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

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

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-en

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,

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. Federal University of São Carlos, Physical Therapy Department, Postgraduate Program in Rehabilitation and Functional Performance, Health Sciences Department, Ribeirão Preto School of Medicine, São Carlos, SP, Brazil.
2. Federal University of São Carlos, Ribeirão Preto School of Medicine, Postgraduate Program in Physical Therapy, Physical Therapy Department, São Carlos, SP, Brazil.
3. University of São Paulo, Ribeirão Preto School of Medicine, Health Sciences Department, Postgraduate Program in Rehabilitation and Functional Performance, Ribeirão Preto, SP, Brazil.

Submitted on September 6, 2023.
Accepted for publication on October 10, 2023.
Conflict of interests: none - Sponsoring sources: none.

HIGHLIGHTS
1. Within the validity domain it is possible to analyze: structural validity and construct validity-hypothesis testing.
2. Responsiveness is the ability of an instrument to detect changes over time.
3. Interpretability is the ability to extract meaning from the results obtained by PROMs.
4. The minimum clinically important change (MIC) is one of the interpretability measures of PROMs.
5. A checklist with 20 items was proposed to assist in determining the PROM with the best quality.

Responsible associate editor: Luciana Buin
https://orcid.org/0000-0002-1824-5749

Correspondence to:
Thaís Cristina Chaves
E-mail: thaischaves@ufscar.br

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 na 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 elaborada a o presente estudo, 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 os 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.

INTRODUCTION

PROM stands for Patient Reported Outcome Measure[1]. Another commonly used acronym is OMI (Outcome Measurement Instrument)[2]. PROM-type instruments were developed to assess constructs or concepts that cannot be directly measured or that would be difficult to measure in practice (e.g. kinesiophobia or fear of movement)[3]. There are numerous PROMs or OMIs available in the literature, however one of the great challenges for clinicians and researchers is to define which available instrument is the most appropriate[4]. Understanding the measurement properties of a PROM or OMI can help clinicians and researchers to make a decision about which instrument to use. Thus, PROMs or OMIs whose majority of measurement properties have been tested and whose properties meet the quality criteria described by international initiatives (such as the COnsensus-based Standards for the selection of health Measurement INstruments - COSMIN)[5,6] should be preferred.

As described in part I of this series of two articles, measurement properties are obtained by studying the characteristics of a given measure (for example, by establishing relationships/comparisons between the score of an instrument and the scores of other instruments), with the aim of identifying whether the measure (e.g. PROM or OMI score) has adequate qualities. Part I presented the measurement properties of the reliability domain. Part II described the properties within the validity, responsiveness and interpretability domains.
The validity domain of an instrument brings together the measurement properties that try to identify whether the instrument "measures what it purports to measure"[2]. The following measurement properties are described in this domain, according to COSMIN: (I) content validity, (II) structural validity, (III) hypotheses testing for construct validity, (IV) cross-cultural validity and criterion validity.
The responsiveness domain brings together only one measurement property, which has the same name as the domain: responsiveness. Responsiveness is aligned with an instrument's ability to detect changes in PROM or OMI scores over time[7] in a valid way. It is a type of validity (the validity of score change), which has been removed from the validity domain (by COSMIN) to avoid confusion.
Finally, the interpretability of a PROM is related to the ease of interpretation and the attribution of meaning to the score of an instrument for its application in practice[8]. Although it is not considered a measurement property, interpretability is a fundamental characteristic of measurement instruments, although it is commonly neglected by researchers.
Considering the difficulty of operationalizing knowledge about measurement properties, the proposal of a guideline or checklist can help gather the necessary information to help professionals and researchers make a decision about choosing the most suitable PROMs, in relation to the quality of their measurement properties. Thus contributing to the translation of scientific knowledge into practice.
Considering these challenges, the objectives of part II of this narrative review (didactically divided into two parts) were: (I) to describe the main measurement properties of the validity and responsiveness domains, (2) to describe the interpretability of PROMs and OMIs, and (3) to provide a checklist which, when completed, can help researchers and clinicians to operationalize/gather information on the quality of the PROM measurement properties available in the literature and thus facilitate the decision-making process.

METHODS

This narrative review was based on studies published by the COSMIN consensus. Of the 31 references cited in this article, 10 are articles from the COSMIN initiative[2,5-8,13-15,21,25].

VALIDITY DOMAIN

Content validity

Content validity is defined by the empirical (subjective) evidence which demonstrates that the items and domains of an instru-

ment are appropriate and comprehensive in relation to the measurement concepts, population and intended use[9]. To this end, it is important that the construct to be assessed is well defined and interpretable. The instrument's questions must be designed in such a way that they can adequately capture people's perception of the construct. In addition, a precise and well-founded definition of the construct should underpin the creation of the instrument's items. A key point in the development of an instrument is the clear definition of the construct to be measured, and the construction of a conceptual model can be of great value in determining which questions/items should be included in the PROM or OMI[9].

For example, the Lower Extremity Functional Scale (LEFS) was developed to specifically assess functioning related to the lower limbs[10]. Its questions only cover functional activities involving the lower limbs and its scale was designed to assess functioning. Thus, the higher the LEFS score, the greater the functioning of the patient with lower limb disorders. Questionnaires such as the Neck Disability Index (NDI), which assesses disability related to neck pain, whose construct is disability but includes questions about pain intensity and headache intensity, have limitations regarding their content. The questions that should be asked in this context are: what is the definition of disability considered by the authors? Is pain intensity a construct that should be included in a questionnaire intended to measure disability? In one of the analyzed articles, the authors indicated that the construct of the NDI is "to measure the limitation of activities due to neck pain"[11]. However, would pain intensity be an activity?

The first stage in creating a PROM or OMI is developing the instrument. It is common to use qualitative studies (focus groups) to carry out the "eliciting content" or content generation phase[9]. It is of the utmost importance that the target audience is involved in this stage and that the participants describe what content should be included in the instrument. At the end of the content generation process, a draft version of the questionnaire can be created, and this version should be evaluated again by the tool's target audience[9,12].

This stage can be considered content validity itself and should preferably involve the participation of the target audience intended by the PROM or OMI and experts. This stage can be carried out through Delphi type studies or through qualitative studies. Three aspects should be considered at this stage: comprehension, comprehensiveness and relevance of the OMI items[8]. COSMIN describes, in one of its articles, a 10-item criterion (table 1) to guide the quality assessment of the content validity of a PROM or OMI[8]. For translated instruments, content validity is not usually described in the literature, since the content of an instrument cannot be modified in the translation process, only culturally adapted without affecting equivalence with the original version.

Table 1. 10-item criterion for assessing the quality of content validity suggested by the COnsensus-based Standards for the selection of health Measurement INstruments (COSMIN)

Relevance
1. Are the items/questions included relevant to the construct of interest?
2. Are the items/questions included relevant to the target population?
3. Are the items/questions included relevant to the PROM or OMI application context?
4. Are the answer options appropriate?
5. Can the memory recall period be considered appropriate?
Scope
6. Is there an important concept for evaluating the construct that is missing from PROM or OMI?
Comprehension
7. Are the PROM or OMI instructions easy to understand for the PROM or OMI target population?
8. Are the PROM or OMI response options easy to understand for the PROM or OMI target population?
9. Are the questions or items in the PROM or IMO properly worded?
10. Are the answer options in line with the questions asked?

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

Structural validity

Structural validity is defined as "the degree to which the scores of a measuring instrument are an adequate reflection of the dimensionality of the construct being measured"[13]. Thus, structural validity assesses how many factors or domains are present in an instrument, and which items are part of each dimension/domain/factor. Thus, structural validity can define the dimensionality of an instrument. Identifying the dimensions is important not only for determining how the PROM or OMI score will be obtained, but also for interpreting the results[7]. Multidimensional questionnaires should have separate scoring systems for each domain, making interpretation and clinical decision-making more precise than when only the total score of an instrument is used[13].

Considering Classical Test Theory, Factor Analysis (FA), based on the correlation of items, is the most widely used method for determining the dimensionality of PROMs or OMIs[13]. The basic principle is that highly correlated items are grouped together in the same factor/domain, while poor correlated items are loaded in different factors, i.e. items belonging to different factors correlate to a lesser extent[13]. A questionnaire with 3 domains, for example, should have its 3-factor model confirmed by factor analysis.

Exploratory Factor Analysis (EFA) is generally applied if there are no clear hypothesis about the number of dimensions of a scale; it is a method that is not very robust and is only suitable for generating a preliminary theory for confirmation *a posteriori*. Therefore, it should preferably be used in the development phase of an instrument[13].

Confirmatory Factor Analysis (CFA) is recommended if a priori hypotheses about the dimensions of the construct are available, based on theory or previous analyses. Therefore, for validation purposes, CFA is more robust and is recommended by COSMIN[13,14].

For CFA, the fit indices are used to test whether the data fit the hypothesized factor structure. COSMIN considers good measurement property of structural validity if the analysis meets the

following criteria: (I) Comparative Fit Index (CFI) or Tucker-Lewis Index (TLI) or comparable measure is > 0.95; and (II) Root Mean Square Error of Approximation (RMSEA) < 0.06; or (III) Standardized Root Mean Square Residual (SRMR) < 0.08[2]. Considering Item Response Theory, Rasch analysis can be used as a mathematical model for evaluating one-dimensional questionnaires, i.e. checking whether the items on a scale that represent a construct are represented by a single dimension[2]. COSMIN provides a detailed description of the quality criteria that should be considered for Rasch analysis: absence of violation of unidimensionality, local independence and monotocity, and adequate model fit (e.g. infit and outfit between 0.5 and 1.5)[15].

The structural validity of the Tampa Kinesiophobia Scale for Temporomandibular Disorders (TSK-TMD/Br) translated into Brazilian Portuguese was verified through a CFA which confirmed the two-factor structure demonstrated for the original English version of the scale, with a Comparative Fit Index (CFI) = 0.97, which meets the criteria for good measurement property of structural validity (Figure 1). Questions 1, 2, 10, 15, 17 and 18 fall into the "Activity avoidance" (AA) domain and questions 8-12 fall into the "Somatic focus" (SF) domain. Figure 1 illustrates the structure of the TSK-TMD/Br.

CONSTRUCT VALIDITY - HYPOTHESIS TESTING

Construct validity is the degree to which the scores of a PROM or OMI are consistent with hypotheses based on the assumption that the PROM or OMI measures the construct it intended to measure[2]. To assess construct validity, hypotheses must be formulated about how the scores of an instrument relate to other instruments that measure similar or different constructs, including not only the direction but also the magnitude of the correlations. Therefore, the hypothesis testing tool is used to test these hypotheses[13].

These hypotheses can be made by internal correlations and external correlations. Internal correlations are comparisons between the scores of the domains of a given PROM or OMI. External correlations are comparisons between different PROMs or OMIs (which measure the same construct or not). There is also the possibility of correlations between hypotheses of differences between the scores obtained to define relevant groups (e.g. when the score of an instrument is able to differentiate groups according to levels of disability). For this reason, the current recommendation is that construct validity be called Hypotheses testing for Construct Validity[13]. A possible statistical test to verify the Hypotheses testing for Construct Validity is the Pearson or Spearman correlation tests.

Table 2 shows guiding questions that can help direct the construction of hypotheses for Hypotheses testing for Construct Validity. For example, if two questionnaires measure the construct "perception of disability related to low back pain", such as the Roland-Morris Disability Questionnaire (RMDQ)[17] and the Oswestry Disability Index (ODI)[18], a correlation between the scores of the scales can be expected, as both measure the same construct.

In addition, understanding the logic of the scale scores is fundamental to guaranteeing the expected direction of the

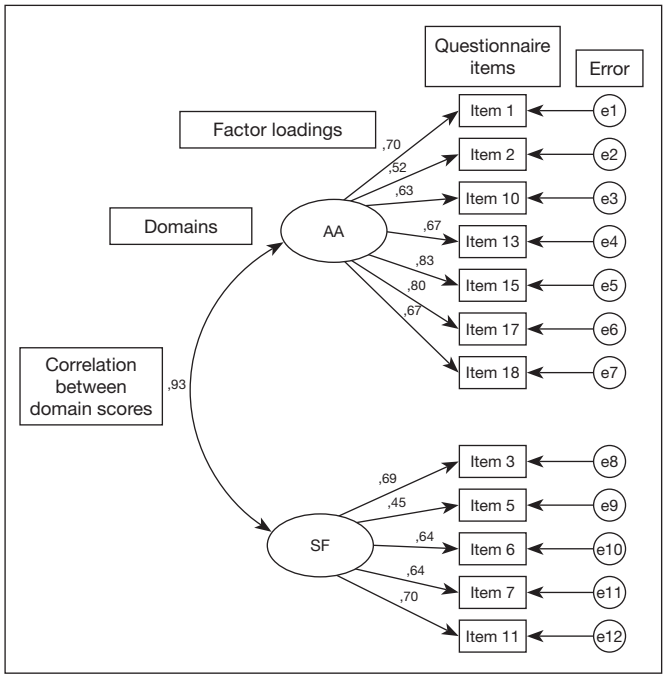

Figure 1. Diagram representing the model tested through the Confirmatory Factor Analysis of the Tampa Kinesiophobia Scale for Temporomandibular Dysfunction. Adapted[16].

AA = Activity Avoidance domain; SF= Somatic Focus domain

Table 2. Guiding questions for defining the hypotheses of the Hypotheses testing for Construct Validity using the example of the disability construct and the PROMs Roland-Morris Disability Questionnaire (RMDQ) and Oswestry Disability Index (ODI) to illustrate the construction of hypotheses.

1. What construct does each PROM measure?	"Both assess the perception of disability related to low back pain"
2. What is the logic behind the PROMs score?	"The higher the score, the worse the disability"
3. Is a correlation expected between PROMs scores? Why?	"Yes, because they both measure the same construct"
4. What is the expected magnitude of the correlation? Ex: weak, moderate or intense?	"At least r > 0.50"
5. What is the direction of the expected correlation? Positive or negative?	"Positive, because for both instruments the higher the score, the greater the disability"*
So, formulate the complete hypothesis of correlation between the scores of the scales:	"A moderate and positive correlation is expected between the RMDQ and ODI scores. Based on the hypothesis that the instruments measure the same construct"

*The positive direction refers to a direct or proportional relationship between the variables, which means that when one variable increases in value, the other also increases. On the other hand, the negative direction refers to an inverse or inversely proportional relationship, in which when the scores of a variable increases, the score of the other decreases.

established hypothesis. In this case, if for both scales the score increases as the individual's disability worsens, then the direction of the correlation will be positive.

The recommendation, according to COSMIN, is that 75% of the hypotheses should be confirmed, i.e. if 4 hypotheses are raised, at least 3 should be confirmed to ensure that the instrument meets the criterion for good quality of Hypotheses Testing for Construct Validity. In addition, the authors must first define the hypotheses based on a conceptual model of the construct in question. The hypotheses should be described in sufficient detail to allow the reader to assess the plausibility of the hypothesis, in order to ensure that the hypotheses are tested objectively and that the results are interpreted correctly[9].

Criterion validity

Criterion validity is defined by COSMIN as the measurement property that indicates the degree to which scores on an instrument are an adequate reflection of a "gold standard"[14]. The term gold standard refers to a reference exam/test that represents the best available option with well-established results[19] for diagnosing/identifying a dysfunction, disorder or disease. But what would be a gold standard for PROM-type instruments? What could be called the gold standard when considering a person's perception of their disability, for example? The scientific community may assume that the gold standard for assessing quality of life is the SF-36 (Medical Outcomes Short-Form Health Survey)[20], but this determination would only be a consensus and does not imply that SF-36 represents the "best available measure" for assessing quality of life.

Based on the results of the COSMIN Delphi study panels, it was recommended that only long-form versions of PROMs, when compared to short-form versions, can be considered the gold standard. Thus, criterion validity consists of comparing/correlating the score of a long-form version with the score of the short-form version of the instrument. An example in this case would be comparing the Brief Pain Inventory[21] - short version (9 items) with the Brief Pain Inventory - long version (17 items). The aim of this comparison is to identify whether it is possible to replace the long version with the short version. Criterion validity is a measurement property that aims to make short versions of questionnaires available, which can favor the use of PROMs or OMIs in practice and research due to the reduction of the burden on patients in the time spent answering long questionnaires[22].

According to the criteria established by COSMIN for a good quality criterion validity measurement property, the correlation between the score of the short version and the score of the "gold standard" (long version) is: $r \geq 0.70$ or AUC (Area Under Curve - statistical test of accuracy) ≥ 0.70[23].

A previous study[24] created a reduced 2-item version of the Pain Self-Efficacy Questionnaire (PSEQ-2), which originally consisted of 10 items (PSEQ-10). The short version was adequately tested according to the COSMIN guidelines in individuals with upper limb pain and showed an acceptable correlation ($r = 0.76$) with the original 10-item version of the tool[24].

RESPONSIVENESS DOMAIN

Responsiveness

According to COSMIN, responsiveness is defined as the ability of an instrument to detect changes over time in the construct being measured, when the change actually occurs[14]. This measurement property is applicable to PROMs or OMIs with evaluative purposes[25]. Longitudinal studies are needed to assess responsiveness. The measurement property of responsiveness is related to the validity of "score change".

This change can occur through the simple flare-up of symptoms or pre/post an intervention that has effects recognized in the literature to treat the specific condition that is the target of the PROM or OMI. Responsiveness can be assessed by comparing, for example, the PROM score with the measure of global perceived effect of improvement. If the statistical accuracy test (AUC) indicates that the PROM score was able to correctly identify the outcome (improvement or worsening) of the majority of the sample evaluated (70%) using the global perceived effect of improvement scale, the PROM is considered to meet adequate responsiveness.

Thus, a PROM that assesses functioning can be considered responsive if the change in its score (pre- and post-treatment) follows the result of the change score on the global perceived effect of improvement scale, i.e. if the global perception of improvement is positive in a specific case, the PROM's change score should show an improvement in functionality. On the other hand, if the global perceived effect of improvement shows a negative score for a given patient then the change score of the PROM must show a functioning worsening For COSMIN, adequate responsiveness values for continuous score instruments are those that can confirm at least 75% of the hypotheses previously established or that have an AUC ≥ 0.70[2,14].

A previous study[26] demonstrated the responsiveness of the PSEQ-10, PSEQ-4 and PSEQ-2 scales (which assess self-efficacy) in patients with chronic low back pain who underwent a physiotherapy program. The scales were applied pre- and post-treatment and the global perceived effect of improvement scale was applied after treatment. The scores of the PSEQ-10, PSEQ-4 and PSEQ-2 scales showed the following accuracy values (AUC), respectively: 0.79, 0.81 and 0.75. These results show that the score change obtained through the self-efficacy scales, both the long and short versions, demonstrated an adequate ability to detect change when the global perceived effect of improvement scale was used as an anchor or reference. Did all patients improve after treatment? This is not relevant, as long as the PROM or OMI change score tested is able to identify what the anchor score (Global Perception of Improvement Scale) detected.

Interpretability

The interpretability of measuring instruments is the ability to understand and extract meaning from the results obtained by these instruments[13,27]. Collecting data using PROMs or OMIs generates results in the form of numerical data, i.e. quantitative data[13,27]. The researcher must be able to interpret

the quantitative data obtained in order to make sense of this information[13,27]. Interpretability is of the utmost importance to encourage the use of PROMs or OMIs in clinical practice and research. The difficulty of interpreting the result of a questionnaire score is one of the barriers cited for the use of these instruments[28].

The application of interpretability is part of the daily life of anyone who undergoes annual health checkups. When an individual receives test results, quantitative indices can be seen, such as their platelet rate in a blood count. Generally, the parameters of normality are described next to the index obtained for that individual. If the platelet count is above or below the values described as normal, the individual needs to know what that count means. The doctor who evaluates the test results knows how to interpret these outcomes in order to establish a diagnostic hypothesis and direct the patient towards treatment, if necessary. Without proper interpretation, the results have no meaning.

Considering PROMs or OMIs, knowing the value of the Minimal Important Change (MIC) of their score, especially for PROMs or OMIs formulated for evaluative purposes, can help researchers and clinicians identify whether the patient has improved or not after treatment. So, the MIC is a parameter for the interpretability of PROMs for evaluative purposes. Is a reduction of 2 units in the intensity of low back pain pre- and post-treatment considered an acceptable MIC value? How to define whether the observed change is actually relevant or just measurement error? For this, the reader will need data on measurement error and the MIC of the pain intensity scale, which must be found in the literature and is condition-specific.

The HIT-6 (Headache Impact Questionnaire)[29] assesses the impact of headaches. The higher the HIT-6 score, the greater the impact of the headache on the patient's life. Thus, for a patient who had an initial pre-treatment HIT-6 score of 65 and a post-treatment HIT-6 score of 40 (Figure 2), it is possible to infer that there has been improvement based on the fact that the patient reported an improvement when answering the anchor instrument (global perceived effect of improvement scale). The patient's change score was 65 - 40 = 15 points. The second important question is: what is the measurement error of the HIT-6 score? The measurement error (SDC) of the HIT-6 described in the literature for the Brazilian Portuguese is SDC = 4.38[29]. The third question is: can this change/improvement be considered clinically relevant? The MIC described for the HIT-6 in the literature is 8 points[30]. For the change to be considered clinically relevant, SDC < MIC, in this case 4.38<15 and 4.38<8, then the change can be considered clinically relevant and not just measurement error.

Other questions related to interpretability could be the following: (I) is there an expected PROM or OMI score for subgroups of patients (e.g. levels of disability)? (II) Is there a cut-off score for a PROM score to define a prognosis? (e.g. what score on the pain catastrophizing scale predicts a high risk of chronic pain?

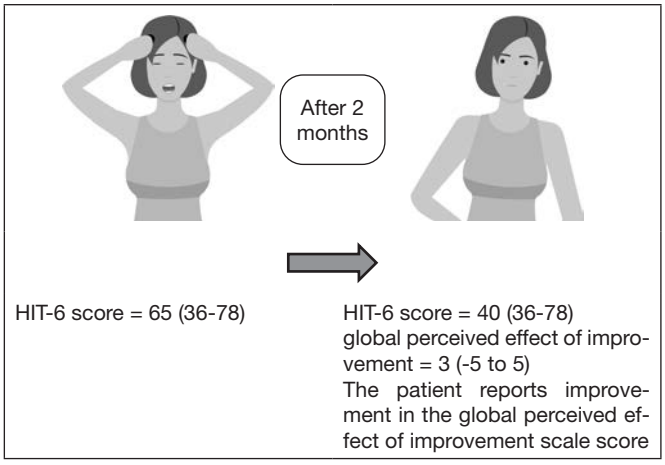

Figure 2. The change in the score between initial assessment and two months after the headache treatment was administered. There was a change of 15 points (65-40 = 15) between the two evaluations and the patient reported improvement when questioned using global perceived effect of improvement scale. This is an indication that the PROM (HIT-6) showed adequate responsiveness, being able to identify improvement over time when there was improvement.

What criteria should be used when deciding which PROM to use in research or clinical practice? Proposal for a checklist based on measurement properties

One of the biggest difficulties reported by clinicians and researchers was: commonly reported is criteria should be followed to determine whether a PROM or OMI is the best available option for assessing a given construct? Searching for information on the quality of the measurement properties of a PROM or OMI in the literature is a fundamental part of the process. Systematic reviews of measurement properties can help by bringing all this information together in one place. However, it is no easy task to interpret the results found in the literature in order to make a decision on the most appropriate PROM or OMI.

This review proposes a *Checklist for Characterizing the Quality of PROMs and OMIs* (Table 3), which can help researchers and clinicians in their decision-making. It is recommended that the table below be filled in taking into account data extracted from systematic reviews of studies on measurement properties. However, when no systematic reviews are available, it is recommended to at least apply the checklist to the original version of the article and to the translated/adapted version. In practical terms, it is recommended that when clinicians and researchers come across an instrument in the literature, they consult the article that cross-culturally translated and validated the instrument and complete the *Checklist for Characterizing the Quality of PROMs and OMIs*. If the instrument meets at least part of the measurement properties described, according to the COSMIN criteria, this may be an indication that the instrument demonstrates good quality measurement properties, and its use is encouraged. However, the use of instruments that have not been adequately tested can lead to biases in decision-making in clinical practice and research, since it is not possible to trust the results obtained.

An analysis of the Brief Pain Inventory (short-form) was carried out considering the data from a systematic review[31]. The Brief

Table 3. Checklist for Characterizing the Quality of Patient Reported Outcome Measure (PROM) and Outcome Measurement Instrument (OMI)

Items	Judging criteria	Classification
Validity of Content	1 - Has the construct measured by PROM or OMI been adequately described/defined?	Yes ☐ No ☐ Not described ☐ Cannot be determined ☐
	2 - Has a conceptual model of the construct measured by PROM or OMI been described?	Yes ☐ No ☐ Not described ☐ Cannot be determined ☐
	3 - Is it clearly described to which target population the PROM or OMI applies?	Yes ☐ No ☐ Not described ☐ Cannot be determined ☐
Relevance	4 - Do the PROM or OMI questions seem relevant to the target population?	Yes ☐ No ☐ Not described ☐ Cannot be determined ☐
	5 - Is the reporting period for recalling the construct adequate and clearly described?	Yes ☐ No ☐ Not described ☐ Cannot be determined ☐
Understanding	6 - Are the questions, answer options and instructions in the PROM or OMI easy for the target population to understand?	Yes ☐ No ☐ Not described ☐ Cannot be determined ☐
Scope	7 - Does the PROM or OMI cover all the fundamental concepts that should be considered when evaluating the construct?	Yes ☐ No ☐ Not described ☐ Cannot be determined ☐
Structural validity - PROM or OMI domains or sub-scales	8- Does a description through appropriate analysis show that the scale is unidimensional or multidimensional?	Yes ☐ No ☐ Not described ☐ Cannot be determined ☐
	a – For confirmatory factor analysis the following aspects are described: CFI or TLI > 0.95 or RMSEA or SRMR < 0.06	Yes ☐ No ☐ Not described ☐ Cannot be determined ☐
	b - For Rasch analysis is described: no violation of unidimensionality, no violation of local independence, no violation of monotocity, and adequate model fit (infit and outfit values between ≥ 0.5 and ≤ 1.5)?	
	c -For exploratory factor analysis, Factor loadings > 0.30 and only 10% of items loading on more than 1 factor and explained variance of at least 50% or scree plot results or Kaiser criterion (Eigenvalues > 1) aligned with the PROM or OMI conceptual model?	
Internal consistency	9 - Has structural validity been verified?	Yes ☐ No ☐ Not described ☐ Cannot be determined ☐
	10 - Cronbach's alpha > 0.70?	Yes ☐ No ☐ Not described ☐ Cannot be determined ☐
Reliability	11 – Did the reliability of the scale show adequate values? Such as ICC or weighted Kappa or r ≥ 0.70?	Yes ☐ No ☐ Not described ☐ Cannot be determined ☐
	12 - Can the test-retest period be considered adequate?	Yes ☐ No ☐ Not described ☐ Cannot be determined ☐
	13 – Was a clear description offered that the patients were stable in the test-retest period?	Yes ☐ No ☐ Not described ☐ Cannot be determined ☐
Error measurement	14 – SDC or LoA < MIC	Yes ☐ No ☐ Not described ☐ Cannot be determined ☐
Construct validity - hypothesis testing	15 - Have at least 75% of the hypotheses raised been confirmed?	Yes ☐ No ☐ Not described ☐ Cannot be determined ☐
Criteria Validity	16 - Was a correlation r > 0.70 observed between the score of the long and short version of the PROM or OMI? Or AUC ≥ 0.70?	Yes ☐ No ☐ Not described ☐ Cannot be determined ☐
Responsivity	17 - Were at least 75% of the hypothetical comparisons confirmed or AUC ≥ 0.70?	Yes ☐ No ☐ Not described ☐ Cannot be determined ☐
Interpretability	18 - Is data/values described for the PROM or OMI that allow interpretation of the scores obtained? Ex1: Minimum Clinically Important Change (MIC)? Ex2: cut-off value for determining subgroups? Ex3: how to interpret the score: for example, what does a high or low score mean?	Yes ☐ No ☐ Not described ☐ Cannot be determined ☐
Cross-cultural adaptation	19 - Is there a version of the PROM or OMI available in Brazilian Portuguese that has followed an appropriate method of cross-cultural adaptation?	Yes ☐ No ☐ Not described ☐ Cannot be determined ☐
	20 - Have the measurement properties of PROM or OMI been tested on a sample of Brazilians?	Yes ☐ No ☐ Not described ☐ Cannot be determined ☐

Pain Inventory Brief Pain Inventory met the criterion "yes" for 14 out of 20 items (70%). Thus, most of the measurement properties were met in several international studies.

CONCLUSION

This study (part II) looked at the measurement properties of validity and responsiveness domains, as well as interpretability. In addition, a checklist was proposed to facilitate the operationalization of knowledge about measurement properties. To meet Structural Validity, the PROM must be submitted to a factor or Rasch-type analysis. For Hypotheses Testing for Construct Validity, at least 75% of the hypotheses raised *a priori* must be confirmed. Responsiveness should be assessed through accuracy analyses (AUC≥0.70) and the Minimal Important Change (interpretability) can be used to determine whether a patient has achieved a clinically relevant improvement. Thus, this research encourages the application of the proposed checklist, which can help obtain reliable and valid data to support and assist clinicians and researchers in choosing the most appropriate instrument to support decision-making.

AUTHORS' CONTRIBUTIONS

Thais Cristina Chaves
Conceptualization, Resource Management, Project Management, Methodology, Writing - Preparation of the Original, Writing - Review and Editing, Supervision
Thamiris Costa Lima
Writing - Preparation of the Original, Writing - Review and Editing
Juliana H. Padilha Spavieri
Methodology, Writing - Preparation of the Original, Writing - Review and Editing
Ana Carolina de Jacomo Claudio
Methodology, Writing - Preparation of the Original, Writing - Review and Editing
Roger Berg Rodrigues Pereira
Methodology, Writing - Preparation of the Original, Writing - Review and Editing
Mariana Romano de Lira
Methodology, Writing - Preparation of the Original, Writing - Review and Editing

ABBREVIATIONS

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: Standardized Root Mean Square Residual
AUC: Area Under the Curve (accuracy analysis)
SDC: Smallest Detectable Change
MIC: Minimal Important Change
LoA: Limits of Agreement (Bland & Altman)

REFERENCES

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.