que estas duas tarefas não são realizadas pelos psicólogos e outros, assumindo-se, levianamente, que as variáveis são quantitativas. Sobre isto Barrett (2003), afirma que a utilização da aritmética e de operações algébricas com números que são assumidos como "medidas", e em que os resultados são tratados como tal, é usual, mas a validade das conclusões que são daí tiradas fica comprometida e as conclusões são, provavelmente, falsas.

Críticas à medição em psicologia

A questão da medição em psicologia não é nova e assumiu uma posição importante na primeira metade do século XX. As propostas de Stevens (1946) referidas acima, a sua definição de medição, de escalas e das suas propriedades, constituem uma referência básica em psicologia, e foram formuladas em resposta à Comissão Ferguson com o nome original de British Ferguson Committee (Ferguson et al. 1938; 1940). Esta comissão, que incluía físicos e psicólogos, foi formada em 1932 pela *British Association for the Advancement of Science* para investigar a possibilidade de se avaliar quantivamente os eventos sensoriais.

Um dos principais críticos atuais da utilização irrefletida da medida na psicologia é Michell (1990, 1997). Afirma que a "psicologia quantitativa moderna está mais preocupada com a implementação de programas quantitativos do que com a resposta a questões científicas fundamentais sobre essas hipotéticas quantidades" (Michell, 1997, p.362). Vários dos títulos deste autor são elucidativos ("Normal science, pathological science and psychometrics", Michell, 2000; "Is psychometrics pathological science?" Michell, 2008; "The psychometricians' fallacy: too clever by half", Michell, 2009). Seguido por muitos outros autores como Barret (2003; 2008) que refere que a ênfase da atual teoria dos testes na estatística em vez de nos modelos psicológicos é inacessível à maioria dos utilizadores, e não espelha preocupações psicológicas, ou Kline (1997). Paul Kline exemplifica com uma escala de locus de control na saúde dizendo (Kline,1997, p.386):

> Here items which have face validity, e.g. 'When I get sick, I am to blame' and 'No matter what I do, I am likely to get sick', are factored and items loading a particular factor are regarded as scales named from the high-loading items. With such a scale the unit of measurement is unknown. Often with only six items per scale it is difficult to see what universe of items they might purport to represent. That they factor together indicates nothing more than that they mean the same thing. This type of blind factoring

> is bound to yield factors if enough items which are essentially paraphrases of each other are included in a test. With this methodology, there is literally no end to factors which can be produced.

Conclui dizendo que "*This kind of psychometrics in which the scales are the variables, simply because their items load a factor, does seem to be measurement gone mad*" (Kline, 1997, p.386). Este aspecto que Kline critica é conhecido de todos os psicólogos, e expressa uma abordagem leviana da psicologia e da avaliação psicológica.

Avaliação psicológica

A avaliação psicológica pode definir-se como a actividade científica e profissional que consiste em recolher, integrar e avaliar dados, acerca de um sujeito, com recurso, sempre que possível, a diferentes fontes, de acordo com um plano estabelecido previamente, de modo a responder às questões colocadas pelo cliente: junta-se o desenvolvimento, construção e avaliação de meios adequados para juntar e processar informação apropriada para a avaliação. Integra duas componentes: Processo e procedimentos. O processo de avaliação refere-se à sequência de passos que o avaliador deverá seguir para responder às questões do cliente; os procedimentos de avaliação referem-se aos instrumentos, testes, e outras técnicas de medida, incluindo métodos qualitativos, de juntar dados (Fernández-Ballesteros, et al. 2001).

A avaliação psicológica avalia construtos que, como refere Kane (2001), são ideias desenvolvidas para organizar e explicar aspetos do conhecimento existente. A literatura psicológica mistura o termo conceito com construto. Conceito é uma invenção, construto uma descoberta, dizem Maraun e Peters, (2005): Conceito é um elemento da linguagem e construto um elemento empírico explicam. Markus (2008) também discute a diferença entre estes termos explicando que os construtos se referem a casos reais, enquanto os conceitos abrangem tanto os casos reais como os casos possíveis. Os construtos seriam dependentes da população para compreender o seu significado, enquanto o significado dos conceitos seria independente da população, aplicando-se a qualquer população. De qualquer modo não se discutirão aqui estes termos, remetendo os interessados para estes autores.

Desde Cronbach e Meheel (1955) que a validade é um processo complexo: Ele salienta a desadequação de muitos procedimentos de validação que, p.ex. se suportam num coeficiente simples (frequentemente duvidoso), ou se baseia, simplesmente, na opinião de especialistas (Cronbach, 1971).

Validade

A validação exige uma análise extensa de evidência, baseada em afirmações explícitas sobre as interpretações, e envolve a tomada em consideração de aspetos vários e contraditórios. Aborda-se a validade como uma hipótese e recorre-se à teoria, à lógica e ao método científico para recolher e juntar dados que suportem ou recusem as interpretações num dado momento, como explica Downing (2003).

Pasquali (2007) identifica mais de 30 termos utilizados na literatura psicológica para definir diferentes tipos de validade. Focando a definição de validade, qualquer psicólogo ou estudante de psicologia dirá que é a propriedade de uma técnica de avaliação que garante que ela mede o que se propõe medir, definição produzida no início do século XX (Kelly, 1927). Esta definição não está errada mas é pueril: a validade é mais complexa do que esta simples definição poderia sugerir. Identificar a validade é um processo complexo que integra números (ideia de quantidade quando contamos, ordenamos e medimos), numerais (representação de

um número, seja ela escrita, falada ou indigitada), ou algarismos (símbolo numérico usado para formar os numerais escritos). Estes constituem (ou não) escalas, utilizadas num processo de medição que deve ser definido na avaliação psicológica como quantitativa ou qualitativa.

Messick (1995, p. 741) explica que "*Validity is an overall evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions on the basis of test scores or other modes of assessment*".

A validade, continua este autor, não é uma propriedade do teste ou da avaliação enquanto tal, mas sim do significado das pontuações do teste ou instrumento. Estas pontuações são função, não apenas dos itens ou estímulos, mas também das pessoas que respondem e do contexto onde ocorre a avaliação. Essencialmente, o que necessita ser válido é o significado ou interpretação da pontuação, em paralelo com as implicações para a acção. Esta definição é a adotada nos manuais de avaliação psicológica de referência, nomeadamente está explicitada na página nove da edição em vigor dos *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999). É a perspetiva dominante, que não se afasta da formulação original de Cronbach e Meheel (1955) que afirmavam que "*One does not validate a test, but only a principle for making inferences*"(p.297). A literatura discute se a validade é um atributo da técnica de avaliação ou das inferências que se tiram do seu resultado (p.ex. Borsboom, Mellenbergh, & van Heerden, 2004; McGrath, 2005).

A validade, desde a segunda metade do século passado, tornou-se um conceito unificado (Elosua, & Iliescu, 2012; Kane, 2001; 2013). Loevinger (1957) defendia que "*since predictive, concurrent, and content validities are all essentially ad hoc, construct validity is the whole of validity from a scientific point of view*" (p. 636). A validade tornou-se então uma abordagem global, integrada, à validade incluindo dados sobre conteúdo, critério, construto, fidelidade e muitos outros parâmetros associados à teoria dos testes, incluindo as suas consequências, como tem sido defendido por Messick (1995) e pelos Standards for Educational and Psychological Testing (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999) entre outros. No entanto a psicologia continua a utilizar os procedimentos de validação dos anos 80, recorrendo à inspeção de dados e de validações parciais (Elosua, & Iliescu, 2012). Bornstein (2011) e Hubley e Zumbo (2011) falam da validade consequencial, como uma parte fundamental da validade, nomeadamente as consequências pessoais e sociais da avaliação. Bornstein (2011) fala ainda da validade evidencial como uma variante da anterior. Hunsley (2002) explica que:

> muito pouco se sabe acerca da validade e da utilidade da avaliação psicológica. Isto não significa que a avaliação psicológica não tenha mérito; antes indica que, tal como muitos outros aspectos da prática psicológica, há falta de evidência científica que sustente a validade da avaliação. Os psicólogos têm que construir uma ciência da avaliação, e não somente um corpo de investigação sobre testes e sub-escalas. Se a avaliação psicológica é para promover com base científica, deverá sê-lo com base em estudos relevantes de avaliação e não a partir de extrapolações em literatura sobre a validade dos testes (p.140).

A validade requer a utilização de um plano forte (*strong program*), linguagem utilizada por Cronbach e Meehl (1955) em oposição a um plano fraco (*weak program*) (Kane, 2001; 2013). O plano fraco é uma simples aplicação empírica, do tipo encontrar uma correlação com outro

teste semelhante. O plano forte implica a explicitação clara das ideias subjacentes ao conceito e construto. Um plano forte de validação começa pelo desenvolvimento de uma teoria forte, e é impossível de aplicar sem esta (Kane, 2001; 2013).

É reconhecido que um programa da validação forte é dispendioso em termos de tempo e recursos e, na presença de dificuldades de aplicar um plano forte, a tendência mais comum, é utilizar um plano fraco, ou até, de passar ao lado da validação. Como se entende, o plano forte não tem sido muito utilizado nos últimos 60 anos (Kane, 2001; 2008). A validade é, então, um processo complexo e, dado que a ciência psicológica gera novos conceitos e reorganiza os estabelecidos, "*validation is never finished"* (Cronbach, 1988, p.5- em itálico no original).

Como dizem Thorndike e Hagen (1977) a evidência da validade é complementarmente racional e empírica. A consideração racional ocupa o centro do processo quando estamos a pensar no produto final (a descrição do individuo) e o modo como a validade da medida é um ato, é um exercício racional. Groth-Marnat (2003) também defende que não há uma abordagem única para determinar a validade, mas sim uma variedade de possibilidades diferentes: Uma abordagem básica da validade, que consiste em avaliar em que medida o teste mede um coceito/construto teórico ou traço, deve envolver três etapas gerais. Primeiro, deve-se fazer uma análise cuidadosa do traço; Segue-se uma consideração do modo como ele deve/pode estar ligado a outras variáveis; Finalmente, deve-se testar se essas ligações hipotéticas existem. Esta proposta de Groth-Marnat (2003) assemelha-se à rede nomológica.

Cronbach e Meehl (1955) recomendam a utilização de uma rede nomológica para evidenciar se a medida é válida. Uma rede nomológica consiste na representação dos conceitos (e dos construtos) em estudo, na observação do modo como se manifestam, e na inter e intra-relação entre eles: Uma rede nomológica deve relacionar quantidades ou propriedades observáveis umas com a as outras ou; relacionar construtos observáveis com os teóricos; ou relacionar construtos teóricos uns com os outros. A menos que, com recurso a rede nomológica, se façam observações, se exibam inferências, e se esclareçam os passos utilizados para as fazer, a existência da validade não pode ser declarada. Não chega racionalizar acerca do construto ou medi-lo. É necessário estabelecer uma cadeia de inferências para confirmar que uma técnica mede um conceito/construto. Para isso têm que existir operações relativamente complexas como as de uma rede nomológica.

Sobre os métodos de validação Cronbach e Meehl (1955) recomendavam as seguintes acções: procurar diferenças entre grupos que seja esperado serem diferentes; procurar a correlação entre testes: se é suposto eles medirem o mesmo construto então deve haver correlação entre eles; inspeção da homogeneidade dos itens que, se medem o mesmo construto devem evidenciar correlação entre eles; investigar a estabilidade entre momentos de passagem diferentes: esta estabilidade deve estar de acordo com a teoria subjacente e assim, devem evidenciar mais ou menos estabilidade consoante se tratar de um traço ou de um estado, por exemplo; investigar o processo de responder ou do modo como a pessoa responde.

Na procura da validade encontraremos sempre autores com posições extremadas, uns pretendem exprimir a validade num único ou poucos índices, sejam a resultante de uma análise em componentes principais ou da análise fatorial confirmatória, mais Realistas, enquanto outros ignoram os números e exprimem-na de forma teórica. Muitos outros, provavelmente a maioria utilizam os números numa perspetiva de Realismo Crítico, a par com outros indicadores, próximo de uma análise da rede nomológica.

A validade das técnicas de avaliação são comuns a todas as ciências e disciplinas que fazem avaliação. Por exemplo, a avaliação na medicina exige evidência de validade para ser

interpretada de modo significativo (Downing, 2003). A medicina tende a interessar-se pelo conteúdo dos itens mais numa perspetiva clinimétrica, enquanto a psicologia com uma orientação mais psicométrica tende a orientar-se mais pela estatística explicam de Vet, Terwee, e Bouter (2003). Wijsman, Hekster, Keyser, Renier e Meinardi (1991) definem clinimética como "a ciência de quantificação dos fenómenos clínicos com particular atenção para a validação das variáveis de resultado"(p182). Feinstein (1994) discute a psicometria e a clinimetria considerando que se diferenciam em vários aspectos, nomeadamente: os instrumentos psicométricos contêm inúmeros itens, agregados, ou não em sub-escalas, que mascaram ou obscurecem sintomas particulares que são significativos do ponto de vista médico, como é o caso da dispneia ou da dor numa articulação que deve ser o foco da intervenção; a psicometria suporta-se em procedimentos, ou em conceitos, que não são familiares para o contexto médico tais como, validade, fidelidade, consistência interna, etc; um instrumento que possua valores elevados para um contexto clínico pode não o ter para outro contexto; a agregação dos itens em sub-escalas produzem indicadores que são pouco sensíveis às mudanças; embora na avaliação baseada em considerandos psicométricos os doentes expressem os seus sentimentos, valores e crenças, o resultado final é tratado por procedimentos matemáticos que devem ser robustos.

De fato as preocupações que diferenciam as duas atividades não são grandes. Afinal muita da instrumentação utilizada pela psicologia, principalmente em contexto de saúde, é clinimétrica.

Podemos utilizar procedimentos matemáticos complexos com medidas que não permitem esses procedimentos?

Michell (2002), defende que tratar atributos ordinais como se fossem estruturas intervalares conduz a conclusões inválidas. Os atributos ordinais não se podem relacionar quantitativamente (e.g., linearmente ou multiplicativamente) a outros atributos: assim, métodos que estudam relações presumivelmente quantitativas, como a análise fatorial, têm valor duvidoso na identificação de atributos subjacentes diz Michell (2002). Em resumo, com este tipo de medidas deveríamos utilizar métodos concebidos para trabalhar com atributos ordinais.

Por outro lado, outros autores, como Nunnally (1967) designam a perspetiva de Michell de "fundamentalista", defendendo que não existem escalas intervalares intrínsecas. Para ele qualquer escala seria uma convenção entre cientistas, e uma boa escala seria aquela que se concordava ser uma boa escala, e que funcionasse bem na prática. Nunnally e Bernstein (1994) explicavam que um item individual é claramente ordinal, mas que, quando se somam vários itens para obter uma pontuação total, o resultado se aproxima de uma escala intervalar, assumindo intervalos iguais.

Há, então, dois grupos extremos, os *fundamentalistas* e, em oposição, os *levianos*, para quem qualquer procedimento estatístico se pode aplicar a qualquer grupo de números. E esta última é o que de fato se faz usualmente na medição, seja em psicologia seja em medicina, educação, criminologia, organizações, e em todas as ciências sociais em geral.

Se de um ponto de vista técnico (matemático) está obviamente errado como Michell (1990, 1997, 1999), entre muitos outros têm defendido; então como compreender a adesão a estes "procedimentos errados" que todos fazem? O próprio Stevens (1946) quando definiu as escalas que se tem utilizado neste tipo de medição, explicava:

> …for this 'illegal' statisticizing there can be invoked a kind of pragmatic sanction: In numerous instances it leads to fruitful results. While the outlawing of this procedure would probably serve no good purpose, it is proper to point out that means and standard deviations computed on an ordinal scale are in error to the extent that the successive intervals on the scale are unequal in size (p.679).

Ou seja, sendo errado, é normalmente utilizado por razões pragmáticas e frutuosas. O mesmo é afirmado por Kline (1997) e Barret (2003) entre outros. Kline (1997) explica que a maioria dos autores clássicos recorre à estatística como ponto de partida para a sua análise. Como contornar esta aparente contradição? Com o recurso ao processo nomológico como Cronbach e Meehl (1955) propuseram, e onde, na complexidade da rede nomológica o recurso a uma estatística "ilegal" pode ser útil, pode ajudar, à compreensão da massa de dados disponíveis e do atributo.

No desenvolvimento de qualquer técnica de avaliação começa-se pela teoria. Quando se propõe avaliar um conceito, tal como a inteligência ou a personalidade, ele deve ser integrado na teoria. Não há uma inteligência, não há uma personalidade: há inúmeras. Com a teoria como moldura principal, passa-se ao conceito, ao construto, ao desenvolvimento da técnica de avaliação (itens, escalas, etc.), à análise de dados (ligação do conteúdo dos itens à teoria e ao conceito; de seguida inspeccionam-se as diferenças entre grupos que devam, ou não, ser diferentes; a correlação entre testes que medem o mesmo conceito/construto; a homogeneidade dos itens; a estabilidade entre momentos de passagem diferentes; o modo como a pessoa responde e, finalmente, as consequências). Se a técnica resistir a todo este processo está dado o primeiro passo para publicar e divulgar a técnica assim como todo o processo que lhe deu origem de modo a que outros estudos possam refutar os dados que foram encontrados.

Velleman e Wilkinson (1993), suportando-se na filosofia de Thomas Kuhn defendem que a anomalia é um elemento importante na consciência da violação do paradigma em vigor e, por isso, importante em ciência. Por isso, diz, uma análise de dados responsável, que persiga o desenvolvimento científico, deve estar aberta à anormalidade. Assim não se deve, dizem, recusar o uso de estatísticas (mesmo que tecnicamente inapropriadas) que facilitem a identificação de anormalidades

Avaliação psicológica ou testagem psicológica

Meyer et al. (2001) salientam a distinção entre testagem psicológica e avaliação psicológica: A testagem psicológica é um processo linear básico em que uma escala é aplicada para obter uma pontuação específica à qual pode ser dado um significado, com base em dados normativos e nomotéticos. Pelo contrário, a avaliação psicológica respeita a interacção clínica, idiográfica, em que se recolhe variada informação, obtida geralmente através de múltiplos testes e instrumentos, considera os dados no contexto da história, a informação proveniente de avaliações complementares, e o comportamento observado, visando compreender a pessoa que está a ser avaliada, para responder a questões colocadas por outros clínicos, e para depois comunicar a informação ao doente/cliente, aos outros significativos, ou às entidades legitimas que pediram a avaliação.

O cliente, em contexto de saúde, é geralmente a equipa que pede a avaliação, equipa esta que inclui o psicólogo. Ou seja, uma equipa de saúde, de instituições mais ou menos especializadas, incluem inúmeros profissionais, desde os médicos (cirurgiões, anestesistas, imageologistas, analistas, entre outros), farmacêuticos, nutricioniatas, enfermeiros (de várias

especialidades), fisioterapeutas, psicólogos, etc. A informação a recolher para tomar decisões apropriadas é vasta e complexa, e cada profissional deve estar apetrechado para identificar e descrever aspectos importantes que contribuam para esse todo.

O psicólogo, através da avaliação psicológica, deverá estar apto a identificar e descrever aspectos psicológicos que são susceptíveis de facilitar ou embaraçar, quer a reacção ao diagnóstico quer a reacção e ajustamento à doença, aos tratamentos (incluindo a adesão), a curto ou a longo prazo.

A propósito da avaliação psicológica Fernández-Ballesteros, et al. (2001) explicam que:

1) O processo de avaliação implica um processo de tomada de decisão, ou seja, visa a utilização de procedimentos (no processo) úteis para a tomada de decisão visando a resolução de problemas práticos importantes;

2) O processo de avaliação implica resolução de problemas, ou seja, é um processo de constante questionamento, e implica, entre outros, um conjunto de fases a) de clarificação do problema, b) planificação, c) desenvolvimento, d) implementação, e) encontrar um resultado e, f) disseminação;

3) O processo de avaliação requer a produção de hipóteses, inerente ao processo clínico.

Estes autores explicam ainda a existência de um duplo significado do termo "avaliação" que, proveniente do inglês "*assessment*" e "*evaluation*", se refere, respectivamente, a uma avaliação (*assessment*) que foca as pessoas, o sujeito humano, e a avaliação (*evaluation*) que se refere a um objecto concreto que está a ser avaliado (o conceito ou construto). Ou seja, enquanto o foco científico da avaliação (*assessment*) psicológica é uma pessoa (ou grupo de pessoas) o foco científico da avaliação enquanto (*evaluation*) é um programa ou um grupo de acções.

A avaliação psicológica tem sempre uma dimensão clínica, no seu sentido mais lato, tal como é utilizado em educação, em saúde, organizações, ou outra (Pais Ribeiro & Leal, 1996) e, por isso, a interpretação da pontuação e do processo que lhe deu origem, de uma qualquer técnica de avaliação, deve ser feita por quem conhece bem a teoria subjacente, o processo de validação, e as implicações da decisão que se retira com essa interpretação. Meyer et al. (2001) explicam que a avaliação psicológica consiste na combinação que é feita, seguindo um método clínico, recolhendo uma larga variedade de pontuações e informações com recurso a diferentes métodos, à sua ligação ao contexto, à história de vida, e a outras informações que foram enviadas, e a observação do comportamento, entre outras, para compreender a pessoas que está a ser avaliada. Ou seja quando se utiliza avaliação, seja em que contexto for, a validade é um aspeto essencial: sem validade a avaliação é irrelevante e mesmo fonte de erro de compreensão do fenómeno em observação.

Ora, hoje não é claro onde começa e onde termina a avaliação psicológica. Por um lado os instrumentos de avaliação que recorrem a procedimentos e processos de avaliação que eram próprios da psicologia ou que nasceram com ela, são utilizados por muitos outros profissionais. Se alguns conceitos/construtos são mais facilmente conotados com a psicologia, como sejam a inteligência e a personalidade não patológica, muitos outros (auto-estima, auto-eficácia, locus de controlo, coping, esperança, espiritualidade, etc) são utilizados quer na saúde, na educação, nas organizações, por outros profissionais, principalmente na investigação e, frequentemente, na avaliação dos resultados da intervenção.

A avaliação psicológica deve ter valor preditivo ou diagnóstico e faz-se com recursos a instrumentos técnicos que, por isso, devem ser utilizados por profissionais com treino no seu uso. Com efeito, a utilização adequada de um teste psicológico requer um treino longo, por

várias razões: em primeiro lugar porque cada teste mede um conceito, um construto: os psicólogos devem conhecer profundamente a teoria subjacente ao instrumento, e os conceitos e construtos que ele avalia. Só assim se pode compreender e explicar os resultados, conhecer o modo como é aplicado, compreender o modo como o respondente se comporta, saber cotar e reportar os resultados, e os procedimentos éticos inerentes à utilização do teste, entre outros.

O que acontece quando alguém usa um teste psicológico (testagem psicológica) e chega a uma pontuação como resultado? Nada. Tal como qualquer pessoa pode utilizar um esfigmomanómetro para medir a pressão arterial, pode pesar-se numa balança, pode contar as pulsações, etc, também qualquer pessoa pode utilizar com a mesma facilidade um teste psicológico. Para que o número obtido com o teste tenha significado, tem que se garantir, pelo menos, dois aspectos: primeiro que foi obtido de modo correcto. Só com treino adequado se pode garantir que o resultado a que se chega é o resultado correto (seja com um teste psicológico ou de pressão arterial), dado que as condições de aplicação são restritas e, por isso, objeto de treino; em segundo lugar porque o resultado a que se chega só tem significado à luz da teoria e dos conceitos que avalia e que os leigos não dominam. Este aspeto é decisivo para que uma técnica de avaliação seja válida (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999; Messick, 1995). As consequências da avaliação são um elemento decisivo na avaliação psicológica. Assim, dizer que tem um QI de 98 não significa nada porque, não só depende de como foi passado o teste, como a partir de que teste foi obtido.

Pelas razões expostas os testes ou instrumentos de avaliação psicológica, não estão acessíveis a qualquer um. Ao pretender adquirir um teste nas empresas que os vendem, é exigido que o comprador demonstre que é idóneo, que é licenciado em psicologia e que pode utilizar o teste.

Construção, adaptação ou adoção de instrumentos de avaliação psicológica

As três hipóteses que se colocam quando se pretende utilizar uma técnica, um instrumento de avaliação são: a construção, a adaptação ou a adoção da técnica. No primeiro caso - construção -, a forma mais apropriada se o objetivo principal não for a comparação entre nações ou entre culturas, devem-se respeitar as exigências técnicas que se apresentaram antes, e que tornam o processo dispendioso. Deve-se começar por identificar e definir com clareza o objetivo da avaliação que se pretende realizar, segue-se a explicitação do quadro teórico onde se inclui o conceito a avaliar, a definição clara do conceito e a sua aplicação à população (construto), a escolha e análise das dimensões que compõem o conceito e dos itens que as integram, a escolha da forma de responder e a definição da escala de resposta, a reflexão falada (*cognitive debriefing*), assim como a aplicação dos restantes procedimentos implícitos e explícitos na rede nomológica.

A adaptação é um processo semelhante à construção. Tem a vantagem de a teoria, o conceito, dimensões e itens já estarem definidos numa qualquer língua, mas tem como dificuldade, dado o principal objetivo ser que os resultados possam ser comparados nas línguas/culturas diferentes, garantir que as técnicas ou instrumentos de avaliação forneçam os mesmos resultados (Wild, et al. 2009). A adaptação é, por isso, provavelmente o processo mais complexo destes todos. Mesmo que dois países usem a mesma língua, os instrumentos necessitam ser adaptados por razões lexicais (as palavras não significam o mesmo), gramaticais (as frases não se juntam do mesmo modo e cada língua tem a sua gramática), culturais (cada cultura dá um significado próprio a conceitos). Por isso é necessário fazer

adaptações do português europeu para o do brasil, do espanhol europeu para os vários sul americanos, do inglês europeu para o norte americano. Van de Vijver e Tanzer (2004) salientam que é difícil garantir que as pontuações que se obtêm numa cultura possam ser comparadas com as obtidas noutras culturas e que essas pontuações podem ter significados completamente diferentes.

Mas as diferenças não dizem respeito somente à língua. A equivalência linguística pode não ser possível por o conceito/construto não ser idêntico ou não existir na cultura para onde se está a adaptar o instrumento (Hambleton, & Patsula, 1999). Herdman, et al. (1998), salientam que há a possibilidade de haver variações na natureza de construtos multidimensionais em diferentes culturas, pelo que é importante investigar diversos aspetos, como: que domínios são importantes para o construto na cultura em jogo, e quais são as relações entre eles (equivalência conceptual); examinar de modo crítico os itens utilizados para avaliar esses domínios, e se a relevância desses itens é idêntica nas duas culturas (equivalência do item); assegurar que a tradução respeita a equivalência semântica dos itens (equivalência semântica); assegurar que os métodos de medição utilizados são adequados para a cultura em questão (equivalência operacional); inspecionar as propriedades psicométricas do instrumento (equivalência de medida) e finalmente; examinar o resultado do processo em termos de comportamento do instrumento (equivalência funcional).

Hambleton, e Patsula, (1999) afirmam que quando a comparação inter-culturas não é importante, pode ser mais relevante e mais fácil desenvolver um novo instrumento na nova língua. Em último caso a adoção de um instrumento pode ser a forma mais prática de resolver o problema expica Van de Vijver, (2003). Van de Vijver, e Hambleton, (1996) afirmam que quando se traduz um instrumento há três opções: aplicar o instrumento com uma tradução literal; adaptar partes do instrumento; ou construir um novo instrumento. Cada uma destas ações podem tornar o instrumento mais adequado para o novo grupo cultural.

A International Test Comission (2010) e Muñiz, Elosua, e Hambleton, (2013) propõe linhas orientadoras a ser utilizadas na tradução e adaptação de testes e instrumentos de avaliação psicológica, e para identificar a equivalência de pontuação entre línguas e grupos culturais.

CONCLUSÃO

A avaliação psicológica, na sua vertente prática, tem três grandes fases; os procedimentos, os processos, e o uso dos resultados. Qualquer destes inclui detalhes fundamentais. No procedimento e processo deve-se garantir que as técnicas de avaliação e/ou os instrumentos utilizados são os mais válidos para responder à questão para que a avaliação foi concebida; ao mesmo tempo, a aplicação das técnicas deve respeitar critérios éticos que estão universalmente definidos para a avaliação psicológica: o último detalhe mais importante diz respeito ao uso dos resultados que incluem: as consequências da avaliação, que devem responder apropriadamente à questão que deu origem à avaliação psicológica, e o respeito ético no uso desses resultados.

O uso dos resultados, porque são informados pela escolha das técnicas de recolha de dados, as quais estão associadas a uma teoria psicológica à luz da qual esses resultados vão ser interpretados, é o que define e exige a função de psicólogo. O psicólogo está obrigado a garantir a adequação da avaliação psicológica e é por isto que a avaliação é psicológica e só

pode ser realizada por psicólogos credenciados, que tenham que responder perante estruturas de vigilância éticas, pela adequação da sua prática (neste caso a ordem dos psicólogos).

Quando se estuda um instrumento para utilizar na avaliação psicológica deve-se fazer uma validação forte. Utilizar a tradução mais uns procedimentos estatísticos simples é uma validação fraca. Chega, é útil? Poderá servir sempre para apoiar todos os procedimentos de validação já feitos ou que outros fizeram. Se vários investigadores ou clínicos publicarem os seus estudos fracos sobre a validação que fizeram, e se eles forem no sentido de que o instrumento é estável e, principalmente, se a utilização dos seus resultados são úteis, essa é uma boa contribuição para a validade do instrumento. Deve-se ter consciência que os procedimentos utilizados são normalmente parciais e, por isso, devem ser utilizados com prudência.

REFERÊNCIAS

Almeida, O. (2009). *De Marx a Darwin - A desconfiança das ideologias.* Lisboa: Gradiva.

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

Barrett, P.(2003). Beyond psychometrics: Measurement, non-quantitative structure, and applied numerics. *Journal of Managerial Psychology, 18,* 421 – 439. doi:org/10.1108/02683940310484026

Barrett, P. (2008). The Consequence of Sustaining a Pathology: Scientific Stagnation - a Commentary on the Target Article "Is Psychometrics a Pathological Science?" by Joel Michell. *Measurement, 6,* 78–83. doi:10.1080/15366360802035521

Becher, T. (1994). The Significance of Disciplinary Differences. *Studies in Higher Education, 19,* 151-161. doi**:**10.1080/03075079412331382007

Blinkhorn, S. (1997). Past imperfect, future conditional: Fifty years of test theory. *British Journal of Mathematical and Statistical Psychology, 50,* 175–186. doi:org/10.1111/j.2044-8317.1997.tb01139.x

Bornstein, R. (2011). Toward a Process-Focused Model of Test Score Validity: Improving Psychological Assessment in Science and Practice. *Psychological Assessment, 23,* 532–544. doi:10.1037/a0022402

Borsboom, D. (2006). The attack of the psychometricians. *Psychometrika, 71,* 425–440. doi:10.1007/s11336-006-1447-6

Borsboom, D. (2008). Latent Variable Theory. *Measurement, 6,* 25–53. doi:10.1080/15366360802035497.

Borsboom, D., Mellenbergh, G., & van Heerden, J. (2004). The Concept of Validity. *Psychological Review , 111,* 1061–1071. doi:10.1037/0033-295X.111.4.1061

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

Cronbach, L. J. (1988). Five perspectives on validity argument. In: H. Wainer & H. Braun (Eds.), *Test validity* (pp. 3-17). Hillsdale, NJ: Lawrence Erlbaum.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52,* 281-302. doi:org/10.1037/h0040957

de Vet, H.C., Terwee, C., & Bouter, L. (2003). Clinimetrics and psychometrics: two sides of the same coin. *Journal of Clinical Epidemiology, 56,* 1146–1147. doi:10.1016/j.jclinepi.2003.08.010

Downing, S.(2003). Validity: on the meaningful interpretation of assessment data. *Medical Education, 37,*830–837. doi:org/10.1046/j.1365-2923.2003.01594.x.

Elosua, P., & Iliescu, D. (2012).Tests in Europe: Where We Are and Where We Should Go. *International Journal of Testing, 12,* 157-175. doi:org/10.1080/15305058.2012.657316

Fechner, G. T. (1860). *Elemente der Psychophysik*. Leipzig: Breitkopf and Hartel

Feinstein, A. (1994). Clinical judgement revisited: the distraction of quantitative models. *Annals of Internal Medicine, 120,* 799-805. doi:org/10.7326/0003-4819-120-9-199405010-00012

Ferguson, A., Myers, C.S., Bartlett, R.J., Banister, H., Bartlett, F.C., Brown, W.,… & Tucker, W.S. (1938). Quantitative estimates of sensory events: Interim report of the committee appointed to consider and report upon the possibility of quantitative estimates of sensory events. *British Association for the Advancement of Science, 108,* 277–334.

Ferguson, A., Myers, C.S., Bartlett, R.J., Banister, H., Bartlett, F.C., Brown, W.,…& Tucker, W.S. (1940). Quantitative estimates of sensory events: Final report of the committee appointed to consider and report upon the possibility of quantitative estimates of sensory events. *Advancement of Science, 1,* 331–349.

Fernández-Ballesteros, R., De Bruyn, E., Godoy, A., Hornke, L., Ter Laak, J., Vizcarro, C., ... &, Zaccagnini, J. (2001). Guidelines for the Assessment Process (GAP): A Proposal for Discussion. *European Journal of Psychological Assessment, 1*, 187–200. doi:org/10.1027//1015-5759.17.3.187

Ferris. T. (2004). A new definition of measurement. *Measurement, 36,* 101–109. doi:10.1016/j.measurement.2004.03.001.

Groth-Marnat, G. (2003). Handbook of psychological assessment (4th Ed.). Hoboken, NJ. John Wiley & Sons, Inc.

Guba, E. G., & Lincoln, Y. S. (1998). Competing paradigms in qualitative research', in N. K. Denzin and Y S Lincoln (eds.), *The landscape of qualitative research: theories and issues* (pp.195-220). Thousand Oaks, Ca: Sage.

Guilford, J.P. (1952). When not to factor analyze. *Psychological Bulletin, 49,* 26-37. doi:org/10.1037/h0054935

Hambleton, R. K (2001). The Next Generation of the ITC Test Translation and Adaptation Guidelines. *European Journal of Psychological Assessment, 17,* 164–172. doi:10.1027//1015-5759.17.3.164.

Hambleton, R. K., & Patsula, L. (1999). Increasing the validity of adapted tests: Myths to be avoided and guidelines for improving test adaptation practices. *Journal of Applied Testing Technology, 1,* 1–30.

Herdman, M., Fox-Rushby, J., & Badia, X., (1998). A model of equivalence in the cultural adaptation of HRQoL instruments: the universalist approach. *Quality of Life Research,7,* 323-335. doi:10.1023/A:1024985930536

Hubley, A., & Zumbo, B. (2011). Validity and the Consequences of Test Interpretation and Use. *Social Indicators Research, 103,* 219–230. doi 10.1007/s11205-011-9843-4

Hunsley, J. (2002). Psychological testing and psychological assessment: A closer examination. *American Psychologist, 57,* 139-140. doi:org/10.1037/0003-066X.57.2.139.

International Test Commission (2010). International Test Commission Guidelines for Translating and Adapting Tests. Retirado em Março de 2013 de http://www.intestcom.org

Kane, M. (2001). Current Concerns in Validity Theory. *Journal of Educational Measurement, 38,* 319-334. doi:org/10.1111/j.1745-3984.2001.tb01130.x

Kane, M. (2008). Terminology, Emphasis, and Utility in Validation. *Educational Researcher, 37,* 76–82. doi: 10.3102/0013189X08315390

Kane, M. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement,50,*1-73. doi:org/10.1111/jedm.12000

Kelly, E. L. (1927). *Interpretation of educational measurements*. New York, NY: Macmillan.

Kline, P.(1997). Commentary on Michell, quantitative science and the definition of measurement in psychology. *British journal of Psychology, 88,* 385-387. doi:org/10.1111/j.2044-8295.1997.tb02642.x

Loevinger, J. (1957). Objective tests as instruments of psychological theory. *Psychological Reports, Monograph Supplement, 3,*635-694. doi:org/10.2466/pr0.1957.3.3.635

Maraun, M., & Peters, J. (2005).What Does It Mean That an Issue Is Conceptual in Nature? *Journal of Personality Assessment, 85,* 128–133. doi:org/10.1207/s15327752jpa8502_04

Markus, K. (2008). Constructs, Concepts and the Worlds of Possibility: Connecting the Measurement, Manipulation, and Meaning of Variables. *Measurement, 6,* 54–77. doi:10.1080/15366360802035513

McGrath, R. (2005). Conceptual Complexity and Construct Validity. *Journal of Personality Assessment, 85,* 112–124. doi:org/10.1207/s15327752jpa8502_02.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50,* 741-749. doi: 10.1037/0003-066X.50.9.741

Meyer, G., Finn, S., Eyde, L., Kay, G., Moreland, K., Dies, R.… Reed, G. (2001).Psychological testing and psychological assessment: A review of evidence and issues. *American Psychologist, 56,* 128-165. doi:10.1037//0003-066X.56.2.128

Michell, J. (1990). *An Introduction to the Logic of Psychological Measurement*. Hillsdale, NJ: Erlbaum

Michell, J. (1997). Quantitative science and the definition of measurement in psychology. *British Journal of Psychology, 88,* 355- 383. doi:org/10.1111/j.2044-8295.1997.tb02641.x

Michell, J. (1999). *Measurement in Psychology: A Critical History of a Methodological Concept.* New York, NY: Cambridge University Press.

Michell, J. (2000). Normal science, pathological science and psychometrics. *Theory & Psychology, 10,* 639–667. doi: 10.1177/0959354300105004

Michell, J. (2002). Stevens's theory of scales of measurement and its place in modern psychology. Australian *Journal of Psychology, 54,* 99 – 104. doi: 10.1080/00049530210001706563

Michell, J. (2005). The logic of measurement: A realist overview. *Measurement, 38,* 285–294. doi:10.1016/j.measurement.2005.09.004

Michell, J. (2008). Is psychometrics pathological science? *Measurement, 6*, 7–24. doi:10.1080/15366360802035489

Michell, J. (2009). The psychometricians' fallacy: too clever by half? *British Journal of Mathematical and Statistical Psychology, 62*, 41–55. doi:org/10.1348/000711007X243582

Michell, J. (2013). Constructs, inferences, and mental measurement. *New Ideas in Psychology, 31*, 13–21. doi:org/10.1016/j.newideapsych.2011.02.004.

Muñiz, J. Elosua, P., & Hambleton, R. (2013). Directrices para la traducción y adaptación de los tests: segunda edición. *Psicothema, 25,* 151-157.doi: 10.7334/psicothema2013.24

Nature (2005). In praise of soft science. *Nature, 435*, 1003. doi:10.1038/4351003a;

Nunnally, J.C. (1967). *Psychometric theory*. New York, NY: McGraw-Hill.

Nunnally, J.C., & Bernstein, I.H. (1994). *Psychometric Theory* (3rd ed.). New York, NY: McGraw- Hill Book Company.

Nussbaum, M. (2010, Setembro). Uma crise planetária da educação. *Courrier Internacional*, pp 60-65.

Pais Ribeiro, J., & Leal, I. (1996). Psicologia clínica da saúde. *Análise Psicológica, XIV*, 67-77

Pasquali, L. (2007). Validade dos Testes Psicológicos: Será Possível Reencontrar o Caminho? *Psicologia: Teoria e Pesquisa, 23*, 99-107

Robson, C. (2002). *Real World Research. A Resource for Social Scientists and Practitioner-Researchers* (2nd. Ed.). Oxford: Blackwell.

Stevens, S. (1935). The operational definition of psychological concepts. *Psychological Review, 42*, 517-527. doi:10.1037/h0056973

Stevens, S. (1946). On the Theory of Scales of Measurement. *Science, 103*, 677-680. doi:org/10.1126/science.103.2684.677

Stevens, S. S. (1951). Mathematics, measurement and psychophysics. In S. S. Stevens (Ed.), *Handbook of experimental psychology* (pp. 1–49). New York, NY: Wiley.

Thorndike, R., & Hagen, E. (1977). *Measurement and evaluation in psychology and education*. New York, NY: John Wiley & Sons.

van de Vijver, F. (2003). Test adaptation/translation methods. In R. Fernandez-Ballesteros (Edt.). *Encyclopedia of Psychological Assessment*. (pp.960-964). Thousand Oaks, Ca.: SAGE Publications Inc.

van de Vijver, F., & Hambleton, R. (1996). Translating tests: some practical guidelines. *European Psychologist,1,* 89-99. doi:org/10.1027/1016-9040.1.2.89

van de Vijver, F., & Tanzer, N. (2004). Bias and equivalence in cross-cultural assessment: an overview. *Revue Européenne de Psychologie Appliquée, 54,* 119–135. doi:10.1016/j.erap.2003.12.004

Velleman, P., & Wilkinson, L. (1993). Nominal, Ordinal, Interval, and Ratio Typologies are Misleading**.** *The American Statistician, 47,* 65-72. doi**:**10.1080/00031305.1993.10475938.

Wild, D., Eremenco,S., Mear,I., Martin, M., Houchin, C., Gawlicki, M., … Molsen, E. (2009). Multinational Trials—Recommendations on the Translations Required, Approaches to Using the Same Language in Different Countries, and the Approaches to Support Pooling the Data: The ISPOR Patient-Reported Outcomes Translation and Linguistic Validation Good Research Practices Task Force Report. *Value in Health, 12*, 430-440. doi**:**10.1111/j.1524-4733.2008.00471.x

Wijsman, D., Hekster,Y., Keyser, A., Renier, W., & Meinardi, H. (1991). Clinimetrics and epilepsy care. Pharmaceutisch Weekblad (Scientific) *13*, 182-188. doi:org/10.1007/BF01957744