COSMIN methodology for systematic reviews of Patient-Reported Outcome Measures (PROMs)

user manual

Version 1.0 dated February 2018

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

Contact
LB Mokkink, PhD
VU University Medical Center
Department of Epidemiology and Biostatistics
Amsterdam Public Health research institute
1081 BT Amsterdam
The Netherlands
Website: www.cosmin.nl
E-mail: w.mokkink@vumc.nl

The development of the COSMIN guideline for systematic reviews of Patient-Reported Outcome Measures (PROMs) and the COSMIN Risk of Bias checklist for systematic reviews of PROMs was financially supported by the Amsterdam Public Health research institute, VU University Medical Center, Amsterdam, the Netherlands.

Table of contents

Foreword

Research performed with outcome measurement instruments of poor or unknown quality constitutes a waste of resources and is unethical (3). Unfortunately this practice is widespread (4). Selecting the best outcome measurement instrument for the outcome of interest in a methodologically sound way requires: (1) high quality studies that document the evaluation of the measurement properties (in total nine different aspects of reliability, validity, and responsiveness) of relevant outcome measurement instruments in the target population; and (2) a high quality systematic review of studies on measurement properties in which all information is gathered and evaluated in a systematic and transparent way, accompanied by clear recommendations for the most suitable available outcome measurement instrument. However, conducting such a systematic review is quite complex and time consuming, and it requires expertise within the research team on the construct to be measured, on the patient population, and on the methodology of studies of measurement properties.

High quality systematic reviews can provide a comprehensive overview of the measurement properties of Patient-Reported Outcome Measures (PROMs) and supports evidence-based recommendations in the selection of the most suitable PROM for a given purpose. For example, for selecting the most suitable PROM for an outcome included in a core outcome set (COS)(5). These systematic reviews can also identify gaps in knowledge on the measurement properties of PROMs, which can be used to design new studies on measurement properties.

The COSMIN (COnsensus-based Standards for the selection of health Measurement INstruments) initiative aims to improve the selection of outcome measurement instruments in research and clinical practice by developing tools for selecting the most suitable instrument for the situation at issue.
Recently, a comprehensive methodological guideline for systematic reviews of PROMs was developed by the COSMIN initiative(2). In addition, the COSMIN checklist was adapted for assessing the risk of bias in studies on measurement properties in systematic reviews of PROMs(6). Also, a Delphi study was performed to develop standards and criteria for assessing the content validity of PROMs(7).

The present manual is a supplement to the COSMIN guideline for systematic reviews of measurement instruments (2) including the use of the COSMIN Risk of Bias checklist(6). This manual contains additional detailed information on how to perform each of the proposed ten steps to conduct a systematic review of PROMs, on how to use the COSMIN Risk of Bias checklist, and on how to come to recommendations on the most suitable instrument for a given.

In the COSMIN methodology we use the word patient. However, sometimes the target population of the systematic review or the PROM is not patients, but e.g. healthy individuals (e.g. for generic PROMs) or caregivers (when a PROM measures caregiver burden). In these cases, the word patient should be read as e.g. healthy person or caregiver.

Throughout the manual we provide some examples to explain the COSMIN Risk of Bias checklist. We would like to emphasize that these are arbitrarily chosen and used for illustrative purposes.

The COSMIN methodology focusses on PROMs used as outcome measurement instruments (i.e. evaluative application). The methodology can also be used for other types of measurement instruments (like clinician-reported outcome measures or performance-based outcome measures), or other applications (e.g. diagnostic or predictive applications), but the methodology may need to be adapted for these other purposes. For example, structural validity may not be relevant for some types of instruments and responsiveness is less relevant when an instrument is used as a diagnostic instrument. New standards for assessing the quality of a study on the reliability of a MRI scan should be developed.

This manual aims to facilitate the understanding and practice performance of a systematic review on PROMs. We give detailed instructions on each step to take, and we will provide many examples and suggestions on how to perform each step.

We aim to continue updating this manual when deemed necessary, based on experience and suggestions by users of the COSMIN methodology. If you have any suggestions or questions, please contact us (w.mokkink@vumc.nl).

1. Background information

1.1 The COSMIN initiative

The COSMIN initiative aims to improve the selection of health measurement instruments both in research and clinical practice by developing tools for selecting the most suitable instrument for a given situation. COSMIN is an international initiative consisting of a multidisciplinary team of researchers with expertise in epidemiology, psychometrics, and qualitative research, and in the development and evaluation of outcome measurement instruments in the field of health care, as well as in performing systematic reviews of outcome measurement instruments.

COSMIN steering committee

Lidwine B Mokkink[1]
Cecilia AC Prinsen[1]
Donald L Patrick[2]
Jordi Alonso[3]
Lex M Bouter[1]
Henrica CW de Vet[1]
Caroline B Terwee[1]

[1] Department of Epidemiology and Biostatistics, Amsterdam Public Health research institute, VU University Medical Center, De Boelelaan 1089a, 1081 HV, Amsterdam, the Netherlands;

[2] Department of Health Services, University of Washington, Thur Canal St Research Office, 146 N Canal Suite 310, 98103, Seattle, USA;

[3] IMIM (Hospital del Mar Medical Research Institute), CIBER Epidemiology and Public Health (CIBERESP), Dept. Experimental and Health Sciences, Pompeu Fabra University (UPF), Doctor Aiguader 88, 08003, Barcelona, Spain.

1.2 How to cite this manual

This manual was based on three articles, published in peer-reviewed scientific journals. If you use the COSMIN methodology, please refer to these articles:

Prinsen, C. A.C., Mokkink, L. B., Bouter, L. M., Alonso, J., Patrick, D. L., De Vet, H. C., et al. (2018). COSMIN guideline for systematic reviews of Patient-Reported Outcome Measures. Qual Life Res, accept.

Mokkink, L.B., de Vet, H.C.W., Prinsen, C.A.C., Patrick, D.L., Alonso, J., Bouter, L.M., Terwee, C.B. (2017). COSMIN Risk of Bias checklist for systematic reviews of Patient-Reported Outcome Measures. Qual Life Res. doi: 10.1007/s11136-017-1765-4. [Epub ahead of print].

Terwee CB, Prinsen CAC, Chiarotto A, Westerman MJ, Patrick DL, Alonso J, Bouter LM, de Vet HCW, Mokkink LB. COSMIN methodology for evaluating the content validity of Patient-Reported Outcome Measures: a Delphi study, submitted.

1.3 Development of a comprehensive COSMIN guideline for systematic reviews of Patient-Reported Outcome Measures (PROMs)

In the absence of empirical evidence, the COSMIN guideline for systematic reviews of PROMs (2) was based on our experience that we (that is: the COSMIN steering committee) have gained over the past years in conducting systematic reviews of PROMs (8, 9), in supporting other systematic reviewers in their work (10, 11) and in the development of COSMIN methodology (12, 13). In addition, we have studied the quality of systematic reviews of PROMs in two consecutive reviews (14, 15), and in reviews that have used the COSMIN methodology we have specifically searched for the comments made by review authors relating to the COSMIN methodology. Further, we have had iterative discussions by the COSMIN steering committee, both at face-to-face meetings (CP, WM, HdV and CT) and by email. We gained experience from results of a recent Delphi study on the content validity of PROMs (7), and from results of a previous Delphi study on the selection of outcome measurement instruments for outcomes included in core outcome sets (COS) (5).
Further, the guideline was developed in concordance with existing guidelines for reviews, such as the Cochrane Handbook for systematic reviews of interventions (16), and for diagnostic test accuracy reviews (17), the PRISMA Statement(18), the Institute of Medicine (IOM) standards for systematic reviews of comparative effectiveness research(19), and the Grading of Recommendations Assessment, Development and Evaluation (GRADE) principles (20).

This guideline focuses on the methodology of systematic reviews of existing PROMs, for which at least a PROM development study or some information on its measurement properties is available, and that are used for evaluative purposes (e.g. to measure the effects of treatment or to monitor patients over time). Instruments that are being used for discriminative or predictive purposes are not being considered in the guideline. The methods can also be used for systematic reviews of other types of outcome measurement instruments such as clinician reported outcome measurement

instruments (ClinROMs) or performance-based outcome measurement instruments (PerBOMs), for reviews of one single outcome measurement instrument, or for a predefined set of outcome measurement instruments. However, some adaptations may be needed in some of the steps or in the COSMIN Risk of Bias checklist.

1.4 COSMIN taxonomy and definitions

In the literature different terminology and definitions for measurement properties are continuously being used. The COSMIN initiative developed a taxonomy of measurement properties relevant for evaluating PROMs. In the first COSMIN Delphi study, conducted in 2006-2007, consensus was reached on terminology and definitions of all included measurement properties in the COSMIN checklist(21). This taxonomy (Figure 1) formed the foundation on which the guideline was based.

Figure 1. The COSMIN taxonomy(21)

Taxonomy of measurement properties

The COSMIN taxonomy of measurement properties is presented in Figure 1. It was decided that all measurement properties included in the taxonomy are relevant and should be evaluated for PROMs used in an evaluative application.

In assessing the quality of a PROM we distinguish three domains, i.e. reliability, validity, and responsiveness. Each domain contains one or more measurement properties, i.e. quality aspects of measurement instruments. The domain reliability contains three measurement properties: internal consistency, reliability, and measurement error. The domain validity also contains three measurement properties: content validity (including face validity), structural validity, hypotheses testing for construct validity, cross-cultural validity and criterion validity. The domain responsiveness contains only one measurement property, which is also called responsiveness.

Definitions of measurement properties

Consensus-based definitions of all included measurement properties in the COSMIN checklist are presented in Table 1.

Table 1. COSMIN definitions of domains, measurement properties, and aspects of measurement properties(21)

Term			Definition
Domain	**Measurement property**	**Aspect of a measurement property**	
Reliability			The degree to which the measurement is free from measurement error
Reliability (extended definition)			The extent to which scores for patients who have not changed are the same for repeated measurement under several conditions: e.g. using different sets of items from the same PROM (internal consistency); over time (test-retest); by different persons on the same occasion (inter-rater); or by the same persons (i.e. raters or responders) on different occasions (intra-rater)
	Internal consistency		The degree of the interrelatedness among the items
	Reliability		The proportion of the total variance in the measurements which is due to 'true'† differences between patients
	Measurement error		The systematic and random error of a patient's score that is not attributed to true changes in the construct to be measured
Validity			The degree to which a PROM measures the construct(s) it purports to measure
	Content validity		The degree to which the content of a PROM is an adequate reflection of the construct to be measured
		Face validity	The degree to which (the items of) a PROM indeed looks as though they are an adequate reflection of the construct to be measured

	Construct validity		The degree to which the scores of a PROM are consistent with hypotheses *(for instance with regard to internal relationships, relationships to scores of other instruments, or differences between relevant groups)* based on the assumption that the PROM validly measures the construct to be measured
		Structural validity	The degree to which the scores of a PROM are an adequate reflection of the dimensionality of the construct to be measured
		Hypotheses testing	Idem construct validity
		Cross-cultural validity	The degree to which the performance of the items on a translated or culturally adapted PROM are an adequate reflection of the performance of the items of the original version of the PROM
	Criterion validity		The degree to which the scores of a PROM are an adequate reflection of a 'gold standard'
Responsiveness			The ability of a PROM to detect change over time in the construct to be measured
	Responsiveness		Idem responsiveness
Interpretability*			Interpretability is the degree to which one can assign qualitative meaning - that is, clinical or commonly understood connotations – to a PROM's quantitative scores or change in scores.

† *The word 'true' must be seen in the context of the CTT, which states that any observation is composed of two components – a true score and error associated with the observation. 'True' is the average score that would be obtained if the scale were given an infinite number of times. It refers only to the consistency of the score, and not to its accuracy (22)*
* *Interpretability is not considered a measurement property, but an important characteristic of a measurement instrument*

1.5 The COSMIN Risk of Bias Checklist

It was decided to create three separate versions of the original COSMIN checklist (1) to be used for different purposes when assessing the methodological quality of studies on measurement properties, i.e. the COSMIN Design checklist, the COSMIN Risk of Bias checklist, and the COSMIN Reporting checklist. The Risk of Bias checklist was exclusively developed for assessing the methodological quality of single studies included in systematic reviews of PROMs. The purpose of assessing the methodological quality of a study in a systematic review is to screen for risk of bias in the included studies. The term 'risk of bias' is in compliance with the Cochrane methodology for systematic reviews of trials and diagnostic studies(16). It refers to whether the results based on the methodological quality of the study are trustworthy.
The checklist contains standards referring to design requirements and preferred statistical methods of studies on measurement properties. For each measurement property a COSMIN box was developed containing all standards needed to assess the quality of a study on that specific measurement property. Both preferred statistical methods based on Classical test Theory (CTT) and Item Response Theory (IRT) or Rasch analyses are included in the standards.

COSMIN standards and criteria

It is important to note that COSMIN makes a distinction between "standards" and "criteria": **Standards** refer to design requirements and preferred statistical methods for evaluating the methodological **quality of studies** on measurement properties. **Criteria** refer to what constitutes good measurement properties (**quality of PROMs**). In the first COSMIN Delphi study (1), only standards were developed for evaluating the quality of studies on measurement properties. Criteria for what constitutes good measurement properties were not developed. However, such criteria are needed in systematic reviews to provide evidence-based recommendations for which PROMs are good enough to be used in research or clinical practice. Therefore, criteria were developed for good measurement properties (Table 4) (2).

Boxes from the checklist
The COSMIN Risk of Bias checklist contains ten boxes with standards for PROM development (box 1) and for nine measurement properties: Content validity (box 2), Structural validity (box 3), Internal consistency (box 4), Cross-cultural validity\measurement invariance (box 5), Reliability (box 6), Measurement error (box 7), Criterion validity (box 8), Hypotheses testing for construct validity (box 9), and Responsiveness (box 10).

Modular tool

In one article one or more studies (often on different measurement properties) can be described. The methodological quality of each study should be assessed separately by rating all standards included in the corresponding box. Therefore, the COSMIN checklist should be used as a modular tool. This means that it may not be necessary to complete the whole checklist when evaluating the quality of studies described in an article. In accordance with the COSMIN taxonomy(21), the measurement properties evaluated in an article determine which boxes need to be completed. Each assessment of a measurement property is considered to be a separate study. For example, if in an article the internal consistency and construct validity of an instrument were assessed, only two boxes need to be completed, i.e. box 4 Internal consistency and box 9 Hypotheses testing for construct validity. This modular system was developed because not all measurement properties are assessed in all articles.

COSMIN considers each subscale of a (multi-dimensional) PROM separately

The measurement properties of a PROM should be rated separately for each set of items that make up a score. This can be a single item if this item is used as a standalone score, a set of items making up a subscale score within a multi-dimensional PROM, or a total PROM score if all items of a PROM are being summarized into a total score. Each score is assumed to represent a construct and is therefore considered a separate PROM. **In the remaining of this manual when we refer to a PROM, we mean a PROM score or subscore.**
For example, if a multidimensional PROM consists of three subscales and one single item, each scored separately, the measurement properties of the three subscales and the single item need to be rated separately. If the subscale scores are also summarized into a total score, the measurement properties of the total score should also be rated separately.

Four-point rating system

For each study, an overall judgement is needed on the quality of the particular study. We therefore developed a four-point rating system where each standard within a COSMIN box can be rated as 'very good', 'adequate', 'doubtful' or 'inadequate' (6). The overall rating of the quality of each study is determined by taking the lowest rating of any standard in the box (i.e. "the worst score counts" principle) (12). This overall rating of the quality of the studies can be used in grading the quality of the evidence (see Chapter 3), taking into account that results of 'inadequate' quality studies will decrease the trust one can put on the results and the overall conclusion about a measurement property of a PROM (as used in a specific context).

Each box contains a standard asking if there were any other important methodological flaws that are not covered by the checklist, but that may lead to biased results or conclusions. For example, in a reliability study, the administrations should be independent. Independent administrations imply that the first administration has not influenced the second administration, i.e. at the second administration the patient should not have been aware of the scores on the first administration (recall bias).

1.6 What has changed in the COSMIN methodology?

Inadequate studies are no longer ignored
In our previous protocol for performing systematic reviews of PROMs, we recommended to ignore the results of inadequate (previously called 'poor') quality studies because the results of these studies might be biased. In the current methodology, we recommend that results from inadequate studies may be included when pooling or summarizing the results from all studies is possible, or if the results of inadequate studies are consistent with the results from studies of better quality. The reasoning behind this is that if the results from inadequate studies are included, it should be considered to downgrade the quality of the evidence because of risk of bias (see Chapter 3). Herewith, the current methodology is more in line with recommendations from the Cochrane Collaboration for systematic reviews of intervention studies (16) and of diagnostic test accuracy studies (17).

Removing standards about a reasonable gold standard for criterion validity and responsiveness
We decided to delete the standards about deciding whether a gold standard used can be considered a reasonable gold standard. We now recommend the review team to determine before assessing the methodological quality of studies, which outcome measurement instrument can be considered a reasonable gold standard. When an included study uses this particular gold standard instrument to asses validity, the study can be considered as a study on criterion validity. The COSMIN panel reached consensus that no gold standard exist for PROMs (13). The only exception is when a shortened instrument is compared to the original long version. In that case, the original long version can be considered the gold standard. Often, authors consider their comparator instrument incorrectly a gold standard, for example when they compare the scores of a new instrument to a widely used and well-known instrument. In this situation, the study is considered a study on construct validity and box 9 Hypotheses testing for construct validity should be completed.

Removing standards on formulating hypotheses for hypotheses testing for construct validity and responsiveness
We decided to delete all standards about formulating hypotheses a priori from the boxes Hypotheses testing for construct validity and Responsiveness. Although we consider it majorly important to define hypotheses in advance when assessing construct validity or responsiveness of a PROM, results of studies without these hypotheses can –in many cases– still be used in a systematic review on PROMs because the presented correlations or mean differences between (sub)groups are not necessarily biased and thus can be evaluated. The conclusions of the authors though may be biased when a priori hypotheses are lacking. We recommend that the review team formulates a set of hypotheses about the expected direction and magnitude of correlations between the PROM of interest and other PROMs and of mean differences in scores between groups (23). This way, all results are compared to the same relevant hypothesis. If construct validity studies do include hypotheses, the review team can adopt these hypotheses if they consider them adequate. This way, the results from many studies can still be used in the systematic review as studies without hypotheses will no longer receive a 'inadequate' (previously called 'poor') quality rating. A detailed explanation for completing these boxes can be found in Chapter 5.

Removal of standards on sample size

We decided to remove the standard about adequate sample size for single studies from those boxes where it is possible to pool the results (i.e. the boxes Internal consistency, Reliability, Measurement error, Criterion validity, Hypotheses testing for construct validity, and Responsiveness) to a later phase of the review, i.e. when drawing conclusions across studies on the measurement properties of the PROM (Chapter 3). This was decided because several small high-quality studies can together provide sufficient evidence for the measurement property. Therefore, we recommend to take the aggregated sample size of the available studies into account when assessing the overall quality of evidence for a measurement property in a systematic review. This is in compliance with Cochrane Handbooks (16, 17). However, the standard about adequate sample size for single studies was maintained in the boxes Content validity, Structural validity, and Cross-cultural validity\measurement invariance, because the results of these studies cannot be pooled. In these boxes factor analyses, IRT or Rasch analyses are included as preferred statistical methods and these methods require sample sizes that are sufficiently large to obtain reliable results. The suggested sample size requirements should be considered as general guidance; in some situations, dependent on type of model, number of factors or items, more nuanced criteria might be applied.

Removal of standards on missing data and handling missing data

Each box of the original COSMIN checklist, except for box on content validity, contains standards about whether the percentage missing items was reported, and how these missing items were handled. Although we consider information on missing items very important to report, we decided to remove these standards from all boxes in the COSMIN Risk of Bias checklist, as it was agreed upon within the COSMIN steering committee that lack of reporting on number of missing items and on how missing items were handled would not necessarily lead to biased results of the study. Furthermore, at the moment there is little evidence and no consensus yet about what the best way is to handle missing items in studies on measurement properties.

New order of evaluating the measurement properties

A new order of evaluating the measurement properties is proposed, as shown in Table 2. Content validity is considered to be the most important measurement property because first of all it should be clear that the items of the PROM are relevant, comprehensive, and comprehensible with respect to the construct of interest and target population (2). Therefore, we recommend to first evaluate the development and content validity studies of the PROMs. PROMs with high quality evidence of inadequate content validity can be excluded from further assessment in the systematic review.

Next, we recommend to evaluate the internal structure of PROMs, including the measurement properties structural validity, internal consistency, and cross-cultural validity\measurement invariance. Internal structure refers to how the different items in a PROM are related, which is important to know for deciding how items might be combined into a scale or subscale. Evaluating the internal structure of the instrument is relevant for PROMs that are based on a reflective model. In a reflective model the construct manifests itself in the items, i.e. the items are a reflection of the construct to be measured (24). Finally, the remaining measurement properties are considered, i.e. reliability, measurement error, criterion validity, hypotheses testing for construct validity and responsiveness.

Table 2. Order in which the measurement properties are assessed

Content validity
1. PROM development*
2. Content validity
Internal structure
3. Structural validity
4. Internal consistency
5. Cross-cultural validity\measurement invariance
Remaining measurement properties
6. Reliability
7. Measurement error
8. Criterion validity
9. Hypotheses testing for construct validity
10. Responsiveness

* not a measurement property, but taken into account when evaluating content validity

The COSMIN methodology is specifically developed for use in reviews of outcome measurement instruments. The proposed order of evaluating the measurement properties is therefore based on use/application as an outcome measure, i.e. in an evaluative purpose. This order is also reflected in the ordering of boxes in the COSMIN Risk of Bias checklist. We think that some of the included measurement properties are less relevant for other purposes. For example, responsiveness is less relevant when an instrument is used as a diagnostic tool.

More information on removal of other standards and adaptations to individual standards and how these changes were decided upon, can be found elsewhere (6).

1.7 Ten-step procedure & outline of the manual

In a systematic review of PROMs an overview is given on available evidence of each measurement property of each included PROM to come to overall conclusions per measurement property and to give recommendation for the most suitable PROM for a given purpose. A guideline, consisting of a sequential ten-step procedure was developed by the COSMIN steering committee, subdivided into three parts: A, B and C (Figure 2) (2).

Part A consists of steps 1-4 and generally, these steps are standard procedures when performing systematic reviews, and are in agreement with existing guidelines for reviews (16, 17): preparing and performing the literature search, and selecting relevant publications. The methodology of part A is presented in Chapter 2.

Part B consists of steps 5-7 and concerns the evaluation of the measurement properties of the included PROMs. These steps were particularly developed for systematic reviews of PROMs. Part B is described in Chapter 3. We first describe the general methodology, and next, we explain step 5 (evaluation of content validity), step 6 (evaluation of internal structure), and step 7 (evaluation of the remaining measurement properties) in more detail.

Part C consists of steps 8-10 and concerns the evaluation of the interpretability and feasibility of the PROMs (step 8), formulating (step 9) and the reporting of the systematic review (step 10). Part C is described in Chapter 4.

In Chapter 5 we elaborate on how to rate each standard included in the COSMIN Risk of Bias checklist. We close the manual with providing examples of a search strategy, a flowchart, and all tables that need to be presented in a review (appendices 1-8).

We aim to continue updating this manual when deemed necessary, based on experience and suggestions by users of the COSMIN methodology. If you have any suggestions or questions, please contact us.

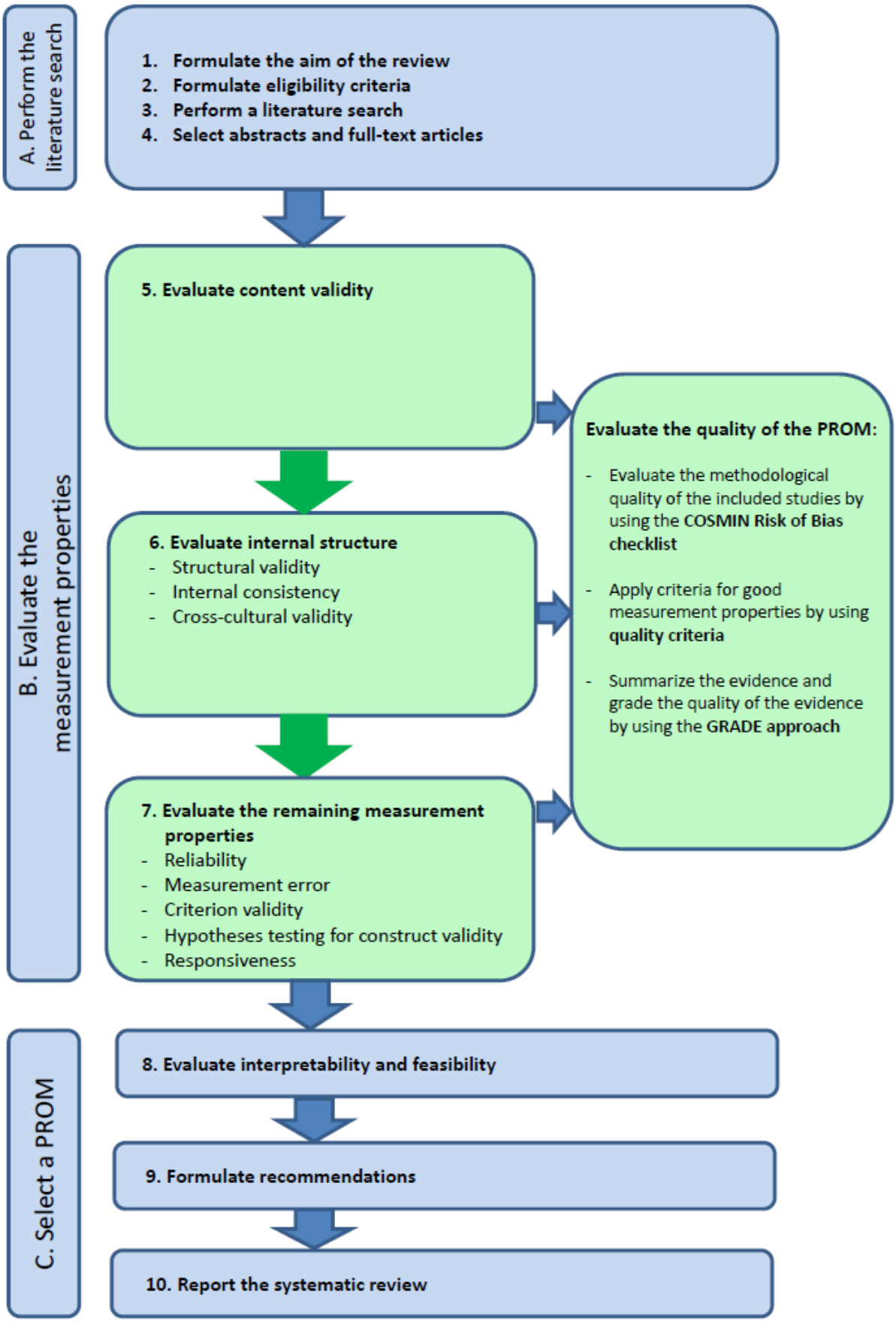


Figure 2. Ten steps for conducting a systematic review of PROMs (2)

2. Part A Steps 1-4: Perform the literature search

Steps 1-4 are standard procedures when performing a systematic reviews (2). It concerns formulating the research aim (step 1) and eligibility criteria (step 2), and performing the literature search (step 3) and the selection of articles (step 4). These are in agreement with existing guidelines for reviews (16, 17).

2.1 Step 1: Formulate the aim of the review

The aim of a systematic review of PROMs focuses on the quality of the PROMs. It should include the following four key elements: 1) the construct; 2) the population(s); 3) the type of instrument(s); and 4) the measurement properties of interest. An example of a research aim could be: "our aim is to critically appraise, compare and summarize the quality of the measurement properties of all self-report fatigue questionnaires for patients with multiple sclerosis (MS), Parkinson's disease (PD) or stroke" (25). In the aim of this review the construct of interest is 'fatigue', the population of interest is 'patients with MS, PD or stroke', the type of instrument of interest is 'self-report questionnaire', and 'all' measurement properties are explored in the review.

We also recommend that these four key elements are included in the title of the review. For example: "Self-report fatigue questionnaires in multiple sclerosis, Parkinson's disease and stroke: a systematic review of measurement properties." The four key elements will also inform both the eligibility criteria in Step 2 and the search strategy to be conducted in Step 3.

2.2 Step 2: Formulate eligibility criteria

The eligibility criteria should be in agreement with the four key elements of the review aim: 1) the PROM(s) should aim to measure the construct of interest; 2) the study sample (or an arbitrary majority, e.g. ≥50%) should represent the population of interest; 3) the study should concern PROMs; and 4) the aim of the study should be the evaluation of one or more measurement properties, the development of a PROM (to rate the content validity), or the evaluation of the interpretability of the PROMs of interest (e.g. evaluating the distribution of scores in the study population, percentage of missing items, floor and ceiling effects, the availability of scores and change scores for relevant (sub)groups, and the minimal important change or minimal important difference(26)).

We recommend to exclude studies that only use the PROM as an outcome measurement instrument, for instance, studies in which the PROM is used to measure the outcome (e.g. in randomized controlled trials), or studies in which the PROM is used in a validation study of another instrument. Including all studies that use a specific PROM would require a much more extended search strategy because the search filter for studies on measurement properties could not be used (see step 3). It would also be extremely time-consuming to find all studies using a specific PROM because the PROMs used in a study may not be reported in the abstract. This basically means that all studies performed in the population of interest should be screened full-text. As it might be

unlikely that this can be done in a standardized way, we therefore consider this approach not feasible.
Further, we recommend to include only full text articles because often very limited information on the design of a study is found in abstracts. Information found in abstracts will hamper the quality assessment of the study and the results of the measurement properties in steps 5 through 7.

2.3 Step 3: Perform a literature search

In agreement with the Cochrane methodology (16, 17), and based on consensus (5), MEDLINE and EMBASE are considered to be the minimum databases to be searched. In addition, it is recommended to search in other (content-specific) databases, depending on the construct and population of interest, for example Web of Science, Scopus, CINAHL, or PsycINFO.

Search strategy
In the guideline and in this manual we focus on a systematic review including *all* PROMs measuring a specific construct which have - at least some extent - been validated. An adequate search strategy therefore consists of a comprehensive collection of search terms (i.e. index terms and free text words) for the four key elements of the review aim: 1) construct; 2) population(s); 3) type of instrument(s), and 4) measurement properties. It is recommended to consult a clinical librarian as well as experts on the construct and study population of interest.

A comprehensive PROM filter has been developed for PubMed by the Patient Reported Outcomes Measurement Group, University of Oxford, that can be used as a search block for 'type of measurement instrument(s)', and is available through the COSMIN website. Regarding search terms for measurement properties we recommend to use a highly sensitive validated search filter for finding studies on measurement properties, which is available for PubMed and EMBASE and can be found on the COSMIN website (27, 28). An example of a PubMed search strategy is included in Appendix 1 (25).

Additional sources for search blocks can be found on the website https://blocks.bmi-online.nl/. Here a group of Dutch medical information specialists have compiled a number of 'Search Blocks'. These building blocks are intended for use when performing omplex search strategies in medical and health bibliographic databases, like PubMed, Embase, PsycInfo, Web of Science and others. These blocks can be used by experts in the field of searching for literature, e.g. information specialists, clinical librarians, health librarians and professionals in closely related areas.

As, in principle, the aim is finding *all* PROMs, in agreement with the Cochrane methodology, it is recommended to search databases from the date of inception till present (16, 17). The use of language restrictions depends on the inclusion criteria defined in Step 2. In general, it is recommend not to use language restrictions in the search strategy, even if there are no resources to translate the articles for the review. In this way, review authors are at least able to report their existence.

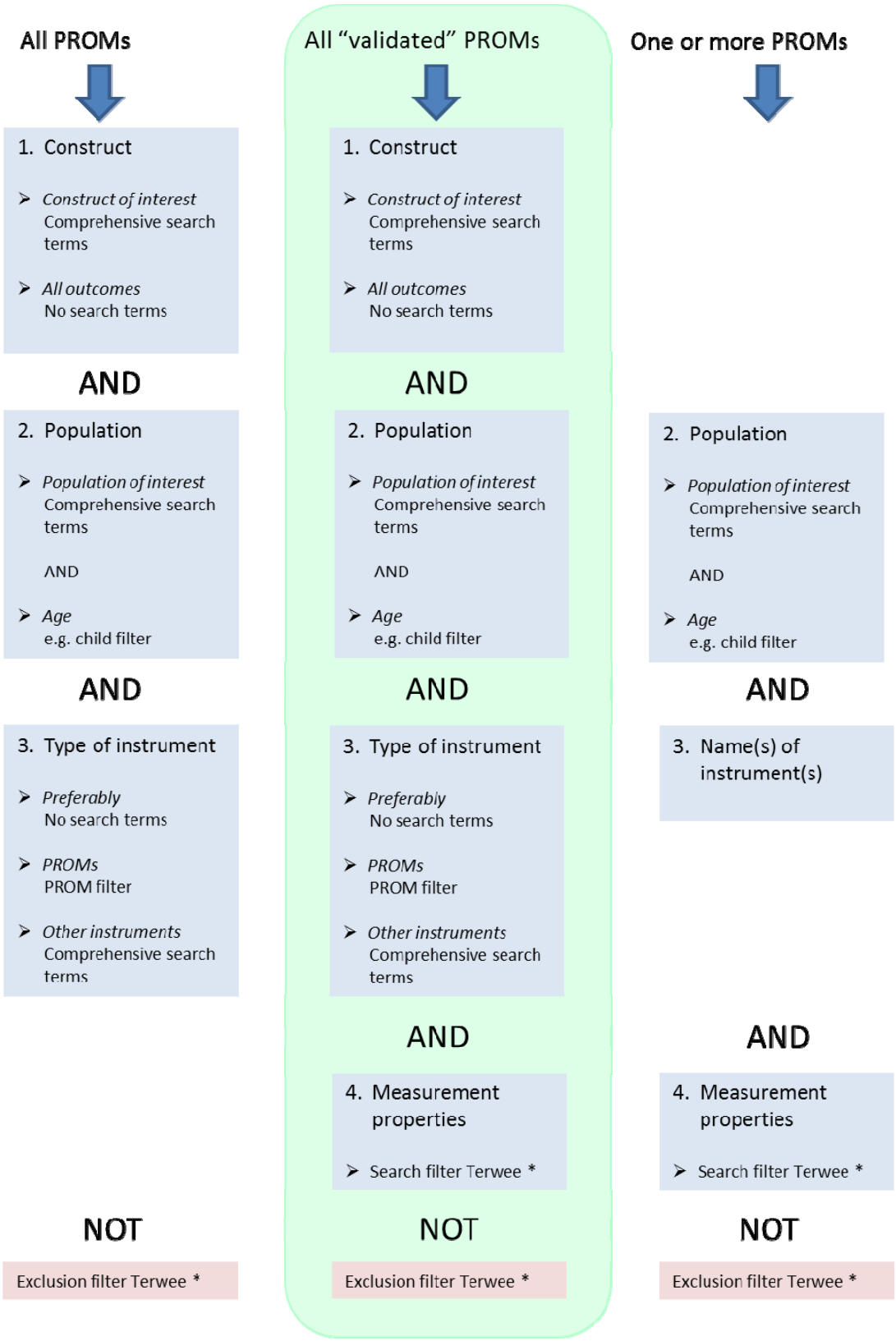
Next, it is recommended to use software such as Endnote or Reference Manager to manage references. Covidence (ww.covidence.org) could be of use when managing the

review screening performance. Covidence is an online tool that supports systematic reviewers in uploading search results, screening abstracts and full texts, resolve disagreements, and export data into RevMan or Excel. It is recommended by the Cochrane Collaboration.

Searches must be as up to date as possible when the review is published, so it may be necessary to update the search before submitting (or before acceptation) of the manuscript.

In addition to searching for *all* PROMs, Figure 3 shows the search strategy for other types of reviews. For example, if the aim of a systematic review is to identify all PROMs that have been *used*, search terms for measurement properties should not be used (first column of Figure 3). If the aim of a systematic review is to report on the measurement properties of *one particular PROM* (or a selection of predefined PROMs), search terms for the construct of interest should not be used and search terms for the type of instrument can be replaced by search terms for the names of the instruments of interest (third column of Figure 3).

As in all systematic reviews the literature search should be carefully documented in the study protocol. We recommend to document the following: names of the databases that were searched, including the interface used to search the databases (for example, PubMed was used for searching MEDLINE); all search terms used; any restrictions that were used (for example, human studies only, not animals); the number of records retrieved from each database; and the number of unique records.
In accordance with the PRISMA statement, it is recommend to add the documentation of the selection process to a flow diagram. An example of the PRISMA flow diagram can be found in Appendix 2 (29). This flow diagram includes information on the total number of abstracts selected, the total number of full-text articles that were selected, and the main reasons for excluding full-text articles. Note that the included articles can described one or more studies (on one or more different measurement properties). Therefore, we recommend to describe in the flow diagram the total number of articles included, the total number of studies described in those articles, and the total number of different (versions of) PROMs found.



* Search filter described by Terwee et al (27)

Figure 3. Search strategy for different types of systematic reviews of measurement instruments

2.4 Step 4: Select abstracts and full-text articles

It is generally recommended to perform the selection of abstracts and full-text articles by two reviewers independently (16, 17). If a study seems relevant by at least one reviewer based on the abstract, or in case of doubt, the full-text article needs to be retrieved and screened. Differences should be discussed and if consensus between the two reviewers cannot be reached, it is recommended to consult a third reviewer. It is also recommended to check all references of the included articles to search for additional potentially relevant studies. If many new articles are found, the initial search strategy might have been insufficiently comprehensive and may need to be improved and redone.

Note that Cochrane reviews use various additional sources in finding relevant studies, such as dissertations, editorials, conference proceedings, and reports. The probability of finding additional relevant articles for systematic reviews of PROMs in these type of additional sources, however, appears to be small.

3. Part B steps 5-7: Evaluating the measurement properties of the included PROMs

Part B consists of steps 5-7 of the guideline on conducting a systematic review of PROMs and concerns the evaluation of the measurement properties of the included PROMs. In Chapter 3.1 we will start with explaining the general methodology of Part B. In Chapter 3.2 we discuss step 5 in which the content validity is assessed. Next, in Chapter 3.3 step 6 is explained, which concerns evaluating the internal structure of a PROM (i.e. structural validity, internal consistency, and cross-cultural validity\measurement invariance). Lastly, in Chapter 3.4 we describe step 7 of the guideline that concerns the evaluation of the remaining measurement properties (i.e. reliability, measurement error, criterion validity, hypotheses testing for construct validity, and responsiveness).

3.1 General methodology

The evaluation of each measurement property includes three sub steps (see Figure 4):

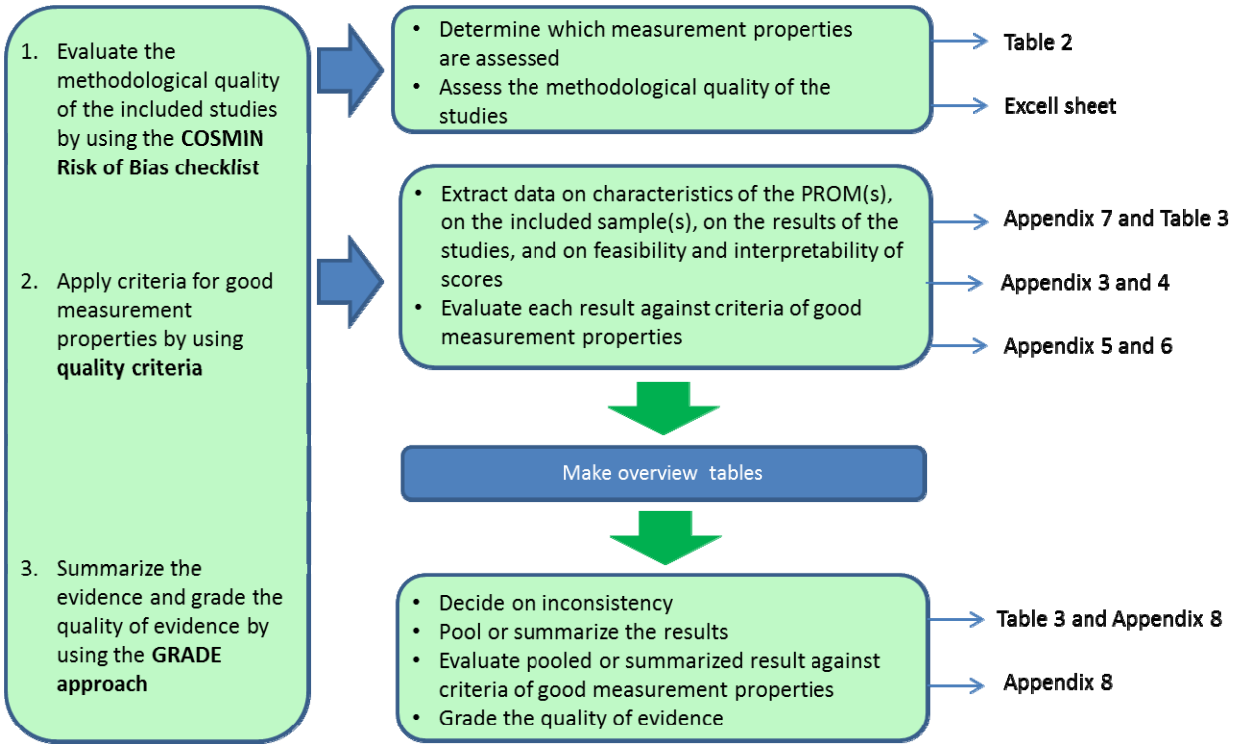


Figure 4. Practical outline for performing a systematic review

- First, the methodological quality of each single study on a measurement property is assessed using the COSMIN Risk of Bias checklist. Detailed instructions for how each standard included in the COSMIN Risk of Bias checklist should be rated are described in Chapter 5.
- Second, the result of each single study on a measurement property is rated against the updated criteria for good measurement properties (+ / - / ?) (Table 4);
- Third, the evidence is summarized per measurement property per PROM, the overall result is rated against the criteria for good measurement properties,

(Table 4), and the quality of the evidence will be graded by using the GRADE approach (2).

3.1.1 Evaluating the methodological quality of studies

To evaluate the methodological quality of the included studies using the COSMIN Risk of Bias checklist, it should first be determined which measurement properties are assessed in each article.

Determine which measurement properties are assessed

Often multiple studies on different measurement properties are described in one article (i.e. one study for each measurement property, e.g. a study on internal consistency, construct validity, and reliability, each with its own specific design requirements). The quality of each study is separately evaluated using the corresponding COSMIN box (Table 3). The COSMIN Risk of Bias checklist should be used as a modular tool; only those boxes should be completed for the measurement properties that are evaluated in the article.

Table 3. Boxes of the COSMIN Risk of Bias checklist

Mark the measurement properties that have been evaluated in the article*.	
Content validity	
	Box 1. PROM development
	Box 2. Content validity
Internal structure	
	Box 3. Structural validity
	Box 4. Internal consistency
	Box 5. Cross-cultural validity\measurement invariance
Remaining measurement properties	
	Box 6. Reliability
	Box 7. Measurement error
	Box 8. Criterion validity
	Box 9. Hypotheses testing for construct validity
	Box 10. Responsiveness

* If a box needs to be completed more than once two or more marks can be placed.

Sometimes the same measurement property is reported for multiple (sub)groups in one article. For example, when an instrument is validated in two different countries and the measurement properties are reported for both countries separately. In that case, the same box may need to be completed multiple times if the design of the study was different among countries. The review team should decide which boxes should be completed (and how many times).

In general, we recommend reviewers to consider how the designs and analyses presented in the article relate to the COSMIN taxonomy (see Figure 1) (21), and

subsequently complete the corresponding COSMIN box. This facilitates a consistent evaluation of the measurement properties across the included studies, regardless of the terminology used by the authors of the different articles.

Determining which box need to be completed sometimes requires a subjective judgement because the terms and definitions for measurement properties used in an article may not be similar to the terms used in the COSMIN taxonomy. For example, authors may use the term reliability when they present limits of agreement. While according to the COSMIN taxonomy, this would be considered measurement error. In that case, we recommend to complete box 7 (Measurement error). Also, authors often use the term criterion validity for studies in which a PROM is compared to other PROMs that measure a similar construct. In most cases, this would be considered evidence for construct validity, rather than criterion validity according to the COSMIN terms and definitions. In that case, we recommend to complete box 9 (Hypotheses testing for construct validity).

Sometimes results presented in studies on measurement properties actually do not (or hardly) provide any evidence on the measurement properties of a PROM, even though they are presented as such. For example, a comparison of different modes of administration (e.g. paper versus computer) does not provide information on the reliability or validity of the PROM. Also, sometimes correlations of a PROM with other variables (e.g. correlations with demographic variables) are reported and considered as evidence for construct validity however, these correlations provide very limited evidence for construct validity. Reviewers may decide to ignore such results in their review.

Assess the methodological quality of the studies

The quality of each study on a measurement property should be assessed separately, using the corresponding COSMIN box. Each study is rated as very good, adequate, doubtful or inadequate quality. To determine the overall rating of the quality of each single study on a measurement property, the lowest rating of any standard in the box is taken (i.e. "the worst score counts" principle)(12). For example, if the lowest rating of all eight items of the reliability box is 'inadequate', the overall methodological quality of *that* specific reliability study is rated as 'inadequate'.

In Chapter 5 we will provide more details and examples for how each standard in the COSMIN Risk of Bias checklist should be rated. We also explain in more detail how to come to the overall conclusion about the methodological quality of a study, i.e. how to apply the worst score counts principle. The COSMIN Risk of Bias checklist can be found on our website, along with a working document (created in Excel) that can be used to document the COSMIN ratings for each box.

3.1.2 Applying criteria for good measurement properties

Data extraction

We recommend to subsequently extract the data on the characteristics of the PROM(s), on characteristics of the included sample(s), on results on the measurement properties, and on information about interpretability and feasibility of the score(s) of the PROM(s). This information is needed to decide whether the results of different studies are sufficiently similar to be pooled or qualitatively summarized. This information can be

presented in overview tables. We recommend to extract the required information from the articles and directly paste it into the overview tables. Examples of these tables are given in Appendices 3-7.

In the Cochrane Handbook for systematic reviews of interventions it is recommended that the data extraction is done by two reviewers independently to avoid missing relevant information (16).

Evaluate each result against criteria of good measurement properties
Next, the result of each study on a measurement property should be rated against the updated criteria for good measurement properties (Table 4) (2). Each result is rated as either sufficient (+), insufficient (–), or indeterminate (?). The result of each measurement property and its quality rating can directly be added to the applicable table (Appendix 7).

Table 4. 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
	?	CTT: 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 validity[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 validity[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^2 < 0.02)
	?	No multiple group factor analysis OR DIF analysis performed
	–	Important differences between group factors OR DIF was found
Criterion validity	+	Correlation with gold standard ≥ 0.70 OR AUC ≥ 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

The criteria are based on e.g. Terwee et al.(30) and Prinsen et al.(5)

AUC = area under the curve, CFA = confirmatory factor analysis, CFI = comparative fit index, CTT = classical test theory, DIF = differential item functioning, ICC = intraclass correlation coefficient, IRT = item response theory, LoA = limits of agreement, MIC = minimal important change, RMSEA: Root Mean Square Error of Approximation, SEM = Standard Error of Measurement, SDC = smallest detectable change, SRMR: Standardized Root Mean Residuals, TLI = Tucker-Lewis index

[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

The criteria provided in Table 4 are the preferred criteria for each measurement property. However, for some measurement properties alternative criteria might also be appropriate. For example, additional criteria could be used for assessing the results of studies using exploratory factor analysis (EFA) or testing bi-factor confirmatory factor

analysis (CFA) models. If the review team has the expertise to assess the results of these types of studies, additional criteria could be used.

From this table, there is one aspect on which we would like to elaborate in more detail. This concerns the 'indeterminate' rating for the result of a study on internal consistency, i.e. when the criterion 'at least low evidence for sufficient structural validity' is not met. This means that either (1) there is only very low evidence for sufficient structural validity (e.g. because there was only one study on structural validity with a very low sample size), (2) there was (any) evidence for *in*sufficient structural validity, (3) there are inconsistent results for structural validity which cannot be explained, or (4) there is no information on the structural validity available.

3.1.3 Summarize the evidence and grade the quality of the evidence

Whereas in Chapter 3.1.1 and 3.1.2 we focused on the quality of single studies on measurement properties of a PROM, in the Chapters 3.1.3 and 3.1.4 we will focus on the quality of the PROM as a whole.

To come to an overall conclusion of the quality of the PROM, one should first decide whether the results of all available studies per measurement property are consistent. If the results are consistent, the results of studies can be quantitatively pooled or qualitatively summarized, and compared against the criteria for good measurement properties to determine whether overall, the measurement property of the PROM is sufficient (+), insufficient (–), inconsistent (±), or indeterminate (?). Finally, the quality of the evidence is graded (high, moderate, low, very low evidence), using a modified GRADE approach, as explained below (page 32-36).

If the results are inconsistent, there are several strategies that can be used: (a) find explanations and summarize per subgroup; (b) do not summarize the results and do not grade the evidence; or (c) base the conclusion on the majority of consistent results, and downgrade for inconsistency. It is up to the review team to decide on which strategy is most appropriate to use in their specific situation. Below each strategy is explained in more detail.

Handling inconsistent results

Based on the quality ratings of single results (as discussed in Chapter 3.1.2.), one decides whether or not all results on one measurement property of a PROM are consistent. If all results are consistent, the overall rating will be sufficient (+) or insufficient (-). If the results are inconsistent (e.g. both sufficient and insufficient results for reliability or appropriate model fit found but for different factor structures across studies), explanations for inconsistency should be explored. For example, inconsistency could be due to different populations or methods used. If an explanation is found, the results can be summarized e.g. per subgroup of consistent results. For example, if different results are found in studies performed in acute patients versus chronic patients, results per subgroup should be separately summarized. The overall rating for the specific measurement property (e.g. reliability) may be sufficient (+) in acute patients, but insufficient (-) in chronic patients.

High quality studies could be considered as providing more evidence than low quality studies and can be considered decisive in determining the overall rating when ratings are inconsistent. If inconsistent results can be explained by the quality of the studies, one may decide to summarize the results of very good or adequate studies only, and to ignore the results of doubtful an inadequate quality studies. This should then be explained in the manuscript.

In some cases, more recent evidence can be considered more important than older evidence. This can also be taken into account when determining the overall rating.

If no explanation for inconsistent results is found, there are two possibilities: (1) the overall rating will be inconsistent (±); or (2) one could decide to base the overall rating on the majority of the results (e.g. if the majority of the (consistent) results of studies are sufficient, an overall rating of sufficient could be considered) and downgrade the quality of the evidence for inconsistency (see Chapter 3.5).

Summarize the evidence

The results can be quantitatively pooled or qualitatively summarized. We recommend to report these pooled or summarized results per measurement property per PROM in Summary of Finding (SoF) Tables, accompanied by a rating of the pooled or summarized results (+ / ― / + / ?), and the grading of the quality of evidence (high, moderate, low, very low). These SoF tables (i.e. one per measurement property) will ultimately be used in providing recommendations for the selection of the most appropriate PROM for a given purpose or a particular situation. See Appendix 8 for an example.

Quantitatively pooling the results

If possible, the results from different studies on one measurement property should be statistically pooled in a meta-analysis. Pooled estimates of measurement properties can be obtained by calculating weighted means (based on the number of participants included per study) and 95% confidence intervals (e.g. (31)). We recommend to consult a statistician for performing meta-analyses.

For test-retest reliability, for example, weighted mean intraclass correlation coefficients (ICCs) and 95% confidence intervals can be calculated using a standard generic inverse variance random effects model (32). ICC values can be combined based on estimates derived from a Fisher transformation, $z = 0.5 \times \ln((1+ICC)/(1-ICC))$, which has an approximate variance, $(Var(z) = 1/(N-3))$, where N is the sample size. This method was applied in a study by Collins et al. (8).

For construct validity, for example, it can be considered to pool all correlations of a PROM with other PROMs that measure a similar construct. For example, when evaluating the construct validity of a physical function subscale one could pool all extracted correlations of this subscale with other comparison instruments measuring physical function.

Qualitatively summarizing into a summarized result

If it is not possible to statistically pool the results, the results of studies should be qualitatively summarized to come to a summarized result, for example by providing the range (lowest and highest) of MIC values found for interpretability, the percentage of confirmed hypotheses for construct validity, or the range of each model fit parameter on a consistently found factor structure in structural validity studies.

Applying criteria for good measurement properties to the pooled or summarized result

The pooled or summarized result per measurement property per PROM should again be rated against the same quality criteria for good measurement properties (Table 4). The overall rating for the pooled or summarized result can be sufficient (+), insufficient (–), inconsistent (±), or indeterminate (?). This rating can be added to the pooled or summarized result per PROM for each measurement property in the Summary of Findings Tables (Appendix 8).

If the results per study are all sufficient (or all insufficient), the overall rating will also be sufficient (or insufficient). To rate the qualitatively summarized results as sufficient (or insufficient), in principle 75% of the results should met the criteria. For example, for structural validity the criteria is that 'at least 75% of the CFA studies found the same factor structure'. The criteria for hypotheses testing and responsiveness (construct approach) for summary results is that 'at least 75% of the results should be in accordance with the hypotheses' to rate the overall results as 'sufficient' and 'at least 75% of the results are not in accordance with the hypotheses' to rate the overall results as 'insufficient''.

If the results of single studies which can be pooled are inconsistent and the inconsistency is unexplained, the results could be pooled anyway, and this pooled result could be rated as either sufficient or insufficient, and subsequently be downgraded due to inconsistency (see also Chapter 3.1.3 and the next section). If the results of single studies which cannot be pooled (e.g. results of CFAs) are inconsistent and the inconsistency is unexplained, the overall result will be rated as 'inconsistent'. In this case, the results are actually not summarized, and the evidence will not be graded. If the results per study are all indeterminate (?), the overall rating will also be indeterminate (?).

Grading the quality of evidence

After pooling or summarizing all evidence per measurement property per PROM, and rating the pooled or summarized result against the criteria for good measurement properties, the next step is to grade the quality of this evidence. The quality of the evidence refers to the confidence that the pooled or summarized result is trustworthy. The grading of the quality is based on the Grading of Recommendations Assessment, Development, and Evaluation (GRADE) approach for systematic reviews of clinical trials (20). Using a modified GRADE approach, the quality of the evidence is graded as high, moderate, low, or very low evidence (for definitions, see Table 5).

The GRADE approach uses five factors to determine the quality of the evidence: risk of bias (quality of the studies), inconsistency (of the results of the studies), indirectness (evidence comes from different populations, interventions or outcomes than the ones of interest in the review), imprecision (wide confidence intervals), and publication bias (negative results are less often published). For evaluating measurement properties in systematic reviews of PROMs, the following four factors should be taken into account: (1) risk of bias (i.e. the methodological quality of the studies), (2) inconsistency (i.e. unexplained inconsistency of results across studies), (3) imprecision (i.e. total sample size of the available studies), and (4) indirectness (i.e. evidence from different populations than the population of interest in the review).

The fifth factor, i.e. publication bias, is difficult to assess in studies on measurement properties, because of a lack of registries for these type of studies. Therefore, we do not take this factor into account in this methodology.

Table 5. Definitions of quality levels

Quality level	Definition
High	We are very confident that the true measurement property lies close to that of the estimate* of the measurement property
Moderate	We are moderately confident in the measurement property estimate: the true measurement property is likely to be close to the estimate of the measurement property, but there is a possibility that it is substantially different
Low	Our confidence in the measurement property estimate is limited: the true measurement property may be substantially different from the estimate of the measurement property
Very low	We have very little confidence in the measurement property estimate: the true measurement property is likely to be substantially different from the estimate of the measurement property

** Estimate of the measurement property refers to the pooled or summarized result of the measurement property of a PROM.*
These definitions were adapted from the GRADE approach (20)

The GRADE approach is used to downgrade evidence when there are concerns about the quality of the evidence. The starting point is always the assumption that the pooled or overall result is of high quality. The quality of evidence is subsequently downgraded by one or two levels per factor to moderate, low, or very low evidence when there is risk of bias, (unexplained) inconsistency, imprecision (low sample size), or indirect results. The quality of evidence can even be downgraded by three levels when the evidence is based on only one inadequate study (i.e. extremely serious risk of bias). The grading of the quality of each measurement property can directly be added to the applicable table (Appendix 8).

Below we explain in more detail how the four GRADE factors can be interpreted and applied in evaluating the measurement properties of PROMs.

How to apply GRADE
For each pooled result or for the summarized result for each measurement property per PROM, the quality of the evidence will be determined by using Table 6.

If in summarizing the evidence and determining the overall rating of the pooled or summarized result for a measurement property, the results of some studies are ignored, these studies should also be ignored in determining the quality of the evidence. For example, if only the results of high quality studies are considered in determining the overall rating, then only the high quality studies determine the grading of the evidence (in this case we would not downgrade for risk of bias).

Table 6. Modified GRADE approach for grading the quality of evidence

Quality of evidence	Lower if
High	Risk of bias -1 Serious -2 Very serious -3 Extremely serious Inconsistency -1 Serious -2 Very serious Imprecision -1 total n=50-100 -2 total n<50 Indirectness -1 Serious -2 Very serious
Moderate	
Low	
Very low	

n=sample size

Below we explain in more detail how the four GRADE factors can be interpreted and applied in evaluating the measurement properties of PROMs.

(1) Risk of bias can occur if the quality of the study is doubtful or inadequate, as assessed with the COSMIN Risk of Bias checklist, or if only one study of adequate quality is available. The quality of evidence should be downgraded with one level (e.g. from high to moderate evidence) if there is serious risk of bias, with two levels (e.g. from high to low) if there is very serious risk of bias, or with three levels (i.e. from high to very low) of there is extremely risk of bias. In Table 7 we explain what we consider as serious, very serious or extremely serious risk of bias.

Table 7. Instructions on downgrading for Risk of Bias

Risk of bias	Downgrading for Risk of Bias
No	There are multiple studies of at least adequate quality, or there is one study of very good quality available
Serious	There are multiple studies of doubtful quality available, or there is only one study of adequate quality
Very serious	There are multiple studies of inadequate quality, or there is only one study of doubtful quality available
Extremely serious	There is only one study of inadequate quality available

(2) Inconsistency: Inconsistency may already have been solved by pooling or summarizing the results of subgroups of studies with similar results and provide overall ratings for these subgroups. In this case, one doesn't need to downgrade. If no explanation for inconsistency is found, the review team can decide not to pool or summarize results, and rate the results as 'inconsistent'. In this case, no

quality level for the evidence will be given. However, an alternative solution could be to rate the pooled or summarized result (e.g. based on the majority of results) as sufficient or insufficient and downgrade the quality of the evidence for inconsistency with one or two levels. The reviewers should also decide what will be considered as serious (i.e. -1 level) or very serious (i.e. -2 levels) inconsistency, because this is context dependent. It is up to the review team to decide which seems to be the best solution in each situation.

(3) Imprecision refers to the total sample included in the studies. We recommend to downgrade with one level when the total sample size of the pooled or summarized studies is below 100, and with two levels when the total sample size is below 50. Note that this principle should not be used for measurement properties in which a sample size requirement is already included in the COSMIN Risk of Bias box, i.e. content validity, structural validity, and cross-cultural validity.

(4) Indirectness can occur if studies are included in the review that were (partly) performed in another population or another context of use than the population or context of use of interest in the systematic review. For example, if only part of the study population consists of patients with the disease of interest, the review team can decide to downgrade with one or two levels for serious or very serious indirectness. One can consider downgrading for indirectness for construct validity or responsiveness when the evidence is considered weak. For example when the evidence is only based on comparisons with PROMs measuring different constructs or only based on differences between extremely different groups. How to decide on what should be considered as serious or very serious indirectness is context dependent, and should be decided on by the review team.

To determine the grading for the quality of evidence, the concerns about the quality of the evidence should be added up. Therefore, it is helpful to consider the GRADE factors one by one by using the consecutive order as specified in Table 6. First, the risk of bias is considered (see Table 7). For example, when three studies are found with sufficient (i.e. '+') results, but all of doubtful quality, the level of evidence will be downgraded for risk of bias from high to moderate (i.e. -1). Second, further downgrading for other factors should be considered. After risk of bias, inconsistency should be considered. If the results of the three studies in the example above are all rated as sufficient, no downgrading for inconsistency is required. Otherwise, downgrading should be considered. Next, the sample size should be taken into account. For example, when the sample size of the three studies together is more than 100, there will be no further downgrading. If the (total) sample size is between e.g. 50-100, one should downgrade with -1. Lastly, the evidence could be downgraded because of indirectness. For example, consider a systematic review that focusses on pain and comfort scales for infants, and the inclusion criteria is 'children between 0-18 years' because a lack of studies in infants only (i.e. below 1 year) was expected. Studies including children of all ages, including infants, may lead to downgrade the quality of evidence by one level, and studies including only children between 4 and 12 years may even lead to downgrade the quality by two levels, due to indirectness of the results. If, in our example, the three studies found all include children between 0-4, but only very few infants, one may decide to downgrade one level (i.e. from moderate to low). In this example, the overall quality of

the evidence is now considered as 'low', so the conclusion will be that there is low quality evidence that the measurement property of interest is sufficient.

We recommend that grading is done by two reviewers independently and that consensus among the reviewers is reached, if necessary with help of a third reviewer.

3.2. Step 5: Evaluating content validity

Content validity (i.e. the degree to which the content of a PROM is an adequate reflection of the construct to be measured (21)) is considered to be the most important measurement property, because it should be clear that the items of the PROM are relevant, comprehensive, and comprehensible with respect to the construct of interest and study population (7). The evaluation of content validity requires a subjective judgment by the reviewers. In this judgement, the PROM development study, the quality and results of additional content validity studies on the PROMs (if available), and a subjective rating of the content of the PROMs by the reviewers is taken into account. As this step is very important but rather extensive we developed a separate users' manual for guidance on how to evaluate the content validity of PROMs. This manual can be found on the COSMIN website(33).

If there is high quality evidence that the content validity of a PROM is insufficient, the PROM will not be further considered in Steps 6-8 of the systematic review and one can directly draw a recommendation for this PROM in Step 9 (i.e. recommendation 'C': "PROMs that should not be recommended (i.e. PROMs with high evidence of insufficient content validity)). In all other cases, the PROM can be further taken into consideration in the systematic review.

3.3 Step 6. Evaluation of the internal structure of PROMs: structural validity, internal consistency, and cross-cultural validity\measurement invariance

Internal structure refers to how the different items in a PROM are related, which is important to know for deciding how items might be combined into a scale or subscale (6). This step concerns an evaluation of structural validity (including unidimensionality) using factor analyses or IRT or Rasch analyses; internal consistency; and cross-cultural validity and other forms of measurement invariance (using differential item functioning (DIF) analyses or Multi-Group Confirmatory Factor Analyses (MGCFA)). Here we are referring to testing of existing PROMs; not further refinement or development of new PROMs. These three measurement properties focus on the quality of items and the relationships between the items in contrast to the remaining measurement properties discussed in Step 7, who mainly focus on the quality of scales or subscales. We recommend to evaluate internal structure directly after evaluating the content validity of a PROM. As evidence for structural validity (or unidimensionality) of a scale or subscale is a prerequisite for the interpretation of internal consistency analyses (i.e. Cronbach's alpha's), we recommend to first evaluate structural validity, to be followed by internal consistency and cross-cultural validity\measurement invariance.

Step 6 is only relevant for PROMs that are based on a reflective model that assumes that all items in a scale or subscale are manifestations of one underlying construct and are expected to be correlated. An example of a reflective model is the measurement of anxiety; anxiety manifests itself in specific characteristics, such as worrying thoughts, panic, and restlessness. By asking patients about these characteristics, we can assess the degree of anxiety (i.e. all items are a reflection of the construct) (23). If the items in a scale or subscale are not supposed to be correlated (i.e. a formative model), these analyses are not relevant and Step 6 can be omitted. In other words, if factor analysis, or IRT or Rasch analysis is performed on a PROM based on a formative model, these results can be ignored, as the results are not interpretable.
If it is not reported whether a PROM is based on a reflective or formative model, the reviewers need to decide on the content of the PROM whether it is likely based on a reflective or a formative model. Unfortunately, it is not always possible to decide afterwards if the instrument is based on a reflective or formative model and thus whether structural validity is relevant. If a study included in the review presents a factor analysis, IRT or Rasch analysis and you are in doubt whether the (sub) scale is reflective, or whether it might be mixed (both reflective and formative items within a subscale), we recommend to consider it a reflective model. Subsequently, the quality of the study and the quality of the results are rated.

The evaluation of structural validity, internal consistency and cross-cultural validity\measurement invariance consists of three sub steps as described in Chapter 3.1 (see Figure 4). First, the risk of bias in each study on structural validity, internal consistency and cross-cultural validity\measurement invariance is assessed using the COSMIN Risk of Bias checklist (see Chapter 5 for detailed instructions for how to rate the standards in each box). Second, data is extracted on the characteristics of the PROM(s), on characteristics of the included patient sample(s), and on the results of the measurement properties (see Appendices 3-7). The result per study is then evaluated against the criteria for good measurement properties. Third, all results per measurement property of a PROM are quantitatively pooled or qualitatively

summarized, and this pooled or summarized result is again evaluated against the criteria for good measurement properties to get an overall rating (Table 4). Finally, the quality of the evidence is graded using the modified GRADE approach, as described in Chapter 3.1.3.

The risk of bias in a study on internal consistency depends on the available evidence for structural validity because unidimensionality is a prerequisite for the interpretation of internal consistency analyses (i.e. Cronbach's alpha's). Therefore, the quality of evidence for internal consistency cannot be higher than the quality of evidence for structural validity. Note that in a systematic review of PROMs, the conclusion about structural validity may come from different studies, even from studies that were conducted after studies on internal consistency.

We recommend to take the quality of evidence for structural validity as a starting point for determining the quality of evidence for internal consistency, and to downgrade further for risk of bias, inconsistency (i.e. unexplained inconsistency of results across studies), imprecision (i.e. total sample size of the available studies), and indirectness if needed.
A Cronbach's alpha based on a scale which is not unidimensional is difficult to interpret. We therefore recommend to ignore results of studies on internal consistency of scales when there is evidence that these scales are not unidimensional. For example, if results on structural validity show that a scale has three factors, internal consistency of each of those three subscales is relevant. Cronbach's alpha's on a total score can be ignored (and clinicians and researchers should be encouraged not to use these total scores) unless there is prove of unidimensionality of the total score, e.g. by a higher order or bi-factor CFA.

3.4 Step 7. Evaluation of the remaining measurement properties: reliability, measurement error, criterion validity, hypotheses testing for construct validity, and responsiveness

Subsequently, the remaining measurement properties (reliability, measurement error, criterion validity, hypotheses testing for construct validity, and responsiveness) should be evaluated. Unlike content validity and internal structure, the evaluation of these measurement properties mainly assess the quality of the scale or subscale as a whole, instead of the items.

Issues regarding validity to be decided upon a priori by the review team
For assessing the quality of hypotheses testing for construct validity and for the construct approach for responsiveness, the review team should a priori decide on two issues:

First, the review team should decide whether there is a gold standard available for measuring the construct of interest in the target population. If there is no gold standard (in principle, there is no gold standard available of a PROM, only the long version of a shortened PROM can be considered as a gold standard), the review team should not use the box for criterion validity and the box for the criterion approach for responsiveness in the systematic review. All evidence for validity presented in the included studies will be considered as evidence for construct validity or the construct approach for responsiveness. If an outcome measurement instrument is considered a gold standard, studies comparing the PROM to this particular instrument are considered as evidence for criterion validity or criterion approach for responsiveness.

Second, for interpreting the results of studies on hypotheses testing for construct validity, and on studies using a construct approach for evaluating responsiveness, the review team should formulate a set of hypotheses about expected relationships between the PROM under review and other well-defined and high quality comparator instruments which are used in the field. If possible, also some hypotheses about expected differences between subgroups can be formulated in advance. This way, the study results are compared against the same hypotheses.
The expected direction (positive or negative) and magnitude (absolute or relative) of the correlations or differences should be included in the hypotheses (for examples we refer to other publications(34-37)). Without this specification it is difficult to decide afterwards whether the results are in accordance with the hypothesis or not. For example, review team could state that they expect a correlation of at least 0.50 between the PROM under study and the comparator instrument that intend to measure the same construct. Or that they expected that chronically ill patients score on average score 10 points higher compared to acute patients on the PROM under consideration. The hypotheses may also concern the relative magnitude of correlations. For example, the review team could state that they expect that the score on PROM A correlates at least 0.10 points higher with the score on instrument (e.g. a PROM) B than with the score on instrument C. In Table 8 some generic hypotheses are presented that could be considered as a starting point for formulating hypotheses for the review (23).

Table 8. Generic hypotheses to evaluate construct validity and responsiveness

Generic hypotheses*:	
1.	Correlations with (changes in) instruments measuring similar constructs should be ≥0.50.
2.	Correlations with (changes in) instruments measuring related, but dissimilar constructs should be lower, i.e. 0.30-0.50.
3.	Correlations with (changes in) instruments measuring unrelated constructs should be <0.30.
4.	Correlations defined under 1, 2, and 3 should differ by a minimum of 0.10.
5.	Meaningful changes between relevant (sub)groups (e.g. patients with expected high vs low levels of the construct of interest)
6.	For responsiveness, AUC should be ≥ 0.70

*Reproduced with approval from De Vet et al. (23)
AUC = Area Under the Curve with an external measure of change used as the 'gold standard'

Hypotheses can be based on literature (including the included studies in the review) and (clinical) experiences of the review team members. In the original COSMIN checklist, one of the standards included in the box on hypotheses testing for construct validity and responsiveness was about whether hypotheses were stated in the study, which often lead to an 'inadequate' rating. By recommending the review team to formulate hypotheses in the new methodology, the methodological quality of an included study is no longer dependents upon whether or not hypotheses were formulated in these studies. All results found in included articles can now be compared against this same set of hypotheses, and more results can be used to build a conclusion about the construct validity of a PROM.
Additional hypotheses may need to be formulated during the review process, depending on the observed results in the included studies. We consider it not a problem if the hypotheses are not all defined a priori, as long as the hypotheses are reported in the review, making the methodology transparent.
If it is impossible to formulate a good hypothesis for a specific analysis (e.g. about an expected magnitude of difference between two subgroups), and the authors of the study did not state their expectations, the review team can consider ignoring these results, as it is unknown how the results should be interpreted.

When assessing responsiveness, one of the most difficult tasks is formulating challenging hypotheses. By challenging hypotheses we aim to show that the instrument truly measures changes in the construct(s) it purports to measure. This means that the instrument should measure changes in the purported construct(s), but also that it should measure the right amount of change, i.e. it should not under- or overestimate the real change in the construct that has occurred. This latter aspect is often overlooked in assessing responsiveness. For example, the hypotheses can concern expected mean differences between changes in groups or expected correlations between changes in the scores on the instrument and changes in other variables, such as scores on other instruments, or demographic or clinical variables. Hypotheses about expected effect size (ES) or similar measures such as standardized response mean (SRM) can also be used, but only when an explicit hypothesis (and rationale) for the expected magnitude of the

effect size is given. The hypotheses may also concern the relative magnitude of correlations, for example a statement that change in instrument A is expected to correlate higher with change in instrument B than with change in instrument C.

Evaluating the measurement properties and grading the quality of the evidence

The evaluation of the measurement properties consists of the three sub steps, as described earlier in Chapter 3.1 (Figure 4).

First, the risk of bias in each study on these measurement properties is assessed using the COSMIN Risk of Bias checklist (see Chapter 5 for detailed instructions for how to rate the standards).

Second, data is extracted on the characteristics of the PROM(s), on characteristics of the included study population(s), and on results on the measurement properties (see Appendices 3-6), and the result per study is evaluated against the criteria for good measurement properties (see Table 4).

Third, all results per measurement property of a PROM are quantitatively pooled or qualitatively summarized, and this pooled or summarized result is evaluated against the criteria for good measurement properties to get an overall rating for the measurement property (Table 4). Finally, the quality of the evidence is graded using the modified GRADE approach, as described in Chapter 3.1.

Specific aspects to consider for measurement error

When applying the criteria for good measurement error, information is needed on the Smallest Detectable Change (SDC) or Limits of Agreement (LoA), as well as on the Minimal Important Change (MIC). This information may come from different studies. The MIC should have been determined using an anchor-based longitudinal approach (38-41). The MIC is best calculated from multiple studies and by using multiple anchors. It is up to the review team to decide whether the quality of the evidence should be downgraded (e.g. one level) when there is information available on the MIC value from only one study, or when the study on the MIC is not adequately performed (e.g. insufficient validity of the anchor)(42, 43). When a MIC value is determined using a distribution-based method, in fact, the MIC value does not reflect what patients consider important, but is rather an indicator of the measurement error of the PROM. Results found in such studies should either be ignored or considered as evidence on measurement error (23). If not enough information is available to judge whether the SDC or LoA is smaller than the MIC, we recommend to just report the information that is available on the SDC or LoA without grading the quality of evidence (note that information on the MIC alone provides information on the interpretability of a PROM, see Chapter 4.1).

Specific aspects to consider for hypotheses testing and responsiveness

In studies on construct validity and responsiveness, different study designs may have been used, i.e. comparisons with other outcome measurement instrument, comparisons between groups, or comparisons before and after an intervention. Therefore, the boxes Hypotheses testing for construct validity and Responsiveness both consist of two and four parts, respectively (see Chapter 5.7 and 5.8). The methodological quality of each part will be rated separately, using the "worst score counts" method (12).

In general, each result (e.g. a correlation between scores of two outcome measurement instruments, or an ES) could be considered as a single study. When for example, the PROM under study is compared to four different outcome measurement instruments,

four correlations are computed, and this could be considered as four single 'studies'. Basically, the box Hypotheses testing should be completed four times. However, when each standard for all four 'studies' will be rated the same (e.g. the constructs measured by all four comparison outcome measurement instruments are clear and all comparison instruments are of good quality) the ratings can be combined into one risk of bias rating (i.e. the methodological quality of the studies is 'very good'). Next, each result is compared against the criterion (i.e. whether or not the hypothesis was confirmed), and reported in the results table (see Appendix 7). If the methodological quality of each 'study' is not similar, it could be assessed separately. In Appendix 7, an example is given (called 'ref 6' under hypotheses testing) of an article in which three hypotheses were tested in adequate studies and the results are in accordance with the hypotheses, and three other hypotheses were tested in inadequate studies, and one of the results was in accordance with the hypotheses and two results were not in accordance with the hypotheses.
Moreover, in one article both hypotheses about comparison with other instruments could be tested as well as hypotheses about expected differences between subgroups. In that case, different parts of the box Hypotheses testing for construct validity will be used (part a and b, respectively). However, when the ratings are the same (for methodological quality), this could be reported together.
Next, the quality of the evidence should be determined. We recommend to determine this similarly as for the other measurement properties. In this step, we consider all results reported in one article together as one study. We do not grade the evidence per hypothesis or per study design, because otherwise high evidence for hypotheses testing can easily be reached.
An example is provided in Appendix 7. Here, 11 out of 13 results are rated as sufficient. We conclude that we have consistent findings for sufficient hypotheses testing for construct validity, as 85% of the results are in line with the hypotheses. We have one very good study. Therefore, we do not downgrade for risk of bias. Depending on imprecision (i.e. total sample size of the available studies), and indirectness we could downgrade, if needed.

In studies on construct validity and responsiveness some results may provide stronger evidence than other results. For example, correlations with PROMs measuring similar constructs (i.e. convergent validity) can be considered as providing more evidence than correlations with PROMs measuring dissimilar constructs. This can be taken into account when determining the pooled or summarized result for construct validity, especially when results are inconsistent. For example, when results about comparisons with PROMs measuring similar constructs are consistently not in accordance with the hypotheses, while results about comparison with dissimilar constructs tend to be in accordance with the hypotheses, the review team could decide to put more emphasis on the results of the hypotheses concerning comparisons with instruments measuring similar constructs.

4. Part C steps 8-10: Selecting a PROM

Part C steps 8-10 concerns the description of data on interpretability and feasibility of the PROMs (step 8), formulating recommendations based on all evidence (step 9) and the reporting of the systematic review (step 10).

4.1 Step 8: Describe interpretability and feasibility

Interpretability is defined as the degree to which one can assign qualitative meaning (that is, clinical or commonly understood connotations) to a PROM's quantitative scores or change in scores (21). Both the interpretability of single scores and the interpretability of change scores is informative to report in a systematic review. The interpretation of single scores can be outlined by providing information on the distribution of scores in the study population or other relevant subgroups, as it may reveal clustering of scores, and it can indicate floor and ceiling effects. The interpretability of change scores can be enhanced by reporting MIC values and information on response shift (referring to changes in the meaning of one's self-evaluation of a target construct as a result of: (a) a change in the patient's internal standards of measurement (i.e. scale recalibration); (b) a change in the patient's values (i.e. the importance of component subdomains constituting the target construct); or (c) a redefinition of the target construct (i.e. reconceptualization)(44)) (see Appendix 5).

Feasibility is defined as the ease of application of the PROM in its intended context of use, given constraints such as time or money (45), for example, completion time, cost of an instrument, length of the instrument, type and ease of administration (see Appendix 6 for all aspects of feasibility) (26). Feasibility applies to patients completing the PROM (self-administered) and researchers or clinicians who interview or hand over the PROM to patients. The concept 'feasibility' is related to the concept 'clinical utility', whereas feasibility focusses on PROMs, clinical utility refers to an intervention (46).
Interpretability and feasibility are not measurement properties because they do not refer to the quality of a PROM. However, they are considered important aspects for a well-considered selection of a PROM. In case there are two PROMs that are very difficult to differentiate in terms of quality, it is recommended that feasibility aspects (e.g. time aspects and budget restrictions) should be taken into consideration in the selection of the most appropriate instrument. Floor and ceiling effects can indicate insufficient content validity, and can result in insufficient reliability.
Interpretability and feasibility are only described, and not evaluated, as these are no formal measurement properties (i.e. they do not tell us something about the quality of an instrument, rather they are important aspects that should be considered in the selection of the most suitable instrument for a specific purpose).

4.2 Step 9: formulate recommendations

Recommendations on the most suitable PROM for use in an evaluative application are formulated with respect to the construct of interest and study population. To come to an evidence-based and fully transparent recommendation, we recommend to categorize the included PROMs into three categories:

(A) PROMs with evidence for sufficient content validity (any level) AND at least low quality evidence for sufficient internal consistency;

(B) PROMs categorized not in A or C.

(C) PROMs with high quality evidence for an insufficient measurement property

PROMs categorized as 'A' can be recommended for use and results obtained with these PROMs can be trusted. PROMs categorized as 'B' have potential to be recommended for use, but they require further research to assess the quality of these PROMs. PROMs categorized as 'C' should not be recommended for use. When only PROMs categorized as 'B' are found in a review, the one with the best evidence for content validity could be the one to be provisionally recommended for use, until further evidence is provided.

Justifications should be given to as why a PROM is placed in a certain category, and direction should be given on future validation work, if applicable. To work towards standardization in outcome measurement (e.g. core outcome set development) and to facilitate meta-analyses of studies using PROMs, we recommend to subsequently advise on one most suitable PROM (5). This recommendation does not only have to be based on the evaluation of the measurement properties, but may also depend on interpretability and feasibility aspects.

4.3 Step 10: report the systematic review

In accordance with the PRISMA Statement (18), we recommend to report the following information:
(1) the search strategy (for example on a website or in the (online) supplemental materials to the article at issue), and the results of the literature search and selection of the studies and PROMs, displayed in the PRIMSA flow diagram (including the final number of articles and the final number of PROMs included in the review) (Appendix 2);
(2) the characteristics of the included PROMs, such as name of the PROMs, reference to the article in which the development of the PROM is described, constructs being measured, language and study population for which the PROM was developed, intended context(s) of use, available language version of the PROM, number of scales or subscales, number of items, response options, recall period, interpretability aspects, and feasibility aspects (Appendix 3);
(3) the characteristics of the included study populations, such as geographical location, language, important disease characteristics, target population, sample size, age, gender, and setting (Appendix 4);
(4) the methodological quality ratings of each study per measurement property per PROM (i.e. very good, adequate, doubtful, inadequate), the results of each study, and the

accompanying ratings of the results based on the criteria for good measurement properties (sufficient (+) / insufficient (-) / indeterminate (?)) (Appendix 7). This table could be published for example as Appendix or supplemental material on the website of the journal only;
(5) a Summary of Findings (SoF) table per measurement property, including the pooled or summarized results of the measurement properties, its overall rating (i.e. sufficient (+) / insufficient (-) / inconsistent (±)/ indeterminate (?)), and the grading of the quality of evidence (high, moderate, low, very low). An example of a SoF table can be found in Appendix 8. These SoF tables (i.e. one per measurement property) will ultimately be used in providing recommendations for the selection of the most appropriate PROM for a given purpose or a particular context of use. Note that these tables can be used in the data extraction process throughout the entire review.

5. COSMIN Risk of Bias checklist

In this chapter we elaborate on all standards included in boxes 3 to 10 of the COSMIN Risk of Bias Checklist. An elaboration of all standards included in box 1 PROM Development and box 2 Content validity is provided in the users' manual for guidance on how to evaluate the content validity of PROMs. This manual can be found on the COSMIN website (33).
Each box contains an item asking if there were any other important other methodological flaws that are not covered by the checklist, but that may lead to biased results or conclusions. For some of the boxes we will provide examples of such flaws below.

5.1 Assessing risk of bias in a study on structural validity (box 3)

Structural validity refers to the degree to which the scores of a PROM are an adequate reflection of the dimensionality of the construct to be measured(21) and is usually assessed by factor analysis.

Box 3. Structural validity	
Does the scale consist of effect indicators, i.e. is it based on a reflective model?	yes / no
Does the study concern unidimensionality or structural validity?	unidimensionality / structural validity

The box on structural validity starts with two questions that can help the reviewer to decide whether this measurement property is relevant for the instrument under study (i.e. question on reflective model), or to become aware of the specific purpose of the study (i.e. question on unidimensionality or structural validity). Both questions are not standards and they are not used in the worst score counts method.

Structural validity is only relevant for instruments that are based on a reflective model. A reflective model is a model in which all items are a manifestation of the same underlying construct. These kind of items are called effect indicators. These items are expected to be highly correlated and interchangeable. Its counterpart is a formative model, in which the items together form the construct. These items do not need to be correlated. Therefore, structural validity is not relevant for items that are based on a formative model(24, 47, 48). For example, stress could be measured by asking about the occurrence of different situations and events that might lead to stress, such as job loss, death in a family, divorces etc. (49). These events do not need to be correlated, thus structural validity is not relevant for such an instrument.

Often, authors do not explicitly describe whether their instrument is based on a reflective or formative model. To decide afterwards which model is used, one can do a simple "thought test". With this test one should consider whether all item scores are expected to change when the construct changes. If yes, the construct can be considered a reflective model. If not, the PROM is probably based on a formative model (24, 47).

It is not always possible to decide afterwards if the instrument is based on a reflective or formative model and thus whether structural validity is relevant. In this case, we recommend to complete the box to assess the quality of the analyses that were performed in the included studies.

The measurement property Structural validity refers to the model fit of a factor analysis on all items in an outcome measurement instrument, e.g. to confirm a 3-factor model for an instrument with three subscales. Unidimensionality refers to whether the items in a scale or subscale measure a single construct. It is an assumption for internal consistency or IRT/Rasch analyses. For the purpose of checking unidimensionality each subscale from a multidimensional PROM can separately be evaluated with a factor analysis or IRT/Rasch analyses. An evaluation of unidimensionality is sufficient for the interpretation of an internal consistency analysis and IRT/Rasch analysis, but it does not provide evidence for structural validity as part of construct validity of a multidimensional PROM. That is, because it does not provide evidence, for example, that an instrument with three subscales indeed measures three different constructs. A question was added to the box on Structural validity about whether the aim of the study was to assess unidimensionality or to assess structural validity. Although the standards for assessing unidimensionality and structural validity are the same, the purpose is different and the conclusion about the PROM can be different and reviewers should take this into account when reporting the results of such studies in a systematic review. This question is not a standard and it is not used in the worst score counts method. It is only a help for the reviewers to be aware of which purpose was served in the study.

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 applicable

Standard 1. To determine the structure of the instrument, a factor analysis is the preferred statistic when using CTT. Although confirmatory factor analysis is preferred over explorative factor analysis, both could be useful for the evaluation of structural validity. Confirmative factor analysis tests whether the data fit a premeditated factor structure(50). Based on theory or previous analyses a priori hypotheses are formulated and tested. Explorative factor analysis can be used when no clear hypotheses exist about the underlying dimensions (50).

Statistical methods					
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	NA

Standard 2: An example of a model that does not fit the research question is when follow-up data are combined but not analysed using a multi-level IRT model.

Statistical methods	**very good**	**adequate**	**doubtful**	**inadequate**	**NA**
3 Was the sample size included in the analysis adequate?	FA: 7 times the number of items, and ≥100	FA: at least 5 the times number of items, and ≥100; OR at least 6 times the 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	

Standard 3. Factor analyses and IRT/Rasch analyses require large sample sizes. The proposed requirements were based on Comrey (51), Brown (chapter 10) (52), Embretson and Reise (53), Edelen and Reeve (54), Reise and Yu (55), Reise et al. (56), and Linacre (57).

These sample size requirements are rules of thumb, and are context related. Sample size requirements increase as a result of the complexity of the mode l (Edelen et al. 2007). Moreover, depending on the research question different levels of precision may be acceptable, which relate again to the sample size needed (Edelen et al. 2007). Another related consideration is the sampling distribution of the respondents. The sample should reflect the population of interest and contain enough respondents with extreme scores so that items even at extreme ends of the construct continuum will have reasonable standard errors associated with their estimated parameters. (Edelen et al. 2007).

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)	

We have not defined specific requirements for factor analyses, such as the choice of the explorative factor analysis (principal component analysis or common factor analysis),

the choice and justification of the rotation method (e.g. orthogonal or oblique rotation), or the decision about the number of relevant factors. Such specific requirements are described by e.g. Floyd & Widaman (50) and de Vet et al. (58). When there are serious flaws in the quality of the factor analysis, we recommend to rate item 4 with "doubtful" or "inadequate".

5.2 Assessing risk of bias in a study on internal consistency (box 4)

Internal consistency refers to the degree of interrelatedness among the items and is often assessed by Cronbach's alpha. Cronbach's alpha's can be pooled across studies when the results are sufficiently similar. For an example, we refer to a study performed by Collins and colleagues (8).

For an appropriate interpretation of the internal consistency parameter, the items together should form a unidimensional scale or subscale. Internal consistency and unidimensionality are not the same. Internal consistency is the relatedness among the items (59). Unidimensionality refers to whether the items in a scale or subscale measure a single construct. It is a prerequisite for a clear interpretation of the internal consistency statistics (59, 60), and can be investigated for example by factor analysis (61) or IRT methods (see structural validity, box 3).

Box 4. Internal consistency
Does the scale consist of effect indicators, i.e. is it based on a reflective model? [1] yes / no

The first question in the Box Internal consistency concerns the relevance of the assessment of the measurement property internal consistency for the PROM under study. The internal consistency statistic only gets an interpretable meaning, when the interrelatedness among the items is determined of a set of items that together form a reflective model (59, 60). See also box 3 for an explanation about reflective models.

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 on unidimensional scale	

Standard 1. The internal consistency coefficient should be calculated for each unidimensional subscale separately. Information on unidimensionality (or structural validity) can be either found in the same study or in other studies. Based on the at least low quality evidence for structural validity we determine whether or not the items from the (sub) scales form an unidimensional scale.
When an internal consistency parameter is calculated per subscale as well as for a multidimensional total scale (for example the total score of four subscales) this latter result can be ignored, as it cannot be interpreted. If an internal consistency parameter is only reported for a multidimensional total scale, this standard should be rated

'inadequate'. If no information is found in the literature on the structural validity or unidimensionality of a PROM, this standard can be rated with 'doubtful'. In this case, we recommend to report the results found on internal consistency, without drawing a conclusion.

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	Na
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	Na
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	Na

Standards 2 and 3. Studies based on CTT should report Cronbach's alpha or Omega values (62).

Standard 4. Several reliability indices are available that are based on IRT/Rasch analyses. These indices are based on one score. Examples are the standard error of theta, and index of person or subject separation.

5.3 Assessing risk of bias in a study on cross-cultural validity\ measurement invariance (box 5)

Cross-cultural validity refers to the degree to which the performance of the items on a translated or culturally adapted instrument are an adequate reflection of the performance of the items of the original version of the instrument. To assess this measurement property data from at least two different groups is required, e.g. a Dutch and an English sample, or males and females. We recommend to complete this box for studies that have evaluated cross-cultural validity for PROMs across culturally different populations. We interpret 'culturally different population' broadly. We do not only consider different ethnicity or language groups as different cultural populations, but also other groups such as different gender or age groups, or different patient populations. Cross-cultural validity is evaluated by assessing whether the scale is measurement invariant or whether or not Differential Item Functioning (DIF) occurs. Measurement Invariance (MI) and non-DIF refer to whether respondents from different groups with the same latent trait level (allowing for group differences) respond similarly to a particular item.

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	

Standard 1. When evaluating cross-cultural validity, scores of two (or more) groups are directly compared in one statistical model. This could be language groups (e.g. when the Dutch version of a PROM is being compared to the English version of the same PROM), or groups based on another variable, such as males versus females, or healthy and chronically ill people. Except from the group variable, the two groups should be similar for relevant characteristics, such as disease severity, age, etc. In one study gender could be the group variable, and in another study (e.g. comparing two language groups) gender is considered a relevant characteristic of the groups and is expected to be similarly distributed across the groups. It is up to the review team to judge whether all relevant characteristics are similarly distributed across the groups.

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

Standard 2. Adequate approached for assessing cross-cultural validity using CTT are regression analyses or confirmatory factor analysis (CFA). For an explanation of the use of ordinal regression models, we refer to Crane et al. (63) or Petersen et al. (64). For an explanation of CFA to investigate measurement invariance we refer to Gregorich (65). An adequate approach for assessing cross-cultural validity using IRT methods is Differential Item Functioning (DIF) analyses (66).

Statistical methods					
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 items, 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	

Standard 3. CFA, IRT analysis or regression analysis require high sample sizes to obtain reliable results. As the results of studies cannot be pooled, a sample size requirement was included as a standard. These recommendations are based on a publication by Scott et al. (67).

5.4 Assessing risk of bias in a study on reliability (box 6)

Reliability refers to the proportion of the total variance in the measurements which is due to 'true' differences between patients. The word 'true' must be seen in the context of CTT, which states that any observation is composed of two components – a true score and error associated with the observation. 'True' is the average score that would be obtained if the scale was administered an infinite number of times to the same person. It refers only to the consistency of the score, and not to its accuracy (22). Reliability can also be explained as the ability of a PROM to distinguish between patients. Within a homogeneous group, it is hard to distinguish between patients (23). An important assumption made in a reliability study (and in a study on measurement error) is that patients are stable on the construct to be measured between the repeated measurements.

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 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

Standard 1. Patients should be stable with regard to the construct to be measured between the administrations. What "stable patients" are depends on the construct to be measured and the target population. Evidence that patients were stable could be, for example, an assessment of a global rating of change, completed by patients or physicians. When an intervention is given in the interim period, one can assume that (many of) the patients have changed on the construct to be measured. In that case, we recommend to rate Standard 2 as "inadequate".

Standard 2. The time interval between the administrations must be appropriate. The time interval should be long enough to prevent recall bias, and short enough to ensure that patients have not been changed on the construct to be measured. What an appropriate time interval is, depends on the construct to be measured and the target population. A time interval of about 2 weeks is often considered appropriate for the evaluation of PROMs (22).

Standard 3. A last requirement is that the test conditions should be similar. Test conditions refer to the type of administration (e.g. a self-administered questionnaire, interview, performance-test), the setting in which the instrument was administered (e.g. at the hospital, or at home), and the instructions given. These test conditions may influence the responses of a patient. The reliability may be underestimated if the test conditions are not similar. However, in clinical practice, different test condition might be used disorderly, and a specific research question could be whether the reliability of a PROM is still appropriate when using the PROM under different test conditions. In this case, the test conditions do not need to be similar as they are supposed to vary as part of the research aim of the study. In that case, this item can be rated with 'very good'. See for example, Van Leeuwen(68).

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

The preferred reliability statistic depends on the type of response options.

Standard 4. For continuous scores the intraclass correlation coefficient (ICC) is preferred (22, 69). To assess the reliability of a PROM, often a straightforward test-retest reliability study design is chosen. The preferred ICC model in that case is the two-way random effects model (70), in which in addition to the variance within patients the variance (i.e. systematic differences) between time point is taken into account (23). The use of the Pearson's and Spearman's correlation coefficient is considered doubtful when it is not clear whether there are systematic differences occurred, because these correlations do not take systematic error into account.

Standards 5, 6, and 7. For dichotomous scores or nominal scores the Cohen's kappa is the preferred statistical method (22). For ordinal scales partial chance agreement should be considered, and therefore a weighted kappa (22, 71, 72) is preferred. A description of the weights (e.g., linear or quadratic weights (73)) should be given. Proportion agreement is considered not adequate as it is a parameter for measurement error (see box 7).
Unweighted kappa considers any misclassification equally inappropriate. However, a misclassification of two adjacent categories may be less erroneous as a misclassification of categories that are more apart from each other. A weighted kappa takes this into account. However, the aim of a study could be to assess the reliability of any misclassification, making a unweighted kappa as an appropriate parameter. In such a study, Standard 7 can be rated as 'very good'.

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

Standard 8: An example of another important flaw is when the administrations were not independent. Independent administrations imply that the first administration has not influenced the second administration. At the second administration the patient should not have been aware of the scores on the first administration.
Another example could be, when the PROM was administered in an interview. Suppose one experienced interviewer administers all first administrations of the interviews of all patients, and the second administration was either done by an experienced interviewer or by an unexperienced interviewer, and it was not clear which patient was interviewed by which interviewer. Subsequently, this was not taken into account in the analysis. Suppose a low ICC was found, it is unknown whether this is due to the different characteristics of the interviewers (that could be improved by standardizing requirements for interviewers), or due to an insufficient quality of the PROM.

5.5 Assessing risk of bias in a study on measurement error (box 7)

Measurement error refers to the systematic and random error of an individual patient's score that is not attributed to true changes in the construct to be measured.

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	

Standards 1-3: see reliability

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

Standard 4. The preferred statistic for measurement error in studies based on CTT is the Standard Error of Measurement (SEM) based on a test-retest design. Note that the calculation of the SEM based on Cronbach's alpha is considered not appropriate, because it does not take the variance between time points into account (74). Other appropriate statistics for assessing measurement error are the Limits of Agreement (LoA) and the Smallest Detectable Change (SDC) (23). Both parameters are directly related to the SEM. Changes within the LoA or smaller than the SDC are likely to be due to measurement error and changes outside the LoA or larger than the SDC should be considered as real change in individual patients. Note that this does not indicate that these changes are also meaningful to patients. This depends on what change is considered important, which is an issue of interpretability.

Standard 5. Often kappa is considered as a measure of agreement, however, kappa is a measure of reliability (72). An appropriate parameter of measurement error (also called agreement) of dichotomous/nominal/ordinal PROM scores is percentage agreement.

5.6 Assessing risk of bias in a study on criterion validity (box 8)

Criterion validity refers to the degree to which the scores of a PROM are an adequate reflection of a 'gold standard'. In a systematic review, the review team should determine what reasonable 'gold standards' are for the construct to be measured. All studies comparing a PROM to this accepted gold standard can be considered 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

Standards 2 and 3. When both the PROM and the gold standard have continuous scores, correlation is the preferred statistical method. When the instrument scores are continuous and scores on the gold standard are dichotomous the area under the receiver operating characteristic (ROC) is the preferred method, and when both scores are dichotomous sensitivity and specificity are the preferred methods to use.

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	

Standard 3: Bias may occur, for example, when a long version of a questionnaire is compared to a short version, while the scores of the short version were computed using the responses obtained with the longer version. In that case, this standard could be rated as inadequate.

5.7 Assessing risk of bias in a study on hypotheses testing for construct validity (box 9)

Hypotheses testing for construct validity refers to the degree to which the scores of a PROM are consistent with hypotheses (for instance with regard to internal relationships, relationships to scores of other instruments, or differences between relevant groups) based on the assumption that the PROM validly measures the construct to be measured. Hypotheses testing is an ongoing, iterative process (34). The more specific the hypotheses are and the more hypotheses are being tested, the more evidence is gathered for construct validity. Many types of hypotheses can be tested to evaluate construct validity of a PROM. In general, these hypotheses concern comparisons with other outcome measurement instruments (that are not being considered as a gold standard), or differences in scores between 'known' groups (e.g. patients with multiple sclerosis (MS) and spasticity were expected to score statistically significantly higher on an arm and hand functioning scale, as compared with patients with MS without spasticity, because spasticity is a causative for activity limitations due to impairments of the arm and hand (68)). The box on hypotheses testing is therefore structured in two sections. To assess the risk of bias of studies comparing the PROM to comparison instruments, part a of the box (items 1-4) should be completed. For assessing the risk of bias of known group validity studies, part b of the box (items 5-7) should be completed. The risk of bias for studies in which MI or DIF was investigated can be assessed using Box 5 Cross-cultural validity.

The overall rating is the lowest rating given on any applicable item per part of the box. If in an article the PROM under study is being compared to both (a) comparison instrument(s), as well as different subgroups are compared, the box should be completed twice, and results are handled as two sub studies. However, when the ratings are the same, they can be combined. For example, in Appendix 7 (overview Table) reference 5 tested 5 hypotheses, and was rated as 'adequate'. These 5 hypotheses concerned both comparisons with other instruments as well as expected differences between groups.

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 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 instrument(s), OR evidence of insufficient measurement properties of the comparator instrument(s)	

Standards 1 and 2: When the PROM under review is compared to another instrument, it should be known what the construct is that the comparator instruments measures, and the instrument itself should be of sufficient quality. If not, it is difficult to interpret the results of (e.g.) correlations.
This information can be reported in the article included in the review. However, when this is not described, the review team could try to find additional information to decide on rating these standards.

When multiple comparator instruments are being used in a study, and for one instrument the construct is clearly described, and for the other instrument it is not clearly described, we recommend to consider each analyses as a separate study, and complete the box multiple times. We recommend to do the same in case some of the comparator instruments are of adequate quality, while other are of doubtful or inadequate quality.

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	

Standard 3. Appropriate statistical methods could be correlations between the PROM and the comparator instrument. When, for example, Pearson correlations were calculated, but the distribution of scores or mean scores (SD) were not presented, it could be considered 'assumable' that the statistical methods were appropriate. P-values should not be used in testing hypotheses, because it is not relevant to examine whether

correlations statistically differ from zero (75). The validity issue is about whether the direction and magnitude of a correlation is similar to what could be expected based on the construct(s) that are being measured.
Another example of an appropriate method is when CFA or Structural Equation Modeling (SEM) is performed over multiple scales or subscales which are considered to measure similar or different constructs to examine whether subscales measuring similar constructs are more related to each other than subscales measuring different constructs.

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		

Standard 5. When PROM scores between two subgroups are compared, important characteristics of the groups should be reported, and it should be clear on which characteristics the groups differ, for example age, gender, disease characteristics, language etc..

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	

Standard 6. Many different hypotheses can be formulated and tested. The users of the COSMIN Risk of Bias checklist have to decide whether or not the statistical methods used in the article are adequate for testing the stated hypotheses. P-values should be avoided in testing hypotheses, because it is not relevant to examine whether correlations statistically differ from zero (75). The validity issue is about whether the direction and magnitude of a correlation is similar to what could be expected based on the construct(s) that are being measured. When assessing differences between groups, it is also less relevant whether these differences are statistically significant (which depends on the sample size) than whether these differences are as large as expected.

5.8 Assessing risk of bias in a study on responsiveness (box 10)

Responsiveness refers to the ability of a PROM to detect change over time in the construct to be measured. Although responsiveness is considered to be a separate measurement property from validity, the only difference between cross-sectional (construct and criterion) validity and responsiveness is that validity refers to the validity of a single score, and responsiveness refers to the validity of a change score (21). Therefore, the standards for responsiveness are similar to the standards of construct and criterion validity. Although the approach is similar to construct and criterion

validity, we do not to use the terms construct and criterion responsiveness, because these terms are unfamiliar in the literature.
Similarly as for construct and criterion validity, the design requirements for assessing responsiveness are divided in situations in which a gold standard is available (part a standards 1-5), situations in which hypotheses are tested between change scores on the PROM under review and change scores on comparator instruments which are not gold standards (part b standards 6-9), and situations in which hypotheses are tested about expected differences in changes between subgroups (part c standards 10-12). In addition, standards are included for situations in which hypotheses are tested about the expected magnitude of an intervention (part d standards 13-15).

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 Area under the ROC Curve (AUC) calculated			Correlations or AUC 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

Standards 1-2 are similar as the standards for criterion validity.

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 instrument(s) OR evidence of insufficient quality of comparator instrument(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	

Standards 4-7 are similar as the standards for hypotheses testing for construct validity part a.

10c. Construct approach: (i.e. hypotheses testing: comparison between subgroups)					
Design requirements	**very good**	**adequate**	**doubtful**	**inadequate**	**NA**
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	

Standards 8 – 9 are similar as the standards for hypotheses testing for construct validity part b.

10d. Construct approach: (i.e. hypotheses testing: before and after intervention)					
Design requirements	**very good**	**adequate**	**doubtful**	**inadequate**	**NA**
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	

An effect size only has meaning as a measure of responsiveness if we know (or assume) beforehand what the magnitude of the effect of the intervention is. If, for example, we expect a large effect of the intervention on the construct measured by the PROM we can test the hypothesis that the intervention results in an effect size of 0.8 or higher. But if we expect a small effect of the intervention, we would not expect such a high effect size. This example shows that a high effect size does not necessarily indicates a good responsiveness. So, if the effectiveness of an intervention or the change in health status on the construct of interest is known, one could use effect sizes (mean change score / SD baseline)(76), and related measures, such as standardised response mean (mean change score / SD change score) (77), Norman's responsiveness coefficient (σ^2 change/ σ^2 change + σ^2 error) (78), and relative efficacy statistic ((t-statistic$_1$ / t-statistic$_2$)^2) (79) to evaluate responsiveness.

When several instruments are compared in the same study, this could give evidence for the relative responsiveness of the instruments. But again, only if a hypothesis is being tested including the expected magnitude of the treatment effect. Let us propose that we have three measurement instruments (A, B, and C), all measuring the same construct. The intervention given is expected to moderately affect the construct measured by the three instruments. Results show that instrument A has an effect size of 0.8, instrument B of 0.40 and instrument C of 0.15. Based on our hypothesis of a moderate effect we should conclude that instrument B appears to best measure the construct of interest. Instrument A seems to over-estimates the treatment effect (e.g. because it shows change in persons who do not really change), and instrument A seems to under-estimates it. This example shows that it may not always be appropriate to conclude that the instrument with the highest effect size is the most responsive.

<u>Inappropriate measures for responsiveness</u>
Guyatt's responsiveness ratio (MIC/SD change score of stable patients)(80) is considered to be an inappropriate measure of responsiveness, because it takes the minimal important change into account. Minimal important change concerns the interpretation of the change score, not the validity of the change score. The paired t-test is also considered to be an inappropriate measure of responsiveness because it is a measure of significant change instead of valid change, and it is dependent on the sample size of the study (75).

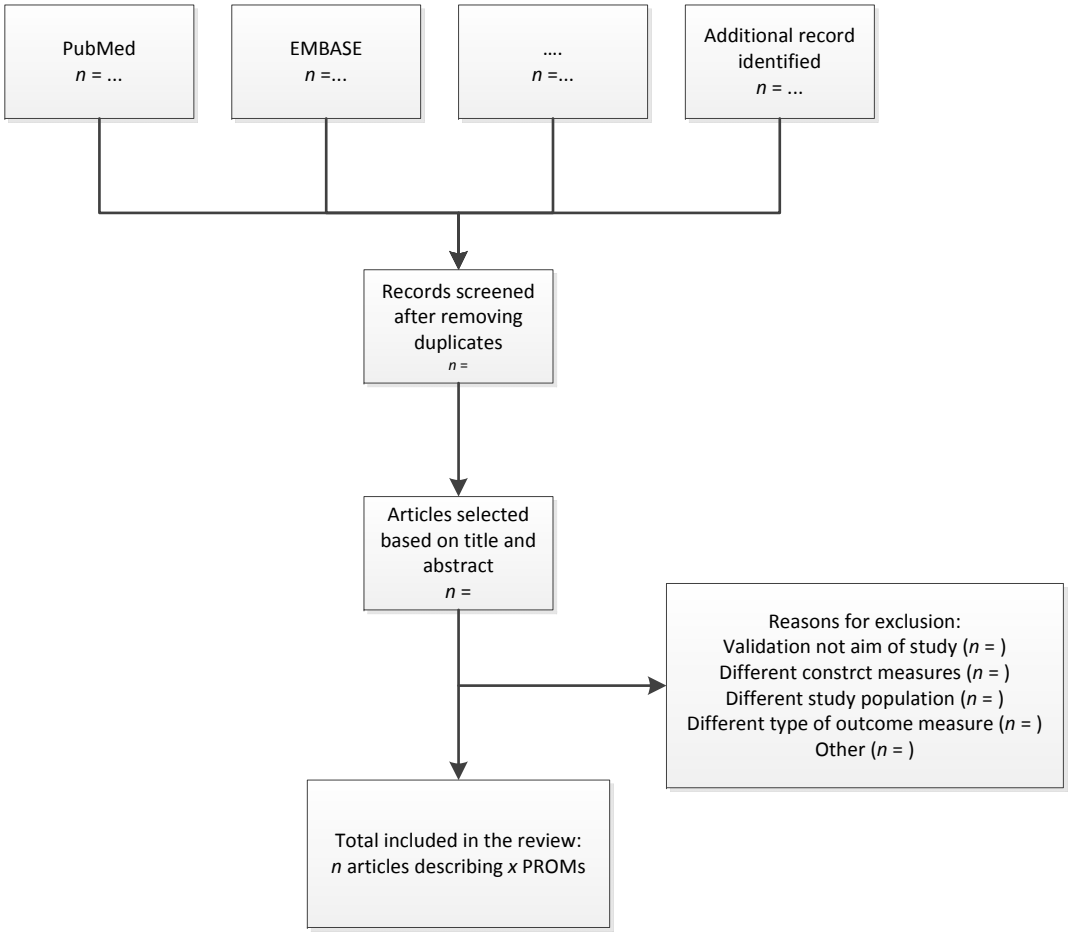
Appendix 1. Example of a search strategy (81)

1) construct
#1: (Depressive disorder[mh] OR depression[mh] OR (depress*[tiab] NOT medline[sb]))

(2) population
#2: (Diabet*[tiab])

(3) sensitive search filter developed by Terwee et al. (27)

Appendix 2. Example of a flowchart



Appendix 3. Example of a table on characteristics of the included PROMs

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.

Other characteristics which may be extracted are: 'administration time'; 'year of development'; 'conceptual model used'; 'recommended by standardization initiatives for [a specific patient population or for the construct to be measures]', 'number of studies evaluating the instrument'; 'completion time (minutes)'; 'Full copy available'; 'licensing information and costs'.

Appendix 4. Example of a table on characteristics of the included study populations

		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										

Other characteristics which may be extracted are 'study design', 'patient selection'.

Appendix 5. Information to extract on interpretability of PROMs

The content of this table is based on the Box Interpretability from the original COSMIN Checklist (1)

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)						
…						

Appendix 6. Information to extract on feasibility of PROMs

The content of this table is based on the guideline for selecting PROMs for Core Outcome Sets (5)

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				

Appendix 7. Table on results of studies on measurement properties

Fictional example of results of the measurement properties

PROM (ref)	**Country (language) in which the questionnaire was evaluated**	**Structural validity**			**Internal consistency**			**Cross-cultural validity\ measurement invariance**			**Reliability**		
		n	Meth qual	Result (rating)	n	Meth qual	Result (rating)	n	Meth qual	Result (rating)	n	Meth qual	Result (rating)
PROM A (ref 1)	US	878	Very good	Unidimensional scale (CFI 0,97, TLI 0,97) (+)	878	Very good	Cronb. Alpha = 0.92 (+)						
PROM A (ref 2)	Dutch				573	Very good	PSI = 0.94 (+)						
PROM A (ref 3)	Spanish				43	Very good	Cronb alpha = 0.91 (+)				84	Very good	ICC (agreement) = 0.85 (+)
PROM A (ref 4)	Dutch										113	adequate	ICC = 0.93 (+)
PROM A (ref 5)	UK										78	inadequate	Spearman rho = 0.94 (+)
Pooled or summary result (overall rating)		**878**		**1 factor (+)**	**1494**		**0.91-0.94 (+)**				**275**		**0.85-0.94 (+)**

PROM	**Country (language) in which the questionnaire was evaluated**	**Measurement error**			**Criterion validity**			**Hypotheses testing**			**Responsiveness**		
		n	Meth qual	Result (rating)	n	Meth qual	Result (rating)	n	Meth qual	Result (rating)	n	Meth qual	Result (rating)
PROM A (ref 2)	Dutch							573	Very good	Results in line with 2 hypo's (2+)			
PROM A (ref 5)	UK							78	adequate	Results in line with 5 hypo's (5+)			
PROM A (ref 6)	US							154	Adequate	Results in line with 3 hypo's (3+)			
									inadequate	Result in line with 1 hypo (1+) Results not in line with 2 hypo's (2-)			
Pooled or summary result (overall rating)								**805**		**11+ and 2- overall (+)**			

Appendix 8. Summary of Findings Tables

Structural validity	**Summary or pooled result**	**Overall rating**	**Quality of evidence**
PROM A	Unidimensional score	sufficient	High (as there is one very good study available)
PROM B			
PROM C			

Internal consistency	**Summary or pooled result**	**Overall rating**	**Quality of evidence**
PROM A	summarized Cronbach alpha / PSI = 0.91-0.94; total sample size: 1494	sufficient	High: multiple very good studies, consistent results
PROM B			
PROM C			

Cross-cultural validity\measurement invariance	**Summary or pooled result**	**Overall rating**	**Quality of evidence**
PROM A	No info available	No info available	
PROM B			
PROM C			

Reliability	**Summary or pooled result**	**Overall rating**	**Quality of evidence**
PROM A	ICC range 0.85 – 0.93; consistent	sufficient	High: one very good study, and

	results; sample size: 275		consistent results
PROM B			
PROM C			

Measurement error	**Summary or pooled result**	**Overall rating**	**Quality of evidence**
PROM A	No info available	No info available	
PROM B			
PROM C			

Hypotheses testing	**Summary or pooled result**	**Overall rating**	**Quality of evidence**
PROM A	11 out of 13 hypotheses confirmed	sufficient	High: as the unconfirmed hypotheses come from inadequate studies, we ignore these results.
PROM B			
PROM C			

Responsiveness	**Summary or pooled result**	**Overall rating**	**Quality of evidence**
PROM A			
PROM B			
PROM C			

6. References

1. Mokkink LB, Terwee CB, Patrick DL, Alonso J, Stratford PW, Knol DL, et al. The COSMIN checklist for assessing the methodological quality of studies on measurement properties of health status measurement instruments: an international Delphi study. Qual Life Res. 2010;19(4):539-49.
2. Prinsen CA, Mokkink LB, Bouter LM, Alonso J, Patrick DL, Vet HC, et al. COSMIN guideline for systematic reviews of outcome measurement instruments. Qual Life Res. 2018;[Epub ahead of print].
3. Ioannidis JP, Greenland S, Hlatky MA, Khoury MJ, Macleod MR, Moher D, et al. Increasing value and reducing waste in research design, conduct, and analysis. Lancet. 2014;383(9912):166-75.
4. Walton MK, Powers JA, Hobart J, al. e. Clinical outcome assessments: A conceptual foundation – Report of the ISPOR Clinical Outcomes Assessment Emerging Good Practices Task Force. Value Health. 2015;18:741-52.
5. Prinsen CA, Vohra S, Rose MR, Boers M, Tugwell P, Clarke M, et al. How to select outcome measurement instruments for outcomes included in a "Core Outcome Set" - a practical guideline. Trials. 2016;17(1):449.
6. Mokkink LB, Vet HC, Prinsen CA, Patrick D, Alonso J, Bouter LM, et al. COSMIN Risk of Bias checklist for systematic reviews of Patient-Reported Outcome Measures. Qual Life Res. 2018;[Epub ahead of print].
7. Terwee CB, Prinsen CA, Chiarotto A, Vet HC, Westerman MJ, Patrick DL, et al. COSMIN standards and criteria for evaluating the content validity of health-related Patient-Reported Outcome Measures: a Delphi study. 2018.
8. Collins NJ, Prinsen CA, Christensen R, Bartels EM, Terwee CB, Roos EM. Knee Injury and Osteoarthritis Outcome Score (KOOS): systematic review and meta-analysis of measurement properties. Osteoarthritis Cartilage. 2016;24(8):1317-29.
9. Gerbens LA, Chalmers JR, Rogers NK, Nankervis H, Spuls PI, Harmonising Outcome Measures for Eczema i. Reporting of symptoms in randomized controlled trials of atopic eczema treatments: a systematic review. Br J Dermatol. 2016;175(4):678-86.
10. Chinapaw MJ, Mokkink LB, van Poppel MN, van MW, Terwee CB. Physical activity questionnaires for youth: a systematic review of measurement properties. Sports Med. 2010;40(7):539-63.
11. Speksnijder CM, Koppenaal T, Knottnerus JA, Spigt M, Staal JB, Terwee CB. Measurement Properties of the Quebec Back Pain Disability Scale in Patients With Nonspecific Low Back Pain: Systematic Review. Phys Ther. 2016;96(11):1816-31.
12. Terwee CB, Mokkink LB, Knol DL, Ostelo RW, Bouter LM, de Vet HC. 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):651-7.
13. Mokkink LB, Terwee CB, Knol DL, Stratford PW, Alonso J, Patrick DL, et al. The COSMIN checklist for evaluating the methodological quality of studies on measurement properties: a clarification of its content. BMC Med Res Methodol. 2010;10:22.
14. Mokkink LB, Terwee CB, Stratford PW, Alonso J, Patrick DL, Riphagen I, et al. Evaluation of the methodological quality of systematic reviews of health status measurement instruments. Qual Life Res. 2009;18(3):313-33.
15. Terwee CB, Prinsen CA, Ricci Garotti MG, Suman A, de Vet HC, Mokkink LB. The quality of systematic reviews of health-related outcome measurement instruments. Qual Life Res. 2016;25(4):767-79.

16. Higgins JP, Green S. Cochrane Handbook for Systematic Reviews of Interventions Version 5.1.0 [updated March 2011]. The Cochrane Collaboration, 2011. Available from: www.handbook.cochrane.org.
17. Cochrane Hanbook for Systematic reviews of Diagnostic Test Accuracy Reviews 2013 [Available from: http://methods.cochrane.org/sdt/handbook-dta-reviews.
18. Peterson DAB, P.; Jabusch, H. C.; Altenmuller, E.; Frucht, S. J. Rating scales for musician's dystonia: the state of the art. Neurology. 2013;81(6):589-98.
19. Herr KB, K.; Decker, S. Tools for assessment of pain in nonverbal older adults with dementia: A state-of-the-science review. Journal of Pain and Symptom Management. 2006;31(2):170-92.
20. GRADE. GRADE Handbook - Handbook for grading the quality of evidence and the strength of recommendations using the GRADE approach. 2013.
21. Mokkink LB, Terwee CB, Patrick DL, Alonso J, Stratford PW, Knol DL, et al. The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes. J Clin Epidemiol. 2010;63(7):737-45.
22. Streiner DL, Norman G. Health Measurement Scales. A practical guide to their development and use. 4th edition ed. New York: Oxford University Press; 2008.
23. de Vet HC, Terwee CB, Mokkink L, Knol DL. Measurement in Medicine: a practical guide: Cambridge University Press; 2010.
24. Fayers PM, Hand DJ, Bjordal K, Groenvold M. Causal indicators in quality of life research. Qual Life Res. 1997;6(5):393-406.
25. Elbers RG, Rietberg MB, van Wegen EE, Verhoef J, Kramer SF, Terwee CB, et al. Self-report fatigue questionnaires in multiple sclerosis, Parkinson's disease and stroke: a systematic review of measurement properties. Qual Life Res. 2012;21(6):925-44.
26. Griffiths C, Armstrong-James L, White P, Rumsey N, Pleat J, Harcourt D. A systematic review of patient reported outcome measures (PROMs) used in child and adolescent burn research. Burns. 2015;41(2):212-24.
27. Terwee CB, Jansma EP, Riphagen, II, de Vet HC. Development of a methodological PubMed search filter for finding studies on measurement properties of measurement instruments. Qual Life Res. 2009;18(8):1115-23.
28. Clanchy KMT, S. M.; Boyd, R. Measurement of habitual physical activity performance in adolescents with cerebral palsy: a systematic review. DevMed Child Neurol. 2011;53(6):499-505.
29. PRISMA Statement. http://prisma-statement org/ [Internet]. 2016 10/24/2016. Available from: http://prisma-statement.org/.
30. Terwee CB, Bot SD, de Boer MR, van der Windt DA, Knol DL, Dekker J, et al. Quality criteria were proposed for measurement properties of health status questionnaires. J Clin Epidemiol. 2007;60(1):34-42.
31. Ernst MY, Albert J.; Kriston, Levente; Schonfeld, Michael H.; Vettorazzi, Eik; Fiehler, Jens. Is visual evaluation of aneurysm coiling a reliable study end point? Systematic review and meta-analysis. Stroke. 2015;46(6):1574-81.
32. DerSimonian R, Laird N. Meta-analysis in clinical trials revisited. Contemp Clin Trials. 2015;45(Pt A):139-45.
33. Terwee CB, Prinsen CA, de Vet HCW, Bouter LM, Alonso J, Westerman MJ, et al. COSMIN methodology for assessing the content validity of Patient-Reported Outcome Measures (PROMs). User manual. 2017.
34. Strauss ME, Smith GT. Construct Validity: Advances in Theory and Methodology. Annu Rev Clin Psychol. 2008.

35. Cronbach LJ, Meehl PE. Construct validity in psychological tests. Psychological Bulletin. 1955;52:281-302.
36. McDowell I, Jenkinson C. Development standards for health measures. J Health Serv Res Policy. 1996;1:238-46.
37. Messick S. The standard problem. Meaning and values in measurement and evaluation. . American Psychologist. 1975;oct:955-66.
38. de Vet HC, Terwee CB. The minimal detectable change should not replace the minimal important difference. J Clin Epidemiol. 2010;63(7):804-5.
39. de Vet HC, Ostelo RW, Terwee CB, van der Roer N, Knol DL, Beckerman H, et al. Minimally important change determined by a visual method integrating an anchor-based and a distribution-based approach. Qual Life Res. 2007;16(1):131-42.
40. de Vet HC, Terwee CB, Ostelo RW, Beckerman H, Knol DL, Bouter LM. Minimal changes in health status questionnaires: distinction between minimally detectable change and minimally important change. Health Qual Life Outcomes. 2006;4:54.
41. Crosby RD, Kolotkin RL, Williams GR. Defining clinically meaningful change in health-related quality of life. J Clin Epidemiol. 2003;56(5):395-407.
42. Yost KJ, Eton DT, Garcia SF, Cella D. Minimally important differences were estimated for six Patient-Reported Outcomes Measurement Information System-Cancer scales in advanced-stage cancer patients. J Clin Epidemiol. 2011;64(5):507-16.
43. van Kampen DA, Willems WJ, van Beers LW, Castelein RM, Scholtes VA, Terwee CB. Determination and comparison of the smallest detectable change (SDC) and the minimal important change (MIC) of four-shoulder patient-reported outcome measures (PROMs). Journal of orthopaedic surgery and research. 2013;8:40.
44. Sprangers MA, Schwartz CE. Integrating response shift into health-related quality of life research: a theoretical model. Soc Sci Med. 1999;48(11):1507-15.
45. Boers M, Kirwan JR, Tugwell P, Beaton D, Bingham CO, III, Conaghan PG, et al. The OMERACT handbook: OMERACT; 2015.
46. Smart A. A multi-dimensional model of clinical utility. International journal for quality in health care : journal of the International Society for Quality in Health Care. 2006;18(5):377-82.
47. Fayers PM, Hand DJ. Factor analysis, causal indicators and quality of life. Qual Life Res. 1997;6(2):139-50.
48. Streiner DL. Being inconsistent about consistency: when coefficient alpha does and doesn't matter. J Pers Assess. 2007;80:217-22.
49. Wirth RJ, Edwards MC. Item factor analysis: Current approaches and future directions. Psychological Methods. 2007;12:58-79.
50. Floyd FJ, Widaman FK. Factor analysis in the development and refinement of clinical assessment instruments. . Psychological Assessment. 1995;7:286-99.
51. Comrey AL, Lee HB. A first course in factor analysis. 2nd ed. ed. Hillsdale, NJ: Erlbaum; 1992 1992.
52. Brown T. Confirmatory Factor Analysis for applied research. New York: The Guilford Press; 2015.
53. Embretson SE, Reise SP. Item response theory for psychologists. Mahwah, New Jersey: Lawrence Erlbaum Associates; 2000.
54. Edelen MO, Reeve BB. Applying item response theory (IRT) modeling to questionnaire development, evaluation, and refinement. Qual Life Res. 2007;16 Suppl 1:5-18.
55. Reise SPY, J. Parameter recovery in the graded response model using MULTILOG. JEM. 1990;27:133-44.

56. Reise SPM, T.M.; Maydeu-Olivares, A. Targeted bifactor rotations and assessing the impact of model violations on the parameters of unidimensional and bifactor models. Journal of Educational and Psychological Measurement. 2011;71:684-711.
57. Linacre JM. Sample size and item calibration stability. Rasch Measurement Transactions. 1994;7(4):328.
58. de Vet HC, Ader HJ, Terwee CB, Pouwers F. Are factor analytical techniques appropriately used in the validation of health status questionnaires? A systematic review on the quality of factor analyses of the SF-36. . Quality of Life Research. 2005;14:1203-18.
59. Cortina JM. What is coefficient alpha? An examination of theory and applications. Appl Psychology. 1993;78:98-104.
60. Cronbach LJ. Coefficient Alpha and the Internal Structure of Tests. Psychometrika. 1951;16:297-334.
61. Streiner DL. Figuring out factors: The use and misuse of factor analysis. Can J Psychiatry. 1994;39:135-40.
62. Revelle WZRE. Coefficients Alpha, Beta, Omega, and the GLB: Comments on Sijtsma. Psychometrika. 2009;74(1):145-54.
63. Crane PK, Gibbons LE, Jolley L, Van Belle G. Differential item functioning analysis with ordinal logistic regression techniques. DIFdetect and difwithpar. Med Care. 2006;44(11 Suppl 3):S115-23.
64. Petersen MA, Groenvold M, Bjorner JB, Aaronson N, Conroy T, Cull A, et al. Use of differential item functioning analysis to assess the equivalence of translations of a questionnaire. Qual Life Res. 2003;12(4):373-85.
65. Gregorich SE. Do Self-Report Instruments Allow Meaningful Comparisons Across Diverse Population Groups? Testing Measurement Invariance Using the Confirmatory Factor Analysis Framework. Medical Care. 2006;44(11 Suppl 3):S78-S94.
66. Teresi JA, Ocepek-Welikson K, Kleinman M, Eimicke JP, Crane PK, Jones RN, et al. Analysis of differential item functioning in the depression item bank from the Patient Reported Outcome Measurement Information System (PROMIS): An item response theory approach. Psychol Sci Q. 2009;51(2):148-80.
67. Scott NW, Fayers PM, Aaronson NK, Bottomley A, De Graeff A, Groenvold M, et al. A simulation study provided sample size guidance for differential item functioning (DIF) studies using short scales. Journal of Clinical Epidemiology. 2009;62:288-95.
68. Van Leeuwen LM, Mokkink LB, Kamm CP, de Groot V, van den Berg P, Ostelo RW, et al. Measurement properties of the Arm Function in Multiple Sclerosis Questionnaire (AMSQ): a study based on Classical Test Theory. Disabil Rehabil. 2016:1-8.
69. Shrout PE, Fleiss JL. Intraclass Correlations: Uses in assessing rater reliability. Psychological Bulletin. 1979;86:420-8.
70. McGraw KOW, S.P. Forming inferences about some intraclass correlation coefficients. Psychological Methods. 1996;1:30-46.
71. Cohen J. Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. Psychological Bulletin. 1968;70:213-20.
72. de Vet HC, Mokkink LB, Terwee CB, Hoekstra OS, Knol DL. Clinicians are right not to like Cohen's kappa. BMJ. 2013;346:f2125.
73. Fleiss JL, Cohen J. The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. . Educational and Psychological Measurement. 1973;33:613-9.
74. de Vet HC, Terwee CB, Knol DL, Bouter LM. When to use agreement versus reliability measures. . Journal of Clinical Epidemiology. 2006;59:1033-9.
75. Altman DG. Practical statistics for medical research. London: Chapman and Hall; 1991.

76. Cohen J. Statistical power analysis for the behavioural sciences. . 2nd ed. ed. Hillsdale, NJ: Lawrence Erlbaum Associates; 1988.
77. McHorney CAT, A.R. Individual-patient monitoring in clinical practice: are available health status surveys adequate? Qual Life Res. 1995;4:293-307.
78. Norman GR. Issues in the use of change scores in randomized trials. J Clin Epidemiol. 1989;42(11):1097-105.
79. Stockler MR, Osoba D, Goodwin P, Corey P, Tannock IF. Responsiveness to change in health-related quality of life in a randomized clinical trial: a comparison of the Prostate Cancer Specific Quality of Life Instrument (PROSQOLI) with analogous scales from the EORTC QLQ-C30 and a trial specific module. European Organization for Research and Treatment of Cancer. J Clin Epidemiol. 1998;51(2):137-45.
80. Guyatt G, Walter S, Norman G. Measuring change over time: assessing the usefulness of evaluative instruments. J Chronic Dis. 1987;40(2):171-8.
81. van Dijk SEA, M.C.; van der Zwaan, L.; Bosmans, J.E.; van Marwijk, H.W.J.; van Tulder, M.W.; Terwee, C.B. Measurement properties of questionnaires for depressive symptoms in adult patients with type 1 or type 2 diabetes: a systematic review. Qual Life Res. 2018; doi: 10.1007/s11136-018-1782-y. [Epub ahead of print].