4 Methodology for Systematic Reviews on Measurement Properties of Patient Reported Outcome Measures (PROMS)

Orestis Argyriou, Michail Chatzikonstantinou, Vanash Patel, and Thanos Athanasiou

Patient Reported Outcome Measures
Measurement Properties
Validity
Reliability
Responsiveness
Systematic Review
To evaluate
To compare
To recommend
To identify the need for new tools
Methodology

O. Argyriou (✉)
Health Education England – North West London, London, UK

The Hillingdon Hospitals NHS Foundation Trust, Uxbridge, UK
e-mail: orestis.argyriou@nhs.net

M. Chatzikonstantinou
Bariatric Centre for Weight Management and Metabolic Surgery, University College Hospital, London, UK
e-mail: m.chatzikonstantinou@nhs.net

V. Patel
West Hertfordshire Teaching Hospitals NHS Trust, Watford, UK

Imperial College London, London, UK
e-mail: vanash.patel06@imperial.ac.uk

T. Athanasiou
Department of Surgery and Cancer, Imperial College London, London, UK

Imperial College Healthcare NHS Trust, London, UK
e-mail: t.athanasiou@imperial.ac.uk

T. Athanasiou et al. (eds.), *Patient Reported Outcomes and Quality of Life in Surgery*,
https://doi.org/10.1007/978-3-031-27597-5_4

Patient Reported Outcome Measures can be assessed by evaluating their **Measurement Properties**.

A **systematic review** can be performed in order **to compare and evaluate** PROMs, to **make recommendations** regarding their use, and to identify any gaps or the **need for the design of a new instrument**.

The **COSMIN initiative** (Consensus-based Standards for the selection of health Measurement Instruments) has provided thorough **methodological guides** for performing such a systematic review.

This involves a step-wise approach, to assess separately **content validity, internal structure and the remaining measurement properties**.

Following the current advancements and increased scientific interest in research relating to **quality of life**, particularly with the use of patient reported outcome tools, clinicians are frequently involved in relevant studies.

A clinician may be interested to investigate which tool is **more appropriate for their practice**, and this is the purpose of this methodological overview.

Nevertheless, although a clinician can massively benefit from a more in-depth understanding of this methodology, it is **strongly advised** that such studies should be undertaken in **close collaboration with Epidemiologists and Biostatisticians**.

Introduction

Aim of the Chapter

This chapter aims to discuss and present the currently used methodology for performing studies and systematic reviews on the measurement properties of PROMs.

It aims to initially provide some insight into the most common terms utilised in the fields of designing and interpreting reported papers and results on PROMs.

The process of PROMs design, and generation of a new PROM is beyond the scope of this chapter and is only discussed as part of the assessment and evaluation of studies for a systematic review.

What Are Patient Reported Outcomes (PROs) and Patient Reported Outcome Measures (PROMs)

Patient-reported Outcomes (PROs) have long been established in current medical research, as both primary and secondary outcomes of studies.

According to the FDA, a ***Patient-Reported Outcome (PRO)*** is 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 [1].

As ***Patient-Reported Outcome Measures (PROMs)*** or, alternatively ***PRO instruments***, we define the instruments that are utilised to measure PROs or capture PRO data, such as questionnaires that are completed by patients [1].

In the relevant literature, when referring to a PROM or a PROM instrument, authors may be discussing a questionnaire as a whole or single question.

What Are the Measurement Properties of PROMs

Measurement properties are essential criteria in the design and evaluation of a PROM.

Broadly, these are Validity, Reliability, Responsiveness and Interpretability. Detailed definitions will be discussed below.

Why Perform Systematic Reviews on Measurement Properties of PROMs

Provided that PROMS, looking at an area of interest, exist already (developed and/or validated), a systematic review may be performed, in order to **compare the measurement properties** of these PROMs, **evaluate the quality** of each PROM, **identify advantages and disadvantages** of each PROM, and ultimately, **recommend which PROMs should be used in future studies**.

In addition, if the results indicate a rather low quality of the available PROMs, or inadequate measurement of the area of interest, then the systematic review may **inform and guide the design of a new PROM**.

Current Methodology: The COSMIN Initiative

The vast majority of guidance and tools on PROMs interpretation, has been provided by the **Consensus-based Standards for the selection of health Measurement Instruments (COSMIN) initiative** [2].

The COSMIN initiative, after initially identifying the lack of clear definitions and widely accepted methodology [3], has specified the definitions of the measurement properties of PROMs [4], and also provides comprehensive guidance for performing a systematic review of outcome measurements, as well as handbooks for the interpretation and assessment of each measurement property in PROMs.

Definitions and Taxonomy

In order to perform a systematic review on measurement properties of PROMs, the researcher must be familiar with the measurement properties, and their definitions.

As mentioned previously, the COSMIN initiative, following a Delphi study, has recommended definitions for the measurement properties [4].

Most importantly, the initiative agreed on a taxonomy, incorporating the measurement properties [4].

According to this taxonomy, COSMIN identifies three main domains of measurement properties in assessing the quality of a PROM; Validity, Reliability and Responsiveness with Interpretability being considered as a fourth domain (Fig. 4.1 and Table 4.1). A fourth domain, Interpretability, is also considered [4].

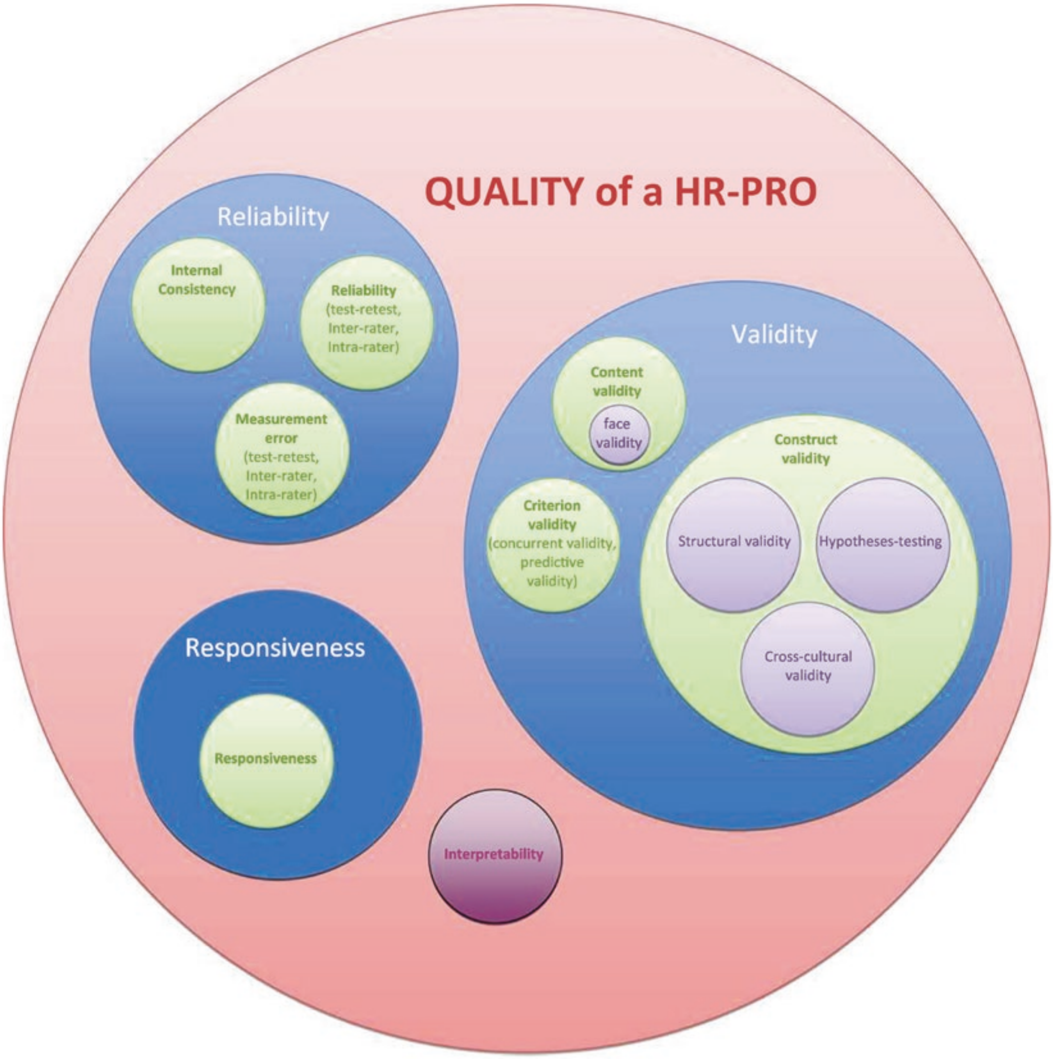


Fig. 4.1 Three plus one domains of assessment of a quality of a PROM. Mokkink, L. B. et al. The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes. J. Clin. Epidemiol. 63, 737–745 (2010)

Table 4.1 Definitions of the three domains of the assessment of a PROM

Definitions	
Validity	The degree to which an HR-PRO instrument measures the construct(s) it purports to measure
Content validity	The degree to which the content of an HR-PRO instrument is an adequate reflection of the construct to be measured
*Face validity	The degree to which (the items of) an HR-PRO instrument indeed looks as though they are an adequate reflection of the construct to be measured
Construct validity	The degree to which the scores of an HR-PRO instrument 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 HR-PRO instrument validly measures the construct to be measured
Structural validity	The degree to which the scores of an HR-PRO instrument are an adequate reflection of the dimensionality of the construct to be measured
Cross-cultural validity	The degree to which the performance of the items on a translated or culturally adapted HR-PRO instrument are an adequate reflection of the performance of the items of the original version of the HR-PRO instrument
Criterion validity	The degree to which the scores of an HR-PRO instrument are an adequate reflection of a "gold standard"
Reliability	The degree to which the measurement is free of measurement error i.e. the extent to which scores for patients who have not changed are the same for repeated measurement under several conditions
Internal consistency	The degree of the interrelatedness among the items
Reliability	The proportion of the total variance in the measurements which is because of "true" differences among 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
Responsiveness	The ability of an HR-PRO instrument to detect change over time in the construct to be measured
Interpretability	The degree to which one can assign qualitative meaning that is, clinical or commonly understood connotations—to an instrument's quantitative scores or change in scores

Performing a Systematic Review

General

A systematic review on measurement properties of PROMs shares some common methodological features with any other systematic review. We will focus more on discussing the process of assessing the measurement properties.

The COSMIN initiative has provided summarising guidelines for performing a systematic review [5] as well as a more detailed user manual, describing the methodology in more depth [6].

In this section, we will present and discuss the processes recommended in these documents. All tables and figures are adopted from these sources.

The overall process and the steps that need to be followed, can be shown in the following flowchart [5].

As shown in the flowchart, a systematic review consists of three stages (Fig. 4.2).

Initially, as per routine practice, a literature search is performed followed by a thorough assessment of the measurement properties. Finally, recommendations can be exported and formed, and the review is reported.

Literature Search

The initial stage consists of the standard steps (steps 1–4) for performing systematic reviews.

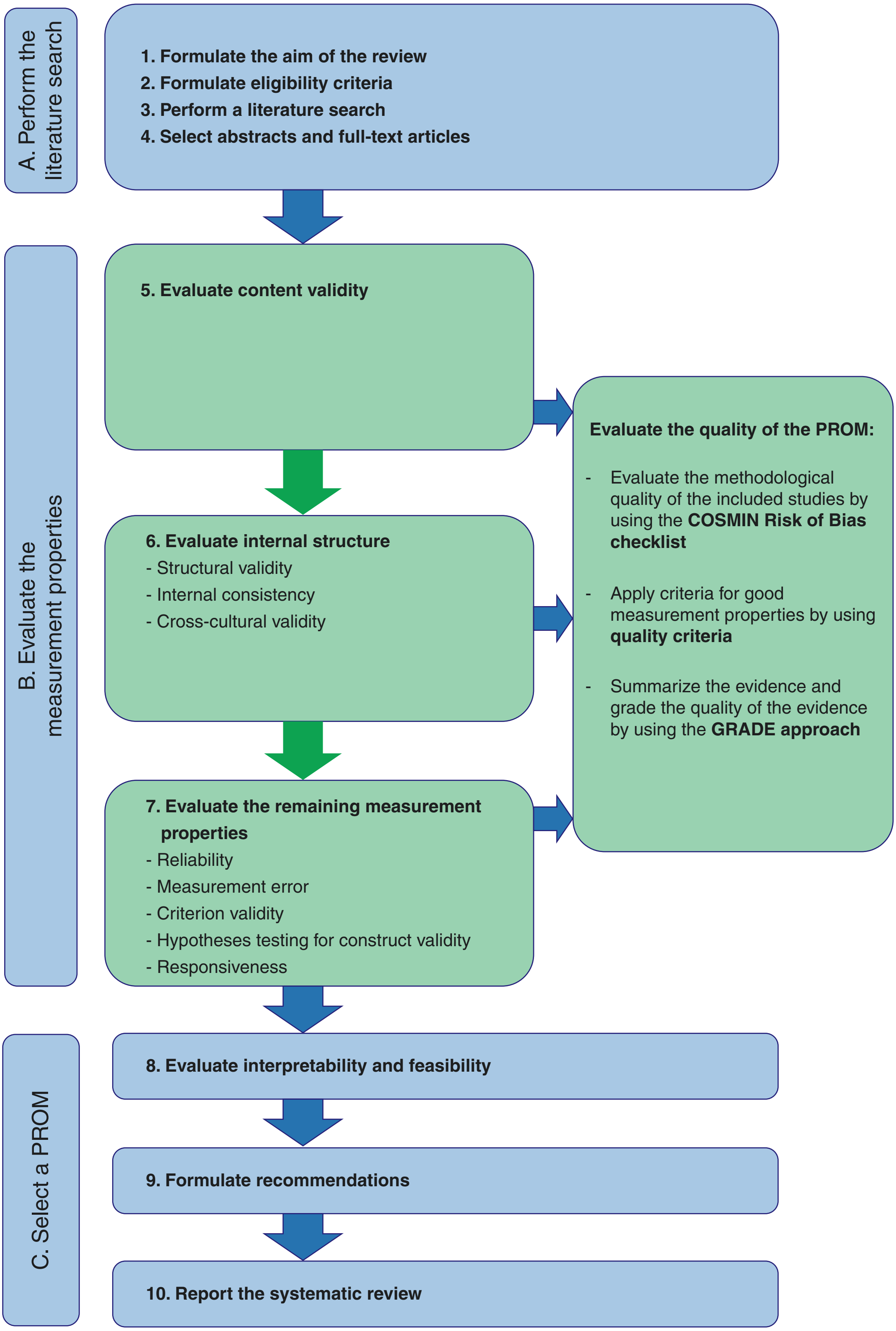

A. Perform the literature search
1. Formulate the aim of the review
2. Formulate eligibility criteria
3. Perform a literature search
4. Select abstracts and full-text articles
B. Evaluate the measurement properties
5. Evaluate content validity
6. Evaluate internal structure
- Structural validity
- Internal consistency
- Cross-cultural validity
7. Evaluate the remaining measurement properties
- Reliability
- Measurement error
- Criterion validity
- Hypotheses testing for construct validity
- Responsiveness
Evaluate the quality of the PROM:
- Evaluate the methodological quality of the included studies by using the COSMIN Risk of Bias checklist
- Apply criteria for good measurement properties by using quality criteria
- Summarize the evidence and grade the quality of the evidence by using the GRADE approach
C. Select a PROM
8. Evaluate interpretability and feasibility
9. Formulate recommendations
10. Report the systematic review

Fig. 4.2 The first four stages of a literature search. Prinsen, C. A. C. et al. COSMIN guideline for systematic reviews of patient-reported outcome measures. Qual. Life Res. 27, 1147–1157 (2018)

- **Step 1: Formulating the aim**
  When deciding and developing the aim of the review, the four key elements that need to be included are the construct of interest, the population, the type of the instrument and the measurement properties of interest.
- **Step 2: Formulating the Eligibility Criteria**
  Not all studies mentioning the PROMs of interest are to be included. Eligible studies should fulfil the aforementioned four key elements. Most importantly, given the large amount of studies on different PROMs, the main focus should be studies looking at the assessment and evaluation of one (or more) of the measurement properties of the PROM, and certainly not studies just using the PROM as an outcome measurement.
- **Step 3: Performing the literature search**.
  Standard Cochrane methodology should be followed for performing the literature search. The four key elements of the aim need to be included, as can be shown in the following flowchart, depicting the search strategy and terms, as described by the COSMIN initiative [5] (Fig. 4.3)
- **Step 4: Selection of abstracts and full-text articles**
  Selection and review of the abstracts and full texts is performed in a routine manner with the general recommendation for this to be performed by two reviewers independently.

Evaluation of Measurement Properties

As demonstrated in the flowchart in Fig. 4.2, this is done in three main stages. Given the significance of content validity and internal structure, these are assessed separately, followed by assessment of the remaining properties.

1. Content Validity
2. Internal Structure
3. Remaining Properties (Reliability, Measurement error, Criterion validity, Hypotheses testing for construct validity, Responsiveness)

Evaluation of Content Validity

The COSMIN initiative, given the significance and complexity of the evaluation of content validity, provides a separate user manual, with the relevant methodology [7].

According to the COSMIN recommendations, there are three aspects of content validity in a PROM:

- Relevance
- Comprehensiveness
- Comprehensibility

In order to assess these, COSMIN recommends ten criteria for good content validity, which have been formulated following a Delphi study [8], as shown in Table 4.2.

To assess the above, we are using a stepwise process:

Step 1—Evaluation of the **quality of the PROM development**
Step 2—Evaluation of the **quality of content validity studies on the PROM**
Step 3—Evaluation of the **content validity of the PROM**

A more detailed description of the steps is provided below, but not in its full length and detail. For each step, COSMIN has very comprehensively provided relevant boxes, summarising the process in a rather succinct manner. These will also be presented below.

Step 1: Evaluating the Quality of the PROM Development

This step is further subdivided into steps 1a and 1b.

In **step 1a**, the quality of the PROM design is assessed (evaluating ***relevance***).

In **step 1b**, the quality of any cognitive interview studies or pilot studies assessing the PROM, are examined (evaluating ***comprehensibility and comprehensiveness***) (Table 4.3).

To perform the above steps, a number of items/questions need to be answered, as per the flowchart shown below (Fig. 4.4).

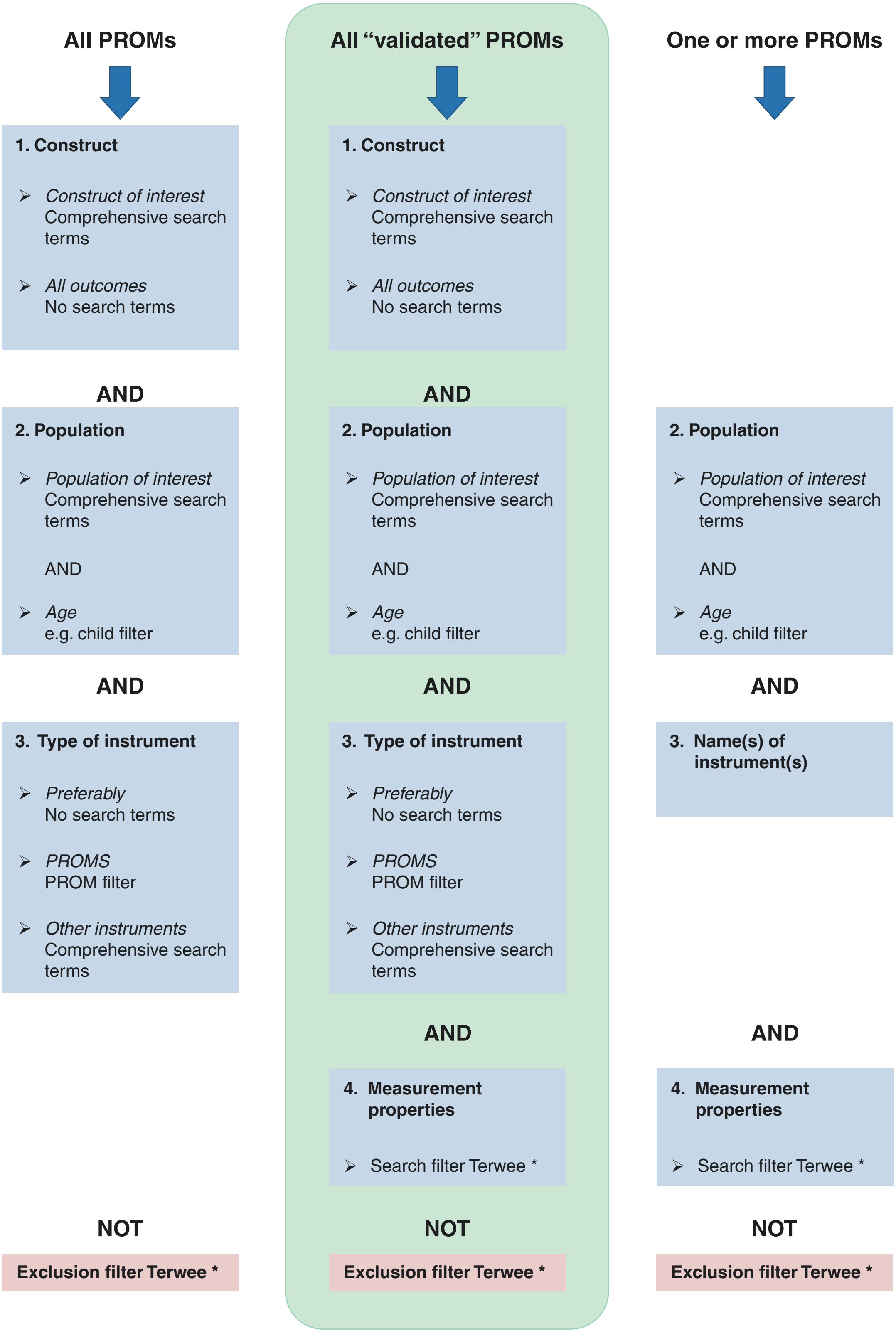
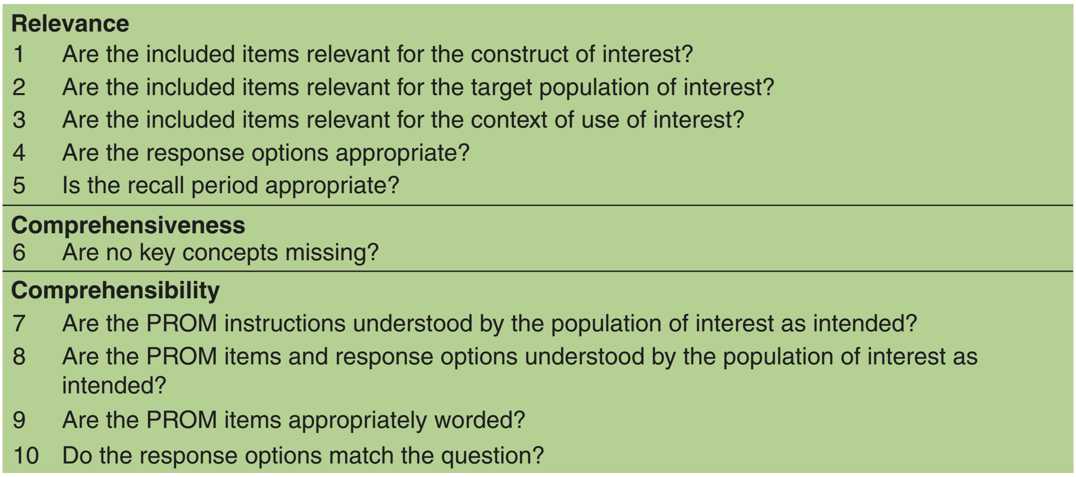


Fig. 4.3 Step 3, performing the literature search. Prinsen, C. A. C. et al. COSMIN guideline for systematic reviews of patient-reported outcome measures. Qual. Life Res. 27, 1147–1157 (2018)

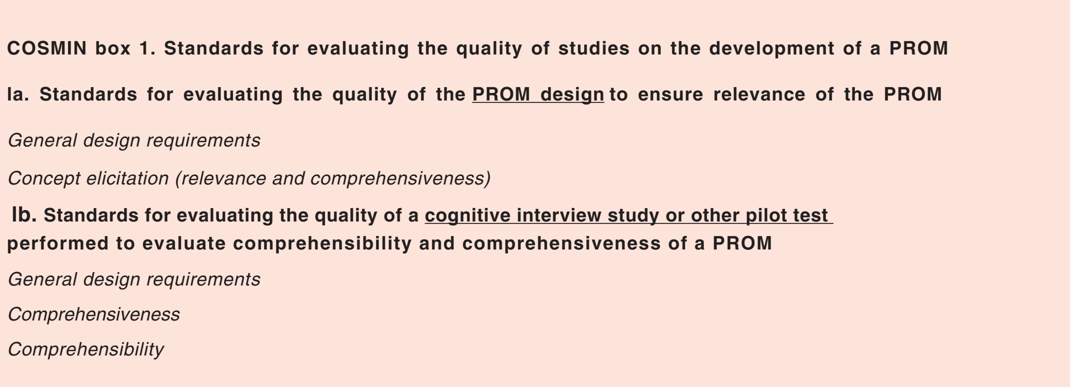
Table 4.2 Criteria for good content validity



Relevance	
1	Are the included items relevant for the construct of interest?
2	Are the included items relevant for the target population of interest?
3	Are the included items relevant for the context of use of interest?
4	Are the response options appropriate?
5	Is the recall period appropriate?
Comprehensiveness	
6	Are no key concepts missing?
Comprehensibility	
7	Are the PROM instructions understood by the population of interest as intended?
8	Are the PROM items and response options understood by the population of interest as intended?
9	Are the PROM items appropriately worded?
10	Do the response options match the question?

Terwee, C. B. et al. COSMIN methodology for evaluating the content validity of patient-reported outcome measures: a Delphi study. Qual. Life Res. 27, 1159–1170 (2018)

Table 4.3 COSMIN box 1



COSMIN box 1. Standards for evaluating the quality of studies on the development of a PROM

Ia. Standards for evaluating the quality of the PROM design to ensure relevance of the PROM

General design requirements

Concept elicitation (relevance and comprehensiveness)

Ib. Standards for evaluating the quality of a cognitive interview study or other pilot test performed to evaluate comprehensibility and comprehensiveness of a PROM

General design requirements

Comprehensiveness

Comprehensibility

This describes 13 items/questions for Part 1a, and 22 items/questions for Part 1b. The detailed items are not presented here, and we would recommend reading the full manual, where the items are presented, along with further explanations and examples.

Step 2: Evaluating the Quality of Content Validity Studies on the PROM

In this step, we assess how patients and professionals were asked about the relevance, comprehensibility and comprehensiveness, either as part of the PROM design process, or as a separate content validity study (Table 4.4).

This can also be widely separated in Steps 2a, 2b and 2c (asking patients about relevance, comprehensiveness and comprehensibility), and steps 2d and 2e (asking professionals about relevance and comprehensiveness), as shown in the respective flowchart. Overall, there are 31 items/questions to be assessed (Fig. 4.5).

For Steps 1–2

As mentioned previously, the exact items that are utilised in each step are not presented here.

What is important to note is how ratings are provided for each item. A 4-point rating scale is utilised, as shown here.

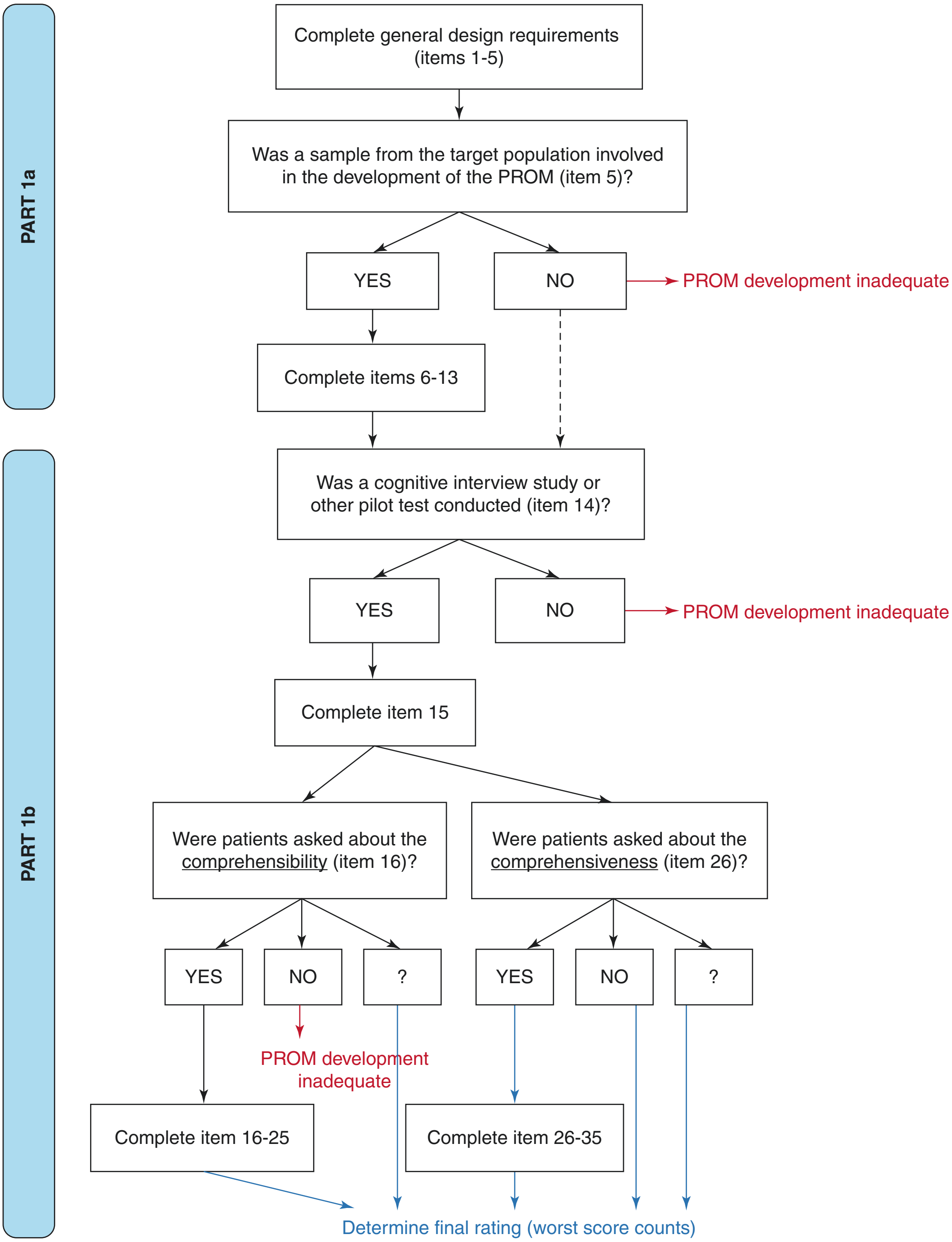

PART 1a
PART 1b
Complete general design requirements (items 1-5)
Was a sample from the target population involved in the development of the PROM (item 5)?
YES
NO
PROM development inadequate
Complete items 6-13
Was a cognitive interview study or other pilot test conducted (item 14)?
YES
NO
PROM development inadequate
Complete item 15
Were patients asked about the comprehensibility (item 16)?
Were patients asked about the comprehensiveness (item 26)?
YES
NO
?
YES
NO
?
PROM development inadequate
Complete item 16-25
Complete item 26-35
Determine final rating (worst score counts)

Fig. 4.4 Evaluating the quality of the PROM development. Caroline B Terwee et al. COSMIN methodology for assessing the content validity of PROMs

Table 4.4 COSMIN box 2: Standards for evaluating the quality of studies on the content validity of a PROM

COSMIN box 2. Standards for evaluating the quality of studies on the content validity of a PROM

2a. Asking patients about the relevance of the PROM items

2b. Asking patients about the comprehensiveness of the PROM

2c. Asking patients about the comprehensibility of the PROM

2d. Asking professionals about the relevance of the PROM items

2e. Asking professionals about the comprehensiveness of the PROM

Caroline B Terwee et al. COSMIN methodology for assessing the content validity of PROMs

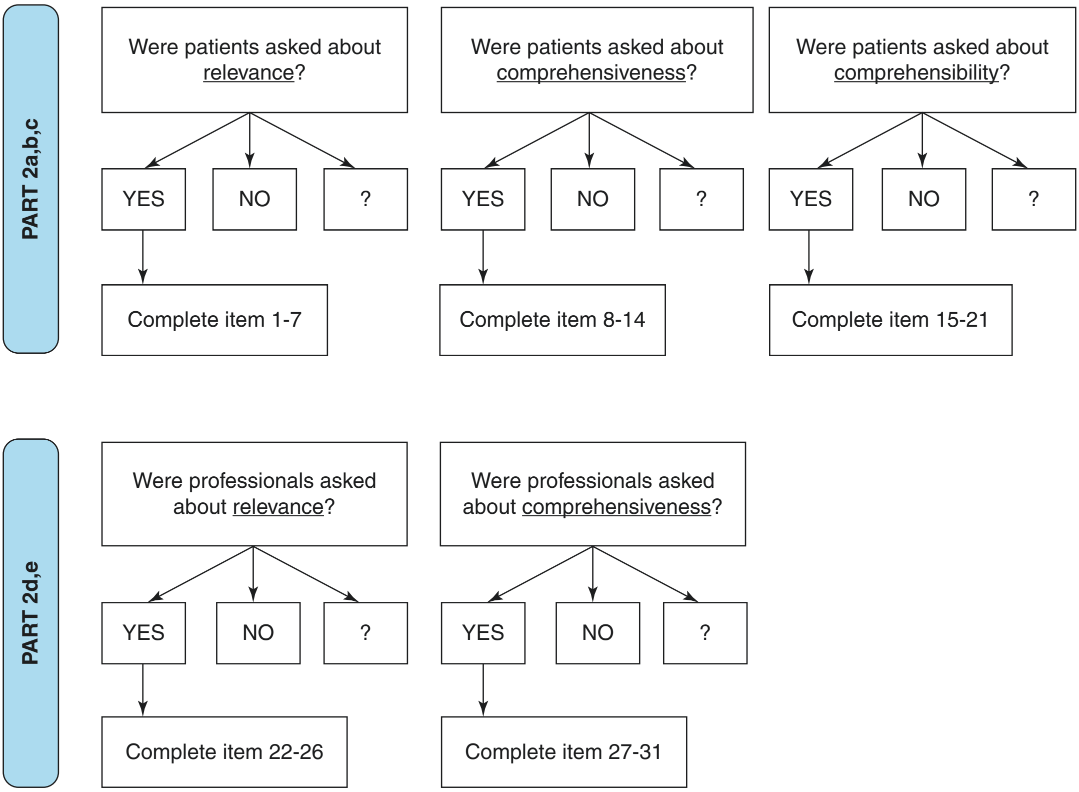

PART 2a,b,c
Were patients asked about relevance?
YES
NO
?
Complete item 1-7
Were patients asked about comprehensiveness?
YES
NO
?
Complete item 8-14
Were patients asked about comprehensibility?
YES
NO
?
Complete item 15-21
PART 2d,e
Were professionals asked about relevance?
YES
NO
?
Complete item 22-26
Were professionals asked about comprehensiveness?
YES
NO
?
Complete item 27-31

Fig. 4.5 Evaluating the quality of content validity studies on the PROM. Caroline B Terwee et al. COSMIN methodology for assessing the content validity of PROMs

- Very good
- Adequate
- Doubtful
- Inadequate

For each item, the COSMIN manuals provide detailed examples of what criteria should be fulfilled to achieve is rating. Below we provide an example, of Item 5, from step 1a (Table 4.5).

To ensure high quality, COSMIN recommends using a 'worst score counts' method, where the lowest rating is utilised as an overall rating.

For Step 1, the lowest rating in the respective items will correspond to the overall rating for the PROM development.

Table 4.5 Example of the COSMIN manuals

		Very good	Adequate	Doubtful	Inadequate	Not applicable
5	Was the PROM development study performed in a sample representing the target population for which the PROM was developed?	Study performed in a sample representing the target population	Assumable that the study was performed in a sample representing the target population, but not clearly described	Doubtful whether the study was performed in a sample representing the target population	Study not performed in a sample representing the target population **(SKIP standards 6-12)**	

Caroline B Terwee et al. COSMIN methodology for assessing the content validity of PROMs

Table 4.6 COSMIN criteria and rating system for evaluating the content validity of PROM

Name of the PROM or subscale:.................................		PROM development study	Content validity study 1	Content validity study 2[2]	Rating of reviewers	OVERALL RATINGS PER PROM[3] (see step 3b)	QUALITY OF EVIDENCE (see step 3c)
Criteria (see Table 2)		+ / - / ± /?[1]	+ / - / ± /?	+ / - / ± /?	+ / - / ± /?	+ / - / ±	High, moderate, low, very low
Relevance							
1	Are the included items relevant for the construct of interest?[4]						
2	Are the included items relevant for the target population of interest?[4]						
3	Are the included items relevant for the context of use of interest?[4]						
4	Are the response options appropriate?						
5	Is the recall period appropriate?						
	RELEVANCE RATING (see Table 3)						
Comprehensiveness							
6	Are all key concepts included?						
	COMPREHENSIVENESS RATING (see Table 3)						
Comprehensibility							
7	Are the PROM instructions understood by the population of interest as intended?						
8	Are the PROM items and response options understood by the population of interest as intended?						
9	Are the PROM items appropriately worded?						
10	Do the response options match the question?						
	COMPREHENSIBILITY RATING (see Table 3)						
	CONTENT VALIDITY RATING (see Table 4)						

[1] Ratings for the 10 criteria can only be + / - /?. The RELEVANCE, COMPREHENSIVENESS, COMPREHENSIBILITY, AND CONTENT VALIDITY ratings can be + / - / ± /?

[2] Add more columns if more content validity studies are available

[3] If ratings are inconsistent between studies, consider using separate tables for subgroups of studies with consistent results.

[4] These criteria refer to the construct, population, and context of use of interest in the systematic review.

Caroline B Terwee et al. COSMIN methodology for assessing the content validity of PROMs

For Step 2, the lowest rating in the respective items will correspond to the overall rating of the content validity studies on the PROM.

Step 3: Evaluating the Content Validity of the PROM

For this step, content validity of the PROM is evaluated by examining the quality and results of already performed studies on the PROM. This, again, is further subdivided in three steps.

For **step 3a**, ratings need to be provided for relevance, comprehensiveness and comprehensibility, using the ten criteria for good content (presented previously), for three different aspects, as per the table shown below.

- Methods and results of PROM development study
- Content validity studies on the PROM
- Reviewers' own ratings of the PROM (Table 4.6)

Table 4.7 GRADE criteria

Study design	Quality of evidence	Lower if
At least 1 content validity study	High	Risk of bias -1 Serious -2 Very serious -3 Very serious Inconsistency -1 Serious -2 Very serious Indirectness -1 Serious -2 Very serious
No content validity studies	Moderate	
	Low	
	Very low	

https://www.gradeworkinggroup.org/

Essentially, the ratings for the methods and results of the PROM development studies, and the content validity studies, are the ones already assessed in steps 1 and 2, according to the respective COSMIN boxes, and are utilised in this table.

With regards to the potential ratings of each criterion, these can be:

- **Sufficient (+)**: ≥85% of the items of the PROM (or sub-scale) fulfil the criterion
- **Insufficient (−)**: <85% of the items of the PROM (or sub-scale) does fulfil the criteria
- **Indeterminate (?)**: No(t enough) information available or quality of (part of a) the study inadequate

After ratings have been provided for each criterion, a final rating can be generated for relevance, comprehensiveness and comprehensibility. These three ratings are then combined to provide the Overall Content Validity Rating.

For these processes, COSMIN provides further tables and guidance in the manual, which are not presented here.

Importantly, given the individual importance of relevance, comprehensiveness and comprehensibility, it is recommended to report on them separately, if found relevant (different ratings/different importance), and not only as an Overall Content Validity Rating.

For **step 3b**, a qualitative summary of available studies is performed, providing a rating for relevance, comprehensiveness and comprehensibility, resulting in an overall rating for each domain, which will be added in the respective boxes of the aforementioned table.

Lastly, for **step 3c**, the ratings achieved from step 3b, are assessed with regards to the quality of the evidence that generated them, to determine how reliable these ratings are.

To do this, the GRADE approach is, as shown in the table below [9] (Table 4.7).

Summary of Content Validity Assessment

In summary, as per the COSMIN guidelines and the methodology to assess content validity, a structured and step-by-step approach was presented.

Sequentially, a number of aspects are being examined systemically, and the relevant outcomes need to be reported in a systematic review:

- Quality of PROM development process (step 1)
- Quality of content validity studies on the PROM (step 2)
- Overall ratings for relevance, comprehensiveness and comprehensibility, as well as a summative overall content validity rating (step 3)

Evaluation of Internal Structure

When evaluating internal structure, the properties that need to be assessed include structural validity, internal consistency and cross-cultural validity, as defined previously.

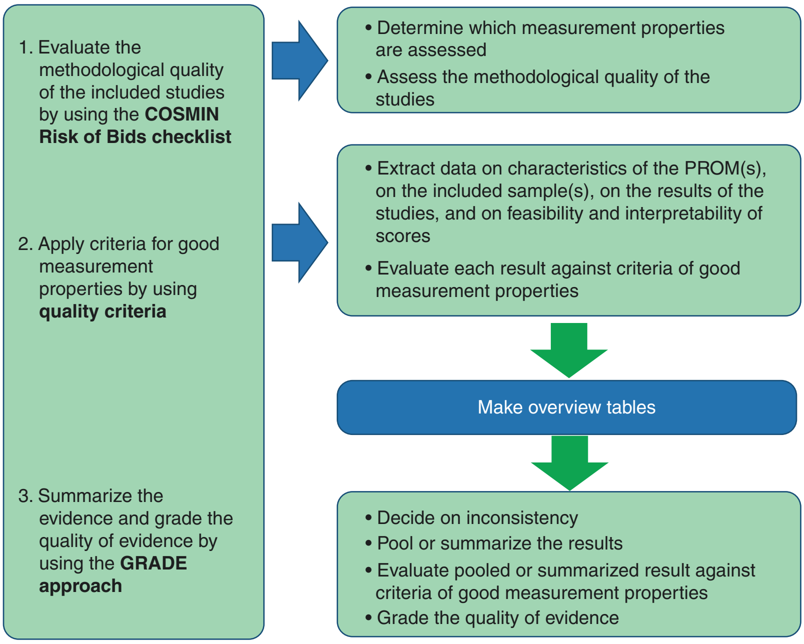
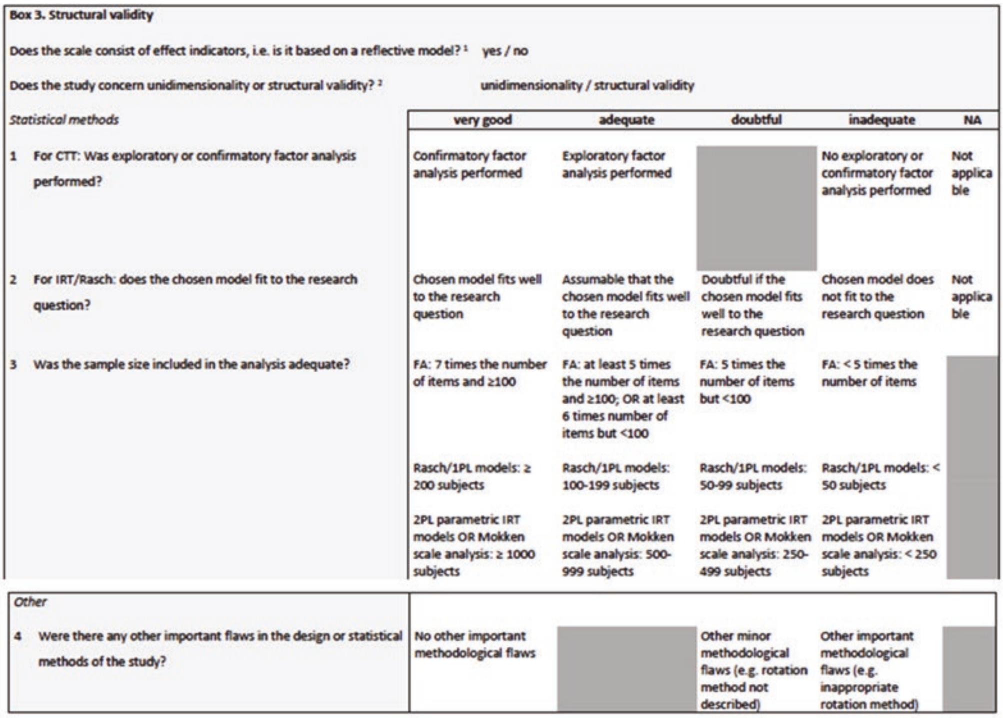

1. Evaluate the methodological quality of the included studies by using the **COSMIN Risk of Bids checklist**
2. Apply criteria for good measurement properties by using **quality criteria**
3. Summarize the evidence and grade the quality of evidence by using the **GRADE approach**
• Determine which measurement properties are assessed
• Assess the methodological quality of the studies
• Extract data on characteristics of the PROM(s), on the included sample(s), on the results of the studies, and on feasibility and interpretability of scores
• Evaluate each result against criteria of good measurement properties
Make overview tables
• Decide on inconsistency
• Pool or summarize the results
• Evaluate pooled or summarized result against criteria of good measurement properties
• Grade the quality of evidence

Fig. 4.6 Evaluation of internal structure. Lidwine B Mokkink, Cecilia AC Prinsen, Donald L Patrick, Jordi Alonso, Lex M Bouter, Henrica CW de Vet, Caroline B Terwee. COSMIN manual for systematic reviews of PROMs COSMIN methodology for systematic reviews of Patient-Reported Outcome Measures (PROMs) user manual. (2018)

As per the definition of internal structure, at this stage, reviewers need to evaluate if the items in a scale or sub scale are appropriately correlated manifestations of the same one underlying construct. Subsequently, this step is relevant for studies based on such a reflective model (not formative).

COSMIN recommends three steps for assessing internal structure, which are summarised in the following table (Fig. 4.6).

In the first step, the COSMIN Risk of Bias Checklist is utilised, by answering the relevant boxes for structural validity, internal consistency and cross-cultural validity/measurement Invariance, as demonstrated below [10].

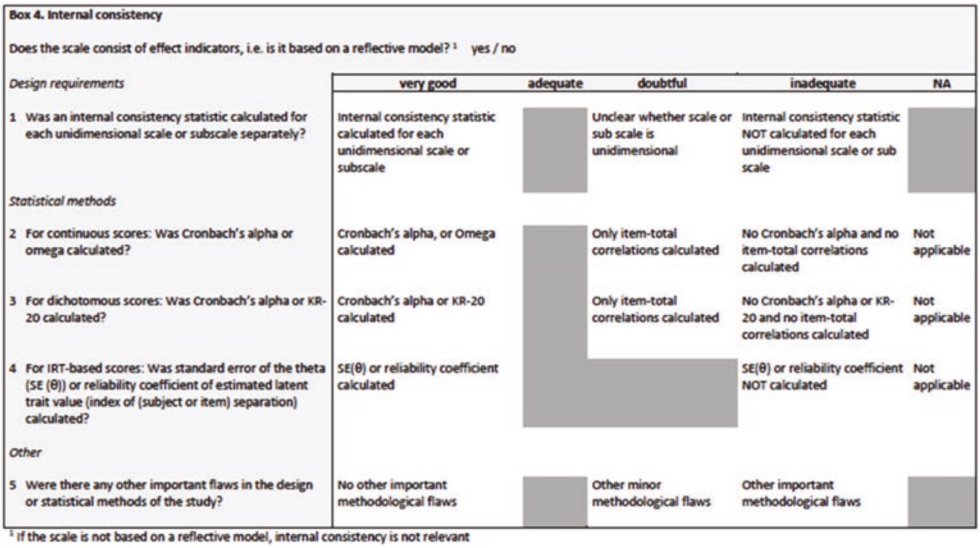


Box 3. Structural validity

Does the scale consist of effect indicators, i.e. is it based on a reflective model? [1] yes / no

Does the study concern unidimensionality or structural validity? [2] unidimensionality / structural validity

Statistical methods	very good	adequate	doubtful	inadequate	NA
1 For CTT: Was exploratory or confirmatory factor analysis performed?	Confirmatory factor analysis performed	Exploratory factor analysis performed		No exploratory or confirmatory factor analysis performed	Not applica ble
2 For IRT/Rasch: does the chosen model fit 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 applica ble
3 Was the sample size included in the analysis adequate?	FA: 7 times the number of items and ≥100	FA: at least 5 times the number of items and ≥100; OR at least 6 times number of items but <100	FA: 5 times the number of items but <100	FA: < 5 times the number of items	
	Rasch/1PL models: ≥ 200 subjects	Rasch/1PL models: 100-199 subjects	Rasch/1PL models: 50-99 subjects	Rasch/1PL models: < 50 subjects	
	2PL parametric IRT models OR Mokken scale analysis: ≥ 1000 subjects	2PL parametric IRT models OR Mokken scale analysis: 500-999 subjects	2PL parametric IRT models OR Mokken scale analysis: 250-499 subjects	2PL parametric IRT models OR Mokken scale analysis: < 250 subjects	
Other					
4 Were there any other important flaws in the design or statistical methods of the study?	No other important methodological flaws		Other minor methodological flaws (e.g. rotation method not described)	Other important methodological flaws (e.g. inappropriate rotation method)	

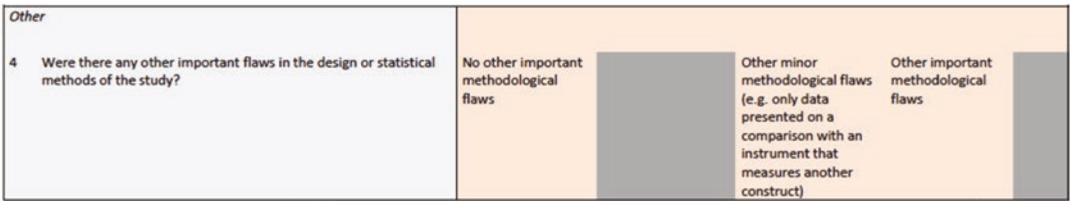


Box 4. Internal consistency

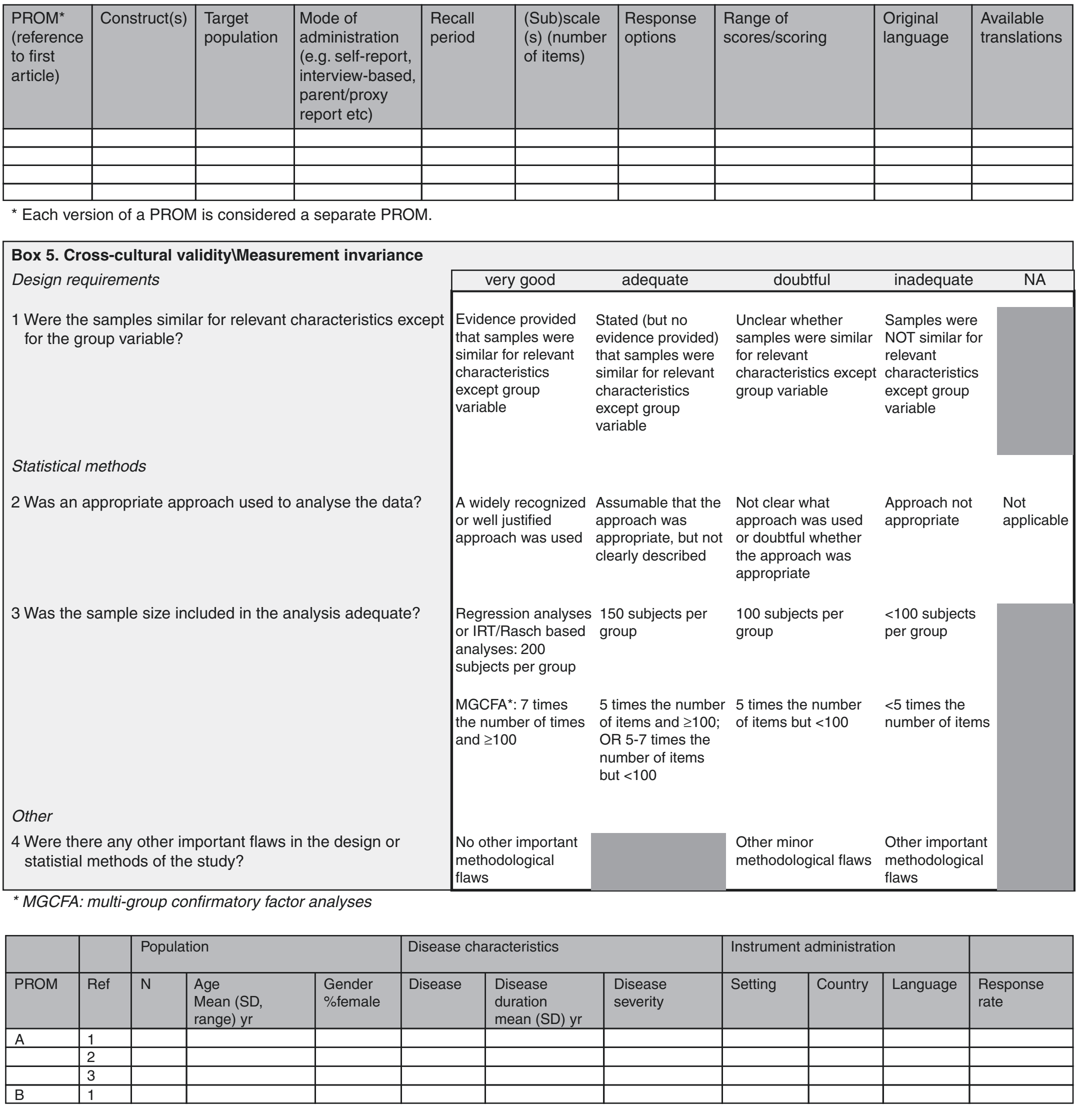
Does the scale consist of effect indicators, i.e. is it based on a reflective model? [1] yes / no

Design requirements	very good	adequate	doubtful	inadequate	NA
1 Was an internal consistency statistic calculated for each unidimensional scale or subscale separately?	Internal consistency statistic calculated for each unidimensional scale or subscale		Unclear whether scale or sub scale is unidimensional	Internal consistency statistic NOT calculated for each unidimensional scale or sub scale	
Statistical methods					
2 For continuous scores: Was Cronbach's alpha or omega calculated?	Cronbach's alpha, or Omega calculated		Only item-total correlations calculated	No Cronbach's alpha and no item-total correlations calculated	Not applicable
3 For dichotomous scores: Was Cronbach's alpha or KR-20 calculated?	Cronbach's alpha or KR-20 calculated		Only item-total correlations calculated	No Cronbach's alpha or KR-20 and no item-total correlations calculated	Not applicable
4 For IRT-based scores: Was standard error of the theta (SE (θ)) or reliability coefficient of estimated latent trait value (index of (subject or item) separation) calculated?	SE(θ) or reliability coefficient calculated			SE(θ) or reliability coefficient NOT calculated	Not applicable
Other					
5 Were there any other important flaws in the design or statistical methods of the study?	No other important methodological flaws		Other minor methodological flaws	Other important methodological flaws	

[1] If the scale is not based on a reflective model, internal consistency is not relevant



Other					
4 Were there any other important flaws in the design or statistical methods of the study?	No other important methodological flaws		Other minor methodological flaws (e.g. only data presented on a comparison with an instrument that measures another construct)	Other important methodological flaws	



PROM* (reference to first article)	Construct(s)	Target population	Mode of administration (e.g. self-report, interview-based, parent/proxy report etc)	Recall period	(Sub)scale (s) (number of items)	Response options	Range of scores/scoring	Original language	Available translations

* Each version of a PROM is considered a separate PROM.

Box 5. Cross-cultural validity\Measurement invariance					
Design requirements	very good	adequate	doubtful	inadequate	NA
1 Were the samples similar for relevant characteristics except for the group variable?	Evidence provided that samples were similar for relevant characteristics except group variable	Stated (but no evidence provided) that samples were similar for relevant characteristics except group variable	Unclear whether samples were similar for relevant characteristics except group variable	Samples were NOT similar for relevant characteristics except group variable	
Statistical methods					
2 Was an appropriate approach used to analyse the data?	A widely recognized or well justified approach was used	Assumable that the approach was appropriate, but not clearly described	Not clear what approach was used or doubtful whether the approach was appropriate	Approach not appropriate	Not applicable
3 Was the sample size included in the analysis adequate?	Regression analyses or IRT/Rasch based analyses: 200 subjects per group	150 subjects per group	100 subjects per group	<100 subjects per group	
	MGCFA*: 7 times the number of times and ≥100	5 times the number of items and ≥100; OR 5-7 times the number of items but <100	5 times the number of items but <100	<5 times the number of items	
Other					
4 Were there any other important flaws in the design or statistial methods of the study?	No other important methodological flaws		Other minor methodological flaws	Other important methodological flaws	

** MGCFA: multi-group confirmatory factor analyses*

		Population			Disease characteristics			Instrument administration			
PROM	Ref	N	Age Mean (SD, range) yr	Gender %female	Disease	Disease duration mean (SD) yr	Disease severity	Setting	Country	Language	Response rate
A	1										
	2										
	3										
B	1										

Fig. 4.7 Structural validity. Mokkink, L. B. COSMIN Risk of Bias checklist [PDF File]. Amsterdam Public Heal. Res. Inst. 1–37 (2018)

In the second step, data extraction is performed from studies on PROMS, focusing on patient characteristics, methods and timings of administration, interpretability, feasibility and results on measurement properties.

COSMIN provides the relevant tables that can facilitate and guide this data extraction (Fig. 4.7). The outcomes of theses will be evaluated against the criteria of good measurement properties (Table 4.8).

In the third step, reviewers should perform a quantitative pooled analysis or qualitative summary, and evaluated against the criteria for good measurement properties. Lastly, as described previously, grading of the evidence with the GRADE criteria, needs to be performed (Table 4.9).

These tables are presented as examples, with the intention to provide the research with an initial overview of the process. The thorough and extensive work done by the COSMIN initiative

Table 4.8 Mokkink Cecilia AC Prinsen Donald L Patrick Jordi Alonso Lex M Bouter Henrica CW de Vet Caroline B Terwee Contact LB Mokkink, L. B. COSMIN manual for systematic reviews of PROMs COSMIN methodology for systematic reviews of Patient-Reported Outcome Measures (PROMs) user manual. (2018)

		Population			Disease characteristics			Instrument administration			
PROM	Ref	N	Age Mean (SD, range) yr	Gender %female	Disease	Disease duration mean (SD) yr	Disease severity	Setting	Country	Language	Response rate
A	1										
	2										
	3										
B	1										

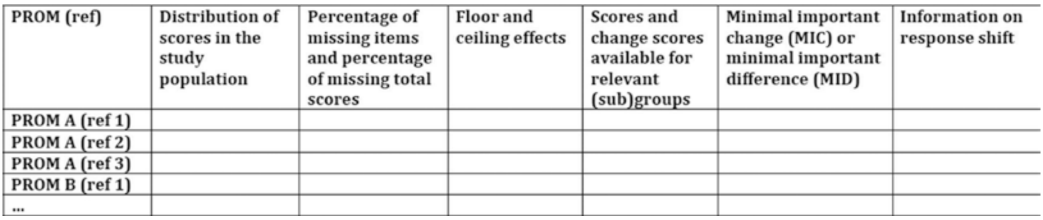
has given us a very precise methodology, which we would be duplicating if we were to describe these processes in more detail. Therefore, we strongly recommend that researchers refer to the relevant manuals and checklists, as cited throughout the chapter—that can also be found on the COSMIN website.

Evaluation of Reliability, Measurement Error, Criterion Validity, Hypotheses Testing for Construct Validity and RESPONSIVENESS

The remaining measurement properties, are once again assessed in a similar process, with the use of the respective COSMIN Risk of Bias Checklist boxes, which are indicatively shown below (Tables 4.10, 4.11, 4.12, 4.13, and 4.14).

Report and Selection of Most Suitable PROM

This final stage consists of evaluating interpretability and feasibility, formulating the recommendations and reporting the systematic review.



PROM (ref)	Distribution of scores in the study population	Percentage of missing items and percentage of missing total scores	Floor and ceiling effects	Scores and change scores available for relevant (sub)groups	Minimal important change (MIC) or minimal important difference (MID)	Information on response shift
PROM A (ref 1)						
PROM A (ref 2)						
PROM A (ref 3)						
PROM B (ref 1)						
…						

- **Evaluation of Interpretability and Feasibility** (Fig. 4.8)
  These are assessed with the use of the relevant tables
- **Formulation of Recommendations**
  COSMIN suggests dividing PROMs into three categories, according to the quality of evidence. In that way, the reviewers can assess and define which of the PROMs they assessed would be recommended for further use in the field, which require further studies and improvements, and which should not be used. The categories are shown below.
  (A) **Recommended**
  PROMs with evidence for sufficient content validity (any level) AND at least low quality evidence for sufficient internal consistency
  (B) **Further research required**
  PROMs categorised not in A or C
  (C) **Not recommended**
  PROMs with high quality evidence for an insufficient measurement property
- **Reporting the Systematic Review**
  Reporting should be performed following PRISMA guidelines [11], and it is suggested it follows the flowchart that was presented initially (Fig. 4.9).

Table 4.9 Updated criteria for good measurement properties

Measurement property	Rating[1]	Criteria
Structural validity	+	**CTT:** CFA: CFI or TLI or comparable measure >0.95 OR RMSEA <0.06 OR SRMR <0.08**[2]** **IRT/Rasch:** No violation of unidimensionality**[3]**: CFI or TLI or comparable measure >0.95 OR RMSEA <0.06 OR SRMR <0.08 *AND* no violation of local independence: residual correlations among the items after controlling for the dominant factor < 0.20 OR Q3's < 0.37 *AND* no violation of monotonicity: adequate looking graphs OR item scalability >0.30 *AND* adequate model fit: IRT: χ^2 >0.01 Rasch: infit and outfit mean squares ≥ 0.5 and ≤ 1.5 OR Z-standardized values > -2 and <2
	?	CTI: Not all information for '+' reported IRT/Rasch: Model fit not reported
	–	Criteria for '+' not met
Internal consistency	+	At least low evidence[4] for sufficient structural validicy[5] AND Cronbach's alpha(s) ≥ 0.70 for each unidimensional scale or subscale[6]
	?	Criteria for "At least low evidence[4] for sufficient structural validity[5]" not met
	–	At least low evidence[4] for sufficient structural validicy[5] AND Cronbach's alpha(s) < 0.70 for each unidimensional scale or subscale[6]
Reliability	+	ICC or weighted Kappa ≥ 0.70
	?	ICC or weighted Kappa not reported
	–	ICC or weighted Kappa < 0.70
Measurement error	+	SDC or LoA < MIC[5]
	?	MIC not defined
	–	SDC or LoA > MIC[5]
Hypotheses testing for construct validity	+	The result is in accordance with the hypothesis[7]
	?	No hypothesis defined (by the review team)
	–	The result is not in accordance with the hypothesis[7]
Cross-cultural validity\measurement invariance	+	No important differences found between group factors (such as age, gender, language) in multiple group factor analysis OR no important DIF for group factors (McFadden's R^5 < 0.02)
	?	No multiple group factor analvsis OR DIF analvsis performed
	–	Important differences between group factors OR DIF was found
Criterion validity	+	Correlation with gold standard ≥ 0.70 ORAUC ≥ 0.70
	?	Not all information for '+' reported
	–	Correlation with gold standard < 0.70 OR AUC < 0.70
Responsiveness	+	The result is in accordance with the hypothesis[7] OR AUC ≥ 0.70
	?	No hypothesis defined (by the review team)
	–	The result is not in accordance with the hypothesis[7] OR AUC < 0.70

[1] "+" = sufficient, " -" = insufficient, "?" = indeterminate

[2] To rate the quality of the summary score, the factor structures should be equal across studies

[3] unidimensionality refers to a factor analysis per subscale, while structural validity refers to a factor analysis of a (multidimensional) patient-reported outcome measure

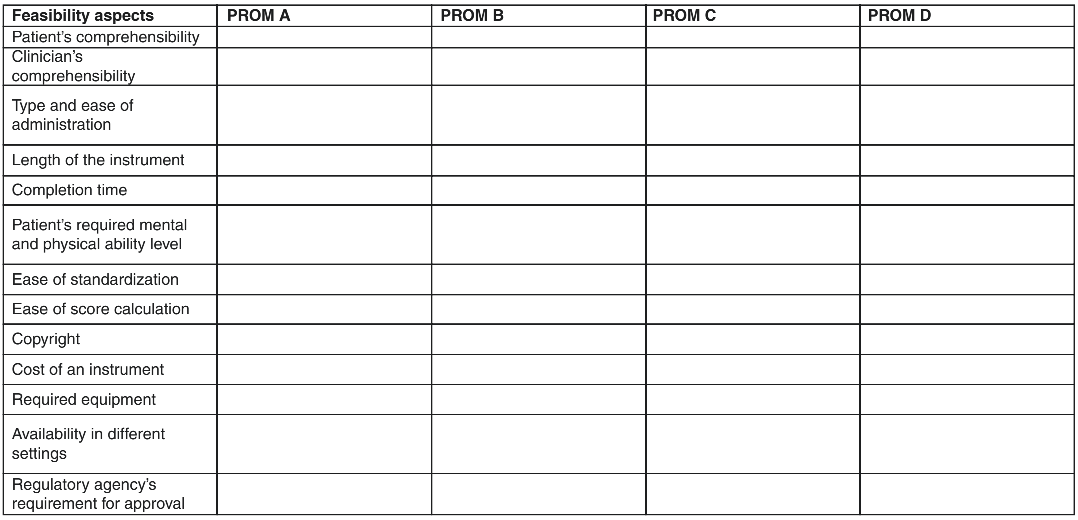
[4] As defined by grading the evidence according to the GRADE approach

[5] This evidence may come from different studies

[6] The criteria 'Cronbach alpha < 0.95' was deleted, as this is relevant in the development phase of a PROM and not when evaluating an existing PROM.

[7] The results of all studies should be taken together and it should then be decided if 75% of the results are in accordance with the hypotheses

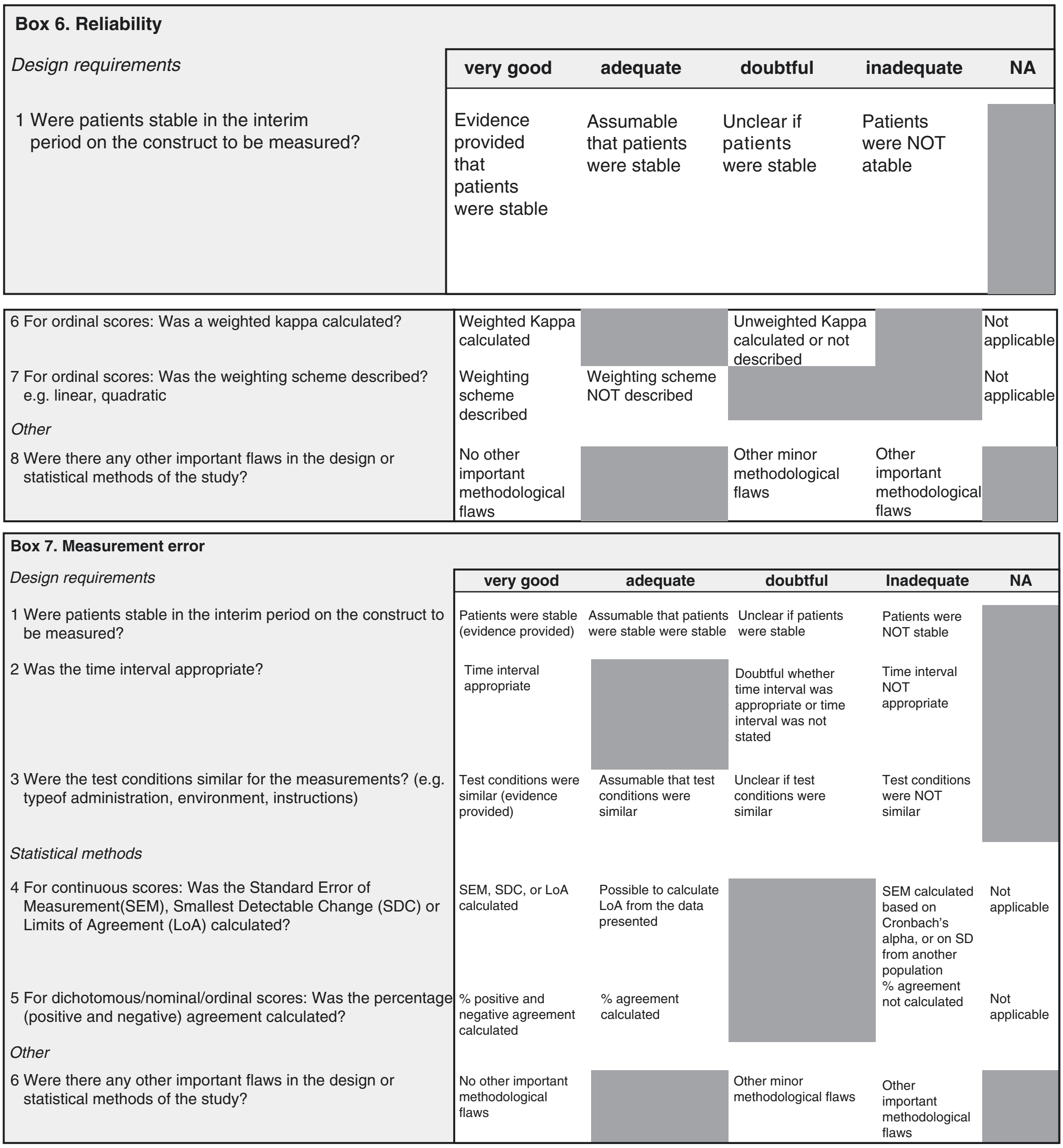
Lidwine B Mokkink, Cecilia AC Prinsen, Donald L Patrick, Jordi Alonso, Lex M Bouter, Henrica CW de Vet, Caroline B Terwee. COSMIN manual for systematic reviews of PROMs COSMIN methodology for systematic reviews of Patient-Reported Outcome Measures (PROMs) user manual. (2018)



Feasibility aspects	**PROM A**	**PROM B**	**PROM C**	**PROM D**
Patient’s comprehensibility				
Clinician’s comprehensibility				
Type and ease of administration				
Length of the instrument				
Completion time				
Patient’s required mental and physical ability level				
Ease of standardization				
Ease of score calculation				
Copyright				
Cost of an instrument				
Required equipment				
Availability in different settings				
Regulatory agency’s requirement for approval				

Fig. 4.8 Lidwine B Mokkink, Cecilia AC Prinsen, Donald L Patrick, Jordi Alonso, Lex M Bouter, Henrica CW de Vet, Caroline B Terwee. COSMIN manual for systematic reviews of PROMs COSMIN methodology for systematic reviews of Patient-Reported Outcome Measures (PROMs) user manual. (2018)

Table 4.10 Evaluation of reliability



Box 6. Reliability

Design requirements	very good	adequate	doubtful	inadequate	NA
1 Were patients stable in the interim period on the construct to be measured?	Evidence provided that patients were stable	Assumable that patients were stable	Unclear if patients were stable	Patients were NOT atable	
6 For ordinal scores: Was a weighted kappa calculated?	Weighted Kappa calculated		Unweighted Kappa calculated or not described		Not applicable
7 For ordinal scores: Was the weighting scheme described? e.g. linear, quadratic	Weighting scheme described	Weighting scheme NOT described			Not applicable
Other					
8 Were there any other important flaws in the design or statistical methods of the study?	No other important methodological flaws		Other minor methodological flaws	Other important methodological flaws	

Box 7. Measurement error

Design requirements	very good	adequate	doubtful	Inadequate	NA
1 Were patients stable in the interim period on the construct to be measured?	Patients were stable (evidence provided)	Assumable that patients were stable were stable	Unclear if patients were stable	Patients were NOT stable	
2 Was the time interval appropriate?	Time interval appropriate		Doubtful whether time interval was appropriate or time interval was not stated	Time interval NOT appropriate	
3 Were the test conditions similar for the measurements? (e.g. typeof administration, environment, instructions)	Test conditions were similar (evidence provided)	Assumable that test conditions were similar	Unclear if test conditions were similar	Test conditions were NOT similar	
Statistical methods					
4 For continuous scores: Was the Standard Error of Measurement(SEM), Smallest Detectable Change (SDC) or Limits of Agreement (LoA) calculated?	SEM, SDC, or LoA calculated	Possible to calculate LoA from the data presented		SEM calculated based on Cronbach's alpha, or on SD from another population	Not applicable
5 For dichotomous/nominal/ordinal scores: Was the percentage (positive and negative) agreement calculated?	% positive and negative agreement calculated	% agreement calculated		% agreement not calculated	Not applicable
Other					
6 Were there any other important flaws in the design or statistical methods of the study?	No other important methodological flaws		Other minor methodological flaws	Other important methodological flaws	

Caroline B Terwee et al. COSMIN methodology for assessing the content validity of PROMs. Available at: https://www.cosmin.nl/wp-content/uploads/COSMIN-methodology-for-content-validity-user-manual-v1.pdf

Table. 4.10 (continued)

2 Was the time interval appropriate?	Time interval appropriate		Doubtful whether time interval was appropriate or time interval was not stated	Time interval NOT appropriate	
3 Were the test conditions similar for the measurements? e.g. type of administration, environment, instructions	Test conditions were similar (evidence provided)	Assumable that test conditions were similar	Unclear if test conditions were similar	Test conditions were NOT similar	

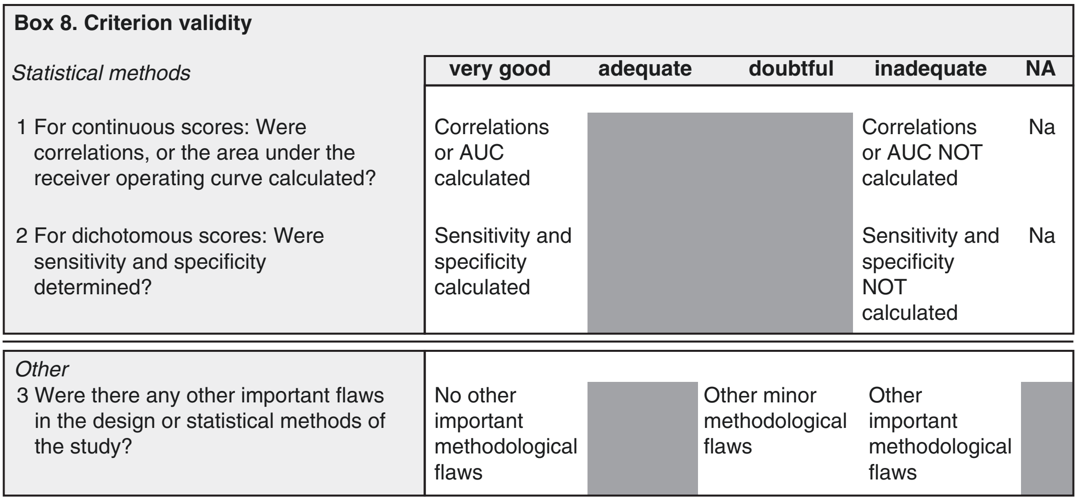
Statistical methods					
4 For continuous scores: Was an intraclass correlation coefficient (ICC) calculated?	ICC calculated and model or formula of the ICC is described	ICC calculated but model or formula of the ICC not described or not optimal. Pearson or Spearman correlation coefficient calculated with evidence provided that no systematic change has occurred	Pearson or Spearman correlation coefficient calculated WITHOUT evidence provided that no systematic change has occurred or WITH evidence that systematic change has occurred	No ICC or Pearson or Spearman correlations calculated	Na
5 For dichotomous/nominal/ordinal scores: Was kappa calculated?	Kappa calculated			No kappa calculated	Na
6 For ordinal scores: Was a weighted kappa calculated?	Weighted Kappa calculated		Unweighted Kappa calculated or not described		Na
7 For ordinal scores: Was the weighting scheme described? e.g. linear, quadratic	Weighting scheme described	Weighting scheme NOT described			Na

Table 4.11 Assessing risk of bias in a study on measurement error

Box 7. Measurement error					
Design requirements	**very good**	**adequate**	**doubtful**	**Inadequate**	**NA**
1 Were patients stable in the interim period on the construct to be measured?	Patients were stable (evidence provided)	Assumable that patients were stable	Unclear if patients were stable	Patients were NOT stable	
2 Was the time interval appropriate?	Time interval appropriate		Doubtful whether time interval was appropriate or time interval was not stated	Time interval NOT appropriate	
3 Were the test conditions similar for the measurements? (e.g. type of administration, environment, instructions)	Test conditions were similar (evidence provided)	Assumable that test conditions were similar	Unclear if test conditions were similar	Test conditions were NOT similar	
Statistical methods					
4 For continuous scores: Was the Standard Error of Measurement(SEM), Smallest Detectable Change (SDC) or Limits of Agreement (LoA) calculated?	SEM, SDC, or LoA calculated	Possible to calculate LoA from the data presented		SEM calculated based on Cronbach's alpha, or on SD from another population	Not applicable
5 For dichotomous/nominal/ ordinal scores: Was the percentage (positive and negative) agreement calculated?	% positive and negative agreement calculated	% agreement calculated		% agreement not calculated	Not applicable

Caroline B Terwee et al. COSMIN methodology for assessing the content validity of PROMs. Available at: https://www.cosmin.nl/wp-content/uploads/COSMIN-methodology-for-content-validity-user-manual-v1.pdf

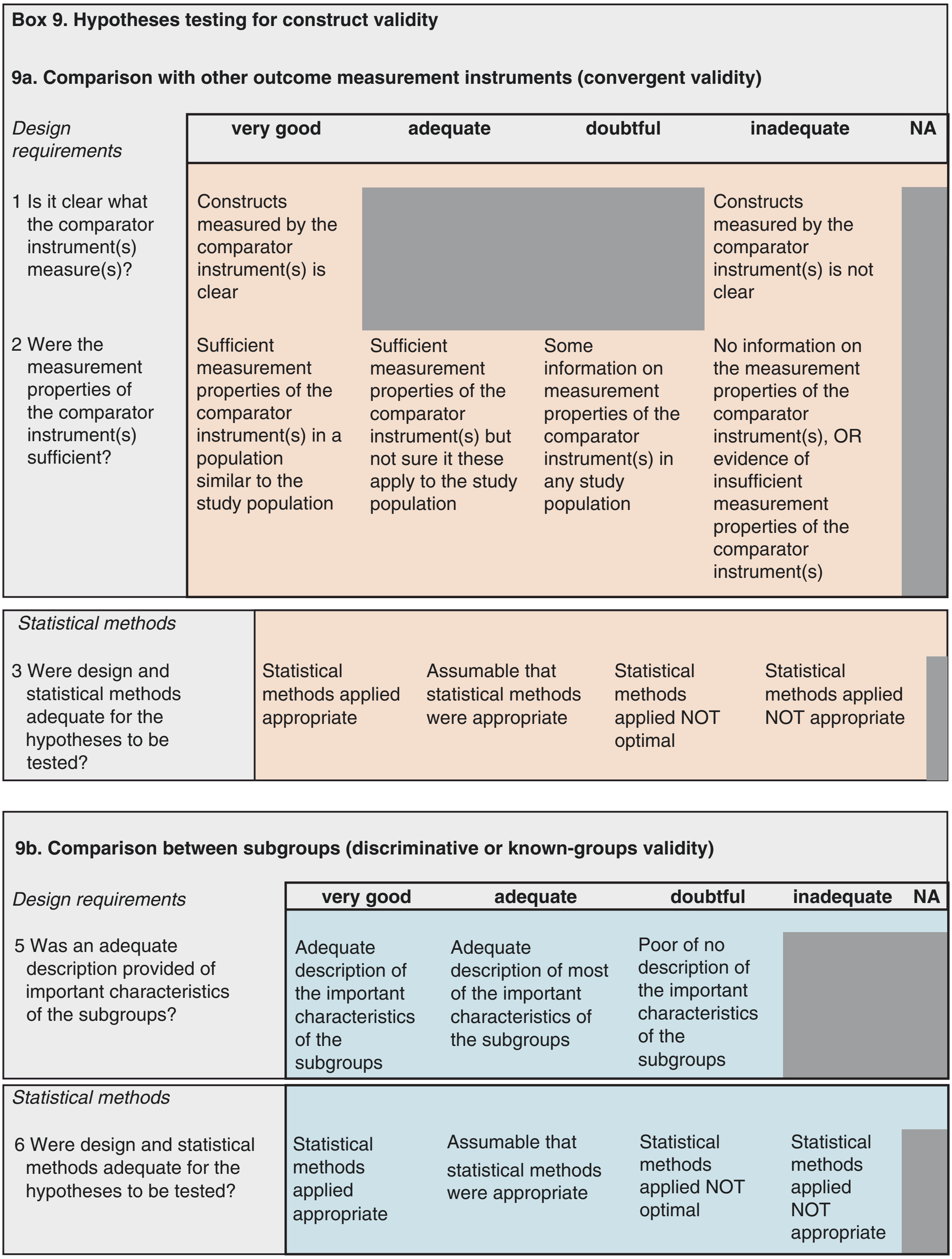
Table 4.12 Assessing risk of bias in a study on criterion validity



Box 8. Criterion validity					
Statistical methods	**very good**	**adequate**	**doubtful**	**inadequate**	**NA**
1 For continuous scores: Were correlations, or the area under the receiver operating curve calculated?	Correlations or AUC calculated			Correlations or AUC NOT calculated	Na
2 For dichotomous scores: Were sensitivity and specificity determined?	Sensitivity and specificity calculated			Sensitivity and specificity NOT calculated	Na
Other 3 Were there any other important flaws in the design or statistical methods of the study?	No other important methodological flaws		Other minor methodological flaws	Other important methodological flaws	

Caroline B Terwee et al. COSMIN methodology for assessing the content validity of PROMs. Available at: https://www.cosmin.nl/wp-content/uploads/COSMIN-methodology-for-content-validity-user-manual-v1.pdf

Table 4.13 Assessing risk of bias in a study on hypotheses testing for construct validity



Box 9. Hypotheses testing for construct validity

9a. Comparison with other outcome measurement instruments (convergent validity)

Design requirements	**very good**	**adequate**	**doubtful**	**inadequate**	**NA**
1 Is it clear what the comparator instrument(s) measure(s)?	Constructs measured by the comparator instrument(s) is clear			Constructs measured by the comparator instrument(s) is not clear	
2 Were the measurement properties of the comparator instrument(s) 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 it these apply to the study population	Some information 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)	
Statistical methods					
3 Were design and statistical methods adequate for the hypotheses to be tested?	Statistical methods applied appropriate	Assumable that statistical methods were appropriate	Statistical methods applied NOT optimal	Statistical methods applied NOT appropriate	

9b. Comparison between subgroups (discriminative or known-groups validity)

Design requirements	**very good**	**adequate**	**doubtful**	**inadequate**	**NA**
5 Was an adequate description provided of important characteristics of the subgroups?	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		
Statistical methods					
6 Were design and statistical methods adequate for the hypotheses to be tested?	Statistical methods applied appropriate	Assumable that statistical methods were appropriate	Statistical methods applied NOT optimal	Statistical methods applied NOT appropriate	

Caroline B Terwee et al. COSMIN methodology for assessing the content validity of PROMs. Available at: https://www.cosmin.nl/wp-content/uploads/COSMIN-methodology-for-content-validity-user-manual-v1.pdf

Table 4.14 Assessing risk of bias in a study on responsiveness

Box 10. Responsiveness

10a. Criterion approach (i.e. comparison to a gold standard)

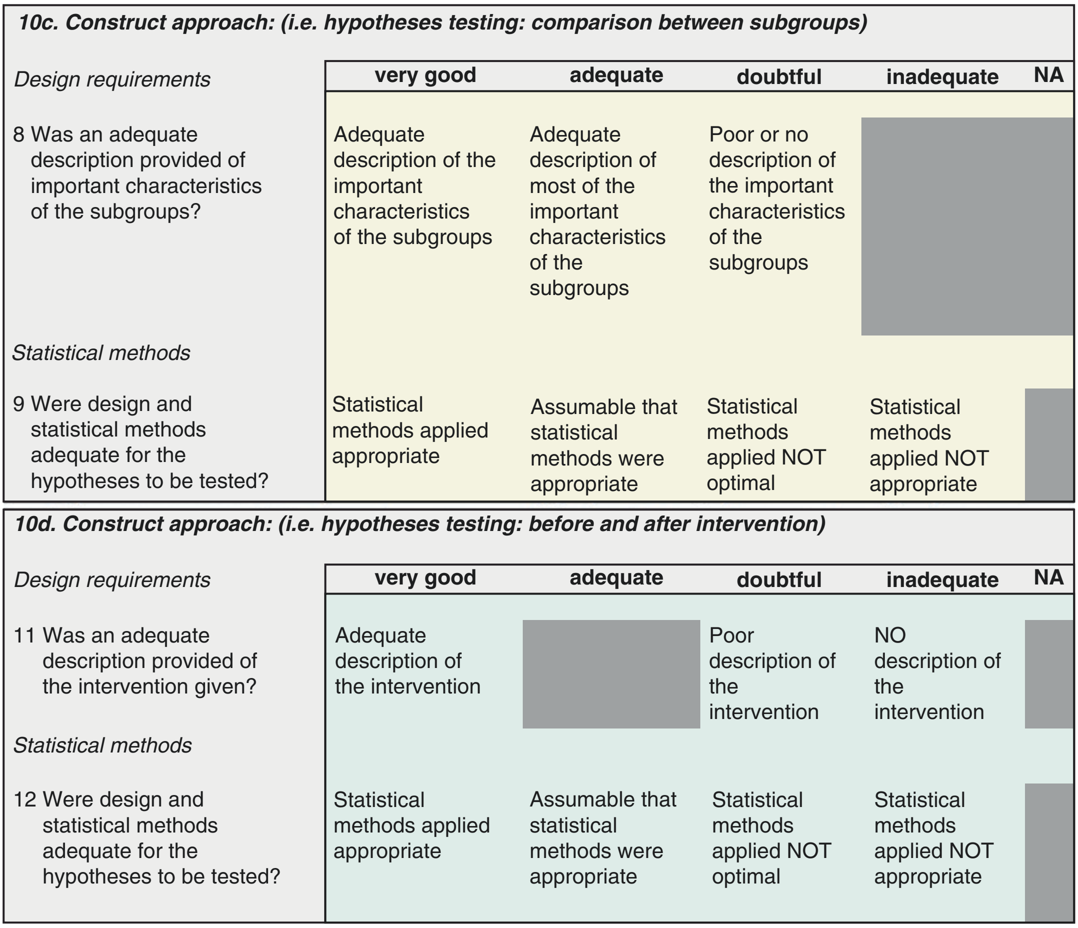
Statistical methods	very good	adequate	doubtful	inadequate	NA
1 For continuous scores: Were correlations between change scores, or the area under the Receiver Operator Curve (ROC) curve calculated?	Correlations or Are under the ROC Curve (AUC) calculated			Correlations or Are NOT calculated	na
2 For dichotomous scales: Were sensitivity and specificity (changed versus not changed) determined?	Sensitivity and specificity calculated			Sensitivity and specificity NOT calculated	na

10b. Construct approach (i.e. hypotheses testing; comparison with other outcome measurement instruments)

Design requirements	very good	adequate	doubtful	inadequate	NA
4 Is it clear what the comparator instrument(s) measure(s)?	Constructs measured by the comparator instrument(s) is clear			Constructs measured by the comparator instrument(s) is not clear	
5 Were the measurement properties of the comparator instrument(s) 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 of the comparator instrument(s) in any study population	NO information on the measurement properties of the comparator instruments(s) OR evidence of insufficient quality of comparator instruments(s)	
Statistical methods					
6 Were design and statistical methods adequate for the hypotheses to be tested?	Statistical methods applied appropriate	Assumable that statistical methods were appropriate	Statistical methods applied NOT optimal	Statistical methods applied NOT appropriate	
Other					
7 Were there any other important flaws in the design or statistical methods of the study?	No other important methodological flaws		Other minor methodological flaws	Other important methodological flaws	

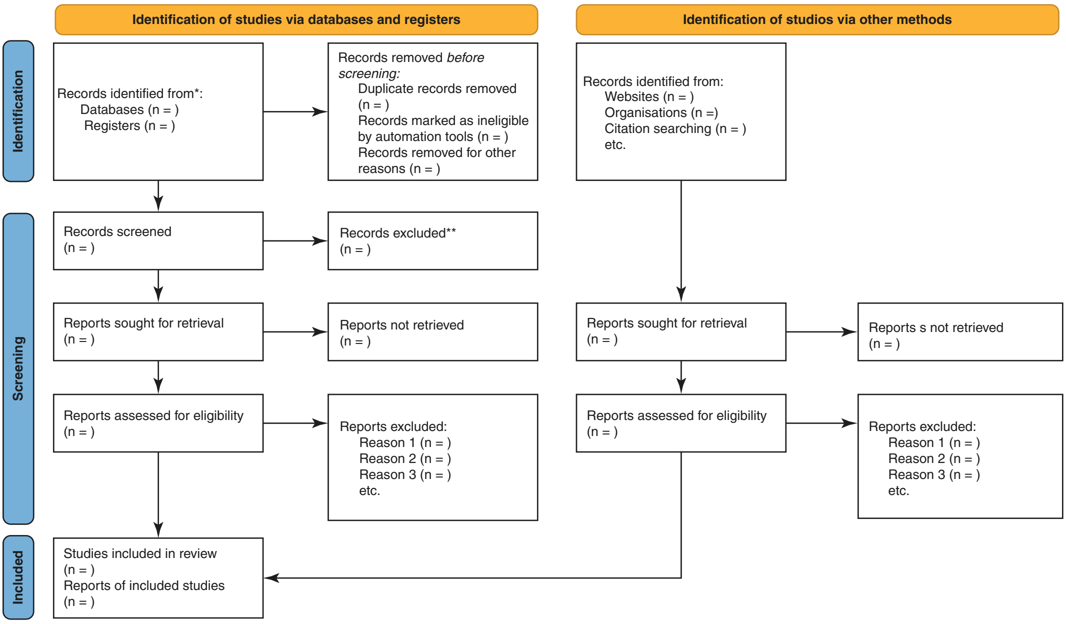
Caroline B Terwee et al. COSMIN methodology for assessing the content validity of PROMs. Available at: https://www.cosmin.nl/wp-content/uploads/COSMIN-methodology-for-content-validity-user-manual-v1.pdf

Table. 4.14 (continued)



10c. Construct approach: (i.e. hypotheses testing: comparison between subgroups)	very good	adequate	doubtful	inadequate	NA
Design requirements					
8 Was an adequate description provided of important characteristics of the subgroups?	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		
Statistical methods					
9 Were design and statistical methods adequate for the hypotheses to be tested?	Statistical methods applied appropriate	Assumable that statistical methods were appropriate	Statistical methods applied NOT optimal	Statistical methods applied NOT appropriate	

10d. Construct approach: (i.e. hypotheses testing: before and after intervention)	very good	adequate	doubtful	inadequate	NA
Design requirements					
11 Was an adequate description provided of the intervention given?	Adequate description of the intervention		Poor description of the intervention	NO description of the intervention	
Statistical methods					
12 Were design and statistical methods adequate for the hypotheses to be tested?	Statistical methods applied appropriate	Assumable that statistical methods were appropriate	Statistical methods applied NOT optimal	Statistical methods applied NOT appropriate	


Identification of studies via databases and registers
Identification of studios via other methods
Identification
Screening
Included
Records identified from*:
Databases (n =)
Registers (n =)
Records removed *before screening:*
Duplicate records removed (n =)
Records marked as ineligible by automation tools (n =)
Records removed for other reasons (n =)
Records identified from:
Websites (n =)
Organisations (n =)
Citation searching (n =)
etc.
Records screened
(n =)
Records excluded**
(n =)
Reports sought for retrieval
(n =)
Reports not retrieved
(n =)
Reports sought for retrieval
(n =)
Reports s not retrieved
(n =)
Reports assessed for eligibility
(n =)
Reports excluded:
Reason 1 (n =)
Reason 2 (n =)
Reason 3 (n =)
etc.
Reports assessed for eligibility
(n =)
Reports excluded:
Reason 1 (n =)
Reason 2 (n =)
Reason 3 (n =)
etc.
Studies included in review
(n =)
Reports of included studies
(n =)

*Consider, it feasible to do so, reporting the number of records identified from each database or register searcned (rather than the total number across all databases/registers).
**It automation tools were used, indicate how many records were excluded by a human and how many were excluded by automaton tools.

Fig. 4.9 PRISMA 2020 flow diagram for new systematic reviews which included searches of databases, registers and other sources. Page, M. J. et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 372, (2021)

Limitations and Considerations

We have chosen to present the COSMIN methodology as a roadmap for performing systematic reviews on measurement properties of PROMs, mainly due to the structured approach and detailed recommended process.

Researchers that are interested in performing a systematic review on measurement properties of PROMs, need to be aware of potential limitations, prior committing to following this methodology.

On a recent article by McKenna and Heaney, several points have been raised and we consider it useful to briefly mention them here [12].

According to this, the authors claim that there is lack of evidence to support the COSMIN recommendations. It is discussed that the guidelines have been produced based on empirical evidence, and the experience of the COSMIN steering committee.

In addition to that, while performing Delphi studies to agree and produce recommendations in a scientifically robust manner, there may be concerns about the inclusivity of the participating professionals.

A further point raised, concerns the omission of several aspects in the assessment of the PROM, that the authors consider significant, such as the construct theories, the fundamental measurements, unidimensionality, item generation and reduction.

Moreover, it is identified that there has been no actual evaluation of the COSMIN guidelines themselves. As an overall concept, the critique concludes that the COSMIN guidelines and recommendations are not evidence-based.

Lastly, the most significant point relates to who utilises and attempts to follow the COSMIN methodology.

As the vast majority of the researchers performing these reviews are clinicians, and given the complexity of the COSMIN guidance, it may be extracted that they lack the necessary expertise and ability to interpret and evaluate the relevant information, hence producing inaccurate reviews and recommendations.

Overall, we feel that through this chapter, a researcher may be introduced to the basics of performing systematic reviews on measurement properties of PROMs, and the COSMIN methodology and guidelines can be used as they introduce a step-wise approach and thorough approach.

Nevertheless, the limitations discussed bear some value—particularly with regards to the researcher's expertise and background in the field. These should be meticulously taken into account, and the research team should certainly consider the involvement of professionals with a strong background in measurement, psychometrics, statistics and health-related quality of life research.

References

1. Chapter 18: Patient-reported outcomes. Cochrane training. Available at https://training.cochrane.org/handbook/archive/v6/chapter-18. Accessed 4 Oct 2022.
2. COSMIN. Improving the selection of outcome measurement instruments. Available at https://www.cosmin.nl/. Accessed 4 Oct 2022.
3. Mokkink LB, et al. Evaluation of the methodological quality of systematic reviews of health status measurement instruments. Qual Life Res. 2009;18:313–33.
4. Mokkink LB, 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.
5. Prinsen CAC, et al. COSMIN guideline for systematic reviews of patient-reported outcome measures. Qual Life Res. 2018;27:1147–57.
6. Mokkink LB, Prinsen CAC, Patrick DL, Alonso J, Bouter LM, de Vet HCW, Terwee CB. COSMIN manual for systematic reviews of PROMs COSMIN methodology for systematic reviews of Patient-Reported Outcome Measures (PROMs) user manual. 2018.
7. Terwee CB, et al. COSMIN methodology for assessing the content validity of PROMs. Available at https://www.cosmin.nl/wp-content/uploads/COSMIN-methodology-for-content-validity-user-manual-v1.pdf. Accessed 4 Oct 2022.
8. Terwee CB, et al. COSMIN methodology for evaluating the content validity of patient-reported outcome measures: a Delphi study. Qual Life Res. 2018;27:1159–70.
9. GRADE Home. Available at https://www.gradeworkinggroup.org/. Accessed 4 Oct 2022.
10. Mokkink LB. COSMIN Risk of Bias checklist [PDF File]. The Amsterdam Public Health Research Institute; 2018. p. 1–37.
11. Page MJ, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ. 2021;372:n71.
12. McKenna SP, Heaney A. Setting and maintaining standards for patient-reported outcome measures: can we rely on the COSMIN checklists? 2021;24:502–11. https://doi.org/10.1080/13696998.2021.1907092.