COSMIN Study Design checklist for Patient-reported outcome measurement instruments

Version July 2019

Lidwine B Mokkink
Cecilia AC Prinsen
Donald L Patrick
Jordi Alonso
Lex M Bouter
Henrica CW de Vet
Caroline B Terwee

Contact
L.B. Mokkink, PhD
Department of Epidemiology and Biostatistics
Amsterdam Public Health research institute
Amsterdam University Medical Centers, location VUmc
P.O. box 7057
1007 MB Amsterdam
The Netherlands
Website: www.cosmin.nl
E-mail: w.mokkink@amsterdamumc.nl

Table of Content

List of abbreviations

CTT:	classical test theory
IRT/Rasch:	Item Response Theory and Rasch analyses
NA:	not applicable
Original CC:	original COSMIN checklist [1]
PROM:	patient-reported outcome measure
RoB:	Risk of Bias; it refers to the COSMIN Risk of Bias checklist [2]

Instructions

The COSMIN Study Design checklist is recommended for designing studies to evaluate measurement properties of existing patient-reported outcome measures (PROMs). It can be used by researchers and clinicians or other professionals who are designing a study to evaluate measurement properties of an existing PROM, or by e.g. scientific committees and medical ethics committees who are appraising protocols of studies on measurement properties, or by reviewers for scientific journals that will publish study protocols of studies on measurement properties of PROMs.

The COSMIN Study Design checklist is based on the original version of the COSMIN checklist [1] [3], as well as on the recently developed COSMIN Risk of Bias checklist for PROMs [2]. Decisions on adaptations were made based on iterative discussions by the COSMIN steering committee, both at face-to-face meetings (LM, CP, HdV and CT) and by email discussions (entire COSMIN steering committee, i.e. all authors).

The COSMIN Study Design checklist consists of ten boxes. The first box, i.e. General recommendations for designing a study on a measurement properties, is relevant for all studies. It contains general standards which should be considered in the design of a study on any measurement property. The remaining boxes contain standards for specific studies on each of the nine measurement properties, i.e. content validity, structural validity, internal consistency, cross-cultural validity\measurement invariance, reliability, measurement error, criterion validity, hypotheses testing for construct validity, and responsiveness [2 4]. In addition, we provide standards for translating an existing PROM in the box Translation process.

In this checklist, each standard is also accompanied with a 4-point rating scale. This rating scale is based on the COSMIN Risk of Bias checklist [2]. The 4-point rating scale is added for illustrative purposes to better understand the consequences of choices made in the design of a study for the methodological quality of the study; it is not intended to be used to actually provide an overall rating (i.e. based on the worst-score counts principle) for your study design. The purpose of this checklist is only to check whether all important issues are considered when designing a study on measurement properties. Details on how to design and analyse these studies are described in the book Measurement in Medicine [5]. A clarification and explanation of most of the individual standards can be found in the user manuals (www.cosmin.nl) accompanying the COSMIN Risk of Bias checklist [6 7]. References, e.g. for sample size requirements, are also provided in the COSMIN user manuals [6 7].

Standards refer to potential risk of bias issues, reporting issues, or sample size issues. In this document for each standard a justification is added, referring to the number of the box (number between brackets refer to number of the specific standard in the specific box) from the COSMIN Risk of Bias checklist [2] (RoB), or the original COSMIN checklist [1] (CC); or by indicating that the standard is about sample size or it is a newly added standard.

General recommendation for the design of a study on measurement properties

The box General recommendations for designing a study on measurement properties is relevant for all studies on measurement properties. The aim of a study evaluating a measurement property of a PROM is to investigate (one or more aspects of) the quality of the PROM at issue. These studies require a clear research aim (i.e. referring to the measurement properties of interest), a clear description of the PROM and a clear description of the study population. The quality of a PROM should be determined in the target population in which the PROM will be used, because the results of studies on measurement properties depend on the sample included in the study.

General recommendations for the design of a study on measurement properties					
Research aim	**very good**	**adequate**	**doubtful**	**inadequate**	**justification**
1 Provide a clear research aim, including (1) the name and version of the PROM, (2) the target population, and (3) the measurement properties of interest	Research aim clearly described			Research aim not clearly described	New
PROM					
2 Provide a clear description of the construct to be measured	Construct clearly described			Construct not clearly described	RoB Box 1
3 Provide a clear description of the development process of the PROM, including a description of the target population for which the PROM was developed	Development process clearly described		Development process clearly described		RoB Box 1
4 The origin of the construct should be clear: provide a theory, conceptual framework (i.e. reflective or formative model) or disease model used or clear rationale to define the construct to be measured	Origin of the construct clear		Origin of the construct not clear		RoB Box 1

5	Provide a clear description of the structure of the PROM (i.e. the number of items and subscales included in the PROM, instructions given and response options) and its scoring algorithm	Structure and scoring algorithm clearly described			Structure and scoring algorithm not clearly described	RoB Box 1
6	Provide a clear description of existing evidence on the quality of the PROM	Existing evidence on the quality of the PROM clearly described		Existing evidence on the quality of the PROM not clearly described		New
7	Provide a clear description of the context of use*	Context of use clearly described		Context of use not clearly described		RoB Box 1
Target population						
8	Provide a clear description of in- and exclusion criteria to select patients, e.g. in terms of disease condition and characteristics like age, gender, language or country, and setting (e.g. general population, primary care or hospital/rehabilitation care)	In- and exclusion criteria for patients clearly described			In- and exclusion criteria for patients not clearly described	Characteristic of study population [6]
9	Provide a clear description of the method used to select the patients for the study (e.g. convenience, consecutive, or random)	Method for patient selection clearly described		Method of patient selection not clearly described		New
10	Describe whether the selected sample is representing the target population in which the PROM will be used in terms of age, gender, important disease characteristics (e.g. severity, status, duration)	Study sample representing the target population clearly described	Assumable that the study sample is representing the target population, but not clearly described	Unclear whether the study sample is representing the target population	Study will not be performed in a sample representing the target population	RoB Box 1

* The context of use refers to the intended application of the PROM (e.g. for research or clinical practice), to a specific setting for which the PROM was developed (e.g. for use in a hospital or at home) or to a specific administration mode (e.g. paper or computer-administered). If the PROM was developed for use across multiple contexts, this should be described.

Content Validity

Content validity of existing PROMs can be assessed by asking patients and professionals about the relevance, comprehensiveness and comprehensibility of the items, response options, and instructions. In the box on content validity, standards are given for studies on content validity in which patients are involved, as well as studies in which professionals are involved.

Content validity	**very good**	**adequate**	**doubtful**	**inadequate**	**NA**	**Justification**
Design requirements						
1 From the perspective of the patients: use an appropriate method for assessing (1) the relevance of each item for the patients' experience with the condition, **AND** (2) the comprehensiveness of the PROM, **AND** (3) the comprehensibility of the PROM instructions, items, response options, and recall period	Widely recognized or well justified method for qualitative research will be used to assess the three aspects	Only quantitative (survey) method(s) will be used or assumable that the method used will be appropriate but not clearly described, but all three aspects will be assessed	Not clear if patients will be asked whether each item is relevant AND comprehensible AND whether items together are comprehensive, or doubtful whether the method will be appropriate	Method used are not appropriate or patients will not be asked about the relevance, comprehensiveness or comprehensibility of all items		RoB Box 2 (1, 8, 15)
2 From the perspective of professionals: use an appropriate method for assessing (1) the relevance of each item for the construct of interest, **AND** (2) the comprehensiveness of the PROM	Widely recognized or well justified method for qualitative research will be used to assess the two aspects	Only quantitative (survey) method(s) will be used or assumable that the method used will be appropriate but not clearly described, but both aspects will be assessed	Not clear if professionals will be asked whether each item is relevant AND items together are comprehensive, or doubtful whether the method will be appropriate	Method used are not appropriate or professionals will not be asked about the relevance or comprehensiveness of all items	Not applicable	RoB Box 2 (22, 27)

3	Include professionals from all relevant disciplines	Professionals from all required disciplines will be included	Assumable that professionals from all required disciplines will be included, but not clearly described	Doubtful whether professionals from all required disciplines will be included or relevant professionals will not be included		Not applicable	RoB Box 2 (28, 23)
4	Evaluate each item in an appropriate number of patients or professionals For qualitative studies For quantitative (survey) studies	 ≥7 ≥50	 4-6 30 - 49	 <4 or not clear <30 or not clear			RoB Box 2 (2, 9, 16, 24, 29)
5	Use skilled group moderators or interviewers	Skilled group moderators or interviewers will be used	Group moderators or interviewers have limited experience or will be trained specifically for the study	Not clear if group moderators or interviewers will be trained or group moderators /interviewers are not trained and have no experience		Not applicable	RoB Box 2 (3, 10, 17)
6	Base the group meetings or interviews on an appropriate topic or interview guide	Appropriate topic or interview guide will be used	Assumable that the topic or interview guide will be appropriate, but not clearly described	Not clear if a topic guide will be used or doubtful if topic or interview guide will be appropriate or no guide used		Not applicable	RoB Box 2 (4, 11, 18)
7	Record and transcribe verbatim the group meetings or interviews	All group meetings or interviews will be recorded and transcribed verbatim	Assumable that all group meetings or interviews will be recorded and transcribed verbatim, but not clearly described	Not clear if all group meetings or interviews will be recorded and transcribed verbatim or recordings will not be transcribed verbatim or only notes will be made during the group meetings/ interviews	No recording and no notes	Not applicable	RoB Box 2 (5, 12, 19)

Analyses						
8 Use an appropriate approach to analyse the data	A widely recognized or well justified approach will be used	Assumable that the approach will be appropriate, but not clearly described	Not clear what approach will be used or doubtful whether the approach will be appropriate	Approach not appropriate		RoB Box 2 (6, 13, 20, 25, 30)
9 Involve at least two researchers in the analysis	At least two researchers will be involved in the analysis	Assumable that at least two researchers will be involved in the analysis, but not clearly described	Not clear if two researchers will be included in the analysis or only one researcher will be involved in the analysis		Not applicable	RoB Box 2 (7, 14, 21, 26, 31)

Structural validity

A PROM can be based on a reflective or on a formative model [8-10]. A reflective model is a model in which all items are a manifestation of the same underlying construct. These items are called effect indicators and are expected to be highly correlated and interchangeable. In a formative model – its counterpart – the items together form the construct. These items do not need to be correlated. It should be described in the protocol whether the PROM is based on a reflective or a formative model. Structural validity is only relevant for PROMs that are based on a reflective model.

When the aim of the study is to assess the structural validity of a multidimensional PROM, a factor analysis should be performed on the whole scale. However, when the aim is to additionally assess unidimensionality of subscales, a factor analysis could also be performed on each subscale separately.

Structural validity	**very good**	**adequate**	**Doubtful**	**inadequate**	**NA**	**Justification**
Statistical methods						
1 For CTT: perform confirmatory factor analysis	Confirmatory factor analysis will be performed	Exploratory common factor analysis will be performed		No exploratory or confirmatory factor analysis will be performed	Not applicable	RoB Box 3 (1)
2 For CTT: provide clear information on how the factor analysis will be performed, e.g. software program, method of estimation, whether and how assumptions will be checked, rotation method, criteria for model fit.	Clear information on the performance of the analysis is provided		Clear information on some of the aspects for performing the analysis is provided	Unclear how the analysis will be performed	Not applicable	Original CC
3 For IRT/Rasch: choose a model that fits to the research question	Chosen model fits well to the research question	Assumable that the chosen model fits well to the research question	Doubtful if the chosen model fits well to the research question	Chosen model does not fit to the research question	Not applicable	RoB Box 3 (2)
4 For IRT/Rasch: provide clear information on how the IRT or Rasch analysis will be performed, e.g. software program, which IRT or Rasch model used, method of estimation, whether and how assumptions will be checked, criteria for model fit.	Clear information on the performance of the analysis is described		Clear information on some of the aspects for performing the analysis is described	Unclear how the analysis will be performed	Not applicable	Original CC

5	Perform the analysis in a sample with an appropriate number of patients (taking into account expected number of missing values)	FA: 7 times number of items and ≥100 Rasch/1PL models: ≥ 200 patients 2PL parametric IRT models OR Mokken scale analysis: ≥ 1000 patients	FA: at least 5 times number of items and ≥100 OR at least 6 times number of items but <100 Rasch/1PL models: 100-199 patients 2PL parametric IRT models OR Mokken scale analysis: 500-999 patients	FA: 5 times number of items but <100 Rasch/1PL models: 50-99 patients 2PL parametric IRT models OR Mokken scale analysis: 250-499 patients	FA: < 5 times number of items Rasch/1PL models: < 50 patients 2PL parametric IRT models OR Mokken scale analysis: < 250 patients	RoB Box 3 (3)
6	Provide a clear description of how missing items will be handled	The way missing items will be handled is clearly described		The way missing items will be handled is not clearly described		Original CC

Internal consistency

Like structural validity, internal consistency is only relevant for PROMs based on a reflective model. Furthermore, internal consistency should be assessed for unidimensional (sub)scales. Therefore, unidimensionality or structural validity using e.g. factor analysis should be assessed for each scale or subscale in the study or evidence for structural validity obtained in a previous study in a sample from a comparable target population should be available.

Internal consistency	**very good**	**adequate**	**doubtful**	**inadequate**	**NA**	**Justification**
Design requirements						
1 Check whether a scale or a subscale is unidimensional	Evidence provided that each scale or subscale is unidimensional		Unclear whether each scale or subscale is unidimensional	the scale or subscale is NOT unidimensional		RoB Box 4 (1)
2 Perform the analysis in a sample with an appropriate number of patients (taking into account expected number of missing values)	≥100 patients	50-99 patients	30-49 patients	<30 patients		Sample size
3 Provide a clear description of how missing items will be handled	The way missing items will be handled is clearly described		The way missing items will be handled is not clearly described			Original CC
Statistical methods						
4 For continuous scores: calculate Cronbach's alpha or Omega for each unidimensional scale or subscale	Cronbach's alpha, or Omega will be calculated		Only item-total correlations will be calculated	No Cronbach's alpha and no item-total correlations will be calculated	Not applicable	RoB Box 4 (2)
5 For dichotomous scores: calculate Cronbach's alpha or KR-20 for each unidimensional scale or subscale	Cronbach's alpha or KR-20 will be calculated		Only item-total correlations will be calculated	No Cronbach's alpha or KR-20 and no item-total correlations will be calculated	Not applicable	RoB Box 4 (3)

6	For IRT-based scores: calculate standard error of theta (SE (θ)) or reliability coefficient of estimated latent trait value (index of (subject or item) separation) for each unidimensional scale or subscale	SE(θ) or reliability coefficient will be calculated		SE(θ) or reliability coefficient will NOT be calculated	Not applicable	RoB Box 4 (4)

Cross-cultural validity\measurement invariance

This measurement property aims to investigate whether items of a PROM behaves similarly in different populations, for example in different ethnicity or language groups, different gender or age groups or different disease populations. Therefore, data is needed from multiple groups (e.g. multiple language groups).

Cross-cultural validity\Measurement invariance	**very good**	**adequate**	**doubtful**	**inadequate**	**NA**	**Justification**
Design requirements						
1 Provide a clear description of the group variable(s), including dichotomization or categorization	The group variable(s) is/are clearly described	Assumable how the group variable will be dichotomized or categorized but not clearly described		The group variable(s) is/are not clearly described		RoB Box 5 (1)
2 Provide a clear description of the relevant characteristics of the patients that should be similar in both subgroups, such as demographic or disease characteristics	Relevant characteristics are clearly described			Relevant characteristics are clearly described		RoB Box 5 (2)
3 Perform the analysis in a sample with an appropriate number of patients (taking into account expected number of missing values)	Regression analyses or IRT/Rasch based analyses: ≥200 patients per group OR MGCFA*: 7 times number of items and ≥100 patients	150-199 patients per group OR 5 times number of items and ≥100 OR 5-7 times number of items but <100 patients	100-149 patients per group OR 5 times number of items but <100 patients	<100 patients per group OR <5 times number of items		RoB Box 5 (3)

Statistical methods						
4 For CTT: perform a multi-group confirmatory factor analysis (MGCFA)	MGCFA will be performed			No confirmatory factor analysis will be performed	Not applicable	RoB Box 5 (2)
5 For CTT: provide clear information on how the factor analysis will be performed, e.g. software program, method of estimation, criteria for model fit, and whether and how assumptions were checked	Clear information on the performance of the analysis is provided		Clear information on some of the aspects for performing the analysis is provided	Unclear how the analysis will be performed	Not applicable	RoB Box 5 (2)
6 For IRT/Rasch: perform differential item functioning (DIF) analyses	DIF will be performed			DIF will not be performed	Not applicable	RoB Box 5 (2)
7 For IRT/Rasch: provide clear information on how the IRT or Rasch analysis will be performed, e.g. software program, which IRT or Rasch model used, method of estimation, criteria for model fit, whether and how assumptions were checked	Clear information on the performance of the analysis is provided		Clear information on some of the aspects for performing the analysis is provided	Unclear how the analysis will be performed	Not applicable	RoB Box 5 (2)
8 Provide a clear description of how missing items will be handled	The way missing items will be handled is clearly described		The way missing items will be handled is not clearly described			Original CC

Measurement error and reliability

Measurement error and reliability can be calculated based on the same study design and data collection. Basically, two measurements are needed in a group of people who are all assumed to be stable on the construct to be measured. As the design and the data collected can be used for both measurement properties, we present the standards in one box. Only the statistical parameters are different. We strongly encourage researchers who use such a design to report measurement error in addition to a reliability parameter.

Measurement error and reliability	**very good**	**adequate**	**doubtful**	**inadequate**	**NA**	**Justification**
Design requirements						
1 Use at least two measurements	At least two measurements			Only one measurement		Original CC
2 Ensure that the administrations will be independent	Independent measurements	Assumable that the measurements will be independent	Doubtful whether the measurements will be independent	measurements NOT independent		New
3 Ensure that the patients will be stable in the interim period on the construct to be measured	Patients will be stable (evidence provided)	Assumable that patients will be stable	Unclear if patients will be stable	Patients will NOT be stable		RoB Box 6/7 (1)
4 Use an appropriate time interval between the two measurements, which is long enough to prevent recall, and short enough to ensure that patients remain stable	Time interval appropriate		Doubtful whether time interval is appropriate or time interval is not stated	Time interval NOT appropriate		RoB Box 6/7 (2)
5 Ensure that the test conditions will be similar for the measurements (e.g. type of administration, environment, instructions)	Test conditions similar (evidence provided)	Assumable that test conditions similar	Unclear if test conditions will be similar	Test conditions will NOT be similar		RoB Box 6/7 (3)
6 Perform the analysis in a sample with an appropriate number of patients (taking into account expected number of missing values)	≥100 patients	50-99 patients	30-49 patients	<30 patients		Sample size

Statistical methods for measurement error							
7	For continuous scores: calculate the Standard Error of Measurement (SEM), Smallest Detectable Change (SDC) or Limits of Agreement (LoA)	SEM, SDC, or LoA will be calculated, and model or formula is clearly described*	SEM or SDC will be calculated, but model or formula of the SEM or SDC is not described or not optimal**		SEM will be calculated based on Cronbach's alpha, or on SD from another population	Not applicable	RoB Box 7 (4)
8	For dichotomous/nominal/ordinal scores: calculate the percentage (positive and negative) agreement	% positive and negative agreement will be calculated	% agreement will be calculated		% agreement will not be calculated	Not applicable	RoB Box 7 (5)
9	Provide a clear description of how missing items will be handled	The way missing items will be handled is clearly described		The way missing items will be handled is not clearly described			Original CC
Statistical methods for reliability							
7	For continuous scores: calculate an intraclass correlation coefficient (ICC)	ICC will be calculated, and model or formula of the ICC is clearly described*	ICC will be calculated, but model or formula of the ICC not described or not optimal**	Pearson or Spearman correlation coefficient will be calculated	No ICC or Pearson or Spearman correlations calculated	Not applicable	RoB Box 6 (4)
8	For dichotomous/nominal/ordinal scores: calculate kappa	Kappa will be calculated			No kappa will be calculated	Not applicable	RoB Box 6 (5)
9	For ordinal scores: calculate a weighted kappa	Weighted Kappa calculated and weighting scheme is described		Unweighted Kappa will be calculated or not described if Kappa will be weighted		Not applicable	RoB Box 6 (6, 7)
10	Provide a clear description of how missing items will be handled	The way missing items will be handled is clearly described		The way missing items will be handled is not clearly described			Original CC

* The model (i.e. one-way random effect model or two-way random or mixed effect model), type (i.e. for single or multiple measurement) and definition (i.e. for consistency or absolute agreement) of the ICC that will be calculated is appropriately chosen and described (see [11]); ** ICC formula does not correspond to the research question

Criterion validity

As PROMs measure constructs that can only be reported by the patients themselves, no gold standards exists for these measures. The only exception is the long version when investigating a short version of the same PROM.

Criterion validity	**very good**	**adequate**	**doubtful**	**inadequate**	**NA**	**Justification**
Design requirements						
1 Describe whether the proposed criterion can be considered as a reasonable 'gold standard'	Criterion can be considered an adequate 'gold standard' (evidence provided)	No evidence provided, but assumable that the criterion can be considered an adequate 'gold standard'	Unclear whether the criterion can be considered an adequate 'gold standard'	Criterion can NOT be considered an adequate 'gold standard'		Original CC
2 Perform the analysis in a sample with an appropriate number of patients (taking into account expected number of missing values)	≥ 50 patients in the smallest group	30-50 patients in the smallest group	<30 patients in biggest group			Sample size
3 Use an appropriate time schedule for assessments of the PROM of interest and 'gold standard'	PROM and gold standard will be administered at the same time	PROM and gold standard not administered at the same time, but assumable that patient will not change in the interim period	PROM and gold standard will not be administered at the same time, but unclear if patients will changed	PROM and gold standard will not be administered at the same time, and patients are expected to change		New
Statistical methods						
4 For continuous scores: calculate correlations, or the area under the receiver operating curve	Correlations or AUC will be calculated			Correlations or AUC will NOT be calculated	Not applicable	RoB Box 8 (1)

5	For dichotomous scores: determine sensitivity and specificity	Sensitivity and specificity will be determined			Sensitivity and specificity will NOT be determined	Not applicable	RoB Box 8 (2)
6	Provide a clear description of how missing items will be handled	The way missing items will be handled is clearly described		The way missing items will be handled is not clearly described			Original CC

AUC = Area under the Receiver Operator Curve (ROC) curve

Hypotheses testing for construct validity

As no 'gold standards' exist for PROMs, the commonly used way to investigate validity of PROMs is to test hypotheses about 1) expected relationships with other outcomes measures of good quality (Part A), and/or 2) expected differences between relevant groups (Part B). It is of major importance to define hypotheses in advance when assessing construct validity of a PROM, to enable the authors to draw unbiased conclusions after data collection and analyses.

Hypotheses testing for construct validity **A. Comparison with other outcome measurement instruments (convergent validity)**	**very good**	**adequate**	**doubtful**	**inadequate**	**NA**	**Justification**
Design requirements						
1 Formulate hypotheses about expected relationships between the PROM under study and other outcome measurement instrument(s)	Hypotheses formulated including the expected direction and magnitude of the correlations stated		Hypotheses vague or not formulated but possible to deduce what was expected	Unclear what is expected		Original CC
2 Provide a clear description of the construct(s) measured by the comparator instrument(s)	Construct(s) measured by the comparator instrument(s) is/are clearly described			Construct(s) measured by the comparator instrument(s) is/are not clearly described		RoB Box 9a (1)
3 Use comparator instrument(s) with sufficient measurement properties	Sufficient measurement properties of the comparator instrument(s) in a population similar to the study population	Sufficient measurement properties of the comparator instrument(s) but not sure if these apply to the study population	Some information on measurement properties (or a reference to a study on measurement properties) of the comparator instrument(s) in any study population	No information on the measurement properties of the comparator instrument(s), or evidence of insufficient measurement properties of the comparator instrument(s)		RoB Box 9a (2)

4	Perform the analysis in a sample with an appropriate number of patients (taking into account expected number of missing values)	≥100 patients	50-99 patients	30-49 patients	<30 patients	Sample size
5	Use an appropriate time schedule for assessments of the PROM of interest and comparison instruments	PROM and comparison instrument(s) will be administered at the same time	PROM and comparison instrument(s) not administered at the same time, but assumable that patient will not change in the interim period	PROM and comparison instrument(s) will not be administered at the same time, but unclear if patients will changed	PROM and comparison instrument(s) will not be administered at the same time, and patients are expected to change	New
Statistical methods						
6	Use statistical methods that are appropriate for the hypotheses to be tested	Statistical methods will be appropriate	Assumable that statistical methods will be appropriate	Statistical methods will not be optimal	Statistical methods will NOT be appropriate	RoB Box 9a (3)
7	Provide a clear description of how missing items will be handled	The way missing items will be handled is clearly described		The way missing items will be handled is not clearly described		Original CC

B. Comparison between subgroups (discriminative or known-groups validity)	**very good**	**adequate**	**doubtful**	**inadequate**	**NA**	**Justification**
Design requirements						
1 Formulate hypotheses regarding mean differences between subgroups	Hypotheses formulated including the expected directions and magnitude of the mean differences stated		Hypotheses vague or not formulated but possible to deduce what was expected	Unclear what was expected		Original CC
2 Provide an adequate description of important characteristics of the subgroups, such as disease or demographic characteristics	Adequate description of the important characteristics of the subgroups	Adequate description of most of the important characteristics of the subgroups	Poor of no description of the important characteristics of the subgroups			RoB Box 9b (5)
3 Perform the analysis in a sample with an appropriate number of patients (taking into account expected number of missing values)	≥100 patients per group	50-99 patients per group	30-49 patients per group	<30 patients per group		Sample size
Statistical methods						
4 Use statistical methods that are appropriate for the hypotheses to be tested	Statistical methods will be appropriate	Assumable that statistical methods will be appropriate	Statistical methods will not be optimal	Statistical methods will NOT be appropriate		RoB Box 9b (6)
5 Provide a clear description of how missing items will be handled	The way missing items will be handled is clearly described		The way missing items will be handled is not clearly described			Original CC

Responsiveness

Responsiveness is considered to indicate longitudinal validity. When a 'gold standard' is available, the criterion approach Part A of this box can be used. When testing hypotheses about change scores of PROMs compared to other outcome measurement instruments, Part B can be used; for comparison of changes scores of PROMs between subgroups Part C can be used; and when testing hypotheses about expected change scores of PROMs before and after intervention, Part D can be used. It is of major importance to define hypotheses in advance when assessing responsiveness of a PROM, to enable the authors to draw unbiased conclusions after data collection and analyses.

Responsiveness ***A. Criterion approach (i.e. comparison to a 'gold standard')***	**very good**	**adequate**	**doubtful**	**inadequate**	**NA**	**Justification**
Design requirement						
1 The proposed criterion can be considered as a reasonable 'gold standard'	Criterion can be considered an adequate 'gold standard' (evidence provided)	No evidence provided, but assumable that the criterion can be considered an adequate 'gold standard'	Unclear whether the criterion can be considered an adequate 'gold standard'	Criterion can NOT be considered an adequate 'gold standard'		Original CC
2 Use an appropriate time schedule for assessments of the PROM of interest and the gold standard	PROM and gold standard will be administered at the same time at all occasions	PROM and gold standard not administered at the same time, but assumable that patient will not change in the interim period at all occasions	PROM and gold standard will not be administered at the same time, but unclear if patients will changed	PROM and gold standard will not be administered at the same time, and patients are expected to change		New
3 Use an appropriate time interval between first and second measurements	Time interval will be appropriate			Time interval will NOT be appropriate		New

4	Describe anything likely to occur in the interim period (e.g. intervention, in case of progressive disease, other relevant events)	Anything likely to occur during the interim period (e.g. treatment) is adequately described		Unclear or NOT described what will likely to occur during the interim period			Original CC
5	Ensure that a proportion of the patients is likely to change (i.e. improvement or deterioration) on the construct to be measured	Part of the patients is likely to change (evidence provided)	NO evidence provided, but assumable that part of the patients will change	Unclear if part of the patients will change	Patients will likely NOT change		Original CC
6	Perform the analysis in a sample with an appropriate number of patients (taking into account expected number of missing values)	≥50 patients in the smallest group	30-50 patients in the smallest group	<30 patients in biggest group			Sample size
	Statistical methods						
7	For continuous scores: calculate correlations between change scores, or the area under the Receiver Operator Curve (ROC) curve (AUC)	Correlations or AUC will be calculated			Correlations or AUC will NOT be calculated	Not applicable	RoB Box 10a (1)
8	For dichotomous scales: calculate sensitivity and specificity (changed versus not changed)	Sensitivity and specificity will be calculated			Sensitivity and specificity will NOT be calculated	Not applicable	RoB Box 10a (2)
9	Provide a clear description of how missing items will be handled	The way missing items will be handled is clearly described		The way missing items will be handled is not clearly described			Original CC

B. Construct approach (i.e. hypotheses testing; comparison with other outcome measurement instruments)	**very good**	**adequate**	**doubtful**	**inadequate**	**NA**	**Justification**
Design requirements						
1 Formulate hypotheses about expected relationships between the change scores on the PROM under study and (change scores on) other outcome measurement instrument(s)	Hypotheses will be formulated including the expected direction and magnitude of the correlations stated		Hypotheses vague or will not be formulated but possible to deduce what was expected	Unclear what is expected		Original CC
2 Provide a clear description of the construct(s) measured by the comparator instrument(s)	Constructs measured by the comparator instrument(s) is/are clearly described			Constructs measured by the comparator instrument(s) is/are not clearly described		RoB Box 10b (4)
3 Provide information that the measurement properties of the comparator instrument(s) are sufficient	Sufficient measurement properties of the comparator instrument(s) in a population similar to the study population	Sufficient measurement properties of the comparator instrument(s) but not sure if these apply to the study population	Some information on measurement properties (or a reference to a study on measurement properties) of the comparator instrument(s) in any study population	No information on the measurement properties of the comparator instrument(s), or evidence of insufficient measurement properties of the comparator		RoB Box 10b (5)

4	Use an appropriate time schedule for assessments of PROM of interest and comparison instruments	PROM and comparison instrument will be administered at the same time at all occasions	PROM and comparison instrument not administered at the same time, but assumable that patient will not change in the interim period at all occasions	PROM and comparison instrument will not be administered at the same time, but unclear if patients will changed	PROM and comparison instrument will not be administered at the same time, and patients are expected to change	New
5	Use an appropriate time interval between first and second measurements	Time interval appropriate			Time interval NOT appropriate	New
6	Describe anything likely to occur in the interim period (e.g. intervention, other relevant events)	Anything likely to occur during the interim period (e.g. treatment) is adequately described		Unclear or NOT described what will likely to occur during the interim period		Original CC
7	Ensure that a proportion of the patients is likely to change (i.e. improvement or deterioration) on the construct to be measured	Part of the patients is likely to change (evidence provided)	NO evidence provided, but assumable that part of the patients will change	Unclear if part of the patients will change	Patients will likely NOT change	Original CC
8	Perform the analysis in a sample with an appropriate number of patients (taking into account expected number of missing values)	≥100 patients	50-99 patients	30-49 patients	<30 patients	Sample size
Statistical methods						
9	Ensure that the statistical methods are adequate for the hypotheses to be tested	Statistical methods are appropriate	Assumable that statistical methods are appropriate	Statistical methods are not optimal	Statistical methods are NOT appropriate	RoB Box 10b (6)
10	Provide a clear description of how missing items will be handled	The way missing items will be handled is clearly described		The way missing items will be handled is not clearly described		Original CC

C. Construct approach: (i.e. hypotheses testing: comparison between subgroups)	**very good**	**adequate**	**doubtful**	**inadequate**	**NA**	**Justification**
Design requirements						
1 Formulate hypotheses regarding differences between change scores of subgroups a priori (i.e. before data collection)	Hypotheses formulated including the expected differences between change scores stated		Hypotheses vague or not formulated but possible to deduce what was expected	Unclear what was expected		Original CC
2 Provide an adequate description about important characteristics of the subgroups, such as disease or demographic characteristics	Adequate description of the important characteristics of the subgroups	Adequate description of most of the important characteristics of the subgroups	Poor or no description of the important characteristics of the subgroups			RoB Box 10c (8)
3 Use an appropriate time interval between first and second administration of the measurement	Time interval appropriate			Time interval NOT appropriate		New
4 Describe anything likely to occur in the interim period (e.g. intervention, progressive disease, other relevant events)	Anything likely to occur during the interim period (e.g. treatment) is adequately described		Unclear or NOT described what will likely to occur during the interim period			Original CC
5 Ensure that a proportion of the patients is likely to change (i.e. improvement or deterioration) on the construct to be measured	Part of the patients is likely to change (evidence provided)	NO evidence provided, but assumable that part of the patients will change	Unclear if part of the patients will change	Patients will likely NOT change		Original CC
6 Perform the analysis in a sample with an appropriate number of patients (taking into account expected number of missing values)	≥100 patients	50-99 patients	30-49 patients	<30 patients		Sample size

Statistical methods					
7 Ensure that the statistical methods are adequate for the hypotheses to be tested	Statistical methods are appropriate	Assumable that statistical methods are appropriate	Statistical methods are not optimal	Statistical methods are NOT appropriate	RoB Box 10c (9)
8 Provide a clear description of how missing items will be handled	The way missing items will be handled is clearly described		The way missing items will be handled is not clearly described		Original CC

D. Construct approach: (i.e. hypotheses testing: before and after intervention)	**very good**	**adequate**	**doubtful**	**inadequate**	**NA**	**Justification**
Design requirements						
1 Formulate challenging hypotheses regarding expected changes before and after intervention a priori (i.e. before data collection)	Hypotheses formulated including the expected changes stated		Hypotheses vague or not formulated but possible to deduce what was expected	Unclear what was expected		Original CC
2 Provide an adequate description of the intervention to allow replication, including how and when they will be administered	Adequate description of the intervention		Poor description of the intervention	NO description of the intervention		RoB Box 10d (11)
3 Use an appropriate time interval between first and second administration	Time interval appropriate			Time interval NOT appropriate		New
4 Describe anything likely to occur in the interim period (e.g. intervention, progressive disease, other relevant events) is adequately	Anything likely to occur during the interim period (e.g. treatment) is adequately described		Unclear or NOT described what will likely to occur during the interim period			Original CC

5 Ensure that a proportion of the patients is likely to change (i.e. improvement or deterioration) on the construct to be measured	Part of the patients is likely to change (evidence provided)	NO evidence provided, but assumable that part of the patients will change	Unclear if part of the patients will change	Patients will likely NOT change	Original CC
6 Perform the analysis in a sample with an appropriate number of patients (taking into account expected number of missing values)	≥100 patients	50-99 patients	30-49 patients	<30 patients	Sample size
Statistical methods					
7 Ensure that the statistical methods are adequate for the hypotheses to be tested	Statistical methods are appropriate	Assumable that statistical methods are appropriate	Statistical methods are not optimal	Statistical methods are NOT appropriate	RoB Box 10d (12)
8 Provide a clear description of how missing items will be handled	The way missing items will be handled is clearly described		The way missing items will be handled is not clearly described		Original CC

Translation process

The process of translating an existing PROM is not a measurement property. Rather, it is part of the development phase of a new version of a PROM. However, a good translation process will likely result in a more valid version of the PROM in the translated language. In this translation box standards are provided to assess the quality of the translation process. When the cross-cultural validity of a translated PROM will be tested subsequently, we refer to the box Cross-cultural validity\measurement invariance using the group variable 'language'.

Translation process	**very good**	**adequate**	**doubtful**	**inadequate**	**Justification**
Design requirements					
1 Describe both the original language in which the PROM was developed, the source language (if different from the original language) and the language in which the PROM will be translated	Original language, source language and target language will be described			Source language will NOT be described	Original CC
2 Ensure that the items will be translated forward and backward	Multiple forward and multiple backward translations	Multiple forward translations but one backward translation	One forward and one backward translation	Only a forward translation	Original CC
3 Ensure that both forward translators have a mother tongue in the target language in which the PROM will be translated	Both forward translators have a mother tongue in the target language in which the PROM will be translated		Only one of the forward translators a mother tongue in the target language in which the PROM will be translated	Both forward translators don't have a mother tongue in the target language	Original CC
4 Ensure that one of the forward translators has expertise in the diseases involved, and in the construct measured by the PROM; the other forward translators is naïve on the construct measured by the PROM	One of the forward translators has expertise on disease and construct to be measured, other translator is naïve	Unclear what expertise of both forward translators with respect to disease or construct	Both forward translators are either both experts with respect to disease or construct, or both naïve with respect to disease or construct		Original CC

5	Ensure that both backward translators have a mother tongue in the original or source language	Both forward translators have a mother tongue in the source language in which the PROM will be translated		Only one of the forward translators a mother tongue in the source language in which the PROM will be translated	Both forward translators don't have a mother tongue in the source language	Original CC
6	Ensure that both backward translators are naïve in the disease involved and the construct to be measured	Both backward translators will be naïve in the disease involved and the construct to be measured	Unclear if both backward translators will be naïve in the disease involved and the construct to be measured			New
7	Ensure that the translators will work independently from each other	Translators will work independent	Assumable that the translators will work independent	Unclear whether translators will work independent	Translators will NOT work independent	Original CC
8	Provide a clear description on how differences between the original and translated versions will be resolved	Adequate description of how differences between translators will be resolved	Poorly or NOT described how differences between translators will be resolved			Original CC
9	Ensure that the translation will be reviewed by a committee (including the original developers of the PROM)	Translation will be reviewed by a committee (involving other people than the translators, e.g. the original developers)	Translation will NOT be reviewed by (such) a committee			Original CC
10	Write a feedback report of the translation process	Feedback report will be written		No feedback report will be written		New

11	Perform a pilot study (e.g. cognitive interview study) to check (1) the relevance of each item for the patients' experience with the condition, **AND** (2) the comprehensiveness of the PROM, **AND** (3) the comprehensibility of the PROM instructions, items, response options, and recall period	Widely recognized or well justified method for qualitative research will be used to assess the three aspects	Only quantitative (survey) method(s) will be used or assumable that the method used will be appropriate but not clearly described, but all three aspects will be assessed	Not clear if patients will be asked whether each item is relevant AND comprehensible AND whether items together are comprehensive, or doubtful whether the method will be appropriate	Method used are not appropriate or patients will not be asked about the relevance, comprehensiveness or comprehensibility of all items	RoB Box 1
12	Perform the pilot study in a patient population representing the target population	The study will be performed in a sample representing the target population	Assumable that the study will be performed in a sample representing the target population	Doubtful whether the study will be performed in a sample representing the target population	Study will NOT be performed in a sample representing the target population	RoB Box 1

References

1. Mokkink LB, Terwee CB, Patrick 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(4):539-49. doi: 10.1007/s11136-010-9606-8 [doi]
2. Mokkink LB, de Vet HCW, Prinsen CAC, et al. COSMIN Risk of Bias checklist for systematic reviews of Patient-Reported Outcome Measures. *Qual Life Res* 2018;27(5):1171-79. doi: 10.1007/s11136-017-1765-4
3. Terwee CB, Mokkink LB, Knol DL, et al. Rating the methodological quality in systematic reviews of studies on measurement properties: a scoring system for the COSMIN checklist. *Qual Life Res* 2012;21(4):651-57. doi: 10.1007/s11136-011-9960-1 [doi]
4. Prinsen CAC, Mokkink LB, Bouter LM, et al. COSMIN guideline for systematic reviews of patient-reported outcome measures. *Qual Life Res* 2018;27(5):1147-57. doi: 10.1007/s11136-018-1798-3
5. de Vet HC, Terwee CB, Mokkink L, et al. Measurement in Medicine: a practical guide: Cambridge University Press 2010.
6. Mokkink LB, Vet HC, Prinsen CA, et al. COSMIN methodology for systematic reviews of Patient-Reported Outcome Measures (PROMs) - user manual 2018. Available from: www.cosmin.nl.
7. Terwee CB, Prinsen CA, de Vet HCW, et al. COSMIN methodology for assessing the content validity of Patient-Reported Outcome Measures (PROMs). User manual., 2018. Available from: www.cosmin.nl.
8. Fayers PM, Hand DJ. Factor analysis, causal indicators and quality of life. *Qual Life Res* 1997;6(2):139-50.
9. Fayers PM, Hand DJ, Bjordal K, et al. Causal indicators in quality of life research. *Qual Life Res* 1997;6(5):393-406.
10. Streiner DL. Being inconsistent about consistency: when coefficient alpha does and doesn't matter. *J Pers Assess* 2007;80:217-22.
11. McGraw KOW, S.P. Forming inferences about some intraclass correlation coefficients. *Psychological Methods* 1996;1:30-46.