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

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

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

DOI 10.5935/2595-0118.20230093-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 narrative review (part I) were: (I) to introduce basic concepts about PROMs with a focus on the terminology and criteria defined through the COnsensus-based Standards for the selection of health Measurement INstruments (COSMIN), and (2) to describe the measurement properties of the reliability domain.

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

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

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

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

RESUMO

JUSTIFICATIVA E OBJETIVOS: O tipo de questionário que pretende captar a percepção/visão de um paciente sobre um aspecto a ser medido (ex: intensidade da dor) é chamado de Instrumento de Medida Baseado no Relato do Paciente (Patient Reported Outcome Measure - PROM). Um dos maiores desafios que clínicos e pesquisadores costumam enfrentar é quanto a tomar uma decisão sobre qual PROM utilizar para a avaliação de seu paciente com dor, especialmente devido à falta do letramento científico necessário para entender os critérios e termos empregados na área de propriedades de medida. Assim, os objetivos desta revisão narrativa (parte I) foram: (I) introduzir conceitos básicos sobre PROMs com enfoque na terminologia e critérios

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

1. Federal University of São Carlos, Professor in the Physical Therapy Department, Physical Therapy Department, São Carlos, SP, Brazil.
2. Federal University of São Carlos, Master's student in the Postgraduate Program in Physical Therapy, Physical Therapy Department. São Carlos, SP, Brazil.
3. Federal University of São Carlos, Postgraduate Program in Physical Therapy, Department, São Carlos, SP, Brazil.
4. Federal University of São Carlos, Master's student in the Postgraduate Program in Physical Therapy, Physical Therapy Department, São Carlos, SP, Brazil.
5. University of São Paulo, Ribeirão Preto School of Medicine, Master Professor of the Postgraduate Program in Rehabilitation and Functional Performance, São Carlos, SP, Brazil.

Submitted on September 06, 2023.
Accepted for publication on October 10, 2023.
Conflict of interests: none - Sponsoring sources: none.

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

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

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

definidos através do *COnsensus-based Standards for the selection of health Measurement INstruments* (COSMIN), e (2) descrever as propriedades de medida do domínio confiabilidade.
MÉTODOS: Utilizando uma busca voltada para os artigos da iniciativa COSMIN, foi elaborado este estudo. Sendo o assunto muito extenso, os autores dividiram o texto em duas partes.
RESULTADOS: O presente artigo descreveu conceitos básicos sobre PROMs (propósitos e construtos), o processo de adaptação transcultural e as propriedades de medida do domínio confiabilidade (confiabilidade, medida de erro e consistência interna). De forma geral, um instrumento com qualidade adequada de confiabilidade deveria atender a alguns critérios, tais como: coeficiente de correlação intraclasse > 0,70, medida de erro < mínima mudança clinicamente importante e α de Cronbach ≥ 0,70.
CONCLUSÃO: O entendimento sobre como determinar a qualidade da propriedade de medida de confiabilidade pode auxiliar os clínicos e pesquisadores na escolha dos melhores PROMs disponíveis. Um checklist para avaliação da qualidade das propriedades de medida de PROMs está descrita na parte II do artigo.
Descritores: Confiabilidade de dados, Dor crônica, Dor musculoesquelética, Inquéritos e questionários, Psicometria.

INTRODUCTION

Defining basic concepts

Applying a questionnaire/scale to capture an individual's perception/view of an aspect to be measured[1] (e.g. pain intensity) is the same as using a measurement instrument based on the individual's report. PROM (Patient Reported Outcome Measure) is the acronym for a measurement instrument based on the patient's report[1]. Another commonly used acronym is OMI (Outcome Measurement Instrument)[2]. PROM-type instruments were developed in order to assess constructs or concepts that cannot be directly measured or that would be difficult to measure in practice (e.g. performance when carrying out activities of daily living)[3]. The acronym PRO applies to the construct assessed by the instrument, while the acronym PROM applies to the measuring instrument. An example of a PRO construct is pain-related self-efficacy. One of the most famous PROMs for assessing this construct is the Pain Self-Efficacy Questionnaire (PSEQ)[4]. The use of PROMs can be very useful in the functioning of health systems, since they can: 1) be administered serially (longitudinally) to monitor patients' progress and facilitate the identification of problems; 2) help health professionals to carry out patient-centered practice; 3) evaluate and compare the efficiency and performance of practices, processes and interventions adopted; and 4) provide data to evaluate policies implemented in health services and systems[5].

PROMs can be classified, in terms of the purpose for which they were developed, as evaluative, discriminative and prognostic[6]. Instruments with an evaluative purpose were developed with the aim of monitoring changes over time (pre- and post-intervention)[6]. The SF-36 (Medical Outcomes Study 36 - Item Short - Form Health Survey)[7] instrument for assessing quality of life is designed for evaluative purposes, i.e. the SF-36 score should be used longitudinally to monitor changes over time. Instruments with a discriminative purpose are those designed to discriminate between subgroups (e.g. patients with different degrees of disability)[6].

The PHQ-9 (Patient Health Questionnaire)[8], for example, was developed to detect individuals with depression and those without. The cut-off value ≥10 points has shown high sensitivity and specificity (88%) for diagnosing individuals with depression when they are assessed in primary health care services[8]. Questionnaires with a prognostic (or predictive) purpose[6] aim to anticipate a prognosis (e.g. the future course of a disease). An example of this type of instrument is the StarT Back Screening Tool[9] (SBST), which was designed to detect the chances of an unfavorable prognosis for recovery (persistent pain) in patients with low back pain in the acute or subacute phase. The literature describes that a score ≥4 on the SBST psychosocial scale is related to an unfavorable diagnosis for patients with acute low back pain, with a risk of chronic low back pain[9].

Questionnaires can also be classified in terms of their applicability to target populations, in specific or generic cases. Generic questionnaires are those whose target population is broad. For example, the Brief Pain Inventory[10] is a questionnaire designed for chronic pain patients in general. Disability questionnaires are usually specific to a particular condition (condition-specific). Thus, the Oswestry Disability Index (ODI) is specific for assessing disability related to low back pain[11].

Questionnaires/scales are designed to try and transform a subjective aspect or concept (construct) into a quantitative measure. The questionnaire scores were assigned numbers in order to monitor the patient's progress in terms of the attribute being measured. In situations where it is not possible to measure a particular attribute, how should be established a parameter for the patient's improvement or worsening? Measuring is the procedure of identifying values of quantitative variables based on their numerical relationship with other values[12]. In order to be considered adequate and have practical applicability, every measurement needs to meet certain properties.

For example, the measurement of body temperature must be reproducible when assessed twice in a short period of time, in which the stability of the patient's clinical condition is guaranteed. A hypothetical example would be the case of a patient who arrived at the hospital and had his temperature taken, obtaining a value of 39° Celsius. After a short period (3 minutes), if the temperature is taken again, without the patient having received an antipyretic drug or having suffered any fluctuation in their clinical condition, it is expected that the thermometer will be able to obtain the same temperature or a temperature very close to the initial value. This is what is known as reliability. If a measurement is unreliable, how can a clinical decision be made safely on the basis of that measurement?

So, what are measurement properties? Measurement properties are obtained by studying the characteristics of a given measure - for example, by establishing relationships/comparisons between the score of one instrument and the score(s) of

other instrument(s) - in order to identify whether the measure (e.g. PROM or OMI score) has adequate qualities. In other words: whether the measure has adequate consistency when it is repeated (reliability), whether the measure actually measures what it intends to measure (validity) and whether the measure is capable of capturing changes over time (responsiveness). Thus, measurement properties help to identify the quality of a PROM or OMI, and the overall quality of an instrument's measurement properties can help clinicians and researchers make the decision about which PROM or OMI to use in their professional practice.

The problem of the lack of standardization of terms in the field of measurement properties and the importance of the COSMIN initiative

COSMIN is an acronym for COnsensus-based Standards for the selection of health Measurement INstruments[13-15]. As the name implies, COSMIN is an initiative for consensus and standardization of aspects related to the measurement properties of PROMs[13-15]. To prepare the various documents available online on the COSMIN platform (https://www.cosmin.nl/cosmin-tools/), various experts from different parts of the world collaborated to reach consensus on the various definitions proposed.

The COSMIN committee was inspired by a lack of clarity in the literature about the terminology and definitions of measurement properties[16]. There are a staggering number of PROMs and many of them measure the same construct[16]. So, what is the best way to define which PROM is the best to use in research or clinical practice? Helping with this decision-making process is one of the central objectives of the COSMIN initiative.

The first step was to standardize the taxonomy. In the first COSMIN[17] Delphi study (studies aimed at defining consensus on different topics through panels of experts), a consensus was established on the terminology (taxonomy) and definitions of these measurement properties[13-15]. The COSMIN initiative also aims to help determine quality criteria for these properties[2]. For example, what is acceptable as a measure of reliability? Which statistical tests should be used to measure validity?

In addition, the COSMIN initiative has come to help researchers conduct systematic reviews of measurement properties[13] by, for example, providing tools for assessing the methodological quality of studies investigating measurement properties. Systematic reviews of measurement properties help to answer the following question: does a given instrument meet the quality criteria for its measurement properties and can it therefore be used in research and clinical practice? For example, a systematic review[18] tried to determine the best instrument for measuring disability in individuals with low back pain (Oswestry Disability Questionnaire and Roland-Morris Questionnaire). This review showed that both questionnaires have limitations, and it was not possible to determine which one had better quality of measurement properties. Thus, both questionnaires were recommended by the literature. This systematic review also warned of the need for better methodological studies on the measurement properties of both questionnaires.

In the literature, there are several reports by researchers and clinicians, in the health and pain fields, about the difficulty of making decisions regarding the choice of a PROM or OMI, especially due to the difficulty of choosing between the numerous PROM options available for patient assessment, follow-up or even diagnosis, as well as the difficulty of interpreting studies of measurement properties[19]. In view of these aspects, the objectives of part I of this narrative review were: (I) to introduce basic concepts about PROMs with a focus on the terminology and criteria defined by COSMIN and to describe the process of cross-cultural adaptation, and (2) to describe the measurement properties of the reliability domain. The part II of this review, published in a second article, covered the measurement properties of validity, responsiveness and interpretability, as well as proposing a checklist for assessing the quality of PROMs.

METHODS

This study was based on studies published by the COSMIN consensus. Of the 32 references cited in this article, 10 are articles from the COSMIN initiative[2,6,13-17,22,23,31].

THE DOMAINS OF MEASUREMENT PROPERTIES ACCORDING TO COSMIN

The choice of a PROM or OMI for assessing a health condition should be based primarily on the quality of its measurement properties[13]. According to the COSMIN consensus, the quality of a PROM or OMI should be assessed through three main areas of analysis: reliability, validity and responsiveness (figure 1). COSMIN also considers that interpretability is a characteristic that should be taken into account[14].

The reliability domain of a PROM encompasses the measurement properties that describe "how error-free the measurement is"[14]. Within the reliability domain, COSMIN considers the following measurement properties: (I) reliability, (II) measurement error and (III) internal consistency[14].

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) hypothesis testing, (IV) cross-cultural validity and criterion validity.

The responsiveness domain, on the other hand, brings together just one measurement property that has the same name as the domain: responsiveness. Responsiveness is aligned with the ability of an instrument to detect changes in the score (change score) of a PROM or OMI over time[14] and in a valid way. It is a type of validity (it refers to the validity of a change score), 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[15]. Although it is not considered a measurement property, interpretability is a fun-

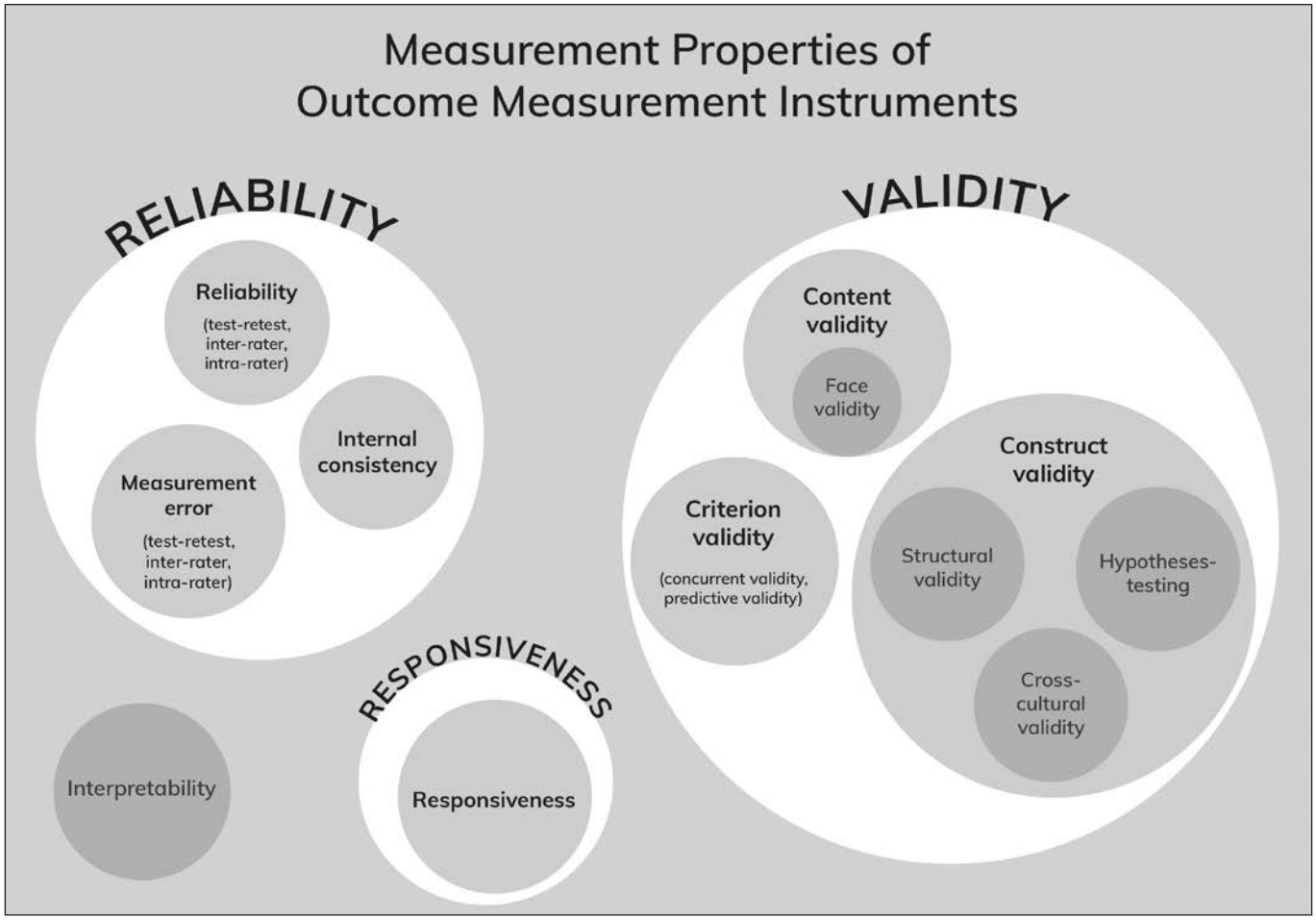

Figure 1. Diagram of the domains of the Measurement properties according to the COnsensus-based Standards for the selection of health Measurement INstrument (COSMIN)
Available in: https://cosmin.nl/wp-content/uploads/COSMIN_taxonomy.pdf.

damental characteristic of measurement instruments, despite being commonly overlooked by researchers. Like interpretability, cross-cultural adaptation is not a property of measurement, but it is an essential process for making instruments available in new languages and ensuring that assessment systems can be internationally interchangeable.

CROSS-CULTURAL ADAPTATION

The term "cross-cultural adaptation" is used to describe a process that combines translation and cultural adaptation in the process of preparing an instrument for use in another setting. Most of the instruments found in the literature were developed in English. Thus, the process of translation and cross-cultural adaptation based on an appropriate method can guarantee the equivalence of a translated instrument in relation to the target version (language into which the instrument was translated)[20].
An example that illustrates the importance of this process is The Activities-Specific Balance Confidence (ABC) Scale, which measures the individual's confidence in your balance when when performing activities. One of the questions of the ABC scale is about the individual's confidence in walking outside on icy sidewalks. This question makes sense in regions with a cold or temperate climate, where the meteorological phenomenon of snowfall is common. However, in countries with a tropical climate, such as Brazil, this question had to go through a process of cross-cultural adaptation. In the Portuguese-Brazilian version[21], this question from the ABC scale was adapted cross-culturally as follows: "walking on wet or slippery sidewalk".
The COSMIN study design checklist for patient-reported outcome measurement instruments[22] (https://www.cosmin.nl/wp-content/uploads/COSMIN-study-designing-checklist_final.pdf) describes 12 recommended items for translation and cross-cultural adaptation processes:

1. Description of the original language and target language of the translation;
2. The PROM or OMI must be translated (forward) and back-translated (backward). Forward translation is translation from the original language into the target language and backward translation is translation from the target language version back into the original language;
3. The translators of the forward translation must have the target language of the translation as their mother tongue;

4. One of the translators in the forward translation stage must be an expert on the construct to be measured and the second forward translator must be a layperson on the subject;
5. Both translators of the back-translation must have the original language of the PROM or OMI as their mother tongue;
6. Both back-translators must be laypeople for the construct to be measured by PROM or OMI;
7. It is necessary to ensure that translators work independently;
8. It is necessary to describe how the differences between the original PROM and the translated PROM were resolved;
9. Ensure that the final translated version is reviewed by a committee that includes the instrument's developers (authors of the original version);
10. Draw up a document reporting on the translation/cross-cultural adaptation process;
11. Carry out a pilot study in which comprehension, comprehensiveness and relevance are assessed in the instrument's target population for the following aspects: items, instructions, response options and memory recall period;
12. Carry out the pilot study on a population that represents the target population of the PROM or OMI.

Following all 12 steps described does not guarantee that the translated instrument is suitable for use in clinical practice and research. Therefore, after translation it is recommended that the instrument's measurement properties be tested[20], since the translated PROM or OMI should be understood as a new PROM or OMI. Therefore, the next sections of the article (part I and part II) described the main measurement properties recommended by the COSMIN initiative.

RELIABILITY DOMAIN

Reliability

Reliability is defined as the proportion of total variation in measurements that can be attributed to true differences between patients[2]. The reliability of instruments assesses the degree to which repeated measurements at different times will provide similar answers, considering clinically stable individuals. The instrument must be able to distinguish the Minimal Important Change (MIC) from the measurement error[1]. This measurement property must be obtained in longitudinal studies in which two applications of the questionnaire must be carried out (test-retest).

For statistical analysis of reliability, COSMIN recommends the following statistical tests: Intraclass Correlation Coefficient (ICC) or weighted Kappa or Spearman/Pearson Correlation ≥0.70[2]. ICC is used for continuous measures. For ordinal scales, the weighted Kappa coefficient is used. The Spearman/Pearson correlation can also be used as long as strict control of the systematic error of the measure is guaranteed in test and retest situations. These measures are interpreted as follows: the results can vary between -1 and 1. The closer to 1, the greater the reliability[23]. Correlations can show positive or negative values, which are an expression of the direction of the correlation and not its magnitude.

To measure reliability, a period is needed between applications of the instrument, which must be long enough to avoid recall bias and at the same time short enough to ensure that there is no change in the patient's stability. A commonly used period is 1 or 2 weeks. Ensuring the patient's clinical stability is essential for measuring reliability, since the aim is to minimize the effects of systematic error in the questionnaire score (which could be related to methodological factors that were not adequately controlled). The same conditions must also be guaranteed in the collection environment during test and retest.

The Western Ontario McMaster Osteoarthritis Index (WOMAC)[24] assesses pain and disability related to knee and hip osteoarthritis. In the example in figure 2, only the disability scale is considered (among the three scales that make up the instrument): the higher the score on the WOMAC disability scale, the greater the disability. The WOMAC score showed adequate reliability, with an ICC value of 0.99. This means that the questionnaire score is consistent (or very similar) in the two applications of the questionnaire. The authors described a test-retest interval of just 24 hours to ensure the stability of the clinical picture. This short time interval can be considered a methodological bias, as it may not be sufficient to minimize recall bias.

It is possible to notice that the WOMAC (Western Ontario McMaster Osteoarthritis Index) score showed good reliability in the two applications of the questionnaire, as the score remained similar when the clinical condition did not change. Some variation in the questionnaire score is always expected and the measurement error will help to identify which variation is acceptable.

Error of measurement

Every measuring instrument has an error, which can be defined as systematic or random. Systematic error is related to methodological aspects[25], such as the position of the patient or the environment of the assessment room (air-conditioned and illuminated). Random error, on the other hand, is the portion of error that cannot be controlled and can help to understand the effect of chance[26] on the PROM or OMI score. It is therefore important to quantify the error in an instrument's score, using a

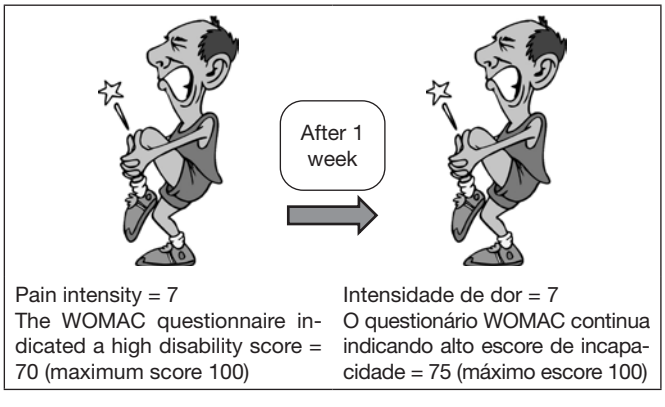

Figure 2. In the figure, we illustrate what would be an adequate reliability of a questionnaire score. We can note that the score of the WOMAC (Western Ontario McMaster Osteoarthritis Index) questionnaire demonstrates good reliability in the two administrations, because the score remains similar when the clinical condition is stable. Some variation in the score of the questionnaire is always expected and the error of the measurement will help to identify which variation is acceptable.

rigorous methodology that minimizes the error that can be controlled (systematic), in an attempt to obtain only the random error (which cannot be predicted or controlled).

The error of measurement helps researchers and clinicians to judge whether the change in the PROM or OMI score actually reflects a change in the clinical picture or whether it can be considered an error. Error of measurement is important in interpreting the score of instruments with an evaluative purpose, i.e. when applying a pre- and post-treatment questionnaire, with the value of the error of measurement it is possible to determine whether the change in the PROM or OMI score for a given patient is clinically relevant (when the value of the change is greater than the error of measurement) or whether the change in score can only be attributed to score error (when the value of the change is less than the error of measurement).

To calculate the error of measurement, the Smallest Detectable Change (SDC) is the most recommended. SDC depends on the Standard Error of Measurement (SEM) and can be calculated using the formula $SDC = 1.96 \times \sqrt{2} \times SEM$[23].

SDC reflects the minimum intra-subject change (i.e. test-retest) in the PROM or OMI score that can be interpreted as a real change above the error for a given individual[17]. In addition, the error of measurement can also be expressed by the Limits of Agreement (LoA) graph[27], which provides the error of measurement through a scatter plot, making it easier to visualize. For evaluation purposes, SDC or LoA should be < MIC of the PROM or OMI score[2].

But what is MIC? It stands for Minimal Clinically Important Change. In the context of PROMs or OMIs, it is the minimum value that identifies the least change in the PROM or OMI score over time that the patient recognizes as important[15]. In other words, it is the PROM or OMI cut-off value, which can be obtained through comparisons with a global perception of change/improvement scale, which assesses the patient's reported perception of their condition (figure 3).

The patient should indicate the numerical value that he/she believes expresses the overall improvement or worsening of their pain. Negative values indicate worsening and positive values indicate a decrease in pain.

In an example, considering that the Numerical Pain Rating Scale was used to assess the pain intensity of an individual with low back pain[28] (the following values were obtained: pre-treatment = 8 and post-treatment = 4), the SDC for Numerical Pain Rating Scale in individuals with low back pain ranged from 2.4 to 3.5 points. If the patient showed a change of 4 points post-treatment (8 - 4 = 4), can the 4-point change in the patient's pain intensity be considered a change or simply an error? Considering that the change was greater than 3.5, then the answer is: the change in score (pre- and post-treatment) is greater than the error and therefore cannot be attributed to error of measurement.

Internal consistency

Internal consistency is defined by COSMIN as "the degree of interrelationship between the items" of a PROM or OMI[2,16]. Internal consistency in the reliability domain considers the extent to which items assess the same construct, i.e. the correlation (homogeneity) between items. This correlation must be high enough for the items in a PROM or OMI to represent the same construct being measured. Internal consistency is an important measurement property for questionnaires that intend to measure a single concept (construct based on a reflective model) using several items that represent it, reflecting that construct (figure 4). For multidimensional scales, internal consistency should be assessed for each domain or subscale, based on the assumption that each domain/factor/subscale is assessing different constructs[2].

The arrows exiting the construct (depression) indicate that each PROM question is a reflection of the construct. Note that the peripheral images (questionnaire items) are very similar to the central image (construct), but different. They are therefore considered to be a non-identical "reflection" of the construct.

In contrast, for questionnaires based on a formative model, in which each item can represent different constructs, and which often do not correlate with each other, internal consistency analysis is not indicated. Internal consistency should only be assessed for unidimensional scales (or for the sub-scales of multidimensional PROMs or OMIs) and those based on a reflective model. The Apgar Scale[29], which assesses the vitality of the newborn and includes the following domains: (I) tone, (II) breathing, (III) coloration, (IV) heart rate and (V) reflexes, is an example of a scale based on the formative model, in which its items contribute to forming a construct (vitality of the baby), but do not correlate with each other, so internal consistency analysis is not applicable. Constructs based on a formative model are considered "artificial constructs", as is also the case with the quality of life construct. There is a lot of controversy on this subject in the literature. For a more in-depth study of the subject, this research recommends one of the articles reviewed[30].

Cronbach's α is the statistical measure used to estimate the internal consistency of PROMs or OMIs[5]. COSMIN considers that the quality of the measurement property is adequate for internal consistency when: (I) at least some evidence of adequate structural validity is available and (II) Cronbach's $\alpha \geq 0.70$ described for each factor/domain/dimension[31].

Compared to when this episode of _____pain first started, how would you describe your_____ these days?

-5	-4	-3	-2	-1	0	1	2	3	4	5
Vastly Worse					Unchanged					Completely Recovered

Figure 3. Global perceived effect of improvement scale. The patient must indicate the numerical value that he/she understands to express the overall improvement or worsening of his pain. Negative values indicate worsening and positive values indicate improvement in pain.

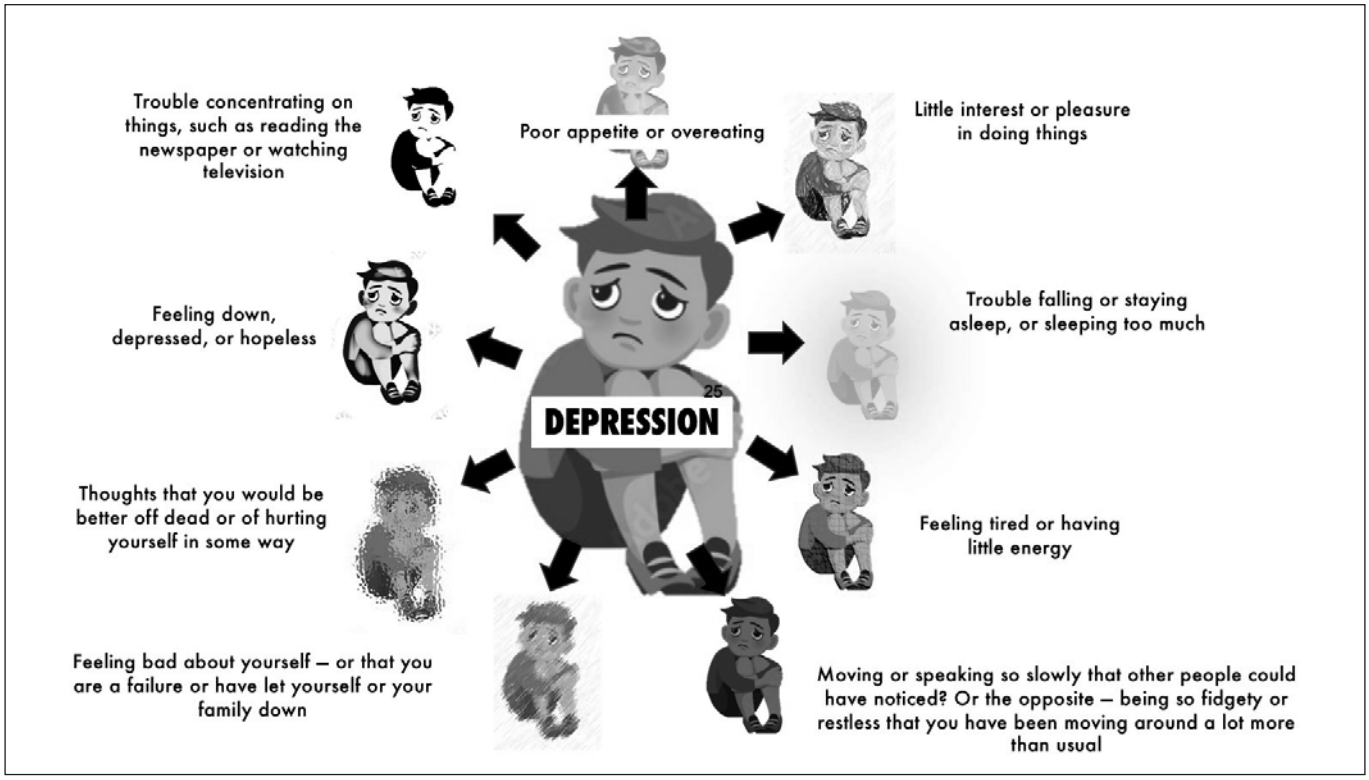

Figure 4. Illustration of the items in the PHQ-9 Questionnaire, which assesses symptoms of depression and is considered a construct based on a reflective model

In the manuscript describing the validation of the Craniofacial Pain and Disability Inventory[32] into Brazilian Portuguese, it is possible to find a practical example of the evaluation of the Cronbach's alpha coefficient. This research indicated (separately, for each factor/domain/dimension below): "Functional and Psychosocial Limitation", $\alpha = 0.86$; "Pain", $\alpha = 0.80$; and "Frequency of Comorbidities", $\alpha = 0.77$; as recommended by COSMIN, with the coefficients being considered adequate ($\alpha \geq 0.70$). It can therefore be concluded that the items contained in each factor/domain/dimension correlate well with each other and adequately represent the dimension they are intended to measure.

CONCLUSION

This narrative review (part I) sought to elucidate basic concepts about questionnaires, bringing up aspects about the importance of the COSMIN initiative for the process of cross-cultural adaptation and about measurement properties within the domain of reliability. Thus, cross-cultural adaptation must follow a method that guarantees equivalence of the translated PROM in relation to the original version, and an instrument with adequate reliability quality must meet certain criteria, such as: ICC ≥ 0.70; error of measurement < minimum clinically important change; and Cronbach's $\alpha \geq 0.70$. The reading of this article should be complemented by the reading of part II, in which the measurement properties of validity, responsiveness and interpretability were addressed, as well as the proposal of a checklist to support decision-making when choosing a suitable PROM.

AUTHORS' CONTRIBUTIONS

Thaís Cristina Chaves
Conceptualization, Resource Management, Project Management, Methodology, Writing - Preparation of the Original, Writing Review and Editing, Supervision
Ana Carolina de Jacomo Claudio
Methodology, Writing - Preparation of the Original, Writing - Review and Editing
Thamiris Costa Lima
Writing - Preparation of the Original, Writing - Review and Editing
Roger Berg Rodrigues Pereira
Methodology, Writing - Review and Editin
Gabriela Zuelli Martins Slva
Writing - preparation of the original, Writing - Review and Editing
Helen Cristina Nogueira Carrier
Writing - Review and Editing

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