Propriedades psicométricas na avaliação de instrumentos: avaliação da confiabilidade e da validade

doi: 10.5123/S1679-49742017000300022

Psychometric properties in instruments evaluation of reliability and validity

Propiedades psicométricas en la evaluación de instrumentos: discusiones sobre la fiabilidad y validez

Ana Cláudia de Souza[1]
Neusa Maria Costa Alexandre[1]
Edinêis de Brito Guirardello[1]

[1]Universidade Estadual de Campinas, Faculdade de Enfermagem, Campinas-SP, Brasil

Resumo

Instrumentos de medida desempenham um importante papel na pesquisa, na prática clínica e na avaliação de saúde. Estudos sobre a qualidade desses instrumentos fornecem evidências de como as propriedades de medida foram avaliadas, auxiliando o pesquisador na escolha da melhor ferramenta para utilização. A confiabilidade e a validade são consideradas as principais propriedades de medida de tais instrumentos. Confiabilidade é a capacidade em reproduzir um resultado de forma consistente, no tempo e no espaço. Validade refere-se à propriedade de um instrumento medir exatamente o que se propõe. Neste artigo são apresentados, discutidos e exemplificados os principais critérios e testes estatísticos empregados na avaliação da confiabilidade (estabilidade, consistência interna e equivalência) e validade (conteúdo, critério e construto) de instrumentos. A avaliação das propriedades de medida de instrumentos é útil para subsidiar a seleção de instrumentos válidos e confiáveis, de modo a assegurar a qualidade dos resultados dos estudos.

Palavras-chave: Estudos de Validação; Reprodutibilidade dos Testes; Inquéritos e Questionários.

Endereço para correspondência:
Ana Cláudia de Souza – Rua Padre Brito, nº 208, apto. 501, Patos de Minas-MG, Brasil. CEP: 38700-172
E-mail: aclau35@gmail.com

Introdução

Atualmente, um número crescente de questionários ou instrumentos de medida que avaliam características psicossociais e diversos desfechos em saúde está disponível para uso em pesquisas, na prática clínica e na avaliação de saúde da população.[1] Apesar da criação de novos instrumentos, muitos não têm sido validados de maneira adequada.[2,3] A literatura vem alertando os pesquisadores para a necessidade de uma avaliação aprofundada das propriedades de medida de questionários.[4,5]

O pesquisador deve permanecer atento para a escolha de um instrumento adequado e preciso, de modo a garantir a qualidade de seus resultados. É necessário conhecer tais instrumentos detalhadamente – itens, domínios, formas de avaliação e, especialmente, propriedades de medida –, antes de utilizá-los. A qualidade da informação fornecida pelos instrumentos depende, em parte, de suas propriedades psicométricas.[6,7]

> *O pesquisador deve permanecer atento para a escolha de um instrumento adequado e preciso, de modo a garantir a qualidade de seus resultados.*

Antes de serem considerados aptos para uso, os instrumentos devem oferecer dados precisos, válidos e interpretáveis para a avaliação de saúde da população.[8] Além disso, as medidas devem fornecer resultados cientificamente robustos.[9] O desempenho dos resultados dessas medidas é, em grande parte, devido à confiabilidade e validade dos instrumentos.[10] Ainda que divergentes em alguns quesitos, pesquisadores são unânimes em considerar como principais propriedades de medida de instrumentos a confiabilidade e a validade.[11,12]

A Figura 1 ilustra as possíveis relações entre confiabilidade e validade. No primeiro alvo representado, os lances foram confiáveis, atingindo o mesmo ponto; porém, não atingiram o centro do alvo, não sendo considerados válidos. O segundo alvo pode ser considerado válido, embora não confiável uma vez que os pontos atingidos não se concentraram em um ponto específico, mas se espalharam por todo o alvo. O terceiro alvo não apresentou confiabilidade e validade, visto que atingiram pontos espalhados apenas na parte superior do alvo. O quarto alvo demonstra o exemplo perfeito de confiabilidade e validade: os lances atingiram o local que pretendiam e o fizeram de forma consistente, bem no centro do alvo. Tais relações também podem ser aplicadas à avaliação das propriedades de medida dos instrumentos.

Com base no que foi apresentado, considera-se relevante a discussão sobre os métodos de análise das propriedades de medida de instrumentos utilizados em pesquisa, na avaliação de saúde e na prática clínica. A seguir, são apresentados, discutidos e exemplificados os aspectos principais da avaliação da confiabilidade e validade de instrumentos de medida, bem como os testes estatísticos mais utilizados.

Confiabilidade

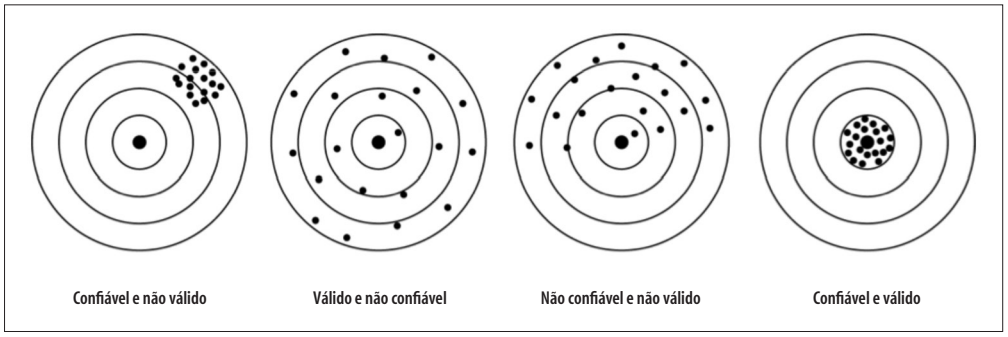
A confiabilidade – ou fidedignidade – é a capacidade em reproduzir um resultado de forma consistente no tempo e no espaço, ou a partir de observadores diferentes, indicando aspectos sobre coerência, precisão, estabilidade, equivalência e homogeneidade. Trata-se de um dos critérios principais de qualidade de um instrumento.[1]

A confiabilidade refere-se, principalmente, à estabilidade, consistência interna e equivalência de uma medida.[14] É importante ressaltar que a confiabilidade não é uma propriedade fixa de um questionário. Pelo contrário, a confiabilidade depende da função do instrumento, da população em que é administrado, das circunstâncias, do contexto; ou seja, o mesmo instrumento pode não ser considerado confiável segundo diferentes condições.[15]

Estimativas de confiabilidade são afetadas por diversos aspectos do ambiente de avaliação (avaliadores, características da amostra, tipo de instrumento, método de administração) e pelo método estatístico utilizado.[7] Portanto, os resultados de uma pesquisa utilizando instrumentos de medida só podem ser interpretados quando as condições de avaliação e a abordagem estatística são apresentadas de maneira clara.[16]

A confiabilidade refere-se a quão estável, consistente ou preciso é um instrumento.[17] A escolha dos testes estatísticos usados para avaliar a confiabilidade pode variar, dependendo do que se pretende medir.[15]

A seguir, serão abordados três critérios da confiabilidade de maior interesse para os pesquisadores, (i) estabilidade, (ii) consistência interna e (iii) equivalência,



Fonte: adaptado de Babbie.[13]

Figura 1 – Combinações possíveis de validade e confiabilidade de instrumentos de medida

bem como os métodos estatísticos mais usuais para avaliação de cada um desses aspectos.

Estabilidade

A estabilidade de uma medida é o grau em que resultados similares são obtidos em dois momentos distintos,[17] ou seja, é a estimativa da consistência das repetições das medidas.

A avaliação da estabilidade pode ser realizada pelo método de teste-reteste. Tal procedimento consiste na aplicação de uma mesma medida em dois momentos[17] O uso desse método requer que o fator a ser medido permaneça o mesmo nos dois momentos dos testes e qualquer mudança no escore pode ser causada por erros aleatórios:[15] por exemplo, se um indivíduo conclui uma pesquisa e a repete em alguns dias, é desejável que os resultados sejam similares.

O coeficiente de correlação intraclasse (*intraclass correlation coefficient*, ICC) é um dos testes mais utilizados para estimar a estabilidade de variáveis contínuas, pois leva em consideração os erros de medida.[18] Outros coeficientes de correlação, como o de Pearson ou o de Spearman, não são os mais adequados para esse tipo de teste de confiabilidade por não considerarem tais erros.[19]

A confiabilidade do teste-reteste tende a diminuir à medida que o tempo de reaplicação do teste é prolongado.[17] O intervalo de tempo entre as medições influenciará a interpretação da confiabilidade do teste-reteste; portanto, considera-se adequado um intervalo de 10 a 14 dias entre o teste e o reteste.[15]

Quanto à amostra, um número de pelo menos 50 sujeitos é considerado adequado.[1] Já quanto à interpretação dos resultados, valores mínimos de 0,70 são considerados satisfatórios.[1,20]

Consistência interna

A consistência interna – ou homogeneidade – indica se todas as subpartes de um instrumento medem a mesma característica.[21] Por exemplo, se um instrumento que avalia a satisfação do indivíduo com seu trabalho possui nove domínios, todos os itens do domínio 'remuneração' devem realmente medir tal construto e não um construto diferente, como 'benefícios', para que o instrumento apresente consistência interna. Trata-se de uma importante propriedade de medida para instrumentos que avaliam um único construto, utilizando, para isso, uma diversidade de itens.[1] Uma estimativa de consistência interna baixa pode significar que os itens medem construtos diferentes ou que as respostas às questões do instrumento são inconsistentes.[15]

A maioria dos pesquisadores avalia a consistência interna de instrumentos por meio do coeficiente alfa de Cronbach.[15,22] Desde a década de 1950,[23] tal medida é a mais utilizada para avaliação da confiabilidade.[24,25] O coeficiente alfa de Cronbach reflete o grau de covariância entre os itens de uma escala. Dessa forma, quanto menor a soma da variância dos itens, mais consistente é considerado o instrumento.[26]

Apesar de o coeficiente alfa de Cronbach ser o mais utilizado na avaliação da consistência interna, ainda não há consenso quanto a sua interpretação. Embora estudos determinem que valores superiores a 0,7 sejam os ideais,[1,20] algumas pesquisas consideram valores abaixo de 0,70 – mas próximos a 0,60 – como satisfatórios.[21,27]

É importante compreender que os valores do coeficiente alfa de Cronbach são fortemente influenciados pelo número de itens do instrumento de medida.[28] Pequeno número de itens por domínio

de um instrumento pode diminuir os valores de alfa, afetando a consistência interna.[29]

Os *softwares* estatísticos apresentam diversos modelos de confiabilidade, além do coeficiente alfa de Cronbach, e geralmente, os pesquisadores apresentam seus resultados juntamente com outros dois modelos de confiabilidade, o alfa se item deletado e a correlação média entre os itens.[21] Valores de alfa se item deletado permitem ao pesquisador avaliar se, ao retirar um item de determinado domínio do instrumento, o valor do coeficiente alfa de Cronbach total do domínio aumenta ou diminui.[28] Dessa forma, o pesquisador pode verificar, previamente, se algum item do instrumento está afetando o valor de alfa de Cronbach.[30]

Quanto à correlação média entre os itens, se esta for baixa, o valor do coeficiente alfa de Cronbach também será baixo. À medida que o coeficiente alfa aumenta, a correlação média acompanha essa elevação. Portanto, se as correlações forem altas, há evidência de que os itens medem o mesmo construto, satisfazendo a avaliação da confiabilidade.[21,28] Pesquisadores consideram que valores médios de correlação entre os itens superiores a 0,30 são considerados adequados e, portanto, medem o mesmo construto.[31]

Ainda, para instrumentos cujas variáveis são dicotômicas, o teste mais adequado é o de Kuder-Richardson e não o coeficiente alfa de Cronbach.[32] Assim como na interpretação dos resultados do coeficiente, valores próximos a 1,00 são considerados ideais.

Equivalência

A equivalência refere-se ao grau de concordância entre dois ou mais observadores quanto aos escores de um instrumento.[17] A forma mais comum de avaliar a equivalência é a confiabilidade interobservadores, que envolve a participação independente de dois ou mais avaliadores.[33] Nesse caso, o instrumento é preenchido pelos avaliadores.[15] Por exemplo, em uma pesquisa com dois avaliadores treinados que preenchem o mesmo instrumento, existe equivalência quando as pontuações obtidas forem as mesmas.

A confiabilidade interobservadores depende, principalmente, de um treinamento adequado dos avaliadores e de uma padronização da aplicação do teste.[34] Quando existe elevada concordância entre os avaliadores, infere-se que os erros de medição foram minimizados.[17]

O coeficiente Kappa é uma medida utilizada para avaliação interobservadores, aplicado a variáveis categóricas. Trata-se de uma medida de concordância entre os avaliadores e assume valor máximo igual a 1,00. Quanto maior o valor de Kappa, maior a concordância entre os observadores. Valores próximos ou abaixo de 0,00 indicam a inexistência de concordância.[35]

A Figura 2 apresenta, de modo resumido, os três tipos de confiabilidade discutidos anteriormente.

Salienta-se que a confiabilidade de um instrumento deve ser sempre discutida em função da população e do propósito do estudo. Um instrumento confiável para um conjunto de situações pode não ter a mesma confiabilidade em circunstâncias diferentes, razão pela qual a confiabilidade e a validade devem ser testadas sempre.[15]

Validade

A validade refere-se ao fato de um instrumento medir exatamente o que se propõe a medir.[36,37] Ressalta-se que a validade não é uma característica do instrumento e deve ser determinada com relação a uma questão particular, uma vez que se refere a uma população definida.[7]

As propriedades de medida – validade e confiabilidade – não são totalmente independentes.[17] Pesquisadores afirmam que um instrumento não confiável não pode ser válido; entretanto, um instrumento confiável pode, às vezes, não ser válido.[17,38] Dessa forma, uma confiabilidade elevada não garante a validade de um instrumento.[17]

Quanto aos tipos de validade, serão abordados no presente estudo os três principais, (i) validade de conteúdo, (ii) validade de critério e (iii) validade de construto:

Validade de conteúdo

A validade de conteúdo refere-se ao grau em que o conteúdo de um instrumento reflete adequadamente o construto que está sendo medido,[39] ou seja, é a avaliação do quanto uma amostra de itens é representativa de um universo definido ou domínio de um conteúdo.[17] Por exemplo, um instrumento que avalia a satisfação no trabalho deve incluir não somente a satisfação como também outras variáveis relacionadas a ela, a exemplo, a remuneração, promoção, relações com colegas de trabalho, entre outras.

Como não existe um teste estatístico específico para avaliação da validade de conteúdo, geralmente utiliza-se

Tipos de confiabilidade	Definição	Exemplo	Testes estatísticos
Estabilidade	Consistência das repetições das medidas, ou seja, o quão estável é a medida ao longo do tempo.[15,17]	Se um indivíduo conclui uma pesquisa e a repete em alguns dias, é esperado que os resultados sejam similares.	Teste-reteste (coeficiente de correlação intraclasse [ICC])
Consistência interna	Avalia se os domínios de um instrumento medem a mesma característica, ou seja, a correlação média entre todos os itens do construto.[21]	Em um instrumento que avalia satisfação no trabalho, todos os itens de um determinado domínio devem, de fato, medir tal construto e não um construto diferente.	Alfa de Cronbach (variáveis contínuas) Kuder-Richardson (variáveis dicotômicas)
Equivalência	É o grau de concordância entre dois ou mais avaliadores quanto aos escores de um instrumento.[17]	Dois avaliadores treinados preenchendo o mesmo instrumento devem obter a mesma pontuação.	Confiabilidade interobservador (Kappa)

Figura 2 – Medidas de confiabilidade de instrumentos

uma abordagem qualitativa, por meio da avaliação de um comitê de especialistas,[38] e após uma abordagem quantitativa com utilização do índice de validade de conteúdo (IVC).[40]

O IVC mede a proporção ou porcentagem de juízes em concordância sobre determinados aspectos de um instrumento e de seus itens.[5] Este método consiste de uma escala de Likert com pontuação de 1 a 4, em que: 1 = item não equivalente; 2 = item necessita de grande revisão para ser avaliada a equivalência; 3 = item equivalente, necessita de pequenas alterações; e 4 = item absolutamente equivalente.[40] Os itens que receberem pontuação de 1 ou 2 devem ser revisados ou eliminados. Para calcular o IVC de cada item do instrumento, basta somar as respostas 3 e 4 dos participantes do comitê de especialistas e dividir o resultado dessa soma pelo número total de respostas, conforme fórmula a seguir:[5,40]

IVC = Nº de respostas 3 ou 4/ Nº total de respostas

O índice de concordância aceitável entre os membros do comitê de especialistas deve ser de no mínimo 0,80 e, preferencialmente, maior que 0,90.[41]

Validade de critério

A validade de critério consiste na relação entre pontuações de um determinado instrumento e algum critério externo.[38] Este critério deve consistir em uma medida amplamente aceita, com as mesmas características do instrumento de avaliação, ou seja, um instrumento ou critério considerado 'padrão-ouro'.[15]

Em avaliações da validade de critério, os pesquisadores testam a validade de uma medida comparando-se os resultados da medida com um 'padrão-ouro' ou critério estabelecido.[7] Se o teste-alvo mede o que pretende medir, então seus resultados devem concordar com os resultados do 'padrão-ouro' ou do critério.[7] Seja qual for o construto avaliado, é considerado válido quando seus escores correspondem aos escores do critério escolhido.[17]

Quando o critério se situa no futuro, tem-se a validade preditiva, e quando é contemporâneo, tem-se a validade concorrente.[38] Ou seja, se um teste é aplicado e seus resultados são comparados com um critério aplicado um tempo depois, obtém-se a validade preditiva, e se ambos os testes são aplicados ao mesmo tempo, tem-se a validade concorrente.[7,17]

Como exemplo de validade preditiva, tem-se estudos sobre avaliação da pressão e níveis de colesterol como fatores preditivos para projetar risco de doença cardiovascular.[38] Para exemplificar a validade concorrente, pode-se citar um estudo no qual pesquisadores buscavam uma alternativa para a aplicação de um instrumento extenso que avalia a depressão e testaram uma única pergunta – *Muitas vezes você se sente triste ou deprimido?* –, confirmando a validade de critério.[42]

Dessa forma, pode-se verificar se a medida investigada possui relação com padrões externos, validados comprovadamente, que avaliam o mesmo construto.[43] Quanto maior a relação entre os dois, maior a validade de critério.[7]

A validade de critério pode ser constatada por um coeficiente de correlação.[17] As pontuações do instrumento de medida são correlacionadas com os escores do critério externo e esse coeficiente é analisado.[15] Valores próximos a 1,00 indicam haver correlação, enquanto valores próximos de 0,00 indicam que não existe

correlação. São desejáveis coeficientes de correlação de 0,70 ou superiores.[17]

Na maioria das vezes, a validação de critério torna-se um desafio para o pesquisador,[38] por exigir uma medida 'padrão-ouro' a ser relacionada com o instrumento escolhido, muitas vezes não encontrada em todas as áreas do conhecimento. Também representa um desafio superar as expectativas de um instrumento reconhecido como 'padrão-ouro'. O pesquisador espera ao menos um instrumento que tenha alguma vantagem sobre o critério escolhido, seja pela maior facilidade de sua utilização, tempo menor de administração ou até mesmo um custo reduzido.[38,43]

Validade de construto

A validade de construto é a extensão em que um conjunto de variáveis realmente representa o construto a ser medido.[44,45] A fim de estabelecer a validade de construto, geram-se previsões com base na construção de hipóteses, e essas previsões são testadas para dar apoio à validade do instrumento.[45] Quanto mais abstrato o conceito, mais difícil é estabelecer a validade de construto.[17]

Dificilmente esse tipo de validade é obtido com um único estudo; geralmente, são realizadas diversas pesquisas sobre a teoria do construto que se pretende medir.[17,44] É essencial que exista uma teoria vinculada ao processo de validação de construto.[44] Dessa forma, quanto mais evidências, mais válida é a interpretação dos resultados.[38,46]

Pesquisadores subdividem a validade de construto em três tipos, teste de hipóteses, validade estrutural ou fatorial e validade transcultural:[37,39]

a) Teste de hipóteses

Existem diversas estratégias para confirmação da validade de construto pelo teste de hipótese. Uma delas é a técnica de grupos conhecidos.[7,17] Nesta abordagem, grupos diferentes de indivíduos preenchem o instrumento de pesquisa e em seguida, os resultados dos grupos são comparados.[17,38] Por exemplo, um instrumento que avalia a qualidade de vida pode ser aplicado a um grupo de pacientes com doença crônica e a um grupo de jovens saudáveis. Espera-se que tais resultados sejam divergentes e o instrumento se mostre sensível a ponto de detectar essas diferenças.[38] Além da verificação da validade de construto pela técnica de grupos conhecidos, também é possível obtê-la de outra forma, pelas avaliações da validade convergente e da validade discriminante do instrumento de pesquisa.[39]

Na ausência de um instrumento 'padrão-ouro', é possível testar a validade convergente por meio da correlação das pontuações do instrumento focal com os escores de outro instrumento que avalie um construto similar.[39] Assim, é possível verificar se o instrumento avaliado está fortemente correlacionado a outras medidas já existentes e válidas. Por exemplo, ao administrar dois instrumentos que avaliam a satisfação no trabalho, espera-se obter fortes correlações entre ambos. Altas correlações entre um novo teste e um teste similar são fortes evidências de que o novo instrumento também mede o mesmo construto que o outro instrumento.[38]

Já a validade discriminante testa a hipótese de que a medida em questão não está relacionada indevidamente com construtos diferentes, ou seja, com variáveis das quais deveria divergir.[39] Por exemplo, um instrumento que avalie a motivação para o trabalho deve apresentar baixas correlações com um instrumento que verifique a autoeficácia no trabalho.[32]

b) Validade estrutural ou fatorial

Outra técnica muito utilizada entre os pesquisadores para verificação da validade de construto estrutural é a análise fatorial. A análise fatorial fornece ferramentas para avaliar as correlações em um grande número de variáveis, definindo os fatores, ou seja, as variáveis fortemente relacionadas entre si.[17,45]

Pesquisadores recomendam que seja verificada a validade fatorial utilizando-se a análise fatorial confirmatória (*confirmatory factor analysis* [CFA]) ao invés da análise fatorial exploratória (*exploratory factor analysis* [EFA]).[37] A EFA proporciona ao pesquisador a quantidade de fatores necessários para representar os dados, ou seja, é uma ferramenta para explorar a dimensionalidade de um conjunto de itens. Já a análise fatorial confirmatória (CFA) é um modo de confirmar quão bem as variáveis analisadas representam um número menor de construtos;[45] ela também é utilizada para confirmar o modelo estrutural de um instrumento[37]

Na EFA, as variáveis produzem cargas para todos os fatores, enquanto na CFA as variáveis só produzem cargas nos fatores indicados no modelo. Dessa forma, o modelo confirmatório é muito mais rigoroso e muito mais restritivo, motivo pelo qual é fortemente indicado para validação de questionários.[39] Por exemplo, pesquisadores pretendem testar se algumas características do ambiente de trabalho – como autonomia e *feedback* – são preditoras da satisfação profissional. Para testar tal hipótese, os pesquisadores realizam uma análise fatorial confirmatória.

Uma técnica bastante utilizada entre os pesquisadores para testar a validade de construto é a modelagem de equações estruturais (*structural equation modeling* [SEM]), considerada uma mistura de CFA com análise de caminhos.[45] Tal método busca explicar as relações entre múltiplas variáveis.[45] Um modelo convencional em SEM consiste, na realidade, de dois modelos: o modelo de mensuração, que representa como as variáveis medidas se unem para representar os construtos; e o modelo estrutural, que demonstra como os construtos estão associados.[47]

Para avaliação do modelo de mensuração é comum verificar as validades de construto convergente e discriminante. Na validade convergente, os itens indicadores de um construto específico devem possuir uma elevada proporção de variância em comum. Já a validade discriminante é o grau em que um construto se difere dos demais.[45]

Existem diversas maneiras de estimar a validade convergente, entre elas a avaliação das cargas fatoriais. Cargas fatoriais altas são um indicativo de que convergem para um ponto comum, ou seja, existe validade convergente. A literatura indica que as cargas fatoriais devem ser de pelo menos 0,5 e idealmente superiores. Se um item apresentar valores inferiores a 0,5 torna-se um forte candidato a deixar o modelo fatorial.[45]

Outra medida é a avaliação da variância média extraída (*average variance extracted* [AVE]), que verifica a proporção da variância dos itens que são explicados pelo construto ao qual pertencem. Assim como na avaliação das cargas fatoriais, quando os valores de AVE são iguais ou superiores a 0,5 assume-se que o modelo converge para um resultado positivo.[48,49]

Por fim, para confirmação da validade convergente é usual avaliar a confiabilidade composta, que é uma estimativa de consistência interna, porém mais adequada ao método SEM porque prioriza as variáveis de acordo com suas confiabilidades – e não como o alfa de Cronbach, fortemente influenciado pelo número de variáveis nos construtos.[50]

Quanto à verificação da existência de validade discriminante, o pesquisador pode realizar a análise das cargas cruzadas. Para confirmar esse tipo de validade, os itens do instrumento avaliado devem apresentar cargas fatoriais mais elevadas nos construtos que foram previamente designados do que nos demais.[51]

Outro critério utilizado para avaliar a validade discriminante é a comparação das raízes quadradas das AVE com os valores de correlação entre os construtos. Para que exista validade discriminante, as raízes quadradas das AVE devem ser maiores do que a correlação entre os construtos.[48,49]

Concluída a avaliação das validades convergente e discriminante, parte-se para a análise do modelo estrutural ou modelo teórico. Trata-se da representação conceitual das relações entre os construtos. Para testar o modelo estrutural, deve-se concentrar no ajuste geral do modelo e nas relações entre os construtos.[50]

Inicialmente, para verificar as relações entre construtos e itens do modelo, procede-se o teste t de Student e o teste do qui-quadrado em que se verifica se os parâmetros são significativamente diferentes de zero. A qualidade de ajuste do modelo pode ser avaliada pelos coeficientes de determinação de Pearson (R2): valores iguais a 2% são classificados como efeito pequeno, 13% como efeito médio e 26% como efeito grande.[50] Também é possível avaliar a raiz do erro quadrático médio (*root mean square error of approximation* [RMSEA] <0,08), o índice de qualidade de ajuste (*goodness-of-fit* [GFI] >0,9), o índice de Tucker-Lewis (*Tucker-Lewis index* [TLI] >0,9), o índice de ajuste comparativo (*comparative fit index* [CFI] >0,95) e o índice de ajuste normalizado (*normed fit index* [NFI] >0,95).[45]

Outros dois indicadores de qualidade de ajuste também podem ser avaliados, a relevância ou validade preditiva (Q^2) e o tamanho do efeito (f^2). O Q^2 avalia quanto o modelo se aproxima do que se esperava dele e valores maiores que 0 são considerados adequados.[48] O f^2 avalia o quanto cada construto é importante para o ajuste do modelo e é obtido por meio da inclusão e exclusão de construtos do modelo. Valores de 2% são considerados como efeito pequeno do construto no ajuste do modelo, 15% efeito médio e 35% efeito grande.[48]

c) Validade transcultural

O terceiro tipo de validade de construto, a validade transcultural, diz respeito à medida em que as evidências suportam a inferência de que o instrumento original e um adaptado culturalmente são equivalentes.[39] Por exemplo, um instrumento que avalia a satisfação no trabalho e que foi traduzido e adaptado para um outro contexto cultural, deve possuir um desempenho similar ao da versão original.[51]

Para avaliar a validade transcultural, o grupo *Consensus-based Standards for the Selection of Health Measurement Instruments* (COSMIN), uma

equipe multidisciplinar internacional dedicada à melhoria da seleção de instrumentos de medida utilizados na pesquisa e na prática clínica, a partir de ferramentas mais adequadas,[52] lista alguns itens a serem avaliados. Por exemplo, se os itens foram traduzidos e retrotraduzidos por tradutores independentes, se a tradução foi revisada por um comitê de especialistas e se o instrumento foi pré-testado, entre outras questões.[53]

Além dessa lista, é possível encontrar outras com padrões para avaliação das propriedades de medida dos instrumentos. Tais listas podem ser utilizadas para testar a qualidade metodológica dos estudos sobre propriedades de medida.[53]

Em suma, a validade de construto é verificada por meio de procedimentos lógicos e empíricos. A Figura 3 apresenta as principais características dos três tipos de validade abordados anteriormente.

Tipos de validade	Definição	Exemplo	Testes estatísticos
Validade de conteúdo	É o grau em que um teste inclui todos os itens necessários para representar o conceito a ser medido.[17]	Um instrumento que avalia a satisfação no trabalho deve incluir não somente a satisfação, senão também outras variáveis relacionadas a ela como remuneração, promoção, relações com colegas de trabalho, entre outras.	- Abordagem qualitativa (comitê de especialistas) - Abordagem quantitativa (índice de validade de conteúdo [IVC])
Validade de critério	É avaliada quando um resultado pode ser comparado a um 'padrão-ouro'.		
Validade concorrente	Pode ser verificada aplicando-se ambos, o teste-alvo e o 'padrão-ouro', ao mesmo tempo.	Na investigação de depressão, aplica-se um novo instrumento e com ele uma questão supostamente 'padrão-ouro': *Você se sente frequentemente triste ou deprimido?*[38]	Testes de correlações
Validade preditiva	Primeiro o teste-alvo é aplicado e posteriormente o 'padrão-ouro'.[38]	Resultados de pressão arterial e níveis de colesterol são embasados em sua validade preditiva para projetar risco de doença cardiovascular.[38]	Testes de correlações
Validade de construto	É a extensão em que um conjunto de variáveis representa, de fato, o construto que foi projetado para medir.[44]		
Técnica de grupos conhecidos	Grupos diferentes de indivíduos realizam o preenchimento do instrumento de pesquisa e depois os resultados dos grupos são comparados.[38]	Um teste que avalia qualidade de vida pode ser aplicado a um grupo de pacientes com doença crônica e a um grupo de jovens saudáveis. Espera-se encontrar diferenças nos escores de qualidade de vida entre esses grupos.[38]	Testes de hipótese
Validade convergente	É obtida pela correlação do instrumento focal com outro instrumento que avalie um construto similar, esperando resultados de altas correlações entre os dois.[39]	Na aplicação de dois instrumentos que avaliam a satisfação no trabalho, espera-se obter fortes correlações.	Testes de correlações
Validade discriminante	Testa a hipótese de que a medida-alvo não está relacionada indevidamente, com construtos diferentes, ou seja, com variáveis das quais deveria divergir.[39]	Um instrumento que avalia motivação deve apresentar baixas correlações com um instrumento que mede auto-eficácia.[32]	Testes de correlações
Validade estrutural ou fatorial	Testa se uma medida capta a dimensionalidade hipotética de um construto.[39]	Pesquisadores querem testar se algumas características do ambiente de trabalho como a autonomia e o *feedback* são preditoras da satisfação profissional.	Análise fatorial e modelagem de equações estruturais
Validade transcultural	Medida em que as evidências suportam a inferência de que o instrumento original e um adaptado culturalmente são equivalentes.[39]	Um instrumento que avalia a satisfação no trabalho que foi traduzido e adaptado culturalmente para um outro contexto deve possuir um desempenho similar à versão original do instrumento.[51]	- Tradutores e retrotradutores independentes - Comitê de especialistas - Pré-teste[51]

Figura 3 – Medidas de validade de instrumentos

Considerações finais

O presente estudo buscou discutir os aspectos principais na avaliação das propriedades de medida de instrumentos utilizados em pesquisa, na prática clínica e na avaliação de saúde. Determinar quão rigorosamente os aspectos de confiabilidade e validade foram abordados em um estudo é essencial para garantia da qualidade dos instrumentos utilizados e na implementação prática dos resultados dos estudos.

Estudos de qualidade fornecem evidências de como todos esses fatores foram abordados, o que auxilia o pesquisador a decidir se deve ou não aplicar os resultados em sua área de pesquisa ou prática clínica. Ressalta-se que a confiabilidade e a validade não são propriedades fixas e, portanto, variam de acordo com as circunstâncias, população, tipo e finalidade do estudo.

Compreendendo que os instrumentos de medida integram a prática clínica e a pesquisa em diferentes áreas do conhecimento, a avaliação de sua qualidade é fundamental para a seleção de instrumentos que forneçam medidas válidas e confiáveis.

Contribuição das autoras

De Souza AC participou da revisão de literatura, discussão dos achados e redação do manuscrito. Guirardello EB colaborou na discussão e redação do manuscrito. Alexandre NMC contribuiu na redação e revisão do conteúdo.

Referências

1. Terwee CB, Bot SD, Boer MR, van der Windt, Knol DL, Dekker J, et al. Quality criteria were proposed for measurement properties of health status questionnaires. J Clin Epidemiol. 2007 Jan;60(1):34-42.
2. Kosowski T, McCarthy C, Reavey PL, Scott AM, Wilkins EG, Cano SJ, et al. A systematic review of patient-reported outcome measures after facial cosmetic surgery and/or nonsurgical facial rejuvenation. Plast Reconstr Surg. 2009 Jun;123(6):1819-27.
3. Chen CM, Cano SJ, Klassen AF, King T, McCarthy C, Cordeiro PG, et al. Measuring quality of life in oncologic breast surgery: A systematic review of patient-reported outcome measures. Breast J. 2010 Nov-Dec;16(6):587-97.
4. Salmond SS. Evaluating the reliability and validity of measurement instruments. Orthop Nurs. 2008 Jan-Feb;27(1):28-30.
5. Alexandre NMC, Coluci MZO. Validade de conteúdo nos processos de construção e adaptação de instrumentos de medidas. Cienc Saude Coletiva. 2011 jul;16(7):3061-68.
6. Fitch E, Brooks D, Stratford PW, et al. Physical rehabilitation outcome measures: a guide to enhanced clinical decision making. 2nd Ed. Hamilton, Ontario: Lippincott Williams & Wilkins; 2002.
7. Roach KE. Measurement of health outcomes: reliability, validity and responsiveness. J Prosthet Orthot. 2006 Jan;18(1S):8-12.
8. Alexandre NMC, Gallasch CH, Lima MHM, Rodrigues RCM. A confiabilidade no desenvolvimento e avaliação de instrumentos de medida na área da saúde. Rev Eletr Enf. 2013 jul-set;15(3):802-9.
9. Cano SJ, Hobart JC. The problem with health measurement. Patient Prefer Adherence. 2011;5:279-90.
10. Salmond SS. Evaluating the reliability and validity of measurement instruments. Orthop Nurs. 2008 Jan-Feb;27(1):28-30.
11. Cook DA, Beckman TJ. Current concepts in validity and reliability for psychometric instruments: theory and application. Am J Med. 2006 Feb;119(2):166.
12. Pittman J, Bakas T. Measurement and instrument design. J Wound Ostomy Continence Nurs. 2010 Nov-Dec;37(6):603-7.
13. Babbie E. The practice of social research. 4th Ed. Belmont: Wadsworth Publishing Company; 1986.
14. Martins GA. Sobre confiabilidade e validade. RBGN. 2006 jan-abr;8(20):1-12.
15. Keszei AP, Novak M, Streiner DL. Introduction to health measurement scales. J Psychosom Res. 2010 Apr;68(4):319-23.
16. Kottner J, Audigé L, Brorson S, Donner A, Gajewski BJ, Hróbjartsson A, et al. Guidelines for Reporting Reliability and Agreement Studies (GRRAS) were proposed. J Clin Epidemiol. 2011 Jan;64(1):96-106.
17. Polit DF, Beck CT. Fundamentos de pesquisa em enfermagem: métodos, avaliação e utilização. 7 ed. Porto Alegre: Artmed; 2011.

18. Vet HC, Terwee CB, Knol DL, Bouter LM. When to use agreement versus reliability measures. J Clin Epidemiol. 2006 Oct;59(10):1033-9.

19. Terwee CB, Schellingerhout JM, Verhagen AP, Koes BW, Vet HC. Methodological quality of studies on the measurement properties of neck pain and disability questionnaires: a systematic review. J Manipulative Physiol Ther. 2011 May;34(4):261-72.

20. Nunnally JC, Bernstein IH. Psychometric theory. 3rd Ed. New York: McGraw-Hill; 1994.

21. Streiner DL. Starting at the beginning: an introduction to coefficient alpha and internal consistency. J Pers Assess. 2003 Feb;80(1):99-103.

22. Streiner DL, Kottner J. Recommendations for reporting the results of studies of instrument and scale development and testing. J Adv Nurs. 2014 Sep;70(9):1970-9.

23. Cronbach LJ. Coefficient alpha and the internal structure of tests. Psychometrika 1951 Sep;16(3):297-334.

24. Beeckman D, Defloor T, Demarre L, Van Hecke A, Vanderwee K. Pressure ulcer prevention: development and psychometric validation of a knowledge assessment instrument. Int J Nurs Stud. 2010 Apr;47(4):399-410.

25. Bonett DG, Wright TA. Cronbach's alpha reliability: interval estimation, hypothesis testing, and sample size planning. J Organ Behav. 2015 Jan;36(1):3-15.

26. Pasquali L. Psicometria: teoria dos testes na psicologia e na educação. Rio de Janeiro: Vozes; 2013.

27. Balbinotti MAA, Barbosa MLL. Análise da consistência interna e fatorial confirmatório do IMPRAFE-126 com praticantes de atividades físicas gaúchos. Psico-USF. 2008 jan-jun;13(1):1-12.

28. Cortina JM. What is coefficient alpha? An examination of theory and applications. J Appl Psychol. 1993;78(1):98-104.

29. Sijtsma K. On the use, the misuse, and the very limited usefulness of Cronbach's alpha. Psychometrika. 2009 Mar;74(1),107-20.

30. Allen K, Reed-Rhoads T, Terry R, Murphy TJ, Stone AD. Coefficient Alpha: an engineer's interpretation of test reliability. JEE. 2008;97(1):87-94.

31. Streiner DL, Norman GR. Health measurement scales: a practical guide to their development and use. 4th Ed. Oxford University Press; 2008.

32. Aaronson N, Alonso J, Burnam A, Lohr KN, Patrick DL, Perrin E, et al. Assessing health status and quality-of-life instruments: attributes and review criteria. Qual Life Res. 2002 May;11(3):193-205.

33. Heale R, Twycross A. Validity and reliability in quantitative studies. Evid Based Nurs. 2015 Jul;18(3):66-7.

34. Rousson V, Gasser T, Seifert B. Assessing intrarater, interrater and test–retest reliability of continuous measurements. Statist Med. 2002 Nov;21(22):3431-46.

35. Salmond SS. Evaluating the Reliability and Validity of Measurement Instruments. Orthop Nurs. 2008 Jan-Feb;27(1):28-30.

36. Roberts P, Priest H. Reliability and validity in research. Nurs Stand. 2006 Jul;20(44):41-5.

37. Mokkink LB, Terwee CB, Patrick DL, Alonso J, Stratford PW, Knol DL, et al. The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes. J Clin Epidemiol. 2010 Jul;63(7):737-45.

38. Kimberlin CL, Winterstein AG. Validity and reliability of measurement instruments used in research. Am J Health Syst Pharm. 2008 Dec;65(23):2276-84.

39. Polit DF. Assessing measurement in health: beyond reliability and validity. Int J Nurs Stud. 2015 Nov;52(11):1746–53.

40. Coluci MZO, Alexandre NMC, Milani D. Construção de instrumentos de medida na área da saúde. Cienc Saude Coletiva. 2015 mar;20(3):925-36.

41. Polit DF, Beck CT. The content validity index: are you know what's being reported? Critique and recommendations. Res Nurs Health. 2006 Oct;29(5): 489-97.

42. Watkins C, Daniels L, Jack C, Dickinson H, van Den Broek M. Accuracy of a single question in screening for depression in a cohort of patients after stroke: comparative study. BMJ. 2001 Nov;323(7322):1159.

43. Fayers PM, Machin D. Quality of life. Assessment, analysis, and interpretation. The assessment, analysis, and interpretation of patient-reported outcomes. 2nd Ed. Chichester: John Wiley & Sons; 2007.

44. Martins GA. Sobre confiabilidade e validade. RBGN. 2006 jan-abr;8(20):1-12.

45. Hair Junior JF, Black WC, Babin BJ, Anderson RE, Tathan RL. Análise multivariada de dados. 6 ed. Porto Alegre: Bookman; 2009.

46. Lamprea JA, Gómez-Restrepo C. Validez en la evaluación de escalas. Rev Colomb Psiquiatr. 2007;36(2):340-8.

47. Chin WW, Newsted PR. Structural equation modelling analysis with small samples using partial least

squares. In.: Hoyle RH. Statistical strategies for small sample research. Thousand Oaks, CA: Sage; 1999. p. 307-41.

48. Hair Junior JF, Hult GTM, Ringle CM, Sarstedt M. A Primer on Partial Least Squares Structural Equation Modeling (PLS-SEM). Los Angeles: SAGE, 2014.

49. Fornell C, Larcker DF. Evaluating structural equation models with unobservable variable and measurement error. J Mark Res. 1981 Feb;18(1):39-50.

50. Ringle CM, Silva D, Bido DS. Modelagem de equações estruturais com utilização do SmartPLS. REMark. 2014 mai;13(2):54-71.

51. Chin WW. The partial least squares approach for structural equation modeling. In: Marcoulides, GA (editor). Modern methods for business research. London: Lawrence Erlbaum Associates Publishers; 1998. p. 295-336.

52. Mokkink LB, Prinsen CAC, Bouter LM, Vet HCW, Terwee CB. The COnsensus-based Standards for the selection of health Measurement Instruments (COSMIN) and how to select an outcome measurement instrument. Braz J Phys Ther. 2016 Mar-Apr;20(2):105-13.

53. Mokkink LB, Terwee CB, Patrick DL, Alonso J, Stratford PW, Knol DL, et al. COSMIN checklist manual. Amsterdam: COSMIN; 2012 [Cited 2016 Nov 2]. Available from: http://www.cosmin.nl/images/upload/files/COSMIN%20checklist%20manual%20v9.pdf

Abstract

Measurement instruments play an important role in research, clinical practice and health assessment. Studies on the quality of these instruments provide evidence of how the measurement properties were assessed, helping the researcher choose the best tool to use. Reliability and validity are considered the main measurement properties of such instruments. Reliability is the ability to reproduce a result consistently in time and space. Validity refers to the property of an instrument to measure exactly what it proposes. In this article, the main criteria and statistical tests used in the assessment of reliability (stability, internal consistency and equivalence) and validity (content, criterion and construct) of instruments are presented, discussed and exemplified. The assessment of instruments measurement properties is useful to subsidize the selection of valid and reliable tools, in order to ensure the quality of the results of studies.

Keywords*: Validation Studies; Reproducibility of Results; Surveys and Questionnaires.*

Resumen

Instrumentos de medición juegan un papel importante en la investigación, la práctica clínica y la evaluación de la salud. Estudios sobre la calidad de instrumentos proporcionan evidencia de cómo se evaluaron las propiedades de medición, lo que ayuda al investigador a elegir la mejor herramienta. La fiabilidad y validez son las principales propiedades de medición de los instrumentos. La fiabilidad es la capacidad de reproducir un resultado de forma consistente en el tiempo y espacio. La validez se refiere a la propiedad de un instrumento de medir exactamente lo que se propone. En este artículo son presentados, discutidos y ejemplificados los principales criterios y las pruebas estadísticas utilizadas en la evaluación de la fiabilidad (estabilidad, consistencia interna y equivalencia) y validez (contenido, criterio y constructo). La evaluación de las propiedades de los instrumentos de medición es útil para apoyar la selección de instrumentos válidos y fiables para garantizar la calidad del estudio.

Palabras-clave*: Estudios de Validación; Reproducibilidad de Resultados; Encuestas y Cuestionarios.*

Recebido em 12/12/2016
Aprovado em 27/12/2016