

**Instrumentos Psicométricos Baseados em Evidências: *Standards*, Evidências de Validade,
Desenvolvimento e Adaptação**

Jorge Sinval^{1,2,3,4}

¹ Department of Evidence-Based Health, Paulista School of Medicine. Federal University of São Paulo,
São Paulo, Brazil.

² Business Research Unit (BRU-IUL), Instituto Universitário de Lisboa (Iscte-IUL), Lisbon, Portugal.

³ Faculty of Philosophy, Sciences and Languages of Ribeirão Preto, University of São Paulo, Ribeirão
Preto, São Paulo, Brazil.

⁴ National Institute of Education, Nanyang Technological University, Singapore, Singapore.

Nota do Autor

A correspondência relativa a este capítulo deverá ser endereçada a Jorge Sinval. Email:
jorgesinval@gmail.com

Introdução

A importância do uso de instrumentos psicométricos de qualidade na psicologia baseada em evidências reside na necessidade de garantir um grau de confiança maior nos dados recolhidos e interpretados (Flake et al., 2017). Os instrumentos psicométricos são uma ferramenta importante na prática profissional tanto de psicólogos como de outros profissionais (Furr, 2020), nomeadamente na área da pesquisa. Instrumentos psicométricos bem desenvolvidos e robustos (i.e., com boas evidências de validade) são essenciais para a obtenção de dados precisos e relevantes, que servirão como base para o diagnóstico, a intervenção e a avaliação das práticas psicológicas. Ao utilizar instrumentos psicométricos com boas propriedades de medida, os profissionais de psicologia podem tomar decisões mais informadas e baseadas em melhores evidências, contribuindo assim para um potencial aumento da qualidade dos serviços prestados e possível promoção do bem-estar dos indivíduos e comunidades.

Além disso, a utilização de instrumentos psicométricos adequados é crucial para a pesquisa em psicologia, desempenhando um papel fundamental na reprodutibilidade e replicabilidade dos resultados das pesquisas e na confiança que os pesquisadores podem ter nos seus resultados. Com instrumentos psicométricos apropriados, é possível — entre outros — realizar comparações de scores entre grupos e selecionar itens mais adequados de acordo com o nível de aptidão ou atitude que se deseja discriminar. Estes tipos de informações conduzem a melhores evidências e a uma maior qualidade naquilo no processo de mensuração e análise, o que contribui para o avanço do conhecimento científico na área. Ao utilizar instrumentos com boas propriedades psicométricas, os pesquisadores podem aumentar a qualidade dos resultados obtidos, o que poderá potenciar a identificação de padrões, a generalização de descobertas e a construção de teorias mais sólidas e fundamentadas.

Por fim, a adoção de instrumentos psicométricos de qualidade na psicologia baseada em evidências também contribui para o fortalecimento da ética profissional e da confiança pública na área. Ao utilizar instrumentos psicométricos com evidências de validade apropriadas, os profissionais de

psicologia demonstram seu compromisso com a integridade científica e a responsabilidade perante os indivíduos e comunidades atendidas. O uso de instrumentos psicométricos adequados garante que os processos de avaliação e intervenção sejam realizados com rigor científico e respeito às necessidades e direitos dos indivíduos, promovendo a justiça e a equidade na prática psicológica (American Educational Research Association et al., 2014).

Os instrumentos são geralmente publicados inicialmente numa primeira língua, normalmente o inglês (Elosua & Iliescu, 2012) e posteriormente, quando se revelam ferramentas úteis e com evidências de validade, são adaptados transculturalmente a outros contextos (países e/ou culturas). Para este efeito, existem várias orientações (Beaton et al., 2000; Gudmundsson, 2009; Hambleton, 1996; Hambleton et al., 2005; He & van de Vijver, 2012; Ohrbach et al., 2013; van de Vijver & Hambleton, 1996) as definidas pela *International Test Commission* (ITC) merecem uma atenção especial (International Test Commission, 2018). Poortinga (1989) alertou para a necessária cautela na comparação de constructos e resultados obtidos em diferentes versões linguísticas de um mesmo instrumento psicométrico. Outros autores acrescentam preocupações sobre o tipo de análises que podem ser realizadas com dados transculturais (van de Vijver, 2016; van de Vijver & Fischer, 2009). Para além da necessidade de versões multilinguísticas dos mesmos instrumentos, há também a consciência de que tais esforços devem ser desenvolvidos com rigor metodológico. Este capítulo começa por descrever a utilização dos *standards* para os instrumentos psicológicos, discutindo o desenvolvimento de novos instrumentos e, em seguida, abordando a adaptação de instrumentos.

Avaliação Psicológica

A primeira avaliação psicológica extensiva ocorreu na China há mais de 3000 anos, medindo a criatividade, a percepção espaço visual, o pensamento divergente e a resolução de problemas (Oakland, 2004). Desde então e até à atualidade, a avaliação psicológica tem conquistado o seu espaço em diversos contextos. A avaliação psicológica é um constituinte tão importante da prática psicológica que a

sua utilidade é, sem dúvida, essencial (Hunsley & Mash, 2008). A avaliação é vista por muitos psicólogos profissionais como a característica única e definidora da sua especialidade (Diniz et al., 2007; Krishnamurthy et al., 2004). A capacidade de efetuar uma avaliação psicológica é vista como uma competência essencial (Hunsley & Mash, 2008). Tanto no exercício da sua profissão como em pesquisa, os psicólogos recorrem a diversas técnicas e tecnologias, embora as mais utilizadas sejam, sem dúvida, os questionários, índices, inventários, escalas e testes (Muñiz & Bartram, 2007); muitos são os termos utilizados para designar os instrumentos psicométricos (Cohen et al., 2022; Streiner, 2003). São a expressão mais impactante e conhecida da psicometria na sociedade, tendo influenciado a tomada de decisões em contextos educativos, sociais, legais e clínicos (Elosua & Iliescu, 2012). A atenção a questões pragmáticas e conceptuais na medição tem sido a pedra angular da ciência psicológica (Hunsley & Mash, 2008). Os psicólogos devem fundamentar as suas opções nos instrumentos que utilizam, particularmente quando se tratam de testes psicológicos, e cujas consequências podem ser muito mais profundas na vida dos inquiridos (Hutz et al., 2016). Para se aprofundar em tais actividades de avaliação, é essencial o conhecimento dos procedimentos e dos princípios psicométricos (Hunsley et al., 2015). A testagem, a utilização dos testes e a avaliação continuam a crescer a nível mundial (Zumbo, 2014) algo que torna importante o incentivo a utilização ética dos testes e apresentação de directrizes para poder avaliar a qualidade das práticas de testagem (Plake & Wise, 2014).

Para utilizar corretamente os testes psicológicos e os questionários, estes instrumentos de medida devem fornecer propriedades psicométricas adequadas, tais como uma dimensionalidade, a fiabilidade dos resultados e a invariância da medida. Além disso, os profissionais devem possuir competência suficiente para os utilizar (Muñiz et al., 2012).

Como qualquer outra técnica, os testes podem ter uma boa ou má utilização, dependendo do profissional que os emprega e do contexto em que são aplicados. A utilização de normas e directrizes que foram originalmente desenvolvidas para outro país pode ser questionável (Zumbo, 2014) embora seja,

sem dúvida, melhor do que não utilizar quaisquer diretrizes ou *standards*. Infelizmente, a nível internacional, são poucos os códigos de ética que abordam a utilização correta dos testes e da avaliação (Oakland & Iliescu, 2016). No entanto, os testes são utilizados em praticamente todos os países e em muitos grupos diferentes (Laher, 2022) embora a sua disponibilidade e utilização difiram consideravelmente entre os diferentes países (Oakland, 2004). A prática psicométrica nem sempre acompanha os avanços da teoria psicométrica, e é isso que acontece frequentemente a nível europeu (Elosua & Iliescu, 2012). As adaptações de testes de outras línguas (principalmente do inglês) são muito comuns. Mesmo testes desenvolvidos recentemente podem ter evidências de validade insuficientes nos seus manuais (Gonçalves et al., 2017). Para melhorá-los, mais conhecimento de psicometria deverá ser desenvolvido e aplicado, e mais estudo mais abrangentes sobre as adaptações a contextos específicos deverão ser produzidos (Muñiz & Bartram, 2007).

No que diz respeito ao Brasil, a pesquisa em avaliação psicológica vem ganhando cada vez mais atenção de profissionais e organizações (Machado & Morona, 2007; Pereira & Bandeira, 2009) e o interesse pelo estudo das propriedades psicométricas dos instrumentos tem sido revitalizado (Vieira & Campos, 2011) tendo atingido um nível de maturidade considerável (Wechsler et al., 2019). A sua utilização está a ganhar um estatuto crescente entre os profissionais, sobretudo graças ao número de estudos recentes e à maior facilidade de acesso aos mesmos (Chiodi & Wechsler, 2008; Wechsler et al., 2019). O Brasil foi pioneiro na implantação de um sistema certificado com base em critérios internacionais de qualidade dos testes (American Educational Research Association et al., 2014) com o objetivo de agrupar/classificar todos os instrumentos existentes em um país (Primi, 2010). O *Sistema de Avaliação de Testes Psicológicos* (SATEPSI) do Conselho Federal de Psicologia (Conselho Federal de Psicologia, 2010) criado em 2001, é um marco na prática profissional da Psicologia (Mansur-Alves et al., 2016; Primi, 2010).. A criação do Instituto Brasileiro de Avaliação Psicológica (IBAP), o apoio do Conselho Federal de Psicologia para a organização de comissões de avaliação psicológica e o crescente número de

grupos científicos (espalhados por diferentes estados) dedicados à criação/adaptação e pesquisa de instrumentos psicométricos mostram um claro desenvolvimento da avaliação psicológica no Brasil (Wechsler et al., 2019). Duas outras associações merecem destaque: a Sociedade de Avaliação Psicológica de Minas Gerais (SAPSI-MG) e a Associação Brasileira de Rorschach e Métodos Projetivos (ASBRo) (Wechsler et al., 2022). A maioria dos estudos de adaptação transcultural de instrumentos de avaliação psicológica no Brasil segue diretrizes internacionais e nacionais (Cassepp-Borges et al., 2010; Manzi-Oliveira et al., 2011). Dito isso, algumas preocupações devem ser levantadas, alguns profissionais ainda utilizam instrumentos que não foram considerados adequados pelo Conselho Federal de Psicologia (Baumgartl et al., 2010). Infelizmente, há testes muito utilizados que não cumprem os critérios de evidências de validade adequadas (Baumgartl & Primi, 2006) porém tem-se registado uma melhoria nos últimos anos (Wechsler et al., 2019). É consensual a necessidade de mais estudos sobre as evidências de validade dos instrumentos utilizando uma taxonomia comum (Baumgartl & Primi, 2006; Manzi-Oliveira et al., 2011; Primi et al., 2009). Sempre que possível, esses estudos devem ser realizados em conjunto com estudos regionais para produzir a padronização necessária para cada território. O treinamento/formação em avaliação psicológica deve continuar a melhorar (Chiodi & Wechsler, 2008; Mäder, 2016; Nunes et al., 2012) como tem acontecido nas últimas duas décadas, com o auxílio do crescente número de programas de pós-graduação e linhas de pesquisa dedicadas ao tema da avaliação psicológica. Apesar dos desenvolvimentos existentes, podem dar-se melhorias através do desenvolvimento/adaptação de instrumentos para sujeitos com necessidades especiais e do aumento do uso de testes adaptativos computadorizados (Wechsler et al., 2019). Nunca é demais lembrar que a avaliação psicológica deve ser realizada por um profissional de psicologia devidamente reconhecido (Primi, 2010) algo que nem sempre acontece. Entre os países da América do Sul, o Brasil parece aquele que tem a organização mais robusta do movimento de desenvolvimento de testes (Wechsler et al., 2022).

Em Portugal, a avaliação psicológica está a ter uma evolução promissora, uma vez que o número de estudos que envolvem a avaliação das propriedades psicométricas dos instrumentos está a aumentar (Almeida et al., 2013). Existem mais instrumentos do que nunca, que podem ser utilizados pelos profissionais e estudados pelos pesquisadores. Este é um progresso bem-vindo, embora se deva ser cauteloso na seleção de instrumentos para uso profissional, uma vez que a utilização de instrumentos sem as devidas evidências de validade para um determinado contexto representa uma prática profissional insatisfatória (Hunsley et al., 2015). Algumas opiniões têm sido expressas, nomeadamente a Ordem dos Psicólogos Portugueses (2015) refere que seria útil ter uma definição de instrumentos de avaliação psicológica que distinguisse os que são exclusivos da psicologia, dos que são transversais a outras áreas científicas. Tal preocupação foi também demonstrada por um inquérito recente a psicólogos portugueses, que indicou que a utilização de testes psicológicos deveria ser da responsabilidade exclusiva dos psicólogos (Seabra-Santos et al., 2019). Além disso, foi registada a necessidade de treinamento/formação adequada para a utilização de instrumentos psicológicos (Diniz et al., 2007; Seabra-Santos et al., 2019) bem como a consciência dos limites, fragilidades e pontos fortes dos instrumentos (Simões, 2005). Além das evidências de validade nem sempre terem sido avaliadas como deveriam, há casos em que o mesmo instrumento está a ser estudado de forma autónoma por diferentes autores (Gonçalves et al., 2017) sem que um tenha conhecimento do outro. Outro problema que ocorre é a utilização de diferentes versões dos instrumentos (i.e., utilização de itens com diferentes traduções para a mesma população) e o desconhecimento dos *standards* e diretrizes internacionais relativas à aplicação de testes (Diniz et al., 2007; Muñiz et al., 1999). Esta fragmentação pode tornar alguns destes esforços menos eficazes do que poderiam ser. Há a necessidade de mais estudos de adaptação que avaliem corretamente as propriedades psicométricas dos instrumentos (Almeida et al., 2006) sobretudo no caso dos testes internacionais e dos testes destinados a grupos minoritários, bem como a utilização de instrumentos com evidências de validade insuficientes (Almeida et al., 2013). O

desenvolvimento de legislação e *standards* na avaliação psicológica, juntamente com mais formação/treinamento e supervisão pós-graduada, foram algumas das necessidades atuais identificadas pelos psicólogos portugueses (Seabra-Santos et al., 2019). No entanto, apesar de partir de uma posição de fragilidade e atraso em termos de avaliação psicológica, Portugal tem sido capaz de adotar algumas das mais modernas tendências de avaliação psicológica, alcançando posições preeminentes em alguns tópicos (Brand, 2022). Todas as avaliações — independentemente de se destinarem à investigação, à prática ou à tomada de decisões com impacto na vida dos sujeitos — deverão existir evidências de validade para o fim a que se destinam; este é o princípio mais importante do panorama psicométrico (Campbell, 2013). Para utilizar um teste ou um questionário, este deve apresentar evidências de validade adequadas. A avaliação das evidências de validade de um instrumento psicométrico consiste num processo de acumulação de evidências relevantes que é demonstrativo do quão apropriada e adequadamente o instrumento satisfaz os objetivos para os quais foi desenvolvido (Sireci & Padilla, 2014). Assim, "são as interpretações dos resultados dos testes para as utilizações propostas que são avaliadas, e não o teste em si" (American Educational Research Association et al., 2014, p. 11).

Standards para Testes Educacionais e Psicológicos

Existem vários *standards* e directrizes para avaliar as propriedades de mensuração dos instrumentos psicométricos, que são adoptadas por várias organizações e associações profissionais importantes (Chan, 2014). As *Directrizes da ITC para a Tradução e Adaptação de Testes* (International Test Commission, 2018) a *checklist* COSMIN (Gagnier et al., 2021; Mokkink et al., 2016) os *Princípios para a Validação e Utilização de Procedimentos de Seleção de Pessoal* da SIOP (Society for Industrial and Organizational Psychology, 2018) a *Avaliação da Medição de Desfechos Relatados pelo Paciente* (Valderas et al., 2008) e o *Modelo de Revisão da Federação Europeia de Associações de Psicólogos* (Evers et al., 2013) são exemplos. As *Standards para os Testes Educacionais e Psicológicos* (American Educational Research Association et al., 2014) é a referência mais amplamente aceite e utilizada em

termos de *standards* e diretrizes para os testes psicológicos, sendo o resultado do consenso de três grandes organizações poderosas (Zumbo, 2014): a *Associação Americana de Pesquisa Educacional* (AERA; American Educational Research Association), a *Associação Americana de Psicologia* (APA; American Psychological Association) e o *Conselho Nacional de Mensuração na Educação* (NCME; National Council on Measurement in Education). Hunsley et al. (2015) exortam os psicólogos a utilizar a edição mais recente dos *Standards para Testes Educacionais e Psicológicos* (American Educational Research Association et al., 2014) como referência, uma vez que apresenta *standards* para uma avaliação cientificamente fundamentada. Apesar de algumas críticas (Zumbo, 2014) relativamente ao seu custo — o que dificultou a sua divulgação — e ao seu enfoque no contexto estado-unidense, continua a ser a referência mais conhecida em termos de diretrizes de avaliação de evidências de validade. No entanto, recentemente, os *Standards para Testes Educacionais e Psicológicos* de 2014 foram disponibilizadas gratuitamente em inglês e espanhol para todos os interessados (ver www.testingstandards.net), ajudando a disseminar as melhores práticas em termos de testes.

Os *Standards* fornecem critérios para avaliar as práticas de testagem, os testes e os efeitos da sua utilização (Sireci & Padilla, 2014). Além disso, fornecem informações aos editores, aos utilizadores e aos criadores de testes; informações que estão relacionadas com elementos essenciais de um programa de testagem que deve informar o desenvolvimento, a seleção e a utilização dos testes (Plake & Wise, 2014). No entanto, Zumbo (2014) observa que, uma vez que os autores dos *Standards* provêm de três organizações estado-unidenses que não são, por definição ou conceção, "organizações internacionais," pelo que se deve ter cuidado ao adotar os *Standards* a nível global. Dito isto, os *Standards* desempenham um papel fundamental na comunidade de avaliação a nível mundial para o propósito de desenvolvimento/adaptação de testes e questionários (Zumbo, 2014) e são igualmente citadas em revistas estado-unidenses e internacionais (Zumbo & Chan, 2014).

Desde a sua publicação original (American Psychological Association, 1954) passaram-se mais de 60 anos e ocorreram várias alterações, resultando na mais recente e 5ª revisão (American Educational Research Association et al., 2014). Uma das evoluções mais importantes do conteúdo dos *Standards* foi a definição de validade. A definição de validade não é consensual, mas é considerada por muitos como o critério mais importante para avaliar os resultados dos testes (Royal, 2017). A definição de validade passou por um processo de reformulação desde a década de 1970, acabando por conduzir à definição apresentada na versão de 1999 dos *Standards*, que ainda hoje se mantém (Shepard, 2016). Esta evolução transformou a validade num conceito mais vasto e bastante complexo (Wolming & Wikström, 2010). A discussão em torno da validade foi objeto de vários contributos importantes, como os de Messick (1989) que propôs a validade como um conceito unitário, ou a proposta de Kane (2006) que — através de uma abordagem argumentativa — enfatizou a viabilidade prática. Apesar de existirem definições complexas, também foram sugeridas definições mais simples. É o caso da definição de Nunnally (1973) que propõe que a validade existe se um instrumento de medida fizer aquilo para que foi concebido. Se um instrumento cumpre bem a função para que foi concebido, tem validade (Nunnally, 1972). Nos *Standards* atuais, a validade é um conceito unitário e é definida como "o grau em que as evidências e a teoria apoiam as interpretações dos resultados dos testes para a utilização prevista dos testes" (American Educational Research Association et al., 2014, p. 11).

A abordagem atual dos *Standards* consiste em construir um argumento de validade baseado em cinco tipos diferentes de fontes de validade (American Educational Research Association et al., 2014): conteúdo do teste (Sireci & Faulkner-Bond, 2014) processos de resposta (Padilla & Benítez, 2014) estrutura interna (Rios & Wells, 2014) relações com outras variáveis (Oren et al., 2014), e consequências do teste (Lane, 2014).

As evidências de validade baseadas no conteúdo do teste, designada "validade de conteúdo" nas versões anteriores dos *Standards*, diz respeito à relação entre o construto a que o teste se destina a

medir e o conteúdo do próprio teste (American Educational Research Association et al., 2014). O conteúdo de um teste deve representar o construto proposto e deve ser adequado aos objetivos do teste (Sireci & Faulkner-Bond, 2014). A avaliação deste tipo de evidências pode incluir análises que demonstrem se o conteúdo do teste está a representar o domínio do conteúdo e a relevância do domínio do conteúdo em relação à interpretação proposta para os resultados do teste (American Educational Research Association et al., 2014). Sireci (1998) apresentou um quadro com quatro elementos para avaliar a evidências de validade baseadas no conteúdo do teste: definição do domínio, representação do domínio, relevância do domínio e adequação dos procedimentos de construção do teste. A validade de conteúdo do teste é uma questão central, não podendo ser ignorada na concepção do teste, sob pena de criar um problema quando se está na busca de uma medida critério (Ebel, 1956). Esta fonte de evidências de validade pode ser avaliada de várias formas: a representatividade dos itens em relação ao domínio (e.g., utilizando o Rácio de Validade de Conteúdo ou o Índice de Validade de Conteúdo), a relação lógica/empírica do conteúdo do teste com o domínio de desempenho, a qualidade dos itens do teste, a análise de sensibilidade e outras (Downing, 2003).

As evidências de validade baseadas no processo de resposta podem fornecer pistas sobre a relação entre a natureza da resposta ou o desempenho do examinando e o construto (American Educational Research Association et al., 2014). Tratam-se de evidências sobre a integridade dos dados, no sentido de que o erro associado à aplicação do teste é controlado ou eliminado o mais possível (Downing, 2003). Para obter evidências de validade sobre os processos de resposta, é possível questionar os participantes no teste sobre as suas estratégias de resposta, registar o desempenho numa tarefa de escrita, os tempos de resposta, a familiaridade com o formato, a validação da matriz de *scores* preliminares (no caso de um teste de aptidões), a exatidão na combinação de diferentes formatos de pontuação, descrições e interpretações compreensíveis/precisas dos *scores* (Downing, 2003). A relação entre este tipo de evidências de validade e as que se baseiam no conteúdo do teste é coerente com a

visão unitária da validade (Padilla & Benítez, 2014). No entanto, esta tipo de evidências validade é menos representada do que outras fontes de evidências de validade (Beckman et al., 2005).

As evidências de validade baseadas na estrutura interna são descritas como o nível a que as dimensões e itens do teste estão em conformidade com o construto proposto (American Educational Research Association et al., 2014). Pode ser dividida em três componentes básicas: dimensionalidade, confiabilidade dos *scores* e invariância da medida (Rios & Wells, 2014). A dimensionalidade centra-se nas inter-relações entre os itens que sustentam os *scores* pretendidos para o teste. A invariância da medida é útil para testar se a estrutura dos itens e dos instrumentos é idêntica e comparável em diferentes grupos (Davidov et al., 2014).. Por último, a confiabilidade indica em que medida os *scores* são consistentes em utilizações repetidas. Para obter evidências de validade para a dimensionalidade é possível utilizar a Análise Fatorial Confirmatória (AFC; Brown, 2015) e os modelos *bifactor* (Holzinger & Swineford, 1937). No que diz respeito à invariância de medida, pode-se utilizar a análise fatorial confirmatória de multigrupos (Byrne, 2016; Millsap, 2011) e o funcionamento diferencial dos itens (Penfield & Camilli, 2006). Relativamente à confiabilidade dos *scores*, é possível estimar a consistência interna, as formas paralelas, a concordância interavaliadores e a confiabilidade teste-reteste. Embora habitualmente a estimativa mais utilizada da consistência interna seja o coeficiente α (Cronbach, 1951) este coeficiente não é o mais adequado, a menos que os itens sejam *tau*-equivalentes. Foram propostas várias outras alternativas, como o coeficiente de ρ de Jöreskog (1971), a confiabilidade composta/compósita (CR) de Fornell e Larcker (1981); o ρ de Dillon e Goldstein (1984), o β de Revelle (1979), o ω de McDonald (1999), ou ω de Raykov (2001); alguns deles são também úteis quando os erros de medição (resíduos dos itens) estão correlacionados, uma vez que o coeficiente α subestimar a fiabilidade.

A fonte de evidências de validade baseadas nas relações com outras variáveis advém da análise da relação dos *scores* instrumento selecionado com variáveis externas, especialmente as relações teste-

critério (McCoach et al., 2013). É definida como uma importante fonte de evidências, uma vez que se espera que a utilização de uma determinada medida esteja relacionada com algumas outras variáveis externas (American Educational Research Association et al., 2014). Um exemplo é a utilização de notas universitárias em relação à validade do sistema de seleção (Oren et al., 2014). As relações entre os *scores* dos instrumentos medidos e outras variáveis podem fornecer evidência convergente (relações entre os mesmos construtos ou construtos semelhantes) e/ou discriminante (relações entre os *scores* dos testes e as medidas de construtos supostamente diferentes) (Beckman et al., 2005). Se existir interesse em relacionar os *scores* do teste com uma variável critério, os profissionais/investigadores podem optar por duas abordagens para obter este tipo de informação: preditiva (relação entre os *scores* do teste e os *scores* critério que são obtidos mais tarde) ou concorrente (relação entre os *scores* do teste e os *scores* critério obtidos sensivelmente ao mesmo tempo). A variável de critério é uma medida de um resultado ou de um atributo que é operacionalmente diferente daquele que é medido pelo teste (American Educational Research Association et al., 2014). Assim, os resultados do teste funcionam como um preditor hipotético desse critério-alvo. Isto refere-se às formas tradicionais de evidências de validade relacionadas com o critério (Sireci & Parker, 2006) e inclui conceitos de validade convergente e discriminante, que estavam contidos na validade de constructo em versões anteriores dos *Standards* (McCoach et al., 2013). Ajuda a elucidar o que é o construto, esclarecendo a *rede nomológica* (Cronbach & Meehl, 1955). As evidências de validade para este tipo de fonte podem ser obtidas através da correlação com outras variáveis relevantes, correlações com testes que mensurem construtos semelhantes ou iguais, correlações com medidas dissemelhantes; correlações teste-critério, ou generalização da evidência — isto é — validade externa (Downing, 2003).

Por último, as evidências de validade baseadas nas consequências da testagem são uma fonte de evidências que pretende avaliar os efeitos da avaliação para apoiar ou contestar a validade das interpretações dos *scores* (Beckman et al., 2005). Os testes podem ter um efeito desejado sobre os

inquiridos e algumas consequências podem resultar diretamente da interpretação pretendida dos *scores* — como a melhoria do desempenho num domínio específico — mas também podem conduzir a efeitos indesejados (American Educational Research Association et al., 2014). Esta fonte de evidências tenta avaliar se os efeitos positivos pretendidos estão a ser alcançados e se estão a ser produzidos alguns efeitos negativos (Lane, 2014). Este tipo de evidências de validade tem especial utilidade para programas de avaliação e responsabilização que podem influenciar políticas em programas educativos, como por exemplo o PISA — *Program for International Student Assessment* (Marôco et al., 2016). Alguns exemplos de como obter este tipo de evidências incluem: impacto dos *scores*/resultados dos testes nos alunos ou na sociedade, consequências na aprendizagem, falsos positivos, falsos negativos, razoabilidade do método de estabelecimento do score de aprovação/reprovação (corte), confiabilidade da decisão de aprovação/reprovação — precisão da classificação; erro padrão condicional da medição do score de aprovação (Downing, 2003).

Para aceder corretamente aos diferentes tipos de evidências de validade, devem ser utilizadas técnicas estatísticas adequadas (Sireci et al., 2006). Foi desenvolvido um grupo sofisticado de princípios e métodos (psicometria) para construir e avaliar testes (Andriola, 2009; Hutz et al., 2015; Pasquali, 2011; Wood et al., 2007). O rigor científico e a utilidade clínica da psicometria levaram-na a ser considerada uma das tradições nobres (Meehl, 1997). Nos últimos anos, a psicometria sofreu melhorias substanciais e, hoje em dia, oferece contributos relevantes, pelo que já é tempo de reduzir o fosso entre a teoria psicométrica e a prática psicométrica utilizada pelos profissionais/pesquisadores (tanto pelos que elaboram os testes como pelos que os utilizam) (Elosua & Iliescu, 2012). Isto permitirá a contínua procura por evidências sólidas que sustentem a utilidade dos instrumentos utilizados na avaliação psicológica (Hunsley & Mash, 2008). Várias referências fornecem perspectivas e informações adicionais sobre estas questões, tanto em português (Damásio & Borsa, 2017; Faiad et al., 2021; Hutz et al., 2015) como em inglês (Downing & Haladyna, 2006; Irwing et al., 2018; Lane et al., 2016; Leong et al., 2016;

Mair, 2018; McDonald, 1999; Nunnally & Bernstein, 1994; Rust et al., 2021; Schweizer & DiStefano, 2016).

Desenvolver um Novo Instrumento ou Adaptar um Instrumento Existente?

Antes de se enveredar no desenvolvimento de um novo instrumento psicométrico seja para pesquisa ou a prática, é aconselhável procurar por um instrumento que já tenha sido desenvolvido e que preencha os quesitos desejados (Gall et al., 2014). A escolha apropriada do instrumento de medida é essencial, pois estes possuem um valor importante na tomada de decisões sobre indivíduos, organizações e instituições. Portanto, é crucial que a escolha dos instrumentos seja meticulosa. Há várias fontes de informação que podem ser consultadas (e.g., *Mental Measurements Yearbook*) e que devem fazer parte do processo de avaliação antes de adotar um instrumento.

A escolha do instrumento deve ter em conta os indivíduos que utilizarão o instrumento e que interpretarão os resultados do instrumento. Para isso dever-se-á examinar e avaliar cuidadosamente o instrumento antes da sua adoção. Nitko (2015) sugere uma abordagem sistemática para realizar que se realize uma revisão e avaliação rigorosa dos instrumentos. Uma vez que não há instrumentos que atendam perfeitamente às necessidades específicas do pesquisador/profissional, a estratégia deverá passar pela comparação das vantagens de um instrumento em relação a outros, sendo este um passo vital na seleção da opção mais favorável (Johnson & Morgan, 2016). A estratégia de revisão e avaliação sugerida por Nitko (2015) para a escolha de instrumentos contempla seis passos:

1. Esclarecer o propósito: Definir os objetivos e as necessidades específicas para as quais o instrumento será usado, considerando a população-alvo e os resultados pretendidos.

2. Usar o contexto local como base para avaliar um novo instrumento: Considerar a forma como o instrumento será utilizado no contexto específico e identificar as necessidades, desafios e recursos disponíveis.

3. Estudar análises profissionais dos materiais do instrumento: Examinar as análises de fontes de informação sobre instrumentos psicométricos como o *Mental Measurements Yearbook* ou o website *Test Collection* (<https://www.ets.org/test-collection.html>) da Educational Testing Service, artigos científicos, manuais técnicos e guias de utilizador. Esta etapa tem em conta o Standard 9.2 dos *Standards for Educational and Psychological Testing* (American Educational Research Association et al., 2014). Este Standard menciona que antes da adoção e uso de um instrumento, o pesquisador/profissional que utilizará o instrumento deverá estudar e avaliar os materiais fornecidos pelo criador do instrumento. São de particular importância aqueles que resumem os propósitos do instrumento psicométrico, especificam os procedimentos de administração do instrumento, definem as populações-alvo a quem o instrumento pode ser aplicado e discutem as interpretações de scores para as quais as evidências de validade estão disponíveis.

4. Obter cópias do instrumento sob avaliação: Rever os materiais do instrumento, incluindo itens de teste, manuais de administração e pontuação e quaisquer materiais suplementares.

5. Resumir os pontos fortes e fracos do instrumento para o contexto específico em causa: Com base em na revisão e avaliação, identificar os pontos fortes e fracos do instrumento em relação ao contexto específico e às necessidades.

6. Chegar a uma conclusão sobre o instrumento e sustentar essa conclusão: Tomar uma decisão final sobre a adequabilidade do instrumento para os propósitos desejados e preparar a justificação da escolha com base nas evidências recolhidas durante o processo de avaliação.

Johnson e Morgan (2016) fornecem um guia de avaliação para documentar de forma estruturada a revisão de instrumentos aqui descrita (ver pp. 22-23). No caso das evidências encontradas sobre os instrumentos existentes ser insatisfatória, dever-se-á então ponderar o complexo processo de criação de um novo instrumento psicométrico (Gall et al., 2014).

Desenvolvimento de Novos Instrumentos Psicométricos

Quando existe a necessidade de medir um construto latente com uma abordagem quantitativa, o pesquisador deve utilizar instrumentos psicométricos. Para tal, existe a opção de criar um instrumento novo ou adaptar um instrumento já existente, levantando questões como: qual das opções vale mais a pena? Quais são as vantagens e desvantagens de cada procedimento? Em que situações é que cada opção é mais pertinente?

Ao criar um instrumento, deve haver uma razão subjacente suficientemente forte para justificar a construção de um novo instrumento (Moreau & Wiebels, 2022). Por exemplo, quando o construto ou construtos que serão abordados não estão contemplados num instrumento psicométrico existente, é pertinente que se desenvolva um novo instrumento. Ou quando se reconhece que é improvável uma adaptação de um determinado instrumento a outro contexto seja bem-sucedida, pode ser adequado construir um novo instrumento. Este processo pode ser diferente tanto ao nível da análise estatística, como ao nível dos procedimentos de recolha de dados em relação à adaptação do instrumento. Existem várias diretrizes para desenvolver novos instrumentos psicométricos (Bandalos, 2018; Boateng et al., 2018; Clark & Watson, 2019; Cohen et al., 2022; Cooper, 2023; Irwing & Hughes, 2018; Lane et al., 2016; Muñiz & Fonseca-Pedrero, 2019; Wilson, 2023). DeVellis e Thorpe (2021) apresentaram uma proposta para o desenvolvimento de novas escalas em nove etapas:

Etapa I: É necessário determinar claramente o que se quer medir — os custos serão muito maiores se apenas se fizer este passo depois de gerar os itens e recolher os dados. Assim, este deve ser o primeiro passo. Este processo inclui, entre outras questões, ser-se sensível a quão específica será a medida, à ênfase que deve ser dada a um fenómeno e não a outros, e se deve basear-se na teoria existente ou abrir novas propostas teóricas.

Etapa II: Gerar um conjunto de itens — é necessário criar um grupo extenso de itens que sejam possíveis candidatos para a inclusão na escala (Bejar, 2010; Morrison & Embretson, 2018; Pacico, 2015). Não existe um número específico de itens iniciais ou mesmo uma relação ideal entre o número inicial e

o número final de itens no desenvolvimento da escala. No entanto, alguns autores sugerem a criação de um mínimo inicial de itens que seja o dobro do número de itens considerado para a versão final do instrumento (Muñiz & Fonseca-Pedrero, 2019). Sendo também necessário ter sensibilidade para introduzir (ou não) itens formulados de forma negativa ou positiva - ainda que esta opção não seja consensual (Salazar, 2015; Suárez-Alvarez et al., 2018).

Etapa III: Determinar o formato da escala de resposta — existem vários tipos de escalas de resposta, e o tipo de itens deve ser compatível com o formato a ser utilizado (Cabooter et al., 2016; Cohen et al., 2022). De um ponto de vista puramente estatístico, as escalas de Likert (Likert, 1932) oferecem vantagens para posterior análise, uma vez que permitem uma métrica específica, simetria, ponto central neutro, mensuração da percepção da concordância e intervalos homogêneos entre os pontos da escala, mas também apresentam algumas desvantagens (Hartley, 2014). A extensão da escala de resposta é um tema em que não existe um consenso claro, no entanto parece que a maioria das recomendações aponta para uma escala de cinco ou sete pontos (Colman et al., 1997; Dawes, 2008; Leung, 2011; Revilla et al., 2014).

Etapa IV: Fazer com que o conjunto inicial de itens seja revisto por peritos — dever-se-á pedir a outros pesquisadores (com *expertise* nos construtos visados) que avaliem a pertinência da inclusão de cada item. O objetivo é maximizar as evidências de validade baseadas no conteúdo do instrumento. Métricas como o Índice de Validade de Conteúdo (IVC; Polit et al., 2007) ou o Rácio de Validade de Conteúdo (RVC; Gilbert & Prion, 2016; Lawshe, 1975) podem ser usadas para esta finalidade.

Etapa V: Entrevista cognitiva — Esta etapa qualitativa tem por objetivo compreender o processo de raciocínio subjacente à interpretação e compreensão dos novos itens por parte dos inquiridos (Willis, 2004). O mecanismo básico consiste em perguntar aos participantes o que entenderam de um item e como conceberam a resposta ao mesmo. Este processo de entrevista cognitiva pode ajudar a identificar eventuais problemas com a redação dos itens ou com a escala de resposta utilizada. Uma estratégia

comum consiste em pedir aos inquiridos que pensem em voz-alta enquanto leem e respondem a cada item.

Etapa VI: Considerar a inclusão de itens de validação — Pode ser possível, e relativamente conveniente, incluir alguns itens adicionais no mesmo questionário com o objetivo de ajudar a refinar a evidência final de validade da escala. Há dois tipos de itens a considerar, os itens que detectam falhas ou problemas na escala, como a desejabilidade social, e os itens que medem constructos relacionados e que, por isso, podem ser utilizados *vis-à-vis* ao desempenho desses itens para servir de evidência de validade concorrente ou divergente. Há também a possibilidade de adicionar itens para detetar outros problemas como o esforço insuficiente na resposta (Huang et al., 2015) ou a aquiscência (Wetzel et al., 2015).

Etapa VII: Aplicar os itens a uma amostra para desenvolvimento — depois de obter o conjunto de itens a aplicar, será necessário aplicar esses itens em conjunto a uma amostra representativa da população pretendida. Normalmente, a análise utilizada na próxima fase do desenvolvimento é a Análise Fatorial Exploratória (AFE), que deve ser utilizada se não existirem fortes indícios sobre a estrutura que o instrumento deve ter (i.e., dimensionalidade). No entanto, se existir uma noção firme *a priori* do número de fatores e de quais fatores estão relacionados com quais fatores, dever-se-á utilizar a Análise Fatorial Confirmatória (AFC) na etapa seguinte (Brown, 2015). Não existe uma orientação clara quanto à dimensão da amostra, mas Nunnally e Bernstein (1994) salientaram que, com uma amostra de mais de 300 sujeitos, já é possível efetuar análises substancialmente robustas. As recomendações, no entanto, variam: alguns autores apontam para cinco sujeitos por item, pelo menos para a AFE (Hair et al., 2009) outros recomendam um mínimo de 100 sujeitos para esta mesma análise (Sapnas & Zeller, 2002). Para a AFC vários autores apresentam diferentes heurísticas, no entanto abordagens mais robustas podem — e devem — ser utilizadas (Kelley & Lai, 2018).

Etapa VIII: Avaliar os itens — este é o momento em que as propriedades distribucionais de cada item são avaliadas. Mesmo com um conjunto generoso de itens, não há garantia de que as variáveis latentes desejadas estejam se manifestem nesses itens. É nesta etapa que se realizam as análises fatoriais, quer seja AFC ou AFE, para que a dimensionalidade seja investigada. Adicionalmente, é possível analisar os itens pelas suas médias, variâncias, correlações entre itens que teoricamente pertencem ao mesmo construto, e pela confiabilidade dos coeficientes dos *scores*, por exemplo utilizando os estimadores α (Cronbach, 1951) ou o ω (McDonald, 1999).

Passo IX: Otimizar a extensão da escala — os instrumentos curtos são menos trabalhosos para os inquiridos, mas as escalas mais longas tendem a apresentar melhor índices de confiabilidade. Assim, estes dois aspetos apresentam uma relação inversa e o pesquisador deve escolher o balanço ideal entre brevidade e confiabilidade. Ou seja, se a confiabilidade da escala for demasiado baixa, a brevidade deixará de ser uma virtude. Existem algumas melhorias que podem ser realizadas do ponto de vista estatístico, como eliminar itens com cargas fatoriais mais baixas, investigar quais os itens que menos contribuem para a confiabilidade da escala, dividir a amostra para testar as modificações impostas.

Adaptação de Instrumentos Psicométricos

Quando se opta por não construir um novo instrumento, a adaptação transcultural pode ser vantajosa, uma vez que já existe uma base teórica e uma estrutura fatorial dos itens para outra cultura/contexto. Caberá ao pesquisador avaliar se essa estrutura se mantém, ou se, por outro lado, as diferenças transculturais exigem que sejam feitas alterações à estrutura original para este novo contexto. Seja qual for a decisão, é sempre necessário tentar compreender até que ponto a teoria inerente à estrutura original dos instrumentos é robusta às diferenças culturais existentes na nova amostra. Mais uma vez, existem vários procedimentos para o efeito (Gudmundsson, 2009; Hambleton et al., 2005; Heggstad et al., 2019) a ITC através da sua segunda edição das *Diretrizes para a Tradução e*

Adaptação de Testes (International Test Commission, 2018) resumiu-os em 18 diretrizes, divididas em seis categorias:

Categoria I: Diretrizes de pré-condição — a primeira categoria é composta por três diretrizes que mencionam as decisões que devem ser tomadas antes de se iniciar o processo de tradução/adaptação propriamente dito: (1) obter autorização do titular da propriedade intelectual; (2) avaliar se o grau de sobreposição da definição e do conteúdo do construto medido pelo instrumento (i.e., conteúdo dos itens) é suficiente para o uso pretendido; e (3) minimizar a influência de quaisquer diferenças linguísticas e culturais que sejam irrelevantes para o uso pretendido.

Categoria II: Diretrizes para desenvolvimento de testes — a segunda categoria centra-se no processo concreto de adaptação de um instrumento psicométrico (cinco diretrizes), nomeadamente: (1) ter em conta as diferenças linguísticas, psicológicas e culturais através da escolha dos especialistas experientes; (2) utilizar orientações de tradução adequadas; (3) obter evidências de que o conteúdo dos itens e as instruções do teste têm um significado congruente para as populações pretendidas; (4) obter evidências de que as categorias de pontuação, o formato dos itens, as escalas de classificação, os modos de administração e as convenções no uso do teste são adequados para as populações pretendidas; e (5) recolher dados-piloto para realizar uma avaliação em pequena escala das evidências de validade (permitindo atempadamente quaisquer revisões que sejam necessárias).

Categoria III: Diretrizes de confirmação — a terceira categoria inclui quatro diretrizes relacionadas com a recolha de evidências para abordar a equivalência e as evidências de validade de um instrumento psicométrico em várias culturas e línguas: (1) seleccionar uma amostra relevante tanto em termos de representatividade como em termos de dimensão; (2) fornecer — evidência estatística — sobre a equivalência do construto, equivalência de método e de itens entre as versões de origem e de destino (e.g., AFC, funcionamento diferencial dos itens); (3) fornecer evidências que apoiem as normas e as evidências de validade do instrumento (American Educational Research Association et al., 2014); (4)

utilizar uma análise de dados adequada e um desenho que permita equiparabilidade (i.e., “test equating”) ao associar scores de diferentes versões.

Categoria IV: Diretrizes de administração — a quarta categoria contém duas diretrizes relativas à preparação e às condições necessárias para a aplicação do instrumento: (1) preparar instruções e materiais de aplicação para reduzir os problemas relacionados com a língua e a cultura provocados por modos de resposta e processos de aplicação que possam afetar a validade das inferências retiradas dos *scores* obtidos; e (2) estabelecer as condições necessárias para a aplicação do instrumento nas populações pretendidas.

Categoria V: Diretrizes de pontuação e interpretação — a quinta categoria contém duas diretrizes relativas às comparações dos *scores* dos grupos: (1) realizar a interpretação das diferenças dos grupos com base em todas as informações relevantes disponíveis; e (2) estabelecer comparações dos *scores* entre populações quando o nível de invariância da medida assim o permitir (Davidov et al., 2018).

Categoria VI: Diretrizes de documentação — a última categoria contém duas diretrizes relacionadas com documentação de apoio do instrumento: (1) fornecer documentação que tenha em conta qualquer alteração à versão original e evidências que apoiem a equiparação (i.e., equivalência) dos *scores* entre populações; e (2) desenvolver documentação para os utilizadores da medida adaptada que apoie as boas práticas no contexto da nova população.

Hernández e colaboradores (2020) desenvolveram uma lista estruturada com critérios avaliativos que auxiliam os pesquisadores/profissionais que estão no processo de adaptação de um instrumento a implementar as recomendações da ITC e serve como modelo para avaliar a qualidade das adaptações de instrumentos. Nesta lista cada um dos critérios avaliativos pode ter uma classificação (i.e., “Não aplicável”, “Não aceitável”, “Aceitável”, “Excelente”). A lista consiste em 29 critérios que abrangem todas as fases da adaptação de instrumentos: pré-condição, desenvolvimento, confirmação, administração, pontuação e interpretação, e documentação. Pode ser tão dispendioso desenvolver um

teste estandardizado a partir do zero como adaptar um teste transculturalmente. No que respeita aos custos monetários e temporais, é necessário confirmar a utilidade das várias fontes de evidências de validade dos instrumentos adaptados, pelo que o pressuposto de que o desenvolvimento de um novo instrumento é mais dispendioso pode não ser totalmente justificado (Gudmundsson, 2009), deixando a decisão ao critério de cada investigador.

Dima (2018) oferece uma boa referência que apresenta uma perspetiva aplicada e condensada das diferentes orientações, apoiada por um tutorial para um protocolo psicométrico de seis etapas. Providencia a análise necessária — através da gratuita linguagem de programação estatística *R* (R Core Team, 2023) — para efetuar uma avaliação mais quantitativa de uma nova medida desenvolvida ou uma adaptação de um instrumento já existente.

As reflexões sobre a adaptação e desenvolvimento de instrumentos psicométricos fornecem contributos úteis não só para os pesquisadores, mas também para os profissionais. Elas podem permitir uma escolha mais abrangente de instrumentos, fundamentada na existência de estudos de adaptação para a população desejada, bem como oferecer evidências de validade, como enfatizado nas diretrizes do ITC (International Test Commission, 2018). Ou uma lacuna for identificada na mensuração de um determinado construto, criar um instrumento. Assim, mais do que saber aplicar um instrumento, é preciso primeiro saber avaliar a qualidade do instrumento para o uso pretendido e refletir sobre a importância teórica de cada passo, evitando procedimentos puramente cegos baseados apenas em números, já que não são raros os estudos de adaptação sem planeamento (Zumbo & Chan, 2014).

Referências

- Almeida, L. S., Araújo, A., & Diniz, A. (2013). Psychological assessment and testing in Portugal. *PSIENCIA Latin American Journal of Psychological Science*, 5(2), 144–149.
<https://doi.org/10.5872/psiencia/5.2.76>
- Almeida, L. S., Diniz, A. M., Pais, L. G., & Guisande, M. A. (2006). A avaliação psicológica na prática dos psicólogos: As provas psicológicas usadas em Portugal. Em C. Machado, L. S. Almeida, M. Gonçalves, & V. Ramalho (Orgs.), *Actas do Congresso Internacional “Avaliação Psicológica: Formas e Contextos”*: Vol. XI (p. 1091–1097). Psiquilíbrios.
- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for Educational and Psychological Testing*. American Psychological Association.
- American Psychological Association. (1954). Technical recommendations for psychological tests and diagnostic techniques. *Psychological Bulletin*, 51(2), 201–238. <https://doi.org/10.1037/h0053479>
- Andriola, W. B. (2009). Psicometria moderna: Características e tendências. Em *Estudos em Avaliação Educacional* (Vol. 20, Número 43, p. 319–340).
- Bandalos, D. L. (2018). *Measurement theory and applications for the social sciences*. The Guilford Press.
- Baumgartl, V. de O., Pagano, A. P., & Lacerda, J. (2010). A utilização de testes psicológicos em organizações de Minas Gerais. *Arquivos Brasileiros de Psicologia*, 62(2), 178–186.
- Baumgartl, V. de O., & Primi, R. (2006). *Contribuições da avaliação psicológica no contexto organizacional: Um estudo com o BPR-5, BFM-1 e o PMK*. Casa do Psicólogo.
- Beaton, D. E., Bombardier, C., Guillemin, F., & Ferraz, M. B. (2000). Guidelines for the process of cross-cultural adaptation of self-report measures. *Spine*, 25(24), 3186–3191.
<https://doi.org/10.1097/00007632-200012150-00014>

- Beckman, T. J., Cook, D. A., & Mandrekar, J. N. (2005). What is the validity evidence for assessments of clinical teaching? *Journal of General Internal Medicine*, 20(12), 1159–1164.
<https://doi.org/10.1111/j.1525-1497.2005.0258.x>
- Bejar, I. I. (2010). Recent development and prospects in item generation. Em S. E. Embretson (Org.), *Measuring psychological constructs: Advances in model-based approaches*. (p. 201–226). American Psychological Association. <https://doi.org/10.1037/12074-009>
- Boateng, G. O., Neilands, T. B., Frongillo, E. A., Melgar-Quíñonez, H. R., & Young, S. L. (2018). Best practices for developing and validating scales for health, social, and behavioral research: A primer. *Frontiers in Public Health*, 6, 1–18. <https://doi.org/10.3389/fpubh.2018.00149>
- Brand, J. P. (2022). The beginnings of psychological assessment in Spain and Portugal. Em S. Laher (Org.), *International histories of psychological assessment* (p. 134–150). Cambridge University Press.
<https://doi.org/10.1017/9781108755078.009>
- Brown, T. A. (2015). *Confirmatory factor analysis for applied research* (2º ed). The Guilford Press.
- Byrne, B. M. (2016). Multigroup comparisons: Testing for measurement, structural, and latent mean equivalence. Em F. T. L. Leong, D. Bartram, F. M. Cheung, K. F. Geisinger, & D. Iliescu (Orgs.), *The ITC international handbook of testing and assessment* (p. 377–394). Oxford University Press.
<https://doi.org/10.1093/med:psych/9780199356942.003.0026>
- Cabooter, E., Weijters, B., Geuens, M., & Vermeir, I. (2016). Scale format effects on response option interpretation and use. *Journal of Business Research*, 69(7), 2574–2584.
<https://doi.org/10.1016/j.jbusres.2015.10.138>
- Campbell, J. P. (2013). Assessment in industrial and organizational psychology: An overview. Em K. F. Geisinger (Org.), *APA handbook of testing and assessment in psychology: Test theory and testing and assessment in industrial and organizational psychology* (Vol. 1, p. 355–395). American Psychological Association. <https://doi.org/10.1037/14047-022>

- Cassepp-Borges, V., Balbinotti, M. A. A., & Teodoro, M. L. M. (2010). Tradução e validação de conteúdo: Uma proposta para a adaptação de instrumentos. Em L. Pasquali (Org.), *Instrumentação psicológica: Fundamentos e práticas* (p. 506–520). Artmed.
- Chan, E. K. H. (2014). Standards and guidelines for validation practices: Development and evaluation of measurement instruments. Em B. D. Zumbo & E. K. H. Chan (Orgs.), *Validity and validation in social, behavioral, and health sciences* (p. 9–24). Springer. https://doi.org/10.1007/978-3-319-07794-9_2
- Chiodi, M. G., & Wechsler, S. M. (2008). Psychological evaluation: Brazilian contributions. *Boletim Academia Paulista de Psicologia*, 28(2), 197–210.
- Clark, L. A., & Watson, D. (2019). Constructing validity: New developments in creating objective measuring instruments. *Psychological Assessment*, 31(12), 1412–1427.
<https://doi.org/10.1037/pas0000626>
- Cohen, R. J., Schneider, W. J., & Tobin, R. M. (2022). *Psychological testing and assessment: An introduction to tests and measurement* (10^o ed). McGraw-Hill.
- Colman, A. M., Norris, C. E., & Preston, C. C. (1997). Comparing rating scales of different lengths: Equivalence of scores from 5-point and 7-point scales. *Psychological Reports*, 80(2), 355–362.
<https://doi.org/10.2466/pr0.1997.80.2.355>
- Conselho Federal de Psicologia. (2010). *Avaliação psicológica: Diretrizes na regulamentação da profissão*. Author.
- Cooper, C. (2023). *An introduction to psychometrics and psychological assessment: Using, interpreting and developing tests* (2^o ed). Routledge. <https://doi.org/10.4324/9781003240181>
- Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, 16(3), 297–334. <https://doi.org/10.1007/BF02310555>
- Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, 52(4), 281–302. <https://doi.org/10.1037/h0040957>

Damásio, B. F., & Borsa, J. C. (Orgs.). (2017). *Manual de desenvolvimento de instrumentos psicológicos*. Vetor.

Davidov, E., Meuleman, B., Cieciuch, J., Schmidt, P., & Billiet, J. (2014). Measurement equivalence in cross-national research. *Annual Review of Sociology*, 40(1), 55–75.
<https://doi.org/10.1146/annurev-soc-071913-043137>

Davidov, E., Schmidt, P., Billiet, J., & Meuleman, B. (Orgs.). (2018). *Cross-cultural analysis: Methods and applications*. Routledge. <https://doi.org/10.4324/9781315537078>

Dawes, J. (2008). Do data characteristics change according to the number of scale points used? An experiment using 5-point, 7-point and 10-point scales. *International Journal of Market Research*, 50(1), 61–104. <https://doi.org/10.1177/147078530805000106>

DeVellis, R. F., & Thorpe, C. T. (2021). *Scale development: Theory and applications* (5^o ed). SAGE.
<https://us.sagepub.com/en-us/nam/scale-development/book269114>

Dillon, W. R., & Goldstein, M. (1984). *Multivariate analysis: Methods and applications*. John Wiley & Sons.

Dima, A. L. (2018). Scale validation in applied health research: Tutorial for a 6-step R-based psychometrics protocol. *Health Psychology and Behavioral Medicine*, 6(1), 136–161.
<https://doi.org/10.1080/21642850.2018.1472602>

Diniz, A. M., Almeida, L. S., & Pais, L. G. (2007). Contextos profissionais e práticas da avaliação psicológica: Inquérito aos psicólogos portugueses. *Psico-USF*, 12(1), 1–12.

Downing, S. M. (2003). Validity: On the meaningful interpretation of assessment data. *Medical Education*, 37(9), 830–837. <https://doi.org/10.1046/j.1365-2923.2003.01594.x>

Downing, S. M., & Haladyna, T. M. (Orgs.). (2006). *Handbook of test development*. Lawrence Erlbaum Associates.

- Ebel, R. L. (1956). Obtaining and reporting evidence on content validity. *Educational and Psychological Measurement*, 16(3), 269–282. <https://doi.org/10.1177/001316445601600301>
- Elosua, P., & Iliescu, D. (2012). Tests in Europe: Where we are and where we should go. *International Journal of Testing*, 12(2), 157–175. <https://doi.org/10.1080/15305058.2012.657316>
- Evers, A., Muñiz, J., Hagemester, C., Høtmælingen, A., Lindley, P., Sjöberg, A., & Bartram, D. (2013). Assessing the quality of tests: Revision of the EFPA review model. *Psicothema*, 25(3), 283–291. <https://doi.org/10.7334/psicothema2013.97>
- Faiad, C., Baptista, M. N., & Primi, R. (Orgs.). (2021). *Tutoriais em análise de dados aplicados a psicometria*. Editora Vozes. <https://www.livrariavozes.com.br/-prevenda-tutoriaisemanalisededadosaplicadosapsicometria6557132938/p>
- Flake, J. K., Pek, J., & Hehman, E. (2017). Construct validation in social and personality research: Current practice and recommendations. *Social Psychological and Personality Science*, 8(4), 370–378. <https://doi.org/10.1177/1948550617693063>
- Fornell, C., & Larcker, D. F. (1981). Evaluating structural equation models with unobservable variables and measurement error. *Journal of Marketing Research*, 18, 39–50. <https://doi.org/10.2307/3151312>
- Furr, R. M. (2020). Psychometrics in clinical psychological research. Em A. G. C. Wright & M. N. Hallquist (Orgs.), *The Cambridge handbook of research methods in clinical psychology* (p. 54–65). Cambridge University Press. <https://doi.org/10.1017/9781316995808.008>
- Gagnier, J. J., Lai, J., Mokkink, L. B., & Terwee, C. B. (2021). COSMIN reporting guideline for studies on measurement properties of patient-reported outcome measures. *Quality of Life Research*, 30(8), 2197–2218. <https://doi.org/10.1007/s11136-021-02822-4>
- Gall, J. P., Gall, M. D., & Borg, W. R. (2014). *Applying educational research: How to read, do, and use research to solve problems of practice* (6^o ed). Pearson.

- Gilbert, G. E., & Prion, S. (2016). Making sense of methods and measurement: Lawshe's content validity index. *Clinical Simulation in Nursing*, 12(12), 530–531. <https://doi.org/10.1016/j.ecns.2016.08.002>
- Gonçalves, M. M., Almeida, L. S., & Simões, M. R. (Orgs.). (2017). *Psicologia clínica e da saúde: Instrumentos de avaliação*. Pactor.
- Gudmundsson, E. (2009). Guidelines for translating and adapting psychological instruments. *Nordic Psychology*, 61(2), 29–45. <https://doi.org/10.1027/1901-2276.61.2.29>
- Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2009). *Multivariate data analysis* (7th ed). Prentice Hall.
- Hambleton, R. K. (1996). Guidelines for adapting educational and psychological tests. *Annual Meeting of the National Council on Measurement in Education*.
- Hambleton, R. K., Merenda, P. F., & Spielberger, C. D. (Orgs.). (2005). *Adapting educational and psychological tests for cross-cultural assessment*. Lawrence Erlbaum Associates.
- Hartley, J. (2014). Some thoughts on Likert-type scales. *International Journal of Clinical and Health Psychology*, 14(1), 83–86. [https://doi.org/10.1016/S1697-2600\(14\)70040-7](https://doi.org/10.1016/S1697-2600(14)70040-7)
- He, J., & van de Vijver, F. J. R. (2012). Bias and equivalence in cross-cultural research. *Online Readings in Psychology and Culture*, 2(2), 1–19. <https://doi.org/10.9707/2307-0919.1111>
- Heggestad, E. D., Scheaf, D. J., Banks, G. C., Monroe Hausfeld, M., Tonidandel, S., & Williams, E. B. (2019). Scale adaptation in organizational science research: A review and best-practice recommendations. *Journal of Management*, 45(6), 2596–2627. <https://doi.org/10.1177/0149206319850280>
- Hernández, A., Hidalgo, M. D., Hambleton, R. K., & Gómez-Benito, J. (2020). International test commission guidelines for test adaptation: A criterion checklist. *Psicothema*, 32(3), 390–398. <https://doi.org/10.7334/psicothema2019.306>

Holzinger, K. J., & Swineford, F. (1937). The bi-factor method. *Psychometrika*, 2(1), 41–54.

<https://doi.org/10.1007/BF02287965>

Huang, J. L., Bowling, N. A., Liu, M., & Li, Y. (2015). Detecting insufficient effort responding with an Infrequency Scale: Evaluating validity and participant reactions. *Journal of Business and Psychology*, 30(2), 299–311. <https://doi.org/10.1007/s10869-014-9357-6>

Hunsley, J. D., Lee, C. M., Wood, J. M., & Taylor, W. (2015). Controversial and questionable assessment techniques. Em S. O. Lilienfeld, S. J. Lynn, & J. M. Lohr (Orgs.), *Science and pseudoscience in clinical psychology* (2^o ed, p. 42–82). The Guilford Press.

Hunsley, J. D., & Mash, E. J. (2008). Developing criteria for evidence-based assessment: An introduction to assessments that work. Em J. Hunsley & E. J. Mash (Orgs.), *A guide to assessments that work* (p. 3–14). Oxford University Press.

Hutz, C. S., Bandeira, D. R., & Trentini, C. M. (Orgs.). (2015). *Psicometria*. Artmed.

Hutz, C. S., Bandeira, D. R., Trentini, C. M., & Krug, J. S. (Orgs.). (2016). *Psicodiagnóstico*. Artmed.

International Test Commission. (2018). ITC Guidelines for Translating and Adapting Tests (second edition). *International Journal of Testing*, 18(2), 101–134.

<https://doi.org/10.1080/15305058.2017.1398166>

Irwing, P., Booth, T., & Hughes, D. (Orgs.). (2018). *The Wiley handbook of psychometric testing: A multidisciplinary reference on survey, scale and test development* (Vols. 1–2). Wiley.

Irwing, P., & Hughes, D. J. (2018). Test development. Em P. Irwing, T. Booth, & D. Hughes (Orgs.), *The Wiley handbook of psychometric testing: A multidisciplinary reference on survey, scale and test development* (p. 3–47). John Wiley & Sons. <https://doi.org/10.1002/9781118489772.ch1>

Johnson, R. L., & Morgan, G. B. (2016). *Survey scales: A guide to development, analysis, and reporting*. Guilford Press.

Jöreskog, K. G. (1971). Statistical analysis of sets of congeneric tests. *Psychometrika*, 36(2), 109–133.

<https://doi.org/10.1007/BF02291393>

Kane, M. T. (2006). Validation. Em R. L. Brennan (Org.), *Educational measurement* (4º ed, p. 17–64).

ACE/Praeger.

Kelley, K., & Lai, K. (2018). Sample size planning for confirmatory factor models: Power and accuracy for effects of interest. Em P. Irwing, T. Booth, & D. Hughes (Orgs.), *The Wiley handbook of*

psychometric testing: A multidisciplinary reference on survey, scale and test development (p. 113–138). John Wiley & Sons. <https://doi.org/10.1002/9781118489772.ch5>

Krishnamurthy, R., VandeCreek, L., Kaslow, N. J., Tazeau, Y. N., Miville, M. L., Kerns, R., Stegman, R.,

Suzuki, L., & Benton, S. A. (2004). Achieving competency in psychological assessment: Directions for education and training. *Journal of Clinical Psychology*, 60(7), 725–739.

<https://doi.org/10.1002/jclp.20010>

Laher, S. (Org.). (2022). *International histories of psychological assessment*. Cambridge University Press.

<https://doi.org/10.1017/9781108755078>

Lane, S. (2014). Validity evidence based on testing consequences. *Psicothema*, 26(1), 127–135.

<https://doi.org/10.7334/psicothema2013.258>

Lane, S., Raymond, M. R., & Haladyna, T. M. (Orgs.). (2016). *Handbook of test development* (2º ed).

Routledge. <https://doi.org/10.4324/9780203102961>

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

<https://doi.org/10.1111/j.1744-6570.1975.tb01393.x>

Leong, F. T. L., Bartram, D., Cheung, F. M., Geisinger, K. F., & Iliescu, D. (Orgs.). (2016). *The ITC international handbook of testing and assessment*. Oxford University Press.

<https://doi.org/10.1093/med:psych/9780199356942.001.0001>

- Leung, S.-O. (2011). A comparison of psychometric properties and normality in 4-, 5-, 6-, and 11-point Likert scales. *Journal of Social Service Research*, 37(4), 412–421.
- <https://doi.org/10.1080/01488376.2011.580697>
- Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology*, 22(140), 1–55.
- Machado, A. P., & Morona, V. C. (2007). *Manual de avaliação psicológica*. Unificado.
- Mäder, B. J. (Org.). (2016). *Avaliação psicológica: Dimensões, campos de atuação e pesquisa*. Conselho Regional de Psicologia - 8ª região.
- Mair, P. (2018). *Modern psychometrics with R*. Springer. <https://doi.org/10.1007/978-3-319-93177-7>
- Mansur-Alves, M., Silva, R. S., & Fernandes, S. C. de Á. (2016). Impact of the Psychological Testing Assessment System (SATEPSI) for scientific publications in psychological assessment. *Psico-USF*, 21(1), 179–188. <https://doi.org/10.1590/1413-82712016210115>
- Manzi-Oliveira, A. B., Balarini, F. B., Marques, L. A. da S., & Pasian, S. R. (2011). Cross-cultural adaptation of psychological assessment instruments: Brazilian literature review from 2000 to 2010. *Psico-USF*, 16(3), 367–381.
- Marôco, J., Gonçalves, C., Lourenço, V., & Mendes, R. (2016). *PISA 2015 - Portugal: Literacia científica, literacia de leitura & literacia matemática: Vol. I*. IAVE.
- McCoach, D. B., Gable, R. K., & Madura, J. P. (2013). *Instrument development in the affective domain: School and corporate applications* (3º ed). Springer. <https://doi.org/10.1007/978-1-4614-7135-6>
- McDonald, R. P. (1999). *Test theory: A unified treatment*. Routledge.
- <https://doi.org/10.4324/9781410601087>
- Meehl, P. E. (1997). Credentialed persons, credentialed knowledge. *Clinical Psychology: Science and Practice*, 4(2), 91–98. <https://doi.org/10.1111/j.1468-2850.1997.tb00103.x>
- Messick, S. (1989). Validity. Em R. L. Linn (Org.), *Educational measurement* (3º ed, p. 13–103). American Council on Education/Macmillan.

Millsap, R. E. (2011). *Statistical approaches to measurement invariance*. Routledge.

<https://doi.org/10.4324/9780203821961>

Mokkink, L. B., Prinsen, C. A. C., Bouter, L. M., de Vet, H. C. W., & Terwee, C. B. (2016). The COnsensus-based standards for the selection of health measurement INstruments (COSMIN) and how to select an outcome measurement instrument. Em *Brazilian Journal of Physical Therapy* (Vol. 20, Número 2, p. 105–113). Revista Brasileira de Fisioterapia. <https://doi.org/10.1590/bjpt-rbf.2014.0143>

Moreau, D., & Wiebels, K. (2022). Psychological constructs as local optima. *Nature Reviews Psychology* 2022 1:4, 1(4), 188–189. <https://doi.org/10.1038/s44159-022-00042-2>

Morrison, K. M., & Embretson, S. (2018). Item generation. Em P. Irwing, T. Booth, & D. Hughes (Orgs.), *The Wiley handbook of psychometric testing: A multidisciplinary reference on survey, scale and test development* (Vol. 1, p. 75–94). Wiley. <https://doi.org/10.1002/9781118489772.ch3>

Muñiz, J., & Bartram, D. (2007). Improving international tests and testing. *European Psychologist*, 12(3), 206–219. <https://doi.org/10.1027/1016-9040.12.3.206>

Muñiz, J., Fernández-Hermida, J. R., Fonseca-Pedrero, E., Campillo-Álvarez, Á., & Peña-Suárez, E. (2012). Test reviewing in Spain. *International Journal of Testing*, 12(2), 176–184. <https://doi.org/10.1080/15305058.2012.661817>

Muñiz, J., & Fonseca-Pedrero, E. (2019). Ten steps for test development. *Psicothema*, 31(1), 7–16. <https://doi.org/10.7334/psicothema2018.291>

Muñiz, J., Prieto, G., Almeida, L. S., & Bartram, D. (1999). Test use in Spain, Portugal and Latin American countries. *European Journal of Psychological Assessment*, 15(2), 151–157. <https://doi.org/10.1027//1015-5759.15.2.151>

Nitko, A. J. (2015). *Using a mental measurements yearbook review and other materials to evaluate a test*. <https://buros.org/using-mental-measurements-yearbook-review-and-other-materials-evaluate-test>

Nunes, M. F. O., Muniz, M., Reppold, C. T., Faiad, C., Bueno, J. M. H., & Noronha, A. P. P. (2012).

Diretrizes para o ensino de avaliação psicológica. *Avaliação Psicológica*, 11(2), 309–316.

Nunnally, J. C. (1972). *Educational measurement and evaluation* (2º ed). McGraw-Hill.

Nunnally, J. C. (1973). *Introducción a la medición psicológica* (L. H. Donghi, Trad.). Editorial Paidós.

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3º ed). McGraw-Hill.

Oakland, T. (2004). Use of educational and psychological tests internationally. *Applied Psychology: An*

International Review, 53(2), 157–172. <https://doi.org/10.1111/j.1464-0597.2004.00166.x>

Oakland, T., & Iliescu, D. (2016). Ethical standards, guidelines, and related issues pertinent to

international testing and assessment. Em F. T. L. Leong, D. Bartram, F. Cheung, K. F. Geisinger, & D.

Iliescu (Orgs.), *The ITC international handbook of testing and assessment* (p. 425–437). Oxford

University Press. <https://doi.org/10.1093/med:psych/9780199356942.003.0029>

Ohrbach, R., Bjorner, J., Jezewski, M., John, M. T., & Lobbezoo, F. (2013). *Guidelines for establishing*

cultural equivalency of instruments.

Ordem dos Psicólogos Portugueses. (2015). *Parecer 24/CEOPP/2015: Sobre instrumentos e avaliação psicológica*.

Oren, C., Kennet-Cohen, T., Turvall, E., & Allalouf, A. (2014). Demonstrating the validity of three general

scores of PET in predicting higher education achievement in Israel. *Psicothema*, 26(1), 117–126.

<https://doi.org/10.7334/psicothema2013.257>

Pacico, J. C. (2015). Como é feito um teste? Produção de itens. Em C. S. Hutz, D. R. Bandeira, & C. M.

Trentini (Orgs.), *Psicometria* (p. 55–69). Artmed.

Padilla, J.-L., & Benítez, I. (2014). Validity evidence based on response processes. *Psicothema*, 26(1),

136–144. <https://doi.org/10.7334/psicothema2013.259>

Pasquali, L. (2011). *Psicometria: Teoria dos testes na psicologia e na educação* (4º ed). Editora Vozes.

- Penfield, R. D., & Camilli, G. (2006). 5 Differential Item Functioning and Item Bias. Em C. R. Rao & S. Sinharay (Orgs.), *Handbook of Statistics* (Vol. 26, p. 125–167). [https://doi.org/10.1016/S0169-7161\(06\)26005-X](https://doi.org/10.1016/S0169-7161(06)26005-X)
- Pereira, D. F., & Bandeira, D. R. (Orgs.). (2009). *Aspectos práticos da avaliação psicológica nas organizações*. Vetor.
- Plake, B. S., & Wise, L. L. (2014). What is the role and importance of the revised AERA, APA, NCME Standards for Educational and Psychological Testing ? *Educational Measurement: Issues and Practice*, 33(4), 4–12. <https://doi.org/10.1111/emip.12045>
- Polit, D. F., Beck, C. T., & Owen, S. V. (2007). Is the CVI an acceptable indicator of content validity? Appraisal and recommendations. *Research in Nursing & Health*, 30(4), 459–467. <https://doi.org/10.1002/nur.20199>
- Poortinga, Y. H. (1989). Equivalence of cross-cultural data: An overview of basic issues. *International Journal of Psychology*, 24, 737–756. <https://doi.org/10.1080/00207598908247842>
- Primi, R. (2010). Psychological assessment in Brazil: Foundations, current situation and future directions. *Psicologia: Teoria e Pesquisa*, 26, 25–35. <https://doi.org/10.1590/S0102-37722010000500003>
- Primi, R., Muniz, M., & Nunes, C. H. S. S. (2009). Definições contemporâneas de validade de testes psicológicos. Em *Avanços e polêmicas em avaliação psicológica* (p. 243–265). Casa do Psicólogo.
- R Core Team. (2023). *R: A language and environment for statistical computing (version 4.3.1) [Computer software]* (4.3.1). R Foundation for Statistical Computing. <https://www.r-project.org/>
- Raykov, T. (2001). Bias of coefficient α for fixed congeneric measures with correlated errors. *Applied Psychological Measurement*, 25(1), 69–76. <https://doi.org/10.1177/01466216010251005>
- Revelle, W. (1979). Hierarchical cluster analysis and the internal structure of tests. *Multivariate Behavioral Research*, 14(1), 57–74. https://doi.org/10.1207/s15327906mbr1401_4

- Revilla, M. A., Saris, W. E., & Krosnick, J. A. (2014). Choosing the number of categories in agree–disagree scales. *Sociological Methods & Research*, 43(1), 73–97.
<https://doi.org/10.1177/0049124113509605>
- Rios, J., & Wells, C. (2014). Validity evidence based on internal structure. *Psicothema*, 26(1), 108–116.
<https://doi.org/10.7334/psicothema2013.260>
- Royal, K. (2017). Four tenets of modern validity theory for medical education assessment and evaluation. *Advances in Medical Education and Practice*, 8, 567–570.
<https://doi.org/10.2147/AMEP.S139492>
- Rust, J., Kosinski, M., & Stillwell, D. (2021). *Modern psychometrics: The science of psychological assessment* (4^o ed). Routledge. <https://doi.org/10.4324/9781315637686>
- Salazar, M. S. (2015). El dilema de combinar ítems positivos y negativos en escalas. *Psicothema*, 27(2), 192–199. <https://doi.org/10.7334/psicothema2014.266>
- Sapnas, K. G., & Zeller, R. A. (2002). Minimizing sample size when using exploratory factor analysis for measurement. *Journal of Nursing Measurement*, 10(2), 135–154.
<https://doi.org/10.1891/jnum.10.2.135.52552>
- Schweizer, K., & DiStefano, C. (Orgs.). (2016). *Principles and methods of test construction: Standards and recent advances*. Hogrefe Publishing. <https://doi.org/10.1027/00449-000>
- Seabra-Santos, M. J., Almiro, P., Simões, M., & Almeida, L. S. (2019). Psychological tests in Portugal: Attitudes, problems and user profiles. *Revista Iberoamericana de Diagnóstico y Evaluación – e Avaliação Psicológica*, 53(4), 101–112. <https://doi.org/10.21865/RIDEP53.4.08>
- Shepard, L. A. (2016). Evaluating test validity: Reprise and progress. *Assessment in Education: Principles, Policy and Practice*, 23(2), 268–280. <https://doi.org/10.1080/0969594X.2016.1141168>
- Simões, M. R. (2005). Potencialidades e limites do uso de instrumentos no processo de avaliação psicológica. *Psicologia Educação e Cultura*, 9(2), 237–263.

- Sireci, S. G. (1998). The construct of content validity. *Social Indicators Research*, 45(1/3), 83–117.
<https://doi.org/10.1023/A:1006985528729>
- Sireci, S. G., & Faulkner-Bond, M. (2014). Validity evidence based on test content. *Psicothema*, 26(1), 100–107. <https://doi.org/10.7334/psicothema2013.256>
- Sireci, S. G., & Padilla, J.-L. (2014). Validating assessments: Introduction to the special section. *Psicothema*, 26(1), 97–99. <https://doi.org/10.7334/psicothema2013.255>
- Sireci, S. G., & Parker, P. (2006). Validity on trial: Psychometric and legal conceptualizations of validity. *Educational Measurement: Issues and Practice*, 25(3), 27–34. <https://doi.org/10.1111/j.1745-3992.2006.00065.x>
- Sireci, S. G., Yang, Y., Harter, J., & Ehrlich, E. J. (2006). Evaluating guidelines for test adaptations: A methodological analysis of translation quality. *Journal of Cross-Cultural Psychology*, 37(5), 557–567. <https://doi.org/10.1177/0022022106290478>
- Society for Industrial and Organizational Psychology. (2018). Principles for the validation and use of personnel selection procedures. *Industrial and Organizational Psychology*, 11(S1), 1–97.
<https://doi.org/10.1017/iop.2018.195>
- Streiner, D. L. (2003). Being inconsistent about consistency: When coefficient alpha does and doesn't matter. *Journal of Personality Assessment*, 80(3), 217–222.
https://doi.org/10.1207/S15327752JPA8003_01
- Suárez-Alvarez, J., Pedrosa, I., Lozano, L. M., García-Cueto, E., Cuesta, M., & Muñiz, J. (2018). Using reversed items in Likert scales: A questionable practice. *Psicothema*, 30(2), 149–158.
<https://doi.org/10.7334/psicothema2018.33>
- Valderas, J. M., Ferrer, M., MENDÍVIL, J., Garin, O., Rajmil, L., Herdman, M., & Alonso, J. (2008). Development of EMPRO: A tool for the standardized assessment of patient-reported outcome measures. *Value in Health*, 11(4), 700–708. <https://doi.org/10.1111/j.1524-4733.2007.00309.x>

- van de Vijver, F. J. R. (2016). Test adaptations. Em F. T. L. Leong, D. Bartram, F. M. Cheung, K. F. Geisinger, & D. Iliescu (Orgs.), *The ITC international handbook of testing and assessment* (p. 364–376). Oxford University Press.
- van de Vijver, F. J. R., & Fischer, R. (2009). Improving methodological robustness in cross-cultural organizational research. Em R. S. Bhagat & R. M. Steers (Orgs.), *Cambridge handbook of culture, organizations, and work* (p. 491–517). Cambridge University Press.
- <https://doi.org/10.1017/CBO9780511581151>
- van de Vijver, F. J. R., & Hambleton, R. K. (1996). Translating tests: Some practical guidelines. *European Psychologist*, 1(2), 89–99. <https://doi.org/10.1027/1016-9040.1.2.89>
- Vieira, R. de C., & Campos, R. H. de F. (2011). Notes on the introduction, reception and development of psychological measurement in Brazil. *Temas em Psicologia*, 19(2), 417–425.
- Wechsler, S. M., Hutz, C. S., & Primi, R. (2019). The development of psychological assessment in Brazil: Historical advances and challenges. *Avaliacao Psicológica*, 18(2), 121–128.
- Wechsler, S. M., Schulmeyer, M. K., Vinet, E. V., & Livia, J. (2022). Psychological assessment in South America: Perspectives from Brazil, Bolivia, Chile, and Peru. Em *International histories of psychological assessment* (p. 295–323). Cambridge University Press.
- <https://doi.org/10.1017/9781108755078.017>
- Wetzel, E., Lüdtke, O., Zettler, I., & Böhnke, J. R. (2015). The Stability of Extreme Response Style and Acquiescence Over 8 Years. *Assessment*, 23(3), 279–291.
- <https://doi.org/10.1177/1073191115583714>
- Willis, G. B. (2004). *Cognitive interviewing: A tool for improving questionnaire design*. SAGE Publications.
- Wilson, M. (2023). *Constructing measures: An item response modeling approach* (2^o ed). Routledge.
- <https://doi.org/10.4324/9781003286929>

Wolming, S., & Wikström, C. (2010). The concept of validity in theory and practice. *Assessment in Education: Principles, Policy and Practice*, 17(2), 117–132.

<https://doi.org/10.1080/09695941003693856>

Wood, J. M., Garb, H. N., & Nezworski, M. T. (2007). Psychometrics: Better measurement makes better clinicians. Em S. O. Lilienfeld & W. T. O'Donohue (Orgs.), *The great ideas of clinical science: The 17 concepts that every mental health practitioner should understand* (p. 77–92). Routledge.

Zumbo, B. D. (2014). What role does, and should, the Test Standards play outside of the United States of America? *Educational Measurement: Issues and Practice*, 33(4), 31–33.

<https://doi.org/10.1111/emip.12052>

Zumbo, B. D., & Chan, E. K. H. (Orgs.). (2014). *Validity and validation in social, behavioral, and health sciences*. Springer. <https://doi.org/10.1007/978-3-319-07794-9>