Como determinar a qualidade de um questionário de acordo com o *CONsensus-based Standards for the selection of health Measurement INstruments*? Um guia simplificado sobre as propriedades de medida de instrumentos de avaliação - Parte I: conceitos básicos e confiabilidade

How to determine the quality of a questionnaire according to the CONsensus-based Standards for the selection of health Measurement INstruments? A simplified guide to the measurement properties of assessment instruments - Part I: basic concepts and reliability

Thaís Cristina Chaves[1], Ana Carolina de Jacomo Claudio[2], Thamiris Costa de Lima[3], Roger Berg Rodrigues Pereira[4], Gabriela Zuelli Martins Silva[5], Helen Cristina Nogueira Carrer[1]

DOI 10.5935/2595-0118.20230093-pt

RESUMO

JUSTIFICATIVA E OBJETIVOS: O tipo de questionário que pretende captar a percepção/visão de um paciente sobre um aspecto a ser medido (ex: intensidade da dor) é chamado de Instrumento de Medida Baseado no Relato do Paciente (Patient Reported Outcome Measure - PROM). Um dos maiores desafios que clínicos e pesquisadores costumam enfrentar é quanto a tomar uma decisão sobre qual PROM utilizar para a avaliação de seu paciente com dor, especialmente devido à falta do letramento científico necessário para entender os critérios e termos empregados na área de propriedades de medida. Assim, os objetivos deste estudo (parte I) foram: (I) introduzir conceitos básicos sobre PROMs com enfoque na terminologia e critérios definidos através do *COnsensus-based Standards for the selection of health Measurement INstruments* (COSMIN), e (2) descrever as propriedades de medida do domínio confiabilidade.
MÉTODOS: Utilizando uma busca voltada para os artigos da iniciativa COSMIN, foi elaborado este estudo. Sendo o assunto muito extenso, os autores dividiram o texto em duas partes.
RESULTADOS: O presente artigo descreveu conceitos básicos sobre PROMs (propósitos e construtos), o processo de adaptação transcultural e as propriedades de medida do domínio confiabilidade (confiabilidade, medida de erro e consistência interna). De forma geral, um instrumento com qualidade adequada de confiabilidade deveria atender a alguns critérios, tais como: coeficiente de correlação intraclasse ≥ 0,70, medida de erro < mínima mudança clinicamente importante e α de Cronbach ≥ 0,70.
CONCLUSÃO: O entendimento sobre como determinar a qualidade da propriedade de medida de confiabilidade pode auxiliar os clínicos e pesquisadores na escolha dos melhores PROMs disponíveis. Um checklist para avaliação da qualidade das propriedades de medida de PROMs está descrita na parte II do artigo.
Descritores: Confiabilidade de dados, Dor crônica, Dor musculoesquelética, Inquéritos e questionários, Psicometria.

Thaís Cristina Chaves – https://orcid.org/0000-0002-6222-4961;
Ana Carolina de Jacomo Claudio – https://orcid.org/0000-0001-7694-2836;
Thamiris Costa de Lima – https://orcid.org/0000-0002-7371-6232;
Roger Berg Rodrigues Pereira – https://orcid.org/0009-0009-2607-5629;
Gabriela Zuelli Martins Silva – https://orcid.org/0000-0002-2846-9228;
Helen Cristina Nogueira Carrer – https://orcid.org/0000-0001-5821-937X.

1. Universidade Federal de São Carlos, Professor do Departamento de Fisioterapia, Departamento de Fisioterapia, São Carlos, SP, Brasil.
2. Universidade Federal de São Carlos, Mestranda no Programa de Pós-Graduação em Fisioterapia, Departamento de Fisioterapia. São Carlos, SP, Brasil.
3. Universidade Federal de São Carlos, Programa de Pós-Graduação em Fisioterapia, Departamento de Fisioterapia, São Carlos, SP, Brasil.
4. Universidade Federal de São Carlos, Mestrando no Programa de Pós-Graduação em Fisioterapia, Departamento de Fisioterapia, São Carlos, SP, Brasil.
5. Universidade de São Paulo, Faculdade de Medicina de Ribeirão Preto, Mestre do Programa de Pós-Graduação em Reabilitação e Desempenho Funcional, São Carlos, SP, Brasil.

Apresentado em 06 de setembro de 2023.
Aceito para publicação em 10 de outubro de 2023.
Conflito de interesses: não há – Fontes de fomento: não há.

DESTAQUES
- *Patient Reported Outcome Measure* (PROM) é a sigla para instrumento baseado no relato do paciente
- Os PROMs podem ser classificados em propósito: avaliativo, discriminativo e prognóstico
- COSMIN é o acrônimo para uma iniciativa que visa padronizar as propriedades de medida de PROMs
- O COSMIN recomenda 12 passos para o processo de adaptação transcultural
- A Confiabilidade avalia se no teste-reteste o escore do PROM fornece resultados semelhantes, em indivíduos estáveis

Editor associado responsável: Luciana Buin
https://orcid.org/0000-0002-1824-5749

Correspondência para:
Thaís Cristina Chaves
E-mail: thaischaves@ufscar.br

© Sociedade Brasileira para o Estudo da Dor

ABSTRACT

BACKGROUND AND OBJECTIVES: The type of questionnaire that aims to capture a patient's perception/view of an aspect to be measured (e.g. pain intensity) is called Patient Reported Outcome Measure (PROM). One of the biggest challenges that clinicians and researchers often face is making a decision about which PROM to use for the assessment of their patient with pain, especially due to the lack of scientific literacy needed to understand the criteria and terms used in the field of measurement properties. Thus, the objectives of this narrative review (part I) were: (I) to introduce basic concepts about PROMs with a fo-

cus on the terminology and criteria defined through the COnsensus-based Standards for the selection of health Measurement INstruments (COSMIN), and (2) to describe the measurement properties of the reliability domain.

METHODS: This study was produced using a search for articles from the COSMIN initiative. As the subject is very extensive, the authors divided the text into two parts.

RESULTS: This study described basic concepts about PROMs (purposes and constructs), the process of cross-cultural adaptation and the measurement properties of the reliability domain (reliability, error measure and internal consistency). In general, an instrument with adequate reliability quality should meet certain criteria, such as: intraclass correlation coefficient > 0.70, error measure < minimal clinically important change and Cronbach's Cronbach's $\alpha \geq 0.70$.

CONCLUSION: The understanding on how to determine the quality of reliability can assist clinicians and researchers in choosing the best PROMs available. A checklist for assessing the quality of the measurement properties of PROMs will be described in the part II of the manuscript.

Keywords: Chronic pain, Data reliability, Musculoskeletal pain, Psychometrics, Surveys and questionnaires.

INTRODUÇÃO

Definindo conceitos básicos

Aplicar um questionário/escala para captar a percepção/visão de um indivíduo sobre um aspecto a ser medido[1] (ex: intensidade da dor) é o mesmo que utilizar um instrumento de medida baseado no relato do indivíduo. PROM (Patient Reported Outcome Measure) é a sigla em inglês para um instrumento de medida baseado no relato do paciente[1]. Outra sigla utilizada comumente é OMI (Outcome Measurement Instrument) ou Instrumento de Medida[2]. Os instrumentos tipo PROM foram desenvolvidos com o intuito de avaliar construtos ou conceitos que não podem ser diretamente mensurados ou que seriam difíceis de serem medidos na prática (ex: performance ao realizar atividades de vida diária)[3]. A sigla PRO se aplica ao construto avaliado pelo instrumento, enquanto a sigla PROM se aplica ao instrumento de medida. Um exemplo de construto/PRO é a autoeficácia relacionada à dor. Um dos PROMs mais famosos para avaliar esse construto é a Escala de Autoeficácia Relacionada à Dor (Pain Self-Efficacy Questionnaire – PSEQ)[4]. A utilização dos PROMs pode ser muito útil no funcionamento dos sistemas de saúde, uma vez que eles podem: 1) ser administrados em série (longitudinalmente) para monitorar o progresso dos pacientes e facilitar a identificação de problemas; 2) ajudar profissionais da saúde a realizar a prática centrada no paciente; 3) avaliar e comparar a eficiência e desempenho de práticas, processos e intervenções adotadas; e 4) fornecer dados para avaliar políticas implementadas nos serviços e nos sistemas de saúde[5].

Os PROMs podem ser classificados, quanto ao propósito para o qual os instrumentos foram desenvolvidos, em avaliativos, discriminativos e prognósticos[6]. Instrumentos com propósito avaliativo foram elaborados com o objetivo de acompanhar mudanças ao longo do tempo (pré e pós-intervenção)[6]. O instrumento SF-36 (Medical Outcomes Study 36 – Item Short – Form Health Survey)[7] para avaliação da qualidade de vida é um instrumento concebido para fins avaliativos, ou seja, o escore do SF-36 deve ser utilizado de forma longitudinal, para acompanhar mudanças ao longo do tempo. Já instrumentos com propósito discriminativo são aqueles que foram elaborados para discriminar subgrupos (ex: pacientes com diferentes graus de incapacidade)[6].

O PHQ-9 (Patient Health Questionnaire)[8], por exemplo, foi desenvolvido para detectar indivíduos com depressão e sem depressão. O valor de corte ≥10 pontos demonstrou alta sensibilidade e especificidade (88%) para diagnosticar indivíduos com depressão quando são avaliados em serviços de atenção primária à saúde[8]. Já os questionários com propósito prognóstico (ou preditivo)[6] têm por objetivo antecipar um prognóstico (ex: o curso futuro de uma doença). Um exemplo desse tipo de instrumento é o *StarT Back Screening Tool*[9] (SBST), que foi elaborado para detectar as chances de prognósticos desfavoráveis de recuperação (dor persistente) em pacientes com dor lombar na fase aguda ou subaguda. A literatura descreve que um escore ≥4 na escala psicossocial do SBST está relacionado a um diagnóstico desfavorável para pacientes com dor lombar aguda, com risco de cronificação da dor lombar[9].

Os questionários também podem ser classificados quanto à sua aplicabilidade em populações-alvo, em casos específicos ou genéricos. Questionários genéricos são aqueles cuja população-alvo é ampla. Por exemplo, o Inventário Breve de Dor[10] é um questionário que foi elaborado para pacientes com dor crônica em geral. Os questionários de incapacidade costumam ser específicos para uma determinada condição (condição-específica). Assim, o *Oswestry Disability Index* (ODI) é específico para avaliar a incapacidade relacionada à dor lombar[11].

Os questionários/escalas são elaborados com o objetivo de tentar transformar um aspecto subjetivo ou conceito (construto) em uma medida quantitativa. Foram atribuídos números aos escores dos questionários com o objetivo de acompanhar a evolução do paciente quanto ao atributo a ser mensurado. Nas situações em que não é possível mensurar um determinado atributo, como se deve proceder para estabelecer um parâmetro de melhora ou piora do paciente? Mensurar/medir é o procedimento de identificar valores de variáveis quantitativas a partir de sua relação numérica com outros valores[12]. Toda medida, para ser considerada adequada e ter aplicabilidade prática, precisa atender a algumas propriedades.

Por exemplo, a medida de temperatura corporal deve ser reprodutível quando avaliada duas vezes em um intervalo de tempo pequeno, em que esteja garantida a estabilidade do quadro clínico do paciente. Um exemplo hipotético seria o caso de um paciente que chegou ao hospital e sua temperatura foi aferida, obtendo-se o valor de 39º Celsius. Após um período curto (3 minutos), se a temperatura for novamente aferida, sem que o paciente tenha recebido um fármaco antitérmico ou sofrido qualquer oscilação do quadro clínico, é esperado que o termômetro seja capaz de obter a mesma temperatura ou uma temperatura muito próxima do valor inicial. Isso é o que se chama de confiabilidade. Se não for possível confiar em uma medida, como tomar uma decisão clínica de forma segura, baseando-se nessa medida?

Então, o que são as propriedades de medida? As propriedades de medida são obtidas a partir do estudo das características de uma determinada medida – por exemplo, estabelecendo relações/com-

parações do escore de um instrumento com o escore(s) de outro(s) instrumentos – com o intuito de identificar se a medida (ex: escore PROM ou OMI) tem qualidades adequadas. Ou seja: se a medida tem adequada consistência quando é repetida (confiabilidade), se a medida mede mesmo aquilo que pretende medir (validade) e se a medida é capaz de captar alterações ao longo do tempo (responsividade). Assim, as propriedades de medida ajudam a identificar a qualidade de um PROM ou OMI, e a qualidade geral das propriedades de medida de um instrumento pode ajudar clínicos e pesquisadores a tomar a decisão sobre qual PROM ou OMI deve ser utilizado em sua prática profissional.

O problema da falta de padronização dos termos na área de propriedades de medida e a importância da iniciativa COSMIN

COSMIN é um acrônimo de *Consensus-based Standards for the selection of health Measurement Instruments*[13-15]. Como o próprio nome já revela, o COSMIN é uma iniciativa para consenso e padronização de aspectos relacionados a propriedades de medida de PROMs[13-15]. Para a elaboração dos vários documentos disponíveis online na plataforma do COSMIN (https://www.cosmin.nl/cosmin-tools/), houve a colaboração de diversos especialistas de diferentes partes do mundo para se chegar a consensos quanto às várias definições propostas.

O comitê COSMIN foi inspirado por uma falta de clareza na literatura sobre a terminologia e as definições das propriedades de medida[16]. Existe uma quantidade impressionante de PROMs e muitos deles mensuram o mesmo construto[16]. Então, como definir qual o melhor PROM para utilizar em uma pesquisa ou na prática clínica? Auxiliar nesse processo de tomada de decisão é um dos objetivos centrais da iniciativa COSMIN.

O primeiro passo foi padronizar a taxonomia. No primeiro estudo COSMIN[17] do tipo *Delphi* (estudos que visam, através de painéis entre *experts*, definir consensos sobre diferentes temas) foi estabelecido um consenso sobre a terminologia (taxonomia) e definições dessas propriedades de medida[13-15]. A iniciativa COSMIN também tem como objetivo auxiliar na determinação de critérios de qualidade dessas propriedades[2]. Por exemplo, o que é aceitável como medida de confiabilidade? Quais testes estatísticos devem ser empregados para se medir a validade?

Além disso, a iniciativa COSMIN veio para ajudar pesquisadores a conduzir revisões sistemáticas de propriedades de medida[13] através, por exemplo, da disponibilização de ferramentas para avaliação da qualidade metodológica de estudos que investigam as propriedades de medida. As revisões sistemáticas de propriedades de medida ajudam a responder a seguinte pergunta: um determinado instrumento atende os critérios de qualidade das suas propriedades de medida e, dessa forma, pode ser utilizado na pesquisa e na prática clínica? Por exemplo, uma revisão sistemática[18] tentou determinar qual o melhor instrumento para medir incapacidade em indivíduos com dor lombar (Questionário de Incapacidade de Oswestry e Questionário de Roland-Morris). Essa revisão demonstrou que ambos os questionários têm limitações, e não foi possível determinar qual deles apresentou melhor qualidade de propriedades de medida. Assim, ambos os questionários foram recomendados pela literatura. Essa revisão sistemática também alertou sobre a necessidade de estudos de melhor qualidade metodológica sobre as propriedades de medida de ambos os questionários.

Na literatura, pode-se encontrar diversos relatos de pesquisadores e clínicos, na área de saúde e na área de dor, sobre a dificuldade de tomar decisões quanto à escolha de um PROM ou OMI, especialmente devido à dificuldade de escolher entre as inúmeras opções de PROMs disponíveis para avaliação, acompanhamento ou mesmo diagnóstico de pacientes, bem como à dificuldade de interpretar os estudos de propriedades de medida[19]. Tendo em vista esses aspectos, os objetivos da parte I desta revisão narrativa foram: (I) introduzir conceitos básicos sobre PROMs com enfoque na terminologia e critérios definidos pelo COSMIN e descrever o processo de adaptação transcultural, e (2) descrever as propriedades de medida do domínio confiabilidade. Já a parte II da revisão publicada em um segundo artigo, compreenderá o domínio das propriedades de medida de validade, responsividade e interpretabilidade, bem como a proposta de um *checklist* para avaliação da qualidade de PROMs.

MÉTODOS

A elaboração deste estudo foi baseada em estudos publicados pelo consenso COSMIN. Das 32 referências citadas nesse artigo, 10 são artigos da iniciativa COSMIN[2,6,13-17,22,23,31].

OS DOMÍNIOS DAS PROPRIEDADES DE MEDIDA SEGUNDO O COSMIN

A escolha de um PROM ou OMI para a avaliação de uma condição de saúde deve ser baseada prioritariamente na qualidade de suas propriedades de medida[13]. A qualidade de um PROM ou OMI deve ser avaliada, segundo o consenso COSMIN, por meio de três grandes domínios de análise: confiabilidade, validade e responsividade (Figura 1). O COSMIN também considera que a interpretabilidade é uma característica que deve ser considerada[14].

O domínio de confiabilidade de um PROM engloba as propriedades de medida que descrevem o "quanto a medida é livre de erro"[14]. Dentro do domínio de confiabilidade o COSMIN considera as seguintes propriedades de medida: (I) confiabilidade, (II) medida de erro e (III) consistência interna[14].

O domínio de validade de um instrumento reúne as propriedades de medida que tentam identificar se o instrumento "mensura aquilo que ele pretende medir"[2]. As seguintes propriedades de medida estão descritas nesse domínio, de acordo com o COSMIN: (I) validade de conteúdo, (II) validade estrutural, (III) teste de hipóteses, (IV) validade transcultural e validade de critério.

Já o domínio da responsividade reúne apenas uma propriedade de medida que tem o mesmo nome do domínio: responsividade. A responsividade está alinhada com a capacidade de um instrumento em detectar mudanças no escore (*change score*) de um PROM ou OMI ao longo do tempo[14] e de forma válida. É um tipo de validade (a validade da mudança do escore), que foi retirada do domínio da validade (pelo COSMIN) para evitar confusões.

Por fim, a interpretabilidade de um PROM está relacionada com a facilidade na interpretação e a atribuição de significado ao escore de um instrumento para sua aplicação na prática[15]. Ainda que não seja considerada uma propriedade de medida, a interpretabilidade é uma característica fundamental de instrumentos de medida, apesar de ser comumente negligenciada por pesquisadores. Assim como a

Figura 1. Diagrama dos domínios das propriedades de Medida de acordo com o *Consensus-based Standards for the selection of health Measurement Instrument* (COSMIN)

Disponível em https://cosmin.nl/wp-content/uploads/COSMIN_taxonomy.pdf.

interpretabilidade, a adaptação transcultural não é uma propriedade de medida, mas é um processo imprescindível para a disponibilização de instrumentos para novas línguas e garantir que os sistemas de avaliação possam ser internacionalmente intercambiáveis.

ADAPTAÇÃO TRANSCULTURAL

O termo "adaptação transcultural" é usado para descrever um processo que combina a tradução e a adaptação cultural no processo de preparação de um instrumento para uso em outro cenário. A maioria dos instrumentos encontrados na literatura foram desenvolvidos na língua inglesa. Assim, o processo de tradução e adaptação transcultural baseado em um método adequado pode garantir a equivalência de um instrumento traduzido em relação à versão alvo (idioma para o qual o instrumento foi traduzido)[20].

Um exemplo que ilustra a importância desse processo é o do instrumento *The Activities-Specific Balance Confidence Scale* (ou Escala ABC, ou Escala de Confiança no Equilíbrio Baseada em Atividades Específicas), que mensura a confiança no equilibrio do indivíduo para realizar atividades. Uma das questões da escala ABC é sobre a confiança do indivíduo em andar em calçadas cobertas com gelo ("walk outside on icy sidewalks"*).* Essa pergunta faz sentido em regiões de clima frio ou temperado, onde o fenômeno meteorológico de nevadas é comum. Entretanto, em países de clima tropical, como o Brasil, essa questão precisou passar por um processo de adaptação transcultural. Na versão português-Brasil[21], essa questão da escala ABC foi adaptada transculturalmente da seguinte forma: "andar em calçada molhada ou escorregadia".

O documento *COSMIN study design checklist for patient-reported outcome measurement instruments*[22] (https://www.cosmin.nl/wp-content/uploads/COSMIN-study-designing-checklist_final.pdf) descreve 12 itens recomendados para a realização do processo de tradução e adaptação transcultural:

1. A descrição da língua original e língua alvo da tradução;
2. O PROM ou OMI deve ser traduzido (forward) e retrotraduzido (backward). A tradução *forward* é a tradução da língua original para a língua alvo e *backward translation* é a tradução da versão na língua alvo de volta para a língua original;
3. Os tradutores da tradução *forward* devem ter como língua mãe a língua alvo da tradução;
4. Um dos tradutores da etapa de tradução *forward* deve ser *expert* sobre o construto a ser medido e o segundo tradutor *forward* deve ser leigo sobre o assunto;

5. Ambos os tradutores da retrotradução devem ter como língua mãe a língua original do PROM ou OMI;
6. Ambos os tradutores da retrotradução devem ser leigos para o construto a ser mensurado pelo PROM ou OMI;
7. É necessário assegurar que os tradutores trabalhem de forma independente;
8. É necessário descrever como as diferenças entre o PROM original e o PROM traduzido foram resolvidas;
9. Garantir que a versão final traduzida será revisada por um comitê que envolva inclusive os desenvolvedores do instrumento (autores da versão original);
10. Elaborar um documento relatando o processo de tradução/adaptação transcultural;
11. Realizar um estudo piloto em que a compreensão, a abrangência e a relevância sejam avaliadas, na população alvo do instrumento, para os seguintes aspectos: itens, instruções, opções de resposta e período para resgate de memória;
12. Realizar o estudo piloto em uma população que represente a população alvo do PROM ou OMI.

Seguir todas as 12 etapas descritas não garante que o instrumento traduzido está adequado para ser utilizado na prática clínica e pesquisa. Assim, após a tradução é recomendado que as propriedades de medida do instrumento sejam testadas[20], uma vez que o PROM ou OMI traduzido deve ser entendido como um novo PROM ou OMI. Dessa forma, nas próximas sessões do artigo (parte I e parte II) foram descritas as principais propriedades de medidas recomendadas pela iniciativa COSMIN.

DOMÍNIO DA CONFIABILIDADE

Confiabilidade

A confiabilidade é definida como a proporção de variação total nas medições que pode ser atribuída a diferenças verdadeiras entre pacientes[2]. A confiabilidade dos instrumentos avalia o grau em que medidas repetidas em diferentes momentos fornecerão respostas semelhantes, considerando indivíduos clinicamente estáveis. O instrumento deve ser capaz de distinguir a mínima mudança clinicamente importante (Minimal Important Change - MIC) do erro de medição[1]. Essa propriedade de medida deve ser obtida em estudos longitudinais nos quais duas aplicações do questionário devem ser realizadas (teste-reteste).

Para análise estatística da confiabilidade o COSMIN recomenda os seguintes testes estatísticos: Coeficiente de Correlação Intraclasse (ICC) ou Kappa ponderado ou Correlação de Spearman/Pearson ≥0,70[2]. O ICC é utilizado para medidas contínuas. Para escalas ordinais é utilizado o coeficiente de Kappa ponderado. A correlação de Spearman/Pearson também pode ser utilizada desde que se garanta um controle rigoroso do erro sistemático da medida nas situações de teste e reteste. Essas medidas são interpretadas da seguinte maneira: os resultados podem variar entre -1 e 1. Quanto mais próximo de um, maior a confiabilidade[23]. As correlações podem demonstrar valores positivos ou negativos, que são uma expressão de direção da correlação e não da sua magnitude.

Para mensurar a confiabilidade, é necessário um período entre as aplicações do instrumento, que deve ser longo o bastante para evitar o viés de memorização (*recall bias*) e ao mesmo tempo curto o bastante para garantir que não ocorra alteração da estabilidade do paciente. Um período comumente utilizado é de uma ou duas semanas. Garantir a estabilidade clínica do paciente é imprescindível para se mensurar a confiabilidade, uma vez que o objetivo é minimizar os efeitos do erro sistemático no escore do questionário (aquele ocasionado por fatores metodológicos não adequadamente controlados). Deve-se garantir também as mesmas condições no ambiente de coleta no teste e no reteste.

O *Western Ontario McMaster Osteoarthritis Index* (WOMAC)[24] avalia a dor e a incapacidade relacionada à osteoartrite de joelho e quadril. No exemplo da figura 2, considera-se apenas a escala de incapacidade (dentre as três escalas que compõem o instrumento): quanto maior o escore na escala de incapacidade do WOMAC, maior a incapacidade. O escore do WOMAC apresentou confiabilidade adequada, com valor de ICC = 0,99. Esse dado significa que o escore do questionário é consistente (ou muito similar) nas duas aplicações do questionário. Os autores descreveram um intervalo teste-reteste de apenas 24 horas para garantir a estabilidade do quadro clínico. Esse intervalo de tempo curto pode ser considerado um viés metodológico, já que pode não ser suficiente para minimizar o *recall bias*. Pode-se notar que o escore WOMC (*Western Ontario McMaster Osteoarthritis Index*) demonstrou boa confiabilidade nas duas aplicações do questionário, pois o escore se manteve similar quando a condição clínica não sofreu alterações. Alguma variação no escore do questionário é sempre esperada e o erro da medida vai ajudar a identificar qual a variação que é aceitável.

Medida de erro

Todo instrumento de medida apresenta um erro, que pode ser definido como erro sistemático ou aleatório. O erro sistemático é aquele relacionado a aspectos metodológicos[25], como a posição do paciente ou o ambiente da sala de avaliação (climatizada e iluminada). Já o erro aleatório é a parcela de erro que não pode ser controlada e pode ajudar a entender o efeito do acaso[26] no escore do PROM ou OMI. Assim, é importante quantificar o erro do escore de um instrumento, através de um rigor metodológico que

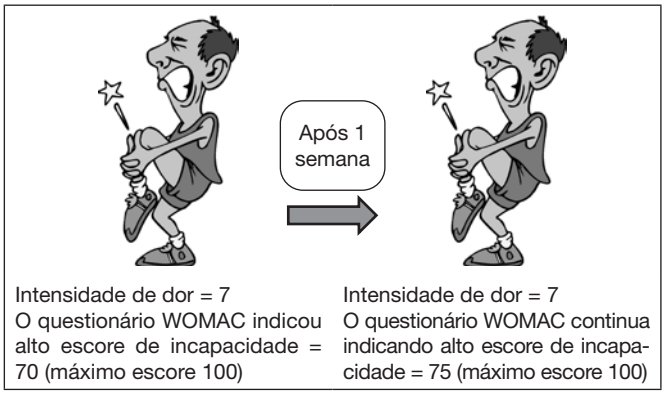

Figura 2. Ilustração do que seria uma confiabilidade adequada do escore de um PROM ou OMI. Pode-se notar que o escore do questionário WOMAC ((Western Ontario McMaster Osteoarthritis Index) demonstrou boa confiabilidade nas duas aplicações do questionário, pois o escore se manteve similar quando a condição clínica não sofreu alterações. Alguma variação no escore do questionário é sempre esperado e o erro da medida vai ajudar a identificar qual a variação que é aceitável

minimize o erro que pode ser controlado (sistemático), na tentativa de se obter apenas o que é erro aleatório (que não é possível de predizer ou controlar).

A medida de erro ajuda pesquisadores e clínicos a julgarem se a mudança no escore do PROM ou OMI reflete de fato uma mudança no quadro clínico ou se pode ser considerada um erro. A medida de erro é importante na interpretação do escore de instrumentos com propósito avaliativo, ou seja, ao se aplicar um questionário pré e pós-tratamento, com o valor da medida de erro é possível determinar se a alteração no escore do PROM ou OMI para um dado paciente é clinicamente relevante (quando o valor da mudança é maior que a medida de erro) ou se a mudança no escore pode ser apenas atribuída ao erro do escore (quando o valor da mudança é menor que a medida de erro).

Para calcular a medida de erro a Mínima Diferença Detectável (Smallest Detectable Change - SDC) é a mais recomendada. A SDC depende da Medida de Erro Padrão (Standard Error of Measurement - SEM) e pode ser calculada através da fórmula $SDC = 1.96 \times \sqrt{2} \times SEM$[23].

A SDC reflete a mínima mudança intra-sujeito (ou seja, teste-reteste) no escore do PROM ou OMI que pode ser interpretada como mudança real e acima do erro para um dado indivíduo[17]. Além disso, a medida de erro também pode ser expressa pelo gráfico de Limites de Concordância (*Limits of Agreement* - LoA)[27], que fornece a medida de erro através de um gráfico de dispersão, facilitando sua visualização. Para fins de avaliação, o SDC ou LoA deve ser < MIC do escore do PROM ou OMI[2].

Mas o que é a MIC? É a sigla para Mínima Mudança Clinicamente Importante. No contexto dos PROMs ou OMIs é o valor mínimo que identifica a mínima mudança no escore do PROM ou OMI ao longo do tempo, que o paciente reconhece como importante[15]. Ou seja, é o valor de corte do PROM ou OMI, que pode ser obtido através de comparações com uma escala de percepção global de mudança/melhora, que avalia a percepção relatada pelo paciente sobre sua condição (figura 3).

O paciente deve indicar o valor numérico que ele/ela entende que expressa a melhora ou piora global da sua dor. Valores negativos indicam piora e valores positivos indicam melhora da dor.

Em um exemplo, considerando que a Escala Numérica de Dor (END) foi utilizada para a avaliação da intensidade da dor de um indivíduo com dor lombar[28] (os seguintes valores foram obtidos: pré-tratamento = 8 e pós-tratamento = 4), a SDC para a END em indivíduos com dor lombar variou de 2,4 a 3,5 pontos. Se o paciente apresentou uma variação de 4 pontos pós-tratamento (8 - 4 = 4), será que a mudança de 4 pontos na intensidade de dor do paciente pode ser considerada mudança ou simplesmente erro? Considerando que a mudança foi maior que 3,5, então a resposta é: a mudança do escore (pré e pós-tratamento) é maior que o erro e, portanto, não pode ser atribuída à medida de erro.

Consistência interna

A consistência interna é definida pelo COSMIN como "o grau de inter-relação entre os itens" de um PROM ou OMI[2,16]. A consistência interna contemplada no domínio de confiabilidade considera a extensão em que os itens avaliam um mesmo construto, ou seja, a correlação (homogeneidade) entre os itens. Essa correlação deve ser alta o suficiente para que os itens de um PROM ou OMI representem um mesmo construto a ser medido. A consistência interna é uma propriedade de medida importante para questionários que pretendem medir um único conceito (construto baseado em modelo reflexivo) utilizando vários itens que o representam, refletindo esse construto (figura 4). Para escalas multidimensionais a consistência interna deve ser avaliada para cada domínio ou subescala, partindo-se do pressuposto de que cada domínio/fator/subescala está avaliando construtos diferentes[2].

As setas saindo do construto (depressão) indicam que cada questão do PROM é um reflexo do construto. Nota-se que as imagens periféricas (itens do questionário) são muito parecidas com a imagem central (construto), mas diferentes. Por isso são consideradas um "reflexo" não idêntico do construto.

Em contraste, para questionários baseados em um modelo formativo, em que cada item pode representar construtos diferentes, e que muitas vezes não estabelecem correlação entre si, a análise de consistência interna não é indicada. A consistência interna deve ser avaliada apenas para escalas unidimensionais (ou para as subescalas de PROMs ou OMIs multidimensionais) e baseadas em modelo reflexivo. A Escala de Apgar[29], que avalia a vitalidade do recém-nascido e que contempla os domínios: (I) tônus, (II) respiração, (III) coloração, (IV) frequência cardíaca e (V) reflexos, é um exemplo de escala baseada no modelo formativo, em que seus itens contribuem para formar um construto (vitalidade do bebê), mas não apresentam correlação entre si, sendo que a análise de consistência interna não se aplica. Os construtos baseados em um modelo formativo são considerados "construtos artificiais", como também é o caso do construto qualidade de vida. Há bastante controvérsia sobre esse tema na literatura. Para um maior aprofundamento no tema, esta pesquisa recomenda um dos artigos aferidos[30].

O coeficiente α de Cronbach é a medida estatística indicada para estimar a consistência interna dos PROMs ou OMIs[5]. O COSMIN considera que a qualidade de propriedade de medida está adequada para a consistência interna quando: (I) pelo menos alguma evidência de adequada validade estrutural está disponível e (II) α de Cronbach ≥ 0.70 descrito para cada fator/domínio/dimensão[31].

Comparado a quando este episódio de dor_____ começou, como você descreveria sua dor_____ nos últimos dias?

-5	-4	-3	-2	-1	0	1	2	3	4	5
Extremamente pior					Sem modificação					Completamente recuperado

Figura 3. Escala de percepção global de melhora. O paciente deve indicar o valor numérico que ele/ela entende que expressa a melhora ou piora global da sua dor. Valores negativos indicam piora e valores positivos indicam melhora da dor.

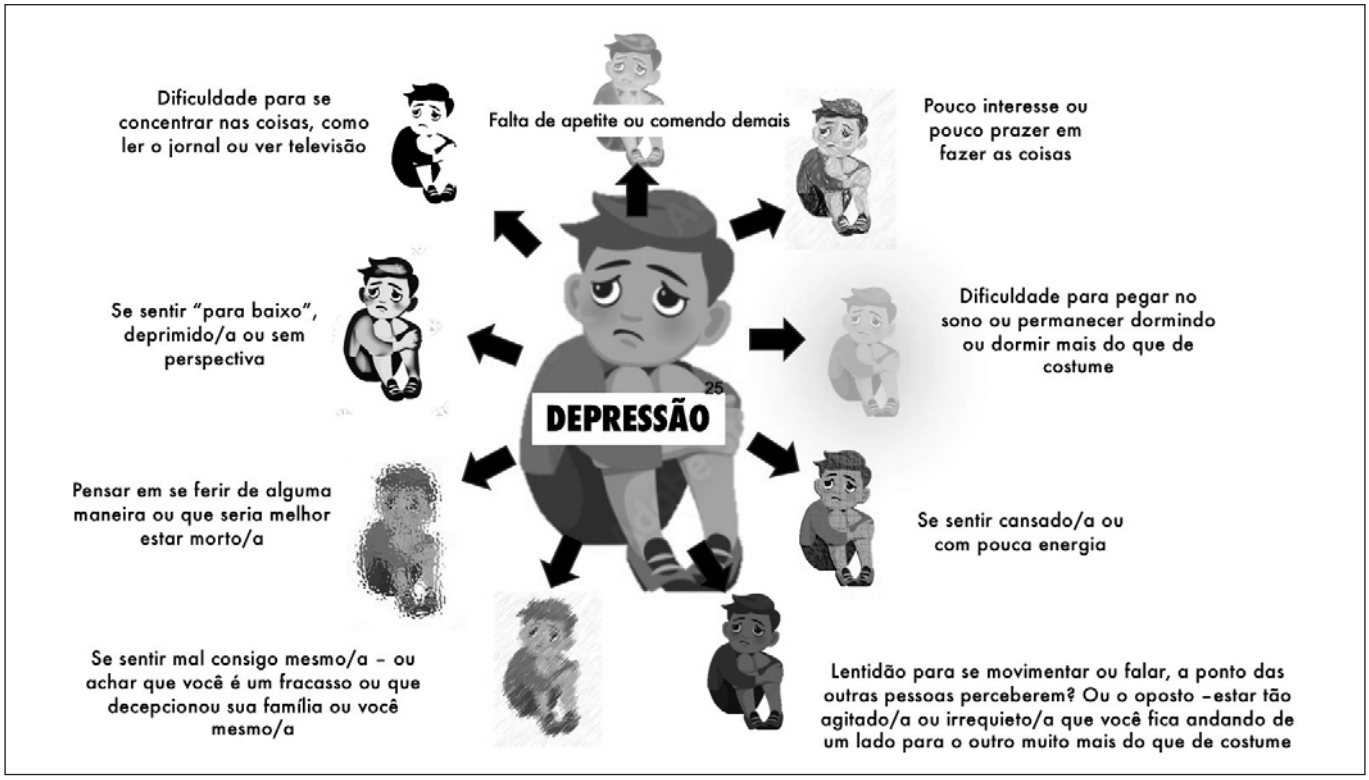

Figura 4. Ilustração dos itens do Questionário PHQ-9 que avalia sintomas de depressão e é considerado um construto baseado em um modelo reflexivo

No manuscrito que descreve a validação para o português Brasil do *Craniofacial Pain and Disability Inventory*[32], é possível encontrar um exemplo prático da avaliação do coeficiente alfa de Cronbach. Essa pesquisa indicou (separadamente, para cada fator/domínio/dimensão a seguir): "Limitação Funcional e Psicossocial", $\alpha = 0,86$; "Dor", $\alpha = 0,80$; e "Frequência de Comorbidades", $\alpha = 0,77$; conforme recomendado pelo COSMIN, com os coeficientes sendo considerados adequados ($\alpha \geq 0,70$). Portanto, pode-se concluir que os itens contidos em cada fator/domínio/dimensão, possuem uma boa correlação entre si e representam adequadamente a dimensão que se propõem a medir.

CONCLUSÃO

O presente estudo (parte I) buscou elucidar conceitos básicos sobre os questionários, trazendo aspectos sobre a importância da iniciativa COSMIN para o processo de adaptação transcultural e sobre as propriedades de medida dentro do domínio da confiabilidade. Assim, a adaptação transcultural deve seguir um método que garanta equivalência do PROM traduzido em relação à versão original, e um instrumento com qualidade adequada de confiabilidade deverá atender a alguns critérios, tais como: ICC ≥ 0,70; medida de erro < mínima mudança clinicamente importante; e α de Cronbach ≥ 0.70. A leitura deste artigo deve ser complementada pela leitura da parte II, na qual foram abordadas as propriedades de medida de validade, responsividade e interpretabilidade, bem como a proposta de um *checklist* para apoio à tomada de decisão na escolha de um PROM adequado.

CONTRIBUIÇÕES DOS AUTORES

Thaís Cristina Chaves
Conceitualização, Gerenciamento de Recursos, Gerenciamento do Projeto, Metodologia, Redação - Preparação do Original, Redação - Revisão e Edição, Supervisão
Ana Carolina de Jacomo Claudio
Metodologia, Redação - Preparação do Original, Redação - Revisão e Edição
Thamiris Costa Lima
Redação - Preparação do Original, Redação - Revisão e Edição
Roger Berg Rodrigues Pereira
Metodologia, Redação - Preparação do Original, Redação - Revisão e Edição
Gabriela Zuelli Martins Silva
Redação - preparação do original, Redação - Revisão e Edição
Helen Cristina Nogueira Carrer
Redação - preparação do original, Redação - Revisão e Edição

REFERÊNCIAS

1. Øvretveit J, Zubkoff L, Nelson EC, Frampton S, Knudsen JL, Zimlichman E. Using patient-reported outcome measurement to improve patient care. Int J Qual Health Care. 2017;29(6):874-9.
2. Elsman EBM, Mokkink LB, Langendoen-Gort M, Rutters F, Beulens J, Elders PJM, Terwee CB. Systematic review on the measurement properties of diabetes-specific patient-reported outcome measures (PROMs) for measuring physical functioning in people with type 2 diabetes. BMJ Open Diabetes Res Care. 2022;10(3):e002729.
3. Davidson M, Keating J. Patient-reported outcome measures (PROMs): how should I interpret reports of measurement properties? A practical guide for clinicians and researchers who are not biostatisticians. Br J Sports Med. 2014;48(9):792-6.

4. Sleijser-Koehorst MLS, Bijker L, Cuijpers P, Scholten-Peeters GGM, Coppieters MW. Preferred self-administered questionnaires to assess fear of movement, coping, self-efficacy, and catastrophizing in patients with musculoskeletal pain-A modified Delphi study. Pain. 2019;160(3):600-6.
5. Black N. Patient reported outcome measures could help transform healthcare. BMJ. 2013;346:f167.
6. De Vet HCW, Terwee CB, Mokkink LB, Knol DL. Measurement in Medicine - A practical guide. 1st edition. New York: Cambridge University Press; 2011.
7. Ware JE, Sherbourne CD. The MOS 36-Item Short-Form Health Survey (SF-36®): I. conceptual framework and item selection. Med Care 1992; 30(6):473-83.
8. Kroenke K, Spitzer RL, Williams JB. The PHQ-9: validity of a brief depression severity measure. J Gen Intern Med. 2001;16(9):606-13.
9. Hill JC, Dunn KM, Lewis M, Mullis R, Main CJ, Foster NE, Hay EM. A primary care back pain screening tool: identifying patient subgroups for initial treatment. Arthritis Rheum. 2008;59(5):632-41.
10. Cleeland CS, Ryan KM. Pain assessment: global use of the Brief Pain Inventory. Ann Acad Med. Singapore. 1994;23(2):129-38.
11. Vigatto R, Alexandre NM, Correa Filho HR. Development of a Brazilian Portuguese version of the Oswestry Disability Index: cross-cultural adaptation, reliability, and validity. Spine (Phila Pa 1976). 2007;32(4):481-6.
12. Michell J. An Introduction to the Logic of Psychological Measurement. Hillsdale, NJ: Lawrence Erlbaum Associates. 1990. 190p.
13. Prinsen CAC, Mokkink LB, Bouter LM, Alonso J, Patrick DL, de Vet HCW, Terwee CB. COSMIN guideline for systematic reviews of patient-reported outcome measures. Qual Life Res. 2018;27(5):1147-57.
14. Mokkink LB, de Vet HCW, Prinsen CAC, Patrick DL, Alonso J, Bouter LM, Terwee CB. COSMIN risk of bias checklist for systematic reviews of patient-reported outcome measures. Qual Life Res. 2018 May;27(5):1171-9.
15. Terwee CB, Prinsen CAC, Chiarotto A, Westerman MJ, Patrick DL, Alonso J, Bouter LM, de Vet HCW, Mokkink LB. COSMIN methodology for evaluating the content validity of patient-reported outcome measures: a Delphi study. Qual Life Res. 2018;27(5):1159-1170.
16. 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. 2016;20(2):105-13.
17. Mokkink LB, Terwee CB, Patrick DL, Alonso J, Stratford PW, Knol DL, Bouter LM, de Vet HC. The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes. J Clin Epidemiol. 2010;63(7):737-45.
18. Chiarotto A, Ostelo RW, Boers M, Terwee CB. A systematic review highlights the need to investigate the content validity of patient-reported outcome measures for physical functioning in patients with low back pain. J Clin Epidemiol. 2018;95:73-93.
19. Swinkels RA, van Peppen RP, Wittink H, Custers JW, Beurskens AJ. Current use and barriers and facilitators for implementation of standardised measures in physical therapy in the Netherlands. BMC Musculoskelet Disord. 2011;22;12:106.
20. Beaton, Dorcas E. BScOT, MSc, PhD; Bombardier, Claire MD, FRCP; Guillemin, Francis MD, MSc; Ferraz, Marcos Bosi MD, MSc, PhD. Guidelines for the Process of Cross-Cultural Adaptation of Self-Report Measures. Spine. 2000;25:3186-91.
21. Marques AP, Mendes YC, Taddei U, Pereira CA, Assumpção A. Brazilian-Portuguese translation and crosscultural adaptation of the activities-specific balance confidence (ABC) scale. Braz J Phys Ther. 2013; 17(2):170-8.
22. Mokkink LB, Prinsen CAC, Patrick DL, Alonso J, Bouter LM, de Vet HCW, Terwee CB. COSMIN Study Design checklist for Patient-reported outcome measurement instruments. Documento disponível em https://www.cosmin.nl/wp-content/uploads/COSMIN-study-designing-checklist_final.pdf Acesso em 7/10/2023.
23. Terwee CB, Bot SD, de Boer MR, van der Windt DA, Knol DL, Dekker J, Bouter LM, de Vet HC. Quality criteria were proposed for measurement properties of health status questionnaires. J Clin Epidemiol. 2007;60(1):34-42.
24. Kottner J, Gajewski BJ, Streiner DL. Guidelines for Reporting Reliability and Agreement Studies (GRRAS). Int J Nurs Stud. 2011;48(6):659-60.
25. Bellamy N, Buchanan WW, Goldsmith CH, Campbell J, Stitt LW. Validation study of WOMAC: a health status instrument for measuring clinically important patient relevant outcomes to antirheumatic drug therapy in patients with osteoarthritis of the hip or knee. J Rheumatol. 1988;15:1833-40.
26. Barraza F, Arancibia M, Madrid E, Papuzinski C. General concepts in biostatistics and clinical epidemiology: Random error and systematic error. Medwave. 2019;19(7):e7687.
27. Altman DG, Bland JM. Measurement in Medicine: The analysis of method comparison studies. Statistician. 1983;32(3):307-17
28. Chiarotto A, Maxwell LJ, Ostelo RW, Boers M, Tugwell P, Terwee CB. Measurement Properties of Visual Analogue Scale, Numeric Rating Scale, and Pain Severity Subscale of the Brief Pain Inventory in patients with low back pain: a systematic review. J Pain. 2019;20(3):245-63.
29. Apgar V. A proposal for a new method of evaluation of newborn infants. Anesth Analg. 1953;32:260-7.
30. Guyon H. The fallacy of the theoretical meaning of formative constructs. Front Psychol. 2018;15;9:179.
31. Prinsen CA, Vohra S, Rose MR, Boers M, Tugwell P, Clarke M, Williamson PR, Terwee CB. How to select outcome measurement instruments for outcomes included in a "Core Outcome Set" - a practical guideline. Trials. 2016;17(1):449.
32. Greghi SM, Dos Santos Aguiar A, Bataglion C, Ferracini GN, La Touche R, Chaves TC. Brazilian Portuguese Version of the Craniofacial Pain and Disability Inventory: Cross-Cultural Reliability, Internal Consistency, and Construct and Structural Validity. J Oral Facial Pain Headache. 2018;32(4):389-99.

This is an open-access article distributed under the terms of the Creative Commons Attribution License.