



Texto & Contexto Enfermagem

ISSN: 0104-0707

textoecontexto@contato.ufsc.br

Universidade Federal de Santa Catarina
Brasil

Echevarría-Guanilo, Maria Elena; Gonçalves, Natália; Priscila Juceli Romanoski, Priscila x
PROPRIEDADES PSICOMÉTRICAS DE INSTRUMENTOS DE MEDIDAS: BASES
CONCEITUAIS E MÉTODOS DE AVALIAÇÃO–PARTE I
Texto & Contexto Enfermagem, vol. 26, núm. 4, 2017, pp. 1-11
Universidade Federal de Santa Catarina
Santa Catarina, Brasil

Disponível em: <http://www.redalyc.org/articulo.oa?id=71453540026>

- Como citar este artigo
- Número completo
- Mais artigos
- Home da revista no Redalyc

redalyc.org

Sistema de Informação Científica

Rede de Revistas Científicas da América Latina, Caribe, Espanha e Portugal

Projeto acadêmico sem fins lucrativos desenvolvido no âmbito da iniciativa Acesso Aberto

PROPRIEDADES PSICOMÉTRICAS DE INSTRUMENTOS DE MEDIDAS: BASES CONCEITUAIS E MÉTODOS DE AVALIAÇÃO - PARTE I

Maria Elena Echevarría-Guanilo¹, Natália Gonçalves², Priscila Juceli Romanoski³

¹ Doutora em Enfermagem. Professora do Departamento de Enfermagem da Universidade Federal de Santa Catarina (UFSC). Florianópolis, Santa Catarina, Brasil. E-mail: elena_meeg@hotmail.com

² Doutora em Enfermagem. Professora do Departamento de Enfermagem da UFSC. Florianópolis, Santa Catarina, Brasil. E-mail: nataliasjbv@gmail.com

³ Doutoranda do Programa de Pós-Graduação em Enfermagem da UFSC. Bolsista CAPES. Florianópolis, Santa Catarina, Brasil. E-mail: priscila.romanoski@gmail.com

RESUMO

Objetivo: apresentar e discutir bases conceituais e métodos de avaliações que fundamentam importantes propriedades de instrumentos de medidas.

Método: estudo teórico embasado na literatura internacional e nacional e nos instrumentos *Consensus-based Standards for the selection of health Measurement Instruments* e *Evaluating the Measurement of Patient-Reported Outcomes* que contemplam conceitos de avaliação de instrumentos para apreciação de resultados relatados pelo paciente. Inicialmente são apresentados e discutidos os conceitos de confiabilidade, responsividade e interpretabilidade, citados exemplos das principais formas de avaliação dessas propriedades.

Resultados: pode-se perceber que ainda há divergências em algumas descrições conceituais. Entretanto, os autores ressaltam a importância da confiabilidade para avaliar o instrumento de medida. Destaca-se a importância do conhecimento do Modelo Conceitual, das propriedades de medidas e dos diferentes métodos de avaliação para garantir, principalmente em estudo de validação de instrumentos, resultados confiáveis e válidos.

Conclusões: a discussão apresentada sobre a confiabilidade, responsividade e interpretabilidade contribui para os profissionais de saúde no conhecimento teórico e senso crítico na escolha de instrumentos e na condução de análises sobre essas propriedades de medida.

DESCRIPTORES: Confiabilidade dos dados. Psicometria. Estudos de validação. Inquéritos e questionários. Precisão da medição dimensional.

PSYCHOMETRIC PROPERTIES OF MEASUREMENT INSTRUMENTS: CONCEPTUAL BASES AND EVALUATION METHODS - PART I

ABSTRACT

Objective: to present and discuss conceptual bases and evaluation methods that support important properties of measurement instruments.

Method: a theoretical study based on the international and national literature and the Consensus-based Standards for the selection of health Measurement Instruments and Evaluating the Measurement of Patient-Reported Outcome which contemplates concepts of evaluation of instruments for the evaluation of results reported by the patient. Initially, the concepts of reliability, responsiveness and interpretability are mentioned and discussed, as well as the main ways of evaluating the properties

Results: it can be seen that there are still differences in some conceptual descriptions. However, the authors emphasize the importance of reliability in order to evaluate the measuring instrument. It is important to note the importance of knowing and understanding the Conceptual Model, the properties of measurements and the different evaluation methods that guarantee reliable and valid results, especially in the instrument validity studies.

Conclusions: the discussion presented on reliability, responsiveness and interpretability which contributes to health professionals with theoretical knowledge, and a critical sense in the choice of instruments and in conducting analyzes on these measurement properties.

DESCRIPTORS: Data accuracy. Psychometry. Validation studies. Surveys and questionnaires. Precision of dimensional measurement.

PROPIEDADES PSICOMÉTRICAS DE INSTRUMENTOS DE MEDIDAS: BASES CONCEPTUALES Y MÉTODOS DE EVALUACIÓN - PARTE I

Objetivo: presentar y discutir bases conceptuales y métodos de evaluaciones que fundamentan importantes propiedades de instrumentos de medidas. En esta primera parte, se presentan y discute los conceptos de confiabilidad, responsividad e interpretabilidad, citados ejemplos de las principales formas de evaluación de esas propiedades.

Método: estudio teórico basado en la literatura internacional y nacional y en los instrumentos *Consensus-based Standards for the selection of health Measurement Instruments* y *Evaluating the Measurement of Patient-Reported Outcomes* que contemplan conceptos de evaluación de instrumentos para la evaluación de los resultados de los pacientes.

Resultados: en este enfoque, se puede percibir que todavía hay divergencias en algunas descripciones conceptuales. Sin embargo, los autores resaltan la importancia de la confiabilidad para evaluar el instrumento de medida. Se destaca la importancia del conocimiento del Modelo Conceptual, de las propiedades de medidas y de los diferentes métodos de evaluación para garantizar, principalmente en estudio de validación de instrumentos, resultados confiables y válidos.

Conclusiones: se concluye que la discusión presentada sobre la confiabilidad, responsividad e interpretabilidad contribuye a los profesionales de la salud en el conocimiento teórico y sentido crítico en la elección de instrumentos y en la conducción de análisis sobre esas propiedades de medida.

DESCRIPTORES: Confiabilidad de los datos. Psicometría. Estudios de validación. Encuestas y cuestionarios. Precisión de la medición dimensional.

INTRODUÇÃO

Estudos voltados para a avaliação de propriedades psicométricas de instrumentos devem ser desenvolvidos sob importante rigor metodológico, de forma que garantam resultados adequados e conclusões apropriadas sobre as propriedades de medida do instrumento. Torna-se, dessa forma, necessário um consenso sobre conceitos, taxonomia, terminologia e definições sobre as propriedades de medida e o que elas representam.¹

Esses instrumentos de medida devem ser fundamentados em teorias, ser utilizados de forma apropriada² e apresentar determinadas características que justifiquem a confiabilidade dos dados que produzem. Portanto, qualquer instrumento de medida deverá ser calibrado para produzir resultados com o mínimo de erro possível.²⁻³

Entre o amplo número de instrumentos de medida propostos, encontram-se aqueles que incluem a medida de um conjunto de dimensões para cada construto teórico que pretende ser estudado ou para o qual se pretende conferir um valor numérico,⁴ isto é, associar conceitos subjetivos a referências numéricas. A utilização dessas estratégias de medida tem se tornado mais intensa nas últimas décadas, motivando a avaliação da validade interna e externa dos instrumentos.⁵

Usualmente, as propriedades de medida mais avaliadas são a validade e a confiabilidade do instrumento.⁶ A validade refere-se à qualidade de um instrumento para medir o construto para o qual foi construído, enquanto a confiabilidade diz respeito ao grau em que um instrumento permite reprodução e consistência de resultados quando aplicado em ocasiões diferentes.^{4,7} Ademais, outros autores têm

descrito que, além das duas propriedades supracitadas, a qualidade da medida dos instrumentos pode ser avaliada pela responsividade que é definida como a capacidade que o instrumento possui de detectar mudanças do estado de saúde dos pacientes ao longo do tempo.^{6,8}

A adaptação, para idiomas e culturas diferentes, de instrumentos para avaliação de construtos subjetivos também tem sido objeto de grande número de investigações, incluindo discussões sobre o processo metodológico apropriado para garantir que o instrumento preserve suas propriedades de validade e confiabilidade após a adaptação.¹ Na última década, esse tipo de investigação tem produzido um número considerável de pesquisas na área da enfermagem,⁹⁻¹³ refletindo a preocupação desses profissionais em identificar o instrumento de medida mais adequado para uma determinada situação ou condição, considerando aquele que aborda a monitorização de pacientes na prática clínica e que contemple a percepção dos próprios indivíduos na avaliação de seu estado de saúde.¹⁴

A partir da análise do aumento de instrumentos de medida disponível na literatura científica, autores identificaram dois aspectos importantes a serem superados: a necessidade de identificar os questionários disponíveis para uma utilização específica nos diversos temas, de forma que, possam ser de conhecimento dos profissionais interessados e, a necessidade de conhecer as medidas que seriam consideradas mais adequadas (válidas e confiáveis) entre os diversos instrumentos disponíveis.¹⁴ Ainda, destacam louváveis esforços em relação à disponibilidade de distintas ferramentas de busca – como livros, sites e bibliotecas on-line – que surgiram na tentativa de agrupar esses questionários.

Apesar do elevado número de instrumentos de medidas (versões originais e adaptadas) disponíveis na literatura, ainda se evidencia a necessidade de haver um consenso nos parâmetros de julgamento que permitam a identificação dos questionários, munidos de medidas adequadas para o construto proposto.¹⁴⁻¹⁷

Compreende-se que medidas de alta qualidade obtidas por meio de instrumentos são importantes para avaliar os benefícios de tratamentos, sejam esses farmacológicos ou não farmacológicos.³ Por essa razão, o estudo da qualidade teórica e das medidas torna-se relevante.

Na literatura identificam-se referenciais para avaliar as propriedades de medida,^{1,3,7,13,18-19} entre os quais, no presente estudo, destacam-se: *Consensus-based Standards for the selection of health Measurement Instruments (COSMIN)*,¹ o qual trata de normas baseadas em consensos para a seleção dos instrumentos de medida de saúde e o *Evaluating the Measurement of Patient-Reported Outcomes (EMPRO)* que contempla conceitos de avaliação de instrumentos para a apreciação de resultados relatados pelo paciente.^{13,18} Ambos foram propostos a fim de contribuir para a identificação de instrumentos de medida em saúde, cujas propriedades apresentassem dados consistentes, além de propor um consenso sobre as propriedades de medida de instrumentos que incorporam a perspectiva do paciente – *Patient-Reported Outcome (PRO)*.

O COSMIN consiste em uma lista de verificação que avalia: consistência interna; confiabilidade; erro de medida; validade de conteúdo (incluindo a validade de face); validade de construto (subdividida em três métodos, sobre validade estrutural, testes de hipóteses e validade transcultural); validade de critério; responsividade e, a interpretabilidade – que embora não seja considerada como uma propriedade de medida, é um requisito importante para a adequação de um instrumento de investigação ou na prática clínica.¹

O EMPRO é uma ferramenta cujo objetivo de avaliação seria a recomendação, ou não, de questionários propostos ou adaptados, disponíveis para aplicação. A ferramenta contempla oito atributos de avaliação, a saber: modelo conceitual e de medição, confiabilidade, validade, responsividade, interpretabilidade, encargos administrativos, modelos alternativos de administração e adaptações linguísticas e transculturais. Ainda, ao final do instrumento, o revisor provê uma avaliação global de recomendação da medida e do questionário avaliado.¹⁸

O objetivo deste estudo é apresentar e discutir bases conceituais e métodos de avaliações que fundamentam importantes propriedades de instru-

mentos de medidas. Na primeira parte do estudo são apresentados e discutidos os diversos conceitos encontrados na literatura nacional e internacional, que fundamentam importantes propriedades dos instrumentos de medida, como a confiabilidade, responsividade e interpretabilidade, além disto, são citados exemplos das principais formas de avaliação dessas propriedades.

Na segunda parte são apresentados os resultados do estudo teórico realizado. A busca dos dados ocorreu na literatura internacional e nacional e nos instrumentos COSMIN e EMPRO que contemplam conceitos de avaliação de instrumentos para apreciação de resultados relatados pelo paciente.

ASPECTOS CONCEITUAIS GERAIS DAS MEDIDAS

O estudo das propriedades de medidas de construtos relacionados à saúde envolve a identificação *a priori* de diversos aspectos:

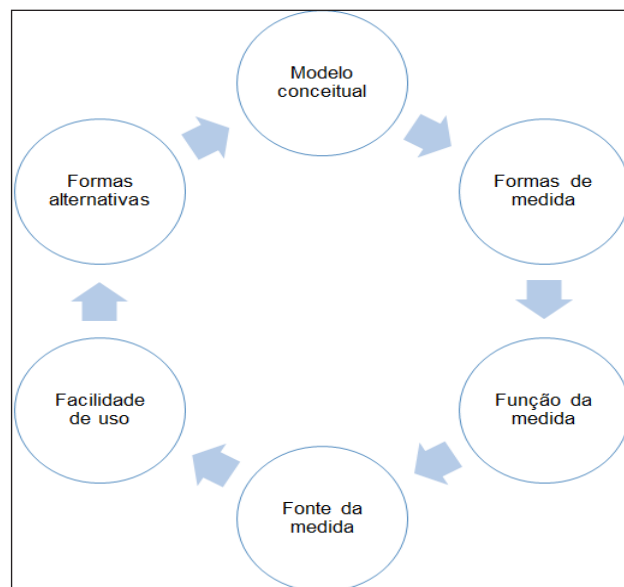


Figura 1 – Aspectos básicos conceituais na escolha do instrumento de medida. Florianópolis-SC, 2017

O modelo conceitual e a função da medida devem ser expressos claramente no instrumento e conhecido pelos pesquisadores, uma vez que estes critérios de escolha representam aspectos importantes no processo de construção e de validação de instrumentos de medida em saúde. Estes representam o caminho escolhido pelos autores para avaliar, nas populações-alvo, determinado construto. Dessa forma, ambos, os conceitos e a função da medida, devem apresentar relação e argumentação adequada.^{1,3,14}

Na área da saúde a função da medida pode não somente apresentar diferentes proposições como também ser utilizada em diversos tipos de estudos, a depender dos objetivos dos pesquisadores.²⁰ Na classificação de Kirshner e Guyatt,²⁰ a medida é considerada “discriminante” quando ela identifica diferenças nos resultados dos próprios indivíduos estudados e/ou entre grupos de interesse. Se a proposta é prever resultados de saúde, por exemplo, em estudos que abordam diagnósticos ou condições de saúde que as pessoas podem vir a desenvolver, chama-se medida “preditiva”; já, se a proposta é avaliar os benefícios e/ou os resultados de tratamentos de saúde, as medidas são consideradas como “avaliativas”.²⁰ Conhecer aspectos como estes poderá beneficiar a escolha do pesquisador sob a melhor forma de utilização da medida, assim como auxiliar no planejamento da adaptação desta em populações distintas à da população de origem do instrumento.

Outra característica importante a ser considerada na escolha do instrumento é a forma de medida, isto é, se o questionário apresenta uma avaliação unidimensional, multidimensional ou se ele propõe, ainda, o estudo de um construto genérico (aplicado em qualquer situação ou condição de saúde) ou específico (aplicado em condições específicas de saúde ou população).^{3,14}

A fonte da medida representa outro aspecto relevante, uma vez que pode gerar resultados a partir “dos relatos da própria pessoa” (*Patient-Reported Outcomes*), os quais contemplam a aplicação por entrevistador ou por autoaplicação; por uma versão *proxy*, quando o participante apresenta alguma incapacidade/dificuldade e o resultado é dado por um familiar ou cuidador; e “observacionais”, representados por instrumentos nos quais o papel principal para o preenchimento é do observador.³ A depender da população-alvo, do objetivo da pesquisa e das condições cognitivas e clínicas dos participantes, essas diferentes fontes de medida devem ser levadas em consideração.

A facilidade do uso do instrumento também representa um aspecto relevante no conhecimento das medidas em saúde, uma vez que esta contempla os recursos necessários para administrar o instrumento, tais como: tempo de aplicação, objetividade e facilidade. Entretanto, os autores destacam que não há padrões de referência para a avaliação desta propriedade,¹⁴ tratando-se de uma avaliação mais qualitativa e baseada na experiência e vivência dos conhecedores da temática. A isto, soma-se o conhecimento das formas alternativas de administração,¹⁴

como a obtenção da medida por meio de *softwares*, aplicativos ou dispositivos móveis e tecnológicos, os quais também devem ser considerados na escolha de um instrumento, uma vez que podem afetar a resposta do indivíduo ou a adesão da população à pesquisa. A respeito, no mundo, há um evidente acesso ascendente à internet, passando a ser a preferência, mesmo entre a população de faixas etárias mais avançadas; entretanto as respostas podem sofrer alterações dependendo da facilidade com que o respondente interage com a forma de apresentação de questionário *on-line* ou compreende a questão.²¹ Desse modo, caso a aplicação do instrumento seja via internet, o acesso a esta poderá influenciar tanto no número de pessoas respondentes quanto na qualidade das respostas em decorrência da falta de habilidade e/ou dificuldade de compreensão do sistema de respostas.²² Importa mencionar que, nas versões impressas, respondidas junto ao entrevistador, a entonação da voz ou a relação dele na rotina de cuidados do respondente poderá influenciar nas respostas, e quanto mais longo o instrumento mais cansativo poderá ser para o entrevistado.³

Os referidos aspectos, quando não propostos na versão original, podem ocasionar mudanças importantes na medida. Por isso, as formas alternativas de administração devem estar suficientemente descritas e justificadas nas pesquisas.¹⁴

As diferentes formas de apresentação dos instrumentos de medida (impresso, via computador e telefone), as características dos indivíduos – como o grau de escolaridade ou a preparação requerida do indivíduo; o tempo necessário para o preenchimento do instrumento e o esforço necessário para a compreensão e preenchimento, por parte do avaliador e do entrevistado, precisam ser considerados durante a fase de desenvolvimento e da adaptação do instrumento. Essas particularidades devem ser analisadas dentro do conjunto de informações oferecidas no momento da proposta de cada instrumento, assim como, na apresentação de uma nova versão (resumida ou adaptada).²³

CONFIABILIDADE (*RELIABILITY*)

A confiabilidade refere-se ao grau de consistência com que os itens do instrumento medem o atributo proposto livre de erro de medida e o grau em que o instrumento permite a reprodução e a obtenção de resultados consistentes, quando aplicados em diferentes ocasiões, exceto por erros aleatórios.^{1,3} Se há ausência de erros na medida ou se estes estão minimizados, a medida seria considerada confiável.

Na literatura, a confiabilidade também é referida como: precisão, concordância, equivalência, consistência, objetividade, fidedignidade, constância, reprodutibilidade, estabilidade, confiança e homogeneidade, sendo expressões também utilizadas para designar a confiabilidade do instrumento. O uso desses termos varia de acordo com o aspecto do teste que se quer realçar e com a literatura utilizada.^{1,3,24}

O estudo da confiabilidade contempla três importantes aspectos: a consistência interna, a confiabilidade propriamente dita e o erro de medida (Figura 2).

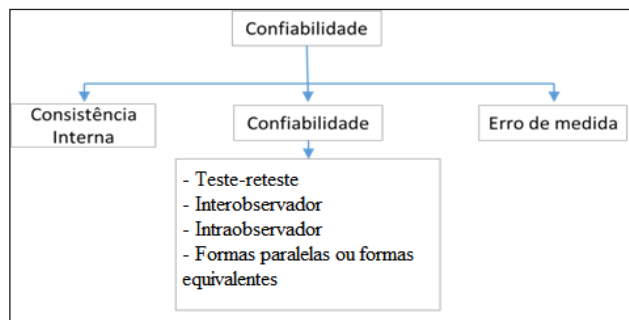


Figura 2 – Propriedades de medida: confiabilidade.
Florianópolis-SC, 2017

A confiabilidade também pode ser avaliada por meio do teste-reteste (*test-retest reliability*), interobservador (*inter-rater reliability*), intraobservador (*intra-rater reliability*) e formas paralelas (*parallel test*)^{1,3,24} ou formas equivalentes (*equivalent-forms*).⁷ Alguns autores referem-se ao teste-reteste e interobservador como reprodutibilidade.⁷

Um importante aspecto a ser considerado sobre a confiabilidade é que ela não é uma propriedade da medida fixa. Além disso, a confiabilidade pode variar de uma população para outra e por diferentes contextos. Dessa maneira, é recomendável que os pesquisadores avaliem quão similares são as populações dos estudos para decidir a necessidade de avaliação desta propriedade em seu estudo.³

Consistência interna (*internal consistency*)

Consistência interna refere-se à homogeneidade dos itens, ou seja, ao quanto os itens medem o mesmo atributo e produzem resultados consistentes. A análise da consistência interna torna-se possível para instrumentos compostos por múltiplos itens aplicados em apenas uma oportunidade.^{3,25} Para tanto, pode ser avaliada a consistência interna da totalidade dos itens (instrumentos unidimensionais) ou segundo subescalas que compõem o instrumento (instrumentos multidimensionais).

Dentre as formas de análises mais utilizadas para calcular a consistência interna de um instrumento de medida temos os testes de duas metades (*split-half*) ou bipartição, Kuder-Richardson e o coeficiente alpha de Cronbach.^{3,26}

A técnica de duas metades consiste na aplicação de um instrumento dividido (de forma aleatória ou não) em duas partes equivalentes (itens pares *versus* ímpares ou randomização, por exemplo), em uma única ocasião, sendo calculadas as correlações entre as pontuações das duas metades.^{24,27} Se os conteúdos das duas partes são consistentes, espera-se que a correlação seja positiva e próxima de 1, o que significaria que os itens do instrumento têm consistência interna, a qual, quanto mais próxima de um, representará maior força de correlação e, portanto, maior consistência interna.²⁷

Na proposta da análise de consistência por meio da técnica de Kuder-Richardson (KR-20), cada item é analisado individualmente, não requerendo subdivisão para a análise da consistência interna. A técnica é baseada na existência de correlação linear entre as respostas aos itens. O teste é recomendado para escalas aplicadas apenas uma vez e cujas respostas são dicotômicas, por exemplo, certo e errado.^{3,27} Não é recomendado sua aplicação para escalas que oferecem formatos de múltipla escolha; para estas, recomenda-se uma análise equivalente: o alpha de Cronbach.^{3,27}

O alpha de Cronbach é uma técnica que pode ser considerada uma extensão do método de Kuder-Richardson. Enquanto no método de KR-20 as variâncias dos itens são embasadas em valores para respostas dicotômicas, no alpha de Cronbach eles se baseiam em escores numéricos discretos que representam as diferentes possibilidades para cada item do instrumento.^{3,27} Parte-se do pressuposto de que a escala está composta por elementos homogêneos selecionados aleatoriamente da população e que os elementos evidenciam uma mesma característica. O alpha de Cronbach é recomendado para instrumentos de medida que adotam as escalas tipo Likert ou de múltipla escolha e cujas categorias apresentam uma ordem crescente ou decrescente de valores.²³

Quando se usa o alpha de Cronbach, é preciso considerar várias de suas características: o alpha confere um valor único para qualquer conjunto de dados e confere o valor para a média da distribuição de todos os possíveis coeficientes das partes que formam o instrumento, representando, assim, uma associação para o conjunto de dados determinado. Ademais, não depende somente da magnitude da correlação entre os itens, mas também do

número de itens da escala. E, se aumentarmos o número de itens de um instrumento, o valor alpha também será aumentado. Consequentemente os itens de dois instrumentos combinados em uma única escala aumentam o valor do alpha; e valores altos do alpha podem sugerir a existência de um nível alto de itens redundantes.²⁷

Para exemplificar os conceitos dos testes acima descritos, foi utilizada a técnica de duas metades em um estudo que objetivou adaptar culturalmente e validar uma versão em português da *HIV Antibody Testing Attitude Scale*. No estudo, o primeiro grupo de itens obteve um valor de 0,766 e o segundo grupo o valor de 0,750 resultando em uma correlação entre as duas formas de 0,819.²⁸ Para exemplificar a técnica Kuder-Richardson, o estudo de desenvolvimento e avaliação das propriedades métricas do instrumento *Knowledge of Malnutrition – Geriatric* (KoM-G) foi aplicado à enfermagem em lares de idosos austríacos. Nesse estudo, os autores avaliaram a consistência interna por meio da técnica Kuder-Richardson obtendo, para o total dos itens (20 itens), um valor de 0,69, considerado aceitável para a consistência interna.²⁸

Como exemplo do uso do método alpha de Cronbach, há um estudo de confiabilidade e validade da escala *Impact of Event* (IES) da versão brasileira, o qual obteve no coeficiente alpha de Cronbach, consistência interna para a escala total de 0,87. Esse coeficiente pode ser interpretado como uma alta consistência interna entre os itens do instrumento.²⁹ Outro exemplo que deve ser citado acerca de propriedades psicométricas da versão portuguesa é o da escala *Positive and Negative Affect Schedule*, aplicado em pessoas com doença renal crônica em programa de hemodiálise, no qual o alpha de Cronbach global na primeira avaliação foi de 0,80 e na segunda avaliação foi de 0,91, o que demonstra uma alta consistência interna entre os itens do instrumento.³⁰

Confiabilidade (*reliability*)

Conforme ilustra a figura 2, a confiabilidade pode ser avaliada pela consistência interna, confiabilidade em si e erro de medida, descritos a seguir.

Teste-reteste (test-retest), interobservador (inter-rater) e intraobservador (intra-rater)

A reprodutibilidade (*reproducibility*) refere-se ao grau em que o instrumento produz os mesmos resultados quando aplicado em momentos distintos não muito distantes um do outro. A depender do tipo de variável e escala utilizada na pesquisa,

pode-se optar pela análise da confiabilidade pelo teste-reteste, interobservador ou, mais raramente, pela medida intraobservador. No teste-reteste, o pesquisador aplica o instrumento de medida duas vezes em um mesmo grupo de pessoas considerado um determinado tempo entre as aplicações. Quando a aplicação é realizada por diferentes observadores em uma mesma população, em um mesmo período ou momento, determina-se a confiabilidade interobservador.^{3,27} Na medida intraobservador, a confiabilidade é obtida pela classificação ou medida do mesmo observador em duas ocasiões distintas.⁶

Com a aplicação do teste-reteste, é possível verificar a estabilidade e a reprodutibilidade da medida.³ Entretanto, há ressalvas sobre sua utilização considerando a natureza da variável investigada, a desejabilidade social e a memória que pode influenciar as respostas da segunda medida.³¹ A possibilidade de modificação de traços de interesse durante o período de tempo decorrido entre o teste e o reteste³ pode ser apontada como uma desvantagem do teste. As atitudes, o humor, o conhecimento de certo tópico são exemplos de traços que podem sofrer alteração com o passar de um período de tempo curto. Portanto, o cálculo da estabilidade é mais indicado para características mais estáveis como a personalidade e a capacidade funcional, entre outros,³¹ uma vez que estados de humor podem ser influenciados por acontecimentos, tais como, o diagnóstico de uma doença e, assim, apresentar baixa estabilidade. Da mesma forma, no decorrer de duas medidas, o participante pode incorporar novos conhecimentos. Então, por exemplo, “traços” tendem a ser mais estáveis do que “estados”, e aspectos como estes seriam importantes parâmetros de determinação do tempo transcorrido entre as aplicações.³

Formas paralelas (parallel test) ou formas equivalentes (equivalent-forms)

Refere-se à concordância entre dois ou mais instrumentos que medem o mesmo atributo, cuja aplicação ocorreu em tempos distintos, em intervalo curto de tempo³² e aplicados aos mesmos indivíduos.^{1,3,27}

Trata-se da administração de formas alternativas da mesma medida para o mesmo grupo ou a grupos diferentes. O questionário original pode ter uma reformulação para medir o mesmo atributo ou construto, ou seja, tanto as perguntas quanto as respostas podem ser reformuladas ou sua ordem ser alterada, de forma a produzir dois itens cujo objeto de avaliação seja semelhante, mas não idêntico.³² Por conseguinte, quanto maior o grau de correlação

entre as duas formas, mais equivalentes seriam consideradas as medidas.

Em um estudo sobre desenvolvimento e validação do Questionário do Tratamento de Osteoporose (OSTREq) – que teve como objetivo avaliar os critérios dos médicos na escolha do tratamento da osteoporose –, os autores estudaram a confiabilidade do questionário a partir da consistência interna, teste-reteste e formas paralelas. A confiabilidade do paralelismo das formas foi examinada por meio de uma amostra aleatória de 40 médicos. Os escores das duas versões diferentes do questionário OSTREq foram altamente correlacionados para todos os fatores ($r > 0.989$), constatando-se a consistência dos resultados através de versões alternativas.³³

Contudo, há críticas quanto ao uso desse método, uma vez que, em estudos que avaliam variáveis subjetivas, por exemplo, a qualidade de vida na população geral, não é comum haver dois instrumentos considerados equivalentes⁷ e a construção de formas equivalentes/semelhantes tornaria o processo ainda mais dispendioso.³²

Entretanto, uma situação na qual a equivalência pode ser aplicável é quando o processo de medida envolve julgamentos subjetivos e devem ser realizados por mais de uma pessoa. Assim como no teste-reteste, o teste de formas paralelas envolve a administração para as mesmas pessoas em duas ocasiões distintas. Para estas medidas, é estimado um parâmetro de confiabilidade. Contrariamente ao reteste, a confiabilidade por formas paralelas é considerada adequada apenas para escalas multi-item.³

Para calcular a estabilidade de um instrumento, seja considerando o teste-reteste, interobservador ou intraobservador e formas paralelas, os métodos mais usados são: índice Kappa para variáveis binárias, índice Kappa ponderado (para variáveis categóricas ordinais) e o Coeficiente de Correlação Intraclasse (CCI) para variáveis contínuas.^{3,27,34}

O CCI é matematicamente equivalente aos índices Kappa e Kappa ponderado.⁷ É utilizado para quantificar a confiabilidade das medidas (duas ou mais) ou avaliar a concordância geral entre dois ou mais métodos, medidas ou observações diferentes em variáveis quantitativas contínuas; porém, em algumas situações pode ser usado para dados categóricos ou que têm mais do que quatro ou cinco categorias de respostas.^{3,27} Esse coeficiente é obtido por meio da análise de variância (ANOVA) com medidas repetidas e definido como a proporção da variabilidade total devido à variabilidade dos indivíduos.³⁵ Os valores podem variar entre zero (0) a um (1), com o valor zero indicando ausência de

concordância e o valor um indicando concordância absoluta. Por convenção, consideram-se valores abaixo de 0,4 como baixa confiabilidade, de 0,4 a 0,75 como confiabilidade regular ou boa, e valores superiores a 0,75 como confiabilidade excelente.^{35,36}

Embora o coeficiente de correlação de Pearson seja, às vezes, considerado como uma medida de reprodutibilidade, sua aplicação nessa avaliação não é recomendada por ser uma medida de associação linear e não de concordância.³⁴ Considera-se que esse valor reflete unicamente a intensidade da associação linear entre duas medidas e não proporciona informações sobre a concordância entre os valores.³⁷ A correlação de Pearson não avalia a magnitude da diferença entre as observações de um mesmo indivíduo, o que pode superestimar a confiabilidade.³⁴

Como exemplo, é relevante mencionar um estudo que buscou desenvolver um questionário para avaliar a Qualidade de Vida (QOL) e Qualidade de Cuidado (QOC) em pacientes com câncer. Esse estudo contou com amostra de 329 pacientes ambulatoriais e 239 pacientes internados. Os coeficientes de correlação intraclasse para todos os itens do questionário foram 0,79 e 0,89 em cada cenário de aplicação, demonstrando uma excelente validade interna.³⁸

Para variáveis não contínuas, como as medidas dicotômicas, a confiabilidade interobservador, intraobservador e teste-reteste pode ser medida pelo coeficiente de Kappa, ou Cohen's kappa,³ embora, seja recomendado avaliar a confiabilidade pelo coeficiente de correlação intraclasse quando possível.

Em um estudo realizado com enfermeiros, cujo objetivo foi avaliar a confiabilidade interobservadores de instrumento de triagem pediátrica, os autores utilizaram diferentes maneiras de calcular o coeficiente Kappa, como o k quadrático, o k linear e o k ponderado. Os autores também fizeram comparações entre faixas etárias, visto que avaliar sinais e sintomas de crianças menores de um ano pode ser mais difícil, e concluíram que o instrumento foi considerado confiável para a triagem pediátrica na urgência.³⁹

Apesar do uso frequente na área da saúde, os valores do coeficiente de kappa devem ser avaliados com cautela, uma vez que o k é fortemente influenciado pela distribuição da classificação das categorias. Autores ainda sugerem que não se deve avaliar os valores obtidos com base em classificações rígidas sobre um coeficiente bom, regular ou ruim.³

Erro da medida (*measurement error*)

Erros de medidas podem acontecer e sua presença tem como principal consequência a redução

da confiabilidade de um instrumento. Os erros de medida podem ocorrer de forma sistemática ou aleatória. O erro sistemático ou viés poderá afetar igualmente todas as medições seja pela influência do entrevistado ou por mudanças de avaliadores com treinamentos distintos. O erro aleatório poderá estar presente em algumas situações, por exemplo, no registro de informações quantitativas semelhantes (registro de pontuações de 66 para o valor 99).³

Outro aspecto que poderá influenciar a confiabilidade da medida é o tempo (curto ou amplo demais) transcorrido entre as medidas, as quais – embora apresentem características estáveis – podem acidentalmente passar a sofrer influência por distintas situações, como já explanado anteriormente. Com relação a este aspecto, é importante que o pesquisador compreenda que a confiabilidade não se trata de uma propriedade de medida fixa de um instrumento, já que pode variar entre populações e entre situações das populações.³

Os conceitos de erro da medida e confiabilidade estão relacionados, no entanto, como o coeficiente de confiabilidade sofre influência da variabilidade da amostra e o erro de medida não, os parâmetros de erro de medida fornecem melhores informações sobre os escores individuais,³ auxiliando na explicação e entendimento do achado. Na literatura, recomenda-se que os pesquisadores descrevam, nos estudos metodológicos, pelo menos um erro de medida do instrumento testado.⁶

Erros de medida refletem em menor confiabilidade, entretanto, a confiabilidade pode ser razoavelmente elevada mesmo quando o erro de medida não é aceitável. Inversamente, baixos erros de medida não garantem um parâmetro de medida confiável.³

O índice mais amplamente utilizado para avaliar o erro da medida é o “erro padrão” de medida (*standard error of measurement* – SEM), também considerado como erro típico da medida. É um índice que pode ser calculado junto com as estimativas de confiabilidade no teste reteste, interobservador, intraobservador e no teste paralelo. Ao contrário da confiabilidade, o índice SEM não é influenciado pela variabilidade da amostra. Ele representa a unidade da medida em si, e seu valor também não é afetado pelo coeficiente de confiabilidade da amostra com a qual este foi calculado.³

O SEM pode ser utilizado para o cálculo de Intervalos de Confiança (IC) das pontuações obtidas. Para o cálculo, seriam necessárias duas aplicações próximas para cada indivíduo, no qual o SEM seria representado pelo desvio-padrão de toda as pontuações.^{3,27}

Os conceitos discutidos anteriormente dizem respeito à Teoria Clássica dos Testes (*Classical Test Theory*). Apesar de não ser o objetivo de estudo desta reflexão teórica, cabe destacar que Teoria da Resposta ao Item (TRI) é um método mais recente para avaliação da consistência interna de um instrumento. Esse método baseia-se em modelos probabilísticos de um indivíduo em responder a um item de acordo com sua experiência relativa ao item, sua habilidade ou dificuldade para respondê-lo (nível de mobilidade, por exemplo).^{2,7,23}

OUTROS ASPECTOS IMPORTANTES DE AVALIAÇÃO

A sensibilidade ou responsividade está mais ligada às características da estrutura do instrumento e é uma importante propriedade psicométrica de estudos longitudinais, enquanto a interpretabilidade está relacionada com os usuários do instrumento (indivíduo, profissional, sociedade, entre outros).^{3,23,31}

Responsividade (*responsiveness*)

A responsividade é definida como a capacidade de o instrumento detectar diferenças ou mudanças no construto avaliado. Muitos autores ainda não a consideram como uma propriedade psicométrica; todavia, nas atuais classificações, ressaltam a importância dessa medida para avaliar a validade de mudança das pontuações.^{1,3,6}

Vários métodos têm sido utilizados para avaliar a responsividade do instrumento ou do construto. Na literatura, tem-se observado uma divisão entre abordagem baseada no critério, por meio do método de mudança de critério e escala global de avaliação; e, abordagem com base no construto,⁶ com a utilização de teste t, tamanho do efeito, média padrão de resposta e o índice de responsividade de Guyatt.^{1,3}

Dois métodos bastante utilizados para avaliar a mudança de escore ao longo do tempo são teste t e o tamanho do efeito. O uso do teste t pareado tem sido usado com a suposição de que valores maiores de mudança indicariam maior sensibilidade à mudança de um instrumento. Entretanto, esse método não é correto quando as mudanças de escores podem ocorrer de forma sistemática, por exemplo, devido ao efeito de aprendizado que ocorre quando uma pessoa responde ao mesmo instrumento por mais de uma vez.⁴⁰ Além disso, o teste t assume que as observações têm uma distribuição Normal (Gaussiana). Quando a amostra é pequena, nem sempre é possível verificar se essa suposição é correta.

A sensibilidade à mudança também pode ser estimada pelo tamanho do efeito, o qual considera a diferença das médias pelo desvio-padrão da média no tempo zero (primeira avaliação ou antes da intervenção), entre os grupos ou entre os momentos. Por convenção, o tamanho do efeito é interpretado como pequeno (entre 0,2 e 0,5), moderado (entre 0,5 e 0,8), e grande (maior que 0,8) (2, 5).⁴⁰

A média de respostas padronizadas (*Standardized response mean*) é outro método de avaliar a responsividade de situações ou condições entre um mesmo grupo.^{3,41} Em um estudo que avaliou as propriedades psicométricas do *Quality of Recovery-40* (QoR-40) em pacientes submetidos à prostatectomia radical, os autores utilizaram a média de respostas padronizadas para avaliar a responsividade do instrumento e constaram que o questionário conseguiu identificar alterações na recuperação cirúrgica. Assim como nos resultados do tamanho do efeito, os autores classificaram a magnitude da mudança das medidas em pequena, moderada e grande.⁴¹

Interpretabilidade (*interpretability*)

A interpretabilidade é um conceito relacionado à responsividade; contudo, refere-se ao grau em que os valores obtidos por meio da aplicação do instrumento produzem informações relevantes ao indivíduo e ao profissional em relação ao construto a ser medido.^{1,3}

A interpretabilidade do instrumento pode ser baseada em comparações entre populações, por exemplo, quando se compara a qualidade de vida medida por um instrumento entre dois grupos de indivíduos: sadios e com doença cardíaca. A interpretação dos valores de um instrumento também pode ser baseada primeiramente no indivíduo, por exemplo, ao compará-lo a uma população (o indivíduo está dentro ou fora das normas da população a que pertence) ou ao compará-lo a si mesmo (comparação dos valores do instrumento antes e depois de uma intervenção clínica).⁷

A interpretabilidade de uma medida é facilitada pela informação que traduz um valor quantitativo ou mudanças em valores para uma categoria qualitativa ou outra medida externa que tem um significado mais familiar ou fácil de interpretar.^{3,8} Na área da saúde, a interpretabilidade contribui para que se obtenham valores ou pontuações que podem ser aplicadas a situações clínicas de forma significativa. Para isso, além de conhecer se as pontuações são confiáveis e responsivas/válidas, é importante saber se as mudanças das pontuações são triviais ou importantes.³

Cabe salientar que a interpretabilidade da mudança das pontuações é um aspecto complexo. Embora se trate de um conceito bastante discutido, ainda não se conta com um método amplamente aceito.³

Considerando os conceitos discutidos neste trabalho, é importante, ainda, enfatizar a necessidade de um planejamento adequado do projeto de pesquisa. O conhecimento destes conceitos e a sua operacionalização, bem como o entendimento do construto a ser avaliado e da população-alvo são imprescindíveis para minimizar os vieses da pesquisa e, com isso, divulgar resultados válidos e confiáveis.

CONCLUSÃO

A utilização de instrumento de medidas requer, por parte do pesquisador, o conhecimento e domínio dos referenciais para avaliação das propriedades de medidas em saúde, as quais contemplam parâmetros de julgamento para a identificação do questionário mais apropriado para ser utilizado na avaliação do construto de interesse, uma vez que os resultados obtidos contribuem para a avaliação de benefícios de intervenções dos profissionais em saúde e podem determinar mudanças da prática do cuidado.

É importante que se conheça o modelo conceitual ou as teorias por meio dos quais os pesquisadores se fundamentaram para construir o instrumento, as justificativas para sua criação, bem como a população para o qual este foi criado e inicialmente validado.

A evidência de que um instrumento oferece dados confiáveis inicia-se na intencionalidade do pesquisador (proposta ou escolha de um instrumento) e apenas se concretiza na aceitação (compreensão) do instrumento pelo respondente.

Em relação às propostas de adaptação do instrumento em outros idiomas, é importante verificar, se as propriedades métricas do instrumento original, neste caso, se a confiabilidade permanece no novo instrumento, buscando evidências que confirmem a existência dessas propriedades por meio dos métodos descritos na literatura aqui apresentados.

AGRADECIMENTOS

Este trabalho foi realizado com apoio bolsa de estudos da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – CAPES/Agência de Apoio e Avaliação da Educação de Pós-Graduação no Ministério da Educação do Brasil, vinculada a Universidade Federal de Santa Catarina.

REFERÊNCIAS

- Mokkink LB, Terwee CB, Patrick DL, Alonso J, Stratford PW, Knol DL, et al. COSMIN checklist manual. COSMIN manual [Internet]. 2012 [cited 2017 Mar 01]. Available from: http://www.cosmin.nl/cosmin_checklist.html
- Pasquali L. Parâmetros psicométricos dos testes psicológicos. In: Pasquali L, editor. Técnicas de exame psicológico-TEP. São Paulo (SP): Casa do Psicólogo; 2001.
- Polit DF, Yang FM. Measurement and the measurement of change. China: Wolters Kluwer; 2016.
- Kimberlin CL, Winterstein AG. Validity and reliability of measurement instruments used in research. Am J Health Syst Pharm [Internet]. 2008 [cited 2017 Mar 01]; 65(23):2276-84. Available from: <http://dx.doi.org/10.2146/ajhp0703647>
- Wong KL, Ong SF, Kuek TY. Constructing a survey questionnaire to collect data on service quality of business academics. Eur J Soc Sci [Internet]. 2012 [cited 2017 Mar 01]; 29:209-21. Available from: <http://eprints.utar.edu.my/860/1/6343.pdf>
- Polit DF. Assessing measurement in health: beyond reliability and validity. Int J Nurs Stud [Internet]. 2015 [cited 2017 Mar 01]; (52):1746-53. Available from: <https://www.ncbi.nlm.nih.gov/pubmed/26234936>
- Fayers PM, Machin D. Quality of life: the assessment analysis and interpretation of patient-reported outcomes. 2 ed. England (UK): Wiley; 2007.
- Scholtes VA, Terwee CB, Poolman RW. What makes a measurement instrument valid and reliable?. Injury [Internet]. 2011 [cited 2017 Mar 01]; 42(3):236-40. Available from: <https://www.ncbi.nlm.nih.gov/pubmed/21145544>
- Flinkman M, LeinoKilpi H, Numminen O, Jeon Y, Kuokkanen L, Meretoja R. Nurse Competence Scale: a systematic and psychometric review. J Adv Nurs [Internet]. 2016 [cited 2017 Mar 01]. Available from: <https://www.ncbi.nlm.nih.gov/labs/articles/27731918/>
- De J, Wand AP. Delirium screening: a systematic review of delirium screening tools in hospitalized patients. Gerontologist [Internet]. 2015 Dec [cited 2017 Mar 01]; 55(6):1079-99. Available from: <https://doi.org/10.1093/geront/gnv100>
- Cornélio ME, Alexandre NMC, São-João TM. Measuring instruments in cardiology adapted into Portuguese language of Brazil: a systematic review. Rev Esc Enferm USP [Internet]. 2014 [cited 2017 Mar 01] 48(2):368-76. Available from http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0080-62342014000200368
- Garin O, Herdman M, Vilagut G, Ferrer M, Ribera A, Rajmil L, et al. Assessing health-related quality of life in patients with heart failure: a systematic, standardized comparison of available measures. Heart Fail Rev [Internet]. 2014 [cited 2017 Mar 01]; 19(3):359-67. Available from <https://doi.org/10.1007/s10741-013-9394-7>
- Maratía S, Cedillo S, Rejas J. Assessing health-related quality of life in patients with breast cancer: a systematic and standardized comparison of available instruments using the EMPRO tool. Qual Life Res [Internet]. 2016 Oct [cited 2017 Mar 01]; 25(10):2467-80. Available from <https://doi.org/10.1007/s11136-016-1284-8>
- Valderas JM, Ferrer M, Mendivil J, Garin O, Rajmil L, Herdman M, et al. Scientific Committee on "Patient-Reported Outcomes" of the IRYSS Network. Development of EMPRO: a tool for the standardized assessment of patient-reported outcome measures. Value Health [Internet]. 2008 Jul-Aug [cited 2017 Mar 01]; 11(4):700-8. Available from <https://doi.org/10.1111/j.1524-4733.2007.00309.x>
- Mokkink LB, Prinsen CA, Bouter LM, Vet HC, 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 [Internet]. 2016 Jan [cited 2017 Mar 01]; 20(2):105-13. Available from <https://doi.org/10.1590/bjpt-rbf.2014.0143>
- Sinclair S, Russell LB, Hack TF, Kondejewski J, Sawatzky R. Measuring compassion in healthcare: a comprehensive and critical review. Patient [Internet]. 2016 [cited 2017 Mar 01]; 1-17. Available from <https://doi.org/10.1007/s40271-016-0209-5>
- Schmidt S, Garin O, Pardo Y, Valderas J M, Alonso J, Rebollo P, EMPRO Group. Assessing quality of life in patients with prostate cancer: a systematic and standardized comparison of available instruments. Qual Life Res [Internet]. 2014 [cited 2017 Mar 01]; 23(8): 2169-81. Available from <https://doi.org/10.1007/s11136-014-0678-8>
- Valderas JM, Ferrer M, Mendivil J, Garin O, Rajmil L, Herdman M, et al. Development of EMPRO: A tool for the standardized assessment of patient-reported outcome measures. Value Health [Internet]. 2008b [cited 2017 Mar 01]; 11(4):700-8. Available from: <https://www.ncbi.nlm.nih.gov/pubmed/18194398>
- Furr RM, Bacharach VR. Psychometrics: and introduction. 2 ed. Los Angeles (US): SAGE Publications; 2013.
- Kirshner B, Guyatt G. A methodological framework for assessing health indices. J Chronic Dis [Internet]. 1985 [cited 2017 Mar 01]; 38(1):27-36. Available from: <https://www.ncbi.nlm.nih.gov/pubmed/3972947>
- Van Gelder MM, Bretveld RW, Roeleveld N. Webbased questionnaires: the future in epidemiology? Am J Epidemiol [Internet]. 2010 [cited 2017 Aug 01]; 172(11):1292-8. Available from: <https://www.ncbi.nlm.nih.gov/pubmed/20880962>
- Faleiros F, K  ppler C, Pontes FAR, Silva SSC, Goes FSN, Cucick CD. Use of virtual questionnaire and dissemination as a data collection strategy in scientific studies. Texto Contexto Enferm [Internet]. 2016 [cited 2017 Aug 01]; 25(4):e3880014. Available from: <http://dx.doi.org/10.1590/0104-07072016003880014>

23. 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* [Internet]. 2002 May [cited 2017 Mar 01]; 11(3):193-205. Available from: <https://www.ncbi.nlm.nih.gov/pubmed/12074258>
24. Pasquali L. Instrumentação psicológica: fundamentos e práticas. Porto Alegre (RS): Artmed; 2010.
25. Mokkink LB, Terwee CB, Patrick DL, Alonso J, Stratford PW, Knol DL, et al. The COSMIN checklist for assessing the methodological quality of studies on measurement properties of health status measurement instruments: an international Delphi study. *Qual Life Res* [Internet]. 2010 [cited 2017 Mar 01]; 19(4):539-49. Available from doi:10.1007/s11136-010-9606-8
26. 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* [Internet]. 2010 [cited 2017 Mar 01]; 63(7):737-45. Available from: <https://doi.org/10.1007/s11136-010-9606-8>
27. Waltz CF, Strickland OL, Lenz ER. Measurement in Nursing and Health Research. 5 ed. New York (US): Springer Publishing Company, LLC; 2017.
28. Frias AMA, Sim-Sim MMSF, Chora MAFC, Caldeira ECV. Adaptation and validation into Portuguese language of the HIV Antibody Testing Attitude Scale. *Acta Paul Enferm* [Internet]. 2016 [cited 2017 Mar 01]; 29(1):77-83. Available from: <http://dx.doi.org/10.1590/1982-0194201600011>
29. Echevarria-Guanilo ME, Dantas RA, Farina Jr JA, Alonso J, Rajmil L, Rossi LA. Reliability and validity of the Impact of Event Scale (IES): version for Brazilian burn victims. *J Clin Nurs* [Internet]. 2011 [cited 2017 Mar 01]; 20(11-12):1588-97. Available from <https://doi.org/10.1111/j.1365-2702.2010.03607.x>
30. Sousa LMMD, Marques-Vieira CMA, Severino SSP, Rosado JLP, José, HMG. Validation of the positive and negative affect schedule in people with chronic kidney disease. *Texto Contexto Enferm* [Internet]. 2016 [cited 2017 Mar 01]; 25(4). Available from: <http://dx.doi.org/10.1590/0104-07072016005610015>
31. Polit DF. Getting serious about test-retest reliability: a critique of retest research and some recommendations. *Qual Life Res* [Internet]. 2014 [cited 2017 Mar 01]; 23(6):1713-20. Available from: <https://link.springer.com/article/10.1007%2Fs11136-014-0632-9>
32. Bolarinwa AO. Principles and methods of validity and reliability testing of questionnaires used in Social and Health Science Researches. *Niger Postgrad Med J* [Internet]. 2015 [cited 2017 Mar 01]; 22(4):195-201. Available from <http://dx.doi.org/10.4103/1117-1936.173959>.
33. Makras P, Galanos A, Rizou S, Anastasilakis AD, Lyritis GP. Development and validation of an osteoporosis treatment questionnaire (OSTREQ) evaluating physicians' criteria in the choice of treatment. *Hormones (Athens)* [Internet]. 2016 [cited 2017 Mar 01]; 15(3):413-22. Available from: <http://dx.doi.org/10.14310/horm.2002.1684>.
34. Plichta EB, Kelvin EA. Munro's Statistical methods for health care research. 6. ed. Philadelphia (US): Lippincott; 2013.
35. McGraw KO, Wong SP. Forming inference about some intraclass correlation coefficients. *Psychol Methods* [Internet]. 1996 [cited 2017 Mar 01]; 1(1):30-46. Available from: <http://dx.doi.org/10.1037/1082-989X.1.1.30>
36. Fleiss JL. The design and analysis of clinical experiments. New York (US): Wiley; 1986.
37. Pasquali L. Psicometria – teoria dos testes na Psicologia e na Educação. 5ª ed. Florianópolis (SC): Editora Vozes; 2011.
38. Shimizu M, Fujisawa D, Kurihara M, Sato K, Morita T, Kato M, et al. Validation study for the Brief Measure of Quality of Life and Quality of Care. *Am J Hosp Palliat Care* [Internet]. 2017 [cited 2017 Mar 01]; 1:1049909117693576. Available from: <https://www.ncbi.nlm.nih.gov/pubmed/28273759>
39. Karjala J, Eriksson S. Inter-rater reliability between nurses for a new paediatric triage system based primarily on vital parameters: the Paediatric Triage Instrument (PETI). *BMJ Open* [Internet]. 2017 [cited 2017 Mar 01]; 7:e012748. Available from: <https://www.ncbi.nlm.nih.gov/pubmed/28235966>
40. Deyo RA, Diehr P, Patrick DL. Reproducibility and responsiveness of health status measures. Statistics and strategies for evaluation. *Control Clin Trials* [Internet]. 1991 Aug [cited 2017 Mar 01]; 12(4Suppl):142S-58S. Available from: [https://doi.org/10.1016/S0197-2456\(05\)80019-4](https://doi.org/10.1016/S0197-2456(05)80019-4)
41. Eduardo AHA, Santos CBD, Carvalho AMP, Carvalho ECD. Validation of the Brazilian version of the Quality of Recovery-40 Item questionnaire. *Acta Paul Enferm* [Internet]. 2016 [cited 2017 Mar 01]; 29(3):253-9. Available from: <http://dx.doi.org/10.1590/1982-0194201600036>

Correspondência: Maria Elena Echevarría-Guanilo
 Universidade Federal de Santa Catarina
 Centro de Ciências da Saúde, Bl I, Sla 408
 88040-900 - Campus Universitário, Trindade, Florianópolis,
 SC, Brasil
 E-mail: elena_meeg@hotmail.com

Recebido: 20 de março de 2017
 Aprovado: 12 de setembro de 2017