COSMIN

Explanation & Elaboration document

Reporting Guideline

for studies on measurement properties of patient-reported outcome measures

Version:	Version 1.0 September 2024
Written by:	Guilherme T de Arruda Ellen BM Elsman Mariana A Avila Almir V Dibai-Filho Andrew D Firth Ava Mehdipour Christos Mousoulis Conrad J Harrison John D Peipert Karen Matvienko-Sikar Kirstie L Haywood Leonie H Klaufus Letícia B Calixtre Luiz A Brusaca Mark A Dubbelman Nan Luo Patricia Driusso Sandrine Herbelet Caroline B Terwee Joel J Gagnier Lidwine B Mokkink
	The development of the Explanation & Elaboration (E&E) document for the COSMIN Reporting Guideline 2.0 for studies on measurement properties of patient-reported outcome measures was financially supported by the Fundação de Amparo à Pesquisa do Estado de São Paulo (2021/11871-9 and 2022/16124-0).
Corresponding author	LB Mokkink, PhD w.mokkink@amsterdamumc.nl Amsterdam UMC Department of Epidemiology and Data Science Amsterdam Public Health research institute Amsterdam The Netherlands www.cosmin.nl

© G. Tavares de Arruda, L.B. Mokkink, J. Gagnier, and C.B. Terwee. 2025.

This publication is in copyright. Subject to statutory exception and to provisions of relevant collective licensing agreements, no reproduction of any part may take place without the written permission of Amsterdam UMC.

Table of contents

Contents

Background 5
The COSMIN reporting guideline for studies on measurement properties of PROMs 9
Explanation & Elaboration 10
Terminology and abbreviations 11
General Reporting recommendations relevant for all studies on measurement properties 13
Report section: Title 13
T1 - Title 13
Report section: Abstract 15
A1 - Objectives 15
A2 - Design 17
A3 - Methods 18
A4 - Results 20
A5 - Discussion/Conclusions 22
Report section: Introduction 24
I1 - PROM 24
I2 - Target population & context if use 26
I3 - State of knowledge & Rationale 27
I4 - Objectives 30
Report section: General Methods 32
GM1 - Study design 32
GM2 - Participants 33
GM3 - PROM details 35
GM4 - Additional data collection 37
GM5 - Time points procedures 39
GM6 - Justification for sample size 41
GM7 - Statistical analyses 43
GM8 - Missing data 44
GM9 - Unplanned analysis 45
Report section: General Results 46
GR1 - Participant characteristics 46
GR2 - Sample size 48
GR3 - Missing data 49
GR4 - Results 50
Report section: Discussion/conclusions 53
DC1 - Measurement property evidence 53
DC2 - Practical relevance 54
DC3 - Strengths and limitations 56
DC4 - Generalizability 59
DC5 - Instrument changes 61
DC6 - Future research 63

DC7 - Conclusions 65
Report section: Other information 67
O1 - Conflict of Interest 67
Specific Reporting recommendations for each measurement property 68
Content Validity 68
Content Validity: Methods 68
CV1 - Relevance 68
CV2 - Comprehensiveness 71
CV3 - Comprehensibility 73
Content Validity: Results 75
CV4 - Relevance 75
CV5 - Comprehensiveness 77
CV6 - Comprehensibility 79
Structural Validity 80
Structural Validity: Methods 80
SV1 - Rationale for approach 80
SV2 - Statistical analyses 82
Structural Validity: Results 85
SV3 - Statistical analyses 85
Internal Consistency 90
Internal Consistency: Methods 90
IC1 - Statistical analyses 90
Internal Consistency: Results 92
IC2 - Statistical analyses 92
Cross-Cultural Validity\Measurement Invariance 93
Cross-Cultural Validity\Measurement Invariance: Methods 94
CCV1 - Groups variable 94
CCV2 - Rationale for approach 95
CCV3 - Statistical analyses 96
Cross-Cultural Validity\Measurement Invariance: Results 98
CCV4 - Statistical analyses 98
Reliability 101
Reliability: Methods 101
R1 - Stability of patients 101
R2 - Statistical analyses 103
R3 - Source of variation 105
R4 - Time interval 106
Reliability: Results 107
R5 - Stability of patients 107
R6 - Statistical analyses 108
Measurement Error 109
Measurement Error: Methods 109
ME1 - Stability of patients 109
ME2 - Statistical analyses 111

ME3 - Source of variation 112
ME4 - Time interval 113
Measurement Error: Results 114
ME5 - Stability of patients 114
ME6 - Statistical analyses 115
Criterion Validity 116
CriV1 - Criterion 116
Criterion Validity: Methods 117
CriV2 - Statistical analyses 117
Criterion Validity: Results 118
CriV3 - Statistical analyses 118
Hypotheses Testing for Construct Validity 120
Hypotheses Testing for Construct Validity: Methods 120
ConV1 - Hypotheses 120
ConV2 - Statistical analyses 123
Hypotheses Testing for Construct Validity: Results 125
ConV3 - Statistical analyses 125
Responsiveness 127
Responsiveness: Methods 127
Resp1 - Hypotheses 127
Resp2 - Intervention/Exposure 129
Resp3 - Statistical analyses 130
Responsiveness: Results 132
Resp4 - Statistical analyses 132
References 134

Background

The use of Patient-Reported Outcome Measures (PROMs) in clinical practice and research to assess patient-reported health status has increased significantly. PROMs must be sufficiently reliable, valid and responsive to trust its scores obtained. To know the quality of a PROM, you need to conduct studies on measurement properties. Measurement properties are the quality aspects of a measurement instrument, referring to its validity, reliability, and responsiveness [1]. Studies on measurement properties should be of high methodological quality to ensure appropriate and trustworthy conclusions about the measurement properties of a measurement instrument (e.g., PROMs). Nine measurement properties have been defined by the COSMIN (COnsensus-based Standards for the selection of health Measurement INstruments) initiative, as shown in Figure 1 and Table 1 [2].

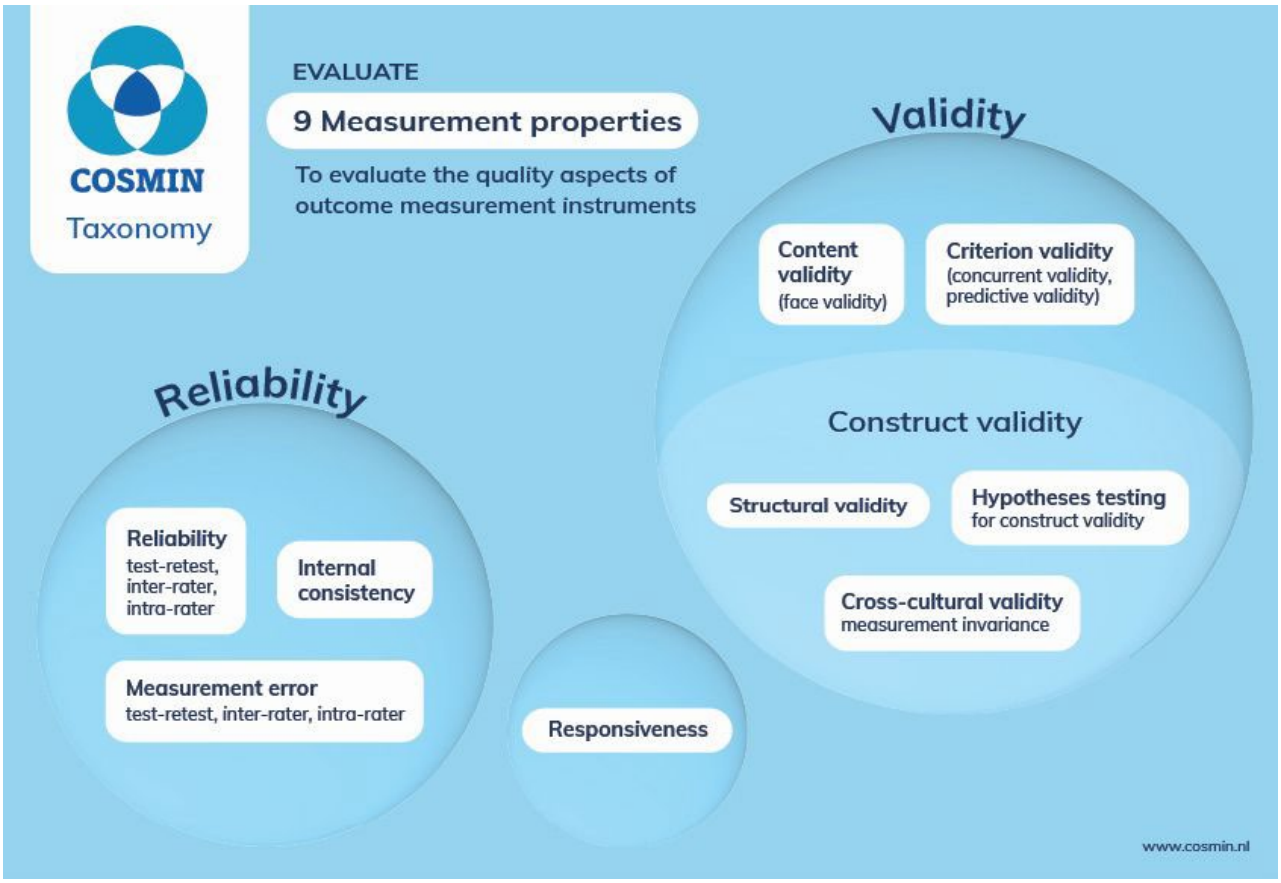

Figure 1 The COSMIN taxonomy [3]

Table 1 COSMIN definitions of domains, measurement properties, and aspects of measurement properties [1]

Domains	Definitions
Reliability	The degree to which the measurement is free from measurement error
Reliability (extended definition)	The extent to which scores for patients who have not changed are the same for repeated measurement under several conditions: e.g., using different sets of items from the same PROM (internal consistency); over time (test-retest); by different persons on the same occasion (inter-rater); or by the same persons (i.e., raters or responders) on different occasions (intra-rater)
Validity	The degree to which a PROM measures the construct(s) it purports to measure
Construct validity[1]	The degree to which the scores of a PROM are consistent with hypotheses (for instance with regard to internal relationships, relationships to scores of other instruments, or differences between relevant groups) based on the assumption that the PROM validly measures the construct to be measured
Responsiveness	The ability of a PROM to detect change over time in the construct to be measured
Measurement properties	**Definitions**
Content validity	The degree to which the content of a PROM is an adequate reflection of the construct to be measured
Face validity[2]	The degree to which (the items of) a PROM indeed looks as though they are an adequate reflection of the construct to be measured
Structural validity	The degree to which the scores of a PROM are an adequate reflection of the dimensionality of the construct to be measured
Internal consistency	The degree of the interrelatedness among the items
Cross-cultural validity	The degree to which the performance of the items on a translated or culturally adapted PROM are an adequate reflection of the performance of the items of the original version of the PROM
Reliability	The proportion of the total variance in the measurements which is due to 'true'[3] differences between patients
Measurement error	The systematic and random error of a patient's score that is not attributed to true changes in the construct to be measured

Criterion validity	The degree to which the scores of a PROM are an adequate reflection of a 'gold standard'
Hypotheses testing for construct validity	The degree to which the scores of a PROM are consistent with hypotheses (with regard to relationships to scores of other instruments, or differences between relevant groups) based on the assumption that the PROM validly measures the construct to be measured
Responsiveness	The ability of a PROM to detect change over time in the construct to be measured
Other relevant properties	**Definitions**
Feasibility[4]	Ease of application of the PROM in its intended context of use, given constraints such as time or money [4].
Interpretability[4]	The degree to which one can assign qualitative meaning - that is, clinical or commonly understood connotations – to a PROM's quantitative scores or change in scores.

[1] construct validity has various forms (i.e., measurement properties), i.e., structural validity, hypotheses testing for construct validity, and cross-cultural validity\measurement invariance; [2] face validity is an unstructured form of content validity; [3] The word 'true' must be seen in the context of the CTT, which states that any observation is composed of two components – a true score and error associated with the observation. 'True' is the average score that would be obtained if the scale were given an infinite number of times. It refers only to the consistency of the score, and not to its accuracy [5] [4] feasibility and interpretability are not considered a measurement property, but an important aspect when selecting measurement instruments.

COSMIN guidelines

COSMIN developed several guidelines. In Figure 2, we show when to use which COSMIN guideline.

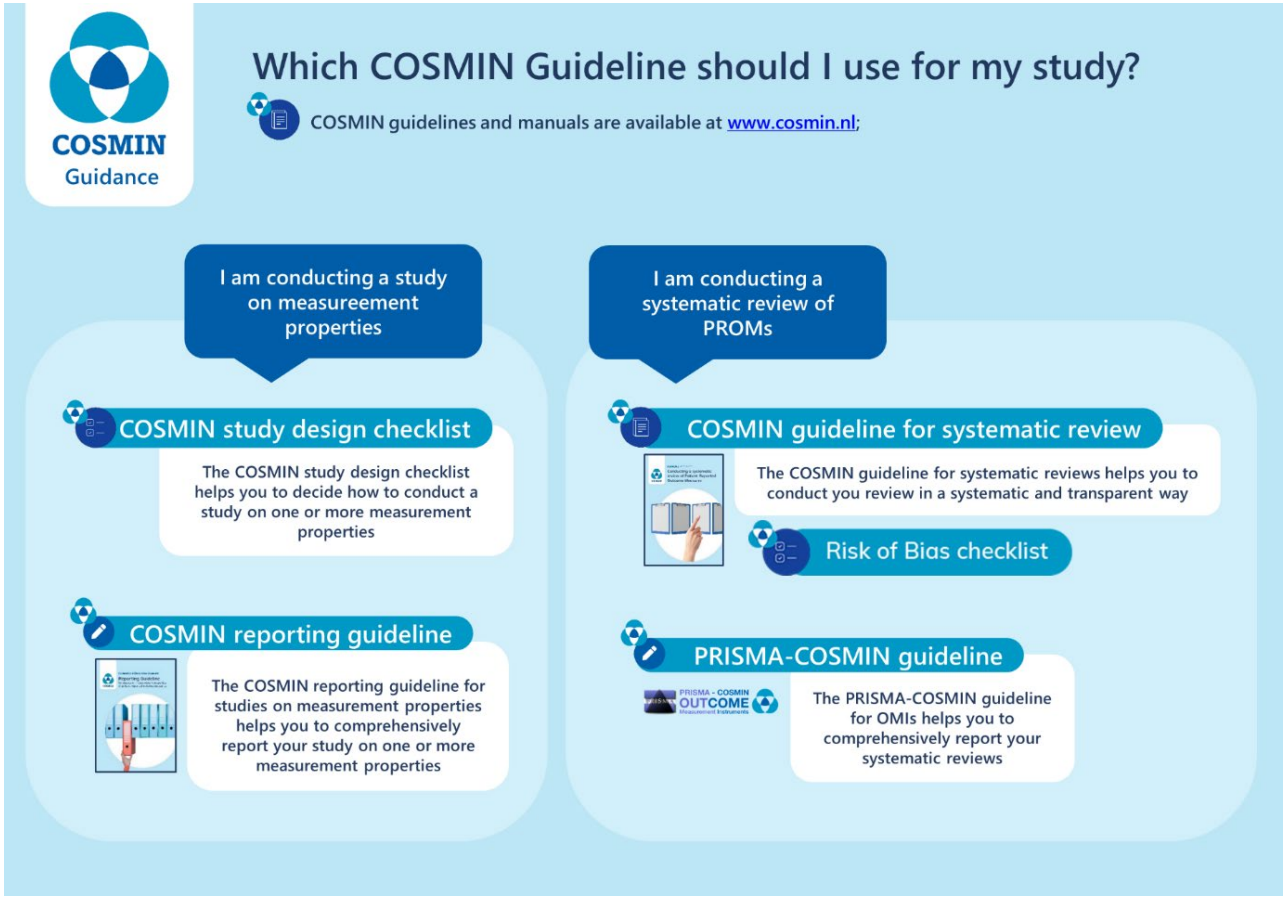

Figure 2 Overview of COSMIN guidelines

The COSMIN Reporting Guideline focuses on the reporting of studies on the evaluation of measurement properties PROMs. It does not focus on the development of PROMs. Moreover, it is not intended to guide the conduct of measurement property studies; therefore, you can use the COSMIN study design checklist. Neither is it intended to assess the methodological quality of studies on measurement properties; therefore, you can use the COSMIN Risk of Bias Checklist, which is part of the COSMIN guideline for systematic reviews [6].

The COSMIN reporting guideline for studies on measurement properties of PROMs

Studies on measurement properties of PROMs must be fully reported so that readers have all the information they need to understand the study and to verify the interpretation and conclusions [7]. The first version of the COSMIN Reporting Guideline for studies on measurement properties of PROMs was developed in 2021 [7], and updated in 2024 [6]. The purpose of this guideline is guiding researchers on what must be reported in studies on measurement properties of PROMs.

Version 2.0 of the guideline includes a set of 68 items and are separated in two sections. The "General reporting recommendations relevant for all studies on measurement properties" section contains 31 items grouped into seven general aspects of a reporting study on measurement properties:

- Title (1 item);
- Abstract (5 items);
- Introduction (4 items);
- General methods (9 items);
- General results (4 items);
- Discussion/conclusions (7 items);
- Other information - conflict of interest (1 item).

The "Specific reporting recommendations for each measurement property" section contains 37 items grouped into nine measurement properties and identifies specific issues on methods and analyses that should be reported for each of the measurement property:

- Content validity (6 items);
- Structural validity (3 items);
- Internal consistency (2 items);
- Cross-cultural validity\measurement invariance (4 items);
- Reliability (6 items);
- Measurement error (6 items);
- Criterion validity (3 items);
- Hypotheses testing for construct validity (3 items);
- Responsiveness (4 items).

The COSMIN Reporting Guideline is developed as a modular tool, meaning when using the guideline, you should always use the items from the “General reporting recommendations relevant for all studies on measurement properties” section, and choose the relevant sections of the “Specific reporting recommendations for each measurement property” referring to the measurement property that you studied and want to report. For example, if you assessed the structural validity and internal consistency of a PROM, you can use the general items, and the specific items of the subsections on structural validity and internal consistency.

How to cite this document

Please refer to this document as Explanation and Elaboration document of the COSMIN Reporting Guideline 2.0 for reporting studies on measurement properties of patient-reported outcome measures:

Arruda GT, et al. Explanation & Elaboration document of the COSMIN Reporting Guideline 2.0 for studies on measurement properties of patient-reported outcome measures. Quality of Life Research, 2025. Doi: 10.1007/s11136-025-03949-4

Explanation & Elaboration

An Explanation & Elaboration (E&E) document has not yet been developed for version 1.0 of the reporting guideline to facilitate its use. An E&E document provides detailed and comprehensive explanations about a particular topic and offers practical examples, ensuring researchers understand the nuances of using the guideline effectively. In the E&E document, each item is explained with reasons *why* reporting of the item is recommended and examples of good reporting from open access published studies. Each item in this document is divided into ‘Item identification’, ‘Explanation’, ‘Essential elements’ and ‘Item examples’. For the examples of good reporting, we have removed all citations within the examples to avoid potential confusion with our citation, removed names of authors and spelt out abbreviations to aid comprehension.

Terminology and abbreviations

Box 1 provides a glossary of terms used throughout this document.

Box 1 Glossary of terms used in E&E document for the COSMIN reporting guideline for studies on measurement properties of PROMs

Measurement properties The quality aspects of a measurement instrument, referring to its validity, reliability, and responsiveness of the instrument's score [1]. Each measurement property requires its own study design and statistical methods for evaluation. Different definitions for measurement properties are being used. COSMIN developed a taxonomy with consensus-based definitions for measurement properties [1]. Another term for measurement properties is psychometric properties.
Outcome domain Refers to *what* is being measured (e.g., fatigue, physical function, blood glucose, pain intensity). Other terms for outcome domain include outcome, construct, concept, latent trait, factor, attribute.
Patient-Reported Outcome "Any report of the status of a patient's health condition that comes directly from the patient, without interpretation of the patient's response by a clinician or anyone else." [8]
Patient-Reported Outcome Measures A measurement instrument (often a questionnaire) completed by patients to measure Patient-Reported Outcomes (PRO). "A PRO instrument can be used to measure the effect of a medical intervention on one or more concepts (i.e., the thing being measured, such as a symptom or group of symptoms, effects on a particular function or group of functions, or a group of symptoms or functions shown to measure the severity of a health condition)". [8]
Measurement properties' results The findings of a study on a measurement property. Measurement properties' results have different formats, depending on the measurement property. For example, reliability results might be the estimate of the intraclass correlation coefficient (ICC), or structural validity results might be the factor loadings of items to their respective scales and the percentage of variance explained.

Box 2 List of acronyms used in E&E document for the COSMIN reporting guideline for studies on measurement properties of PROMs

AUC	Area under the ROC Curve
CAT	Computer Adaptive Test
CFA	Confirmatory Factor Analysis
CFI	Comparative Fit Index
COSMIN	COnsensus-based Standards for the selection of health Measurement INstruments
DIF	Differential Item Functioning
E&E	Explanation & Elaboration
EFA	Exploratory Factor Analysis
ICC	Intraclass Correlation Coefficient
IRT	Item Response Theory
KMO	Kaiser-Meyer-Olkin
LoA	Limits of Agreement
MGCFA	Multigroup Confirmatory Factor Analysis
MI	Modification Indices
MIC	Minimal Important Change
MLR	Maximum Likelihood with Robust standard errors
PRO	Patient-Reported Outcomes
PROM	Patient-Reported Outcome Measures
RMSEA	Root Mean Square Error of Approximation
SD	Standard Deviation
SDC	Smallest Detectable Change
SEM	Standard Error of Measurement
SRMR	Standardized Root Mean Squared Residual
TLI	Tucker-Lewis Index

General Reporting recommendations relevant for all studies on measurement properties

Report section: Title

T1 - Title

Identify the report as a study of one or more measurement properties of a specific PROM to measure a specified construct in a specified population.

Explanation: With a clear and precise title for your article, you can facilitate your reader's understanding of what is described in your article. A descriptive title also facilitates indexing and subsequent retrieval in electronic databases such as Medline and EMBASE [7]. We recommend that you identify your report as a study of one or more measurement properties of a specific PROM to measure a specified construct in a specified population. Details about the specific PROM refer to the name and the version of the PROM. By including the PROM version in the title, especially in cases where different versions exist (e.g., SF-36, SF-12, SF-8a), translations (e.g., Hungarian version of the Female Sexual Function Index – FSFI-H), or versions for specific populations (e.g., FSFI for breast cancer patients – FSFI-BC), you allow for easy identification as to which specific PROM is evaluated in your article. The rationale for this is that each version or modified PROM is considered a unique PROM. If the name of the PROM does not clearly identify the construct being measured, we recommend including in the title the construct the PROM intends to measure (e.g., Numerical Rating Scale to assess the severity of knee pain, Visual Analogue Scale to assess the severity of fatigue related to physical exertion, or Barthel Index for measuring performance in activities of daily living). When examining multiple measurement properties, you could state in the title that "measurement properties were evaluated" or "the quality of [PROM] was evaluated". It is also reasonable to use terms for measurement domains (e.g., validity, reliability, responsiveness) in the title. In that case, it is important to make clear in the abstract and manuscript which measurement properties within each domain are being evaluated. We also recommend that you use the terminology of measurement properties according to the international consensus-based COSMIN terminology [1]. To facilitate understanding of the generalizability of the results, we recommend that you describe the study population in terms of the key disease and demographic characteristics.

Essential elements

- Specific name and version of the PROM.
- The construct the PROM intends to measure.
- Specific measurement properties evaluated, using COSMIN terminology.
- Key disease and demographic characteristics of the study population.

General example

"The [measurement properties/specific measurement property(ies) of interest] of [name/version of the PROM(s)] for measuring [outcome of interest] in [population of interest]"

Example 1

"Structural validity of the Pittsburgh Sleep Quality Index among medical students in Iran" [9]

Example 2

"The Brazilian-Portuguese version of the Upper Extremity Functional Index (UEFI): Translation, cross-cultural adaptation and measurement properties for Brazilian adults after a burn injury" [10]

Report section: Abstract

It is imperative to provide a comprehensive interpretation of the study on measurement properties of PROMs within the broader context of existing evidence. However, the specific guidance and word limitations imposed by journals add complexity to the task of constructing a compelling abstract. Therefore, we recommend that you select the most relevant details to report in your manuscript abstract.

A1 - Objectives

Provide the specific objective(s) of the research, specifying (1) the name (and version, if relevant), and construct(s) of the PROM, (2) the measurement properties being evaluated, and (3) relevant study characteristics.

Explanation: Objectives are specific statements on the various accomplishments the research intends to achieve. We recommend that you provide the specific objective(s) of your study in the abstract. The objectives should include the **measurement properties (or domains) being studied**, the **name (and version, if relevant)** and **construct(s) the PROM intends to measure**, and a general description of the **study sample characteristics**. Note that some of these aspects may also be reported in an introduction sentence or in the methods section of the abstract. Indicating the measurement properties being investigated in the abstract provides information that will allow the readers to decide whether or not the full paper is relevant to them. If the character/word limit allows, we recommend that you clearly state which specific measurement properties are being studied, according to the COSMIN terminology [1]. However, if many measurement properties are studied, the abstract should at least indicate which domains (i.e., reliability, validity, responsiveness) are studied, since details about which specific measurement properties are evaluated in the paper may become clear in the methods or results section of the abstract. Providing key information about the PROM being evaluated (e.g., name, version, and construct in intends to measure) in the abstract, again helps your readers decide whether to access the full report. If you employ an item bank, such as one of the PROMIS item banks, we recommend that you explicitly state the version (e.g., PROMIS item bank v1.0). In situations where the construct to be measured is not part of the PROM's name, or when multiple constructs are being measured (e.g., in a multidimensional PROM), we recommend that you explicitly describe and define each of the constructs the PROM intends to measure in the abstract (e.g., in the introduction sentence of the abstract). A description of the study sample characteristics in terms of relevant disease characteristics and/or relevant demographic characteristics, helps the reader of your study to understand the applicability or generalizability

of the results to specific patients, subgroups or populations. Measurement properties are specific to the population in which they were evaluated. In many cases it may not be reasonable to generalize the results from one study to another population [11].

Essential elements

• Specific objective(s) of the research including the name – and version – of the PROM, the construct being measured, the measurement properties being studied, and the study sample characteristics.

General example

"The aim of this study was to assess the [measurement properties/specific measurement property(ies) of interest] of the [name/version of the PROM(s)] for measuring [outcome of interest] in [study sample characteristics]"

Example 1

"This study evaluated the reliability and validity of the Sinhala version of Katz index of activities of daily living (ADL) in assessing the functional status of older people aged >65 years." [12]

Example 2

"The current study investigated the structural validity and internal consistency of the Strengths and Stressors (SSF) questionnaire (...) used in Swedish habilitation services to measure the positive and negative consequences that the fostering of a child with a developmental disability can have on family functioning." [13]

A2 - Design

Specify (details of the) study design used to evaluate the measurement properties.

Explanation: We recommend reporting key elements of the study design. For example, it should be clear whether the study was a cross-sectional or longitudinal study and whether any interventions were given. For example, if the investigation translated and assessed the content validity of a PROM using a qualitative method, we recommend reporting the design as a "translation and qualitative study". If the study assessed the structural validity, internal consistency, cross-cultural validity, hypothesis testing for construct validity, and/or criterion validity of a PROM, the study design is typically referred to as a "cross-sectional study". If the study assessed test-retest reliability and measurement error of a PROM, it should be reported as an "observational study with repeated measurements in stable patients" or a "longitudinal study". If the study assessed responsiveness, it could be reported it as a "longitudinal study" or "prospective cohort study".

Essential elements

• The study design in terms of number of outcome assessment timing (e.g., cross-sectional, test-retest design, longitudinal), data type (e.g., qualitative), and/or exposure/intervention allocation method (e.g., observational, randomized study).

Example 1

"This is a cross-sectional study in which we used confirmatory factor analysis to identify the best structure of the Roland-Morris Disability Questionnaire (RMDQ)." [14]

Example 2

"A longitudinal study was performed, (…) patients with advanced CKD chronic kidney disease (CKD) were invited by e-mail to complete the PROMs digitally at the KLIK research platform (www.hetklikt.nu) at 3 time points; at inclusion (i.e. baseline), after 2 weeks (for assessing test-retest reliability, as described in a separate paper [ref]) and after 6 months." [15]

A3 - Methods

Specify the methods for evaluating each measurement property.

Explanation: To help your readers to understand how you conducted your study, it is important to report the main methods for evaluating each measurement property, including the statistical approaches that you used in your study. As a minimum, we advise the following be reported: For content validity, whether patients and/or professionals/experts assessed the relevance, comprehensiveness and comprehensibility of the PROM; In structural validity studies, the approach used to analyze data (e.g., factor analysis, Item Response Theory (IRT)/Rasch analysis); For internal consistency, report that internal consistency for each unidimensional scale was assessed; For cross-cultural validity\measurement invariance, provide the characteristics of groups being compared; For test-retest reliability and measurement error, whether intraclass correlation coefficient (ICC) and standard error of measurement (SEM) were calculated; For criterion validity, report which criterion was used; For hypotheses testing for construct validity and responsiveness, which if any hypotheses were tested (e.g., by using a word as 'expectations', 'hypotheses', 'hypothesized' or 'we/as expected' - in the results); In responsiveness studies, report if the patients received any interventions or exposures in the interim period. By clearly specifying the methods in the abstract, a reader can quickly grasp the study's approach and methodology and the article can be indexed accurately and specifically. To avoid redundancy, some details of methods or statistical approaches may not be described in the methods section, but instead can be reported in the results section of the abstract. For example, the range of the time interval in a test-retest design or the specific ICC formula used.

Essential elements

- For content validity, whether professionals and/or patients assessed the relevance, comprehensiveness and comprehensibility of the PROM.
- For structural validity, the approach used to analyze data.
- For internal consistency, calculation for each unidimensional scale.
- For cross-cultural validity\measurement invariance, the characteristics of groups being compared.
- For test-retest reliability and measurement error, whether intraclass correlation coefficient (ICC) and standard error of measurement (SEM) were calculated.
- For criterion validity, which criterion was used.
- For hypotheses testing for construct validity and responsiveness, report that specific hypotheses were tested.

• For responsiveness, if the patients received any interventions or exposures in the interim period.

Example 1

"Qualitative research included 35 SIBO Small Intestinal Bacterial Overgrowth (SIBO) patients in three study stages, using a hybrid concept elicitation (CE)/cognitive interview (CI) method with US patients, ≥ 18 years. Stage 1 included a literature review, clinician interviews, and initial CE interviews with SIBO patients to identify symptoms important to patients for inclusion in the Symptom Measure (SSM). Stage 2 included hybrid CE/CI to learn more about patients' SIBO experience and test the draft SSM. Finally, stage 3 used CIs to refine the instrument and test its content validity." [16]

Example 2

"(…) structural validity was evaluated by exploratory and confirmatory factor analysis; internal consistency by Cronbach's Alpha [for each unidimensional scale]; measurement invariance by multigroup confirmatory factor analysis between geographic regions of Brazil; and construct validity [by testing a hypothesis about the correlation between] WHODAS 2.0 to the Numerical Rating Scale for pain severity." [17]

Example 3

"27 patient-reported outcomes version of the common terminology criteria for adverse events (PRO-CTCAE®) items representing 14 symptomatic adverse events (AEs) were collected using both a 24-hour recall (24 h) and the standard 7 day recall in a sample of patients receiving active cancer treatment (n = 113). Using data captured with a PRO-CTCAE-24h on days 6 and 7, and 20 and 21, we computed intra-class correlation coefficients ($ICC_{[agreement]}$) (…)" [18]

A4 - Results

Provide the main results for all measurement properties evaluated.

Explanation: It is important to provide the main results for all measurement properties being studied. If the word limit does not allow, it is reasonable to report whether the measurement properties are sufficient, insufficient or unknown for the PROM being studied. Briefly reporting the main results for each measurement property can help readers assess whether or not the measurement property(ies) is/are sufficient for the PROM in the population of interest.

Essential elements

- Main results presented for each measurement property investigated.

Example 1

"Patients confirmed Patient-Reported Outcomes Measurement Information System®-Fatigue Short Form 7a (PROMIS-SF-7a) items were relevant to assessing fatigue, instructions and response options were clear, and its 7-day recall period was appropriate." [19]

Example 2

"WHODAS 2.0 presented a single factor by exploratory factor analysis and adequate model by confirmatory factor analysis (CFI = 0.924, TLI = 0.900, RMSEA = 0.038), excellent internal consistency (α = 0.89) for all items and an invariancy across geographic regions (ΔCFI ≤ 0.01 and ΔRMSEA < 0.015)." [17]

Example 3

"Construct and known-groups validity against selected PROs were overall strong and within ranges hypothesized a priori." [20]

Example 4

"Convergent validity was confirmed with high-to-moderate correlations ($r \geq 0.4$) for 12/15 hypothesized relationships between pairs of domains and/or total scores. Discriminant validity was confirmed with low correlations ($r < 0.3$) observed for 5/7 hypothesized relationships. Known-groups validity was confirmed with significant differences ($p \leq 0.05$) in mean Congenital Thrombotic Thrombocytopenic Purpura–Patient Experience Questionnaire (cTTP-PEQ) scores between the two Patient Global Impression of Severity (PGI-S) groups for most domains and items at both timepoints. Cronbach's alpha was 0.88 at baseline and 0.91 at Day 14, confirming internal consistency of the instrument. Test-retest reliability was also confirmed with a high $ICC_{[agreement]}$ (0.96)." [21]

Example 5

"Most results were in accordance with expectations (70–91% of hypotheses confirmed), with some exceptions for PROMIS Anxiety and Ability to Participate (60% and 42% of hypotheses confirmed, respectively). For PROMIS Anxiety and Depression correlations with the Global Rating Scales (GRS) were too low (0.04 and 0.20, respectively)." [15]

A5 - Discussion/Conclusions

Provide a brief statement of the implications of the findings in the context of existing evidence on the PROM.

Explanation: The discussion/conclusions section within the abstract holds significant importance, as this is often the most read part of the paper. The main findings of the study should be clearly communicated related to the study objectives. It is important to discuss the clinical and research implications, as well as the generalizability of the findings. Additionally, space permitting, it is important to discuss what further research is needed related to the evaluated PROM. For example, you may state that the PROM can be used (all measurement properties are evaluated and sufficient), or the PROM should not be used (one of the measurement properties is insufficient, and it first needs to be improved), or the PROM has the potential to be used, the tested measurement properties are sufficient, but some are still lacking.

Essential elements

• Conclusions in the context of existing evidence and need for future research.

Example 1

"Sufficient structural validity, internal consistency and measurement invariance were found, both in the entire sample and in subsamples with and without chronic diseases. Requirements for sufficient evidence for construct validity were (…) met for most subscales. Future studies should investigate test-retest reliability, measurement error, and responsiveness of the PROMIS-29." [22]

Example 2

"An active collaboration with patients and health-care professionals has supported the co-production of a potential new PROM of axial spondyloarthritis (axSpA) fatigue, underpinned by strong evidence of face and content validity." [23]

Example 3

"The study findings show that the items in the SF-36v2® Health Survey Acute (SF-36v2) are applicable to adults with hypoparathyroidism (HP). (...) The SF-36v2 is therefore recommended for usage in clinical trials examining adults with HP, although it is recommended that this generic measure be supplemented with disease-specific instruments such as the recently developed Hypoparathyroidism Patient Experience Scale-Symptom (HPES-Symptom) and Hypoparathyroidism Patient Experience Scale-Impact (HPES-Impact) measures." [24]

Example 4

"The Italian Version of the Palatable Eating Motives Scale (PEMS-IT) appears to be an instrument with promising psychometric properties and potential applications in clinical settings. However, it also has some limitations, and future studies could focus on improving the semantic content of the elements to increase the overall utility and precision of the instrument." [25]

Report section: Introduction

I1 - PROM

Specify the name and, if relevant, the version, and construct(s) of the PROM.

Explanation: A reader of your article needs to clearly understand which PROM you are studying, so it is important to provide key information about the PROM (i.e., the **name** and, if applicable, **the version of the PROM**). The version of the PROM must be provided, especially in cases where multiple versions exist, such as short forms, translations, variations in item banks or versions for specific populations (see A1). The **construct(s)** the PROM intends to measure and the definition of the construct(s) should also be reported. Ideally the definition of the construct(s) should be taken from the original instrument development investigations or subsequent work which clarifies the construct(s). It could be helpful to use a conceptual model, such as the model proposed by Wilson and Cleary [26], the ICF model [27], or the OMERACT filter [28], to define what is (and what is not) measured by the PROM. When the PROM under study consists of multiple subscales, it is important to describe the construct and definition each of these subscales purports to measure. Sometimes in such multidimensional scales, a total score is also used, referring to an overarching or higher order construct measured by the PROM as a whole, which should also be described and defined. In some cases, a citation of the PROM development paper or investigation (and other relevant citations) may be sufficient, or some details (e.g., the definitions of the constructs of the subscales) can be provided in the Methods section, instead of in the Introduction section.

Essential elements

• The name and version of the PROM.

• The name of each construct measured by the PROM, and the definition of each construct, being investigated.

• Citations to the original development paper(s) and other highly relevant citations describing the PROM, its subscales, if any, and the definitions of all constructs.

Example 1

"(...) the Pelvic Floor Distress Inventory (PFDI-20) [ref] is a PROM that is often used in clinical practice and clinical trials to assess the distress caused by the presence of pelvic floor dysfunction (PFD) (...) PFDI-20 assesses the distress of pelvic organ prolapse (POP), anorectal and urinary symptoms in three subscales, respectively, Pelvic Organ Prolapse Distress Inventory (POPDI-6), Colorectal-Anal Distress Inventory (CRADI-8) and Urinary Distress Inventory (UDI-6) [ref]. According to the PFDI-20 development study, (...) [ref]." [29]

Example 2

"Chen et al. [ref] developed the Dysmenorrhea Symptom Interference (DSI) scale. This scale measures dysmenorrhea symptom interference with physical, mental, and social activities." [30]

I2 - Target population & context of use

Specify the target population and context of use that the PROM was designed for.

Explanation: Some PROMs are tailored towards a specific target population during development to best measure a construct of interest within that group [11]. This **target population** for which the PROM was developed may be different from the population of interest in which you subsequently evaluate the PROM (that is, your study population). Therefore, it is important to specify the target population for which the PROM was originally developed. This might include relevant disease (e.g., type, severity, etc.) and demographic (e.g., age, gender, language/country, etc.) characteristics of the target population. In addition, the **context of use** for which the PROM was developed may be different from the context of use for which it now is being evaluated. Therefore, it is important to specify the context of use that the PROM was designed for. Context of use may be described in terms of the purpose of measurement (e.g., evaluative or discriminative), and a particular setting (e.g., hospital, home). Also, specify the country in which the original PROM was developed. Presentation of the characteristics of the target population and the context of use for which the PROM was designed, assists readers of your study in establishing the appropriateness and relevance of the PROM for the intended application and provides a comprehensive understanding of its utility and applicability in the specified context.

Essential elements

• Relevant disease and demographic characteristics of the target population for which the PROM was developed.

• Context of use for which the PROM was designed, in terms of purpose of measurement, and intended setting, including country.

Example 1

"(...) the CLEFT-Q, a condition-specific PROM for people aged 8 to 29 years, born with a cleft lip or palate (CL/P) or other craniofacial conditions." [31]

Example 2

"Initially, the Hospital Anxiety and Depression Scale (HADS) was developed for adults receiving treatment for physical health problems in general hospitals [ref]." [32]

I3 - State of knowledge & Rationale

Provide a description of the current scientific knowledge (what is known and not known) regarding the measurement properties of the PROM. Explain why the new study is necessary. Provide citations for the original development paper(s).

Explanation: It is imperative to provide a description of the current scientific knowledge (what is known and not known) regarding the relevant measurement properties of the PROM being studied, providing citations for the original development paper(s) and other relevant citations of this knowledge. Ideally, if there is a systematic review of the measurement properties of the PROM, it should be summarized and cited. Depending on the measurement property under investigation, it may not be necessary to describe all prior research on the PROMs' measurement properties. Instead, it may be appropriate to only discuss the measurement properties relevant to the aim of the study. For example, if you focus on structural validity, information on content validity is relevant, but details on other measurement properties are less crucial. Similarly, if the aim is to evaluate internal consistency, data on structural validity should be thoroughly described. In addition, it is necessary to provide the rationale for all objectives of your study. The rationale for your study should include the gaps in the literature, why the objectives are important, and why it is necessary to evaluate the specific measurement properties in the specific study sample. Evaluation of measurement properties is relevant because a high-quality PROM allows you to put faith in the score obtained from it. For example, a new PROM is developed, but not all measurement properties have been studied yet; or if your aim is to evaluate a specific measurement property in a new population, and you provide an overview of studies conducted in other population in which the same measurement property is evaluated, you can show that there is a gap in knowledge, as it is not known whether the results from the other studies can be generalized to your population; or a study on specific measurement property of a PROM was conducted because results from a systematic review of the measurement properties of the PROM showed that these specific measurement properties were not evaluated at all, or only in low-quality studies; or there is a high-quality study in the population of interest, but the intent is to repeat the study for reproducibility of the results. As noted above, in addition to the original PROM development paper(s) citations, it is important to provide other highly relevant citations related to the quality of the PROM. These citations give credit to the developer of the PROM, help your readers to find the original development of the PROM, and assist them with relevant information found in other studies on the measurement properties of the PROM.

Essential elements

• Current state of scientific knowledge regarding the relevant measurement properties to the focus of the study.

• Gaps in the literature.

• Rationale of all of the study objectives.

• Citation of the original PROM development paper(s).

• Citation of any relevant paper(s) on the measurement properties of the PROM.

Example 1

"In this study, we assess the measurement invariance of the SF-12 with respect to multiple demographic characteristics that are highly relevant in health-related research: gender, age, educational level, and ethnicity. Previous studies have used such an approach to investigate measurement invariance of the SF-12 or the SF-36 [ref] and some of them indeed indicated a violation of the assumption of measurement invariance, with regard to one or more demographic characteristics [ref]. Most of these studies focused on the SF-36 or were performed among specific patient groups [ref]. (...) [ref] conducted the only study that explored DIF of the SF-12 in a general population sample. In this study, performed in the US, a comparison was made between white Americans, African-Americans and Hispanic Americans. (...) African and Hispanic American participants also gave higher ratings on this item [item measuring feeling calm and peaceful], compared with white Americans with a similar health status. (...) The results for these groups are not necessarily generalizable to European populations, however, no such studies have been performed in a European setting. Given the wide-spread use of the SF-12, there is a need for further studies on its measurement invariance in the general population. Not only do European countries host different ethnic minority groups compared to the US, also educational or other subgroups within the majority population might differ, e.g. with respect to cultural beliefs." [33]

Example 2

"The Dutch-Flemish versions of the PROMIS v1.0 adult sleep item banks showed adequate content validity in adolescents (...), meaning the items were considered relevant and comprehensible for adolescents, parents and sleep experts and no key issues were considered missing. Additional psychometric evaluation in a community sample of over 1,000 Dutch adolescents, however, showed that the one-factor models found in adults could not be replicated. Thus, the items used for adolescents did not reflect the same single construct measured in adults and the item banks were not unidimensional. Adaptation of the item banks to improve the unidimensionality needed for IRT analyses resulted in a shortened version of the Sleep-Related Impairment item bank (11 instead of 16 items) with adequate fit (comparative fit index [CFI] 0.98) and a shortened version of the Sleep Disturbance item bank (23 instead of 27 items) with fit indices just below the recommended value (CFI 0.90, recommended value > 0.95). (…) is necessary to optimize structural validity in order to enable CAT. In contrast, the Sleep-Related Impairment-11 item bank has sufficient structural validity and performed well as CAT. It can properly distinguish between clinical and non-clinical groups of adolescents, but test-retest reliability was just below the recommended criterion." [34]

I4 - Objectives

Provide the specific objective(s) of the research, specifying (1) the name (and version, if relevant) of the PROM, (2) the measurement properties being evaluated, and (3) relevant study sample characteristics.

Explanation: Typically, the objective(s) is/are stated at the end of the Introduction section. In an article about the measurement properties of a PROM, at least one objective should be stated about which measurement properties are being evaluated (however, additional objectives may be investigated, for example, referring to translating or evaluating the interpretability of the PROM). In the objective of your study, it is important to specify all **measurement properties** that are being studied. In contrast to reporting the objective(s) of the study in the abstract (see A1) where we allow reporting the domains (e.g., validity, reliability, and responsiveness), we recommend that you describe all specific (aspects of (e.g., translation, interpretability)) measurement properties in this section. This makes it clear to your readers exactly which (aspect of) measurement properties have been evaluated. As mentioned previously, we advise to use the COSMIN terminology of measurement properties. In line with the objective described in the abstract, also, the **name (and version, if relevant) of the PROM** being studied and a general description of the **characteristics of the study sample** (in terms of disease characteristics and demographic characteristics, if relevant) should also be specified in the objective (see A1).

Essential elements

• Specific objective(s) of the research including name – and version – of the PROM, all measurement properties being studied, and study sample characteristics.
• Additional objective (e.g., on translation or interpretability) could be specified.

Example 1

"This study's aim was to evaluate the test-retest reliability and measurement error of the Danish WHO-Five Well-being Index (WHO-5) and the Danish Problem Areas in Diabetes (PAID) questionnaire used in a PRO-based telehealth solution (DiabetesFlex) among patients with type 1 diabetes." [35]

Example 2

"(…) the aim of this study is to examine measurement invariance of the SF-12 regarding age, gender, educational level, and ethnicity, using a multi-ethnic sample (HELIUS) of over 23,000 participants collected in the Netherlands." [33]

Report section: General Methods

GM1 - Study design

Specify (details of the) study design used to evaluate the measurement properties.

Explanation: We advise clearly presenting the key elements of the study design. Clear reporting of your study design is particularly important when your study is being used in systematic reviews or meta-analysis of measurement instruments. This clarity is crucial for accurately assessing the risk of bias in your study. At item A2, we provided specific recommendations and examples tailored to studies focusing on different measurement properties. Additionally, we suggest reporting these elements both in the abstract and the methods section. If your study utilizes secondary data, it is important to cite the original study or data source in the methods section. For studies based on clinical trials, we advise citing the trial protocol.

Essential elements

• The study design in terms of number of measures (e.g., cross-sectional, test-retest design, longitudinal), data type (e.g., qualitative), and/or exposure/intervention (e.g., observational, clinical trial).

Example 1

"This was a cross-sectional study to investigate the structural validity of the Roland-Morris Disability Questionnaire (RMDQ)." [14]

Example 2

"This was a qualitative study to develop a conceptual model and develop/evaluate a new PRO instrument suitable for use in chronic ocular surface pain (COSP) treatment trials." [36]

Example 3

"De-identified patient responses to the Knee Injury and Osteoarthritis Outcome Score (KOOS) were obtained from the Surgical Outcome System (SOS), a patient-reported outcome database maintained by Arthrex (Naples, FL, USA)." [37]

GM2 - Participants

Specify how the study participants were selected. Specify the inclusion and exclusion criteria.

Explanation: PROMs should be tested in the population and context in which they are intended to be used [11], since the results of measurement properties may vary in different populations. It is important to specify which criteria were used to select participants (e.g., inclusion and exclusion criteria) including who applied these criteria (i.e., their relevant training or experience) and what sampling methods were used to select them. The inclusion and exclusion criteria should make clear from which population the sample was selected. Therefore, typically relevant disease characteristics (e.g., those with the condition of interest or without comorbidities, etc.) and demographic characteristics (e.g., age, gender) are specified. In addition to the sampling strategy (e.g., simple random sampling, stratified sampling, clustered sampling, convenience sampling, etc.), information on study location, or recruitment sites should be described. All of this is necessary to report in order for your readers to assess whether the population of interest was adequately and appropriately sampled, and whether your results are applicable to a population of interest to your reader.

Essential elements

- Inclusion and exclusion criteria.
- Study participants sampling methods.

Example 1

"Sixty adult idiopathic Parkinson disease (PD) patients, according to the criteria of the International Parkinson and Movement Disorder Society and the UK Parkinson's Disease Society Brain Bank that did not suffer from dementia or conditions that would interfere with the study's assessments were consecutively enrolled from June 2019 to December 2019." [38]

Example 2

“Participants were invited to participate in the study through Facebook®, Instagram®, WhatsApp®, emails from universities and schools through a Google Forms link. Brazilian women aged between 14 and 42 years, with a report of dysmenorrhea in the last three months and able to speak, read and write in Brazilian Portuguese were included. The lower age limit is the mean menarche of Brazilian women [ref] and the upper limit decreases the probability of perimenopausal women [ref]. Pregnant women, women with up to 6 months of puerperium and transgender were excluded. We excluded transgender people due to hormonal issues that may interfere with the assessment of dysmenorrhea, which was beyond the scope of this study.” [17]

GM3 - PROM details

Provide details about the original version of the PROM as well as of the PROM version being studied, specify the conceptual framework (reflective/formative model), details on the structure (the number of items and subscales), the language, response scale, recall period, direction of scoring, and scoring algorithm of the PROM. Specify how the PROM was administered (e.g., in what setting, mode of administration (e.g., paper, electronic) what instructions were given), including the country in which it is administered.

Explanation: It is important to provide details both about the original PROM, as well as about the PROM version being studied (or state that the versions are the same, if applicable). With this information, the reader of your study is able to determine the generalizability of the results of your study. Relevant details about the PROM versions are the **language**. If the PROM has been translated, it is important to report details about the translation process (or a citation). Providing details about the language in which the PROM is being evaluated, in combination with **the country** in which the PROM was developed (see I2), is also important to help your readers understand the cultural and linguistic context of the study (e.g., the Dutch version of the PROM was applied in Belgium, or the Spanish version of the PROM was applied in North Chile). Next, details about its **conceptual framework** should be provided. Specifically, it should be reported if the items of a (sub)scale are based on a reflective model or a formative model. In a reflective model, the construct manifests itself in the items and items in a scale are expected to correlate to some extent. In a formative model, the construct is defined by the items, and the items together form the construct, and do not need to be correlated [39]. To decide which model is applicable (reflective or formative), there are some decision rules (e.g., [40]). We recommend that you describe the structure of the PROM in detail, referring to **which and how many items are included (per subscale)**. It is also important to report the **response scale**, which may take the form of a Likert scale, number scale, visual scale, or face scale. The number of categories, or the range of response options and what each level of each response refers to (e.g., zero for none and 10 for always, or 1 for strongly disagree and 5 for strongly agree) should also be described. The **recall period** refers to the specific timeframe across which patients are instructed to reflect on their health-related experience or status when responding to items in a PROM (e.g., now, past 24 hours, past week). The length of the recall period can significantly impact the potential for recall bias, emphasizing the importance of transparency in reporting to ensure accurate interpretation of data and results. Finally, the **scoring rule or algorithm** should be clearly described. The scoring algorithm indicates how the patient's responses are calculated into one or more scores. For example, item scores can be averaged into an

unweighted scale score; or some item scores get more importance by using a weighted sum score; or theta scores based on IRT or Rasch parameters are used. It is also important to report the meaning of the **direction of the score** (i.e., does a higher score represent a better or worse outcome?). Last, the **mode of administration** of the PROM versions is important to report, as this may influence how the PROM is scored. The mode of administration refers to whether a paper version, an electronic version, or an interviewer version is developed or used. Also, the device used to deliver the PROM electronically (e.g., a small-screen smartphone or computer) is relevant to report. If an interviewer was involved, it is important to report whether and how the interviewer was trained or how much experience the interviewer had.

Essential elements

For both the original PROM version, and the PROM version being used:

- Language.
- The conceptual framework of the PROM (e.g., reflective or formative model).
- Structure: number of items per (sub)scale and ordering of items per subscale.
- The response scale and recall period.
- The scoring algorithm and direction of scoring.
- The mode of administration.

General example

"This PROM was developed in [country] in [language] to measure [construct] in [recall period] in [population] and to be used [mode of administration]. The [PROM name] is based on the [conceptual framework] because of [explanation of conceptual framework]. The [PROM name] has [number] items divided into [number] subscales: [subscale A] [items 1 to 4], [subscale B] [items 5 to 7], and [subscale C] [items 8 to 10]. The response scale for each item ranges from x [meaning] to y [meaning], where higher scores indicate [direction of scoring] for [subscale A], (...) [subscale B], (...) [subscale C]. For each subscale, the score is calculated by [scoring algorithm]."

GM4 - Additional data collection

Describe why and how other data was collected (e.g., construct and measurement properties of the comparator instruments, characteristics of groups being compared, and rationale for choosing groups), including mode of administration (e.g., paper, electronic).

Explanation: Additional data are commonly collected in studies on measurement properties and used for a variety of purposes. We recommend reporting why and how you collected additional data. Additional data may be collected to **describe the study sample** (e.g., data collected on disease and demographic characteristics), to **stratify the group** for particular investigations (e.g., to evaluate cross-cultural validity\measurement invariance, known group in hypotheses testing for construct validity, or responsiveness), or to use as a **check whether patients were stable** on the construct between repeated measurements (e.g., in a study on reliability and measurement error), or have changed (e.g., to determine MIC values). The way the data were collected may be based on single-item questions, such as a variable to describe the sample or to divide it into groups (e.g., age, gender), or a single-item scale (e.g., a global perceived change question); or it may be based on a specific measurement instruments (e.g., used as comparator instruments). If another measurement instrument is used, it is also important to describe the measurement properties of this comparator instrument. If a comparator instrument is used as a gold standard, it is important to provide arguments that it can be considered as such for the construct of interest (that is, the construct that is measured by the PROM being studied).

Essential elements

- Description of why and how data on disease and demographic characteristics were collected.
- Description of why and how the sample was separated into groups, if applicable.
- Description of the construct(s) the comparator instrument(s) intends to measure and its measurement properties.
- Description of why the comparator instrument(s) were chosen.
- Description of why the comparator instrument is considered a 'gold standard'.
- Description of why and how patient stability, if applicable, was verified.

Example 1

"Body appreciation scale (BAS-2) validated for Brazilian Portuguese was used to assess body appreciation. This one-dimensional PROM showed excellent test-retest reliability ($ICC_{agreement}$=0.81) (...) in Brazilian adults." [41]

Example 2

"Known-groups validity was evaluated by comparing the scale scores of the Rheumatoid Arthritis Symptom and Impact Questionnaire (RASIQ) between groups of participants with different levels of disease activity according to the following categories: Disease activity score-28 (DAS28) [ref] was used to split participants into four groups: disease remission (DAS28 ≤ 2.6), low disease activity (2.6 < DAS28 ≤ 3.2), moderate disease activity (3.2 < DAS28 ≤ 5.1) and high disease activity (DAS28 > 5.1); (…) Patient's assessment of arthritis pain (measured using a validated visual analog scale [VAS] scale [ref]) was used to split participants into three groups: no pain to mild pain (VAS ≤ 40), moderate pain (40 < VAS ≤ 70) and severe pain (VAS > 70)." [42]

Example 3

"In order to test the invariance of the factor model in the two samples, (...) Chile and Ecuador were selected for comparison. Although both are Spanish-speaking countries, they are nations with significant cultural differences. Based on the cultural dimensions of the Hofstede framework, Hofstede and Bond found differences between Chilean and Ecuadorian citizens for power distance, individualism vs. collectivism, masculinity vs. femininity and in the uncertainty avoidance dimension [ref]. (…) While in Chile the prevalence of malnutrition is very low, and while it is one of the South American countries with the highest proportion of obese adults (about 30%), Ecuador has a moderately high prevalence of malnutrition [ref]." [43]

GM5 - Time points procedures

Provide all time points of all measurements.

Explanation: Time points refer to specific occasions or intervals during the study period when data is collected or measurements are taken. It is important to report all time points for all measurements of the PROM under study, and all other measurements described in item GM5. Furthermore, it is important to provide clear justification for these time points for all investigations. In cross-sectional studies, it is necessary to report whether all measurements were taken within the same time frame (per patient). In a repeated measurements design (for example, to evaluate reliability and measurement error) or in a longitudinal design (for example, to evaluate responsiveness), some or all of the measurement instruments are repeated over time. We recommend that you report which measurements were conducted and at what specific follow-up points. When the PROM under study and other measurements were not taken at the same points in the study timeline, this should clearly be described, and the reasons for these differences also explained.

Essential elements

• All time points of all measurements for all instruments being investigated.

• Number and time interval of repeated measurements taken in reliability and measurement error studies, where applicable.

• Rationale for why measurements were not taken at the same time point, if applicable.

Example 1

"For data collection, an initial assessment was designed to provide the sample's demographical and clinical characteristics, including the following instruments: Widespread Pain Index (WPI), Symptom Severity Scale (SSS), FiRST® Fibromyalgia Rapid Screening Tool (FiRST®), Numerical Rating Scale (NRS), Fibromyalgia Impact Questionnaire-Revised (FIQ-R®) and Brief Pain Inventory (BPI®). Seven to ten days after answering the questionnaires, participants were asked to answer again the translated version of the FiRST® for test-retest reliability and measurement errors. This period of time followed the COSMIN recommendations in which seven days after the test is the minimum time for the patients not to have changed in the measured construct and not to remember their answers to the instrument." [44]

Example 2

“The study included 3 time points. Time 1 (T1) occurred 1 to 2 days before an early course of chemotherapy. Time 2 (T2) occurred 7 to 16 days following chemotherapy initiation at the time when the patient’s nadir was projected. T3 occurred within 1 to 2 days preceding the next course of scheduled chemotherapy or approximately two weeks following T2.” [45]

GM6 - Justification for sample size

Provide a rationale for the sample size for all measurement properties analyses (including subgroups).

Explanation: Reporting the rationale for the sample size ensures that the study is adequately powered, thereby reducing the risk of spurious findings. It is recommended to provide a rationale for the sample size you intend to include – including sample sizes for subgroups – for each measurement property analysis, and add relevant citations for the rationale for the chosen sample size. Often, for studies on measurement properties there are no formal methods for sample size calculations. Therefore, it is reasonable to refer to rules of thumb for all analyses. However, if a formula to calculate the sample size(s) is used it should be reported. In content validity studies, a qualitative approach is often used with the goal of achieving comprehensive item generation. Some authors recommend a minimum number of interviews in a qualitative study (e.g., 12 to 20 [46], 10 to 15 [47]), while others emphasize the important of achieving comprehensive item generation [48]. For a minimum number of participants in a quantitative study different recommendations have been published (e.g., 30 [49], 50 [50]). In structural validity, factor analyses and item response theory (IRT)/Rasch analyses require large sample sizes and can be based on several guidelines about the required number of participants [51–53]. Note that cross-cultural validity\measurement invariance testing requires subgroups to be compared, and the sample size requirements refer to both subgroups. In studies on reliability and measurement error with continuous scores, there is an online application developed to provide recommendations for adequate sample sizes based on simulation studies [54]. For nominal/ordinal scores, we recommend you refer to the recommendations based on kappa statistics [55] and contingency tables [56, 57]. In studies on criterion validity, hypothesis testing for construct validity, and responsiveness, the sample size requirements are based on the statistics used for each hypothesis. When reporting the rationale for the sample size, it increases the transparency of your study.

Essential elements

- Rationale for the sample size for all measurement properties analyses (including subgroups).

Example 1

Content validity: "We recruited at least three experts from each profession because COSMIN recommends that more than seven experts participate [ref]." [58]

Example 2

Structural validity and cross-cultural validity: "For factor analysis, the minimum sample size of 7 times the number of items in the instrument to be validated and greater than 100 individuals is advisable [ref]. Therefore, for this study, a minimum sample of 140 women was appropriate, since the Pelvic Floor Distress Inventory (PFDI-20) has 20 items." [29]

Example 3

Hypotheses testing for construct validity: "(...) for the comparison of mean Oral Health Literacy Instrument - French (OHLI-F) scores with education levels or frequency of dental visits, the two-means formula was applied, with alpha = 0.05, power = 80%, standard deviation = 18 (15), expected difference = 10, and expected dropout rate = 20%, resulting in n = 51 per group. Since level of education and frequency of dental visit were classified into 3 groups, the total sample size required was 191 participants." [59]

Example 4

Test-retest reliability and measurement error: "In this study, the sample size for intraclass correlation coefficient (ICC) was calculated with a correlation coefficient of 0.7 as an effect size [ref], power 0.8, and target width 0.3 of the 95% confidence interval of ICCs (ICC2,1). The required same size was a minimum of 40 for 3 raters with systematic deviation based on the recommendation by [sample size decision assistant developed by Mokkink et al. [ref]]." [60]

Example 5

Responsiveness: "According to the study of [ref], and α:0.05, β :0.2, μ1 i.e. menstrual problems domain mean score before oral contraceptive: 3.9; S1: 1.4, μ2 i.e. menstrual problems domain mean score after oral contraceptive: 5.26; S2: 1.6, estimated 10% drop out the sample size were estimated at least 150 women per group (single and married) as followed [46]:

$$n = \frac{\left(z_{1-\frac{\alpha}{2}} + z_{1-\beta}\right)^2 \left(s_1^2 + S_2^2\right)}{\left(\mu_1 - \mu_2\right)^2}$$ [61]

GM7 - Statistical analyses

Describe the statistical analyses corresponding to all objectives (see measurement properties specific boxes). Describe the criteria for good measurement properties. Name the statistical package used and the version.

Explanation: Providing a detailed description of your statistical analyses will allow your readers to assess the validity of your methods and their implementation aiding in the interpretation of results and allowing for replication of the analyses [62]. We recommend that you report all statistical analyses corresponding to the objectives and measurement properties. Specifically, it is imperative to report the specific statistical tests and procedures used (e.g., Pearson or Spearman correlation, rather than only correlation). We also recommend that you describe the criteria for good measurement properties that you use to assess the findings and formulate conclusions (i.e., a predetermined value used as a reference for interpreting the analysis) with a citation. Next, we recommend that you report the statistical package that you used, include the version numbers (e.g., STATA/MP 14.2) and citations as appropriate. We describe additional details for reporting the analyzes for specific measurement properties in the items for specific measurement properties below.

Essential elements

- Statistical analyses corresponding to all objectives and measurement properties.
- Criteria for good measurement properties for each measurement property evaluated.
- The statistical package used.

Example 1

"For the confirmatory factor analysis, (...) we used the maximum likelihood robust, Root Mean Square Error of Approximation (RMSEA), Standardized Root Mean Squared Residual (SRMR), Comparative Fit Index (CFI) and Tucker-Lewis Index (TLI). The model was considered adequate when RMSEA and SRMR < 0.08, CFI and TLI > 0.90 [ref]. (...) Internal consistency was evaluated by Cronbach's Alpha (α), in which α = 0.70 to 0.95 was considered adequate [ref]. For the hypothesis test for construct validity, Pearson correlations were calculated between the WHODAS 2.0 and numerical rating scale. We followed the correlation magnitudes by Cohen [ref]: weak (r < 0.30), moderate (r = 0.30 to 0.50) and strong (r > 0.50). We hypothesized a positive and moderate correlation between human functioning/disability and dysmenorrhea pain severity [ref]. (...) All analyses were conducted in Psych package in RStudio." [17]

GM8 - Missing data

Describe approaches for dealing with missing data.

Explanation: Missing data is the absence or unavailability of observations or values in a data set that were intended to be collected during a study. In studies on measurement properties, missing data could occur because one of more questions in a multi-item PROM have not been filled-out, or the entire PROM was not completed at some of the occasions. Also, other data that is planned to be collected (see GM5) may be missing. Various reasons can lead to missing data, such as data collection errors, loss of information, participant dropout or other issues encountered during the research process. We recommend that you report and justify any approaches (e.g., imputation or data removal) that will be used to deal with missing data on the measurements of the PROM under study and other collected data. Reporting plans to handle missing data ensures transparency and reproducibility in research or data-driven decision making.

Essential elements

• Approaches for dealing with missing data for each measurement property.

Example 1

"Patients with a missing value for the studied Utrecht Symptom Diary (USD) item were excluded for analysis for that item." [63]

GM9 - Unplanned analysis

Specify analyses that were unplanned and their rationale.

Explanation: In a well-designed study, typically and ideally research questions, and specific statistical analysis are planned before data collection begins. Unplanned analyses can arise for various reasons, such as the discovery of unexpected patterns or new research questions that emerge during the course of the study. We recommend that you specify and provide the rationale for any unplanned analyses on measurement properties (i.e., describe the deviation from the protocol). The purpose of this explanation is to help the readers understand which and why unplanned analyses were done.

Essential elements

• Analyses that were unplanned and their rationale.

Example 1

"An initial confirmatory factor analysis (CFA) was conducted in Mplus on the polychoric correlation matrix with weighted least squares with mean and variance adjustment (WLSMV) estimation for each item bank separately, to examine the expected unidimensionality of the item banks. (...) When model fit for an item bank was not found to meet the recommended criteria of Comparative Fit Index (CFI) > 0.95, Tucker-Lewis Index (TLI) > 0.95, and Root Means Square error or Approximation (RMSEA) < 0.06, [ref] an exploratory approach similar to what was used in the item banks' primary development procedure was adopted [ref]. This provided the opportunity to further assess the dimensionality and to understand the misfit of the model in adolescents." [64]

Report section: General Results

GR1 - Participant characteristics

Provide study participants' characteristics, specified per subgroup if applicable.

Explanation: When describing the study population in terms of relevant disease and demographic characteristics, it is ideal to provide enough information for the readers to determine generalizability and applicability of the results to specific subgroups or populations. Although there is no universal answer to the question of what descriptive characteristics of the study participants should be reported, we recommend that you report **relevant disease characteristics** (e.g., severity, duration of illness) and other **relevant demographic characteristics** (e.g., sex, age). If subgroups of participants are used in some of the analyses, we recommend that you describe the characteristics both for the entire group, as well as for the subgroup(s). For example, when part of the patients was asked to participate in the reliability study, or when the total sample was split for performing exploratory factor analysis on one subgroup and confirmatory factor analysis on the other subgroup. In addition, the PROM (sub)scale scores (baseline, and change scores, if applicable) should also be provided for each (sub)group, in terms of mean (standard deviation), or median (range or interquartile scores). Providing detailed descriptive data helps demonstrate whether the study sample is representative of the intended target population for the PROM, and it also facilitates interpretation of the PROM scores. It is acceptable for descriptive data and PROM scores to be summarized in a table.

Essential elements

- Relevant disease and demographic characteristics.
- Baseline and change PROM scores (i.e., mean and standard deviation, or median and interquartile scores or range.

Example 1

"(...) Table 1 presents sample characteristics. Notably, the majority had a stroke (92%), and time since injury varied substantially across rehabilitation settings (Figure 1). In community-based rehabilitation, median time since injury was 9-months, ranging from 1-161 months. For stroke inpatients, median time since injury was 13 days, ranging from 5-47 days. (...) Statistics on Dutch Multifactor Fatigue Scale (DMFS) subscales are reported in Table 2."

Table 1. Characteristics of participants.

Characteristics	Community-Based (n = 100)	Sub-Acute (n = 49)	Full Sample (N = 149)
Sex, n (%)			
Female	37 (37.0)	19 (38.8)	56 (37.6)
Male	63 (63.0)	30 (61.2)	93 (62.4)
Age, M (SD)	54.3 (10.7)	66.7 (12.2)	58.4 (12.6)
Education, M (SD)[1]	15.0 (3.3)[2]	14.4 (4.0)	14.8 (3.6)[3]
Days since injury, M (SD)	540.8 (801.3)[4]	17.2 (10.5)[5]	367.4 (669.4)[6]
Type of injury, n (%)			
Stroke	89 (89.0)	49 (100.0)	138 (92.6)
Traumatic brain injury	5 (5.0)	0 (0.0)	5 (3.4)
Other[7]	6 (6.0)	0 (0.0)	6 (4.0)
Type of Stroke, n (%)			
Ischemic	59 (66.3)	33 (67.3)	92 (66.7)
Hemorrhagic	26 (29.2)	12 (24.5)	38 (27.5)
Both	3 (3.4)	0 (0.0)	3 (2.2)
Missing data	1 (1.2)	4 (8.2)	5 (3.6)
Previous brain injury, n (%)	16 (16.0)	9 (18.4)	25 (16.8)
Missing data	4 (4.0)	7 (14.3)	11 (7.4)

Note. [1] Reported in years. [2] n = 99. [3] n = 148. [4] n = 97. [5] n = 48. [6] n = 145. [7] Includes central nervous system infection, aneurisma surgery, anoxia, and hydrocephalus.

Table 2. Statistics on subscales of Dutch Multifactor Fatigue Scale.

Scale	Descriptive Statistics		
	N	M (SD)	Range
IF	147	38.7 (10.4)	11-55
SC	149	28.9 (7.6)	9-45
MF	148	25.2 (6.3)	7-35
PF	148	17.4 (5.4)	6-30
CF	148	15.3 (3.9)	7-25

[65]

GR2 - Sample size

Provide the total number of participants included in the study and the sample size for each analysis.

Explanation: Adequate sample sizes are necessary to estimate outcomes with enough precision to draw valid conclusions. We recommend that you provide the total number of participants included in the study, as well as for each subgroup and in each analysis (per measurement properties). For example, in a study on cross-cultural validity/measurement invariance, we recommend that you report the sample size for each group included in the analysis; or the sample size in a reliability study, when only part of the all participants were asked to participate in this sub study. Reporting this information increases the transparency of your study and allow readers to compare the anticipated sample size (as described in the method section) with the actual sample size of your study.

Essential elements

- The total number of participants included in the study.
- The sample size for each analysis and (sub)groups.

Example 1

"A total of 172 atrial fibrillation (AF) patients visited the Cardiology Clinic from January to March 2022, of whom 133 patients met the inclusion criteria. Twelve patients were excluded due to cognitive impairment, recent hospitalization, or physical disabilities. Of 121 patients who consented to participate, 6 were excluded because of acute conditions and 13 others could not be contacted for the retest procedure, leaving 102 patients who were included in the analyses (…)." [66]

Example 2

"Structural validity was carried out on the original sample of 262 patients (...). Internal consistency was calculated using the baseline scores of the 262 patients (...). Test-retest reliability was assessed on a sample of 64 patients from the general sample who completed the Michigan Hand Outcomes Questionnaire (MHQ) for the second time. From the initial sample, 222 patients completed the MHQ, Disability of Arm, Shoulder and Hand (DASH) and DASH-work as well as grip strength and Visual Analogue Scale (VAS) pain at 5 weeks after baseline." [67]

GR3 - Missing data

Provide amount of (proportion or count) and reasons for missing data for each analysis for the PROM, and for any analyses of other outcome measurement instruments.

Explanation: Missing data is the absence or unavailability of certain observations or values in a dataset that were intended to be collected during a study. You should provide the amount and reasons (e.g., data collection errors, loss of information, or other issues encountered during the research process) for missing data for each analysis for the PROM or other outcome measurement instruments per time point. Missing data may occur, for example, if some items were not answered by the patients, or if patients in a longitudinal study dropped out or failed to complete any measurements at a specific time point. Additionally, when reporting reasons for missing data, it is important to specify whether the data were missing at random, missing completely at random, or missing not at random. Understanding the mechanism of missing data is important because it can affect the validity of statistical analyses and the interpretation of results. Reporting missing data is a fundamental aspect of transparency in a study of measurement properties of PROMs, and allows readers to understand the study and PROM's limitations, and make informed decisions about how to interpret the results.

Essential elements

- Amount and reasons for missing data for each analysis for the PROM under study, at any timepoint.
- Amount and reasons for missing data on other outcome measurement instruments.

Example 1

"(...) one patient (0.7%) and 38 patients (26%) did not answer the WHO-Five Well-being Index (WHO-5) or Problem Areas in Diabetes (PAID) questionnaires, respectively. Furthermore, missing data in single items ranged from 0.7% in the general health status item to 25.3% in the decreased feet feeling and pain in feet items. The highest proportion of missing data was found in the erectile dysfunction item (60.3%) (Table 5). Informal conversations with patients related to non-response indicated two primary reasons for not completing the second questionnaire: 1). Some patients did not understand the purpose of the second questionnaire as they had already answered the same questionnaire at test 1, and 2). Some patients expressed a lack of energy to fill in an additional questionnaire." [35]

GR4 - Results

Describe the results corresponding to all objectives (see measurement properties specific boxes).

Explanation: Results corresponding to each objective provide a comprehensive understanding of your study's findings. This facilitates the interpretation of your results in the context of specific research questions or aims, helping your readers discern the relevance and implications of the study. We recommend that you describe all results according to all objectives. Results can be reported in tables, and details in Appendices. For example, for structural validity, model fit indices of the tested models could be reported in a table in the manuscript, while the factor loadings of items from different models can be reported in appendices. Reporting all results helps the readers understand the purpose and results of your study, specifically when your results are used in a future systematic review of measurement instruments to facilitate PROM selection. Details per measurement property are provided below in the boxes per measurement property.

Essential elements

• Description of the results corresponding to all objectives.

• Results can be provided in text and/or in tables, and detailed results can be provided in an appendix.

Example 1

Structural validity and internal consistency: "Structural validity was evaluated through exploratory factor analysis (EFA) and confirmatory factor analysis (CFA) on 400 women of reproductive age (conducting EFA and CFA) on same sample). Kaiser-Meyer-Olkin (KMO) value equaled 0.726. The KMO > 0.7 confirmed the significance of Bartlett's test and model adequacy (P ≤ 0.001). Moreover, the three-factor structure in the EFA process obtained a total variance of 58.15% (Table 3). The first factor was Pelvic Organ Prolapse Distress Inventory-6 (POPDI-6), which comprised 6 items that explained 18.25% of the total variance. It should be mentioned that item 6 of factor POPDI-6 has a factor loading of less than 0.3, but according to the opinion of the research team, this item had a high importance and weight and we could not remove it, but it should be noted that in the CFA, this item was significant (p < 0.001), so it was not removed. Colorectal-Anal Distress Inventory-8 (CRADI-8) was the second factor that covered 8 items that explained 17.10% of the total variance. Finally, the third factor included Urinary Distress Inventory 6 (UDI-6) comprised 6 items that explained 22.80% of the whole variance.
In the CFA phase, three factors were obtained and then used in EFA by CFA. According to results (RMSEA = 0.07, SRMR = 0.07, TLI = 0.97, CFI = 0.99, $\chi 2/df$ (normed chi-square) = 3.19), this model had an optimal fit. Hence, factor structure could be confirmed (Table 4)." [68]

Example 2

Internal consistency, test-retest reliability and measurement error: "Cronbach's alpha ranged between 0.84 and 0.90 (Table 3). The lower limits of all ICC confidence intervals were greater than 0.7. For each of the short forms, 18% to 20% of the test-retest sample had the best possible scores on both test occasions. Both the standard error of measurement (SEM_{agr}) and smallest detectable changes (SDC_{90}) were higher for pain intensity (PAIN) and pain interference (PI) compared to physical function (PF). The effect size based on SEM_{agr} was around 5 to 6 for all three short forms, and smaller than that for the Oxford Knee Score (OKS) (9.5)."

Table 3. Reliability, agreement and smallest detectable change

	Crobanch's α[a]	ICC[a]	SEM_{agr}	SDC90	Effect size based on SEM_{agr}
PROMIS PAIN	0.84 (0.79 to 0.88)	0.93 (0.88 to 0.96)	3.55	8.28	5.51
PROMIS PI	0.90 (0.78 to 0.92)	0.90 (0.83 to 0.94)	3.34	7.78	4.84
PROMIS PF	0.88 (0.74 to 0.91)	0.97 (0.94 to 0.98)	1.72	4.02	6.33
OKS	-	-	1.78	4.15[b]	9.52

ICC intraclass correlation coefficient; SEM_{agr} agreement for T-scores assessed using standard error of measurement from test-retest; SDC90 smallest detectable change for individuals that can be considered above the measurement error with a 90% confidence level; Effect size based on SEM, $_{agr}$ calculated as absolute value of the mean change score divided by SEM_{agr} PROMIS Patient Reported Outcomes Measurement Information System; PAIN pain intensity; PI pain interference; PF physical function; OKS Oxford Knee Score

[a]95% confidence interval in parentheses

[b]According to: Beard DJ, Harris K, Dawson J, Doll H, Murray DW, Carr AJ, et al. Meaningful changes for the Oxford hip and knee scores after joint replacement surgery. J Clin Epidemiol. 2015;68(1):73-9 [69]

Report section: Discussion/conclusions

DC1 - Measurement property evidence

Provide the main findings and if each measurement property is sufficient or insufficient and why.

Explanation: The discussion section of your study plays a crucial role in interpreting and contextualizing the main findings of your study with the literature, helping readers to understand the implications of your results and compare with other studies. We recommend that you state the main findings for each measurement property, if it is sufficient or insufficient, and compare your findings with other studies of the same PROM. Offering a summary of the findings helps your readers quickly grasp the main outcomes of your studies. Knowing whether a PROM has sufficient reliability, validity, and/or responsiveness informs its selection and applicability in different settings. PROMs with insufficient measurement properties should not be recommended for use in clinical practice or research. When you compare your findings to other studies of the same PROM, you help your readers understand the generalizability of the PROM in the same population.

Essential elements

- The main findings for each measurement property.
- If each measurement property is sufficient or insufficient.

Example 1

"Cronbach's alpha values were all around 0.9 or higher, except for sleep disturbance (alpha=0.75), thereby showing sufficient internal consistency." [22]

Example 2

"Of our a priori defined hypotheses 78% could be confirmed, thereby meeting the 75% required for sufficient construct validity according to the COSMIN criteria for good measurement properties [ref]. For most domains, this criterion was also (almost) met. Although we based our hypothesis on analyses with other Dutch datasets [ref] and previous experiences, one should note that a one point difference, as used in some hypotheses, might not (always) be meaningful. It is not yet clear what a minimal important difference in scores between groups is for PROMIS measures, but most studies suggest a within-person change of at least three points to be meaningful [ref]." [22]

DC2 - Practical relevance

Discuss the practical relevance of the findings in terms of recommendations for (not) using the PROM.

Explanation: The Discussion section can provide an opportunity to contextualize your results within real-world applications of the PROM. Based on your findings and results of other studies on the same measurement property of the same PROM, we recommend that you discuss the practical relevance for instrument selection (i.e., explain whether the PROM can be recommended for use, whether more research is needed, or whether the PROM in its current form should not be used in research and/or clinical practice, i.e. because of insufficient or inconsistent results). Describing the practical relevance may guide the selection process for future readers by providing information on how and when the PROM can be applied/used (e.g., as an evaluative instrument to monitor patients over time in healthcare, or in research). By including a discussion of practical relevance, you can highlight how the use of the instruments can contribute to better patient outcomes (e.g., improved symptom monitoring). This information is crucial for informing decisions about PROM use, and ultimately for standardization of PROM use, and improving outcomes in various fields.

Essential elements

- Practical relevance for instrument recommendation/selection and use.

Example 1

"Our research contributes to the field of female sexuality and sexual medicine in Romania, for research, as well as teaching purposes and in clinical practice. The clinical utility of the Romanian version of the female sexual function index (FSFI-RO) lies in its potential to enhance the assessment and understanding of female sexual dysfunction (FSD) in clinical practice and research settings. (...) the Romanian female sexual function index (FSFI) can be used as an outcome measure in clinical trials or intervention studies. Its sensitivity to change allows researchers and clinicians to evaluate the effectiveness of various treatments or interventions for female sexual dysfunction. By administering the Romanian FSFI before and after treatment, researchers can assess the impact of interventions on different aspects of sexual function and quantify the magnitude of change. This not only contributes to the evidence base for treatment effectiveness but also helps guide clinical decision-making. Moreover, the availability of the Romanian FSFI can facilitate cross-cultural and international research collaborations. Researchers and clinicians in Romania can use the same standardized measurement tool as their counterparts in other countries, allowing for direct comparisons and meta-analyses across different cultural and linguistic contexts. This enables a more comprehensive understanding of the global prevalence, determinants, and consequences of female sexual dysfunction." [70]

Example 2

"Developing a reliable and valid measure of nausea/vomiting in this population will facilitate the measurement of these reactions in clinical trials. This study supported many of the psychometric characteristics of a short daily diary designed to measure nausea and vomiting. The Nausea/Vomiting Symptom Assessment (NVSA©) may be used to support study endpoints in clinical trials comparing the nausea and vomiting profile of novel Secondary hyperparathyroidism (SHPT) therapies." [71]

DC3 - Strengths and limitations

Discuss strengths and limitations of each study. For example, discuss if there were any potential biases in the study that could have impacted the results.

Explanation: A study is strengthened when it employs a well-designed and rigorous methodology. This includes appropriate sampling (techniques), data collection methods, statistical analyses, and methods for interpretation of results. On the other hand, the limitations section identifies significant potential biases or evidence of imprecision in your study that could have impacted your results and conclusions. For example, (1) if you conducted a study on the content validity of a PROM and the sample size was predominantly adults with high levels of education, you should discuss that the results should not be generalized to adults with lower levels of education; (2) a reliability study requires repeated measurements in stable patient, but some of the patients included may have changed; or (3) some of the patients did not completed the comparison instruments used to assess validity or responsiveness on the same day. We recommend that you discuss strengths and limitations of each study on measurement properties described in your article in terms of potential biases that could have impacted your results. Transparently discussing the strengths and limitations enhances the credibility of each of the studies that may affect the broader applicability of your results.

Essential elements

- Strengths of the study.
- Limitations of the study.

Example 1

"A strength of this study is the very large sample size, enabling us to perform the analyses for subgroups with and without chronic diseases and to investigate differential item functioning (DIF) for important sociodemographic and clinical characteristics. A limitation of our study is the representativeness of the Lifelines cohort, in which males, younger persons, and persons with an immigration background are underrepresented compared with the general Dutch population. Furthermore, in our sample, 62% reported not having a chronic condition, whereas according to registries in 2019, 43% of the Dutch population had no chronic condition [ref]. Thus, our sample was not representative for the Dutch population, and therefore, reported T-scores should not be interpreted as reference values for the Dutch population. Papers regarding reference values for the Dutch population on the domains included in the PROMIS-29 have recently been or will soon be published ref]. Finally, formulating challenging hypotheses in which both the direction and the magnitude of the difference or relationship are included, is difficult. We based our hypotheses on findings of previous research, to show that PROMIS-29 functions in our population as expected." [22]

Example 2

"Although the numbers of patient participants in the focus groups were limited, the voice of patients was widely represented throughout the development process and further enhanced by the active involvement of our patients as research partners (PRPs) at all key stages (Supplementary Table S1, available at Rheumatology Advances in Practice online). Although all patient participants in stage 3 cognitive interviews were male, the contribution of the PRPs (4 female and 3 male) to the analysis ensured that a gendered view of the data was facilitated. A rigorous approach to all phases of the qualitative research is described, which involved patient, clinical and research experts participating in an iterative approach to item development and refinement. This process increases confidence that our multifaceted approach has minimized the risk that any patient-important outcomes have been omitted from the measurement framework. The involvement of health-care professionals in additional focus groups was essential to enhancing the clinical relevance of the model and is a further strength of the Warwick Axial Spondyloarthritis faTigue and Energy questionnaire (WASTEd)." [23]

Example 3

"Limitations included the post hoc secondary analysis of patients, all of whom had recurrent or metastatic disease and were subject to recall bias. This limitation was mitigated by cognitive interviews with patients actively undergoing radiation therapy (RT) or who had recently completed RT for head and neck squamous cell carcinoma (HNSCC). Sample sizes were limited for both the physician-directed survey and the cognitive interviews but were consistent with other published PRO measure development and content validation studies (ref), and went beyond the standard content validity sample size guidelines (ref)." [72]

DC4 - Generalizability

Discuss the generalizability of the results. For example, discuss whether the results could be generalized to other populations given the sample studied.

Explanation: Generalizability is the applicability of your results to other populations and circumstances. Different populations or settings may have unique characteristics that influence the performance of a PROM. We recommend that you discuss generalizability of the results to other populations given your sample studied, and given construct under study, and the measurement property under study. For example, if the PROM was evaluated in a different population than for which it was developed, you could reflect on differences or similarities in results (if known); or some measurement properties (e.g., reliability and content validity) are specifically dependent on the population in which it is applied. Also, if the PROM measures a more generic outcome, results may be easier to generalize compared to a disease-related outcome. Generalizability may or may not be directly related to the context of use for which the PROM was developed. For example, a measurement instrument was developed to screen patients, but is subsequently used to evaluate patients. Discussing generalizability allows readers to consider contextual factors that may impact the applicability of the PROM in diverse situations. This information is valuable for researchers, clinicians, and policymakers who seek to implement the PROM in different settings.

Essential elements

• Generalizability of the results (to other populations in terms of disease and demographic characteristics, to other settings, to other modes of administration, etc.).

Example 1

"The sample of this study was collected in a referral rehabilitation hospital, therefore, community samples or samples assisted in clinics may have particular characteristics. The Brazilian Portuguese version of the Functional Assessment Questionnaire (FAQ) showed adequate understanding and psychometric properties, but it should be investigated in other languages. Our study only involved patients with cerebral palsy (CP) and, therefore, extrapolations to other populations should be avoided." [73]

Example 2

“Furthermore, since this study was conducted in a heterogeneous sample (i.e., diversity of diagnoses) of chronic hand orthotic users, wearing the three most commonly prescribed types of hand orthoses, we are confident that the results can be generalized to the population of chronic hand orthotic users at large.” [74]

DC5 - Instrument changes

Discuss what modifications are needed to the existing PROM.

Explanation: Results of studies on measurement properties may inform adjustments or improvements of the PROM being studied. We recommend that you discuss the need to modify the existing PROM, and provide specific recommendations for improvements. For example, if the content validity results indicate that there are missing items, we recommend that you indicate which content is missing in the PROM; or if specific items were not comprehensible, suggestions for reformulation of items can be provided. Your study on structural validity may suggest another structure in the PROM than is previously suggested, and you may propose another scoring algorithm for the PROM. Each improvement to a PROM leads to a new version, which requires new evaluation. If you need to propose too many adjustments to the PROM being studies, we recommend that you strongly propose to discontinue the use of this version of the PROM (see also DC2).

Essential elements

• Discuss what modifications are needed to the existing PROM.

Example 1

"end-users should note that the removal of items pertaining to pain, headaches, lifting, reading and sleep should not be taken to suggest that these domains are unimportant. However, for either conceptual or statistical reasons, they either do not fit with the construct of neck-related function, do not function well in their current form, or are statistically redundant and therefore do not add enough additional information to warrant retention. Each of these constructs have been deemed more or less important to the neck pain experience by previous authorship groups." [75]

Example 2

“The results of our study, and of previous studies indicate that this item bank in its present form, has limitations, and it is questionable whether it can be used without restrictions in short forms and in computerized adaptive testing (CAT) applications. (...) The results of our study, together with the results of other studies reporting on psychometric properties could be used to improve the item bank, for example, by removing items with poor scalability. We recommend more validation studies of translated versions in other languages and in populations with other types of pain, especially to study dimensionality and monotonicity.” [76]

DC6 - Future research

Describe new research questions or hypotheses generated from these findings, and provide/ describe the research needed to answer those questions.

Explanation: If a PROM seems promising for use (i.e., no evidence for insufficient measurement properties yet, and no suggestions for improving the PROM in its current from), future research recommendations allow your readers to acknowledge and address the shortcomings or lack of knowledge of the quality of the PROM being studied. By highlighting gaps in knowledge or areas where further investigation is needed, the research community is informed to increase the evidence on the quality of the PROM. We recommend that you indicate what new research questions or hypotheses generated from these findings have arisen or have not yet been addressed, followed by suggestions for future research on the particular measurement property(ies) of the PROM. For example, include longitudinal studies to assess the responsiveness of the PROM. This is particularly important for measuring changes in health status over time. Future research recommendations may also suggest exploring the use of the PROM in specific subpopulations or cultural contexts. For example, in various settings, such as primary care, specialty clinics, or research settings. In summary, suggestions for future research, provide a roadmap for building evidence on the quality of the PROM.

Essential elements

• What new questions have arisen.

• Suggestions for future research on particular measurement properties of the PROM or in various settings.

Example 1

"The evaluation of responsiveness, the ability of an instrument to detect changes over time in the construct to be measured, was beyond our scope. Therefore, we suggest that future studies assess Female Genital Self-image Scale (FGSIS) responsiveness in the Brazilian population." [41]

Example 2

"The results of our study, and of previous studies indicate that this item bank in its present form, has limitations, and it is questionable whether it can be used without restrictions in short forms and in computer adaptive testing (CAT) applications. Further item response theory (IRT) analyses are recommended in a Dutch general population and on a combined set of clinical and general population data to determine the optimal IRT item parameters for use of CAT in The Netherlands and Flanders (Dutch or Flemish speaking part of Belgium). The results of our study, together with the results of other studies reporting on psychometric properties could be used to improve the item bank, for example, by removing items with poor scalability. We recommend more validation studies of translated versions in other languages and in populations with other types of pain, especially to study dimensionality and monotonicity." [76]

DC7 - Conclusions

Provide the overall conclusions for the use of the PROM.

Explanation: Providing overall conclusions of your study offers an opportunity for your readers to interpret the results in the context of the study's objectives, hypotheses, and existing literature. We recommend that you provide your overall conclusions on whether the studied measurement properties were found to be sufficient or insufficient, and conclude whether the PROM can safely be used. Clear conclusions may improve standardization of PROM selection and use. In the body of the Discussion section, we recommend that you explain your conclusion (see DC2), and we recommend that you repeat your conclusion about the applicability of the PROM being studied in the very end of the article.

Essential elements

- Overall conclusions on the studied measurement properties of the PROM.
- Overall conclusion on the use of the PROM.

Example 1

"The Dutch-Flemish PROMIS Pain Behavior item bank showed that all items fitted the item response theory (IRT) model, and our results supported cross-cultural validity. However, the assumptions of unidimensionality and monotonicity were not met. Bifactor analysis indicated a low risk of biased scores when assuming unidimensionality, although the fit of the bifactor model was still suboptimal. Our study showed limitations of the Dutch-Flemish PROMIS Pain Behavior item bank when used in a primary care population with musculoskeletal complaints." [76]

Example 2

"Comparisons of these results with the validation study of the English version are therefore limited to the examination of the convergent validity, which showed similarly high correlations with comparative measurements such as the short form of the Stroke Impact Scale 2.0 (SF-SIS) 3.0, the EQ-5D-5L and the National Institutes of Health Stroke Scale (NIHSS). The analysis of structural validity did not confirm the suggested unidimensionality of the scale and found evidence for an underlying two-factor solution with a physical and cognitive domain. Further research is needed to improve and clarify the underlying factorial structure of the SF-SIS. The discriminative validity had also not been investigated so far, but showed sufficient ability to distinguish between mildly and severely affected stroke survivors. Adequate test-retest reliability was found in stable subjects after one week." [77]

Report section: Other information

O1 - Conflict of Interest

State any conflict of interest you may have related to the PROM. This may include any involvement in the development of the PROM or any commercial funding or profit.

Explanation: A conflict of interest in a study refers to a situation in which the professional or personal interests of individuals involved in the research could potentially compromise the integrity, objectivity, or impartiality of the study. Conflicts of interest can arise in various forms and involve financial, personal, or professional relationships that may influence the design, conduct, analysis, or reporting of research findings. We recommend that you disclose any conflicts of interest relevant to the PROM being studied. For example, if you received funding to conduct the study, we recommend that you describe the funding party; if you are the developer of the PROM, we recommend that you state this; if you get profit from other parties when they use the PROM you developed, this should be declared. Many journals have guidelines and policies in place to manage and mitigate conflicts of interest.

Essential elements

• Conflicts of interest related to the PROM.

Example 1

"Stein Arne Rimehaug serves as a national contact person for PROMIS in Norway, Aaron J Kaat and Stein Arne Rimehaug are both members of the 2020 and 2021 Scientific Advisory Committee for the PROMIS Health Organization International. J E Nordvik, M Klokkerud and H S Robinson have no conflicts of interest to disclose." [78]

Example 2

"The authors declare that they have no competing interest." [29]

Example 3

"(…) is one of the developers of the Arm Function in Multiple Sclerosis Questionnaire (AMSQ)." [79]

Specific Reporting recommendations for each measurement property

Content Validity

Content validity is the degree to which the content of an instrument is an adequate reflection of the construct to be measured [1]. Three aspects of content validity are distinguished: (1) relevance (all items in a PROM should be relevant for the construct of interest, for the target population and for context of use), (2) comprehensiveness (no key aspects of the construct should be missing), and (3) comprehensibility (the items should be understood by patients as intended). These aspects can all be assessed in the same study using the same methods.

Content Validity: Methods

CV1 - Relevance

Specify if and how patients and/or professionals were asked whether the instructions, each of the items, response options, and the recall period were relevant for the construct, population, and context of use.

Explanation: To ensure that the PROM adequately captures the intended construct for the target population and context of use, we recommend that you specify whether and how patients and/or professionals (e.g., clinicians, researchers) were asked whether the instructions, each of the items, the response options, and the recall period were relevant for the construct, the study population, and the context of use. Each of these aspects should be evaluated separately. The recall period can be included though in the instructions or in the items, so it may simultaneously be evaluated. Patients and/or professionals assess the relevance of the PROM based on their experience or expertise with the construct. Relevance of the PROM can be assessed using a qualitative and/or quantitative (survey) study. Qualitative studies involving direct communication with patients and/or professionals, such as through focus groups or individual interviews, are recommended to adequately capture their perspectives on issues of importance relative to the focus of the PROM [80]. When focus groups or individual interviews were conducted, we recommend that you report whether you used a structured, semi-structured, or unstructured approach and describe the topic or interview guide used (the full topic or interview guide is often published as a supplementary document). The purpose of this guide is to establish procedures, maintain consistency in data collection, prevent undue

influence by the moderator/interviewer, and keep the conversation on topic. It is also recommended to report whether open-ended questions were asked to allow for spontaneous reporting and ensure unbiased data collection. If you used interviews or focus groups, we recommend that you report whether you recorded them with audio or video equipment to fully capture the context and content of each session, and whether all interviews or focus groups were transcribed verbatim. We also recommend that you report if skilled moderators/interviewers were involved to interview the patients. To report data analysis in a qualitative study, you should report the method chosen (e.g., content analysis, deductive analysis, framework analysis, grounded theory) along with the type of coding used (e.g., open, axial, selective), and whether at least two people independently coded part of the data. If two people coded the data, we recommend that you report the percentage of codes that were coded repeatedly. We also recommend that you report if and how the qualitative data were collected until no new relevant knowledge was gained from new patients or other approaches used to justifying the sample size. In quantitative (survey) studies, participants are asked to rate the relevance of each aspect of the PROM (e.g., the instructions, each of the items, the response options, and the recall period) to the construct being measured, the study population, and the context of use. Participants may be asked to provide ratings on a Likert scale and/or open-ended questions to gather qualitative feedback on the relevance of aspects of the PROM.

Essential elements

• Whether patients and/or professionals were asked whether instructions, each of the items, the response options and the recall period were relevant for the construct, population, and context of use.
• Method used (qualitative and/or quantitative).
• In a qualitative study, which kind of method (focus groups or individual interviews) and approach (structured, semi-structured, or unstructured) were used.
• In a qualitative study, description of a topic or interview guide; whether open-ended questions were used; whether skilled moderators/interviewers were used; description of the method of data analysis (e.g., content analysis, deductive analysis, framework analysis, grounded theory); type of coding used (e.g., open coding, axial coding, selective coding); and evidence for no new relevant knowledge was gained from new patients.

Example 1

"(...) face-to-face cognitive interviews were held during December 2019 and January 2020. Interviews were held at the De Hoogstraat Rehabilitation center or in the respondents' homes, as they wished. (...) The preliminary 25-item version of the self-regulation assessment (SeRA) was used in the first series of cognitive interviews. Respondents were asked to think aloud and explain all of their thoughts when answering the items. Probing was used when respondents did not spontaneously express their thoughts on the items. Reflective questions on the SeRA in general were asked directly after the cognitive interview, e.g., whether the items fully reflect the concept, what their thoughts were concerning the items in general, reflection on the instructions and the response options, and if they missed items. Interviews were held in the Dutch language and were audiotaped. Interviews lasted between 40 and 90 min. Prior to the interviews, a few test and feedback interviews were conducted to ensure qualification of the interviewer. All eight interviews were held by the same interviewer (KH), and during the first three interviews, her supervisor (TM) was the second interviewer.
All audio recordings were transcribed verbatim and anonymized. Data analyses were performed by using MaxQDA software (Verbi Gmbh MaxQDA 2018.2). Thematic analysis and open coding were used for content analysis. Cognitive interviews were initially analyzed by the interviewer, and next by the first author of this manuscript. The coding system of the analyses was based on three categories in which 10 criteria for good content validity are divided: (a) comprehensibility, (b) comprehensiveness, and (c) relevance (ref). Comprehensibility covers four criteria: instructions understood as intended, items and response options understood as intended, items worded appropriately, and match between questions and response options. Comprehensiveness refers to the criterion whether no key concepts are missing. Finally, relevance contains five criteria: all items relevant for the construct, for the target population, and for the context of use; are response options appropriate; and if the recall period is appropriate (ref). (...) All authors involved in the content analyses were, or were supervised by, experienced qualitative researchers and were from different backgrounds such as psychology, health sciences, and rehabilitation medicine. (...) To ensure the data credibility, each transcript was independently analyzed by at least two researchers." [81]

CV2 - Comprehensiveness

Specify whether and how patients and/or professionals were asked whether all key concepts are included in the PROM.

Explanation: To determine whether the PROM adequately captures the breadth and depth of the target construct, we recommend that you specify whether and how patients and/or professionals were asked whether all key concepts are included in the PROM. Comprehensiveness should be assessed for all subscales. Comprehensiveness of the PROM can be assessed using a qualitative and/or quantitative (survey) study. All recommended methods for assessing the comprehensiveness are similar to those for assessing relevance and are described previously (see CV1).

Essential elements

• Whether and how patients and/or professionals were asked whether all key concepts are included in the PROM.
• Method used (qualitative and/or quantitative).
• In a qualitative study, which kind of method (e.g., focus groups, individual interviews, Delphi method) and approach (structured, semi-structured, or unstructured) were used.
• In a qualitative study, description of a topic or interview guide; whether open-ended questions were made; whether skilled moderators/interviewers were used; the method of data analysis (e.g., content analysis, deductive analysis, framework analysis, grounded theory); type of coding used (e.g., open coding, axial coding, selective coding); evidence for no new relevant knowledge was gained from new patients.

Example 1

"The study objective was to assess the content validity of the Cough and Sputum Assessment Questionnaire (CASA-Q) cough domains and the Shortness-of-Breath Questionnaire (UCSD-SOBQ) for use in patients with Idiopathic Pulmonary Fibrosis (IPF). (…) The cough domains of the CASA-Q include 11-items (...). The 9 questions of the CASA-Q sputum domains were not included for this assessment (...). The study consisted of one-on-one interviews. Participants completed the CASA-Q cough domains and the UCSD-SOBQ, followed by debriefing questions using a semi-structured interview guide. The interview guide contained questions about the participant's understanding of the instructions for each instrument, the recall period, the intended meaning and relevance of the items and response options, and general questions about the overall instrument and missing concepts." [82]

Example 2

"Cognitive debriefing (~40 min) involved the participants completing each Vitiligo Patient Priority Outcome (ViPPO) measure using a 'think aloud' technique in which they spoke their thoughts aloud while reading the questions and selecting an answer. Participants were asked direct and specific questions to establish the content validity of the ViPPO measures (Table 1)."

Table 1. Example cognitive debriefing questions

Concept	Example interview questions
Understanding of item wording	What does the question mean to you? How would you say this in your own words?
Interpretation of the measurement concept	What does [self-conscious] mean to you?
Relevance of the measurement concept	Have you ever experienced this? Can you give me an example?
Item redundancy	Do you think feeling [embarrassed] and [self-conscious] are the same or different?
Relationship between items	Is it important to ask about feeling [sad] and feeling [depressed] separately? Or could be questions be combined?
Comprehensiveness	Was there anything missing from the questionnaire?
Suitability of the different response scales/ options	What are your thoughts on the response scale? Which scale do you prefer?
Relevance of the recall period to the measurement concepts	How easy or difficult was it to answer thinking about the past 7 days? Would a different timeframe be easier to think about?

[83]

CV3 - Comprehensibility

Specify whether and how the comprehensibility of the PROM instructions, items, response options, and recall period was evaluated by patients and/or professionals.

Explanation: If the PROM instructions, items, response options, and/or recall period are unclear, incorrect information may be gathered or patients may get frustrated because they do not understand how to complete the PROM. Assessing the comprehensibility of a PROM ensures that the language, wording, and format of the PROM are clear and accessible to the intended respondents. We recommend that you specify whether and how the comprehensibility of the PROM instructions, items, response options, and recall period was evaluated by patients and/or professionals. For the comprehensibility assessment, you can use widely recognized methods (e.g., Think aloud method, Three-Step Test-Interview, debriefing, probing, or other forms of cognitive interviewing). Assessing comprehensibility can be part of a development process of a PROM (also called the pilot testing phase of the PROM), as a content validity study, or as part of a translation process. In essence, the methods for testing comprehensibility are similar in these different phases. We recommend that you report the phase in which comprehensibility is being tested. We recommend that you described details about the use of a topic or interview guide, recording and transcription, moderators or interviewers, methods to analyze the data and type of code used (including people involved), similar as described for assessing relevance (see CV1).

Essential elements

• Whether and how patients and/or professionals were asked about the comprehensibility of the instructions, each of the items, the response options and the recall period for the construct, population, and context of use; phase in which comprehensibility is being tested (i.e., development phase, content validity phase, or translation phase).
• Method used (qualitative and/or quantitative).
• In a qualitative study, which kind of method and approach (structured, semi-structured, or unstructured) were used.
• In a qualitative study, description of a topic or interview guide; whether open-ended questions were made; whether skilled moderators/interviewers were used; the method of data analysis (e.g., content analysis, deductive analysis, framework analysis, grounded theory); type of coding used (e.g., open coding, axial coding, selective coding); evidence for no new relevant knowledge was gained from new patients.

Example 1

“Interviews took place between November 2020 and March 2021, during the COVID-19 pandemic, and thus were conducted remotely. Interviews were conducted by one of six trained qualitative interviewers (...) All interviews were audio-recorded and transcribed verbatim. Transcripts were analysed using ATLAS.ti v7.5. Each transcript was coded by one analyst, with codes and quotations reviewed by the project lead at regular intervals. The project team revisited the verbatim quotes when summarizing the findings. (…) As recommended in the Best Practice Guidelines for Establishing Content Validity of PRO [patient-reported outcome] Instruments [ref] and FDA PFDD Draft Guidance 2 [ref], interview transcripts were grouped into four equal sets in chronological order, and the elicited concepts were compared between sets. (...) Data obtained via cognitive debriefing were subject to framework coding [ref], whereby a predefined code list was applied to identify the relevance and appropriateness of item wording, response options and the recall period.” [83]

Content Validity: Results

CV4 - Relevance

Specify if the instructions, all items, response options, and recall period were considered relevant, by patients and/or professionals, to the construct, population, and context of use.

Explanation: We recommend that you specify whether the instructions, items, response options, and recall period were considered relevant for the construct, population, and context of use by patients and professionals. If patients and/or professionals identified important issues (e.g., some items are not relevant to the population, or the recall period is not relevant for the construct), you can report the number (percentage) of items that were considered relevant (and not relevant) and indicate which items (or aspects of the PROM) were considered not relevant. If some items are relevant to only a small number of patients, but are necessary to capture the full range of patient experiences, we recommend that you provide justification for their inclusion. If your conclusion is that the content validity is insufficient in terms of relevance and the PROM needs to be improved, we recommend that you give suggestions for improvement based on your results (see DC5).

Essential elements

• Whether patients and/or professionals considered the *instructions*, *all items*, *response options*, and *recall period* relevant for the construct, population, and context of use of interest.
• Number and/or percentage of patients and/or professionals who did not consider the *instructions*, *items*, *response options*, and *recall period* relevant.

General example

"The relevance of the instructions, items, response options, and recall period of the [PROM name] was assessed by *x* professionals and *y* patients. All the professionals considered the instructions, items, response options, and recall period of the [PROM name] relevant for assessing [construct] in [population and context of use of interest]. Most of the patients (n = *z*, *z*%) considered item A irrelevant to the assessment of [construct] in [population and context of use of interest]. The instructions, other items, response options, and recall period were considered relevant for all patients."

Example 1

“Eight respondents were included. Respondents underscored the relevance of the items for a rehabilitation population. They particularly emphasized the items on gaining insights and understanding, and items on determination of own boundaries or possibilities. The respondents marked them as very relevant for medical rehabilitation treatment, and also for the construct of self-regulation. Comments from two respondents were: (...). Respondents also expressed that all items were relevant. An example quote is: (...). Also, items were marked as relevant for the rehabilitation population. Two respondents commented as follows: (...). Four items were marked as overlapping with other items, because they were similar to other items, or due to their wording. Concerning the response options, a few remarks were made that these were not appropriate for each item. Two respondents indicated that answering the questions is not a matter of indicating agree or not agree and missed the option to clarify their answers. Finally, respondents seemed not to have any difficulties regarding the recall period. None of the respondents had problems to remember anything needed to answer the items.” [81]

CV5 - Comprehensiveness

Specify whether patients and/or professionals considered all key concepts to be included in the PROM.

Explanation: We recommend that you specify whether patients and/or professionals considered all key concepts to be included in the PROM. You may report the results of the comprehensiveness in terms of what aspects were missing, and how many patients and/or professionals considered the aspect to be missing. The percentage of patients/professionals who considered the PROM to be comprehensive could also be stated. Patients may have reported missing content that is not related to the intended construct or context of interest of the PROM. You could report this, and explain why it is not included in the PROM. This way, you acknowledge the input of patients, but explain it is beyond the scope of the PROM being studied.

Essential elements

- Whether patients and/or professionals considered all key concepts to be included in the PROM.
- Specify which concepts were considered missing.

Example 1

"[Participants in total: Providers (n = 10) and Family members (n = 9)]. Families suggested adding questions about medication choices (n = 1), feeling comfortable with the provider (n = 1), the ability to stop or refuse treatment (n = 1), addressing crises in a timely fashion (n = 3), role of family involvement (n = 5), and locating services (n = 1). Providers suggested adding questions about accessible information (n = 2), overall service effectiveness (n = 1), evaluating previous experiences (not current experiences, n = 1), the role of family involvement (n = 3), insurance (n = 1), provider preferences (n = 2), and other types of therapeutic approaches (e.g., outdoor therapy, n = 1)." [84]

Example 2

"The Vitiligo Patient Priority Outcome (ViPPO) items were well understood by all participants, including adolescents (Supplementary Material 4). The instructions were largely understood (n = 55/60, 92%). Five participants initially considered their total body when completing the ViPPO facial vitiligo only (ViPPO-F) but understood the intention when the interviewer repeated the instruction.

No participants experienced difficulty understanding either response scale:

"I think it's good. I think it gives you plenty of choices, um, to, like, measure exactly where you feel about it." (Female, 51 years old; 'frequency' responses)

"I think they were a good choice of answers to pick out of." (Female, 13 years old; 'severity' responses)

When asked which response scale they preferred (Fig. 2), over half (n = 33/58, 57%) preferred 'frequency'; 18 (n = 18/58, 31%) preferred 'severity'; six (n = 6/58, 10%) had no preference; while an additional participant (n = 1/58, 2%) felt that their preference changed for each domain.

Fifty-three participants (n = 53/59, 90%) discussed the recall period. Of these, 24 (n = 24/53, 45%) understood and used the recall period correctly throughout:

"I'm thinking in the last seven days […] I kept saying that was the last seven days because some of my answers would be different if you would have asked me in the past, you know, six months." (Female, 37 years old)

Thirty-seven participants (n = 37/59, 63%) were asked whether it would be easy or difficult to recall the past 7 days, regardless of whether they had used the recall period correctly. Most (n = 29/37, 78%) felt that it would be easy. Eight (n = 8/37, 22%) felt that it would be difficult, half of whom (n = 4/8, 50%) suggested that the activities on the ViPPO-F are unlikely to have happened over a 7-day period, particularly the concepts included in the social functioning domain." [83]

CV6 - Comprehensibility

Specify whether patients understood the PROM instructions, items, response options, and recall period as intended and/or whether professionals considered the instructions, items, response options, and recall period were appropriately worded.

Explanation: we recommend that you specify whether patients understood the PROM instructions, items, response options, and recall period as intended by the developers and/or whether professionals considered the instructions, items, response options, and recall period appropriately worded. Specific findings may include insight into the clarity of the wording, any potential ambiguities, and the overall ease of understanding the content of the PROM. If patients and/or professionals identified important issues (e.g., some items have highly technical language for the population), we recommend that you report the number (percentage) of items that were comprehensible (and not comprehensible) and indicate which items (or aspects of the PROM) were considered not comprehensible. If your conclusion is that the content validity is insufficient in terms of comprehensibility (see DC1) and the PROM needs to be improved, you can give suggestions for improvement based on your results (see DC5).

Essential elements

- Whether patients understood the PROM *instructions*, *items*, *response options*, and *recall period*.
- Whether professionals considered PROM *instructions*, *items*, *response options*, and *recall period* appropriately worded.

Example 1

"The intended meaning of the items was clearly understood by most of the participants (89-100%) (Figure 1). Two participants (11%) reported difficulty in understanding Item 10. One participant indicated that they could not differentiate between "falling asleep" and "falling back to sleep." The other participant indicated that "waking up," "preventing you from falling asleep," and "preventing you back to sleep" were the same concept. For Item 1 and 2, one participant was either not asked or did not reply to the question during the interview, resulting in missing data. (…) The Cough and Sputum Assessment Questionnaire (CASA-Q) instructions were well understood by all participants. The recall period was (...) [understood] by 89% (n=16) of the study participants. A majority of the participants (n=13 of 17, 76%) demonstrated understanding [by debriefing interviews] of the response options. The remaining 24% (n=4) [participants] demonstrated difficulty in distinguishing specific response options, such as, distinguishing "rarely" from "sometimes," distinguishing "quite a bit" from "a lot," distinguishing "somewhat" from "a little," and "often" from "always."" [82]

Structural Validity

Structural validity refers to the degree to which the scores of a PROM are an adequate reflection of the dimensionality of the construct to be measured [1]. Two types of measurement models exist, reflective and formative models. In a reflective model the items reflect an underlying latent construct and are expected to be highly correlated and interchangeable. In a formative model, the items together form the construct and do not need to be correlated [85, 86]. Factor analysis and Item Response Theory (IRT) analyses can be used to evaluate structural validity of PROMs based on a reflective model. The factor model (e.g., a unidimensional model) represents the intended structure of the PROM and informs you about its scoring algorithm (e.g., calculating one sum score, or theta score). Standards to evaluate the structural validity of formative models are not yet part of the COSMIN guidelines.

Structural Validity: Methods

SV1 - Rationale for approach

Provide a rationale for the approach (e.g., factor analysis, Item Response Theory (IRT)/Rasch analysis) used.

Explanation: To ensure that your approach is consistent with the objectives of your study, we recommend that you report why you chose the approach you used to assess structural validity (e.g., exploratory or confirmatory factor analysis, Item Response Theory (IRT)/Rasch analysis). These approaches require reflective measurement models. Therefore, we recommend that you state and explain that the PROM being studied is indeed based on a reflective model. Exploratory factor analysis (EFA) is indicated when no clear hypotheses exist about the underlying dimensions. Confirmatory factor analysis (CFA) tests whether the data fit a predetermined factor structure based on theory or previous analyses. A priori hypotheses about the structure should be reported and tested [87]. While factor analysis allows researchers to explore or confirm the underlying factor structure of an instrument, IRT/Rasch analysis can be used to assess the relationship between the patient's responses and the latent trait.

Essential elements

• Rationale for the approach used.

Example 1

"For CFA, three models were compared: (...). These models were based on the results of EFA and compared with results on the internal structure of Pelvic Floor Distress Inventory (PFDI-20) in the literature, assuming that the three subscales of PFDI-20 form 3 dimensions (pelvic organ prolapse, anorectal and urinary symptoms) and the five factors found by Ma et al. [ref]." [29]

Example 2

"Since the Norwegian Fatigue Characteristics and Interference Measure (FCIM) is a new instrument and includes items with different response categories, we applied a Partial Credit Model (PCM) which make no assumptions about the equidistance between thresholds across items [ref]. Then we assessed rating scale functioning according to Linacre's guidelines to determine whether the scale was suitable for Rasch analysis [ref] (Table1). The primary focus in the Rasch analysis was to address two main aims: item reduction and evaluation of FCIM's structural validity and internal consistency." [88]

SV2 - Statistical analyses

Exploratory (EFA) or confirmatory factor analyses (CFA)

Describe the tested model (e.g., number of factors, which items included in which factor), method of estimation, type of correlation matrix, and methods and criteria for good model fit.

IRT/Rasch analysis

Describe the type of IRT/Rasch model, method of estimation, methods for checking assumptions (e.g., dimensionality, local independence, monotonicity), methods and criteria for good item parameters and model fit. Indicator software and version used.

Other approaches

Provide details of the methods if other approaches were used.

Explanation: Factor analysis and IRT/Rasch analysis are important to test the dimensionality of the PROM and to ensure that items are coherently explained by their respective (reflective) dimensions. If EFA or CFA were used, you should provide the method of estimation (e.g., maximum likelihood, weighted least squares, etc.), type of correlation matrix (e.g., polychoric correlations for ordinal item scores), the tested model(s) (i.e., for EFA, how many factors can be extracted; for CFA, a path diagram explaining which items are a reflection of which factor), methods and criteria for good model fit (e.g., factor leadings; scree plot and eigenvalues above 1 for EFA; or goodness-of-fit statistics for CFA (e.g., CFI, TLI or comparable measure, RMSEA or SRMR)). If IRT/Rasch analysis was used, we recommend that you provide the type of IRT model used (e.g., 1 or 2 parameter model (e.g., partial credit model or graded response model)), the estimation method (e.g., maximum likelihood, weighted least squares/ the marginal maximum likelihood (MML) estimation based on the expectation maximization (EM) algorithm, or the fully Bayesian estimation via the use of the Markov chain Monte Carlo (MCMC) simulation techniques), methods for testing assumptions (e.g., dimensionality, local independence, monotonicity), and methods and criteria used for model fit and retaining or excluding items.

Essential elements

For EFA/CFA:

- Method of estimation.
- Type of correlation matrix.
- Tested model (e.g., number of factors, which items included in which factor).

• Methods and criteria for good model fit.

For IRT/Rasch analysis:

• Type of IRT/Rasch analysis (e.g., 1 or 2 parameter model (e.g., partial credit model or graded response model)).

• Method of estimation.

• Methods for testing assumptions (e.g., dimensionality, local independence, monotonicity).

• Methods and criteria used for model fit and retaining or excluding items.

Example 1

"(…) it was necessary to assess the structural validity of this WHODAS 2.0 version in a sample of women with dysmenorrhea by exploratory factor analysis and confirmatory factor analysis. To assess the factorability of the data, Kaiser-Meyer-Olkin (KMO) test and the Bartlett sphericity test were used. KMO ≥ 0.70 and p ≤ 0.05 in the Bartlett sphericity test indicated criterion to perform exploratory factor analysis. We performed exploratory factor analysis on the total sample, and used Minimum rank and Parallel Analysis to retain the number of factors with quartimax rotation. If necessary, items with factor loading < 0.40 were excluded [ref]. For the confirmatory factor analysis, the sample was divided equally and randomly in SPSS 22 [and tested in all models]. Thus, we used the maximum likelihood robust [method of estimation], Root Mean Square Error of Approximation (RMSEA), Standardized Root Mean Squared Residual (SRMR), Comparative Fit Index (CFI) and Tucker-Lewis Index (TLI). The model was considered adequate when RMSEA and SRMR < 0.08, CFI and TLI > 0.90 [ref]. Items with higher modification indices (MI) had error covariances." [17]

Example 2

“We assessed unidimensionality using a scree plot, Kaiser criterion analysis (with a threshold of 1.0 eigenvalues) and a confirmatory factor analysis (CFA) with polychoric correlation and a diagonally weighted least squares estimator in the lavaan package (version 0.6-11). (…) We used the following fit statistics and thresholds to indicate good model fit: root mean squared error of approximation < 0.06, standardized root means square residual ≤ 0.08, comparative fit index ≥ 0.95, and Tucker-Lewis index ≥ 0.95. (…) We took Loevinger's Hi values > 0.3 to indicate monotonicity. (…) We tested for this [local independence] using Yen's Q3 residual correlation statistic, with a threshold of > 0.20 indicating undesirable local dependence between items. (…) Using the mirt package (version 1.36.1), we fitted graded response models (GRMs) to the complete preoperative item response sets in each dataset and used these to calculate IRT scores (specifically, expected a posteriori scores computed with a standard normal prior).” [89]

Structural Validity: Results

SV3 - Statistical analyses

EFA or CFA

For EFA: provide all factor loadings, eigenvalues and % variance explained of the model reflecting the original PROM structure and best-fitting model if applicable.

For CFA: provide all factor loadings and results for model fit indices of the model reflecting the original PROM structure and best-fitting model, if applicable.

IRT/Rasch analyses

Provide item/model fit results, all item parameters, and figures if appropriate (e.g., item characteristic curves, person-item mapping, item and/or test information functions).

Other approaches

Provide relevant results for other approaches used.

Explanation: Providing the results of your factor analysis, or IRT/Rasch analysis is essential for the readers of your study to interpret the dimensionality of the PROM, and understanding item performance. We recommend providing results for both the most commonly used structure of the PROM (e.g., if a total score is always used, we recommend providing results for a unidimensional model), and if the results show that this model does not hold in your data, provide results for other tested models that do hold. If you conducted factor analysis, we recommend that you provide the factor loadings of the tested model(s), the results of which are used to determine model fit: for EFA we recommend that you provide eigenvalues, a scree plot, and the percentage of variance explained; for CFA, we recommend that you provide the results of the goodness-of-fit statistics you choose to use (e.g., CFI, TLI, RMSEA, etc.). In addition, modification indices could be provided and residual correlations. If you conduct IRT, we recommend that you provide information on the assumptions for unidimensionality and local independence, on the item/model fit, and all item parameters. In addition, item characteristic curves, person-item mapping, and test information functions are recommended to be provided. All this information can be presented in tables or figures in the manuscript itself, or in an Appendix.

Essential elements

For EFA/CFA:

• Factor loadings of the model reflecting the original PROM structure.

• Best-fitting model (if applicable).
• Eigenvalues and/or % variance explained.
• Model fit indices.

For IRT/Rasch analysis:

• Item/model fit.
• Item parameters.
• Assumptions for unidimensionality and local independence.
• Figures (if appropriate) (e.g., item characteristic curves, person-item mapping, test information functions).

Example 1

"In the total sample, Bartlett's sphericity test [$\chi 2(df) = 807.58(11)$, $p < 0.0001$] and KMO (0.913) were adequate. A single factor was suggested with an explained variance of 51% for WHODAS 2.0.

Table 2 presents the factor loadings (> 0.50) and fit indexes of the WHODAS 2.0 for the split sample. The confirmatory factor analysis for showed the one-factor structure with the following fit indexes for model 1 [$\chi 2(df) = 388.576(54)$, $p < 0.001$; CFI = 0.462; TLI = 0.342; RMSEA = 0.093 (90%CI 0.089-0.102); SRMR = 0.066]. To improve model fit, error covariances were freed items 8 and 9 (MI = 206.677), 10 and 11 (MI = 124.583), 1 and 7 (MI = 52.526), 1 and 2 (MI = 49.172), 2 and 11 (MI = 48.225), 5 and 6 (MI = 28.442), and 1 and 11 (MI = 22.067). This resulted in an adequately-fitting model 2 [$\chi 2(df) = 93.118(46)$, $p < 0.001$; CFI = 0.924; TLI = 0.900; RMSEA = 0.038 (90%CI 0.027-0.049); SRMR = 0.029]. The path diagram of the model 2 for WHODAS 2.0 is shown in Fig. 1. Although the structure of model 2 contains correlated errors, it should be accepted as a better structure compared to model 1 due to better indexes."

Table 2. Factor loadings and fit indexes of WHODAS 2.0.

Items	Factor loadings Split sample (n=717)	
	Model 1	Model 2
1. Stand up	0.655	0.637
2. Do housework	0.698	0.705
3. Learn a new task	0.695	0.699
4. Participate in community activities	0.704	0.714
5. Be emotionally affected	0.675	0.668
6. Focus for 10 minutes	0.637	0.630
7. Walking for long distances	0.649	0.635
8. Take a shower	0.547	0.505
9. Dress up	0.617	0.586
10. Dealing with strangers	0.602	0.590
11. Keep a friendship	0.592	0.609
12. Work	0.652	0.657
χ^2(df)	388.576 (54)	93.118 (46)
CFI	0.462	0.924
TLI	0.342	0.900
RMSEA (90% CI)	0.093 (0.084 – 0.102)	0.038 (0.027 – 0.049)
SRMR	0.066	0.029

WHODAS 2.0: WHO Disability Assessment Schedule 2.0. CFI: Comparative Fit Index. df: degrees of freedom. RMSEA: Root Mean Square Error of Approximation. SRMR: Standardized Root Mean Squared Residual. TLI: Tucker-Lewis Index.

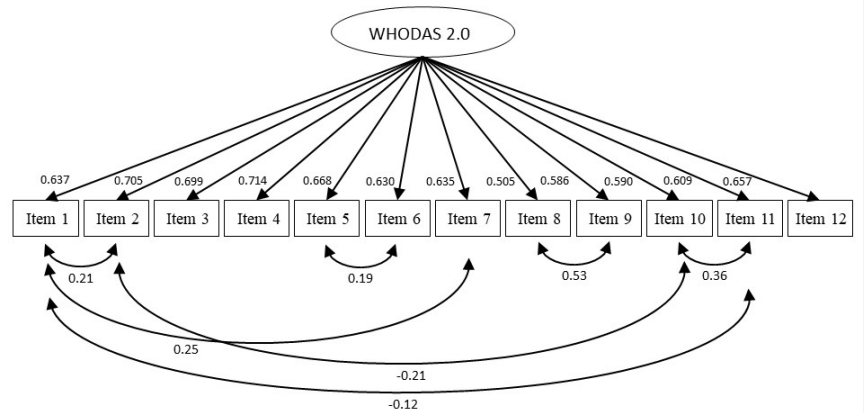

Figure 1. Path diagram of the WHODAS 2.0 for dysmenorrhea.

Table 3. Multigroup Confirmatory Factor Analysis among geographic regions of Brazil for the WHODAS 2.0.

Measurement invariance	χ^2(df)	CFI	RMSEA	$\Delta\chi^2$(df)	ΔCFI	ΔRMSEA
Geographic regions of Brazil (Southeast vs. Northeast vs. South vs. North vs. Midwest)						
Configural invariance	1337.078(270)	0.843	0.119	-	-	-
Metric invariance	1417.469(314)	0.838	0.113	80.391(44)	0.005	0.006
Scalar invariance	1507.029(358)	0.831	0.108	89.56(44)	0.007	0.005

WHODAS 2.0: WHO Disability Assessment Schedule 2.0. CFI: Comparative Fit Index. ΔCFI: Difference in Comparative Fit Index. RMSEA: Root Mean Square Error of Approximation. ΔRMSEA: Difference in Root Mean Square Error of Approximation. df: Degrees of freedom.

[17]

Example 2

"Unidimensionality within the scales was concluded from the three performed single CFAs. (...) The basic IRT assumptions were evaluated and showed monotonicity (H for pain interference items ranged 0.59 to 0.70 [total scale H = 0.68], fatigue items ranged 0.53–0.69 [total scale H = 0.63] and physical activity items ranged 0.48–0.72 [total scale H = 0.65]), and local independence was found among the items. (...) The item parameter estimates and the χ2 mean square item fit statistics are shown in Table 2. In this table the items are sorted in order of decreasing discrimination (a), so the generally best indicators of pain interference, fatigue, and physical activity are near the top of the tables. The best and the worst discriminating items are shown in category characteristic curves, see Fig. 2. (…) For the pain interference items, five of the items exhibited significant lack of fit as indicated by the SS χ2 item fit ($p < 0.001$, χ2 ranged from 503.88 to 754.07, df = 391) (Table 2), after Benjamini-Hochberg correction for multiplicity. For the fatigue items, three of the items showed significant lack of fit ($p < 0.05$, χ2 ranged from 887.04 to 1232.74, df = 636), and for physical activity items, three items showed significant lack of fit ($p < 0.05$, χ2 ranged from 856.52 to 1007.04, df = 662).
The Test Information Function (TIF), Item Information Curves (IIC), and Standard Errors (SE), were satisfactory (see Fig. 3). SE for pain interference items ranged from 0.07 to 0.62 (M = 0.35, SD = 0.68), SE for fatigue items ranged from 0.11 to 0.49 (M = 0.19, SD = 0.70), and SE for physical activity items ranged from 0.16 to 0.52 (M = 0.22, SD = 0.70)."

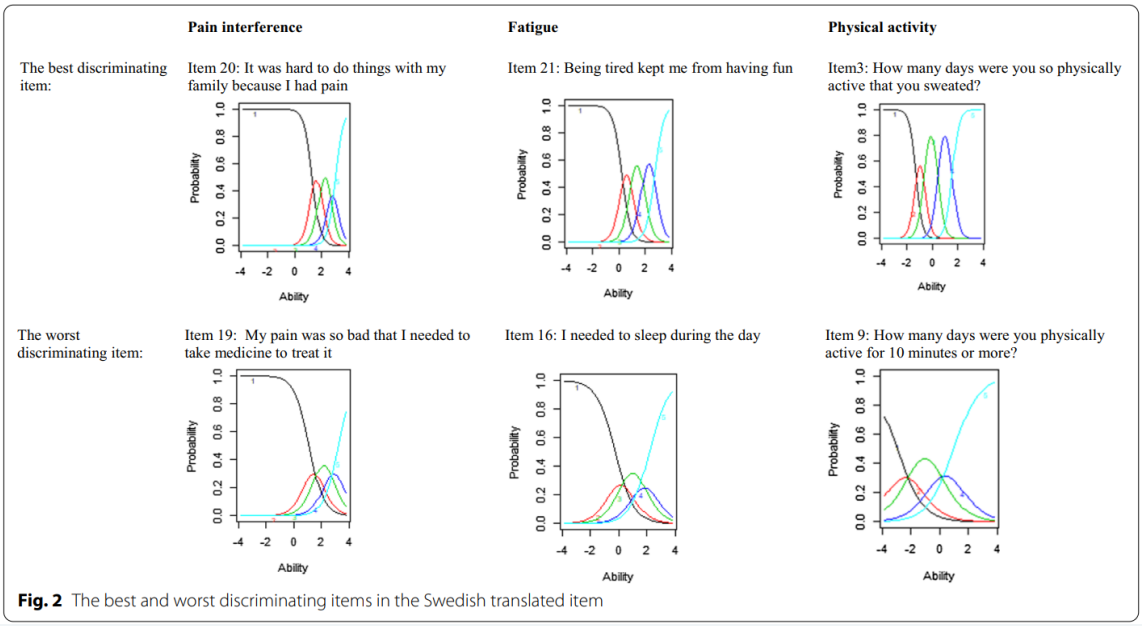

Fig. 2 The best and worst discriminating items in the Swedish translated item

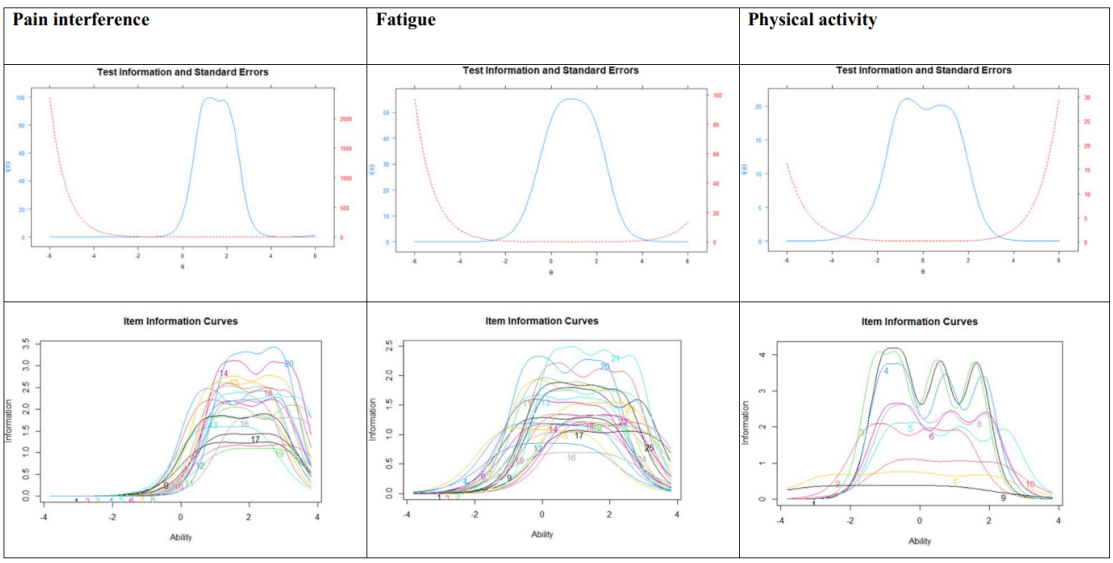

Fig. 3 Test information function, standard error, item information curves of the Swedish translated PROMIS item banks

[90]

Internal Consistency

Internal consistency refers to the degree of interrelatedness among the items [1]. For an appropriate interpretation of the internal consistency parameter, the items together should form a unidimensional scale or subscale. Internal consistency and unidimensionality are not the same. Internal consistency is the relatedness among the items [91]. Unidimensionality refers to whether the items in a scale or subscale measure a single construct. It is a prerequisite for a clear interpretation of the internal consistency statistics [91, 92], and can be investigated for example by factor analysis [93] or IRT methods.

Internal Consistency: Methods

IC1 - Statistical analyses

Provide evidence for the unidimensionality of the PROM (subscales) and provide evidence of lack of local item dependence. Describe statistical methods used to calculate internal consistency.

Explanation: To be able to appropriately interpret the internal consistency statistic (e.g., Cronbach Alpha), the scale needs to be unidimensional. Therefore, we recommend that you check the assumption of unidimensionality, for example, by showing results from a factor analysis or citations to studies in which this is shown. This can come from a study on structural validity (in case of multi-dimensional PROMs), or alternatively, the unidimensionality per subscale could be tested. If a total score is used for a multidimensional PROM, evidence should be provided that this total score also fits one factor (e.g., based on a one-factor model, the general factor in a bifactor model, or the first-order factor in a hierarchical factor model). Internal consistency is subsequently calculated with Cronbach's Alpha, Omega-total statistics, or a standard error (SE) of theta (in an IRT/Rasch study). We recommend that you report which statistic you use.

Essential elements

- Evidence for the unidimensionality of the PROM (subscales).
- Statistics used.

Example 1

“Cronbach's alpha was calculated separately for each construct [unidimensional scale] of the Bilateral Vestibulopathy Questionnaire (BVQ) to measure the internal consistency of each item in its respective construct. A low Cronbach's alpha suggested a low or absent correlation between each item in the same construct, whereas a very high Cronbach's alpha indicated that the items measured identical concepts. Therefore, Cronbach's alpha was considered good when it ranged from 0.70 to 0.95 (ref).” [94]

Internal Consistency: Results

IC2 - Statistical analyses

Provide internal consistency results for each unidimensional scale or subscale separately.

Explanation: Providing the internal consistency results helps to establish the reliability of the instrument. You should provide results for each unidimensional scale or subscale separately. If the internal consistency was calculated for numerous (sub)scales, you can present the results in a table.

Essential elements

• Results for each unidimensional scale or subscale separately.

Example 1

"(...) exploratory factor analyses (EFA) (...) suggested the one-dimensionality of the Pelvic Floor Distress Inventory (PFDI-20) (…) The Cronbach's alpha value of the one-dimensional model was considered adequate (α = 0.93)." [29]

Cross-Cultural Validity\Measurement Invariance

Cross-cultural validity refers to the degree to which the performance of the items of a translated or culturally adapted PROM are an adequate reflection of the performance of the items of the original version of the PROM. When using PROMs, we assume that patients from different groups (e.g., different countries or languages) with the same true score on the construct or latent trait (the same level of, e.g., pain or depression), have the same probability of choosing a certain response option on a PROM item. If this is not the case, the item is measurement non-invariant or differential item functioning (DIF) occurs in this item. Violations of measurement invariance or the occurrence of DIF may preclude meaningful interpretation of measurement data. For example, an item was not well translated, and has another meaning in the new language; or due to cultural differences an item can get another meaning. For example, the item whether you are able to stand up from the toilet may have another meaning to people in the Netherlands than it has to people in Canada, because in general the toilets in Canada are lower, and it is more difficult to stand up from a lower toilet.

Testing for measurement invariance typically involves a series of statistical analyses comparing the structure of the measurement model across groups, using multiple-group confirmatory factor analysis (MGCFA), or DIF tests. Each approach has unique advantages: MGCFA assesses whether the factor structure (configural invariance), and factor loadings (metric invariance) is equivalent across groups; DIF can be assessed by a series of regression analyses and enables detailed item-level analysis. First, testing whether the response on an item (outcome) depends on using either the IRT-based theta score (where theta represents the latent trait being measured) or the sum score (which is based on the total score or sum of responses to individual items on a scale) (model 1); secondly, you test whether a group variable also determines the item response (outcome) (model 2); lastly, you test whether there is even an interaction between theta score or sum score and group variable (model 3). By comparing Model 1 and 2, you test for uniform DIF; by comparing Model 2 and 3, you test for non-uniform DIF. To run any of these analyses, you need to have PROM data from two or more different (sub)groups (e.g., using two different language versions). Subgroups can also be based on any other variable, such as age, gender, and disease characteristic. The principles to test for measurement invariance or DIF are the same regardless of which subgroup variable is used in the analysis.

Cross-Cultural Validity\Measurement Invariance: Methods

CCV1 - Groups variable

Describe the variable that differs between the subgroups that are being compared.

Explanation: To evaluate cross-cultural validity you compare data from two (or more) subgroups that have similar demographic and disease characteristics, except that they differ on one specific characteristic (i.e., the group variable). Here, you describe the variable(s) for which you will test measurement invariance or DIF (e.g., language, country, sex, age, etc.).

Essential elements

• Describe variable that is used to make subgroups on which measurement invariance is being tested.

Example 1

"For measurement invariance, we conducted multigroup CFA (MGCFA) between groups (on-menses and off-menses women) and sociodemographic region of Brazil (Southeast, North East, South, North and Midwest) in total sample (…)." [30]

CCV2 - Rationale for approach

Provide a rationale for the approach (e.g., multi-group confirmatory factor analysis, or logistic regression analysis of DIF using Theta or sum scores, or other) used.

Explanation: Various approaches, such as MGCFA and DIF based on logistic regression analyses, can be used to assess the cross-cultural validity/measurement invariance of a PROM. We recommend that you provide a rationale for the approach used to assess this measurement property. When you provide a rationale for the approach used to analyze data for cross-cultural validity\measurement invariance, you allow the readers of your article to understand the methodological choices made in ensuring the PROM's applicability across different cultural groups.

Essential elements

• Rationale for the approach used.

Example 1

"The influence of gender and age on the factor structures within the models was examined by multiple-group confirmatory factor analysis (CFA) invariance evaluation (ref) to determine whether the factor structure remained consistent across different groups. This analysis enabled the assessment of whether the relationships between the observed indicators and the latent factors are comparable and invariant across various groups." [95]

Example 2

"We used an iterative hybrid approach of ordinal logistic regression (OLR) and item-response theory (IRT) as implemented in the lordif R-package [ref]. This procedure was used to maintain high comparability with other studies that investigated differential item functioning (DIF) in PROMIS® Pain interference (PROMIS PI) items." [96]

CCV3 - Statistical analyses

Multiple-group confirmatory factor analysis (MGCFA)

Provide a description of the tested models, method of estimation, type of correlation matrix, and methods and criteria used to flag items for measurement invariance.

Differential Item Functioning (DIF) analyses

Provide the statistical approach used to calculate the patient's score (e.g., IRT/Rasch based theta score, or (un)weighted sum score), a description of the tested regression models, and criteria used to flag items for DIF.

Other approaches

Provide details of the methods if other approaches were used.

Explanation: If MGCFA was used for evaluating cross-cultural validity\measurement invariance, we recommend that you provide the method of estimation (see SV2), type of correlation matrix used (see SV2), and information on the tested models (in terms of internal structure of the PROM). MGCFA compares models with increasing levels of constraints to determine whether the PROM functions similarly across groups. We recommend that you describe the type of measurement invariance tested (i.e., configural, metric, scalar and residual invariance). In addition, we recommend that you provide information on the methods (e.g., Root Mean Square Error of Approximation - RMSEA, Comparative Fit Index - CFI, Tucker-Lewis Index - TLI, etc.), and criteria for relevant change in model fit (e.g., magnitude of change in goodness-of-fit indices). For regression-based DIF analyses, we recommend that you describe how the patient's score (used as a determinant in the regression model) was calculated (e.g., theta score (and the type of IRT/Rasch model used (see SV2)), or (un)weighted sum score), which subsequent regression models were tested and compared (both for uniform and non-uniform DIF), and the criteria used to flag items for DIF. If other approaches were used, you should provide details of the methods used.

Essential elements

For MGCFA:

- Tested model (e.g., number of factors, which items included in which factor).
- Method of estimation.
- Type of correlation matrix.
- Methods and criteria for relevant change in model fit.

For DIF with IRT/regression analysis:

- Type of IRT/Rasch analysis (e.g., partial credit model, graded response model etc.).
- Description of the tested regression model, if applicable.
- Type of DIF.
- Criteria used to flag items for DIF.

Example 1

"The other main objective of this study is the testing of Measurement invariance (MI) across different cultural backgrounds. For MI testing, the two samples were tested in one model using a multigroup confirmatory factor analysis (MGCFA) for the CPC-12R. The procedure of the MGCFA was inspired by Rudnev et al. (ref) as well as by Chen et al. (ref). The MGCFA was conducted using MLR [Maximum likelihood with robust standard errors] as the estimator. As a prerequisite for the MGCFA on second-order level, the levels of invariance (viz., configural, metric, scalar, and residual invariance) were tested for the first-order factors aforehand. To test the MI, the criteria proposed by Cheung and Rensvold (ref) and those from Chen (ref) were used. Cheung and Rensvold declare that a ΔCFI larger than −0.01 should lead to the null hypothesis of invariance being rejected. (…) In this study, for MI testing with an adequate sample size (total N > 300), unequal sample sizes and mixed lack of invariance, the following cut-offs proposed by Chen (ref) were applied: For testing loading invariance, a change of ≥ −0.01 in the CFI, in addition with a change of ≥ 0.015 in RMSEA, or a change of ≥ 0.03 in SRMR indicates non-invariance. When testing intercept or residual invariance, a change of ≥ −0.01 in CFI, supplemented by a change of ≥ 0.015 in RMSEA, or a change of ≥ 0.01 in SRMR indicates non-invariance." [97]

Example 2

"DIF for gender, age (median split), language (Swedish translated vs US original pediatric PROMIS item banks of pain and fatigue) [ref], were calculated for each item on each scale using the IRT Likelihood Ratio DIF approach [ref], using LR $\chi 2$ item fit statistics, as implemented in the software R package mirt [ref]. The Benjamini-Hochberg procedure [ref] was used to control for multiplicity of comparisons in DIF (see Table 2). McFadden's R^2 was used to evaluate when DIF was detected (> 2%) [ref]. McFadden's R^2 could be interpreted as < 0.035 = negligible DIF, 0.035-0.07 = moderate DIF, and > 0.07 = large DIF [ref]. The level of the effect size was evaluated tabular and graphically using methods outlined by Steinberg and Thissen [ref] for items with significant DIF." [90]

Cross-Cultural Validity\Measurement Invariance: Results

CCV4 - Statistical analyses

Provide results on changes or differences in model fit between tested models (e.g., MGCFA, IRT/Rasch analyses, logistic regression DIF analyses, or other approaches).

Explanation: If MGCFA was used, we recommend that you report model fit indices (e.g., CFI, TLI, RMSEA) per model (i.e., configural, metric, scalar, or residual models), and specifically, the changes in the model fit between models (e.g., ΔCFI, ΔRMSEA). If modifications (e.g., covariance between items) are made to achieve measurement invariance, we recommend that you report the modification indices and justifications for any changes. If regression analyses were conducted to test for DIF, we recommend that you report DIF statistics for each item evaluated (e.g., Mantel-Haenszel index, logistic regression coefficients, or item parameters in IRT models).

Essential elements

• Results for calculated statistics.

Example 1

"The resulting fit indices of the test for configural invariance, metric invariance and scalar invariance are displayed in Table 2. Indices of the second-order MGCFA, that should be emphasized, are the CFI = 0.97; TLI = 0.97; RMSEA = 0.05; 90% CI RMSEA = (0.04, 0.06) and SRMR = 0.05, regarding the metric invariance. Furthermore, the CFI = 0.97; TLI = 0.97; RMSEA = 0.05; 90% CI RMSEA = (0.04, 0.07) and SRMR = 0.06."

Table 2. Confirmatory factor analyses and measurement invariance results

Model	χ^2 (df)	Sig.	CFI	TLI	RMSEA [CI]	SRMR
CFA German Sample	81.69 (50)	0.003	0.97	0.96	0.07 [0.46, 0.83]	0.05
CFA US sample	160.72 (100)	<0.001	0.98	0.97	0.05 [0.03, 0.64]	0.04
Configural invariance	160.72 (100)	<0.001	0.98	0.97	0.05 [0.04, 0.06]	0.04
Metric invariance	184.70 (111)	<0.001	0.97	0.97	0.05 [0.04, 0.07]	0.05
Scalar invariance	199.47 (118)	<0.001	0.97	0.97	0.05 [0.04, 0.07]	0.06

χ^2 refers to the Chi-square difference value with respective degrees of freedom (df). Sig. is used to display the p-value of the Chi-square difference test. The confirmatory fit index is reported as CFI, the Tucker-Lewis index as TLI. RMSEA is the robust Root Mean Square Error of Approximation with respective 90%-confidence intervals [CI]. The SRMR is the Root Mean Square Residual. CFA means confirmatory factor analysis. MGCFA means multigroup confirmatory factor analysis.
[97]

Example 2

"Table 2 shows the results of the DIF analyses. Two items in the anxiety item bank, 'It scared me when I felt nervous' (EDANX03) and 'I felt worried' (EDANX30), showed uniform and non-uniform DIF, respectively. The item 'I felt worried' is present in the PROMIS anxiety 7a short form. The items are present in respectively 1 and 3% of the CAT-based assessments. In the depression item bank, two items showed uniform DIF: 'I felt worthless' (EDDEP04) and 'I felt unhappy' (EDDEP36). (...) For the item 'It scared me when I felt nervous', the threshold parameters for the Dutch population were mostly slightly lower than the thresholds for the US population, indicating that the Dutch population endorses higher response categories at the same level of anxiety. (...) Fig 1 illustrates the impact of DIF on respondents total scores. The plots on the left show the impact of DIF when all items are considered, whereas the plots on the right show the impact of DIF when only DIF items are considered. The plots show that DIF had a minimal impact on the total score when all items are administered in each item bank."

Table 2. McFadden's pseudo R^2 and IRT parameters for items displaying DIF

Item bank	Item with DIF	DIF type	McFadden's pseudo R^2	Slope; and threshold parameters	Included in CAT[d]
Anxiety	EDANX03: It scared me when I felt nervous	Uniform	R^2_{12} = 0.021 R^2_{23} = 0.011	**NL:** 2.62; 0.15, 0.97, 2.02 US: 3.74; 0.59, 1.18, 1.95	1%
	EDANX30: I felt worried[a]	Non-uniform	R^2_{12} = 0.010 R^2_{23} = 0.033	NL: 2.16; -1.12, -0.10, 1.29, 2.64 US: 3.14; -0.57, 0.24, 1.22, 2.12	3%
Depression	EDDEP04: I felt worthless[b]	Uniform	R^2_{12} = 0.024 R^2_{23} = 0.013	**NL:** 2.93; -0.17, 0.58, 1.56, 2.61 US: 4.37; 0.29, 0.88, 1.61, 2.36	8%
	EDDEP36: I felt unhappy[c]	Uniform	R^2_{12} = 0.037 R^2_{23} = 0.001	NL: 4.21; -0.14, 0.61, 1.33, 2.20 **US:** 3.44; -0.64, 0.23, 1.20, 2.17	100%

The bold population had lower thresholds compared to the other population, indicating that this population endorses higher item response categories at the same level of the domain (anxiety, depression)
[a] present in the anxiety 7a short form
[b] present in the depression 4a, 6a, 8a and 8b short form
[c] present in the depression 6a, 8a and 8b short form
[d] Based on 4047 CAT-based assessments for anxiety and 4293 CAT-based assessments for depression

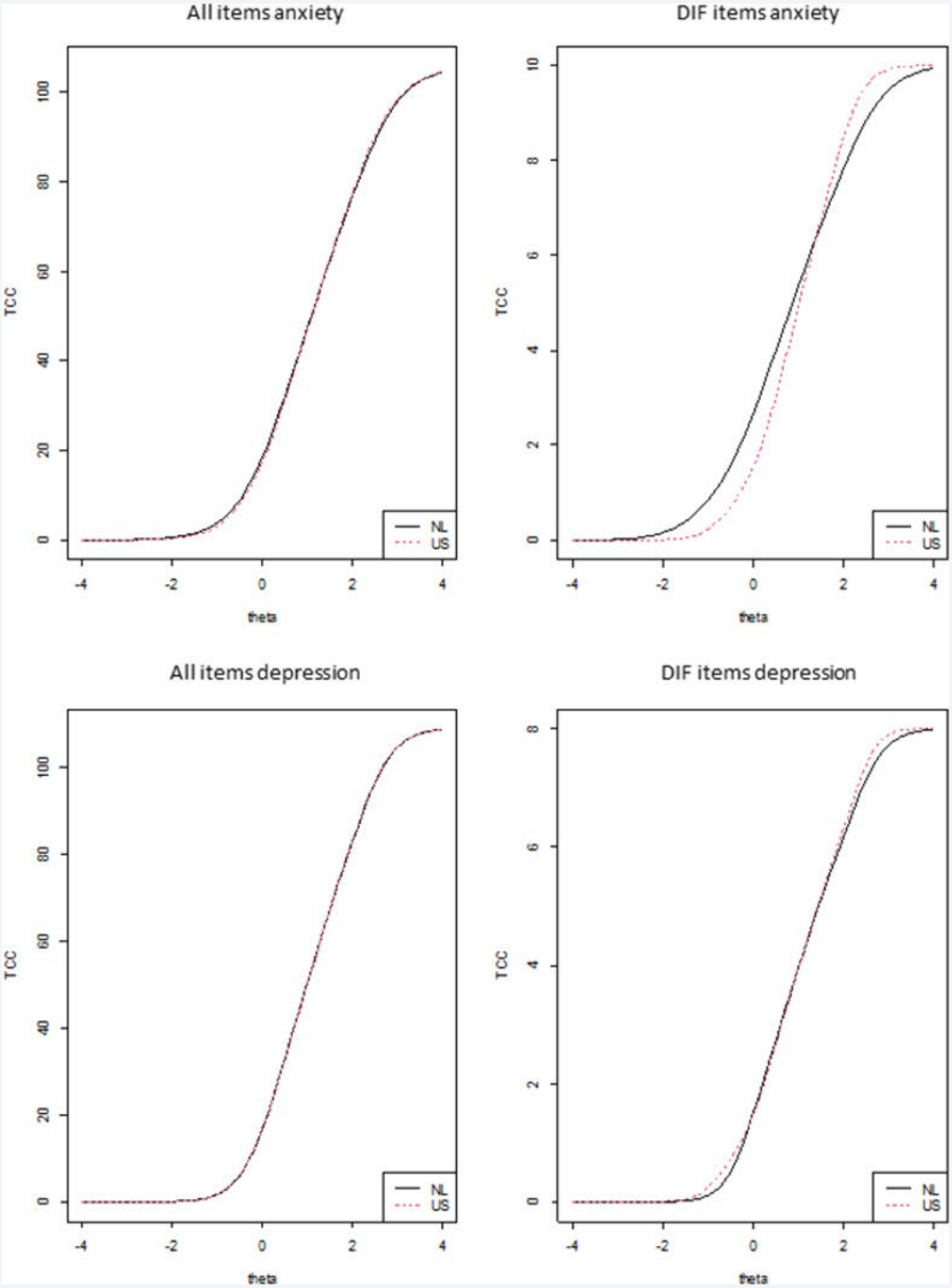

Figure 1. Total impact of DIF on the test characteristic curve (TCC) for anxiety and depression.
[98]

Reliability

Reliability refers to the proportion of the total variance in the measurements which is due to 'true' differences between patients. 'True' is the average score that would be obtained if the scale was administered an infinite number of times to the same person [99]. Reliability can also be explained as the ability of a PROM to distinguish between patients [99]. Within a homogeneous group, it is hard to distinguish between patients [100]. An important assumption made in a reliability study (and in a study on measurement error) is that patients are stable on the construct to be measured between the repeated measurements.

Reliability: Methods

R1 - Stability of patients

Provide arguments for assuming stability of patients on the construct of interest between the repeated measurements.

Explanation: Reliability is evaluated by repeated measurements in stable patients. What "stable patients" are depends on the construct to be measured and the target population. You should provide arguments for assuming stability of patients on the construct of interest between the repeated measurements to ensure the reliability of the repeated measurements is not compromised by changes on the construct to be measured. Data may also be collected on intervention(s) related to the construct being measured in the interim time (e.g., surgery, rehabilitation, psychological treatment, etc.). If an intervention is given in the interim period, it can be assumed that (many of) the patients have changed on the construct to be measured. For example, if you were measuring the impact of low back pain on quality of life, your patients would not undergo rehabilitation for pain relief; or if you were to measure body esteem, psychological therapies could influence the stability of your patients.

Essential elements

• Arguments for assuming stability of patients on the construct of interest between the repeated measurements.

Example 1

"Patients were excluded from the retest when they experienced a change in disease activity from active to quiescent disease or vice versa according to the Harvey-Bradshaw Index/Simple Clinical Colitis Activity Index (HBI/SCCAI)." [101]

Example 2

"The patients completed the questionnaire at two time points. (…) A long interval between test administrations increases the risk of change in patients' health status in a test-retest study, whereas a short interval increases the risk of recall bias [ref]. (…) Patients' mental health was assumed to be stable during the time period from test 1 to test 2, since the health status of epilepsy patients is not likely to change over a period of 2 weeks." [102]

Example 3

"For the retest evaluation, a link was returned with the Female Genital Self-image Scale (FGSIS) and the following question: "Did you undergo treatment (surgery/physical therapy/medication on genitals) between the first and second assessments of this research?". The answer options were "yes" or "no." Only women who answered "no" to this question were included in the test-retest reliability." [41]

R2 - Statistical analyses

Provide the specific intraclass correlation coefficient (ICC) or kappa formula used and justify the choice of the formula used.

Explanation: The preferred reliability statistic depends on the unit of measurement of the PROM. For continuous scores, an intraclass correlation coefficient (ICC) is preferred to evaluate reliability. For ordinal scores, a weighted kappa is preferred, for which the specific weighting scheme should be described [103]. For a dichotomous/nominal score, an unweighted kappa is preferred. For the ICC the specific model should be described, such as a two-way random effects model for agreement ($ICC_{agreement}$ or $ICC_{(2.1)}$) [104] which takes all measurement error into account, including the influence of the source of variation of interest (e.g., a systematic change between time points in a test-retest design) [99]. Another model could be a two-way mixed effects model for consistency ($ICC_{consistency}$ or $ICC_{(3.1)}$) that does not consider a systematic change between the repeated measurements (e.g., the average difference between interviewers in an inter-interviewer reliability design). If this model is used, you should justify why you are not interested in the systematic difference (or change) of the source of variation that you varied across the repeated measurements. A third model that could be used is an one-way random effects model ($ICC_{(1.1)}$) that takes all errors into account [99], but does not investigate the specific influence of another source of variation (e.g. occasions or interviewers). More complex models are described elsewhere [99]. Pearson or Spearman correlation coefficient without evidence whether systematic change over time had occurred may overestimate reliability and are not recommended because the reliability may be overestimated when a systematic change was not taken into account.

Essential elements

• Description of the model or weighting scheme for the specific intraclass correlation coefficient (ICC) or kappa used.

• Rationale/justification for the choice of the formula used.

Example 1

"To investigate test-retest reliability of the Arm Function in Multiple Sclerosis Questionnaire (AMSQ), we calculated one-way ICC for the whole sample due to an incomplete design [ref] (research question 1). (...) In addition, we investigated whether there was a systematic difference between two measurements due to differences in mode of administration (research question 2), and whether there was a systematic difference between two measurements due to different observers (research question 3). These two questions were investigated by calculating ICC two-way ANOVA random effect models for agreement for patients who completed baseline and retest questionnaires in different ways of administration (i.e., paper at baseline and online at follow-up), and for patients who were interviewed.[ref]" [105]

Example 2

"Test-retest reliability of single ordinal items was assessed using weighted kappa statistic with squared weights and corresponding 95% CI estimated using the bootstrap method (1000 replications)." [102]

R3 - Source of variation

Specify which source of variation was purposely varied across the repeated measurements.

Explanation: Each measurement should be conducted under the same measurement conditions, except for the condition (also called a source of variation) of interest that was intentionally varied across the repeated measurements. Sources of variation can be occasion, moments of the day, different interviewers, different settings or environments, different modes of administration, different software, etc. You should specify which source of variation is of your main interest, which you purposely varied across the repeated measurements in your design. For example, if the occasion was varied, you could indicate this as 'test-retest design'. It allows the readers of your article to understand the design used in your study.

Essential elements

• Which source of variation was purposely varied.

General example for test-retest design

"We evaluated the test-retest reliability of the PROM with patients completing the PROM at baseline and 7 days later."

General example for inter-interviewer reliability study

"Two different interviewers administered the PROM to the same patients to assess inter-rater reliability."

General example for different environments

"We used a reliability design with repeated measurements where patients completed the PROM one time at home and one time at the hospital."

Example 1

"Test-retest reliability and measurement error were assessed for women who did not modify complaints of dysmenorrhea by the Global Rating of Change Scale between 7 to 10 days." [106]

R4 - Time interval

Provide arguments for the appropriateness of the time interval.

Explanation: The time interval between the repeated measurements should be long enough to prevent recall bias, and short enough to ensure that patients have not been changed on the construct to be measured. You should provide arguments for the appropriateness of the time interval to ensure the reliability of the repeated measurements is not compromised by the time interval. What an appropriate time interval is, depends on the construct to be measured and the target population. A time interval of about 2 weeks is often considered appropriate for the evaluation of PROMs [5]. In general, arguments for an appropriate time interval can be based on previous studies or expert experience with the construct being measured. The time interval itself has been described at item GM6.

Essential elements

- The time interval between repeated measurements.
- Arguments for the appropriateness of the time interval.

Example 1

"We set two weeks as the interval between the assessments based on previous research [ref], which also mentioned the clinical status is unlikely to change appreciably in the absence of intervention during this time." [107]

Reliability: Results

R5 - Stability of patients

Describe whether patients were stable on the construct of interest between the repeated measurements.

Explanation: Descriptions of patient stability are essential to ensure the credibility to the results in a study on reliability of the repeated measurements. In the results section, you should provide evidence whether your patients indeed were stable on the construct of interest between the repeated measurements. In addition, you should provide the number of stable patients and reasons for unstable patients. The number of stable patients is necessary to verify that it is consistent with the sample size required to ensure statistical power for your study. Reasons for instability may include any intervention related to the construct being measured in the interim time (e.g., surgery, rehabilitation, psychological treatment, etc.).

Essential elements

- Evidence whether patients were stable on the construct of interest between the repeated measurements.
- Number of stable patients.
- Reason for non-instability of patients.

Example 1

"In the total sample, 355 (57.82%) women returned the questionnaires between 14 and 20 days for retest, and 22 (6.20%) women were excluded for having undergone treatment on the genitals. Thus, test-retest reliability analysis was performed with 333 (93.80%) women." [41]

R6 - Statistical analyses

Provide results for calculated statistics (ICC or kappa) and measures of precision (standard errors or confidence intervals). Provide results of all variance components, or contingency tables.

Explanation: In addition to providing the reliability parameters of the PROM under study, all variance components should be reported for ICC values, as well as their 95% confidence interval. This facilitates an understanding of the impact of different sources of variation on the observed score. It also facilitates the comparison and future pooling of results in systematic reviews and allows the readers of your article to understand the precision of the ICC estimates [99]. For kappa values, we recommend reporting contingency tables, in addition to the kappa value itself.

Essential elements

- Results for calculated statistics (ICC or kappa).
- Results of all variance components, or contingency tables.
- Measures of precision: 95% confidence intervals.

Example 1

"The $ICC_{agreement}$ was 0.79 (95% CI 0.70-0.84) indicating sufficient test-retest reliability [for the Patient-Reported Outcomes Measurement Information System (PROMIS) short form designed to measure physical function in geriatric rehabilitation patients (PROMIS-PF-GR)]." [108]

Measurement Error

Measurement error refers to the systematic and random error of an individual patient's score that is not attributed to true changes in the construct to be measured. It refers to the precision of the score [99]. Measurement error (sometimes also called agreement) is evaluated by repeated measurements in stable patients. The same data collected to evaluate reliability can be used.

Measurement Error: Methods

ME1 - Stability of patients

Provide arguments for assuming stability of patients on the construct of interest between the repeated measurements.

Explanation: Similar to reliability, measurement error is evaluated by repeated measurements in stable patients (see R1). You should provide arguments for assuming stability of patients on the construct of interest between the repeated measurements.

Essential elements

• Arguments for assuming stability of patients on the construct of interest between the repeated measurements.

Example 1

"Patients were excluded from the retest when they experienced a change in disease activity from active to quiescent disease or vice versa according to the Harvey-Bradshaw Index/Simple Clinical Colitis Activity Index (HBI/SCCAI)." [101]

Example 2

"The patients completed the questionnaire at two time points. (…) A long interval between test administrations increases the risk of change in patients' health status in a test-retest study, whereas a short interval increases the risk of recall bias [ref]. (…) Patients' mental health was assumed to be stable during the time period from test 1 to test 2, since the health status of epilepsy patients is not likely to change over a period of 2 weeks." [102]

Example 3

"For the retest evaluation, a link was returned with the Female Genital Self-image Scale (FGSIS) and the following question: "Did you undergo treatment (surgery/physical therapy/medication on genitals) between the first and second assessments of this research?". The answer options were "yes" or "no." Only women who answered "no" to this question were included in the test-retest reliability." [41]

ME2 - Statistical analyses

Provide and justify the specific statistical formula used for standard error of measurement (SEM), smallest detectable change (SDC), limits of agreement (LoA), or percentage (specific) agreement.

Explanation: The preferred measurement error statistic depends on the unit of measurement of the PROM. For continuous scores, the standard error of measurement (SEM) for a single score, and the smallest detectable change (SDC) - also referred to as minimal detectable change (MDC) - for a change score are preferred. Alternatively, the limits of agreement (LoA) can be calculated. For dichotomous, nominal or ordinal scores, we recommend that you use percentage (specific) agreement to calculate the measurement error of the PROM. You should also specify the measurement error formula used for the statistic used. Similar to the ICC models (see R2), the specific SEM model should be described (e.g., $SEM_{agreement}$) and its formula [99]. Note that the SDC derived from the LoA are equal to the $SDC_{consistency}$. In addition, if measures of precision (e.g., 95% confidence intervals) were calculated, we recommend that you describe which methods or techniques were used to calculate these.

Essential elements

- The standard error of measurement (SEM), smallest detectable change (SDC), limits of agreement (LoA), or percentage (specific) agreement formula used.
- Rationale for the choice of the formulas used.

Example 1

"For measurement error, standard error of the measurement ($SEM_{agreement}$), smallest detectable change (SDC) at the individual level, and Bland and Altman graph were used. $SEM_{agreement}$ was estimated by the formula [$_{difference}SD/\sqrt{2}$], in which $_{difference}SD$ was the standard deviation (SD) of the difference between the test and retest score of the Female Genital Self-image Scale (FGSIS). SDC was estimated by [$SEM*1.96*\sqrt{2}$]. Bland and Altman graph was estimated by limits of agreement (LoA) using the formula [$d- \pm(1.96*_{difference}SD)$], in which d- is the mean of the differences between the test and retest of the FGSIS." [41]

ME3 - Source of variation

Specify which source of variation was purposely varied across the repeated measurements.

Explanation: Similar to the reliability, each measurement should be conducted under the same measurement conditions, except for the condition (also called a source of variation) of interest that was intentionally varied across the repeated measurements (see R3). We recommend that you specify which source of variation you are most interested in, and which you purposely varied across the repeated measurements in your design.

Essential elements

- Which source of variation was purposely varied.

General example for test-retest design

"We evaluated the measurement error of the PROM with patients completing the PROM at baseline and 7 days later."

General example for inter-interviewer measurement error study

"Two different interviewers administered the PROM to the same patients to assess inter-rater reliability."

General example for different environments

"We used a repeated measurements design where patients completed the PROM one time at home and one time at the hospital."

Example 1

"Test-retest reliability and measurement error were assessed for women who did not modify complaints of dysmenorrhea by the Global Rating of Change Scale between 7 to 10 days." [106]

ME4 - Time interval

Provide arguments on the appropriateness of the time interval.

Explanation: Similar to the reliability, the time interval should be long enough to prevent recall bias, and short enough to ensure that patients have not been changed on the construct to be measured (see R4). We recommend that you provide arguments for the appropriateness of the time interval.

Essential elements

- The time interval between repeated measurements.
- Arguments for the appropriateness of the time interval.

Example 1

"We set two weeks as the interval between the assessments based on previous research [ref], which also mentioned the clinical status is unlikely to change appreciably in the absence of intervention during this time." [107]

Measurement Error: Results

ME5 - Stability of patients

Describe whether patients were stable on the construct of interest between the repeated measurements.

Explanation: Similar to the reliability, descriptions of patient stability are essential to ensure the credibility to the results in a study on measurement error of the repeated measurements (see R5). In the results section, we recommend that you describe whether your patients indeed were stable on the construct of interest between the repeated measurements. In addition, we recommend that you provide the number of stable patients and reasons for unstable patients.

Essential elements

- Description whether patients were stable on the construct of interest between the repeated measurements.
- Number of stable patients.
- Reason for non-instability of patients.

Example 1

"In the total sample, 355 (57.82%) women returned the questionnaires between 14 and 20 days for retest, and 22 (6.20%) women were excluded for having undergone treatment on the genitals. Thus, test-retest reliability analysis was performed with 333 (93.80%) women." [41]

ME6 - Statistical analyses

Provide results for calculated statistics and measures of precision (if applicable). Provide results of all variance components included in the SEM, provide the systematic change or difference between the repeated measurements (when LoA was applied), or contingency tables.

Explanation: In addition to reporting measurement error parameters of the PROM (e.g., SEM, SDC, LoA or percentage (specific) agreement), we recommend reporting all variance components in the SEM and SDC values, as well as the 95% confidence interval of the SEM. This facilitates an understanding of the impact of the different sources of variation on the measurement error. It also facilitates the comparison and pooling of your results in systematic reviews and allows the readers of your article to understand the precision of the SEM estimates [99]. For dichotomous/nominal/ordinal PROM scores, we recommend that you report the percentage (specific) agreement expressed separately for each category of the score (e.g., positive and negative agreement for a dichotomous score), as well as contingency tables.

Essential elements

- Results for calculated statistics (SEM, SDC, and/or LoA, or percentage agreement).
- Results of all variance components, or contingency tables.
- Measures of precision: 95% confidence intervals.

Example 1

"The mean difference (d-) between the test and retest results was −0.285. $SEM_{agreement}$ and SDC at the individual level were 1.469 and 4.071, respectively. Figure 1 shows the Bland and Altman plot with the lower (−4.359) and upper (3.788) limits of agreement (LoA)." [41]

Example 2

"For the measurement error, the percentage of agreement between test and retest was (...) (71.8%), and of disagreement was 28.2%." [44]

Criterion Validity

Criterion validity refers to the degree to which the scores of a PROM are an adequate reflection of a 'gold standard'. A 'gold standard' is a benchmark or criterion that is widely accepted as the best available method or standard against which other methods can be compared or evaluated [109]. A diagnostic gold standard is often not an appropriate gold standard to evaluate criterion validity of a PROM, as the PROM under study often measures another construct than the diagnostic gold standard. Within the first COSMIN study, we reached consensus that there is no gold standard for a PROM (as the patient itself is the gold standard for a PRO). Only, the long version of a shortened PROM can be considered as a 'gold standard' [110]. Often, the comparator instrument is wrongly considered as a gold standard (for example, when the scores of a new instrument is compared to a widely used instrument like the SF-36). In this case, we consider it as hypotheses testing for construct validity, and expected hypotheses about the magnitude and direction of the correlation between (subscales of) the instruments should be formulated and tested (e.g., ≥0.70).

CriV1 - Criterion

Justify the assumption that the comparator instrument is a reasonable gold standard.

Explanation: PROMs do not have 'gold standards' as the patient itself is the 'gold standard'. The only exception is that the long version of a short PROM can be considered the 'gold standard' [1]. In this case, there is an implicit hypothesis that a very high correlation (≥0.90) should be expected. If you consider your study as an evaluation of criterion validity, we recommend that you justify why an instrument is considered to be the 'gold standard' for the construct of interest [111].

Essential elements

• Rationale for assuming that the comparator is an appropriate gold standard.

Example 1

"Criterion validity estimates how well one measure is correlated with a "gold standard". The long-version scores of the questionnaires [Pain Catastrophizing Scale and Tampa Scale for Kinesiophobia] were considered the "gold standard" and correlations with the short-version scores were calculated. Pearson correlation coefficients > 0.7 were expected." [112]

Criterion Validity: Methods

CriV2 - Statistical analyses

Provide and justify the statistics used: correlations when criterion has continuous scores or area under the ROC curve, and sensitivity and specificity when criterion is dichotomous.

Explanation: We recommend that you provide and justify the statistics used to assess criterion validity. When the PROM under study has a continuous score, you can use a Pearson's or Spearman's correlation coefficient when the 'gold standard' is also continuous. When the 'gold standard' is dichotomous, you can calculate the area under the ROC curve [113].

Essential elements

- The statistics used.
- Justify the statistics used.

Example 1

"(...) we used Spearman's correlation coefficient (rho) to correlate the long and short versions, given that the data did not present a normal distribution when analyzed using the Kolmogorov-Smirnov test. Correlation magnitude > 0.70 was considered the appropriate cut-off point for criterion validity [ref]." [14]

Example 2

"For the PROMIS Anxiety Computer Adaptive Test (PROMIS-A-CAT) and PROMIS Anxiety Short Form (PROMIS-A-SF), we calculated sensitivity, specificity, negative predictive value (NPV) and positive predictive (PPV) at the T score cut-points for mild (55) and moderate (65) anxiety proposed by (...). Too few people were classified with severe anxiety to warrant analysis. To examine cut-points we used Youden's Index, a commonly used summary measure of sensitivity and specificity [ref] ranging from zero to one, where one represents perfect performance. Youden's index assumes equal importance of sensitivity and specificity and therefore may not be optimal for anxiety screening programs [ref], which might give higher weight to sensitivity, so that cases are not missed. Nevertheless, we used it to allow comparison with previous studies [ref]." [114]

Criterion Validity: Results

CriV3 - Statistical analyses

Provide results for calculated statistics.

Explanation: In the study of criterion validity, we recommend that you provide results for calculated statistics (see CriV2). When correlation analysis is performed between the PROM score and the 'gold standard', the direction (positive or negative) and magnitude of the Pearson's or Spearman's correlation coefficient should be provided. When performing an area under the ROC curve analysis between the PROM score and the gold standard, it is essential to report the AUC value and its confidence interval (CI), as well as sensitivity (true positive rate) and specificity (true negative rate) at different cutoff points. You can also include the ROC curve graph in your article to illustrate the trade-off between sensitivity and specificity at different cutoff points. These findings improve the understanding of the end-users of your article of how well the instrument performs in comparison to the established 'gold standard'.

Essential elements

- Results for calculated statistics.

Example 1

"Area under the curve (AUC) were in the 'good' range for the PROMIS Anxiety Computer Adaptive Test (0.82, 95%CI = 0.73-0.90) and PROMIS Anxiety Short Form (0.80, 95%CI = 0.71-0.89) (Table2). The AUC for the Generalised Anxiety Disorder-7 (0.84) was also in the good range while the AUC for all other legacy measures fell into the 'useful' range (0.72 to 0.79)."

Table 2. Area under the curve (AUC) for each measure versus diagnosis of any anxiety disorder using the structured clinical interview for DMS disorders (SCID)

Measure	AUC (95% CI)
PROMIS-A-CAT T score	.82 (.73 - .90)
PROMIS-A-SF T score	.80 (.71 - .89)
GAD-7	.84 (.77 - .91)
HADS-A	.79 (.70 - .89)
PSYCH-6	.79 (.70 - .89)
DASS-S	.77 (.68 - .86)
DASS-A	.72 (.62 - .82)
DT	.72 (.62 - .82)

n = 122 with all measures

[114]

Example 2

"Regarding criterion validity, the correlation between the Roland-Morris Disability Questionnaire (RMDQ) with 24 items and 15 items is adequate (rho = 0.954, $p < 0.001$)." [14]

Hypotheses Testing for Construct Validity

Hypotheses testing for construct validity refers to the degree to which the scores of a PROM are consistent with hypotheses based on the assumption that the PROM validly measures the construct(s) to be measured. Many types of hypotheses can be tested to evaluate construct validity of a PROM. In general, these hypotheses concern an expected relationship between the cross-sectional score of the PROM under study, and the cross-sectional score of another well-defined and high-quality comparator instrument(s), or an expected differences in scores between subgroups (also referred to as 'known' groups validity).

Hypotheses Testing for Construct Validity: Methods

ConV1 - Hypotheses

State hypotheses, and provide rationales for each hypothesis.

Explanation: Many types of hypotheses can be tested to evaluate construct validity of a PROM. Hypotheses should include both the direction (positive or negative) and magnitude (absolute or relative) of the expected correlations or the differences in scores [111]. Hypotheses can be formulated around expected correlations between the PROM under study and another measurement instrument, or differences in scores of the PROM between groups. We recommend that you provide all hypotheses including the expected direction and magnitude of correlations between the PROM scores and the scores on comparator instruments; or the expected direction and magnitude of differences in scores of the PROM between groups. For example, if a PROM aims to measure social participation, you may expect a moderate (but not high) correlation with a PROM that aims to measure activity limitations because this is a different construct; or you can expect chronic patients to score an average of 10 points higher on the PROM under study for knee pain intensity than acute patients. Without this specification, it is difficult to decide afterwards whether the results are in accordance with the hypothesis [111]. We also recommend that you explain why you chose your hypotheses, which can be based on the literature (theory and empirical evidence) and/or your (clinical/research) experience [111]. It allows the readers of your article to interpret your results. The more specific the hypotheses are and the more hypotheses are being tested, the more evidence is gathered for construct validity.

Essential elements

• Hypotheses including the expected direction and magnitude of correlations between the PROM scores and the scores on comparator instruments (if applicable).

• Hypotheses including the expected direction and magnitude of differences in scores of the PROM between groups (if applicable).
• Rationales for the hypotheses.

Example 1

"Because Celiac Disease Quality of Life Survey (CD-QOL) items assess social limitations, emotional concerns, and cognitive concerns rather than physical aspects, we hypothesized moderate (r = 0.40–0.69) negative correlations between CD-QOL total and SF-36 social functioning, emotional well-being, and general health subscale scores, and small (r = 0.00–0.39) negative correlations between CD-QOL total and SF-36 physical functioning, role limitations due to physical and emotional problems, energy/fatigue, and bodily pain, such that worse Celiac disease-specific quality of life would be related to worse generic quality of life." [115]

Example 2

"With respect to construct validity, known-group validity was assessed for groups that were expected to differ in score: groups differing in age (three age groups were compared), gender, and chronic diseases (yes/no) were evaluated. The expected direction and magnitude of the differences were based on previous research on other Dutch adults on the same domains [ref]. Furthermore, Pearson correlations between each of the domains and the pain intensity item were calculated. The magnitude and direction of the expected correlation was based on previous knowledge on and experience with the measured constructs. In total, 88 a priori hypotheses were formulated (see Table 6)."

Table 6. Confirmation of priori hypothesis regarding the expected difference between groups and the correlation domains of the PROMIS-89

Hypotheses	Number confirmed
Older participants score better than younger participants on the domains fatigue, sleep disturbance, anxiety and depression; a difference of at least 1 point is expected between each adjacent age group (18-39, 40-64, ≥ 65)[a]	2/8
Younger participants score better than older participants on the domains physical function and pain interference and on pain intensity; a difference of at least 1 point is expected between each adjacent age group (18-39, 40-64, ≥ 65) for physical function and pain interference, and a difference of at least 0.5 point for pain intensity[a]	5/6
The youngest (18-39) and oldest (≥ 65) age group score at least 1 point better than the middle age group (40-64) on the domains ability to participate in social roles and activities[a]	0/2
Males score at least 1 point better than females on all domains, and 0.5 point better on pain intensity[a]	7/8
Participants without chronic diseases score at least 2 points better than people with chronic diseases on all domains, and 1 point better on pain intensity[a]	5/8
The following domains have a correlation between 0.6 and 0.7: physical function and pain interference (negative), physical function and pain intensity (negative), anxiety and depression, sleep disturbance and fatigue	4/8
The domains ability to participate in social roles and activities and physical function have a correlation between 0.4 and 0.6	2/2
The domain pain interference has a correlation of at least 0.6 with pain intensity	2/2
The remaining domains have a correlation of less than 0.5 (depending on the direction, this might be negative)	42/44
Total	69/88 (78%)
Physical function	10/11 (91%)
Ability to participate in social roles and activities	8/11 (73%)
Anxiety	8/11 (73%)
Depression	8/11 (73%)
Fatigue	9/11 (82%)
Sleep disturbance	7/11 (64%)
Pain interference	11/11 (100%)
Pain intensity	8/11 (73%)

[a]Better means higher T-scores for the domains physical function and ability to participate in social roles and activities, and lower T-scores for the domains anxiety, depression, fatigue, sleep disturbance, pain interference; for pain intensity, better means a lower score on the 0-10 numeric scale
[22]

ConV2 - Statistical analyses

Specify all statistical methods used to test the hypotheses.

Explanation: To allow the readers of your article to clearly understand your results, we recommend that you specify the statistical methods used to test your hypotheses. If you compare the score of the PROM under study with other instruments, we recommend that you calculate the correlation between the two instruments [111]; if you compared subgroups, you could provide descriptive information of the two groups (e.g., mean PROM score in both groups).

Essential elements

• Statistical or descriptive methods used to test the hypotheses.

Example 1

"Convergent validity was assessed by computing Spearman's rho correlation coefficients for Celiac Disease Quality of Life Survey (CD-QOL) total and subscale scores and scores on the SF-36 scales, (…). Coefficients r = 0.00-0.39 were considered small, r = 0.40-0.69 were considered moderate, and r = 0.70-1.00 were considered large. Because CD-QOL items assess social limitations, emotional concerns, and cognitive concerns rather than physical aspects, we hypothesized moderate negative correlations between CD-QOL total and SF-36 social functioning, emotional well-being, and general health subscale scores, and small negative correlations between CD-QOL total and SF-36 physical functioning, role limitations due to physical and emotional problems, energy/fatigue, and bodily pain, such that worse Celiac disease-specific quality of life would be related to worse generic quality of life." [115]

Example 2

"To determine construct validity, we assessed the difference in T-scores between groups, testing four hypotheses about the ability of the PROMIS item banks to distinguish between these groups. In line with COSMIN guidelines, construct validity is considered sufficient when ≥75% of the results are in accordance with the hypotheses (...). We expected PROMIS T-scores to be higher (worse) in (a) the sleep-clinic sample compared with the high-school sample, (b) the adolescents with a high risk of sleep problems compared with healthy high school students, and (c) high-school students with health issues compared with healthy high-school students. (...) A mean difference of ≥2 points, with a higher (worse) score in the clinical or health issues sample, was considered clinically relevant (...). The fourth hypothesis was that T-scores would worsen (increase) with more problems (higher response category) on item Sleep20 (I had a sleep problem). This item is included in the original PROMIS Sleep Disturbance item bank, but not in the Sleep Disturbance-23. For the fourth hypothesis we merged the last two response categories ('Quite a bit' and 'Very much') because the separate groups were too small. Differences in T-scores between groups were evaluated using linear regression analysis, with correction for relevant demographic variables based on the results from the analyses of the mean values." [34]

Hypotheses Testing for Construct Validity: Results

ConV3 - Statistical analyses

Provide all results and specify if each result is in accordance with its hypothesis.

Explanation: To help your readers interpret your results according to your hypotheses, we recommend that you provide all results and specify if each result is in accordance with its hypothesis. We also recommend that you report each result in the form of a numeric value for the direction and magnitude of a correlation or mean difference (e.g., r = -0.49, M = 2.55), or in a table (e.g., by highlighting the results that are in accordance with the hypotheses in bold). We recommend that you ignore p-values of correlations and mean change scores because it is not relevant to test whether correlations or mean differences are statistically different from zero. Rather, validity is concerned with whether the direction and magnitude of a correlation or mean difference are similar to what would be expected based on the construct(s) being measured.

Essential elements

- All results of analyses for construct validity.
- Indicate whether each result is in accordance with its hypothesis.

Example 1

"(…) we hypothesized moderate negative correlations between CD-QOL [Celiac Disease Quality of Life Survey] total and SF-36 social functioning, emotional well-being, and general health subscale scores, and small negative correlations between CD-QOL total and SF-36 physical functioning, role limitations due to physical and emotional problems, energy/fatigue, and bodily pain, such that worse CeD-specific quality of life would be related to worse generic quality of life. (…) correlations between CD-QOL total and SF-36 social functioning, emotional well-being, and general health and scores were moderate (r = 0.40-0.69) and correlations between CD-QOL total and SF-36 physical functioning, role limitations, and bodily pain were small (r = 0.00-0.39), as expected (…)" [115]

Example 2

"For both the Sleep Disturbance-23 and the Sleep-Related Impairment-11 all results were in accordance with the hypotheses (Table 4). The Sleep Disturbance-23 showed differences between the different samples of 2.7 to 10.6 points, and the largest difference was found between the high-school sample and the sleep-clinic sample. Additionally, adolescents who reported having more sleep problems on the single item also had worse T-scores. The Sleep-Related Impairment-11 showed differences between the different healthy and non-healthy samples of 4.0 to 8.6 points, corrected for age and gender. Here also, the largest difference was found between the high-school sample and the sleep-clinic sample. Adolescents who reported having more sleep problems on the single item also had worse T-scores."

Table 4. Hypothesis testing Sleep Disturbance-23 and Sleep-related Impairment-11

	Mean difference in T-score (95% confidence interval)[a]	
We expected that:	**Sleep Disturbance−23**	**Sleep-related Impairment−11[b]**
1. The sleep-clinic sample had higher scores than the high-school students	10.6 (8.1-13.1)	8.6 (5.2-11.9)
2. The adolescents with sleep problems and/or relevant health issues had higher scores than healthy high-school students	4.8 (3.5-6.2)	5.3 (3.5-7.1)
3. The high-school students with relevant health issues had higher scores than healthy high-school students	2.7 (1.2-4.2)	4.0 (2.0-6.0)
4. Adolescents who answered item Sleep20 'I had a sleep problem' with a higher response category, would have higher scores		
'Not at all' versus 'A little bit'	6.1 (5.1-7.1)	4.5 (3.0-6.0)
'A little bit' versus 'Somewhat'	2.2 (0.9-3.4)	3.9 (1.7-6.1)
'Somewhat' versus 'Quite a bit/very much'	5.2 (4.0-6.5)	3.2 (0.6-5.7)

[a]A mean difference of ≥2 points was considered clinically relevant.

[b]Corrected for age and gender.

[34]

Responsiveness

Responsiveness refers to the ability of a PROM to detect change over time in the construct to be measured. Although responsiveness is considered to be a separate measurement property from validity, the only difference between (construct and criterion) validity and responsiveness is that validity refers to the validity of a single score (cross-sectional), and responsiveness refers to the validity of a change score (longitudinal validity) [116]. Therefore, the standards for responsiveness are similar to the standards for construct and criterion validity. The criterion approach for responsiveness concerns a comparison of the change score of the PROM and the change score of the gold standards. Three types of hypotheses can be distinguished for using the construct approach to assess the responsiveness of an instrument: (1) hypotheses about the expected direction and magnitude of correlations between change scores on the PROM under study and change scores on comparator instruments; (2) hypotheses about expected differences in change scores between different subgroups; or (3) hypotheses about the magnitude of change in scores that we can expect after receiving a treatment with known efficacy on the construct of interest [117].

Responsiveness: Methods

Resp1 - Hypotheses

State hypotheses and provide rationales for each hypothesis.

Explanation: Similar to construct validity (see ConV1), we recommend that you provide all hypotheses including the expected direction and magnitude of correlations between the PROM change scores and the change scores on comparator instrument(s), or the expected direction and magnitude of differences in change scores of the PROM between groups. In addition, we recommend that you explain the rationales for choosing the hypotheses (see ConV1). Hypotheses about effect sizes (ES), standardized response means (SRM), or area under the ROC Curve (AUC) may also be used, but only if an explicit hypothesis and rationale for the expected magnitude of the ES, SRM or AUC after an intervention with known efficacy, is provided [111]. Without specifying the direction and magnitude of the change score, it is impossible to decide afterwards whether the results are consistent with the hypothesis [111].

Essential elements

• Hypotheses including the expected direction and magnitude of correlations between the PROM change scores and the change scores on comparator instruments (if applicable).

• Hypotheses including the expected direction and magnitude of differences in change scores of the PROM between groups (if applicable).
• Hypotheses including the ES, SRM or AUC.
• Rationales for the hypotheses.

Example 1

"For both the Pediatric Quality of Life Inventory (PedsQL) and the Children's Dermatology Life Quality Index (CDLQI), we formulated 7 hypotheses (Table 1). These hypotheses were based on previous studies assessing responsiveness and methodology guidelines by COSMIN (COnsensus-based Standards for the Selection of Health Measurement INstruments). [ref]

Table 1. Hypotheses for testing responsiveness of the PedsQL and CDLQI total scores. If ≥75% of these hypotheses were confirmed, it was considered responsive to change

PedsQL
1. High positive correlation (>0.5) between PedsQL total change score and the GRC scale
2. High positive correlation (>0.5) between PedsQL total change score and CDLQI total change score
3. Moderate positive correlation (0.3-0.5) between PedsQL total change score and the CDLQI leisure subscale change score
4. Low positive (<0.3) or negative correlation between PedsQL total change score and the CDLQI treatment subscale change score
5. Patients indicating improvement on the GRC scale should have a positive mean change score
6. Patients indicating worsening on the GRC scale should have a negative mean change score
7. The mean change score of patients indicating improvement should be higher than the mean change score of unchanged patients, which in turn should be higher than the mean change score of worsened patients
CDLQI
1. High positive correlation (>0.5) between CDLQI total change score and the GRC scale
2. High positive correlation (>0.5) between CDLQI total change score and PedsQL total change score
3. Moderate positive correlation (0.3-0.5) between CDLQI total change score and the PedsQL physical subscale change score
4. Low positive (<0.3) or negative correlation between CDLQI total change score and the PedsQL school subscale change score
5. Patients indicating improvement on the GRC scale should have a positive mean change score
6. Patients indicating worsening on the GRC scale should have a negative mean change score
7. The mean change score of patients indicating improvement should be higher than the mean change score of unchanged patients, which in turn should be higher than the mean change score of worsened patients

Correlation strength was interpreted as high (≥0.5), moderate (0.3-0.5), and low (<0.3), based on previous studies and guidelines for assessing responsiveness. [ref] Hypotheses 5, 6, and 7 concerned the mean change scores of improved, unchanged, and worsened patients according to the global rating of change (GRC) scale. (...) Patients were categorized as improved or unchanged according to their GRC scale response. With an AUC of ≥0.7, an instrument can be considered responsive. [ref]" [118]

Resp2 - Intervention/Exposure

Provide the intervention given or exposure in the interim period (or report that no intervention was provided).

Explanation: Any interventions or exposures the patients received in the interim period should be described. This allows the readers of your article to identify whether changes in the construct occurred due to that specific intervention or exposure. To describe interventions (e.g., surgery, medicines, physical therapy treatment, psychological therapy, etc.) or exposures (e.g., environmental pollution, traumatic events, natural course of disease, etc.), we recommend that you include all essential information in terms of equipment, supplies, drugs, or software; team members (e.g., surgeon, nurse, physical therapist), including their roles; intervention approach, including number of sessions, intervention progression, instructions for the patient to perform the intervention, etc. [119, 120]. If you used secondary data from another study, you could provide citations to the intervention protocol or primary study for more details of the intervention or exposure. If no intervention or exposures were provided, we recommend that you report this so that your readers can identify that changes in the construct occurred despite no intervention or exposure.

Essential elements

- Intervention given or exposure in the interim period (if any or if not).
- Descriptions of the intervention or exposure.

Example 1

"Participants completed all study assessments both before heart transplant (i.e., "baseline," including any time after the patient was listed for transplant) and after transplant (i.e., "follow-up," 8-12 weeks after surgery, deemed by cardiologist coinvestigators as the minimum time after transplant at which a clinically significant improvement in functioning is typically observed)." [121]

Resp3 - Statistical analyses

Specify all statistical methods used to test the hypotheses.

Explanation: The responsiveness statistic depends on the unit of measurement in accordance with your hypothesis. For continuous scores, correlations or AUCs are preferred to evaluate responsiveness. We recommend that you report expected correlations with direction (positive or negative) and magnitude (absolute or relative) [111] and ignore the p-values of the correlations, and report the expected magnitude of the AUC (see ConV3). If effect sizes or standardized response means are used, their expected magnitudes should be reported. For dichotomous scores, sensitivity and specificity should be used and their expected magnitudes should be specified (see CriV2). Specifying all statistical methods to test each hypothesis allows your readers to interpret the results and replicate your analyses.

Essential elements

- Statistical methods used to test if results are in accordance with the hypothesis.
- For continuous scores, correlations with direction and magnitude.
- For continuous scores, expected magnitude of AUC.
- For dichotomous scores, expected sensitivity and specificity.

Example 1

"Responsiveness was, according to the COSMIN guidelines [ref], assessed by 2 methods: (1) the criterion approach by assessing the area (AUC) under the Receiver Operating Curve (ROC) and (2) the construct approach by hypotheses testing. To assess the criterion approach, the population was dichotomised into an 'improved' group and an 'unchanged' group (...) on the global rating of change scale (GROC) (...). The AUC was calculated as the ability of the Quick Disabilities of the Arm, Shoulder and Hand questionnaire (QuickDASH) and Patient-Specific Functional Scale (PSFS) to discriminate between patients classified as 'improved' and 'unchanged'. An AUC of at least 0.70 was regarded as acceptable responsiveness [ref]. To assess the construct approach, 9 a-priori hypotheses were formulated and tested for both the QuickDASH and PSFS. These hypotheses were (...) described in Table 1. The data were assumed to be normally distributed if there was no or minimal difference between the mean and median value, confirmed by histograms, Q plot and the Shapiro-Wilk test. Pearson correlation coefficient was used if the data were normally distributed, otherwise, a Spearman's rank correlation coefficient was used. A correlation of less than 0.30 indicates a weak correlation, at least 0.30 and less than 0.60 indicates moderate correlation, and a correlation at least 0.60 indicates good correlation [ref]. The standardised response mean (SRM) was calculated by dividing the mean change score by the standard deviation (SD) of the change. The effect size (ES) was calculated by dividing the mean change score by the SD of the baseline scores [ref]. An instrument was considered having acceptable responsiveness, based on the construct approach, if meeting at least 75% of the hypotheses according to the COSMIN guidelines [ref]." [122]

Responsiveness: Results

Resp4 - Statistical analyses

Provide results for calculated statistics, and specify, for the construct approach, if each result is in accordance with its hypothesis.

Explanation: Similarly to ConV3, we recommend that you report your results in relation to the corresponding hypotheses used to assess responsiveness. We recommend that you report each result in the form of a numeric value for the direction and magnitude of a correlation between changes scores; or mean difference of the change scores between groups (see ConV2); or the AUC value and its confidence interval (CI), as well as sensitivity (true positive rate) and specificity (true negative rate) at different cutoff points for criterion validity approach. You can also include the AUC graph to illustrate the trade-off between sensitivity and specificity at different cutoff points (see CriV3).

Essential elements

- Results for calculated statistics on responsiveness analyses.
- Results that are (not) in accordance with its hypothesis (for the construct approach).

Example 1

"Criterion approach of responsiveness. Dichotomisation of the GROC [global rating of change scale] showed that 50 patients (47.6%) improved and 55 patients (52.4%) were stable; 1 patient (0.9%) were excluded in the ROC curves analysis, since he or she had worsened clinical condition. The ROC curves (Fig. 2) were similar for both questionnaires, with an AUC for the Quick Disabilities of the Arm, Shoulder and Hand questionnaire (QuickDASH) of 0.75 (95% CI: 0.66, 0.84) and an AUC for the Patient-Specific Functional Scale (PSFS) of 0.75 (95% CI: 0.65, 0.85). The responsiveness for both questionnaires was therefore considered satisfactory based on the criterion approach.
Construct approach of responsiveness. Responsiveness according to testing 9 a-priori hypotheses (Table 1) were met by both instruments; the QuickDASH met 7 hypotheses (77.8%) and the PSFS met 8 hypotheses (88.9%). The correlations between the QuickDASH/PSFS and comparator instruments are presented in Table 4.

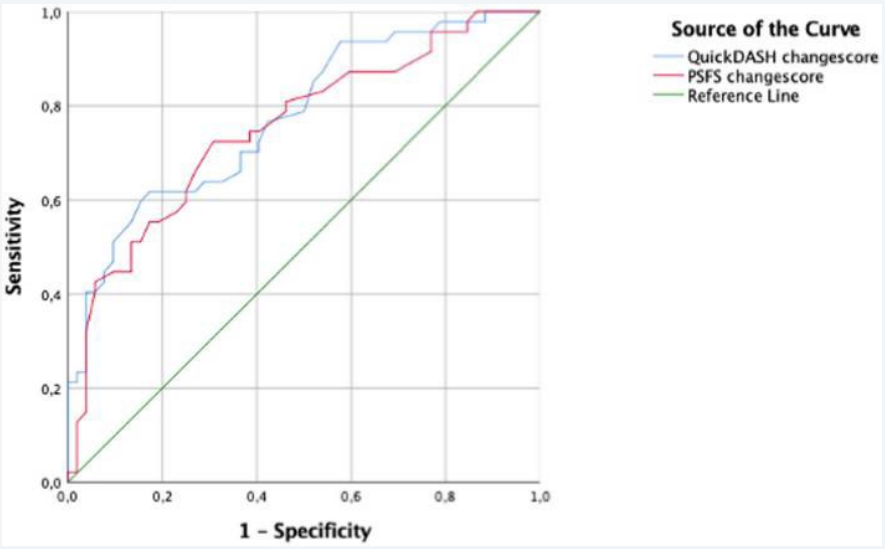

Figure 2. Receiver operating characteristic curves for the QuickDASH and PSFS for 'improved' and 'unchanged' on the GROC. QuickDASH at 3-month follow-up (area under curve = 0.75). PSFS at 3-month follow-up (area under curve = 0.75). Abbreviations: GROC, global rating of change; PSFS, Patient-Specific Functional Scale; QuickDASH, shortened version of the Disabilities of the Arm, Shoulder and Hand questionnaire

Table 4. Correlations among the PROMs' change scores (n = 117)

	QuickDASH[a]	PSFS[a]
QuickDASH (0-100)	-	0.45 (0.28, 0.59)
PSFS (0-10)	0.45 (0.28, 0.59)	-
NRS (0-10)	0.62 (0.49, 0.72)	0.32 (0.14, 0.49)
SQK (0-10)	0.37 (0.20, 0.52)	0.38 (0.20, 0.53)
WA (0-10)	0.44 (0.27, 0.58)	0.46 (0.29, 0.60)
HSCL-25 (0-10)	0.37 (0.20, 0.52)	0.25 (0.06, 0.42)
GROC (1-7)	−0.47 (− 0.61, − 0.31)	−0.50 (− 0.64, − 0.34)

Abbreviations: GROC = global rating of change scale; HSCL-25 = Hopkins Symptom Checklist total score; NRS = numeric pain rating scale; PSFS = Patient-Specific Functional scale; SQK = Single substitute question for kinesiophobia; QuickDASH = Quick Disabilities of the Arm, Hand and Shoulder questionnaire; WA = Workability
n refers to the total sample sizes and may deviate in some of the correlation analysis due to missing data
All correlations were significant at P < 0.01
[a]Values in parentheses are 95% confidence interval
[122]

References

1. 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;63:737–45.

2. Mokkink LB, Prinsen CAC, Bouter LM, Vet HCW de, Terwee CB. The COnsensus-based Standards for the selection of health Measurement INstruments (COSMIN) and how to select an outcome measurement instrument. Brazilian J Phys Ther. 2016;20:105–13.

3. Mokkink LB, Terwee CB, Patrick DL, Alonso J, Stratford PW, Knol DL, et al. The COSMIN checklist for assessing the methodological quality of studies on measurement properties of health status measurement instruments : an international Delphi study. Qual Life Res. 2010;19:539–49.

4. Boers M, Idzerda L, Kirwan JR, Beaton D, Escorpizo R, Boonen A, et al. Toward a Generalized Framework of Core Measurement Areas in Clinical Trials: A Position Paper for OMERACT 11. J Rheumatol. 2014;41:978–85.

5. Streiner DL, Norman GR, Cairney J. Health Measurement Scales. Oxford University Press; 2015.

6. Mokkink LB, Elsman EBM, Terwee CB. COSMIN guideline for systematic reviews of patient-reported outcome measures version 2.0. Qual Life Res. 2024. https://doi.org/10.1007/s11136-024-03761-6.

7. Gagnier JJ, Lai J, Mokkink LB, Terwee CB. COSMIN reporting guideline for studies on measurement properties of patient-reported outcome measures. Qual Life Res. 2021;30:2197–218.

8. FDA. Guidance for industry: patient-reported outcome measures: use in medical product development to support labeling claims: draft guidance. Health Qual Life Outcomes. 2006;4:79.

9. Shadzi MR, Rahmanian M, Heydari A, Salehi A. Structural validity of the Pittsburgh Sleep Quality Index among medical students in Iran. Sci Rep. 2024;14:1538.

10. Itakussu EY, Morita AA, Kakitsuka EE, Kuwahara RM, Anami EHT, Pitta F, et al. The Brazilian-Portuguese version of the Upper Extremity Functional Index (UEFI): Translation, cross-cultural adaptation and measurement properties for Brazilian adults after a burn injury. Burns. 2024;50:219–25.

11. De Vet HCW, Terwee CB, Mokkink LB, Knol DL. Measurement in Medicine - A practical guide. 1st edition. New York: Cambridge University Press; 2011.

12. Rathnayake N, Karunadasa R, Abeygunasekara T, De Zoysa W, Palangasinghe D, Lekamwasam S. Katz index of activities of daily living in assessing functional status of older people: Reliability and validity of Sinhala version. Dialogues Heal. 2023;2:100134.

13. Ivarsson M, Danielsson H, Andersson AK, Gothilander J, Granlund M. Structural validity and internal consistency of the Strengths and Stressors in Parenting (<scp>SSF</scp>) Questionnaire in parents of children with developmental

disabilities. Scand J Psychol. 2023;64:486–94.

14. Frota NT, Fidelis-de-Paula-Gomes CA, Pontes-Silva A, Pinheiro JS, de Jesus SFC, Apahaza GHS, et al. 15-item Roland-Morris Disability Questionnaire (RMDQ-15): structural and criterion validity on patients with chronic low back pain. BMC Musculoskelet Disord. 2022;23:1–8.

15. Terwee CB, van der Willik EM, van Breda F, van Jaarsveld BC, van de Putte M, Jetten IW, et al. Responsiveness and minimal important change of seven PROMIS computerized adaptive tests (CAT) in patients with advanced chronic kidney disease. J Patient-Reported Outcomes. 2023;7:35.

16. Durgam N, Dashputre AA, Moshkovich O, Rezaie A, Martinez N, Enayati P, et al. Content validation of a daily patient-reported outcome measure for assessing symptoms in patients with Small Intestinal Bacterial Overgrowth. Qual Life Res. 2023;32:2573–85.

17. de Arruda GT, de Melo Mantovan SG, Da Roza T, Silva BI da, Tonon da Luz SC, Avila MA. WHODAS measurement properties for women with dysmenorrhea. Health Qual Life Outcomes. 2023;21:55.

18. Lee MK, Basch E, Mitchell SA, Minasian LM, Langlais BT, Thanarajasingam G, et al. Reliability and validity of PRO-CTCAE® daily reporting with a 24-hour recall period. Qual Life Res. 2023;32:2047–58.

19. Feagan BG, Sandborn WJ, Sands BE, Liu Y, Vetter M, Mathias SD, et al. Qualitative and psychometric evaluation of the PROMIS®-Fatigue SF-7a scale to assess fatigue in patients with moderately to severely active inflammatory bowel disease. J Patient-Reported Outcomes. 2023;7:115.

20. Dubinsky MC, Delbecque L, Hunter T, Harding G, Stassek L, Moses RE, et al. Validation of the bowel urgency numeric rating scale in patients with Crohn's disease: results from a mixed methods study. Qual Life Res. 2023;32:3403–15.

21. Oladapo A, Ito D, Rodriguez AM, Philpott S, Krupnick R, Allen V, et al. Psychometric evaluation of a patient-reported outcomes instrument for congenital thrombotic thrombocytopenic purpura. J Patient-Reported Outcomes. 2023;7:68.

22. Elsman EBM, Roorda LD, Smidt N, de Vet HCW, Terwee CB. Measurement properties of the Dutch PROMIS-29 v2.1 profile in people with and without chronic conditions. Qual Life Res. 2022;31:3447–58.

23. Pearson NA, Tutton E, Martindale J, Strickland G, Thompson J, Packham JC, et al. Development of the Warwick Axial Spondyloarthritis faTigue and Energy questionnaire (WASTEd)—a new patient-reported outcome measure. Rheumatol Adv Pract. 2022;6.

24. Brod M, Waldman LT, Shu AD, Smith A. Content validation of the SF-36v2® Health Survey Acute for use in hypoparathyroidism. Qual Life Res. 2023;32:1795–806.

25. Pietrabissa G, Castelnuovo G, Semonella M, Mannarini S, Rossi AA. Measuring Motivations to Eat Palatable Foods: Adaptation and Psychometric Properties of the Italian Version of the Palatable Eating Motives Scale (PEMS-IT). Healthcare. 2024;12:574.

26. Wilson IB, Cleary PD. Linking clinical variables with health-related quality of life. A conceptual model of patient outcomes. JAMA. 1995;273:59–65.

27. World Health Organization (WHO). International Classification of functioning, disability and health: ICF. World Health Organization. 2001.

28. Boers M, Beaton DE, Shea BJ, Maxwell LJ, Bartlett SJ, Bingham CO, et al. OMERACT Filter 2.1: Elaboration of the Conceptual Framework for Outcome Measurement in Health Intervention Studies. J Rheumatol. 2019;46:1021–7.

29. de Arruda GT, de Andrade DF, Virtuoso JF. Internal structure and classification of pelvic floor dysfunction distress by PFDI-20 total score. J Patient-Reported Outcomes. 2022;6:51.

30. Mantovan SG de M, de Arruda GT, Da Roza T, da Silva BI, Avila MA, da Luz SCT. Translation, cross-cultural adaptation, and measurement properties of the dysmenorrhea symptom interference (DSI) scale–Brazilian version. Brazilian J Phys Ther. 2024;28:101065.

31. Harrison C, Apon I, Ardouin K, Sidey-Gibbons C, Klassen A, Cano S, et al. The Development, Deployment, and Evaluation of the CLEFT-Q Computerized Adaptive Test: A Multimethods Approach Contributing to Personalized, Person-Centered Health Assessments in Plastic Surgery. J Med Internet Res. 2023;25:e41870.

32. Sivertsen HE, Helvik A-S, Gjøra L, Haugan G. Psychometric validation of the Hospital Anxiety and Depression Scale (HADS) in community-dwelling older adults. BMC Psychiatry. 2023;23:903.

33. Galenkamp H, Stronks K, Mokkink LB, Derks EM. Measurement invariance of the SF-12 among different demographic groups: The HELIUS study. PLoS One. 2018;13:e0203483.

34. van Kooten JAMC, Terwee CB, Luijten MAJ, Steur LMH, Pillen S, Wolters NGJ, et al. Psychometric properties of the Patient-Reported Outcomes Measurement Information System (PROMIS) Sleep Disturbance and Sleep-Related Impairment item banks in adolescents. J Sleep Res. 2021;30.

35. Schougaard LMV, Laurberg T, Lomborg K, Hansen TK, Hjollund NH, Jensen AL. Test–retest reliability and measurement error of the WHO-5 Well-being Index and the Problem Areas in Diabetes questionnaire (PAID) used in telehealth among patients with type 1 diabetes. J Patient-Reported Outcomes. 2022;6:99.

36. Karpecki PM, Findley A, Sloesen BJ, Hodson N, Bentley S, Arbuckle R, et al. Qualitative Research to Understand the Patient Experience and Evaluate Content Validity of the Chronic Ocular Pain Questionnaire (COP-Q). Ophthalmol Ther. 2024;13:615–33.

37. Quintana DT, Casanova MP, Cady AC, Baker RT. Assessing the Structural Validity of the Knee Injury and Osteoarthritis Outcome Score Scale. Healthcare. 2024;12:414.

38. Stathis P, Papadopoulos G. Evaluation and validation of a patient-reported quality-of-life questionnaire for Parkinson's disease. J Patient-Reported Outcomes. 2022;6:17.

39. Fayers PM, Hand DJ. Factor analysis, causal indicators and quality of life. Qual Life Res. 1997;6:139–50.

40. Jarvis CB, MacKenzie SB, Podsakoff PM. A Critical Review of Construct Indicators and Measurement Model Misspecification in Marketing and Consumer Research. J Consum Res. 2003;30:199–218.

41. Arruda GT de, Silva EV da, Somavilla P, Oliveira MCR de, Braz MM. Female Genital Self-image Scale (FGSIS): cut-off point, reliability, and validation of measurement properties in Brazilian women. Fisioter e Pesqui. 2023;30.

42. Becker B, Bracher M, Chauhan D, Rendas-Baum R, Lin X, Raymond K, et al. Development, psychometric evaluation and cognitive debriefing of the rheumatoid arthritis symptom and impact questionnaire (RASIQ). J Patient-Reported Outcomes. 2021;5:129.

43. Schnettler B, Miranda-Zapata E, Lobos G, Lapo M, Grunert KG, Adasme-Berríos C, et al. Cross-cultural measurement invariance in the satisfaction with food-related life scale in older adults from two developing countries. Health Qual Life Outcomes. 2017;15:113.

44. de Sousa AP, de Arruda GT, Pontes-Silva A, de Souza MC, Driusso P, Avila MA. Measurement properties of the Brazilian online version of the Fibromyalgia Rapid Screening Tool (FiRST). Adv Rheumatol. 2022;62:39.

45. Reeve BB, Edwards LJ, Jaeger BC, Hinds PS, Dampier C, Gipson DS, et al. Assessing responsiveness over time of the PROMIS® pediatric symptom and function measures in cancer, nephrotic syndrome, and sickle cell disease. Qual Life Res. 2018;27:249–57.

46. Guest G, Bunce A, Johnson L. How Many Interviews Are Enough? Field methods. 2006;18:59–82.

47. Francis JJ, Johnston M, Robertson C, Glidewell L, Entwistle V, Eccles MP, et al. What is an adequate sample size? Operationalising data saturation for theory-based interview studies. Psychol Health. 2010;25:1229–45.

48. Malterud K, Siersma VD, Guassora AD. Sample Size in Qualitative Interview Studies. Qual Health Res. 2016;26:1753–60.

49. Browne RH. On the use of a pilot sample for sample size determination. Stat Med. 1995;14:1933–40.

50. Johanson GA, Brooks GP. Initial Scale Development: Sample Size for Pilot Studies. Educ Psychol Meas. 2010;70:394–400.

51. Comrey AL, Lee HB. A First Course in Factor Analysis. Psychology Press; 2013.

52. Edelen MO, Reeve BB. Applying item response theory (IRT) modeling to questionnaire development, evaluation, and refinement. Qual Life Res. 2007;16:5–18.

53. Brown T. Confirmatory Factor Analysis for Applied Research. 2nd editio. The Guilford Press; 2015.

54. Mokkink LB, de Vet H, Diemeer S, Eekhout I. Sample size recommendations for studies on reliability and measurement error: an online application based on simulation studies. Heal Serv Outcomes Res Methodol. 2023;23:241–65.

55. Bujang MA, Baharum N. Guidelines of the minimum sample size requirements for Kappa agreement test. Epidemiol Biostat Public Heal. 2022;14.

56. Fleiss JL, Levin B, Paik MC. Statistical Methods for Rates and Proportions. Wiley; 2003.

57. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics. 1977;33:159–74.

58. Ikeuchi K, Nishida S, Karikawa M, Sakamoto C, Mori F, Tanaka M. Development of an Assessment Tool to Measure the Quality of Life Goal Setting for Cancer Survivors: A Content Validity Study. Cureus. 2024. https://doi.org/10.7759/cureus.71272.

59. Clément C, Lvovschi V-E, Verot E, du Sartz de Vigneulles B, Darlington-Bernard A, Bourgeois D, et al. Supporting health education policies: translation, cross-cultural adaptation and validation of a health literacy instrument, in French. Front Public Heal. 2023;11.

60. Li Y, Lyu L, Fan X, Xu L, Li Y, Song R. Reliability, validity and minimal detectable change of the Chinese Version of the Assessment of Physical Activity in Frail Older People (APAFOP-C). BMC Geriatr. 2024;24:582.

61. Bazarganipour F, Taghavi SA, Montazeri A, Ahmadi F, Derakhshideh Z, Asadikalameh Z. Psychometric properties of the Iranian version of the polycystic ovary syndrome quality of life questionnaire for married and unmarried women with PCOS: PCOSQoL-47 and PCOSQoL-42. BMC Psychol. 2024;12:722.

62. Vandenbroucke JP, von Elm E, Altman DG, Gøtzsche PC, Mulrow CD, Pocock SJ, et al. Strengthening the Reporting of Observational Studies in Epidemiology (STROBE): Explanation and elaboration. Int J Surg. 2014;12:1500–24.

63. van der Baan FH, Koldenhof JJ, de Nijs EJ, Echteld MA, Zweers D, Hesselmann GM, et al. Validation of the Dutch version of the Edmonton Symptom Assessment System. Cancer Med. 2020;9:6111–21.

64. van Kooten JAMC, van Litsenburg RRL, Yoder WR, Kaspers GJL, Terwee CB. Validation of the PROMIS Sleep Disturbance and Sleep-Related Impairment item banks in Dutch adolescents. Qual Life Res. 2018;27:1911–20.

65. Dornonville de la Cour FL, Schow T, Andersen TE, Petersen AH, Zornhagen G, Visser-Keizer AC, et al. Measurement Properties of the Dutch Multifactor Fatigue Scale in Early and Late Rehabilitation of Acquired Brain Injury in Denmark. J Clin Med. 2023;12:2587.

66. Zulmiyusrini P, Yamin M, Muhadi M, Kurniawan J, Salim S. The validity and reliability of Indonesian version of atrial fibrillation effect on quality of life (AFEQT) questionnaire for atrial fibrillation patients. J Patient-Reported Outcomes. 2023;7:133.

67. Martínez-Fernández MV, Sarabia-Cobo CM, Sánchez-Labraca N. Cross-cultural adaptation, reliability, validity and responsiveness of the Michigan Hand Outcomes Questionnaire (MHQ-Sp) in Spain. J Orthop Surg Res. 2024;19:256.

68. Mashayekh-Amiri S, Asghari Jafarabadi M, Rashidi F, Mirghafourvand M. Translation and measurement properties of the pelvic floor distress inventory-short form (PFDI-20) in Iranian reproductive age women. BMC Womens Health. 2023;23:333.

69. Stephan A, Stadelmann VA, Preiss S, Impellizzeri FM. Measurement properties of PROMIS short forms for pain and function in patients receiving knee arthroplasty. J Patient-Reported Outcomes. 2023;7:18.

70. Daescu A-MC, Dehelean L, Navolan D-B, Pop GN, Stoian DL. Psychometric properties of the Romanian version of the female sexual function index (FSFI-RO). BMC Womens Health. 2023;23:528.

71. McHorney CA, Bensink ME, Burke LB, Belozeroff V, Gwaltney C. Development and psychometric validation of the Nausea/Vomiting Symptom Assessment patient-

reported outcome (PRO) instrument for adults with secondary hyperparathyroidism. J Patient-Reported Outcomes. 2018;2:6.

72. Gharzai LA, Mierzwa ML, Peipert JD, Kirtane K, Casper K, Yadav P, et al. Monitoring Adverse Effects of Radiation Therapy in Patients With Head and Neck Cancer. JAMA Otolaryngol Neck Surg. 2023. https://doi.org/10.1001/jamaoto.2023.2177.

73. da Silva APAF, Bassi-Dibai D, Moreira BL, Gava AD, Takahasi HY, Salomão LGPB, et al. Cross-cultural adaptation, reliability and validation of the Gillette Functional Assessment Questionnaire (FAQ) into Brazilian Portuguese in patients with cerebral palsy. BMC Pediatr. 2023;23:165.

74. Oud T, Tuijtelaars J, Schenk J, Nollet F, Brehm M-A. Validity and reliability of the Dutch translation of the OPUS' client satisfaction with device module in chronic users of hand orthoses. Health Qual Life Outcomes. 2023;21:93.

75. Walton DM, MacDermid JC. A brief 5-item version of the Neck Disability Index shows good psychometric properties. Health Qual Life Outcomes. 2013;11:108.

76. Schuller W, Terwee CB, Klausch T, Roorda LD, Rohrich DC, Ostelo RW, et al. Psychometric properties of the Dutch-Flemish Patient-Reported Outcomes Measurement Information System Pain Behavior item bank in patients with musculoskeletal complaints. J Pain. 2019;20:1328–37.

77. Coppers A, Möller JC, Marks D. Psychometric properties of the short form of the Stroke Impact Scale in German-speaking stroke survivors. Health Qual Life Outcomes. 2021;19:190.

78. Rimehaug SA, Kaat AJ, Nordvik JE, Klokkerud M, Robinson HS. Psychometric properties of the PROMIS-57 questionnaire, Norwegian version. Qual Life Res. 2022;31:269–80.

79. Tacchino A, Ponzio M, Pedullà L, Podda J, Bragadin MM, Pedrazzoli E, et al. Italian validation of the Arm Function in Multiple Sclerosis Questionnaire (AMSQ). Neurol Sci. 2020;41:3273–81.

80. Tong A, Sainsbury P, Craig J. Consolidated criteria for reporting qualitative research (COREQ): a 32-item checklist for interviews and focus groups. Int J Qual Heal Care. 2007;19:349–57.

81. Mol T, Scholten E, van Bennekom C, Post M. Development of the self-regulation assessment and content validation using cognitive interviews in a multicultural post-rehabilitation population. Front Rehabil Sci. 2023;4.

82. Gries KS, Esser D, Wiklund I. Content Validity of CASA-Q Cough Domains and UCSD-SOBQ for Use in Patients with Idiopathic Pulmonary Fibrosis. Glob J Health Sci. 2013;5.

83. Kitchen H, Gandhi K, Carmichael C, Wyrwich KW, Lukic T, Al-Zubeidi T, et al. A Qualitative Study to Develop and Evaluate the Content Validity of the Vitiligo Patient Priority Outcome (ViPPO) Measures. Dermatol Ther (Heidelb). 2022;12:1907–24.

84. Kramer JM, Beasley JB, Caoili A, Kalb L, Urquilla MP, Klein AE, et al. Development and content validity of the Person Experiences Interview Survey (PEIS): a measure of the mental health services experiences of people with developmental disabilities. Front Psychiatry. 2023;14.

85. Fayers PM, Hand DJ, Bjordal K, Groenvold M. Causal indicators in quality of life research. Qual Life Res. 1997;6:393–406.

86. Streiner DL. Being Inconsistent About Consistency: When Coefficient Alpha Does and Doesn't Matter. J Pers Assess. 2003;80:217–22.

87. Floyd FJ, Widaman KF. Factor analysis in the development and refinement of clinical assessment instruments. Psychol Assess. 1995;7:286–99.

88. Skogestad IJ, Kottorp A, Larsson P, Moen TM, Gay CL, Borge CR, et al. Development and evaluation of the Norwegian Fatigue Characteristics and Interference Measure (FCIM) for stroke survivors: cognitive interviews and Rasch analysis. Qual Life Res. 2023;32:3389–401.

89. Harrison CJ, Plessen CY, Liegl G, Rodrigues JN, Sabah SA, Beard DJ, et al. Item response theory assumptions were adequately met by the Oxford hip and knee scores. J Clin Epidemiol. 2023;158:166–76.

90. Carlberg Rindestig F, Wiberg M, Chaplin JE, Henje E, Dennhag I. Psychometrics of three Swedish physical pediatric item banks from the Patient-Reported Outcomes Measurement Information System (PROMIS)®: pain interference, fatigue, and physical activity. J Patient-Reported Outcomes. 2021;5:105.

91. Cortina JM. What is coefficient alpha? An examination of theory and applications. J Appl Psychol. 1993;78:98–104.

92. Cronbach LJ. Coefficient Alpha and the Internal Structure of Tests. Psychometrika. 1951;16:297–334.

93. Streiner DL. Figuring Out Factors: The Use and Misuse of Factor Analysis. Can J Psychiatry. 1994;39:135–40.

94. van Stiphout L, Rolfes J, Waardenburg S, Kimman M, Guinand N, Pérez Fornos A, et al. Construct validity and reliability of the Bilateral Vestibulopathy Questionnaire (BVQ). Front Neurol. 2023;14.

95. Sandseter EBH, Sando OJ, Kleppe R, Lorås H, Storli L. Assessment of psychometric properties and measurement invariance of the sensation seeking scale for children in a Norwegian sample. Front Psychol. 2024;15.

96. Obbarius A, Klapproth CP, Liegl G, Christmann PM, Schneider U, Fischer F, et al. Measuring PROMIS pain interference in German patients with chronic conditions: calibration, validation, and cross-cultural use of item parameters. Qual Life Res. 2023;32:2839–52.

97. Lorenz T, Hagitte L, Prasath PR. Validation of the revised Compound PsyCap Scale (CPC-12R) and its measurement invariance across the US and Germany. Front Psychol. 2022;13.

98. Elsman EBM, Flens G, de Beurs E, Roorda LD, Terwee CB. Towards standardization of measuring anxiety and depression: Differential item functioning for language and Dutch reference values of PROMIS item banks. PLoS One. 2022;17:e0273287.

99. Mokkink LB, Eekhout I, Boers M, van der Vleuten CP, de Vet HC. Studies on Reliability and Measurement Error of Measurements in Medicine – From Design to Statistics Explained for Medical Researchers. Patient Relat Outcome Meas. 2023;Volume 14:193–212.

100. de Vet HCW, Terwee CB, Mokkink LB, Knol DL. Measurement in Medicine. Cambridge University Press; 2011.

101. Stoker AMH, Gruters A, Loon MCM van der E, Postulart D, Czuber-Dochan W, Gilissen LPL. Translation, validation and psychometric properties of the Dutch version of the Inflammatory Bowel Disease-Fatigue (IBD-F) self-assessment scale. J Patient-Reported Outcomes. 2023;7:108.

102. Schougaard LMV, de Thurah A, Bech P, Hjollund NH, Christiansen DH. Test-retest reliability and measurement error of the Danish WHO-5 Well-being Index in outpatients with epilepsy. Health Qual Life Outcomes. 2018;16:175.

103. Vach W. The dependence of Cohen's kappa on the prevalence does not matter. J Clin Epidemiol. 2005;58:655–61.

104. Shrout PE, Fleiss JL. Intraclass correlations: Uses in assessing rater reliability. Psychol Bull. 1979;86:420–8.

105. van Leeuwen LM, Mokkink LB, Kamm CP, de Groot V, van den Berg P, Ostelo RWJG, et al. Measurement properties of the Arm Function in Multiple Sclerosis Questionnaire (AMSQ): a study based on Classical Test Theory. Disabil Rehabil. 2017;39:2097–104.

106. Avila MA, Arruda GT de, Godoy AG de, Driusso P. Translation, cross-cultural adaptation to Brazilian Portuguese and measurement properties of the WaLIDD score. Rev Bras Ginecol e Obs / RBGO Gynecol Obstet. 2024;46.

107. Liu S, Liang J, Xu N, Mai S, Wang Q, Zeng L, et al. Reliability and validity of simplified Chinese version of the Italian spine youth quality of life questionnaire in adolescents with idiopathic scoliosis. BMC Musculoskelet Disord. 2021;22:568.

108. Smit EB, Bouwstra H, Roorda LD, van der Wouden J (Hans) C, Wattel E (Lizette) M, Hertogh CMPM, et al. A Patient-Reported Outcomes Measurement Information System Short Form for Measuring Physical Function During Geriatric Rehabilitation: Test-Retest Reliability, Construct Validity, Responsiveness, and Interpretability. J Am Med Dir Assoc. 2021;22:1627-1632.e1.

109. Streiner DL, Norman GR. Health Measurement Scales. Oxford University Press; 2008.

110. Versi E. "Gold standard" is an appropriate term. BMJ. 1992;305:187–187.

111. Prinsen CAC, Bouter LBMLM, Patrick JADL. COSMIN guideline for systematic reviews of patient-reported outcome measures. Qual Life Res. 2018;27:1147–57.

112. Le Carré J, Luthi F, Burrus C, Konzelmann M, Vuistiner P, Léger B, et al. Development and Validation of Short Forms of the Pain Catastrophizing Scale (F-PCS-5) and Tampa Scale for Kinesiophobia (F-TSK-6) in Musculoskeletal Chronic Pain Patients. J Pain Res. 2023;Volume 16:153–67.

113. Moore D, McCabe G, Craig B. Introduction to the Practice of Statistics. 2016.

114. Clover K, Lambert SD, Oldmeadow C, Britton B, Mitchell AJ, Carter G, et al. Convergent and criterion validity of PROMIS anxiety measures relative to six legacy measures and a structured diagnostic interview for anxiety in cancer patients. J Patient-Reported Outcomes. 2022;6:80.

115. Dochat C, Afari N, Arigo D. Psychometric validation of the celiac disease-specific quality of life survey (CD-QOL) in adults with celiac disease in the United States. Qual

Life Res. 2023;32:2195–208.

116. Mokkink LB, Terwee CB, Knol DL, Stratford PW, Alonso J, Patrick DL, et al. The COSMIN checklist for evaluating the methodological quality of studies on measurement properties: A clarification of its content. BMC Med Res Methodol. 2010;10:22.

117. Mokkink L, Terwee C, de Vet H. Key concepts in clinical epidemiology: Responsiveness, the longitudinal aspect of validity. J Clin Epidemiol. 2021;140:159–62.

118. Lokhorst MM, Horbach SER, Waner M, O TM, van der Vleuten CJM, Spuls PI, et al. Responsiveness of quality of life measures in children with peripheral vascular malformations: The OVAMA project. JPRAS Open. 2021;27:70–9.

119. Zhang K, Wu J, Su Z, Ma Y, Shi Q, Barchi LC, et al. The SUPER reporting guideline suggested for reporting of surgical technique: explanation and elaboration. Gland Surg. 2023;12:749–66.

120. Mallinson T, Kozlowski AJ, Johnston M V., Weaver J, Terhorst L, Grampurohit N, et al. Rasch Reporting Guideline for Rehabilitation Research (RULER): the RULER Statement. Arch Phys Med Rehabil. 2022;103:1477–86.

121. Flynn KE, Dew MA, Lin L, Fawzy M, Graham FL, Hahn EA, et al. Reliability and construct validity of PROMIS® measures for patients with heart failure who undergo heart transplant. Qual Life Res. 2015;24:2591–9.

122. Rysstad T, Grotle M, Klokk LP, Tveter AT. Responsiveness and minimal important change of the QuickDASH and PSFS when used among patients with shoulder pain. BMC Musculoskelet Disord. 2020;21:328.