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 II: validade, responsividade, interpretabilidade e *checklist* para caracterização da qualidade dos instrumentos

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 II: validity, responsiveness, interpretability and a checklist for characterizing the quality of instruments

Thaís Cristina Chaves[1], Thamiris Costa de Lima[2], Juliana H. Padilha Spavieri[2], Ana Carolina de Jacomo Claudio[2], Roger Berg Rodrigues Pereira[2], Mariana Romano de Lira[3]

DOI 10.5935/2595-0118.20230092-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 tomada de 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 II) 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 dos domínios validade, responsividade e interpretabilidade e propor um *checklist* para avaliação da qualidade das propriedades de medida de PROMs.

MÉTODOS: Utilizando uma busca voltada para os artigos da iniciativa COSMIN, foi elaborado o presente estudo de revisão, que foi dividido em duas partes para fins didáticos.

RESULTADOS: O presente artigo compreendeu a descrição das propriedades de medida dos domínios de validade (conteúdo, estrutural, construto), responsividade (deve ser avaliada através de análises de acurácia, AUC≥0,70) e interpretabilidade (que fornece a mínima mudança clinicamente importante). Além disso, foi proposto um *checklist* para determinação da qualidade das propriedades de medida de instrumentos de avaliação.

CONCLUSÃO: Este estudo descreveu as propriedades de medida dentro dos domínios validade e responsividade, e a importância da interpretabilidade para a obtenção da mínima diferença clinicamente importante. O *checklist* proposto para avaliação dessas propriedades pode auxiliar clínicos e pesquisadores a determinarem a qualidade de um instrumento e tomar a decisão sobre a melhor opção disponível.

Descritores: Confiabilidade, Dor crônica, Dor musculoesquelética, Inquéritos e questionários, Psicometria.

Thaís Cristina Chaves – https://orcid.org/0000-0002-6222-4961;
Thamiris Costa de Lima – https://orcid.org/0000-0002-7371-6232;
Juliana H. Padilha Spavieri – https://orcid.org/0000-0001-9653-0986;
Ana Carolina de Jacomo Claudio – https://orcid.org/0000-0001-7694-2836;
Roger Berg Rodrigues Pereira – https://orcid.org/0009-0009-2607-5629;
Mariana Romano de Lira – https://orcid.org/0000-0003-4032-5689.

1. Universidade Federal de São Carlos, Departamento de Fisioterapia, Programa de Pós-Graduação em Reabilitação e Desempenho Funcional, Departamento de Ciências da Saúde, Faculdade de Medicina de Ribeirão Preto, São Carlos, SP, Brasil.
2. Universidade Federal de São Carlos, Faculdade de Medicina de Ribeirão Preto, Programa de Pós-Graduação em Fisioterapia, Departamento de Fisioterapia, São Carlos, SP, Brasil.
3. Universidade de São Paulo, Faculdade de Medicina de Ribeirão Preto, Departamento de Ciências da Saúde, Programa de Pós-Graduação em Reabilitação e Desempenho Funcional, Ribeirão Preto, 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
1. Dentro do *domínio validade* é possível analisar: validade estrutural e validade de construto-teste de hipóteses
2. A *responsividade* é a capacidade de um instrumento de detectar mudanças ao longo do tempo
3. *A interpretabilidade* é a capacidade de extrair significado dos resultados obtidos por PROMs.
4. A mínima mudança clinicamente importante (MIC) é umas das medidas de interpretabilidade dos PROMs.
5. Um checklist com 20 itens foi proposto para auxiliar na determinação do PROM com melhor qualidade.

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

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

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 study (part II) were: (I) to introduce basic concepts about PROMs with a focus on the terminology and criteria defined by the COnsensus-based Standards for the selection of health Measurement INstruments (COSMIN) and (2) to describe the measurement properties of the validity, responsiveness and interpretability domains and propose a checklist for assessing the quality of PROMs' measurement properties.
METHODS: This study was produced using a search for articles from the COSMIN initiative. For didactic purposes, the text was divided into two parts.
RESULTS: This article included a description of the measurement properties of the validity (content, structural, construct), responsiveness (must be assessed through accuracy analyses, AUC≥0.70) and interpretability (which provides the minimum clinically important change) domains. In addition, a checklist was proposed for determining the quality of the measurement properties of assessment instruments.
CONCLUSION: This study described the measurement properties within the validity and responsiveness domains, and the importance of interpretability for obtaining the minimum clinically important difference. The proposed checklist for evaluating these properties can help clinicians and researchers to determine the quality of an instrument and make a decision about the best option available.
Keywords: Chronic pain, Psychometrics, Musculoskeletal pain, Reliability, Surveys and questionnaires,

INTRODUÇÃO

PROM (Patient Reported Outcome Measure) é a sigla em inglês para 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: Cinesiofobia ou medo de movimento)[3]. Existem inúmeros PROMs ou OMIs disponíveis na literatura, entretanto um dos grandes desafios para clínicos e pesquisadores é definir qual instrumento disponível é o mais adequado[4]. O entendimento das propriedades de medida de um PROM ou OMI pode auxiliar os clínicos e pesquisadores a tomarem uma decisão sobre qual instrumento utilizar. Assim, PROMs ou OMIs cuja maioria das propriedades de medida foram testadas e cujas propriedades atenderam aos critérios de qualidade descritos por iniciativas internacionais (como o COnsensus-based Standards for the selection of health Measurement INstruments - COSMIN)[5,6] devem ser preferencialmente utilizados.

Como descrito na parte I desta série de dois artigos, as propriedades de medida são obtidas a partir do estudo das características de uma determinada medida (por exemplo, estabelecendo relações/comparações do escore de um instrumento com o(s) escore(s) de outro(s) instrumentos), com o intuito de identificar se a medida (ex: escore do PROM ou OMI) tem qualidades adequadas. Na parte I, foram apresentadas as propriedades de medida do domínio confiabilidade. Na parte II foram descritas as propriedades dentro dos domínios de validade, responsividade e interpretabilidade.
O domínio da 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 de detectar mudanças no escore (*change score*) de um PROM ou OMI ao longo do tempo[7] de forma válida. É um tipo de validade (a validade da mudança do escore), que foi retirada do domínio 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[8]. 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.
Considerando a dificuldade de operacionalizar o conhecimento sobre as propriedades de medida, a proposta de um roteiro ou checklist pode auxiliar a reunir as informações necessárias para auxiliar os profissionais e pesquisadores a tomar uma decisão sobre a escolha dos PROMs mais adequados, relativa à qualidade de suas propriedades de medida. Assim, contribuindo para a translação do conhecimento científico na prática.
Considerando esses desafios, os objetivos da parte II desta revisão narrativa (dividida didaticamente em duas partes) foram: (I) descrever as principais propriedades de medida dos domínios validade e responsividade, (2) descrever a interpretabilidade de PROMs e OMIs, e (3) disponibilizar um *checklist* que ao ser preenchido pode auxiliar pesquisadores e clínicos a operacionalizar/reunir as informações sobre a qualidade das propriedades de medida de PROMs disponíveis na literatura e, dessa forma, facilitar o processo de tomada de decisão.

MÉTODOS

A elaboração deste estudo foi baseada em estudos publicados pelo consenso COSMIN. Das 31 referências citadas neste artigo, 10 são artigos da iniciativa COSMIN[2,5-8,13-15,21,25].

DOMÍNIO DE VALIDADE

Validade de conteúdo

A validade de conteúdo é definida pela evidência empírica (subjetiva) que demonstra que os itens e domínios de um instrumento

são apropriados e abrangentes em relação aos conceitos de medição, população e ao uso pretendido[9]. Para tal, é importante que o construto a ser avaliado seja bem definido e interpretável. As perguntas do instrumento devem ser elaboradas de tal forma que possam captar adequadamente a percepção das pessoas sobre o construto. Além disso, uma definição precisa e fundamentada do construto deve embasar a criação dos itens do instrumento. Um ponto-chave do desenvolvimento de um instrumento é a definição clara do construto a ser mensurado, e a construção de um modelo conceitual pode ser de grande valia na determinação das perguntas/itens que devem ser incluídos no PROM ou OMI[9].

Por exemplo, a *Lower Extremity Functional Scale* (LEFS) foi desenvolvida para avaliar especificamente a funcionalidade relacionada aos membros inferiores[10]. Suas questões contemplam apenas atividades funcionais que envolvem os membros inferiores e sua escala foi projetada para avaliar a funcionalidade. Assim, quanto maior o escore LEFS, maior a funcionalidade do paciente com disfunções do membro inferior. Questionários como o *Neck Disability Index* (NDI), para avaliação de incapacidade relacionada à dor cervical, cujo construto é incapacidade, mas incluem questões sobre intensidade da dor e intensidade da dor de cabeça, apresentam limitações relativas ao seu conteúdo. As perguntas que devem ser feitas nesse contexto são: qual a definição de incapacidade considerada pelos autores? Intensidade da dor é um construto que deve ser incluído em um questionário que pretende medir a incapacidade? Em um dos artigos verificados os autores indicaram que o construto do NDI é "mensurar a limitação de atividades devido à dor cervical"[11]. Entretanto, a intensidade da dor seria uma atividade?

A primeira etapa para criação de um PROM ou OMI é o desenvolvimento do instrumento. É comum o emprego de estudos qualitativos (grupos focais) para realizar a fase de "emergir o conteúdo" ou de geração do conteúdo[9]. É de suma importância que o público-alvo seja envolvido nessa etapa e os participantes descrevam qual o conteúdo que deve ser incluído no instrumento. Ao final do processo de geração do conteúdo, uma versão "*draft*" do questionário pode ser criada, e essa versão deve ser avaliada novamente pelo público-alvo da ferramenta[9,12].

Essa etapa pode ser considerada como validade de conteúdo propriamente dita e deve preferencialmente envolver a participação do público-alvo pretendido pelo PROM ou OMI e de *experts*. Essa etapa pode ser realizada através de estudos tipo *Delphi* ou através de estudos qualitativos. Três aspectos devem ser considerados nessa etapa: compreensão, abrangência e relevância dos itens do OMI[8]. O COSMIN descreve, em um dos seus artigos, um critério de 10 itens (Tabela 1) para nortear a avaliação da qualidade da validade de conteúdo de um PROM ou OMI[8]. Para instrumentos traduzidos, a validade de conteúdo não costuma ser descrita na literatura, já que o conteúdo de um instrumento não pode ser modificado no processo de tradução, apenas adaptado culturalmente sem prejuízos da equivalência em relação à versão original.

Tabela 1. Critério de 10 itens para avaliação da qualidade da validade de conteúdo sugerido pelo *COnsensus-based Standards for the selection of health Measurement INstruments* (COSMIN)

Relevância
1. Os itens/questões incluídos são relevantes para o construto de interesse?
2. Os itens/questões incluídos são relevantes para a população alvo?
3. Os itens/questões incluídos são relevantes para o contexto de aplicação do PROM ou OMI?
4. As opções de resposta são apropriadas?
5. O período de resgate de memória pode ser considerado apropriado?
Abrangência
6. Existe algum conceito importante para avaliação do construto faltando no PROM ou OMI?
Compreensão
7. As instruções do PROM ou OMI são fáceis de serem entendidas pela população alvo do PROM ou OMI?
8. As opções de resposta do PROM ou OMI são fáceis de serem entendidas pela população alvo do PROM ou OMI?
9. As perguntas ou itens do PROM ou OMI estão adequadamente redigidas?
10. As opções de resposta estão alinhadas (condizentes) com as questões formuladas?

PROM = *Patient Reported Outcome Measure*, OMI = *Outcome Measurement Instrument.*

Validade estrutural

A validade estrutural é definida como "o grau em que as pontuações de um instrumento de medida são um reflexo adequado da dimensionalidade do construto a ser medido"[13]. Dessa forma, a validade estrutural avalia quantos fatores ou domínios estão presentes em um instrumento, e quais itens fazem parte de cada dimensão/domínio/fator. Sendo assim, a validade estrutural pode definir a dimensionalidade de um instrumento. A identificação das dimensões é importante não apenas para determinar como o escore do PROM ou OMI será obtido, mas também para a interpretação dos resultados[7]. Questionários multidimensionais devem apresentar sistemas de pontuação separados para cada domínio, tornando a interpretação e a tomada de decisão clínica mais precisa do que quando utilizado apenas o escore total de um instrumento[13].

Considerando a Teoria Clássica dos Testes, a Análise Fatorial (AF), baseada na correlação dos itens, é o método mais utilizado para determinar a dimensionalidade dos PROMs ou OMIs[13]. O princípio básico é que itens altamente correlacionados entre si são agrupados em um mesmo fator/domínio, enquanto que itens que apresentam baixa correlação entre si, carregam em domínios diferentes, ou seja, itens pertencentes a fatores distintos se correlacionam em menor proporção[13]. Um questionário com 3 domínios, por exemplo, deve ter o seu modelo de 3 fatores confirmado pela análise fatorial.

A Análise Fatorial Exploratória (AFE) geralmente é aplicada se não houver ideias claras sobre o número de dimensões de uma escala; é um método pouco robusto e indicado apenas para gerar uma teoria prévia para confirmação *a posteriori.* Portanto, deve preferencialmente ser utilizada na fase de desenvolvimento de um instrumento[13].

A Análise Fatorial Confirmatória (AFC) é recomendada se hipóteses *a priori* sobre as dimensões do construto estiverem disponíveis, com base na teoria ou em análises anteriores. Portanto, para fins de validação, a AFC é mais robusta e recomendada pelo COSMIN[13,14]. Na AFC, os parâmetros de ajuste são utilizados para testar se os dados se ajustam à estrutura fatorial hipotética. O COSMIN considera como qualidade adequada de validade estrutural se a análise aten-

der os seguintes critérios: (I) Índice de Ajuste Comparativo (Comparative Fit Index - CFI) ou *Tucker-Lewis Index* (TLI) ou medida comparável forem > 0,95; e (II) Raiz do Erro Quadrático Médio de Aproximação (Root Mean Square Error of Approximation - RMSEA) < 0,06; ou (III) Raiz Quadrada Média Residual Padronizada (Standardized Root Mean Square Residual - SRMR) < 0,08[2].

Considerando a Teoria de Resposta ao Item, a análise Rasch pode ser utilizada como modelo matemático para a avaliação de questionários unidimensionais, ou seja, verificar se os itens de uma escala que representam um construto são representados por uma única dimensão[2]. O COSMIN faz uma descrição detalhada dos critérios de qualidade que devem ser considerados para a Análise Rasch: ausência de violação da unidimensionalidade, da independência local e da monotocidade, e ajuste do modelo adequado (ex: infit e outfit entre 0,5 e 1,5)[15].

A validade estrutural da Escala Tampa de Cinesiofobia para Disfunção Temporomandibular (TSK-TMD/Br) traduzida para o português Brasil, foi verificada através de uma AFC que confirmou a estrutura de dois fatores demonstrada para a versão original da escala em inglês, com o CFI = 0,97, o que atende aos critérios de qualidade adequada de validade estrutural de acordo com o COSMIN (Figura 1). As questões 1, 2, 10, 15, 17 e 18 se enquadram no domínio "Evitando atividade" (AA) e as questões 8-12 se enquadram no domínio "Foco somático" (SF). A figura 1 ilustra a estrutura da TSK-TMD/Br.

VALIDADE DE CONSTRUTO-TESTE DE HIPÓTESES

A validade de construto é o grau em que os escores de um PROM ou OMI são consistentes com hipóteses com base na suposição de que o PROM ou OMI mede o construto que se propõe a medir[2]. Para se avaliar a validade de construto, deve-se formular hipóteses sobre como os escores de um instrumento se relacionam com outros instrumentos que medem construtos semelhantes ou diferentes, incluindo não apenas a direção, mas também a magnitude das correlações. Portanto, para testar essas hipóteses é utilizada a ferramenta de teste de hipóteses[13].

Essas hipóteses podem ser realizadas por correlações internas e correlações externas. As internas são comparações entre os escores dos domínios de um determinado PROM ou OMI. Já as correlações externas são comparações entre diferentes PROMs ou OMIs (que medem o mesmo construto ou não). Há também a possibilidade de correlações entre hipóteses de diferenças entre os escores obtidos para definição de grupos relevantes (ex: quando o escore de um instrumento é capaz de diferenciar grupos de acordo com os níveis de incapacidade). Por isso, a recomendação atual é que a validade de construto seja denominada Validade de Construto–Teste de Hipóteses[13]. Um possível teste estatístico para verificação da Validade de Construto–Teste de Hipóteses é o teste de correlação de Pearson ou Spearman.

Na tabela 2 são mostradas perguntas norteadoras que podem ajudar a direcionar a construção das hipóteses para Validade de Construto–Teste de Hipóteses. Por exemplo, se dois questionários mensuram o construto "percepção de incapacidade relacionada à dor lombar", como o Questionário de Incapacidade de Roland-Morris (RMDQ)[17] e o Índice de Incapacidade de Oswestry (ODI)[18], pode-se esperar uma correlação entre os escores das escalas, pois ambas medem o mesmo construto.

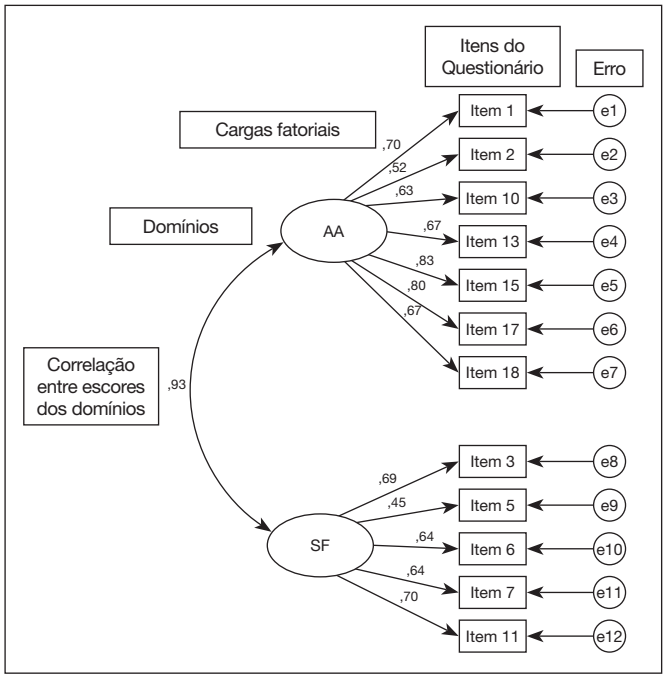

Figura 1. Diagrama representando o modelo testado através da Análise Fatorial Confirmatória da Escala Tampa de Cinesiofobia para Disfunção Temporomandibular. Adaptado[16]

AA = domínio Evitando Atividade; SF= domínio Foco Somático

Tabela 2. Perguntas norteadoras para a definição das hipóteses da Validade de Construto-Teste de Hipóteses utilizando o exemplo do construto incapacidade e os PROMs Questionário de Incapacidade de Roland-Morris (RMDQ) e Índice de Incapacidade de Oswestry (ODI) para ilustrar a construção de hipóteses

1. Qual construto cada PROM mede?	"Ambas avaliam a percepção de incapacidade relacionada à dor lombar"
2. Qual a lógica do escore dos PROMs?	"Quanto maior o escore, pior a incapacidade"
3. Espera-se correlação entre os escores dos PROMs? Por quê?	"Sim, porque ambas medem o mesmo construto"
4. Qual a magnitude de correlação esperada? Ex: fraca, moderada ou intensa?	"Pelo menos r>0,50"
5. Qual a direção da correlação esperada? Positiva ou Negativa?	"Positiva, porque em ambos os instrumentos quanto maior o escore, pior a incapacidade"*
Assim, formule a hipótese completa de correlação entre os escores das escalas:	"Espera-se uma correlação moderada e positiva entre os escores do RMDQ e do ODI. Baseado na hipótese de que os instrumentos medem o mesmo construto"

*A direção positiva refere-se a uma relação direta ou proporcional entre as variáveis, o que significa que quando uma variável aumenta seus valores, a outra também aumenta. Por outro lado, a direção negativa refere-se a uma relação inversa ou inversamente proporcional, na qual quando os valores de variável aumentam, os valores da outra variável diminuem.

Além disso, entender a lógica do escore das escalas é fundamental para garantir a direção esperada da hipótese estabelecida. Nesse caso, se para ambas as escalas o escore aumenta à medida que o indivíduo apresenta pior incapacidade, então a direção da correlação será positiva.

A recomendação, de acordo com o COSMIN, é de que 75% das hipóteses sejam confirmadas, ou seja, se 4 hipóteses forem levantadas, pelo menos 3 devem ser confirmadas para garantir que o instrumento atendeu uma boa qualidade da propriedade de após medida de Validade de Construto–Teste de Hipóteses. Além disso, os autores devem definir previamente as hipóteses com base em um modelo conceitual sobre o construto em questão. As hipóteses devem ser descritas em detalhes suficientes para permitir que o leitor avalie a plausibilidade da hipótese, a fim de garantir que as hipóteses sejam testadas de forma objetiva e que os resultados sejam interpretados corretamente[9].

Validade de critério

A validade de critério é definida pelo COSMIN como a propriedade de medida que aponta o grau em que escores de um instrumento são um reflexo adequado de um "padrão-ouro"[14]. O termo padrão-ouro refere-se a um exame/teste de referência que representa a melhor opção disponível e com resultados bem estabelecidos[19] para diagnosticar/identificar uma disfunção, transtorno ou doença. Mas o que seria um padrão-ouro para instrumentos tipo PROM? O que se poderia chamar de padrão-ouro quando se considera a percepção de uma pessoa sobre sua incapacidade, por exemplo? A comunidade científica pode assumir que o padrão-ouro para avaliar qualidade de vida é o SF-36 (Medical Outcomes Short-Form Health Survey)[20], mas essa determinação seria apenas um consenso e não implica afirmar que o SF-36 representa a "melhor medida disponível" para avaliar a qualidade de vida.

A partir dos resultados dos painéis de estudos tipo *Delphi* do COSMIN, foi recomendado que apenas versões longas (*long-form*) de PROMs, quando comparadas com versão curtas (*short-form*), podem ser consideradas padrão-ouro. Dessa forma, a validade de critério consiste em comparar/correlacionar o escore de uma versão longa com o escore da versão curta do instrumento. Um exemplo, nesse caso, seria a comparação do Inventário Breve de Dor[21] – versão curta (9 itens) com o Inventário Breve de Dor – versão longa (17 itens). O objetivo dessa comparação é identificar se é possível substituir a versão longa pela versão curta. A validade de critério é uma propriedade de medida que visa disponibilizar versões curtas de questionários, o que pode favorecer o uso dos PROMs ou OMIs na prática e na pesquisa devido à redução de sobrecarga para pacientes no tempo gasto para responder questionários longos[22].

De acordo com os critérios estabelecidos pelo COSMIN para uma boa qualidade de propriedade de medida de validade de critério, a correlação do escore da versão curta com o escore do "padrão-ouro" (versão longa) é: r ≥ 0,70 ou AUC (Area Under Curve – teste estatístico de acurácia) ≥ 0,70[23].

Um estudo[24] criou uma versão reduzida de 2 itens do Questionário de Autoeficácia sobre a Dor (PSEQ-2) que, originalmente, é composto por 10 itens (PSEQ-10). A versão reduzida foi testada adequadamente pelas orientações do COSMIN em indivíduos com dor no membro superior e apresentou uma correlação aceitável (r=0,76) com a versão original de 10 itens da ferramenta[24].

DOMÍNIO DA RESPONSIVIDADE

Responsividade

De acordo com as diretrizes estabelecidas pelo COSMIN, responsividade é definida como a capacidade de um instrumento de detectar mudanças ao longo do tempo no construto a ser medido, quando elas realmente ocorrem[14]. Tal propriedade de medida é aplicável para PROMs ou OMIs com propósitos avaliativos[25]. Para avaliação da responsividade são necessários estudos longitudinais. A propriedade de medida de responsividade está relacionada à validade dos "escores de mudança" (score change).

Essa mudança pode ocorrer através da simples flutuação de sintomas ou pré/pós uma intervenção que tenha efeitos reconhecidos na literatura para tratar a condição específica que é alvo do PROM ou OMI. A responsividade pode ser avaliada comparando-se, por exemplo, o escore de mudança do PROM com a medida de percepção global de melhora. Caso o teste estatístico de acurácia (AUC) indique que o escore do PROM foi capaz de identificar corretamente o resultado (melhora ou piora) da maior parte da amostra avaliada (70%) através da Escala de Percepção Global de Melhora, considera-se que o PROM obteve uma responsividade adequada.

Assim, um PROM que avalia funcionalidade pode ser considerado responsivo se a mudança no seu escore (pré e pós-tratamento) acompanhar o resultado do escore de mudança da Escala de Percepção Global de Melhora, ou seja, se a percepção global de melhora for positiva em um caso específico, o escore de mudança do PROM deve mostrar uma melhora da funcionalidade. Mas, por outro lado, se a percepção global de melhora for negativa para um dado paciente, o escore de mudança do PROM deve mostrar uma piora da funcionalidade.

Para o COSMIN, valores adequados de responsividade em instrumentos de escores contínuos são aqueles que conseguem confirmar pelo menos 75% das hipóteses previamente estabelecidas ou que apresentem AUC ≥0,70[2,14].

Um estudo prévio[26] demonstrou a responsividade do escore das escalas PSEQ-10, PSEQ-4 e PSEQ-2 (que avaliam autoeficácia) em pacientes com dor lombar crônica que foram submetidos a um programa de fisioterapia. As escalas foram aplicadas pré e pós-tratamento e a Escala de Percepção Global de Melhora foi aplicada após o tratamento. Os escores das escalas PSEQ-10, PSEQ-4 e PSEQ-2 demonstraram os seguintes valores de acurácia (AUC), respectivamente: 0,79, 0,81 e 0,75. Esses resultados demonstram que os escores de mudança (score change) obtidos através das escalas de autoeficácia, tanto nas versões longas quanto curtas, demonstraram uma capacidade adequada de detectar mudança quando a Escala de Percepção Global de Melhora foi utilizada como âncora ou referência. Todos os pacientes melhoraram após o tratamento? Isso não é relevante, desde que o escore de mudança do PROM ou OMI testado seja capaz de identificar o que o escore da âncora (Escala de Percepção Global de Melhora) detectou.

Interpretabilidade

A interpretabilidade de instrumentos de medida é a capacidade de compreender e extrair significado dos resultados obtidos por esses instrumentos[13,27]. A coleta de dados com o uso de PROMs ou OMIs gera resultados na forma de dados numéricos, ou seja, quantitati-

vos[13,27]. O pesquisador deve ser capaz de interpretar os dados quantitativos obtidos para conseguir elaborar um significado a partir dessas informações[13,27]. A interpretabilidade é de suma importância para estimular o uso dos PROMs ou OMIs na prática clínica e na pesquisa. A dificuldade de interpretabilidade do resultado de um escore de um questionário é uma das barreiras citadas para a utilização desses instrumentos[28].

A aplicação da interpretabilidade está no cotidiano de qualquer indivíduo que se submete a *checkups* anuais de saúde. Quando um indivíduo recebe os resultados de exames, pode-se observar índices quantitativos, como, por exemplo, sua contagem de plaquetas em um hemograma. Geralmente, ao lado do índice obtido para aquele indivíduo, estão descritos os parâmetros de normalidade. Caso a contagem das plaquetas esteja acima ou abaixo dos valores descritos como normais, o indivíduo precisa saber o que aquela contagem significa. O médico que avalia os resultados dos exames sabe interpretar aqueles números para estabelecer uma hipótese diagnóstica e direcionar o paciente para um tratamento, caso seja necessário. Sem a interpretação adequada, os resultados não têm nenhum significado.

Considerando-se os PROMs ou OMIs, conhecer o valor da Mínima Mudança Clinicamente Importante (MIC) do seu escore, especialmente para PROMs ou OMIs formulados com propósito avaliativo, pode auxiliar pesquisadores e clínicos a identificarem se o paciente melhorou ou não melhorou após o tratamento. Então, a MIC é um parâmetro de interpretabilidade de PROMs de propósito avaliativo. Será que uma redução de 2 unidades na intensidade da dor lombar pré e pós-tratamento é considerada um valor de MIC aceitável? Como definir se a mudança observada é de fato relevante ou apenas erro da medida? Para isso, o leitor vai precisar dispor de dados sobre o erro da medida e a MIC da escala de intensidade de dor, que deve ser encontrada na literatura e é condição específica.

O questionário HIT-6 (Headache Impact Questionnaire)[29] avalia o impacto da dor de cabeça. Quanto maior o escore no HIT-6, maior o impacto da dor de cabeça na vida do paciente. Assim, para uma paciente que apresentou um escore inicial de 65 pontos no HIT-6 pré-tratamento e escore 40 no HIT-6 pós-tratamento (Figura 2), é possível inferir que houve melhora com base no fato de que a paciente relatou uma melhora quando respondeu ao instrumento âncora (Escala de Percepção Global de Melhora). O *score change* da paciente foi de 65 – 40 = 15 pontos. A segunda pergunta importante é: qual é o erro da medida do escore do HIT-6? O erro da medida (SDC) do HIT-6 descrito na literatura para a versão português-Brasil é de SDC = 4.38[29]. Já a terceira pergunta é: essa mudança/melhora pode ser considerada clinicamente relevante? A MIC descrita para o HIT-6 na literatura é de 8 pontos[30]. Para que a mudança possa ser considerada clinicamente relevante SDC < MIC, nesse caso 4,38 < 15 e 4,38 < 8, então, a mudança pode ser considerada relevante clinicamente e não apenas erro de medida.

Outras perguntas relacionadas à interpretabilidade podem ser as seguintes: (I) existe um escore do PROM ou OMI esperado para subgrupos de pacientes (ex: níveis de incapacidade)? (II) Existe uma nota de corte do escore de um PROM para definição de um prognóstico? (ex: qual o escore na escala de catastrofização da dor que prediz alto risco de cronificação da dor?).

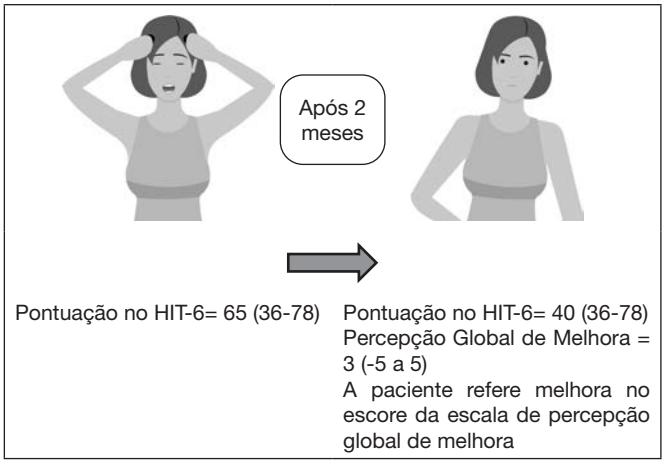

Figura 2. Observa-se a mudança no escore (change score) entre uma avaliação inicial e dois meses após a administração de um tratamento para a paciente com dor de cabeça. Houve uma mudança de 15 pontos (65-40 = 15) entre as duas avaliações e a paciente relatou melhora quando questionada através da escala de percepção global de mudança. Isso é um indicativo de que o PROM (HIT-6) demonstrou adequada responsividade, sendo capaz de identificar a melhora ao longo do tempo quando houve melhora.

Quais critérios devem ser adotados para tomada de decisão quanto ao PROM a ser utilizado na pesquisa ou prática clínica? Proposta de um *checklist* baseado em propriedades de medida

Uma das maiores dificuldades relatadas por clínicos e pesquisadores é: quais critérios devem ser seguidos para determinar se um PROM ou OMI é a melhor opção disponível para avaliar um determinado construto? Buscar informações sobre a qualidade das propriedades de medida de um PROM ou OMI na literatura é parte fundamental do processo. As revisões sistemáticas de propriedades de medida podem ajudar, reunindo em um único lugar todas essas informações. Entretanto, não é tarefa fácil interpretar os resultados encontrados na literatura para a tomada de decisão sobre o PROM ou OMI mais adequado.

Nesta revisão foi proposto um *Checklist para Caracterização da Qualidade dos PROMs e OMIs* (Tabela 3), que pode auxiliar pesquisadores e clínicos na tomada de decisão. Recomenda-se preferencialmente que a tabela 3 seja preenchida levando em consideração dados extraídos de revisões sistemáticas de estudos de propriedades de medida. Entretanto, quando não houver revisões sistemáticas disponíveis, recomenda-se, pelo menos, aplicar o *checklist* para a versão original do artigo e para a versão traduzida/adaptada. Em termos práticos, recomenda-se que clínicos e pesquisadores, ao se deparar com um instrumento na literatura, consultem o artigo de tradução e validação do instrumento e que seja realizado o exercício de preencher o *Checklist para Caracterização da Qualidade dos PROMs e OMIs*. Caso o instrumento atenda, pelo menos parte das propriedades de medida descritas, de acordo com os critérios do COSMIN, isso pode ser um indicativo de que o instrumento demonstra boa qualidade de propriedades de medida, e seu uso é encorajado. Entretanto, o uso de instrumentos não testados adequadamente pode levar a vieses na tomada de decisão na prática clínica e na pesquisa, uma vez que não é possível confiar nos resultados obtidos.

Foi feita uma análise do Inventário Breve de Dor (*short-form*) considerando os dados da revisão sistemática[31]. O *Brief Pain In-*

Tabela 3. Checklist para Caracterização da Qualidade dos *Patient Reported Outcome Measurement* (PROM) e *Outcome Measure Instrument* (OMI)

Itens	Critério de Julgamento	Classificação
Validade de Conteúdo	1 - Foi descrito/definido adequadamente o construto mensurado pelo PROM ou OMI?	Sim ☐ Não ☐ Não foi descrito ☐ Não é possível determinar ☐
	2 - Um modelo conceitual do construto mensurado pelo PROM ou OMI foi descrito?	Sim ☐ Não ☐ Não foi descrito ☐ Não é possível determinar ☐
	3 - Está descrito com clareza para qual população alvo o PROM ou OMI se aplica?	Sim ☐ Não ☐ Não foi descrito ☐ Não é possível determinar ☐
Relevância	4 - As questões do PROM ou OMI parecem ser relevantes para a população alvo?	Sim ☐ Não ☐ Não foi descrito ☐ Não é possível determinar ☐
	5 - O período de relato para resgate de memória sobre o construto é adequado e está claramente descrito?	Sim ☐ Não ☐ Não foi descrito ☐ Não é possível determinar ☐
Compreensão	6 - As questões, as opções de resposta e as instruções do PROM ou OMI são de fácil compreensão pela população alvo?	Sim ☐ Não ☐ Não foi descrito ☐ Não é possível determinar ☐
Abrangência	7 - O PROM ou OMI contempla todos os conceitos fundamentais que deveriam ser considerados para avaliar o construto?	Sim ☐ Não ☐ Não foi descrito ☐ Não é possível determinar ☐
Validade de estrutura – domínios ou subescalas do PROM ou OMI	8- Uma descrição através de análises adequadas demonstra que a escala é unidimensional ou multidimensional?	Sim ☐ Não ☐ Não foi descrito ☐ Não é possível determinar ☐
	a - Para análise fatorial confirmatória há descrição dos seguintes aspectos: CFI ou TLI > 0.95 ou RMSEA ou SRMR < 0.06	
	b - Para análise Rasch é descrita: sem violação da unidimensionalidade, sem violação da independência local, sem violação da monotocidade, e ajuste do modelo adequado (valores *infit* e *outfit* entre ≥ 0.5 e ≤ 1.5)?	Sim ☐ Não ☐ Não foi descrito ☐ Não é possível determinar ☐
	c - Para análise fatorial exploratória, Cargas fatoriais > 0.30 e apenas 10% dos itens carregando em mais de 1 fator e variância explicada de pelo menos 50% ou resultados do scree plot ou critério de Kaiser (*Eigenvalues* > 1) alinhados com o modelo conceitual do PROM ou OMI?	
Consistência Interna	9 - A validade estrutural foi verificada?	Sim ☐ Não ☐ Não foi descrito ☐ Não é possível determinar ☐
	10 - O alpha de Cronbach ≥ 0.70?	Sim ☐ Não ☐ Não foi descrito ☐ Não é possível determinar ☐
Confiabilidade	11 - A confiabilidade da escala demonstrou valores adequados? Tais como ICC ou Kappa ponderado ou r ≥ 0,70?	Sim ☐ Não ☐ Não foi descrito ☐ Não é possível determinar ☐
	12 - O período de teste-reteste pode ser considerado como adequado?	Sim ☐ Não ☐ Não foi descrito ☐ Não é possível determinar ☐
	13 - Foi oferecida uma descrição clara de que os pacientes estavam estáveis no período de teste-reteste?	Sim ☐ Não ☐ Não foi descrito ☐ Não é possível determinar ☐
Medida de Erro	14 - SDC ou LoA < MIC	Sim ☐ Não ☐ Não foi descrito ☐ Não é possível determinar ☐
Validade de Construto – teste de hipóteses	15 - Pelo menos 75% das hipóteses de comparações levantadas foram confirmadas?	Sim ☐ Não ☐ Não foi descrito ☐ Não é possível determinar ☐
Validade de Critério	16- Uma correlação r ≥ 0.70 foi observada entre o escore da versão longa e curta do PROM ou OMI? Ou AUC ≥ 0.70?	Sim ☐ Não ☐ Não foi descrito ☐ Não é possível determinar ☐
Responsividade	17 - Pelo menos 75% das hipóteses de comparações levantadas foram confirmadas ou AUC ≥ 0.70?	Sim ☐ Não ☐ Não foi descrito ☐ Não é possível determinar ☐
Interpretabilidade	18 - É descrito para o PROM ou OMI dados/valores que permitem a interpretação dos escores obtidos? Ex1: Mínima Mudança Clinicamente Importante (MIC)? Ex2: valor de corte para determinação de subgrupos? Ex3: como interpretar o escore: por exemplo o que significa um escore alto ou baixo?	Sim ☐ Não ☐ Não foi descrito ☐ Não é possível determinar ☐
Adaptação transcultural	19 - Existe uma versão do PROM ou OMI disponível em português Brasil que seguiu um método adequado de adaptação transcultural?	Sim ☐ Não ☐ Não foi descrito ☐ Não é possível determinar ☐
	20 - As propriedades de medida de foram testadas do PROM ou OMI foram testadas em uma amostra de brasileiros?	Sim ☐ Não ☐ Não foi descrito ☐ Não é possível determinar ☐

ventory atendeu o critério "sim" para 14 dos 20 itens do critério proposto (70%)". Assim, a maior parte das propriedades de medida foram atendidas em vários estudos internacionais.

CONCLUSÃO

Este artigo (parte II) abordou as propriedades dos domínios da validade e da responsividade, bem como da interpretabilidade. Além disso, foi proposto um *checklist* para facilitar a operacionalização do conhecimento sobre as propriedades de medida. Para atender a Validade de Estrutura, o PROM deve ser submetido a uma análise fatorial ou do tipo Rasch. Já para a Validade de Construto–Teste de Hipóteses é necessário confirmar pelo menos 75% das hipóteses levantadas *a priori*. A responsividade deve ser avaliada através de análises de acurácia (AUC ≥ 0.70) e a Mínima Mudança Importante (interpretabilidade) pode ser utilizada para determinar se um paciente obteve uma melhora clinicamente relevante. Assim, esta pesquisa encoraja a aplicação do *checklist* proposto, o que pode auxiliar a obtenção de dados fidedignos e válidos para dar suporte e auxiliar clínicos e pesquisadores na escolha do instrumento mais adequado para subsidiar a tomada de decisão.

CONTRIBUIÇÕES DOS AUTORES

Thais Cristina Chaves
Conceitualização, Gerenciamento de Recursos, Gerenciamento do Projeto, Metodologia, Redação - Preparação do Original, Redação - Revisão e Edição, Supervisão
Thamiris Costa Lima
Redação - Preparação do Original, Redação - Revisão e Edição
Juliana H. Padilha Spavieri
Metodologia, Redação - Preparação do Original, Redação - Revisão e Edição
Ana Carolina de Jacomo Claudio
Metodologia, Redação - Preparação do Original, Redação - Revisão e Edição
Roger Berg Rodrigues Pereira
Metodologia, Redação - Preparação do Original, Redação - Revisão e Edição
Mariana Romano de Lira
Metodologia, Redação - Preparação do Original, Redação - Revisão e Edição

SIGLAS

PROM: *Patient Reported Outcome Measure*
OMI: *Outcome Measurement Instrument*
CFI: *Confirmatory Fit Index*, *Root Mean Square*
TLI: *Tucker-Lewis Index*
RMSEA: *Root Mean Square Error of Approximation*
SRMR: S*tandardized Root Mean Square Residual*
AUC: *Area Under the Curve* (análise de acurácia)
SDC: Mínima Diferença Detectável/*Smallest Detectable Change*
MIC: Mínima Diferença Clinicamente Importante/*Minimal Important Change*
LoA: *Limits of Agreement* (Bland & Altman)

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. 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.
5. Mokkink LB, Boers M, van der Vleuten CPM, Bouter LM, Alonso J, Patrick DL, De Vet HCW, and Terwee CB. COSMIN Risk of Bias tool to assess the quality of studies on reliability or measurement error of outcome measurement instruments: a Delphi study. BMC Med Res Methodol. 2020;20:293.
6. Mokkink LB, Terwee CB, Patrick DL, Alonso J, Stratford PW, Knol DL, et al. The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes. J Clin Epidemiol. 2010;63(7):737-45.
7. 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;27(5):1171-9.
8. 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):1159-70.
9. Patrick DL, Burke LB, Gwaltney CJ, Leidy NK, Martin ML, Molsen E, et al. Content Validity—Establishing and Reporting the Evidence in Newly Developed Patient-Reported Outcomes (PRO) Instruments for Medical Product Evaluation: ISPOR PRO Good Research Practices Task Force Report: Part 1—Eliciting Concepts for a New PRO Instrument. Value Health. 2011;14(8):967-77.
10. Binkley JM, Stratford PW, Lott SA, Riddle DL. The Lower Extremity Functional Scale (LEFS): scale development, measurement properties, and clinical application. North American Orthopaedic Rehabilitation Research Network. Phys Ther. 1999;79(4):371-83.
11. Ackelman BH, Lindgren U. Validity and reliability of a modified version of the neck disability index. J Rehabil Med. 2002;34(6):284-7
12. Patrick DL, Burke LB, Gwaltney CJ, Leidy NK, Martin ML, Molsen E, Ring L. Content validity--establishing and reporting the evidence in newly developed patient--reported outcomes (PRO) instruments for medical product evaluation: ISPOR PRO good research practices task force report: part 1--eliciting concepts for a new PRO instrument. Value Health. 2011;14(8):967-77.
13. De Vet HCW, Terwee CB, Mokkink LB, Knol DL. Measurement in Medicine - A practical guide. 1st edition. New York: Cambridge University Press; 2011.
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;27(5):1171-9.
15. 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 Sep 13;17(1):449.
16. Aguiar AS, Bataglion C, Visscher CM, Bevilaqua Grossi D, Chaves TC. Cross-cultural adaptation, reliability and construct validity of the Tampa scale for kinesiophobia for temporomandibular disorders (TSK/TMD-Br) into Brazilian Portuguese. J Oral Rehabil. 2017;44(7):500-510.
17. Nusbaum L, Natour J, Ferraz MB, Goldenberg J. Translation, adaptation and validation of the Roland-Morris questionnaire--Brazil Roland-Morris. Braz J Med Biol Res. 2001;34(2):203-10.
18. Vigatto R, Alexandre NMC, Filho HRC. Development of a Brazilian Portuguese Version of the Oswestry Disability Index. Spine. 2007;32(4):481–6.
19. Cardoso JR, Pereira LM, Iversen MD, Ramos AL. What is gold standard and what is ground truth? Dental Press J Orthod. 2014;19(5):27-30.
20. Ware JE, Sherbourne CD. The MOS 36-Item Short Form Health Survey (SF-36) I. Conceptual framework and item selection. Med Care. 1992;30:473-83.
21. Cleeland CS, Ryan KM. Pain assessment: global use of the Brief Pain Inventory. Ann Acad Med Singap. 1994;23(2):129-38.
22. Vaegter HB, Handberg G, Kent P. Brief psychological screening questions can be useful for ruling out psychological conditions in patients with chronic pain. Clin J Pain. 2018;34(2):113-21.
23. 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.
24. Bot AGJ, Nota SPFT, Ring D. The Creation of an Abbreviated Version of the PSEQ: the PSEQ-2. Psychosomatics. 2014;55(4):381-5.
25. Bull C, Teede H, Watson D, Callander EJ. Selecting and Implementing Patient-Reported Outcome and Experience Measures to Assess Health System Performance. JAMA Health Forum. 2022;3(4):e220326.

26. Chiarotto A, Vanti C, Cedraschi C, Ferrari S, de Lima E Sà Resende F, Ostelo RW, Pillastrini P. Responsiveness and Minimal Important Change of the Pain Self-Efficacy Questionnaire and Short Forms in Patients With Chronic Low Back Pain. J Pain. 2016;17(6):707-18.
27. 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
28. Turner GM, Litchfield I, Finnikin S, Aiyegbusi OL, Calvert M. General practitioners' views on use of patient reported outcome measures in primary care: a cross-sectional survey and qualitative study. BMC Fam Pract. 2020;21(1):14.
29. Pradela J, Bevilaqua-Grossi D, Chaves TC, Dach F, Carvalho GF. Measurement properties of the Headache Impact Test (HIT-6TM Brazil) in primary and secondary headaches. Headache. 2021;11;61(3):527-35.
30. Castien RF, Blankenstein AH, Windt DA, Dekker J. Minimal clinically important change on the Headache Impact Test-6 questionnaire in patients with chronic tension-type headache. Cephalalgia. 2012;32(9):710-4
31. Jumbo SU, MacDermid JC, Kalu ME, Packham TL, Athwal GS, Faber KJ. Measurement Properties of the Brief Pain Inventory-Short Form (BPI-SF) and Revised Short McGill Pain Questionnaire Version-2 (SF-MPQ-2) in Pain-related Musculoskeletal Conditions: a Systematic Review. Clin J Pain. 2021;37(6):454-74.