COSMIN methodology for assessing the content validity of PROMs

User manual

version 1.0

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

Contact
CB Terwee, PhD
VU University Medical Center
Department of Epidemiology and Biostatistics
Amsterdam, The Netherlands
Website: www.cosmin.nl
E-mail: cb.terwee@vumc.nl

Table of contents

Preface

Content validity is the most important measurement property of a patient-reported outcome measure (PROM). The aim of this document is to explain how the content validity of PROMs can be evaluated. The methodology described in this manual was primarily developed to evaluate the content validity of PROMs based on evidence from the literature, e.g. in systematic reviews of PROMs. It is not a guideline for designing a study on content validity, although the standards described in this manual indicate the important design requirements of a content validity study.

This methodology was developed in 2016 in a Delphi study among 158 experts from 21 countries [2]. We thank all researchers who participated voluntarily in the Delphi study. Their names are listed in Appendix 1. Please refer to this paper when using the COSMIN methodology for content validity:

Terwee, C. B., Prinsen, C. A. C., Chiarotto, A., De Vet, H. C. W., Westerman, M. J., Patrick, D. L., Alonso, J., Bouter, L. M., Mokkink, L. B. (2017). COSMIN standards and criteria for evaluating the content validity of health-related Patient-Reported Outcome Measures: a Delphi study. Qual Life Res in press.

We hope that this methodology will result in a growing recognition of the importance of content validity of PROMs and in a growth of publications reporting in detail aspects of PROM development and evaluation.

February 2018
The COSMIN steering committee

Background: COSMIN tools

The **CO**nsensus-based **S**tandards for the selection of health **M**easurement **IN**struments (COSMIN) initiative (www.cosmin.nl) aims to improve the selection of health outcome measurement instruments in research and clinical practice, by developing standards and criteria for evaluating the measurement properties of outcome measurement instruments. The COSMIN initiative has developed several tools:

COSMIN taxonomy: Which measurement properties are important and how are they defined?

In the first international COSMIN Delphi study, performed in 2006-2007, COSMIN developed a taxonomy of measurement properties [3]. Nine measurement properties were considered relevant for health-related Patient-Reported Outcome Measures (PROMs), categorized into three broad domains:

- Reliability, containing the measurement properties internal consistency, reliability, and measurement error;
- Validity, containing the measurement properties content validity (including face validity), criterion validity, and construct validity (including structural validity, hypotheses testing, and cross-cultural validity);
- Responsiveness, containing the measurement property responsiveness.

Interpretability was also recognized as an important aspect of a measurement instrument, but not a measurement property. Definitions of the nine measurement properties and interpretability are provided on the COSMIN website.

COSMIN checklist for evaluating the methodological quality of studies on measurement properties

In the Delphi study mentioned above, the COSMIN initiative also reached consensus on standards for evaluating the methodological quality (risk of bias) of studies on measurement properties. For each measurement property, standards for design requirements and preferred statistical analyses were agreed upon. The standards were included in the COSMIN checklist, which can be used in systematic reviews of measurement instruments to evaluate the quality of the included studies [1]. A rating system was also developed to rate the quality of a study on a measurement property, using a 4-point rating scale [4].
The COSMIN checklist was originally developed for rating the quality of studies on the measurement properties of PROMs. However, the checklist has also been used for rating the quality of studies on the measurement properties of other measurement instruments (see examples [5-8]).
Recently, a new version of the COSMIN checklist, the COSMIN Risk of Bias checklist for PROMs was developed [9]. The COSMIN boxes for content validity, presented in this manual, are part of the COSMIN Risk of Bias checklist for PROMs.

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 content validity in the second COSMIN Delphi study (Table 1).

COSMIN database of systematic reviews of outcome measurement instruments

The COSMIN initiative systematically collects systematic reviews of outcome measurement instruments. These systematic reviews are important tools for the selection of outcome measurement instruments for research and clinical practice and for identifying gaps in knowledge on the quality of outcome measurement instruments, i.e. their measurement properties. The reviews are available in a searchable database, which is regularly updated: http://database.cosmin.nl/.

COSMIN guideline for performing systematic reviews of outcome measurement instruments

Systematic reviews of PROMs differ from systematic reviews of health interventions and diagnostic test accuracy studies and are quite complex. In fact, multiple reviews (i.e. one review per measurement property) are performed. COSMIN developed a methodological guideline for performing systematic reviews of PROMs [10]. A sequential ten-step procedure for conducting a systematic review of PROMs is described. Steps 1-4 concern preparing and performing the literature search, and selecting relevant articles. Steps 5-8 concern the evaluation of the measurement properties of the PROMs: first by assessing the risk of bias of the included studies (using the COSMIN standards); second, by applying criteria for good measurement properties; and third, by summarizing the evidence and grading the quality of the evidence on measurement properties. Also an evaluation of interpretability and feasibility aspects is included. Steps 9 and 10 concern formulating recommendations, and publishing the systematic review. Step 5 of this guideline concerns the evaluation of content validity, which is described in this manual. The remaining steps of a systematic review of PROMs are described in the manual "COSMIN methodology for systematic reviews of Patient-Reported Outcome Measures (PROMs) – user manual" published on the COSMIN website.

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.

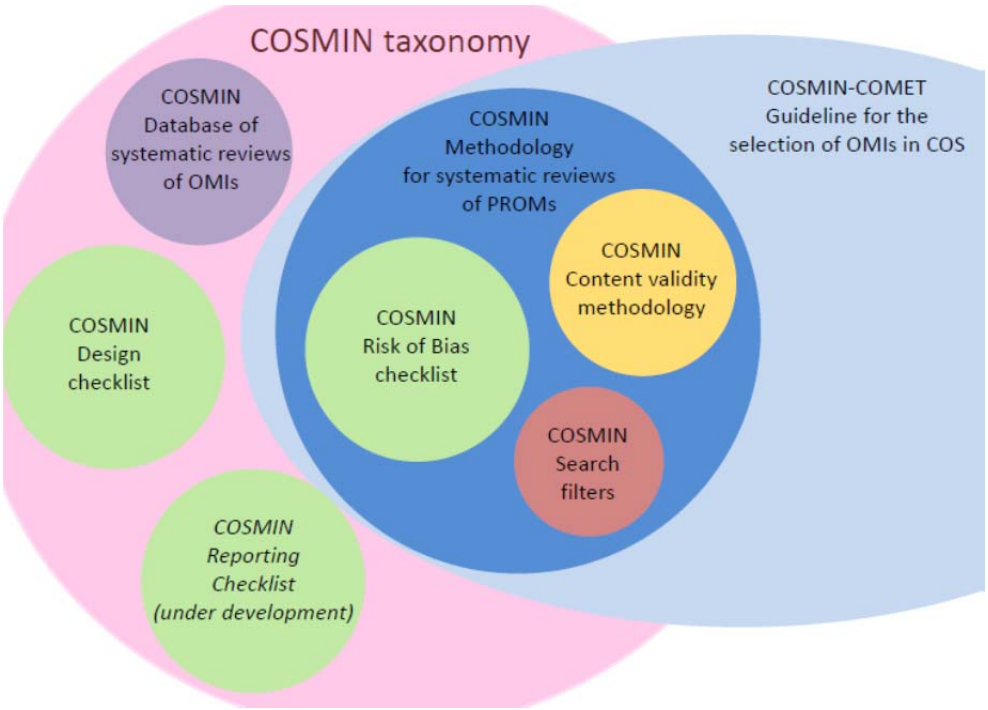
COSMIN/COMET guideline for selection outcome measurement instruments to be included in a Core Outcome Set

In cooperation with the Core Outcome Measures in Effectiveness Trials (COMET) initiative the COSMIN group developed a consensus-based guideline on how to select outcome measurement instruments for outcomes (i.e., constructs or domains) included in a "Core Outcome Set" (COS) [11]. A COS is an agreed minimum set of outcomes that should be measured and reported in all clinical trials of a specific disease or trial population. In this international Delphi study consensus was reached that for an outcome measurement instrument to be selected for a COS, it should have at least high quality evidence of good content validity.

COSMIN methodology for evaluating the content validity of PROMs

The methodology for evaluating the content validity of PROMs is described in this manual. This methodology was developed in 2016 in a Delphi study among 158 experts from 21 countries [2].

More information about COSMIN can be found on the COSMIN website: www.cosmin.nl



Development of the COSMIN methodology for evaluating the content validity of PROMs.

Why was this methodology developed?

Content validity is the degree to which the content of an instrument is an adequate reflection of the construct to be measured [3]. Three aspects of content validity are being distinguished: (1) relevance (all items in a PROM should be relevant for the construct of interest within a specific population and context of use), (2) comprehensiveness (no key aspects of the construct should be missing), and (3) comprehensibility (the items should be understood by patients as intended). Content validity is the most important measurement property of a PROM and the most challenging to assess. Content validity of existing PROMs should be assessed in a content validity study by systematically asking patients and professionals (e.g. clinicians, researchers) about the relevance, comprehensiveness and comprehensibility of the items. The original COSMIN checklist already contained a box for evaluating the quality of content validity studies. However, in this box, the quality of the development of the PROM was not taken into account. Since a well-designed qualitative approach to item construction helps to ensure content validity, new standards were developed for evaluating the quality of PROM development. Furthermore, the original COSMIN standards for content validity only considered *whether* certain things were done, but not *how* they were done. For example, one of the standards addressed whether it was assessed if all items refer to relevant aspects of the construct to be measured. However, there were no standards to evaluate how this should be assessed (e.g. whether a relevant sample of professionals was consulted). Therefore, a second international COSMIN Delphi study was performed in 2016 with the following aims: (1) to develop standards for evaluating the quality of PROM development (box 1); (2) to update the original COSMIN standards for assessing the quality of content validity studies of PROMs (box 2, which replaces the original COSMIN box D); and (3) to develop criteria for what constitutes good content validity of PROMs, and (4) to develop a rating system for summarizing the evidence on a PROM's content validity and grading the quality of the evidence in systematic reviews of PROMs

The COSMIN methodology described in this manual was developed to evaluate the content validity of PROMs, based on the quality and results of the PROM development study, the quality and results of additional content validity studies, and the content of the PROM itself. The methodology was developed for use in a systematic review in which the content validity of a number of PROMs is evaluated and compared in a systematic way, but it can also be used for rating the content validity of single PROMs.

How was this methodology developed?

This methodology was developed in a Delphi study among 158 experts from 21 countries [2]. In each Delphi round, panelists were asked to rate the degree to which they (dis)agreed to proposed standards or criteria on a 5-point rating scale ranging from 'strongly disagree' to 'strongly agree', and to provide arguments for their ratings. Proposed standards and criteria were based on three literature searches: (1) a search used for the development of minimum standards for PROMs used in patient-centered outcomes and comparative effectiveness research [12]; (2) a search on methods for selecting outcome measurement instruments for outcomes included in a COS [11]; and (3) a PubMed search "content validity"[ti], in which 25 methodological papers on content validity were identified [13-37]. In addition, relevant text books and articles were consulted (e.g. ISPOR taskforce papers [30; 31], PROMIS standards for developing Item Response Theory-based item banks (obtained from http://www.healthmeasures.net/explore-measurement-systems/promis/measure-development-research/119-measure-development-research), the BMJ guidelines for qualitative research [38], the APA standards for educational and psychological testing [39] and other relevant papers [40; 41]).

Aim of the methodology: rating the PROM against criteria for content validity

The ultimate aim when evaluating the content validity of a PROM is to judge whether the PROM meets pre-defined criteria for what constitutes good content validity. Ten criteria were developed in the Delphi study, concerning three aspects of content validity of a PROM: relevance, comprehensiveness and comprehensibility (Table 1). The criteria are formulated as questions. When evaluating the content validity of a PROM, these ten questions should be answered and summarized into an overall rating for relevance, comprehensiveness, comprehensibility and overall content validity. Note that the questions should be answered for each score or subscore of a PROM separately. The questions should be answered based on information about how the PROM was developed, evidence from content validity studies, if available, and the PROM content. Hereby, the quality of the PROM development study and the quality of available content validity studies should be taken into account.

Ten criteria for good content validity

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

To answer the ten questions in an evidence-based and transparent way, three steps need to be performed:

Step 1: Evaluate the quality of the PROM development
Step 2: Evaluate the quality of additional content validity studies on the PROM (if available)
Step 3: Evaluate the content validity of the PROM, based on the quality and results of the available studies and the PROM itself

Before these three steps are described in more detail, we will first provide several general recommendations for evaluating the content validity of PROMs in a systematic review of PROMs. We recommend to read this section before starting a systematic review. After that, detailed instructions will be provided for performing the three steps described above, to evaluate the content validity of PROMs.

General recommendations on how to perform a systematic review on the content validity of PROMs

Use the scope of the review as a reference point

Authors of a systematic review of PROMs should clearly define the scope of their review. By scope we mean the construct, target population, and context of use (e.g. discriminative, evaluative, or predictive purpose) of interest in the review. This scope should be the reference point for evaluating content validity of the included PROMs. For example, the scope of a review could be PROMs for assessing physical function (construct) in patients with non-specific acute and chronic low back pain (target population) to be used as outcome measures in clinical trials (context of use: evaluative application).
The content validity of a PROM may be different when using the PROM for measuring different constructs, in different populations, or in different contexts of use. Researchers do not validate an instrument but rather the application of it. The content validity of a PROM may be good in the target population for which the PROM was originally developed, but less good when used in another patient population. Some PROMs may be valid for a wide range of uses in different populations, but each use may require new supporting evidence. PROMs are increasingly applied to populations beyond their original intention, which may eventually lead to inappropriate use of the PROM. When rating the content validity of PROMs in a systematic review, reviewers should rate the content validity of the PROMs for the construct, population, and context of use of interest in their systematic review (see also our recommendations on including indirect evidence below).

Use existing ratings of the quality of PROM development.

One of the steps in a systematic reviews of PROMs is rating the quality of the PROM development (as described in the next sections of this manual). This should be done for each PROM only once. We recommend researchers to share their ratings, because the same PROM can be included in multiple systematic reviews. Once the quality of the PROM development is rated, this information can be used by other researchers and the PROM development does not need to be rated again in subsequent reviews. On the COSMIN website a Table is published with existing ratings of the quality of PROM developments (rated with the COSMIN methodology described in this manual). We encourage reviewers to send their ratings of PROM developments to cb.terwee@vumc.nl to be included in this Table.

Use evidence from studies on translations of PROMs

Studies in which a translation of a PROM is described should be included if a pilot study was performed after translation to evaluate the comprehensibility of the translated PROM. Comprehensibility is one aspect of content validity (next to relevance and comprehensiveness).

Use scoring manuals and other additional information

When rating the content validity of a PROM, we recommend to use additional relevant information such as scoring manuals, development papers, or information from websites. We also recommend to consider contacting the developers of the PROM giving them the opportunity to provide additional information. It is important to have a copy of the PROM because we recommend that reviewers also give their own rating of the content of the PROM (see step 3). It is up to the reviewers to decide whether they are willing to pay for obtaining a copy of a PROM, if that is required. If reviewers do not want to pay for a PROM, they could decide to state in the review that the content of the PROM itself was not evaluated because the PROM was not freely available.

Consider each subscale of a (multi-dimensional) PROM separately

The content validity is rated separately for each set of items of a PROM 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 instrument. For instance, in one subscale all items may be relevant for the construct of the scale, but another subscale may contain irrelevant items for the construct of the scale. And one subscale may be more comprehensively covering the construct of interest than another subscale.
The PROM development probably needs be rated only once, because the quality of the PROM development is likely the same for all subscales. For multidimensional PROMs where subscale scores are added up into a total score, it is also possible to give a rating for the content validity of the total score, by combining the evidence on the content validity of each of the subscales (see step 3b).

Modified PROMs

A modified PROM should, in principle, be treated as a new PROM. However, for rating the content validity of a modified PROM, information on the development of the original version of the PROM and additional content validity studies could be relevant. For example, if a modified PROM is a shortened version of a previously developed PROM, information on the relevance and comprehensibility of the items can be obtained from studies on the original PROM. However, information on the comprehensiveness of the shortened PROM should come from a new study using the shortened PROM. Reviewers should decide which information can be used from studies on the original PROM and which information should come from new studies, using the modified PROM.

Consider indirect evidence

If no content validity studies are performed in the population of interest, one could consider including content validity studies performed in (slightly) different populations. Such studies could provide evidence on the comprehensibility of the PROM, and (although perhaps to a lesser extent) the relevance and comprehensiveness.
Also, if a PROM was developed for a different (or broader) target population than the population of interest of the review, and has good content validity in the original target population, this information could be relevant for the review.

For example: in a systematic review of PROMs for patients with hand Osteoarthritis (OA) the Disabilities of the Arm, Shoulder an Hand (DASH) questionnaire was included as a measure of physical function and symptoms. No content validity study of the DASH has been performed in patients with hand OA. However, the DASH was designed for patients with any or several musculoskeletal disorders of the upper limb and therefore content validity studies performed in other patients with upper extremity problems may provide some evidence on the content validity for hand OA patients. In the data synthesis, described later, this information may be included as indirect evidence of the content validity of the DASH in patients with hand OA (note that indirect evidence will be weighted less than direct evidence obtained in the population of interest).

Also one should consider the target population for which the PROM was developed in relation to the population of interest in the review. In the example above, the DASH was developed for a broader target population (i.e. musculoskeletal disorders of the upper limb) than the population of interest in the review (i.e. hand OA). If only a few patients with hand OA were involved in the PROM development one may not be sure that the items of the DASH are relevant and comprehensive for patients with hand OA. This can be taken into account in the data syntheses (see step 3).

We recommend reviewers to carefully define and report the search strategy and the in- and exclusion criteria of their review, considering which kind of studies may provide direct or indirect evidence and which kind of studies they want to include. We recommend to include all PROMs measuring one or more specific constructs of interest, rather than only the most commonly used PROMs because newer PROMS might be of better quality but less often used [10]. Reviewers should also consider the target audience of the review. By broadening the search (e.g. including all PROMs instead of only PROMs measuring a specific construct of interest) or by including also indirect evidence (from slightly different populations or from a general population), the review may be of interest to a broader audience. The inclusion criteria may be adapted during the review (if possible, considering the search strategy). One could, for example, first consider only the evidence obtained in the population of interest, and then broaden the inclusion criteria to evidence from slightly different populations if limited information is found on a PROM.

Include the required expertise in the review team

We recommend that the review team includes reviewers with at least some knowledge of the construct of interest; experience with the target population; and at least some knowledge or experience with PROM development and evaluation, including qualitative research. Assessing content validity is not an easy task and having reviewers with strong knowledge of PROM development and evaluation is encouraged.
The COSMIN standards and criteria are more detailed, more standardized, and more transparent than earlier published guidelines. Nevertheless, judgment is still needed, for example, about what appropriate qualitative data collection methods are to identify relevant items for a new PROM or for analysing such data. It was considered not possible to define exactly what is considered appropriate due to many possible variations in design and analysis of qualitative studies. For this reason we recommend that the review team includes reviewers with at least some knowledge or experience with qualitative research, who could rate the quality of the PROM development and the quality of qualitative content validity studies. This may make the results of the review also more easily trusted by the scientific and clinical community. Furthermore, professionals with experience with the target population of interest could rate the content of the PROM (see step 3). It may also be considered to include patients as research partners in the review for rating the content of the PROM.

Use two independent reviewers

The review team will need to make judgements when rating the content validity. We therefore recommend that all ratings are done by two reviewers, independently. This is good practice for all kind of systematic reviews. We recommend that reviewers discuss a priori how certain standards or criteria will be rated, based on practicing the ratings with a few articles from the review, and taking the scope of the review into account. All ratings should be compared and consensus reached among the reviewers, if needed, with help of a third reviewer. We recommend to have regular consensus meetings rather than one meeting at the end of the rating process, to discuss rating issues and ensure that issues are rated consistently among papers.

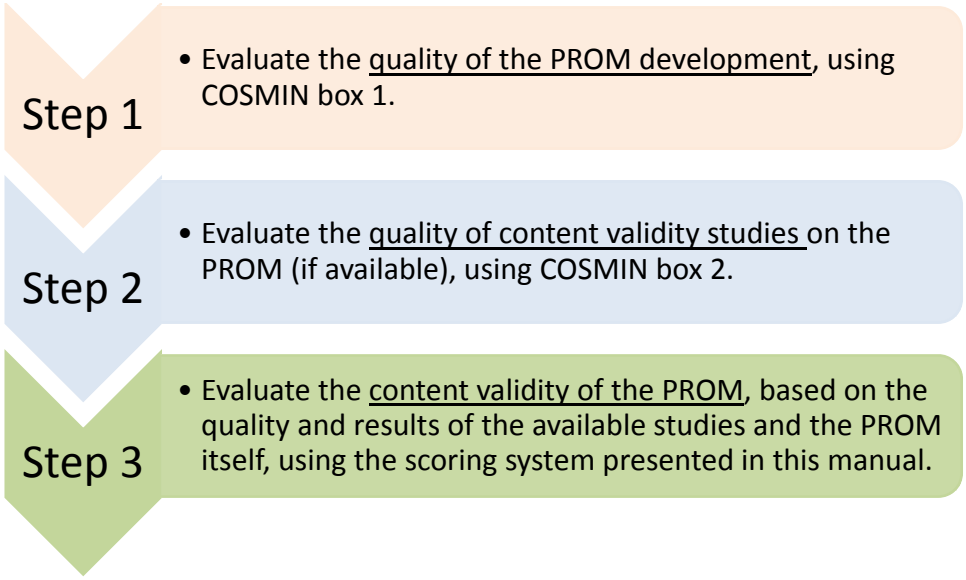
Report a conflict of interest

Sometimes authors of a systematic reviews are also the developers of one of the included PROMs or the authors of one or more of the included content validity studies. This may raise a conflict of interest. We recommend that authors clearly report such potential conflicts of interest in the review paper. We also recommend that the particular PROM or study is rated by an (additional) independent reviewer.

It is important to ensure that the strength of qualitative methods is not lost in an attempt to standardize the evaluation. Therefore, the COSMIN methodology should be used as guidance, leaving the final judgment to the reviewers based on the available evidence and their methodological and clinical expertise.

Instructions for evaluating the content validity of PROMs

The COSMIN methodology for evaluating the content validity of PROMs consists of three steps:



Structure of the COSMIN boxes 1 and 2

The two boxes for rating the risk of bias of the PROM development study and content validity studies (box 1 and 2 respectively), contain of multiple parts each (see boxes below), that should be completed based on the available information. If information on a certain part is lacking (e.g. professionals were not involved in a content validity study), the corresponding part can be skipped.

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

1a. Standards for evaluating the quality of the PROM design to ensure relevance of the PROM

General design requirements

Concept elicitation (relevance and comprehensiveness)

1b. Standards for evaluating the quality of a cognitive interview study or other pilot test performed to evaluate comprehensibility and comprehensiveness of a PROM

General design requirements

Comprehensiveness

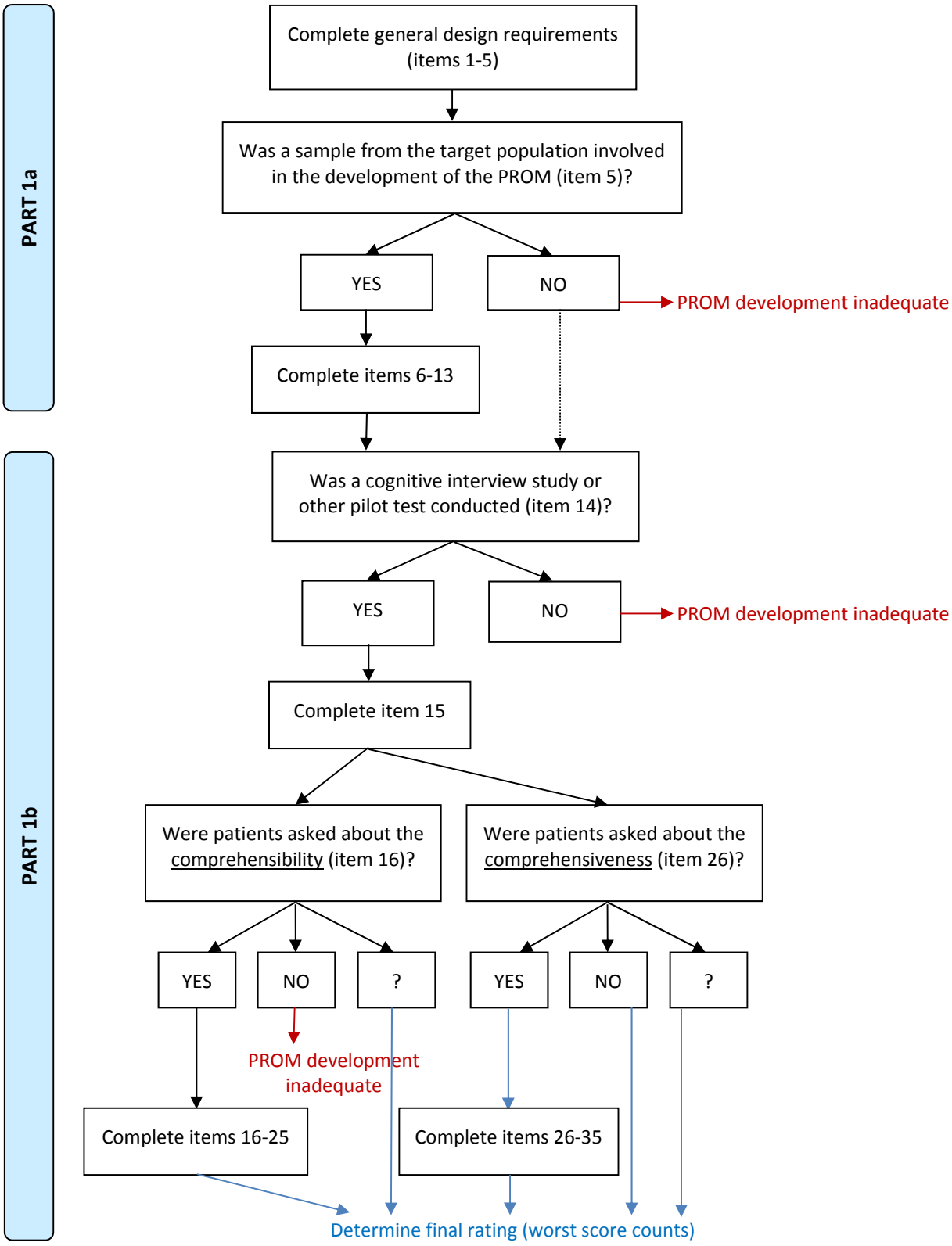
Comprehensibility

The standards in box 1 are divided into two parts: Part 1a concerns standards for evaluating the quality of research performed to identify relevant items for a new PROM. The quality of the concept elicitation study provides information on the relevance and comprehensives of the items in a PROM. Part 1b concerns standards for evaluating the quality of a cognitive interview study or other pilot test (e.g. survey) performed to evaluate comprehensiveness and comprehensibility of the PROM. A cognitive interview study provides additional information on the comprehensiveness and especially the comprehensibility of the items.

Both parts need to be completed for each PROM because all standards of part 1a and part1b will be included in the final rating of the quality of the PROM development (see page 15). If a cognitive interview study or other pilot test was not performed, only the first standard in part 1b needs to be completed and the rest of the box can be skipped.

To decide which parts of box 1 should be completed, the flow chart given below can be used.

Flowchart for completing box 1



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

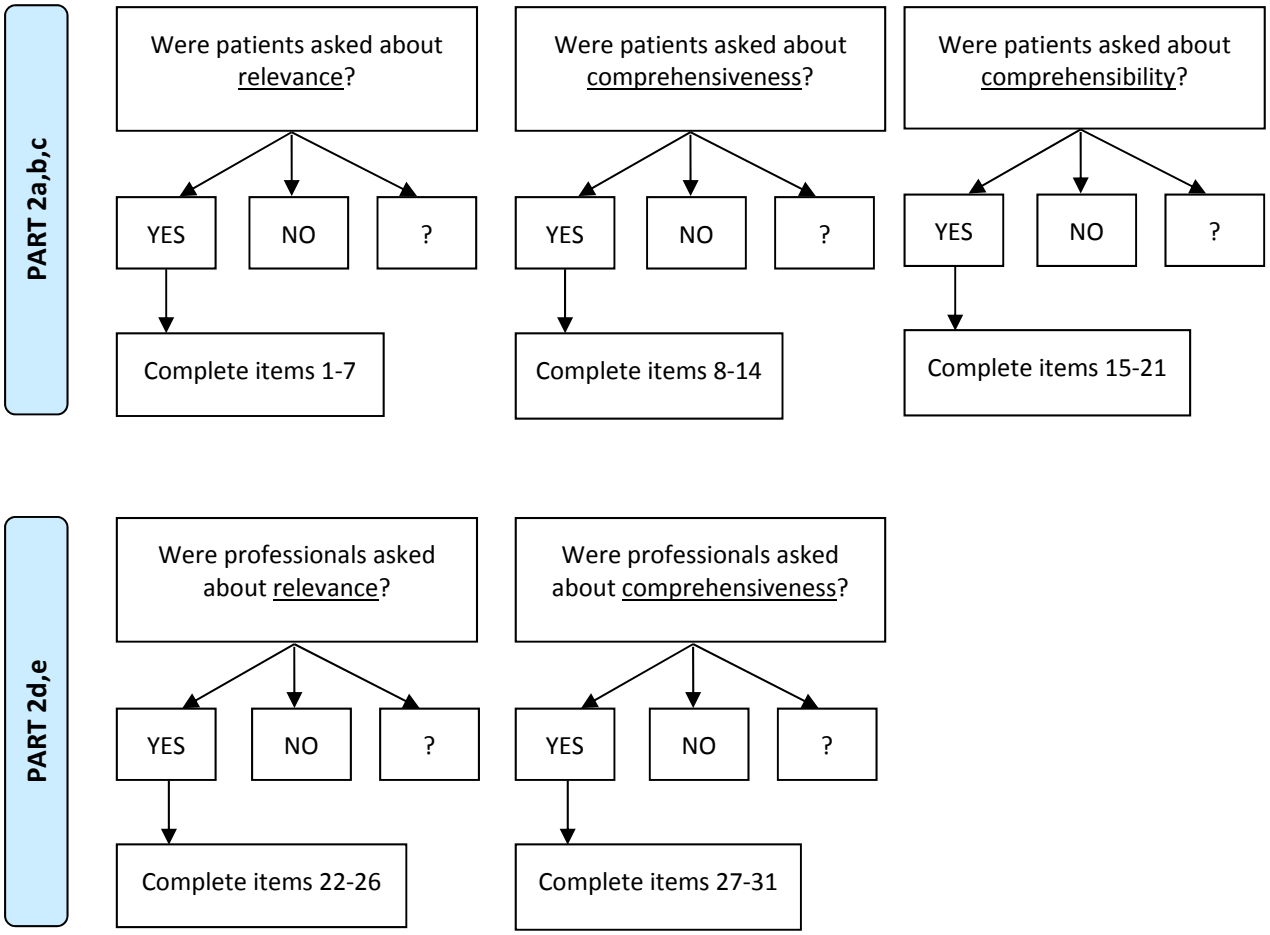
2a. Asking patients about the relevance of the PROM items
2b. Asking patients about the comprehensiveness of the PROM
2c. Asking patients about the comprehensibility of the PROM
2d. Asking professionals about the relevance of the PROM items
2e. Asking professionals about the comprehensiveness of the PROM

The standards in this box are divided into five parts. The review team needs to decide which parts of the box should be completed, depending on the available information in the content validity studies. For example, if professionals were not included in the content validity study, parts 2d and 2e do not need to be completed. If patients were included, but they were only asked about comprehensibility of the PROM items parts 2a and 2b do not need to be completed. Each part can be rated separately, see next page.

There is no part on asking professionals about comprehensibility because comprehensibility should be evaluated by patients, not by professionals. If comprehensibility was assessed by asking professionals, we recommend to ignore this information.

To decide which parts of box 2 should be completed, the flow chart given below can be used.

Flowchart for completing box 2



Determine the final rating (worst score counts) for each part separately (page 15)

Difference between PROM development study and content validity study

PROM development includes concept elicitation and testing of a new PROM. The rating of the PROM development is therefore based on all qualitative or quantitative studies that were performed in order to develop a PROM, including pilot testing of a draft PROM. A content validity study refers to a study on the relevance, comprehensiveness, or comprehensibility of an existing PROM. Such as study can be performed by researchers who were not included in the PROM development, but it can also be performed by the PROM developers, after the final PROM was established. The quality of content validity studies is therefore rated based on studies that were performed after the final version of the PROM was established.

Sometimes it can be unclear if a study should be considered a pilot study of a newly developed PROM (part 1b box 1) or whether it should be considered a content validity study (box 2). If a study was performed in a new patient sample, independent from the sample who participated in the PROM development (including cognitive debriefing), we consider it a content validity study.

Example: Duruöz et al developed a PROM for measuring functional disability for patients with rheumatic hand conditions [42]. The scale was constructed in three steps: 1) a list of hand activities was collected from published PROMs and patients input; 2) the provisional scale was tested. The PROM was administered to 102 patients and some items were deleted based on response frequencies, reliability and factor analysis; 3) the final scale was tested for reliability and validity in a new sample of 96 patients, who were interviewed. Interviewers asked each patient if the questions were comprehensible. When rating the content validity of this PROM, we considered step 2 as part of the PROM development (to be rated with part 1b of box 1). Step 3 was considered a new study because it was performed in a new sample, and was considered a content validity study (rated with box 2).

Rating the COSMIN standards

Consistent with the COSMIN boxes for the other measurement properties [9], a 4-point rating scale is used (i.e. very good, adequate, doubtful, inadequate) to rate each standard. Standards that are considered not applicable can be skipped. An Excel file is provided on the COSMIN website for data entry and calculating overall ratings.

When rating the standards, the following general rules should be applied:

- A standard is rated as **very good** when there is evidence that the quality aspect of the study to which the standard is referring is adequate. For example, if evidence is provided that interviews were based on an appropriate interview guide (e.g. the guide was clearly described or published), standard 8 in box 1 is rated as very good.
- A standard is rated as **adequate** when relevant information is not reported in an article, but it can be assumed that the quality aspect is adequate. For example, if it is assumable that saturation was reached (e.g. because a large number of focus groups or interviews were performed in a diverse sample of patients), but evidence is not provided, standard 12 in box 1 is rated as adequate.
- A standard is rated as **doubtful** if it is doubtful whether the quality aspect is adequate. For example, if it was doubtful whether the cognitive interview study was performed in a diverse sample of patients (e.g. because the characteristics of the patient sample were not clearly described), standard 15 of box 1 is rated as doubtful.
- Finally, a standard is rated as **inadequate** when evidence is provided that the quality aspect is not adequate. For example, if items were not re-tested after substantial adjustments standard 25 of box 1 is rated as inadequate.

An overall rating for the PROM development can be obtained by taking the lowest rating of any of the standards in box 1 (“worst score counts” method). It is also possible to obtain and report a rating for a specific part of the box, e.g. for the concept elicitation study (items 6-13), the total PROM design (items 1-13), the cognitive interview study (items 15-35). In the Excel file all possible overall ratings are shown.

An overall rating for the content validity study can be obtained by taking the lowest rating of any of the standards in box 2. However, often only one or a few parts of box 2 will be completed. In that case we recommend to determine the overall ratings per sub study separately (part 2a, 2b, 2c, 2d, 2e).

The overall ratings are used in step 3 when the relevance, comprehensiveness, comprehensibility, and overall content validity of the PROM are determined, based on the quality and results of the available studies.

The “worst score counts” method is used in all COSMIN boxes because poor methodological aspects of a study cannot be compensated by good aspects. In defining the response options, the ‘‘worst score counts’’ method was taken into consideration. Only fatal flaws in the design or statistical analyses were regarded as inadequate quality. If, for example, an appropriate qualitative data collection method was not used to identify relevant items for a new PROM, this is considered a fatal flaw in the PROM development study and the overall quality of the PROM development study is rated as inadequate. For some standards, the worst possible response option was defined as adequate or doubtful instead of inadequate because we did not want these standards to have too much impact on the quality rating per box.

In the next chapters, recommendations are provided for how each standard should be rated.

Finally, note that in the COSMIN materials 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.

STEP 1: Evaluate the quality of the PROM development, using COSMIN box 1

CHECK EXISTING RATINGS of the quality of the PROM development

Step 1 (evaluating the quality of the PROM development) needs to be done only once per PROM. Ratings of the quality of PROM developments are collected and published on the COSMIN website. We recommend to check the COSMIN website first to see if the quality of the PROM development has already been rated (e.g. in another systematic review). If a rating of the PROM development already exists, we recommend reviewers to consider using this rating instead of rating the quality of the PROM development again.

COSMIN Box 1 consists of two parts:
Part 1: standards for evaluating the quality of the PROM design (item generation)
Part 2: standards for evaluating the quality of a cognitive interview study or other pilot test performed to evaluate comprehensibility and comprehensiveness of a newly developed PROM
Both parts need to be completed when evaluating the quality of the PROM development.

Below, instructions are provided for how each standard should be rated. We recommend to use the Excel file (available from the COSMIN website) to document ratings and determine quality ratings per (part of a) box.

1a. Standards for evaluating the quality of the PROM design to ensure relevance of the PROM

		Very good	Adequate	Doubtful	Inadequate	Not applicable
1	Is a clear description provided of the construct to be measured?	Construct clearly described			Construct not clearly described	

The construct to be measured by the PROM (or subscale) should be clearly described by the PROM developers. One of the greatest threats to content validity is an unclear conceptual match between the PROM and the construct it intends to measure [34].
However, the construct does not necessarily have to be defined before developing the PROM but could also have been defined based on the results of the PROM development study. Sometimes defining the construct could be part of the aim of the study. As long as it is clearly described when the PROM development is finished so that it is clear what the final PROM intends to measure.
What is considered 'clear' should be decided by the reviewers. In any case, the description should be clear enough to judge whether the items of a PROM are relevant for the construct and whether the construct is comprehensively covered by the items. Just a word describing the construct of interest may not be enough.

Example: if the PROM aims to measure pain, it is important to know what aspects of pain the PROM intends to measure, e.g. pain intensity, pain interference, etc. If this is not described, it is hard to rate the relevance and comprehensiveness of the items in the PROM.

Example: if the PROM intends to measure physical function, it is important to know if the PROM intends to measure the capacity to perform certain activities, or the actual performance.[43]

Example of a clear description (very good rating)
" The PROMIS adult Pain Behavior item bank measures self-reported external manifestations of pain: behaviors that typically indicate to others that an individual is experiencing pain. These actions or reactions can be verbal or nonverbal, and involuntary or deliberate. They include observable displays (sighing, crying), pain severity behaviors (resting, guarding, facial expressions, and asking for help), and verbal reports of pain.." (https://www.assessmentcenter.net/documents/PROMIS Pain Behavior Scoring Manual.pdf) The definition is clear enough to judge that an item 'when I was in pain I screamed' is relevant for the construct, while an item 'how would you rate your pain on average' is not relevant for the construct of pain behavior because it does not refer to an action or reaction.

Example of an unclear description (inadequate rating)
The Oswestry Disability Index (ODI) intends to measure disability. "By disability we mean the limitations of a patient's performance compared with that of a fit person" [43]. In this definition it is not clear what is considered 'a patient's performance' or 'a fit person'. It is, for example, not clear whether it refers to physical limitations or (also) mental limitations. Therefore, it is difficult to rate the comprehensiveness of the PROM or to decide if an item on sleeping problems is relevant for the construct or not.

		Very good	Adequate	Doubtful	Inadequate	Not applicable
2	Is the origin of the construct clear: was a theory, conceptual framework or disease model used or clear rationale provided to define the construct to be measured?	Origin of the construct clear		Origin of the construct not clear		

It should be clear on what theoretical ground - if any - the PROM is based. For example, constructs can be based on the International Classification of Functioning (ICF) (http://www.who.int/classifications/icf/en/) , the model of Wilson & Cleary [44], or a specific theoretical model of pain. A model such as ICF defines the construct in a common language. This is important for understanding the context in which the PROM was developed and can assist in defining the construct to be measured. The conceptual model should reflect the current state of the science across disciplines for the construct.
It should also be clear how the construct to be measured is related to similar constructs measured by other (subscales of) PROMs. For example, how is the construct of fatigue related to constructs like sleepiness and vitality? Concept's boundaries need to be determined. The relative weight of patient versus professional input is central to the content validity debate. For example, even when patient input is used to develop item content, someone must decide what remains and what is removed from a questionnaire [26]. Further, if a PROM consists of multiple subscales that are being added up into a total score, it should be clear how these subscales theoretically are related to each other.
If a theory or conceptual model is not available, e.g. for new constructs, a rationale should be provided for why a new construct is proposed and how the proposed construct relates to other existing constructs. However, a PROM can still have good measurement properties without being based on a theory. Therefore we recommend to give a doubtful rating (not inadequate) if the origin of the construct is not clear.

Example of a clear description (very good rating)
The QUALIDEM is a dementia specific quality of life (QOL) questionnaire rated by professionals that can be applied in residential care. The questionnaire was based on the "adaptation-coping model" of Dröes and Van Tilburg (1996) and Finnema et al (2000). The model describes seven adaptive tasks that were interpreted as domains of QOL in dementia and were used for item formulation [45].

Example of a clear description (very good rating)
The conceptual model for the design of the Adolescent Cancer Suffering Scale was based on the components of the quality of life model in cancer survivors described by Ferrell. “Ferrell defined quality of life from a multidimensional perspective including four domains of patient well-being: physical, psychological, social and spiritual. Although distinct, each domain is able to influence the others. This model is particularly relevant to suffering because it recognizes its multidimensional aspect and because suffering could often derive from a poor quality of life” [46].

Example of an unclear description (doubtful rating)
The Menopause-Specific Quality Of Life (MENQOL) questionnaire is a condition-specific quality of life questionnaire, which was defined as “the extent that the physical, emotional and social aspects of an individual’s life are intact and not adversely affected by that condition or treatment”. The researchers attributed each question to one of five domains: physical, vasomotor, psychosocial, sexual and working life. It is unclear why these five domains were chosen and how they relate to the physical, emotional and social aspects of an individual’s life, as described in the definition of the construct [47].

		Very good	Adequate	Doubtful	Inadequate	Not applicable
3	Is a clear description provided of the target population for which the PROM was developed?	Target population clearly described			Target population not clearly described	

It is important to understand for whom the PROM was developed for, to determine the relevance and comprehensiveness of the content for the target population, and to determine the applicability of the PROM in other populations. Relevant information should be provided about the target population(s) with respect to type of disease (e.g. cancer, breast cancer, general population), important disease characteristics (e.g. stage of disease, acute versus chronic disease, with or without comorbidities), demographic characteristics (e.g. age group, gender). If the PROM was developed for use across multiple populations, each should be clearly described.

If the target population was described very broadly, reviewers can still consider giving a very good rating, assuming that the PROM is applicable in all patients with the specific condition. For example, the Quebec Back Pain Disability Scale was developed for patients with “back pain” [48]. It was not described if the questionnaire was developed for patients with acute or chronic back pain, for patients with low back pain or all types of back pain, for non-specific or specific back pain, etc.. Reviewers can give a very good rating, assuming that the PROM is applicable in all patients with back pain, e.g. chronic and acute patients across all settings (“all-inclusive”).

Example of a clear description (very good rating)
The Copenhagen Hip and Groin Outcome Score (HAGOS) was developed for use in “young to middle-aged, physically active patients with long-standing hip and/or groin pain” [49].

		Very good	Adequate	Doubtful	Inadequate	Not applicable
4	Is a clear description provided of the context of use?	Context of use clearly described		Context of use not clearly described		

The context of use refers to the intended application of the PROM. It should be clear for which application(s) the PROM was developed, e.g. for discriminative, evaluative, or predictive applications. For example, for an evaluative application you need to include items covering the entire range of the scale, while for a diagnostic application this may not be necessary. Context of use can also refer to a specific setting for which the PROM was developed (e.g. for use in a hospital or at home) or a specific administration mode (e.g. paper or computer-administered).
If the PROM was developed for use across multiple contexts, this should be described.
The context of use is less crucial than a description of the construct or the target population because a PROM is likely applicable across different contexts. Therefore, if the context(s) of use is not clear the lowest possible rating for this standard is doubtful (not inadequate).

Example of a clear description (very good rating)
The Haemo-QoL Index was developed as a short measure for health-related quality of life assessment in children and adolescents with haemophilia. The instrument was developed to be used as a screening tool in daily clinical routine, to be used in large clinical studies, and for comparing data across ages [50].

Example of an unclear description (doubtful rating)
Shapiro et al. developed a questionnaire for measuring fatigue and sleepiness [51]. In the introduction of the paper they state: "Although researchers and clinicians have long recognized the importance of these deficit states, psychometrically sound measurement instruments have been lacking until recently. The present study was undertaken to develop a psychometrically sound instrument to obtain conceptually distinct measures of common facets of the energy deficit states described variously by terms such as fatigue, sleepiness and tiredness." It is not stated for which application the PROM is to be developed. Some text in the discussion suggest that the PROM can be used for discrimination and evaluation ("Future studies can contribute usefully by investigating whether individuals in treatment for various diagnoses display differential responsiveness across subscales in response to alternative treatments (e.g., insomnia vs. narcolepsy vs. chronic fatigue syndrome). Future research can also contribute importantly by evaluating the sensitivity of the FACES questionnaire to change (e.g., in clinical trials))" , but this is not clearly described.

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

Input from members of the target population for which the PROM was developed is considered essential in the development of a PROM [24; 36]. The sample should be broad enough to capture the many facets of a phenomenon, and limitations to the sample should be clearly justified [38]. The qualitative study should include a diversity of patients with different characteristics to cover the breadth of the construct of interest. One should deliberately include patients with different manifestations of the construct (e.g. high and low levels of depression if the construct to be measured is depression), different disease characteristics that are important for the target population for which the PROM was developed (e.g. patients with acute and chronic disease, and patients with mild and severe disease), and different socio-demographic characteristics as appropriate for the construct and target population (e.g. patients that differ in age, gender, ethnicity, education, and literacy).

A diverse sample can be obtained by purposive sampling (that is: choose specific patients to ensure that all relevant experience is being captured), but other sampling techniques can also be appropriate [52], as long as a diverse sample, representing the target population, is obtained. Note that representativeness in this context is not being used in the statistical sense.

If the target population was described very broadly, e.g. just "back pain", and reviewers gave a 'very good' rating for standard 3 (assuming that the PROM is applicable in all patients with back pain), the sample in which the PROM development study was performed should be representative of all patients with back pain, e.g. chronic and acute patients across all settings.

If the target population for which the PROM was developed was not clearly described (standard 3), we may not be sure if the study was performed in a sample that is representative of the target population. In that case we recommend to give a doubtful rating.

Example of a study performed in a sample representing the target population (very good rating)
To identify how thyroid diseases impact the patients' lives and to select the most relevant quality of life (QoL) issues for a thyroid-specific questionnaire patients were selected "by a nonrandom strategic sampling procedure, which aimed at maximizing the patient variation as regards diagnoses, treatment, disease duration (including both newly diagnosed, untreated patients and treated patients), and age" [53].

Example of a study where it was doubtful whether the study was performed in a sample representing the target population (doubtful rating)
The Uterine Fibroid Symptom and Quality of Life (UFS-QOL) questionnaire was developed as a symptom and health-related quality of life instrument that was specific for women with uterine leiomyomata. Two focus groups (n=17) were held. "Focus group participants were recruited from a newspaper advertisement with all respondents screened for eligibility to ensure the presence of uterine leiomyomata symptoms" [54]. The sample was not further described and therefore it is doubtful whether the study was performed in a sample representing the target population.

		Very good	Adequate	Doubtful	Inadequate	Not applicable
6	Was an appropriate qualitative data collection method used to identify relevant items for a new PROM?	Widely recognized or well justified qualitative method used, suitable for the construct and study population	Assumable that the qualitative method was appropriate and suitable for the construct and study population, but not clearly described	Only quantitative (survey) method(s) used or doubtful whether the method was suitable for the construct and study population	Method used not appropriate or not suitable for the construct or study population	

There are numerous ways to gather qualitative data, so multiple methods can be used. The methods used should be appropriate for identifying relevant items for a new PROM. Examples of widely recognized qualitative methods for PROM development are focus groups, interviews, and concept mapping. If another method was used (e.g. observation), a justification should be provided. Also, the methods should be suitable for the construct of interest (e.g. considering the sensitivity of the topic) and the study population (considering factors like age (e.g. children, elderly), and physical, cognitive or communication abilities).

It is often recommendable to start with open-ended questions to allow spontaneous reporting and ensure unbiased data collection. We recommend to include someone with expertise in qualitative research in the review team to judge the quality of the qualitative methods used in the concept elicitation study.

A combination of qualitative and quantitative methods can also be used. However, we recommend to give a doubtful rating if only written information (e.g. survey) was used because more valuable information is obtained by personal contact with patients because of inter-personal interaction. We also recommend to give a doubtful rating if the method is not described clearly enough or not justified clearly enough to determine whether the approach was appropriate.

Example of a well justified qualitative method (very good rating)
The Influenza Intensity and Impact Questionnaire (FluiiQ™) was developed to measure the symptoms and impact of influenza on individuals with influenza-like illness (ILI) and laboratory-confirmed influenza. Three concept mapping workshops with 16 people were organized. Participants were asked to respond to the following seeding statement: "Thinking as broadly as possible, generate statements to describe how your recent episode of flu affected you and your life." Responses were printed on individual cards, and participants were then required to intuitively sort them into categories. Participants were also asked to rate each response according to two dimensions: impact and duration. The sort data were then analyzed during the workshop using specialized software. The outcome of this process provides a visual map that groups responses into clusters. The final step involves displaying the map to participants who are asked to come to a consensus around the meaningfulness of statement groupings (i.e., clusters) and identify overarching descriptors of the underlying theme of each cluster. The concepts within each broad cluster informed the development of items [55].

Example of an unclear qualitative method (doubtful rating)
"A rigorous item selection process was adopted throughout the development of the 29-item, nystagmus-specific quality-of-life questionnaire (NYS-29). Items for the questionnaire were created by RJM and JM working with data from the 21 individual interviews previously conducted across several writing sessions over a 6-month period" [56]. The interviews were not further described and therefore it is considered doubtful whether the methods were appropriate.

		Very good	Adequate	Doubtful	Inadequate	Not applicable
7	Were skilled group moderators/ interviewers used?	Skilled group moderators/ interviewers used	Group moderators /interviewers had limited experience or were trained specifically for the study	Not clear if group moderators /interviewers were trained or group moderators /interviewers not trained and no experience		Not applicable

Focus groups and interviews require skilled moderators/interviewers to standardize the process across all patients and to ensure that information gained is of relevance to the PROM development. Moderators/interviewers need to be familiar with the use of qualitative methods. They also need to be well informed about the population and patient experiences. We recommend to give a very good rating if the group moderators/interviewers had training and experience with qualitative methods in previous studies. Pilot testing of the interview guide is also helpful.

A well-prepared interviewer or informed student could also do a good job. Therefore we recommend to give an adequate rating if the group moderator/interviewer at least has had some experience or some kind of training. Using untrained interviewers may not necessarily be a fatal flaw in a PROM development study. Therefore we recommend to give a doubtful rating (not inadequate) if it is not clear if group moderators/interviewers were skilled or when they were not trained and had no experience.

		Very good	Adequate	Doubtful	Inadequate	Not applicable
8	Were the group meetings or interviews based on an appropriate topic or interview guide?	Appropriate topic or interview guide	Assumable that the topic or interview guide was appropriate, but not clearly described	Not clear if a topic guide was used or doubtful if topic or interview guide was appropriate or no guide		Not applicable

Group meetings or interviews can be structured, semi-structured or unstructured. In all cases, a topic or interview guide helps to support the procedures to be used, ensure a consistent approach in data collection, avoid undue influence of the group moderator/interviewer, and direct the conversation toward the topic(s) of interest. The FDA, for example, recommends that documentation for PROMs used in clinical trials include the protocol for qualitative interviews and focus groups [57]. Guides describe which instructions to give, which kind of questions to ask (e.g. a seeding question or open questions to open up the conversation), in what sequence, how to pose the questions, how to pose follow-ups, and in what time. The guide is not a strict plan that has to be carried out in the described way, but it leads the process. A single seeding question can be enough. The guide can also be adapted during the study. There should always be time for the patients to describe their experiences without too much restrictions to ensure that no important topics are missed.

We recommend to give a very good rating if the topic guide is available or it is at least described which instructions were given, which kind of questions were asked, and whether there was room for the patients to describe their experiences (e.g. open questions).

We recommend to give a doubtful rating (not inadequate) if no guide was used, because the results of the study do not have to be biased or incomplete if no guide was used.

Very open approaches such as grounded theory approaches may not use a topic guide yet produce very valuable results. If a good reason was provided why a topic guide was not used, one could rate this standard as not applicable.

Example of an appropriate topic or interview guide (very good rating)
"A review of the findings of initial focus groups from private industry, the results from the Prospective Evaluation of Radial Keratotomy study and NEI-VFQ field test, and initial interviews conducted by investigators with patients with myopia who have chosen a variety of corrective techniques provided the initial material for a draft NEI-RQL focus group protocol. The protocol included the following general and specific domains as topics: reading, driving, general vision, adjustment to change, occupation, recreation, vision correction method, general well-being, expectations about future vision, and other comments. Participants were asked first to describe characteristics they associate with their eyes or vision. They then provided answers to open-ended questions about what aspects of their life were most affected by their vision and their vision correction method, and were asked specific questions about how vision affected their day-to-day activities. They also provided their predictions about their visual functioning in the future. Table 2 is an abbreviated version of the focus group script. The focus group script was tested with one group of myopes and one of hyperopes" [58].

Example of an assumable appropriate topic or interview guide (adequate rating)
The Chronic Venous Insufficiency Questionnaire (CIVIQ) was developed based on 20 semi-structured interviews. "An interview guide was drawn up from preliminary information collected from a review of published literature and from interviews with four medical specialists and three general practitioners. The interview guide was designed to ensure that all aspects of the venous insufficiency problem were thoroughly assessed" [59]. We recommend to give an adequate rating in this case because an interview guide was developed, but not clearly described.

		Very good	Adequate	Doubtful	Inadequate	Not applicable
9	Were the group meetings or interviews recorded and transcribed verbatim?	All group meetings or interviews were recorded and transcribed verbatim	Assumable that all group meetings or interviews were recorded and transcribed verbatim, but not clearly described	Not clear if all group meetings of interviews were recorded and transcribed verbatim or recordings not transcribed verbatim or only notes were made during the group meetings/ interviews	No recording and no notes	Not applicable

This standard is based on the ISPOR guidelines for content validity studies [30; 31]. These guidelines describe good methodological practices for ensuring and documenting the content validity of newly developed PROMs. Concept elicitation interviews and focus groups should be recorded by high quality audio or video equipment to fully capture the context and content of each session as well as produce transcripts that form the data for analysis. All interviews or focus groups should be transcribed verbatim.

For some sensitive issues or when patients refuse or there are ethical problems, recording might not be an option. It can also sometimes be even inappropriate to record interviews e.g. on sensitive or illegal issues. Also, video recording can influence behavior. In these situations, it is recommended to make notes. Reviewers can give a very good or adequate rating if a good reason was provided why recording was considered inappropriate. In other cases, we recommend to give a doubtful rating if only notes were made.

With a concept mapping approach, there is no recording or transcription. Also sometimes non-standard response modes are used where there will be no notes, e.g. where people with learning disabilities are given multiple choices to aid their expression op opinions. In that case, one could consider this standard not applicable.

		Very good	Adequate	Doubtful	Inadequate	Not applicable
10	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	

For instrument development, the goal is to understand, organize, and communicate the meaning of the data and translate that meaning into a set of items that can be scored to represent the targeted concept(s) quantitatively. The patient quotes show the relationship between the concepts, the words and phrases, and the final PROM.
In qualitative research, data collection and analysis are interrelated and concurrent, rather than linear processes; analysis begins as soon as the first bit of data is collected. Accordingly, as emergent themes are identified in ongoing data analysis, all potentially relevant issues should be incorporated into the next set of interviews and observations [19].
Different qualitative methods can be used for analyzing the data, such as content analysis (a way of organizing qualitative data by counting and reporting the frequency of concepts/words/behaviors held within the data), deductive analysis (explores know theory/phenomenon/concepts in the data, concerned with testing or confirming hypotheses), framework analysis (a method to classify and organize data according to key themes, concepts and emerging categories), grounded theory (an inductive form of qualitative research where data collection and analysis are conducted together).
All methods share an emphasis on using and including the context of patients' discussions.
Different kinds of coding can be used, e.g. open coding (examining, comparing, conceptualizing, and categorizing data); axial coding (reassembling data into groupings based on relationships and patterns within and among the categories identified in the data); and selective coding (identifying and describing the central phenomenon, or core category) [30]. Computer-assisted qualitative data analysis software programs can also be used. Coding is often an iterative process; codes may be changed or reorganized during the process.

The appropriate method depends on the aim of the study. Often, a number of approaches are used to arrive at useful results. We recommend to include someone with expertise in qualitative research in the review team to judge the quality of the methods used in the qualitative data analysis.

Example of an appropriate approach (very good rating)
"Three members of the study team thoroughly and repeatedly reviewed the transcripts and audiotapes for patterns in health outcomes attributable to dysphagia. Working independently, they each developed a set of codes for qualitative data derived from each question and from each group. Content analysis was used to organize and condense patient and family reports into mutually exclusive and substantively exhaustive inventories of health outcomes" [60].

Example of an inappropriate approach (inadequate rating)
"Focus group discussions and three single interviews began with open-ended questions, followed by semi-structured interviews. Item generation was discontinued when no new items were identified in two interviews. The selected items were combined to form the preliminary questionnaire" [61]. No formal qualitative approach was used.

		Very good	Adequate	Doubtful	Inadequate	Not applicable
11	Was at least part of the data coded independently?	At least 50% of the data was coded by at least two researchers independently	11-49% of the data was coded by at least two researchers independently	Doubtful if two researchers were involved in the coding or only 1-10% of the data was coded by at least two researchers independently	Only one researcher was involved in coding or no coding	Not applicable

Ways of identifying and labeling qualitative data may differ across individuals. Good practice in analyses of qualitative data therefore involves two or more coders thoroughly trained. Each coder completes a few transcripts independently and meets with fellow coders to compare codes assigned, identify areas of consistency and inconsistency, and revise the coding rules. This process is repeated several times throughout the data analyses [30].
It may be advantageous to start independently, then after agreement, move ahead together. For example, a group of researchers can work together on a project, interactively and iteratively, with checks to ensure that nothing was missed. At least part of the data should be coded independently. Also, some evidence should be provided about consensus among the different coders.
If nothing is reported about coding of the data, it is likely that no standardized coding method was used. In that case, we recommend to give an inadequate rating.

		Very good	Adequate	Doubtful	Inadequate	Not applicable
12	Was data collection continued until saturation was reached?	Evidence provided that saturation was reached	Assumable that saturation was reached	Doubtful whether saturation was reached	Evidence suggests that saturation was not reached	Not applicable

Researchers should provide evidence that qualitative data was collected until no new relevant knowledge was being obtained from new patients (data saturation). This will lead to comprehensibility and wide applicability of the PROM [62].

Saturation is especially important for instruments based on a formative model (e.g. symptom scales). Such a model requires that no important aspects of the construct are missing.

We recommend to give a very good rating when it is clear how data saturation was determined and some kind of evidence is provided that saturation was reached. A common approach to analyzing data for saturation is to identify codes in each set of transcripts and compare these with the codes that appeared in previous groups. A saturation table (saturation grid) organized by concept code can be used to document the elicitation of information by successive focus groups or interviews.

Example of a saturation table [30].

Table 4 – Example symptom saturation table for concept elicitation interviews..

Concept codes	Transcript group where concept first appeared			
	Transcript Group 1 (n = 5 transcripts)	Transcript Group 2 (n = 4 transcripts)	Transcript Group 3 (n = 5 transcripts)	Transcript Group 4 (n = 5 transcripts)
Shortness of breath		X		
Difficult to breathe	X			
Not enough air	X			
Gasping			X	
Shallow/quick breathing	X			
Wheezing	X			
Coughing	X			
Chest tightness	X			
Chest pain		X		
Dizziness/lightheadedness	X			
No. of new concept codes appearing in each transcript group	7	2	1	0
% Of total new concept codes (total = 10)	70	20	10	0

We recommend to give an adequate rating if evidence is not provided, but the methods used make it assumable that saturation was reached (e.g. a large number of focus groups or interviews in a diverse population).

Example of evidence provided (very good rating)
"Saturation was evaluated across the sample to determine data sufficiency for breadth and depth of analysis to meet the study's objectives. This evaluation was facilitated by a saturation matrix developed using MAXQDA, which summarized frequency of code application overall and by treatment phase, and by reviewing the descriptive summaries in the concept spreadsheet. The saturation matrix indicated that no new codes were added from the final six focus groups or interviews" [63].

Example of assumable saturation (adequate rating)
The AA/PNH-specific QoL questionnaire (QLQ-AA/PNH) was developed according to EORTC guidelines. Patients in more than 25 German and Swiss cities were interviewed face to face. In phase I, interviews of 19 patients and 8 physicians specialized in AA/PNH treatment resulted in 649 QoL issues; these were condensed to 175 and graded according to their importance by 30 patients and 14 physicians (phase II). Five physicians took part in phases I and II [64]. Given this approach, it can be considered likely that saturation has been reached.

		Very good	Adequate	Doubtful	Inadequate	Not applicable
13	For quantitative studies (surveys): was the sample size appropriate?	≥100	50-99	30-49	<30	Not applicable

If a quantitative (survey) study was used to identify relevant content for a PROM, the sample size of the survey study should be large enough to assume that saturation was reached. In line with previous COSMIN recommendations, we recommend to give a very good rating if the sample size was at least 100.

1b. Standards for evaluating the quality of a cognitive interview study or other pilot test performed to evaluate comprehensibility and comprehensiveness of a PROM

		Very good	Adequate	Doubtful	Inadequate	Not applicable
14	Was a cognitive interview study or other pilot test conducted?	YES			**NO** (SKIP standards 15-34)	

A cognitive interview study or other pilot test should be performed to test the PROM for comprehensibility and comprehensiveness. If a cognitive interview study or any other kind of pilot test was not performed, the rest of the box can be skipped and the total quality of the PROM development study will be rated as inadequate.

		Very good	Adequate	Doubtful	Inadequate	Not applicable
15	Was the cognitive interview study or other pilot test performed in a sample representing the target population?	Study performed in a sample representing the target population	Assumable that the study was performed in a sample representing the target population, but not clearly described	Doubtful whether the study was performed in a sample representing the target population	Study not performed in a sample representing the target population	

See standard 5.

		Very good	Adequate	Doubtful	Inadequate	Not applicable
16	Were patients asked about the comprehensibility of the PROM?	YES		Not clear **(SKIP standards 17-25)**	NO **(SKIP standards 17-25)**	

The PROM instructions, items, response options, and recall period should be understood by patients as intended. If these PROM aspects are unclear, incorrect information may be gathered or patients may get frustrated because they do not understand how to complete the PROM.
If patients were not asked about the comprehensibility of the PROM, this standard will be rated as inadequate, which means that the total quality of the PROM development study will be rated as inadequate (worst score counts). In that case, standards 17 through 25 can be skipped
If it is not clear if patients were asked about the comprehensibility of the PROM, this standard will be rated as doubtful (the total quality of the PROM development study will be rated as doubtful or inadequate, depending on the rest of the standards in this box) and standards 17 through 25 can be skipped.

		Very good	Adequate	Doubtful	Inadequate	Not applicable
17	Were all items tested in their final form?	All items were tested in their final form	Assumable that all items were tested in their final form, but not clearly described	Not clear if all items were tested in their final form	Items were not tested in their final form or items were not re-tested after substantial adjustments	

PROM items should be tested in its final form (final wording, response options and lay-out) to ensure that the comprehensibility of the final item is tested.

Minor adjustments based on cognitive interviews or other pilot tests are allowed but if substantial adjustments are made to an item, its response options, or recall period, the adjusted items needs to be re-tested in their final form.

If in subsequent (psychometric) analyses items were removed from the PROM, this is not a problem for the rating of comprehensibility because it will likely not affect the comprehensibility of the remaining items.

Example of items tested in its final form (very good rating)
"Based on patient feedback from the first set of cognitive debriefing interviews revisions were made to the questionnaire items and response options during an international harmonization meeting. […] To confirm the face and content validity and cultural relevance of the revised HAE PRO, further cognitive debriefing interviews were conducted. Based on comments from the patients in this final set of interviews and considering feedback from earlier rounds of interviews minor revisions were made to the questionnaire to facilitate patient understanding of the instructions and response options on the HAE PRO" [65].

Example where it is not clear if the final set of items was tested (doubtful rating)
"We separately piloted the developed questionnaire with a purposive sample of 10 patients with a confirmed GI condition from a local hospital. Patients were asked to complete the questionnaire and four supplementary questions: Did you find any of the questions difficult to understand? Was there any question you did not want to answer? Were there any specific aspects of your bowel condition that were not covered by these questions? Did you find any of these questions not applicable to you?" [66]. The results of the pilot test were not described in the article so it is unclear whether any problems were found, and whether the items were adjusted and tested again.

		Very good	Adequate	Doubtful	Inadequate	Not applicable
18	Was an appropriate qualitative method used to assess the comprehensibility of the PROM instructions, items, response options, and recall period?	Widely recognized or well justified qualitative method used	Assumable that the method was appropriate but not clearly described	Only quantitative (survey) method(s) used or doubtful whether the method was appropriate or not clear if patients were asked about the comprehensibility of all items, response options, instructions, and recall period or patients not asked about the comprehensibility of the PROM instructions or the recall period	Method used not appropriate or patients not asked about the comprehensibility of all items and response options	

Each PROM instruction, item, response option, instruction, and recall period should be assessed separately. If patients were not asked about the comprehensibility of all items and response options, we recommend to give an inadequate rating. If patients were asked about the comprehensibility of all items and response options but not about the comprehensibility of the instructions or recall period, we recommend to give a doubtful rating (not inadequate). In this case, the risk of bias may be lower than when the items and response options are not assessed because it is expected that most PROM instructions and recall periods are rather easy to understand. Note that the recall period can be included in the instructions, so it may not always be evaluated separately.

Widely recognized methods to assess comprehensibility are e.g. the think aloud method, Three-Step Test-Interview, debriefing, probing, or other forms of cognitive interviewing [67]. Multiple methods can be used.

We recommend to give a doubtful rating if only written information was used because more valuable information is obtained by personal contact with patients because of inter-personal interaction.

We also recommend to give a doubtful rating if the method is not described clearly enough or not justified clearly enough to determine whether the approach was appropriate.

Example of a widely recognized method (very good rating)
The Oxford Participation and Activities Questionnaire is a patient-reported outcome measure in development that is grounded on the World Health Organization International Classification of Functioning, Disability, and Health (ICF). "Two rounds of cognitive interviewing were carried out to explore how respondents understood and answered the candidate items with the aim of improving the validity and acceptability of the questionnaire. Thirteen participants were interviewed. The "verbal probing" method was used during interviewing, which requires participants to complete the questionnaire unaided, followed by a focused interview. Participants then explained the reasons for their answers to each item and commented on any ambiguities. This method of interviewing allowed the interviewer to query a participant's understanding of an item and their interpretation of the instructions and response options" [68].

Example of an assumable appropriate method (adequate rating)
"This pre-final questionnaire was tested in one center with ten patients under the guidance of a psychologist who conducted semi-structured debriefing interviews immediately upon questionnaire completion. This procedure helped assess patients' perception of the questionnaire and its acceptability (i.e., whether it was appropriate to their condition and not intrusive)" [69]. It was not clearly described what questions were asked but it was described that semi-structured debriefing interviews were conducted, which assumes a systematic approach was taken. Also, the interviews were conducted by a psychologist, who probably has some knowledge of cognitive interviewing. Therefore, an adequate rating could be given.

	Very good	Adequate	Doubtful	Inadequate	Not applicable
19 Was each item tested in an appropriate number of patients?					
For qualitative studies	≥7	4-6	<4 or not clear		
For quantitative (survey) studies	≥50	≥30	<30 or not clear		

In a qualitative study, the number of interviews needed is a function of the complexity of the construct of interest, the complexity of the PROM instructions and items, and the characteristics and diversity of the target population. Willis has suggested that seven to 10 interviews are sufficient to confirm patient comprehensibility of the item [70]. However, Blair and Conrad showed that much larger samples are

needed to achieve an acceptable likelihood to detect problems, even if they are prevalent [71]. We consider saturation to be more important than the number of patients interviewed. Therefore, we recommend not to give an inadequate rating. The lowest possible rating is doubtful, if each item was reviewed by less than 4 patients.

For large item banks it is sometimes not possible for a patient to review all items. This is not a problem as long as each item is reviewed by an appropriate number of patients.

If quantitative (survey) methods were used, a sample size of at least 50 patients is considered very good. We assume that saturation will be reached with a survey of 50 patients.

		Very good	Adequate	Doubtful	Inadequate	Not applicable
20	Were skilled interviewers used?	Skilled group moderators/ interviewers used	Group moderators /interviewers had limited experience or were trained specifically for the study	Not clear if group moderators /interviewers were trained or group moderators /interviewers not trained and no experience		Not applicable

See standard 7.

		Very good	Adequate	Doubtful	Inadequate	Not applicable
21	Were the interviews based on an appropriate interview guide?	Appropriate topic or interview guide	Assumable that the topic or interview guide was appropriate, but not clearly described	Not clear if a topic guide was used or doubtful if topic or interview guide was appropriate or no guide		Not applicable

The use of an interview guide is highly relevant for cognitive interview studies. The ISPOR guidelines also recommend the use of a semi-structured interview guide and provide examples of interview questions [30]. We recommend to give a very good rating if it was at least know which questions were asked to the patients. See also standard 8.

		Very good	Adequate	Doubtful	Inadequate	Not applicable
22	Were the interviews recorded and transcribed verbatim?	All group meetings or interviews were recorded and transcribed verbatim	Assumable that all group meetings or interviews were recorded and transcribed verbatim, but not clearly described	Not clear if all group meetings or interviews were recorded and transcribed verbatim or recordings not transcribed verbatim or only notes were made during the group meetings/ interviews	No recording and no notes	Not applicable

Recording or taking notes is important for cognitive interviews for recording facial expressions, puzzlement etc. See also standard 9.

		Very good	Adequate	Doubtful	Inadequate	Not applicable
23	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	

See standard 10.

		Very good	Adequate	Doubtful	Inadequate	Not applicable
24	Were at least two researchers involved in the analysis?	At least two researchers involved in the analysis	Assumable that at least two researchers were involved in the analysis, but not clearly described	Not clear if two researchers were included in the analysis or only 1 researcher involved in the analysis		

Involving at least two researchers in the analyses is ideal to ensure rigor of the analyses and prevent bias. However, it is less essential in the phase of cognitive interviewing than in the phase of item elicitation. Therefore, we do not require independent coding in this phase (for example, two researchers could analyze the results together or discuss issues where the main researcher has doubts). We recommend to give a doubtful rating (not inadequate) if only one researcher was involved in the analyses.

		Very good	Adequate	Doubtful	Inadequate	Not applicable
25	Were problems regarding the comprehensibility of the PROM instructions, items, response options, and recall period appropriately addressed by adapting the PROM?	No problems found or problems appropriately addressed and PROM was adapted and re-tested if necessary	Assumable that there were no problems or that problems were appropriately addressed, but not clearly described	Not clear if there were problems or doubtful if problems were appropriately addressed	Problems not appropriately addressed or PROM was adapted but items were not re-tested after substantial adjustments.	Not applicable

If important problems are identified by patients, the PROM may need to be adapted and re-tested. We recommend to give an inadequate rating if problems were not appropriately addressed, or if the PROM was not re-tested after substantial adjustments.

Example of adapted PROM (very good rating)
"Based on patient feedback from the first set of cognitive debriefing interviews revisions were made to the questionnaire items and response options during an international harmonization meeting". And "To confirm the face and content validity and cultural relevance of the revised HAE PRO, further cognitive debriefing interviews were conducted. Based on comments from the patients in this final set of interviews and considering feedback from earlier rounds of interviews minor revisions were made to the questionnaire to facilitate patient understanding of the instructions and response options on the HAE PRO" [65].

Example of unclear if adaptations were made (doubtful rating)
"We separately piloted the developed questionnaire with a purposive sample of 10 patients with a confirmed GI condition from a local hospital. Patients were asked to complete the questionnaire and four supplementary questions: Did you find any of the questions difficult to understand? Was there any question you did not want to answer? Were there any specific aspects of your bowel condition that were not covered by these questions? Did you find any of these questions not applicable to you?" [66]. The results of the pilot test were not described so it is unclear whether the items were adjusted and tested again.

		Very good	Adequate	Doubtful	Inadequate	Not applicable
26	Were patients asked about the comprehensiveness of the PROM?	YES		NO or not clear **(SKIP standards 27-35)**		

Patients should explicitly be asked whether the items together comprehensively cover the construct the PROM (or subscale) intends to measure.
Although comprehensiveness may have been covered in the concept elicitation phase if saturation was reached, patients in the cognitive interview study may have a different notion of what is important to them, what might have been missed in focus groups or interviews at an earlier stage. However, since the risk that important concepts have been missed is not so big anymore if the concept elicitation phase has been well performed, we recommend to give a doubtful rating (not inadequate) if comprehensiveness was not assessed in the cognitive interview study. The subsequent standards of this part will therefore also not be rated lower than doubtful, with the exception of standard 35.

	Very good	Adequate	Doubtful	Inadequate	Not applicable
27 Was the final set of items tested?	The final set of items was tested	Assumable that the final set of items was tested, but not clearly described	Not clear if the final set of items was tested or not the final set of items was tested or the set of items was not re-tested after items were removed or added		

The final set of PROM items should be tested to evaluate the comprehensiveness of the PROM or subscale. If items are removed or added after pilot testing (or psychometric testing), a new pilot test should be performed.

	Very good	Adequate	Doubtful	Inadequate	Not applicable
28 Was an appropriate method used for assessing the comprehensiveness of the PROM?	Widely recognized or well justified method used	Assumable that the method was appropriate but not clearly described or only quantitative (survey) method(s) used	Doubtful whether the method was appropriate or method used not appropriate		

If only written information (a survey) was used this can be considered adequate for assessing comprehensiveness (not for assessing comprehensibility). See also standard 18.

Example of widely recognized method (very good rating)
In the development phase of the Crohn's Life Impact Scale (CLIQ) semi-structured cognitive debriefing interviews were conducted with CD patients to assess the relevance, comprehensiveness and practicality of the draft questionnaire. "Interviewees completed the printed questionnaire in the presence of the interviewer who noted obvious difficulties or hesitations over items. On completion, interviewees were asked about the reasons for observed problems and encouraged to comment on the suitability of the items, instructions and response categories. They were also asked whether any important issues had been omitted" [72].

Example of an assumable appropriate method (adequate rating)
The Knee OA Pre-Screening Questionnaire (KOPS) was developed for screening for knee osteoarthritis. The first version was tested by performing a pilot study with 15 participants and by consulting the mentioned expert panel. "Self-reported answers were confirmed in an interview to verify the Portuguese grammar and semantics, to check the clarity and relevance of the questions, to assure that all essential concepts were included correctly and to guarantee that all of the items related to the objectives were appropriate and comprehendible" [73]. The interviews were not clearly described but in the results section is was stated than one new risk factor was added based on pilot testing, so it is assumable that comprehensiveness was appropriately assessed.

	Very good	Adequate	Doubtful	Inadequate	Not applicable
29 Was each item tested in an appropriate number of patients? For qualitative studies For quantitative (survey) studies	 ≥7 ≥50	 4-6 ≥30	 <4 or not clear <30 or not clear		

See standard 19.

	Very good	Adequate	Doubtful	Inadequate	Not applicable
30 Were skilled interviewers used?	Skilled interviewers used	Interviewers had limited experience or were trained specifically for the study	Not clear if interviewers were trained or interviewers not trained and no experience		Not applicable

See standard 7.

	Very good	Adequate	Doubtful	Inadequate	Not applicable
31 Were the interviews based on an appropriate interview guide?	Appropriate topic or interview guide	Assumable that the topic or interview guide was appropriate, but not clearly described	Not clear if a topic guide was used or doubtful if topic or interview guide was appropriate or no guide		Not applicable

See standard 8.

	Very good	Adequate	Doubtful	Inadequate	Not applicable
32 Were the interviews recorded and transcribed verbatim?	All group meetings or interviews were recorded and transcribed verbatim	Assumable that all group meetings or interviews were recorded and transcribed verbatim, but not clearly described	Not clear if all group meetings or interviews were recorded and transcribed verbatim or recordings not transcribed verbatim or only notes were made during the group meetings/ interviews or no recording and no notes		Not applicable

See standard 9.

		Very good	Adequate	Doubtful	Inadequate	Not applicable
33	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 or approach not appropriate		

See standard 10.

		Very good	Adequate	Doubtful	Inadequate	Not applicable
34	Were at least two researchers involved in the analysis?	At least two researchers involved in the analysis	Assumable that at least two researchers were involved in the analysis, but not clearly described	Not clear if two researchers were included in the analysis or only 1 researcher involved in the analysis		

See standard 24.

		Very good	Adequate	Doubtful	Inadequate	Not applicable
35	Were problems regarding the comprehensiveness of the PROM appropriately addressed by adapting the PROM?	No problems found or problems appropriately addressed and PROM was adapted and re-tested if necessary	Assumable that there were no problems or that problems were appropriately addressed, but not clearly described	Not clear if there were problems or doubtful if problems were appropriately addressed or PROM was adapted but items were not re-tested after substantial adjustments	Problems not appropriately addressed	Not applicable

Even though a doubtful rating is recommended if comprehensiveness is not assessed at all, we recommend to give an inadequate rating if comprehensiveness is assessed and important problems were found but not appropriately addressed. This is considered a clear lack of content validity.
See also standard 25.

Step 2: Evaluate the quality of content validity studies on the PROM (if available), using COSMIN box 2

COSMIN Box 2 consists of five parts:

2a. Asking patients about the relevance of the PROM items
2b. Asking patients about the comprehensiveness of the PROM
2c. Asking patients about the comprehensibility of the PROM
2d. Asking professionals about the relevance of the PROM items
2e. Asking professionals about the comprehensiveness of the PROM

Each part can be completed and rated separately, depending on the design of the available content validity studies. If patients nor professionals were asked about the relevance, comprehensiveness, or comprehensibility of the PROM items, no parts of the box can be completed and the results of the study will be ignored.

> Example: Content validity of the WHO-QOL-BREF was examined in a study by calculating the skewness and kurtosis of each item. One question was excluded from further analyses because of values deviating too much from prevailing skewness or kurtosis criteria. It was concluded that the content validity of the remaining items was good [74]. Since the study did not ask patients, not professionals about the relevance, comprehensiveness, or comprehensibility of the PROM items, this is not regarded as a content validity study, and the results of the study are ignored.

In a systematic review, we recommend to report separate ratings for the relevance, comprehensiveness, and comprehensibility of a PROM because often more information is available on some aspects of content validity (e.g. comprehensibility), but less on other aspects (e.g. comprehensiveness). In addition, reviewers should report whether their judgments were based on information from patients and/or professionals. It is also possible to report an overall rating for content validity (see step 3: "Evaluate the content validity of the PROM"). We recommend to use the Excel file (available from the COSMIN website) to document rating and determine quality ratings per (sub)box.

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

		Very good	Adequate	Doubtful	Inadequate	Not applicable
1	Was an appropriate method used to ask patients whether each item is relevant for their experience with the condition?	Widely recognized or well justified method used	Only quantitative (survey) method(s) used or assumable that the method was appropriate but not clearly described	Not clear if patients were asked whether each item is relevant or doubtful whether the method was appropriate	Method used not appropriate or patients not asked about the relevance of all items	

The most appropriate way to collect data to support content validity is by conducting qualitative research entailing direct communication with patients to adequately capture their perspective on issues of

importance relative to the focus of the PRO measure. Both focus groups and individual interviews can be conducted to gather this information [19].

Adequate methods could for example be an interview study in which patients were asked to complete a PROM and indicate for each item whether the item is relevant for them. Or, an interview or focus group study in which open questions are used to ask patients about their relevant experiences and their answers are being compared to the items of the PROM.

Survey methods can also be used to ask patients about the relevance of the PROM items for them. We recommend to give an adequate rating (not very good) if survey methods were used.

Patient should be asked about the relevance of the items for them, considering their own experience, instead of the relevance of the items for the target population in general.

It is important that each item is evaluated separately. If the relevance is not asked for each item separately, we recommend to give an inadequate rating.

Example of a widely recognized method (very good rating)
"In order to evaluate the relevance of generic utility measures to the HRQL of RAI-refractory DTC patients concept detail from qualitative analysis results were mapped to the content of two generic utility measures (EQ-5D and SF-6D). Qualitative data were obtained from participants via focus groups and individual interviews, grouped by DTC treatment phase. Individual interviews were conducted with participants who were not available for scheduled focus groups for their treatment phase, or where preferred by the individual. All focus groups and individual interviews were conducted in-person and followed semi-structured discussion/interview guides. Each focus group was conducted by two trained researchers, one leading the discussion with the second in a supportive observational role" [63].

Example of a well justified method (very good rating)
Martin et al. evaluated the content validity of two SF-36 subscales for use in type 2 diabetes and non-dialysis chronic kidney disease-related anemia. "60 patients completed a cognitive interview and were asked to indicate whether they considered a list of limitations in energy and physical function to be ''not relevant at all, somewhat relevant, or highly relevant'' to them personally. They were then asked to identify the frequency with which they experienced the limitation in the past week (''not at all, some of the time, all of the time'') and to indicate which limitations they considered as one of their top six most important ones to be rid of. Data from the relevance, frequency, and importance ratings were summarized using descriptive statistics to assist in the evaluation of conceptual fit between patient descriptions of energy and levels of physical function and the items in the SF-36 VT and PF subscales" [75].

Example of an unclear method (doubtful rating)
Content validity of the Turkish version of the Ankylosing Spondylitis Quality of Life (ASQOL) questionnaire was evaluated "via ''cognitive debriefing'' method interviewing with 2 authors of this study and 15 ankylosing spondylitis patients. Cognitive debriefing showed the new Turkish ASQoL to be clear, relevant, and comprehensive" [76]. The methods were not described in more detail and therefore it is unclear how the relevance of the items was evaluated.

		Very good	Adequate	Doubtful	Inadequate	Not applicable
2	Was each item tested in an appropriate number of patients? For qualitative studies For quantitative (survey) studies	 ≥7 ≥50	 4-6 ≥30	 <4 or not clear <30 or not clear		

Guidelines for qualitative research have suggested 4-10 patients per group to ensure appropriate discussion among group participants [77-79].
See also standard 19, box 1.

		Very good	Adequate	Doubtful	Inadequate	Not applicable
3	Were skilled group moderators/interviewers used?	Skilled group moderators/ interviewers used	Group moderators /interviewers had limited experience or were trained specifically for the study	Not clear if group moderators /interviewers were trained or group moderators /interviewers not trained and no experience		Not applicable

See standard 7, box 1.

		Very good	Adequate	Doubtful	Inadequate	Not applicable
4	Were the group meetings or interviews based on an appropriate topic or interview guide?	Appropriate topic or interview guide	Assumable that the topic or interview guide was appropriate, but not clearly described	Not clear if a topic guide was used or doubtful if topic or interview guide was appropriate or no guide		Not applicable

See standard 8, box 1.

		Very good	Adequate	Doubtful	Inadequate	Not applicable
5	Were the group meetings or interviews recorded and transcribed verbatim?	All group meetings or interviews were recorded and transcribed verbatim	Assumable that all group meetings or interviews were recorded and transcribed verbatim, but not clearly described	Not clear if all group meetings or interviews were recorded and transcribed verbatim or recordings not transcribed verbatim or only notes were made during the group meetings/ interviews	No recording and no notes	Not applicable

See standard 9, box 1.

		Very good	Adequate	Doubtful	Inadequate	Not applicable
6	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	

See standard 10, box 1.

Example of a widely recognized approach (very good rating)
"30 adult people with a physician-diagnosis of asthma were asked to elicit their views on the content validity of three commonly used asthma-specific QoL questionnaires. In-depth interviews then explored individuals' subjective narratives of how the content of the questionnaires related to their experience of living with asthma. Thematic content analysis was performed by coding the verbatim transcripts and then grouping the codes into thematic categories. Data were coded by CA using ATLAS. The emerging themes were discussed regularly within the research team and credibility of the findings was established by seeking agreement among co-researchers" [80].

Example of assumable appropriate approach (adequate rating)
The QoLISSY was adapated to American-English. "Patients and parents separately completed the QoLISSY questionnaire and subsequently participated in a cognitive debriefing exercise. During the debriefing, they were specifically asked to evaluate the items in terms of clarity, sensitivity, importance, and relevance for their personal situation. Data collected from the focus groups were processed using MaxQDA, a qualitative data analysis program (VERBI-Software MaxQDA). Parallel coding involving two trained coding experts was used. For each statement inter-observer agreement was discussed and the few disagreements in statement codings were resolved by consensus" [81]. Although the analysis is not clearly described, a special qualitative data analysis program was used, so it is assumable that the approach was appropriate.

		Very good	Adequate	Doubtful	Inadequate	Not applicable
7	Were at least two researchers involved in the analysis?	At least two researchers involved in the analysis	Assumable that at least two researchers were involved in the analysis, but not clearly described	Not clear if two researchers were included in the analysis or only 1 researcher involved in the analysis		

See standard 24, box 1.

2b. Asking patients about the comprehensiveness of the PROM

		Very good	Adequate	Doubtful	Inadequate	Not applicable
8	Was an appropriate method used for assessing the comprehensiveness of the PROM?	Widely recognized or well justified method used	Only quantitative (survey) method(s) used or assumable that the method was appropriate but not clearly described	Doubtful whether the method was appropriate	Method used not appropriate	

Patients should explicitly be asked whether the items together comprehensively cover the construct the PROM subscale intend to measure, or if the included domains together comprehensively cover the wider construct measured by the PROM total score.
See also standard 28, box 1.

Example of a well justified method (very good rating)
"The study objective was to assess the content validity of the Cough and Sputum Assessment Questionnaire (CASA-Q) cough domains and the UCSD Shortness of Breath Questionnaire (SOBQ) for use in patients with Idiopathic Pulmonary Fibrosis (IPF). The study consisted of one-on-one interviews. Participants completed the CASA-Q cough domains and the UCSD-SOBQ, followed by debriefing questions using a semi-structured interview guide. The interview guide contained questions about the participant's understanding of the instructions for each instrument, the recall period, the intended meaning and relevance of the items and response options, and general questions about the overall instrument and missing concepts" [82]. It is clear that patients were asked about missing concepts and the authors also refer to the ISPOR guidelines [30] and the FDA guideline for PROM development [57].

Example of assumable appropriate approach (adequate rating)
Content validity of the Adolescent Cancer Suffering Scale was assessed by submitting the reduced pool of items to a panel of five health care professionals and four new patients.
"As a final iteration, the same panel of patients and health care professionals was asked to review all the items on the list and to modify, add or delete any items still considered to be irrelevant or unclear" [46]. The word 'add' suggest that comprehensiveness was assessed. Therefore we recommend to rate the quality of the methods as adequate.

		Very good	Adequate	Doubtful	Inadequate	Not applicable
9	Was each item tested in an appropriate number of patients?					
	For qualitative studies	≥7	4-6	<4 or not clear		
	For quantitative (survey) studies	≥50	≥30	<30 or not clear		

See standard 19, box 1.

		Very good	Adequate	Doubtful	Inadequate	Not applicable
10	Were skilled group moderators/interviewers used?	Skilled group moderators/ interviewers used	Group moderators /interviewers had limited experience or were trained specifically for the study	Not clear if group moderators /interviewers were trained or group moderators /interviewers not trained and no experience		Not applicable

See standard 20, box 1.

		Very good	Adequate	Doubtful	Inadequate	Not applicable
11	Were the group meetings or interviews based on an appropriate topic or interview guide?	Appropriate topic or interview guide	Assumable that the topic or interview guide was appropriate, but not clearly described	Not clear if a topic guide was used or doubtful if topic or interview guide was appropriate or no guide		Not applicable

See standard 21, box 1.

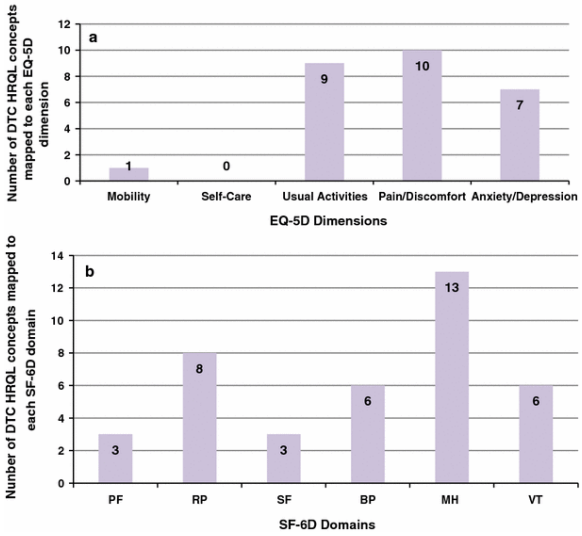
		Very good	Adequate	Doubtful	Inadequate	Not applicable
12	Were the group meetings or interviews recorded and transcribed verbatim?	All group meetings or interviews were recorded and transcribed verbatim	Assumable that all group meetings or interviews were recorded and transcribed verbatim, but not clearly described	Not clear if all group meetings or interviews were recorded and transcribed verbatim or recordings not transcribed verbatim or only notes were made during the group meetings/ interviews	No recording and no notes	Not applicable

See standard 22, box 1.

		Very good	Adequate	Doubtful	Inadequate	Not applicable
13	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	

See standard 23, box 1.

Example of a widely recognized approach (very good rating)
"Of the 50 concepts identified by qualitative analysis, 25 mapped to EQ-5D dimensions/items and 27 mapped to SF-6D domains/items. Figure 2 shows how concepts are captured by each instrument. Some of the concepts mapped to more than one dimension. Concepts that were not mapped to either instrument were mostly symptoms rather than broader HRQL impacts" [63].



Example of an inappropriate approach (inadequate rating)
"In a Danish study on the measurement properties of the DASH it was stated: "We calculated content validity which shows whether a questionnaire has enough items and covers the area of interest adequately. The content validity was high since we found no floor or ceiling effects" [83]. Although floor and ceiling effects may indicate a lack of content validity, the absence of floor and ceiling effects does not guarantee that no important items are missing. This should be evaluated in a qualitative study.

		Very good	Adequate	Doubtful	Inadequate	Not applicable
14	Were at least two researchers involved in the analysis?	At least two researchers involved in the analysis	Assumable that at least two researchers were involved in the analysis, but not clearly described	Not clear if two researchers were included in the analysis or only one researcher involved in the analysis		

See standard 24, box 1.

2c. Asking patients about the comprehensibility of the PROM

		Very good	Adequate	Doubtful	Inadequate	Not applicable
15	Was an appropriate qualitative method used for assessing the comprehensibility of the PROM instructions, items, response options, and recall period?	Widely recognized or well justified qualitative method used	Assumable that the method was appropriate but not clearly described	Only quantitative (survey) method(s) used or doubtful whether the method was appropriate or not clear if patients were asked about the comprehensibility of all items, response options instructions, and recall period or patients not asked about the comprehensibility of the PROM instructions or the recall period	Method used not appropriate or patients not asked about the comprehensibility of all items and response options	

See standard 18, box 1.

We recommend to give an inadequate rating if comprehensibility was not systematically assessed but conclusions about comprehensibility are drawn based on spontaneous comments made (or not made) by respondents, or based on the fact that all patients completed the questionnaire or that there were no missing items.

Example of a widely recognized method (very good rating)
The Gastroparesis Cardinal Symptom Index (GCSI) is a patient-reported outcome for gastroparesis using a two-week recall period. To minimize potential patient recall effects, a daily diary version of the GCSI (GCSI-DD) was developed. “Face-to-face cognitive debriefing interviews were conducted and each interview was approximately 1 h. The interviews followed a semi-structured interview schedule that provided an introduction to the interview session and served as a guide for queries and prompting. The interview was intended to capture information on how the participants describe their symptom experience, on the language they use to describe their condition and symptoms, and how they understood the instructions, individual items and response options on the GCSI-DD. At the conclusion of the interview, participants were asked to complete the PAGI-SYM and demographic questions. All sessions were audio-recorded and later transcribed” [84]. The recall period is not mentioned, but since this concerns a daily diary version, this was considered not relevant.

Example of a widely recognized method (very good rating)
“The aim was to assess face and content validity of the KIDSCREEN-52 questionnaire as a measurement of self- and proxy-reported QoL in children born with gastroschisis. The interviewer went through each of

the 52 items of the questionnaire, asking the child to "think aloud" to capture child's understanding of each question, what they were thinking about when choosing their response to each question and what was unclear. Further probing covered other topics such as the child's interests and what type of things make them happy (in general and in the past week/month), to elicit dimensions of life that might be missing from the items. Children were also asked what they would like to do that they do not do at the moment. Both parents and children were also asked to comment overall on the appropriateness of the items, their acceptability, and the clarity of the content and rating format" [85].

Example of an assumable appropriate method (adequate rating)
Cognitive debriefing interviews (CDIs) with 15 patients with SI were conducted to evaluate the SEAQ_ with respect to their ability to read, understand, and complete the questionnaire. Individual patient CDIs were conducted following interim modifications using a semi-structured cognitive debriefing interview guide [86]. Although the exact questions that patients were asked were not described, cognitive debriefing using a semi-structured interview guide can be considered a widely recognized method, so it can be assumed that the method was appropriate.

		Very good	Adequate	Doubtful	Inadequate	Not applicable
16	Was each item tested in an appropriate number of patients? For qualitative studies For quantitative (survey) studies	 ≥7 ≥50	 4-6 ≥30	 <4 or not clear <30 or not clear		

See standard 19, box 1.

		Very good	Adequate	Doubtful	Inadequate	Not applicable
17	Were skilled group moderators/interviewers used?	Skilled group moderators/ interviewers used	Group moderators /interviewers had limited experience or were trained specifically for the study	Not clear if group moderators /interviewers were trained or group moderators /interviewers not trained and no experience		Not applicable

See standard 20, box 1.

		Very good	Adequate	Doubtful	Inadequate	Not applicable
18	Were the group meetings or interviews based on an appropriate topic or interview guide?	Appropriate topic or interview guide	Assumable that the topic or interview guide was appropriate, but not clearly described	Not clear if a topic guide was used or doubtful if topic or interview guide was appropriate or no guide		Not applicable

See standard 21, box 1.

		Very good	Adequate	Doubtful	Inadequate	Not applicable
19	Were the group meetings or interviews recorded and transcribed verbatim?	All group meetings or interviews were recorded and transcribed verbatim	Assumable that all group meetings or interviews were recorded and transcribed verbatim, but not clearly described	Not clear if all group meetings or interviews were recorded and transcribed verbatim or recordings not transcribed verbatim or only notes were made during the group meetings/ interviews	No recording and no notes	Not applicable

See standard 22, box 1.

		Very good	Adequate	Doubtful	Inadequate	Not applicable
20	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	

See standard 23, box 1.

Example of a widely recognized approach (very good rating)
"Qualitative data were analyzed using ATLAS.ti software (ATLAS.ti Scientific Software Development GmbH, Berlin, Germany). Using ATLAS.ti, qualitative data (interview transcripts) can be systematically analyzed, coded, and compared. First, the codes were developed based on the content of the interview guide. Codes were smaller units and identifying concepts, themes, or recurring regularities that appeared within each interview. Second, participant transcripts were uploaded into ATLAS.ti and participant statements were coded by a researcher. Third, output tables were created using ATLAS.ti toanalyze coded responses. Using the output tables, responses were tallied and trends were identified" [87].

Example of an unclear approach (doubtful rating)
"Based on this first French version of the WOSI, a pilot test was conducted on a sample of patients who had chronic shoulder instability, whether or not it had been treated surgically, to validate that the patients understood the questions. All problems reported went into a new adaptation and evaluation with patients until the final French version of the WOSI (WOSI-Fr)" [88]. It seems that patients were asked about the comprehensibility of the PROM, but it is unclear how the results were analyzed and no results are presented.

	Very good	Adequate	Doubtful	Inadequate	Not applicable
21 Were at least two researchers involved in the analysis?	At least two researchers involved in the analysis	Assumable that at least two researchers were involved in the analysis, but not clearly described	Not clear if two researchers were included in the analysis or only one researcher involved in the analysis		

See standard 24, box 1.

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

It is important that a PROM has 'buy-in' from all stakeholders and that the included items in a PROM are important to clinicians and other professionals (e.g. policy makers, researchers). Professionals can also ensure that the included items are consistent with the theory, conceptual framework or disease model that was used to define the construct of interest [27]. The opinion of professionals about the relevance and comprehensiveness of the PROM is considered especially important for assessing symptoms because professionals may have seen many patients with different symptoms.

	Very good	Adequate	Doubtful	Inadequate	Not applicable
22 Was an appropriate method used to ask professionals whether each item is relevant for the construct of interest?	Widely recognized or well justified method used	Only quantitative (survey) method(s) used or assumable that the method was appropriate but not clearly described	Not clear if professionals were asked whether each item is relevant or doubtful whether the method was appropriate	Method used not appropriate or professionals not asked about the relevance of all items	

A typical approach is to convene a panel of professionals for a specific construct and population of interest, provide them with a list of objectives and the PROM items, and elicit their feedback in a standardized manner.
See also standard 1, box 2.

Example of a well justified method (very good rating)
"We assessed content validity by requesting advice related to the suitability of the HCFS (fatigue questionnaire) first draft from 10 experts (5 oncologists and 5 expert oncology nurses). We asked these experts to evaluate each of the 49 items on the following three points: 1) whether the question expresses perception of fatigue, 2) whether the question expresses different aspects of said perception, including physical, mental and cognitive perceptions" [89].

Example of an unclear method (doubtful rating)
The Swedish Lymphedema Quality of Life Inventory (SLQOLI) was presented to a lymphedema expert group, including four physiotherapists, four enrolled nurses, two occupational therapists and a social worker with extensive experience working with patients with lymphedema and knowledge of questionnaire design. Nine of the expert group members were also lymph therapists. "All were asked to check and relate their experience of their patient's relation to lymphedema and quality of life. The expert

group and the social worker judged the abbreviated scale to have good face validity" [90]. It is not clear what the experts were asked to do and whether they rated each item as relevant.

	Very good	Adequate	Doubtful	Inadequate	Not applicable
23 Were professionals from all relevant disciplines included?	Professionals from all required disciplines were included	Assumable that professionals from all required disciplines were included, but not clearly described	Doubtful whether professionals from all required disciplines were included or relevant professionals were not included		

It is important to include professionals from all relevant disciplines because different health professionals may have different opinions about what is relevant. By relevant professionals we mean researchers, clinicians, and other health care workers with professional expertise in the construct and population of interest. For example, for a study on the content validity of a PROM measuring physical functioning of low back pain patients, researchers with expertise in research on physical functioning of low back pain patients, and clinicians and other relevant health care workers specialized in the treatment of low back pain (e.g. orthopedic surgeons, physiotherapists, rheumatologists, psychologists, and chiropractors) should be included.
We recommend to include someone with experience with the target population of interest in the review team.

Example of all relevant professionals included (very good rating)
"The ABILOCO-Kids is a 10-item questionnaire described for use in children with Cerebral Palsy (CP) to record a parent's perceptions of their child's usual walking performance. To ensure face & content validity of Gujarati version using group consensus method each item was examined by group of experts (n =8) with mean experience of 24.62 years in the field of paediatric, paediatric neurology, paediatric orthopedics and paediatric physiotherapy. Each item was analysed by professionals for content, meaning, wording, format, ease of administration and scoring. Each item was scored as either accepted, rejected or accepted with modification" [91].

Example of unclear whether professionals from all required disciplines were included (doubtful rating)
"The content validity indices for the scale and each item of the Japanese LupusPRO were evaluated by five experts. The mean content validity index for the scale score was 0.99" [92]. It was not described who the five experts were.

	Very good	Adequate	Doubtful	Inadequate	Not applicable
24 Was each item tested in an appropriate number of professionals?					
For qualitative studies	≥7	4-6	<4 or not clear		
For quantitative (survey) studies	≥50	≥30	<30 or not clear		

Even though the inclusion of professionals from all relevant disciplines is considered more important than the number of professionals to achieve saturation, a minimal number of professionals should be included. For example, It has been recommended that approximately 5 content professionals with professional expertise in the construct being measured review the instrument [25; 93]. Alternatively, a focus group of

15-20 professionals who are somewhat knowledgeable in the area has been suggested [17]. We recommend to use the same criteria for the number of professionals as we do for the number of patients. See also standard 19, box 1.

		Very good	Adequate	Doubtful	Inadequate	Not applicable
25	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	

An appropriate approach is to rate the relevance of each item separately. Then the mean rating across professionals could be calculated or the Content Validity Index or a variation.

Example of a well justified approach (very good rating)
"Expert Panel Members were asked to complete a questionnaire in which they were asked to rate the BAMF UEGMS by responding to six standard format questions for each of the 11 PROM items. The questions were: (1) This item should be included; (2) The item is clearly worded; (3) Item should be reordered higher on scale; (4) Item should be reordered lower on scale; (5) This is a functionally relevant motor behavior; (6) This behavior is easily discriminated from others on the scale. The range of possible responses were 1 = Disagree to 4 = Agree. For each BAMF item, a mean value (average agreement) of 3.0 or higher for standard statements 1,2,5,6 was considered high agreement" [94]. The mean agreement score for questions (1) and (5) can be considered evidence for the relevance of the items.

		Very good	Adequate	Doubtful	Inadequate	Not applicable
26	Were at least two researchers involved in the analysis?	At least two researchers involved in the analysis	Assumable that at least two researchers were involved in the analysis, but not clearly described	Not clear if two researchers were included in the analysis or only one researcher involved in the analysis		

See standard 24, box 1.

2e. Asking professionals about the comprehensiveness of the PROM

		Very good	Adequate	Doubtful	Inadequate	Not applicable
27	Was an appropriate method used for assessing the comprehensiveness of the PROM?	Widely recognized or well justified method used	Only quantitative (survey) method(s) used or assumable that the method was appropriate but not clearly described	Doubtful whether the method was appropriate	Method used not appropriate	

See also standard 28, box 1.

Example of an unclear method (doubtful rating)
"Content validity reflects the extent to which the scale samples all the dimensions of the appropriate indicators of the construct. The items were distributed to 5 judges (an academic, a sport physician, a physical therapist, an athletic trainer, and an elite athlete, all with higher degrees in relevant areas) who were not involved in the adaptation process of the questionnaire. The judges rated each of the 10 items on a 5-point rating scale (1, poor; 2, fair; 3, good; 4, very good; 5, excellent match). The purpose was to evaluate and summarize item ratings using quantitative statistical procedures" [95]. Although it seems that the authors are interested in the comprehensiveness of the PROM ("the extent to which the scale samples all the dimensions") it is not clear what exactly the judges were asked to rate. It seems that the ratings refer more to the relevance ('excellent match') rather than the comprehensiveness of the items.

		Very good	Adequate	Doubtful	Inadequate	Not applicable
28	Were professionals from all relevant disciplines included?	Professionals from all required disciplines were included	Assumable that professionals from all required disciplines were included, but not clearly described	Doubtful whether professionals from all required disciplines were included or relevant professionals were not included		

See standard 23, box 2.

		Very good	Adequate	Doubtful	Inadequate	Not applicable
29	Was each item tested in an appropriate number of professionals?					
	For qualitative studies	≥7	4-6	<4 or not clear		
	For quantitative (survey) studies	≥50	≥30	<30 or not clear		

See standard 24, box 2.

		Very good	Adequate	Doubtful	Inadequate	Not applicable
30	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	

See standard 25, box 2.

		Very good	Adequate	Doubtful	Inadequate	Not applicable
31	Were at least two researchers involved in the analysis?	At least two researchers involved in the analysis	Assumable that at least two researchers were involved in the analysis, but not clearly described	Not clear if two researchers were included in the analysis or only one researcher involved in the analysis		

See standard 26, box 2.

Step 3: Evaluate the content validity of the PROM, based on the quality and results of the available studies and the PROM itself, using the rating system presented below.

In this step, the content validity of the PROM is rated, based on a summary of all available evidence on the PROM development and additional content validity studies, if available. In addition, the reviewers rate the content of the PROM themselves. If an independent reviewer cannot see profound concordance between the content of every PROM item and the intended construct, the content validity of the PROM may be inadequate. However, the reviewers' rating is considered as additional to the evidence from the literature. It will be weighted less than the evidence from a content validity study and the PROM development study. The quality of the PROM development and the quality of the available content validity studies are also taken into account.
Remember that each score or subscore of a PROM is rated separately.

Rating the content validity of a PROM consists of three sub-steps:

- **Step 3a**: First, each result of a single study on PROM development and content validity is rated against the 10 criteria for good content validity. In addition, the reviewers rate the content of the PROM themselves;
- **Step 3b**: The results of all available studies are qualitatively summarized to determine whether OVERALL the relevance, comprehensiveness, comprehensibility, and overall content validity is sufficient (+), insufficient (–), inconsistent (±), or indeterminate (?), taking all evidence into account. The focus is here on the PROM, while in the previous sub-steps the focus was on the single studies;
- **Step 3c**: The OVERALL ratings determined in step 3b will be accompanied by a grading for the quality of the evidence, using a modified GRADE approach (high, moderate, low, very low). This indicates how confident we are that the overall ratings are trustworthy.

For each PROM all ratings can be recorded in one Table, presented below (Table 1).

Rating the content validity of PROMs and grading the quality of the evidence requires a subjective judgment of the results of the PROM development, available content validity studies, the PROM itself, and the quality of the total body of evidence. Below some recommendations are provided for how to determine the ratings in different situations and how to use the GRADE principles to determine the quality of the evidence. However, reviewers should keep in mind that this methodology is new and that limited experience exists yet with using these recommendations. Therefore we recommend reviewers to use common sense or to ask advice from reviewers with experience in using GRADE if needed. We also encourage reviewers to provide feedback to the COSMIN steering committee to further improve this methodology.

Step 3a. Rate the result of the single studies on PROM development and content validity against the 10 criteria for good content validity

For each scale or subscale of a PROM, Table 1 should be filled in, rating the relevance, comprehensiveness, and comprehensibility of the PROM using the 10 criteria for good content validity, and summarizing the ratings according to the guidelines presented below. This is done first based on the methods and results of the PROM development study; second, based on each additional available content validity study of the specific PROM; and third, based on the reviewer's own rating of the content of the PROM. Extra columns can be added for additional content validity studies if necessary. Each criterion is rated as sufficient (+), insufficient (-), or indeterminate. The general rule to give a sufficient rating per criterion is:

\+ ≥85% of the items of the PROM (or subscale) fulfill the criterion
– <85% of the items of the PROM (or subscale) does fulfill the criteria
? No(t enough) information available or quality of (part of a) the study inadequate

The quality of the studies should be taken into account
When rating the results of each study against the criteria for good measurement properties, the quality of the study should be taken into account, to account for risk of bias. The criteria cannot be rated if the quality of the study was inadequate. For example, the criterion 'Are the included items relevant for the target population of interest?' can only be rated as sufficient or insufficient for a PROM development study if representative patients were involved in the elicitation of relevant items for the PROM. And the criterion 'Are the PROM instructions understood by the population of interest as intended?' can only be rated as sufficient or insufficient for a content validity study if patients were systematically asked about the comprehensibility of the items and the content validity study was not of inadequate quality. When a content validity study was performed, but it is unclear what exactly was done, or the quality of the study was rated as inadequate, or no results are reported, we recommend to rate the results of the study (or part of a study) as indeterminate (?). For example, if a content validity study was performed, but it is unclear if patients or professionals were asked about the relevance of the items, criteria 1 through 5 will be rated as indeterminate (?).

> Example: After translation of the ODI into Hungarian, the translated version was tested in 8 patients. "The patients could discuss any general or specific questions with the investigator controlling the pilot procedure. Upon completion, the subjects were briefly interviewed about the meaning of each item". Some interpretation difficulties were reported by the patients [96]. This part of the study can be regarded as an assessment of comprehensibility, so criterion 7 and 8 can be completed. It is not clear whether patients were asked about the relevance and comprehensiveness of the items, so these aspects will be rated as indeterminate. In addition, ten clinicians (spine surgeons and physiotherapists) were asked for a clinician's review of the PROM. However, no results were reported for this part of the study. Therefore the study performed in clinicians will be ignored.

Guidance for how each criterion should be rated, taking the quality of the studies into account, is presented in Table 2.

The RELEVANCE RATING, COMPREHENSIVENESS RATING, and COMPREHENSIBILITY RATING per study are determined by summarizing the five criteria for relevance (1-5), one criterion for comprehensiveness (6), and four criteria for comprehensibility, respectively (7-10). Guidance is provided in Table 3. These ratings can be **+ / - / ? / ±.** Finally, a CONTENT VALIDITY RATING per study is determined (**+ / - / ? / ±**). Guidance is provided in Table 4.
The last two columns in Table 1 are used for summarizing the ratings from all available studies on the PROM. This will be discussed in step 3b and step 3c.

Table 1. COSMIN criteria and rating system for evaluating the content validity of PROMs

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

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

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

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

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

Table 2. Guidance for giving a sufficient (+) rating for the 10 criteria for good content validity of a PROM

	PROM development study	Content validity study	Reviewers' rating	Remarks
1	The construct of interest is clearly described (i.e. 'very good' rating of box 1 standard 1), the origin of construct is clear (i.e. 'very good' rating of box 1 standard 2) and there is evidence from concept elicitation, literature, or professionals that at least 85% of the items refer to the construct of interest.	Professionals rated the relevance of the items for the construct of interest in a content validity study that was not inadequate (i.e. 'very good', 'adequate' or 'doubtful' rating for quality of relevance study in box 2d, standards 22-26) and found at least 85% of the items relevant for the construct.	Reviewers consider at least 85% of the items relevant for the construct of interest.	Every PROM item should measure a defined facet of the construct of interest, within the conceptual framework. PROM items should also be specific for the construct of interest, i.e. they should not measure a co-existing, but separate construct. For example, an item in a fatigue questionnaire such as "my muscles are weak" is relevant to fatigue but not specific for fatigue. Someone with MS may answer this question in the affirmative but not experiencing fatigue. There should also be no unnecessary items (too many items, except for a large scale item bank that will be used for computer adaptive testing). When a total PROM score is evaluated, each subscale (domain) should be relevant for the construct that the total PROM intends to measure. Professionals can best ensure that items are consistent with the theory, conceptual framework or disease model that was used to define the construct of interest.
2	The target population of interest is	Patients rated the relevance of the	Reviewers consider at least 85% of	The relevance of the items for the

	clearly described (i.e. 'very good' rating box 1 standard 3) and representative patients were involved in the elicitation of relevant items (i.e. 'very good' or 'adequate' rating box 1 standard 5) and concept elicitation ('worst score counts' box 1a standards 6-13) was not inadequate. If it is doubtful whether the study was performed in a sample representing the target population, we recommend to give an indeterminate (?) rating.	items for them in a content validity study that was not inadequate (rating for quality of relevance study, box 2a standards 1-7) and found at least 85% of the items relevant for them.	the items relevant for the population of interest.	target population can best be judged by patients. Some items may be relevant to only a small number of patients but they are necessary to capture the full range of patient experiences.
3	The context of use of interest is clearly described (i.e. 'very good' rating box 1 standard 4).	Professionals rated the relevance of the items for the context of use of interest in a content validity study that was not inadequate (rating for quality of relevance study, i.e. box 2d, standards 22-26) and found at least 85% of the items relevant for the context of use.	Reviewers consider at least 85% of the items relevant for the context of use of interest.	It should especially be clear whether the PROM is suitable for use in research and/or clinical practice. Professionals are considered to be more knowledgeable about the context of use of the PROM than patients.
4	A justification is provided for the response options.	Patients or professionals rated the appropriateness of the response options in a content validity study that was not inadequate (rating for quality of relevance study, i.e. box 2a standards 1-7 or box 2d standards 22-26) and found at least 85% of the response options	Reviewers consider the at least 85% of the response options appropriate for the construct, population, and context of use of interest.	The response options should be appropriate for the construct, population, and context of use of interest. For example, if the construct is pain intensity, the response options should measure intensity, not frequency. Also, a reasonable range of responses

		relevant.		should be provided for measuring the construct of interest.
5	A justification is provided for the recall period.	Patients or professionals rated the appropriateness of the recall period in a content validity study that was not inadequate (rating for quality of relevance study, i.e. box 2a standards 1-7 or box 2d standards 22-26) and found the recall period appropriate.	Reviewers consider the recall period appropriate for the construct, population, and context of use of interest.	The recall period can be important for measuring the construct, for example, whether there is no recall period (do you feel depressed now?) or whether the recall period is 1 week (did you feel depressed last week?). Different recall periods may be important, depending on the context. However, sometimes is does not matter whether the recall period is e.g. 1 or 2 weeks.
6	Patients were asked about the comprehensiveness of the PROM in the concept elicitation phase or in a cognitive interview study that was not inadequate (rating for quality of comprehensiveness study, i.e. box 1a standards 6-13, or box 1b standards 26-35) and no key concepts were missing.	Patients or professionals were asked about the comprehensiveness of the PROM in a content validity study that was not inadequate (rating for quality of comprehensiveness study, i.e. box 2b standards 8-14, or box 2e standards 27-31) and no key concepts were missing.	Reviewers consider the PROM comprehensive for the construct, population and context of use of interest.	The items should cover the full breadth of the construct of interest. However, there are often good reasons for not including all content suggested by patients in a PROM, for example because an item (or domain) is considered to be outside the scope of the PROM. When a total PROM score is evaluated, the subscales (domain) together should cover the full breath of the construct that the total PROM intends to measure.
7	Patients were asked about the comprehensibility of the instructions (including recall period) in a cognitive interview study that was not inadequate	Patients were asked about the comprehensibility of the instructions (including recall period) in a content validity study that was not inadequate (rating for		

	(rating for quality of comprehensibility study, i.e. box 1b standards 16-25) and problems were adequately addressed.	quality of comprehensibility study, box 2c standards 15-21) and no important problems were found.		
8	Patients were asked about the comprehensibility of the items and response options (including wording of the items and response options) in a cognitive interview study that was not inadequate (rating for quality of comprehensibility study, box 1b standards 16-25) and problems were adequately addressed.	Patients were asked about the comprehensibility of the items and response options in a content validity study that was not inadequate (rating for quality of comprehensibility study, box 2c standards 15-21) and no important problems were found for at least 85% of the items and response options.		
9			Reviewers consider at least 85% of the items and response options appropriately worded.	Consider aspects such as reading level (a scale should not require reading skills beyond that of a 12-year old), ambiguous items, double-barrelled questions, jargon, value-laden words, and length or items [97].
10			Reviewers consider at least 85% of the response options matching the questions.	The response options should be appropriate to the question asked and should be linguistically linked to the item content.

Table 3. Guidance for determining the RELEVANCE RATING, COMPREHENSIVENESS RATING, and COMPREHENSIBILITY RATING per study

	RELEVANCE	**COMPREHENSIVENESS**	**COMPREHENSIBILITY**	
			PROM development study AND content validity studies	**Reviewers rating**
+	At least criteria 1 and 2 are rated + AND at least two of the other three criteria on relevance are rated + Criteria 1 and 2 (relevance for construct and population) are considered the most important criteria and therefore they need to be rated +. A maximum of 1 criterion rated – is allowed, but reviewers can also rate ± in that case.	Rating of criterion 6	At least criterion 8 is rated + and criterion 7 is NOT rated - Criterion 8 is considered the most important, but for a sufficient rating criterion 7 should NOT be rated - (it may be rated ?).	Both criteria 9 and 10 are rated +
-	at least criteria 1 and 2 are rated - AND at least two of the other three criteria on relevance are rated -	Rating of criterion 6	Criterion 8 is rated – (independent of the rating for criterion 7)	Both criteria 9 and 10 are rated -
?	At least two of the criteria are rated ?	Rating of criterion 6	Criterion 8 is rated ? (independent of the rating for criterion 7)	At least one of the criteria is rated ?
±	All other situations	Rating of criterion 6	Criterion 8 is rated + and criterion 7 is rated -	One criterion is rated + and one is rated -

Table 4. Guidance for determining the CONTENT VALIDITY RATING per study

+	The RELEVANCE RATING is +, the COMPREHENSIVENESS RATING is +, and the COMPREHENSIBILITY RATING is +
-	The RELEVANCE RATING is -, the COMPREHENSIVENESS RATING is -, and the COMPREHENSIBILITY RATING is -
±	At least one of the ratings is + and at least one of the ratings is –
?	Two or more of the ratings are rated ?

Note: Relevance, comprehensiveness, and comprehensibility are weighted equally. If reviewers consider one aspect more important than another or if the results for relevance, comprehensiveness, and comprehensibility are very different, we recommend not to determine a CONTENT VALIDITY RATING but to only report the RELEVANCE RATING, COMPREHENSIVENESS RATING, and COMPREHENSIBILITY RATING.

Step 3b. The results of all available studies are qualitatively summarized to determine whether OVERALL, the relevance, comprehensiveness, comprehensibility, and overall content validity of the PROM is sufficient (+), insufficient (–), or inconsistent (±).

In this step, all results from the available studies on PROM development and content validity of the PROM, and the reviewer's rating (the ratings determined per study in step 3a) will be qualitatively summarized into OVERALL RATINGS for the relevance, comprehensiveness, comprehensibility, and overall content validity of the PROM. The focus is now on the PROM, while in the previous step the focus was on the single studies.

These OVERALL RATINGS will be filled in in the next to last column of Table 1.

The OVERALL RATINGS will be sufficient (+), insufficient (-),or inconsistent (±). An indeterminate overall rating (?) is not possible because the reviewer's rating is always available, which will be + or – or ±. If there are no content validity studies, or only content validity studies of inadequate quality, and the PROM development is of inadequate quality, the rating of the reviewers will determine the overall ratings. Indeterminate (?) ratings for development or content validity studies can be ignored.

If the ratings per study are sufficient (+) or insufficient (-), the OVERALL RATINGS will also be sufficient(+) or insufficient (-)
For example, if the RELEVANCE RATING based on the PROM development study is +, the RELEVANCE RATING based on content validity studies of the PROM is also +, and the reviewer's rating of the relevance of the PROM is also +, the OVERALL RELEVANCE RATING for the PROM will be +.
If no content validity studies were performed but the RELEVANCE RATING based on the PROM development study was + and the reviewer's rating of relevance is also +, the OVERALL RELEVANCE RATING will be also be +.

If the ratings per study are inconsistent, explanations for inconsistency should be explored and OVERALL RATINGS may be determined for relevant subgroups of studies with similar results. If no explanation is found, the OVERALL RATINGS will be inconsistent (±)
If the ratings of the PROM development study, the available content validity studies, and the reviewer's rating are inconsistent, reviewers should examine whether there is an explanation for the inconsistency. Explanations may lie in the population (e.g. differences in disease severity), the country in which the study was performed (or language version on the PROM), the year in which the PROM was developed, or the methods used in the study (e.g. study quality or patient versus professionals judgment). If an explanation is found, one should consider making subgroups of studies with similar results and draw conclusions on these subsets of studies.

> Example: If different ratings were given to studies performed in acute patients versus chronic patients, separate OVERALL RATINGS for the content validity of the PROM could be determined for acute and chronic patients. For example, the OVERALL COMPREHENSIVENESS RATING may be sufficient (+) in acute patients, but insufficient (-) in chronic patients (e.g. if evidence exists that key concepts for chronic patients are missing).

Some studies could be considered as providing more evidence than other studies and can be considered decisive in determining the OVERALL RATINGS when ratings are inconsistent:

- A content validity study provides more evidence than the PROM development study (because content validity studies evaluate the relevance, comprehensiveness and comprehensibility of a PROM in patients that were not included in the PROM development)

- A content validity study and the PROM development study provide more evidence than the reviewers' ratings (because evidence from studies should be given more weight than the subjective opinion of the reviewers, even if they can be considered professionals as well)
- A higher quality study provides more evidence than a lower quality study

  Example: If different RELEVANCE RATINGS were given to studies with very good or adequate quality than to studies with doubtful quality, one could consider determining the OVERALL RELEVANCE RATING based on the very good and adequate quality studies only and ignore the results of the doubtful quality studies. Otherwise studies of doubtful or inadequate quality (e.g. older studies) will always influence the overall ratings, even when multiple adequate or very good studies are available (e.g. performed later).

  Example: If a different COMPREHENSIBILITY RATING was given to a content validity study than to the PROM development study, one could consider determining the OVERALL COMPREHENSIBILITY RATING on the content validity study only (if the content validity study is of at least adequate quality).

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 RATINGS.

Example: If a PROM was developed many years ago, it might have been comprehensive at that time, but reviewers may find the PROM not comprehensive for using it nowadays. For example, a PROM measuring limitations in daily activities in patients with rheumatologic hand conditions developed in the last century, may not include questions about using a computer or mobile phone, which is highly relevant nowadays. This may lead to inconsistent COMPREHENSIVENESS RATINGS based on the PROM development study or a content validity study and the reviewers' rating. In that case, one could consider using only the reviewers' rating to determine the OVERALL COMPREHENSIVENESS RATING and ignore the older results.

If no explanation can be found for inconsistent results the OVERALL RATING will be inconsistent (±).

Example: If there are no content validity studies, the PROM development study is of doubtful quality, and the RELEVANCE RATING of the PROM development study is inconsistent with the reviewer's RELEVANCE RATING, the OVERALL RELEVANCE RATING could be rated as inconsistent (±).

Content validity of a total PROM score

As indicated before, we recommend to rate the content validity of each subscale of a (multi-dimensional) PROM separately. However, for multidimensional PROMs where subscale scores are added up into a total score, it is also possible to give a rating for the relevance, comprehensiveness, comprehensibility, and overall content validity of the total score, by combining the evidence on each of the subscales. If the OVERALL RATINGS of all subscales are sufficient (+) or insufficient (-), the OVERALL RATINGS for the total PROM score will also be sufficient(+) or insufficient (-). If the OVERALL RATINGS across the subscales are inconsistent or indeterminate, the OVERALL RATINGS for the total PROM score will also be inconsistent or indeterminate.

Step 3c. The OVERALL RATINGS will be accompanied by a grading for the quality of the evidence

The OVERALL RATINGS determined in step 3b will be accompanied by a rating for the quality of the evidence (i.e. the total body of evidence of the content validity of a PROM). The quality of the evidence indicates how confident we are that the OVERALL RATINGS are trustworthy. The evidence can be of high, moderate, low, or very low quality, depending on the number and quality of the available studies, the results of the studies, the reviewer's rating, and the consistency of the results.

For systematic reviews of clinical trials, the GRADE approach was developed for grading the quality of evidence in systematic reviews of clinical trials (http://www.gradeworkinggroup.org/intro.htm) [98]. For systematic reviews of PROMs we developed a modified GRADE approach, shown in 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 that the ones of interest in the review), imprecision (wide confidence intervals), and publication bias (negative results are less often published). For evaluating content validity, three of these factors are applicable: risk of bias, inconsistency, and indirectness. These factors are most relevant for evaluating content validity. Imprecision and publication bias are not taken into account.

Table 5. Grading the quality of evidence on content validity (modified GRADE approach)

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

For each OVERALL RATING determined in step 3b the quality of the evidence will be determined, using Table 5. The rating for the quality of the evidence will be filled in in the last column of Table 1.

If in step 3b 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 quality of the evidence.

The GRADE approach is meant to downgrade evidence when there are concerns about the quality of the evidence. The starting point is always the assumption that the OVERALL RATING is of high quality. The quality of evidence is subsequently downgraded by one or two levels per factor to moderate, low, or very low when studies' quality is doubtful or inadequate (i.e. risk of bias), or when there is (unexplained) inconsistency or indirect results. Below we explain in more detail how the three GRADE factors can be interpreted and applied in evaluating the content validity of PROMs, and why imprecision and publication bias are not taken into account.

Risk of bias can occur if the quality of the PROM development study or the quality of additional content validity studies is doubtful. Our recommendations for down grading for risk of bias are presented in Figure 1. We recommend to grade down for serious risk of bias (-1 level, e.g. from high to moderate) if the available content validity studies are of doubtful quality. We recommend to grade down for very serious risk of bias (-2 levels) if there are no content validity studies (or only of inadequate quality) and the PROM development study is of doubtful quality. We recommend to grade down -3 levels (to very low evidence) if there are no content validity studies (or only of inadequate quality) and the PROM development study is of inadequate quality. In the latter case, the OVERALL RATING will be based only on the reviewer's rating.

Figure 1 shows that high quality evidence for content validity can be obtained by at least one content validity study of adequate quality, independent of the quality of the PROM development. This means that high quality evidence for content validity study can also be obtained for PROMs that were poorly developed.

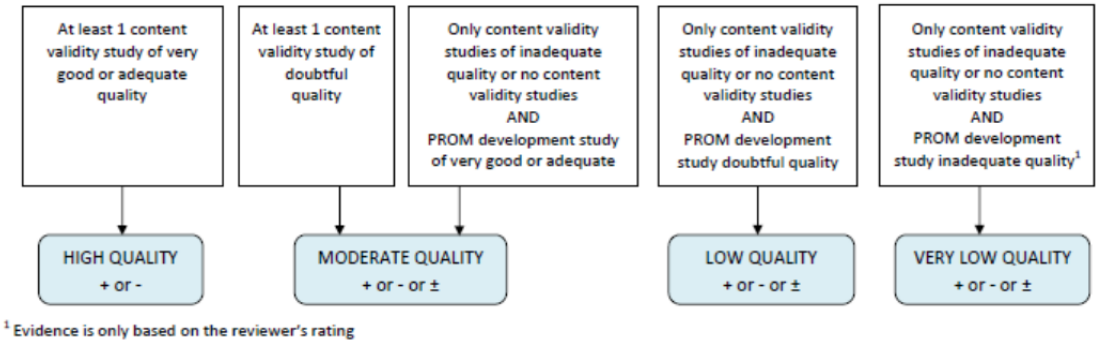


Figure 1. Flow chart for grading the quality of evidence, based on risk of bias.

Inconsistency can occur if the ratings of the PROM development study and additional content validity studies are inconsistent or if the ratings of these studies are inconsistent with the reviewers' ratings of the PROM. Inconsistency may already have been solved in step 3b, by making subgroups of studies with similar results and provide OVERALL RATINGS for these subgroups of studies. However, an alternative solution could be to give one OVERALL RATING, including all studies, even if the ratings per study are inconsistent, and grade down the quality of the evidence for inconsistency. It is up to the review team to decide which seems to be the best solution for their review.

Indirectness can occur if content validity studies are included that were performed in another population or another context of use than the population or context of use of interest in the systematic review, or if the content validity study assessed whether the PROM was valid for measuring another construct of interest than the one in the systematic review. Such studies could provide evidence on the comprehensibility of the PROM, but the evidence for the relevance and comprehensiveness may be considered indirect because this clearly depend on the construct and population of interest. In that case, it is possible to downgrade the evidence for indirectness.

> Example: in a systematic review of PROMs for patients with hand Osteoarthritis (OA) the Disabilities of the Arm, Shoulder an Hand (DASH) questionnaire was included as a measure of physical function and symptoms. No content validity study of the DASH has been performed in patients with hand OA. However, the DASH was designed for patients with any or several musculoskeletal disorders of the upper limb and therefore content validity studies performed in other patients with upper extremity problems may provide some evidence on the content validity

for hand OA patients. Evidence for comprehensibility of the PROM obtained in other populations with upper extremity problems may be considered also relevant for hand OA patients. Evidence on relevance and comprehensiveness are more context-specific and this information may therefore be included as indirect evidence of the content validity of the DASH in patients with hand OA. In that case, the evidence for relevance and comprehensiveness could be downgraded (e.g. from high quality evidence to moderate quality evidence or from moderate to low quality evidence) for indirectness.

Indirectness can also occur when the PROM was developed for a target population that is not the same as the population of interest in the review. In the example above, the DASH was developed for a broader target population (musculoskeletal disorders of the upper limb) than the population of interest in the review (hand OA). If only a few patients with hand OA were involved in the PROM development one may not be sure that the items of the DASH are relevant and comprehensive for patients with hand OA. In that case, reviewers may consider to down grade the quality of the evidence from the PROM development study for indirectness.

Imprecision is less relevant for content validity because PROM development and content validity studies concern qualitative research. Publication bias is difficult to assess because of a lack of registries for PROM development studies and content validity studies.

Reporting a systematic review on the content validity of PROMs

In reporting conclusions on the content validity of a PROM, both the OVERALL RATING and the quality of the evidence should be mentioned.

> Example: “There is high quality evidence for sufficient content validity of PROM X”.
>
> Example: “There is moderate quality evidence for sufficient relevance and comprehensibility of PROM Y, but there is very low quality evidence for insufficient comprehensiveness of PROM Y”.

We recommend to report the following information in a systematic review of the content validity of PROMs:

1. Characteristics of the PROMs (construct, target population, intended context of use, number of scales, number of items, recall period, etc.)
2. Quality of the PROM development (ratings of box 1)
3. Study characteristics of the available content validity studies (study population, sample size, study design, patients and/or professionals involved)
4. Quality of available content validity studies (ratings of box 2)
5. OVERALL RELEVANCE RATING, OVERALL COMPREHENSIVENESS RATING, OVERALL COMPREHENSIBILITY RATING and OVERALL CONTENT VALIDITY RATING per PROM and the associated quality of the evidence (green columns Table 1)

When using the Excel file to rate the quality of the studies, the summary ratings to be presented under point 2 and 3, will be generated in a separate worksheet that can be included in the manuscript.

An example of a systematic review of the content validity of PROMs can be found in Chiarotti et al [99].

References

1. Mokkink, L. B., Terwee, C. B., Patrick, D. L., Alonso, J., Stratford, P. W., Knol, D. L., Bouter, L. M., & de Vet, H. C. (2010). 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, 19(4), 539-549.
2. Terwee, C. B., Prinsen, C. A. C., Chiarotto, A., De Vet, H. C. W., Westerman, M. J., Patrick, D. L., Alonso, J., Bouter, L. M., & Mokkink, L. B. (2017). COSMIN standards and criteria for evaluating the content validity of health-related Patient-Reported Outcome Measures: a Delphi study. Qual Life Res in press.
3. Mokkink, L. B., Terwee, C. B., Patrick, D. L., Alonso, J., Stratford, P. W., Knol, D. L., Bouter, L. M., & de Vet, H. C. (2010). The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes. J Clin Epidemiol, 63(7), 737-745.
4. Terwee, C. B., Mokkink, L. B., Knol, D. L., Ostelo, R. W., Bouter, L. M., & de Vet, H. C. (2012). Rating the methodological quality in systematic reviews of studies on measurement properties: a scoring system for the COSMIN checklist. Qual Life Res, 21(4), 651-657.
5. Balemans, A. C., Fragala-Pinkham, M. A., Lennon, N., Thorpe, D., Boyd, R. N., O'Neil, M. E., Bjornson, K., Becher, J. G., & Dallmeijer, A. J. (2013). Systematic review of the clinimetric properties of laboratory- and field-based aerobic and anaerobic fitness measures in children with cerebral palsy. Arch Phys Med Rehabil, 94(2), 287-301.
6. Dobson, F., Choi, Y. M., Hall, M., & Hinman, R. S. (2012). Clinimetric properties of observer-assessed impairment tests used to evaluate hip and groin impairments: a systematic review. Arthritis Care Res (Hoboken), 64(10), 1565-1575.
7. Saether, R., Helbostad, J. L., Riphagen, II, & Vik, T. (2013). Clinical tools to assess balance in children and adults with cerebral palsy: a systematic review. Dev Med Child Neurol, 55(11), 988-999.
8. Vrijman, C., Linthorst Homan, M. W., Limpens, J., van der Veen, W., Wolkerstorfer, A., Terwee, C. B., & Spuls, P. I. (2012). Measurement properties of outcome measures for vitiligo. A systematic review. Arch Dermatol, 148(11), 1302-1309.
9. 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 assessing the methodological quality of studies on the measurement properties of Patient-Reported Outcome Measures. Qual Life Res 2017, Dec 19 [epub ahead of print].
10. Prinsen, C. A., Mokkink, L. B., Bouter, L. M., Alonso, J., Patrick, D. L., De Vet, H. C. W., & Terwee, C. B. (2017). COSMIN guideline for systematic reviews of outcome measurement instruments. Qual Life Res in press.
11. Prinsen, C. A., Vohra, S., Rose, M. R., Boers, M., Tugwell, P., Clarke, M., Williamson, P. R., & Terwee, C. B. (2016). How to select outcome measurement instruments for outcomes included in a "Core Outcome Set" - a practical guideline. Trials, 17(1), 449.
12. Reeve, B. B., Wyrwich, K. W., Wu, A. W., Velikova, G., Terwee, C. B., Snyder, C. F., Schwartz, C., Revicki, D. A., Moinpour, C. M., McLeod, L. D., Lyons, J. C., Lenderking, W. R., Hinds, P. S., Hays, R. D., Greenhalgh, J., Gershon, R., Feeny, D., Fayers, P. M., Cella, D., Brundage, M., Ahmed, S., Aaronson, N. K., & Butt, Z. (2013). ISOQOL recommends minimum standards for patient-reported outcome measures used in patient-centered outcomes and comparative effectiveness research. Qual Life Res, 22(8), 1889-1905.
13. Abanobi, O. C. (1986). Content validity in the assessment of health status. Health Values, 10(4), 37-40.

14. Armstrong, T. S., Cohen, M. Z., Eriksen, L., & Cleeland, C. (2005). Content validity of self-report measurement instruments: an illustration from the development of the Brain Tumor Module of the M.D. Anderson Symptom Inventory. Oncol Nurs Forum, 32(3), 669-676.
15. Basch, E., Abernethy, A. P., & Reeve, B. B. (2011). Assuring the patient centeredness of patient-reported outcomes: content validity in medical product development and comparative effectiveness research. Value Health, 14(8), 965-966.
16. Beck, C. T. (1999). Content validity exercises for nursing students. J Nurs Educ, 38(3), 133-135.
17. Beck, C. T., & Gable, R. K. (2001). Ensuring content validity: an illustration of the process. J Nurs Meas, 9(2), 201-215.
18. Beckstead, J. W. (2009). Content validity is naught. Int J Nurs Stud, 46(9), 1274-1283.
19. Brod, M., Tesler, L. E., & Christensen, T. L. (2009). Qualitative research and content validity: developing best practices based on science and experience. Qual Life Res, 18(9), 1263-1278.
20. De Champlain, A. F., Grabovsky, I., Scoles, P. V., Pannizzo, L., Winward, M., Dermine, A., & Himpens, B. (2011). Collecting evidence of content validity for the international foundations of medicine examination: an expert-based judgmental approach. Teach Learn Med, 23(2), 144-147.
21. Fehnel, S. (2009). Establishing optimal requirements for content validity: a work in progress. Value Health, 12(8), 1074.
22. Grant, J. S., & Kinney, M. R. (1992). Using the Delphi technique to examine the content validity of nursing diagnoses. Nurs Diagn, 3(1), 12-22.
23. Johnston, M., Dixon, D., Hart, J., Glidewell, L., Schroder, C., & Pollard, B. (2014). Discriminant content validity: a quantitative methodology for assessing content of theory-based measures, with illustrative applications. Br J Health Psychol, 19(2), 240-257.
24. Leidy, N. K., & Vernon, M. (2008). Perspectives on patient-reported outcomes : content validity and qualitative research in a changing clinical trial environment. Pharmacoeconomics, 26(5), 363-370.
25. Lynn, M. R. (1986). Determination and quantification of content validity. Nurs Res, 35(6), 382-385.
26. Magasi, S., Ryan, G., Revicki, D., Lenderking, W., Hays, R. D., Brod, M., Snyder, C., Boers, M., & Cella, D. (2012). Content validity of patient-reported outcome measures: perspectives from a PROMIS meeting. Qual Life Res, 21(5), 739-746.
27. Mastaglia, B., Toye, C., & Kristjanson, L. J. (2003). Ensuring content validity in instrument development: challenges and innovative approaches. Contemp Nurse, 14(3), 281-291.
28. O'Connor, R. (1992). Health-related quality of life measures need content validity. Aust Health Rev, 15(2), 155-163.
29. Olshansky, E., Lakes, K. D., Vaughan, J., Gravem, D., Rich, J. K., David, M., Nguyen, H., & Cooper, D. (2012). Enhancing the Construct and Content Validity of Rating Scales for Clinical Research: Using Qualitative Methods to Develop a Rating Scale to Assess Parental Perceptions of Their Role in Promoting Infant Exercise. Int J Educ Psychol Assess, 10(1), 36-50.
30. Patrick, D. L., Burke, L. B., Gwaltney, C. J., Leidy, N. K., Martin, M. L., Molsen, E., & Ring, L. (2011). Content validity--establishing and reporting the evidence in newly developed patient-reported outcomes (PRO) instruments for medical product evaluation: ISPOR PRO Good Research Practices Task Force report: part 2--assessing respondent understanding. Value Health, 14(8), 978-988.
31. Patrick, D. L., Burke, L. B., Gwaltney, C. J., Leidy, N. K., Martin, M. L., Molsen, E., & Ring, L. (2011). Content validity--establishing and reporting the evidence in newly developed patient-reported outcomes (PRO) instruments for medical product evaluation: ISPOR PRO good research practices task force report: part 1--eliciting concepts for a new PRO instrument. Value Health, 14(8), 967-977.

32. Polit, D. F., & Beck, C. T. (2006). The content validity index: are you sure you know what's being reported? Critique and recommendations. Res Nurs Health, 29(5), 489-497.
33. Polit, D. F., Beck, C. T., & Owen, S. V. (2007). Is the CVI an acceptable indicator of content validity? Appraisal and recommendations. Res Nurs Health, 30(4), 459-467.
34. Rothman, M., Burke, L., Erickson, P., Leidy, N. K., Patrick, D. L., & Petrie, C. D. (2009). Use of existing patient-reported outcome (PRO) instruments and their modification: the ISPOR Good Research Practices for Evaluating and Documenting Content Validity for the Use of Existing Instruments and Their Modification PRO Task Force Report. Value Health, 12(8), 1075-1083.
35. Tilden, V. P., Nelson, C. A., & May, B. A. (1990). Use of qualitative methods to enhance content validity. Nurs Res, 39(3), 172-175.
36. Vogt, D. S., King, D. W., & King, L. A. (2004). Focus groups in psychological assessment: enhancing content validity by consulting members of the target population. Psychol Assess, 16(3), 231-243.
37. Wynd, C. A., Schmidt, B., & Schaefer, M. A. (2003). Two quantitative approaches for estimating content validity. West J Nurs Res, 25(5), 508-518.
38. Kuper, A., Reeves, S., & Levinson, W. (2008). An introduction to reading and appraising qualitative research. BMJ, 337, a288.
39. Association, A. E. R., Association, A. P., & Education, N. C. o. M. i. (2014). Standards for Educational & Psychological Testing.
40. Fitzpatrick, R., Davey, C., Buxton, M. J., & Jones, D. R. (1998). Evaluating patient-based outcome measures for use in clinical trials. Health Technol Assess, 2(14), i-iv, 1-74.
41. Lasch, K. E., Marquis, P., Vigneux, M., Abetz, L., Arnould, B., Bayliss, M., Crawford, B., & Rosa, K. (2010). PRO development: rigorous qualitative research as the crucial foundation. Qual Life Res, 19(8), 1087-1096.
42. Duruoz, M. T., Poiraudeau, S., Fermanian, J., Menkes, C. J., Amor, B., Dougados, M., & Revel, M. (1996). Development and validation of a rheumatoid hand functional disability scale that assesses functional handicap. J Rheumatol, 23(7), 1167-1172.
43. Fairbank, J. C., Couper, J., Davies, J. B., & O'Brien, J. P. (1980). The Oswestry low back pain disability questionnaire. Physiotherapy, 66(8), 271-273.
44. Wilson, I. B., & Cleary, P. D. (1995). Linking clinical variables with health-related quality of life. A conceptual model of patient outcomes. JAMA, 273(1), 59-65.
45. Ettema, T. P., Droes, R. M., de Lange, J., Mellenbergh, G. J., & Ribbe, M. W. (2007). QUALIDEM: development and evaluation of a dementia specific quality of life instrument. Scalability, reliability and internal structure. Int J Geriatr Psychiatry, 22(6), 549-556.
46. Khadra, C., Le May, S., Tremblay, I., Dupuis, F., Cara, C., Mercier, G., Vachon, M. F., & Lachance Fiola, J. (2015). Development of the Adolescent Cancer Suffering Scale. Pain Res Manag, 20(4), 213-219.
47. Hilditch, J. R., Lewis, J., Peter, A., van Maris, B., Ross, A., Franssen, E., Guyatt, G. H., Norton, P. G., & Dunn, E. (1996). A menopause-specific quality of life questionnaire: development and psychometric properties. Maturitas, 24(3), 161-175.
48. Kopec, J. A., Esdaile, J. M., Abrahamowicz, M., Abenhaim, L., Wood-Dauphinee, S., Lamping, D. L., & Williams, J. I. (1996). The Quebec Back Pain Disability Scale: conceptualization and development. J Clin Epidemiol, 49(2), 151-161.
49. Thorborg, K., Holmich, P., Christensen, R., Petersen, J., & Roos, E. M. (2011). The Copenhagen Hip and Groin Outcome Score (HAGOS): development and validation according to the COSMIN checklist. Br J Sports Med, 45(6), 478-491.
50. Pollak, E., Muhlan, H., S, V. O. N. M., & Bullinger, M. (2006). The Haemo-QoL Index: developing a short measure for health-related quality of life assessment in children and adolescents with haemophilia. Haemophilia, 12(4), 384-392.
51. Shapiro, C. M., Flanigan, M., Fleming, J. A., Morehouse, R., Moscovitch, A., Plamondon, J., Reinish, L., & Devins, G. M. (2002). Development of an adjective checklist to measure five

FACES of fatigue and sleepiness. Data from a national survey of insomniacs. J Psychosom Res, 52(6), 467-473.
52. Marshall, M. N. (1996). Sampling for qualitative research. Fam Pract, 13(6), 522-525.
53. Watt, T., Hegedus, L., Rasmussen, A. K., Groenvold, M., Bonnema, S. J., Bjorner, J. B., & Feldt-Rasmussen, U. (2007). Which domains of thyroid-related quality of life are most relevant? Patients and clinicians provide complementary perspectives. Thyroid, 17(7), 647-654.
54. Spies, J. B., Coyne, K., Guaou Guaou, N., Boyle, D., Skyrnarz-Murphy, K., & Gonzalves, S. M. (2002). The UFS-QOL, a new disease-specific symptom and health-related quality of life questionnaire for leiomyomata. Obstet Gynecol, 99(2), 290-300.
55. Osborne, R. H., Norquist, J. M., Elsworth, G. R., Busija, L., Mehta, V., Herring, T., & Gupta, S. B. (2011). Development and validation of the Influenza Intensity and Impact Questionnaire (FluiiQ). Value Health, 14(5), 687-699.
56. McLean, R. J., Maconachie, G. D., Gottlob, I., & Maltby, J. (2016). The Development of a Nystagmus-Specific Quality-of-Life Questionnaire. Ophthalmology, 123(9), 2023-2027.
57. Administration, U. S. D. o. H. a. H. S. a. D., (CDER), C. f. D. E. a. R., (CBER), C. f. B. E. a. R., & (CDRH), C. f. D. a. R. H. (2009). **Guidance for Industry Patient-Reported Outcome Measures:Use in Medical Product Development to Support Labeling Claims**.
58. Berry, S., Mangione, C. M., Lindblad, A. S., & McDonnell, P. J. (2003). Development of the National Eye Institute refractive error correction quality of life questionnaire: focus groups. Ophthalmology, 110(12), 2285-2291.
59. Launois, R., Reboul-Marty, J., & Henry, B. (1996). Construction and validation of a quality of life questionnaire in chronic lower limb venous insufficiency (CIVIQ). Qual Life Res, 5(6), 539-554.
60. McHorney, C. A., Bricker, D. E., Kramer, A. E., Rosenbek, J. C., Robbins, J., Chignell, K. A., Logemann, J. A., & Clarke, C. (2000). The SWAL-QOL outcomes tool for oropharyngeal dysphagia in adults: I. Conceptual foundation and item development. Dysphagia, 15(3), 115-121.
61. Rutishauser, C., Sawyer, S. M., Bond, L., Coffey, C., & Bowes, G. (2001). Development and validation of the Adolescent Asthma Quality of Life Questionnaire (AAQOL). Eur Respir J, 17(1), 52-58.
62. Kerr, C., Nixon, A., & Wild, D. (2010). Assessing and demonstrating data saturation in qualitative inquiry supporting patient-reported outcomes research. Expert Rev Pharmacoecon Outcomes Res, 10(3), 269-281.
63. Gallop, K., Kerr, C., Simmons, S., McIver, B., & Cohen, E. E. (2015). A qualitative evaluation of the validity of published health utilities and generic health utility measures for capturing health-related quality of life (HRQL) impact of differentiated thyroid cancer (DTC) at different treatment phases. Qual Life Res, 24(2), 325-338.
64. Groth, M., Singer, S., Niedeggen, C., Petermann-Meyer, A., Roth, A., Schrezenmeier, H., Hochsmann, B., Brummendorf, T. H., & Panse, J. (2016). Development of a disease-specific quality of life questionnaire for patients with aplastic anemia and/or paroxysmal nocturnal hemoglobinuria (QLQ-AA/PNH)-report on phases I and II. Ann Hematol.
65. Bonner, N., Abetz-Webb, L., Renault, L., Caballero, T., Longhurst, H., Maurer, M., Christiansen, S., & Zuraw, B. (2015). Development and content validity testing of a patient-reported outcomes questionnaire for the assessment of hereditary angioedema in observational studies. Health Qual Life Outcomes, 13, 92.
66. Hutchings, H. A., Cheung, W. Y., Russell, I. T., Durai, D., Alrubaiy, L., & Williams, J. G. (2015). Psychometric development of the Gastrointestinal Symptom Rating Questionnaire (GSRQ) demonstrated good validity. J Clin Epidemiol, 68(10), 1176-1183.
67. Jobe, J. B., & Mingay, D. J. (1989). Cognitive research improves questionnaires. Am J Public Health, 79(8), 1053-1055.

68. Kelly, L., Jenkinson, C., Dummett, S., Dawson, J., Fitzpatrick, R., & Morley, D. (2015). Development of the Oxford Participation and Activities Questionnaire: constructing an item pool. Patient Relat Outcome Meas, 6, 145-155.
69. Kacha, S., Guillemin, F., & Jankowski, R. (2012). Development and validity of the DyNaChron questionnaire for chronic nasal dysfunction. Eur Arch Otorhinolaryngol, 269(1), 143-153.
70. Willis, G. B. (2005). Cognitive interviewing. A tool for improving questionnaire design. Thousend Oaks, California: SAGE Publications, Inc
71. Blair, J., & Conrad, F. G. (2011). Sample size for congitive interview pretesting. Public Opinion Quarterly, 75, 636-658.
72. Wilburn, J., McKenna, S. P., Twiss, J., Kemp, K., & Campbell, S. (2015). Assessing quality of life in Crohn's disease: development and validation of the Crohn's Life Impact Questionnaire (CLIQ). Qual Life Res, 24(9), 2279-2288.
73. Yazigi, F., Carnide, F., Espanha, M., & Sousa, M. (2016). Development of the Knee OA Pre-Screening Questionnaire. Int J Rheum Dis, 19(6), 567-576.
74. Trompenaars, F. J., Masthoff, E. D., Van Heck, G. L., Hodiamont, P. P., & De Vries, J. (2005). Content validity, construct validity, and reliability of the WHOQOL-Bref in a population of Dutch adult psychiatric outpatients. Qual Life Res, 14(1), 151-160.
75. Martin, M. L., Patrick, D. L., Gandra, S. R., Bennett, A. V., Leidy, N. K., Nissenson, A. R., Finkelstein, F. O., Lewis, E. F., Wu, A. W., & Ware, J. E., Jr. (2011). Content validation of two SF-36 subscales for use in type 2 diabetes and non-dialysis chronic kidney disease-related anemia. Qual Life Res, 20(6), 889-901.
76. Duruoz, M. T., Doward, L., Turan, Y., Cerrahoglu, L., Yurtkuran, M., Calis, M., Tas, N., Ozgocmen, S., Yoleri, O., Durmaz, B., Oncel, S., Tuncer, T., Sendur, O., Birtane, M., Tuzun, F., Bingol, U., Kirnap, M., Celik Erturk, G., Ardicoglu, O., Memis, A., Atamaz, F., Kizil, R., Kacar, C., Gurer, G., Uzunca, K., & Sari, H. (2013). Translation and validation of the Turkish version of the Ankylosing Spondylitis Quality of Life (ASQOL) questionnaire. Rheumatol Int, 33(11), 2717-2722.
77. Kitzinger, J. (1995). Qualitative research. Introducing focus groups. Bmj, 311(7000), 299-302.
78. Liamputtong, P. (2009). Qualitative Research Methods: Oxford University Press.
79. Greenbaum, T. L. (1998). The handbook for focus group research Thousand Oakes, California, US: Sage Publications.
80. Apfelbacher, C. J., Jones, C. J., Frew, A., & Smith, H. (2016). Validity of three asthma-specific quality of life questionnaires: the patients' perspective. BMJ Open, 6(12), e011793.
81. Bullinger, M., Sommer, R., Pleil, A., Mauras, N., Ross, J., Newfield, R., Silverman, L., Rohenkohl, A., Fox, J., & Quitmann, J. (2015). Evaluation of the American-English Quality of Life in Short Stature Youth (QoLISSY) questionnaire in the United States. Health Qual Life Outcomes, 13, 43.
82. Gries, K. S., Esser, D., & Wiklund, I. (2013). Content validity of CASA-Q cough domains and UCSD-SOBQ for use in patients with Idiopathic Pulmonary Fibrosis. Glob J Health Sci, 5(6), 131-141.
83. Schonnemann, J. O., & Eggers, J. (2016). Validation of the Danish version of the Quick-Disabilities of Arm, Shoulder and Hand Questionnaire. Dan Med J, 63(12).
84. Revicki, D. A., Camilleri, M., Kuo, B., Norton, N. J., Murray, L., Palsgrove, A., & Parkman, H. P. (2009). Development and content validity of a gastroparesis cardinal symptom index daily diary. Aliment Pharmacol Ther, 30(6), 670-680.
85. Rankin, J., Glinianaia, S. V., Jardine, J., McConachie, H., Borrill, H., & Embleton, N. D. (2016). Measuring self-reported quality of life in 8- to 11-year-old children born with gastroschisis: Is the KIDSCREEN questionnaire acceptable? Birth Defects Res A Clin Mol Teratol, 106(4), 250-256.
86. Jacobson, T. A., Edelman, S. V., Galipeau, N., Shields, A. L., Mallya, U. G., Koren, A., & Davidson, M. H. (2016). Development and Content Validity of the Statin Experience Assessment Questionnaire (SEAQ)(c). Patient.

87. Safikhani, S., Sundaram, M., Bao, Y., Mulani, P., & Revicki, D. A. (2013). Qualitative assessment of the content validity of the Dermatology Life Quality Index in patients with moderate to severe psoriasis. J Dermatolog Treat, 24(1), 50-59.
88. Perrin, C., Khiami, F., Beguin, L., Calmels, P., Gresta, G., & Edouard, P. (2017). Translation and Validation of the French version of the Western Ontario Shoulder Instability Index (WOSI): WOSI-Fr. Orthop Traumatol Surg Res.
89. Hirai, K., Kanda, K., Takagai, J., & Hosokawa, M. (2015). Development of the Hirai Cancer Fatigue Scale: Testing its reliability and validity. Eur J Oncol Nurs, 19(4), 427-432.
90. Klernas, P., Johnsson, A., Horstmann, V., Kristjanson, L. J., & Johansson, K. (2015). Lymphedema Quality of Life Inventory (LyQLI)-Development and investigation of validity and reliability. Qual Life Res, 24(2), 427-439.
91. Diwan, S., Diwan, J., Patel, P., & Bansal, A. B. (2015). Validation of Gujarati Version of ABILOCO-Kids Questionnaire. J Clin Diagn Res, 9(10), YC01-04.
92. Inoue, M., Shiozawa, K., Yoshihara, R., Yamane, T., Shima, Y., Hirano, T., Jolly, M., & Makimoto, K. (2016). The Japanese LupusPRO: A cross-cultural validation of an outcome measure for lupus. Lupus.
93. Gable, R. K., & Wolf, M. B. (1993). Instrument Development in the Affective Domain: Measuring Attitudes and Values in Corporate and School Settings: Springer Netherlands.
94. Cintas, H. L., Parks, R., Don, S., & Gerber, L. (2011). Brief assessment of motor function: content validity and reliability of the upper extremity gross motor scale. Phys Occup Ther Pediatr, 31(4), 440-450.
95. Korakakis, V., Malliaropoulos, N., Baliotis, K., Papadopoulou, S., Padhiar, N., Nauck, T., & Lohrer, H. (2015). Cross-cultural Adaptation and Validation of the Exercise-Induced Leg Pain Questionnaire for English- and Greek-Speaking Individuals. J Orthop Sports Phys Ther, 45(6), 485-496.
96. Valasek, T., Varga, P. P., Szoverfi, Z., Kumin, M., Fairbank, J., & Lazary, A. (2013). Reliability and validity study on the Hungarian versions of the oswestry disability index and the Quebec back pain disability scale. Eur Spine J, 22(5), 1010-1018.
97. Streiner, D. L., & Norman, G. R. (2008). Health measurement scales. A practical guide to their development and use. New York: Oxford University Press.
98. Guyatt, G. H., Oxman, A. D., Vist, G. E., Kunz, R., Falck-Ytter, Y., Alonso-Coello, P., & Schunemann, H. J. (2008). GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. BMJ, 336(7650), 924-926.
99. Chiarotto, A., Terwee, C. B., & Ostelo, R. W. (2017). A systematic review on the content validity and structural validity of questionnaires to measure physical functioning in patients with low back pain. Submitted.

Appendix 1 Names of the Delphi panel members

We would like to thank all panelists who participated in at least one round of the COSMIN content validity Delphi study: ASE Alreni, N Aaronson, A Abedi, F Abma, I Abma, C Acquadro, B Adair, C Ammann-Reiffer, E Andresen, C Apfelbacher, R Arbuckle, S Ashford, MJ Atkinson, KS Bagraith, L Bar-On, G Barrett, B Bartels, D Beaton, M Beattie, KA Benfer, G Bertolotti, C Bingham, J Blazeby, M Boers, E Boger, L Brosseau, R Buchbinder, M Calvert, S Cano, JC Cappelleri, D Cella, TC Chaves, SK Cheong, R Christensen, M Coenen, D Collins, N Collins, A Conijn, CE Cook, A Davis, S Deckert, L Deichmann Nielsen, KJFM Dekkers, F Dobbels, S Donnelly, T Dunning, K Edwards, T Egerton, K Ehrenbrusthoff, R Elbers, CDCM Faria, C Gagnon, B Gandek, AM Garratt, J Geere, M George, C Gibbons, E Gibbons, F Gilchrist, C Granger, JS Grant, J Greenhalgh, CR Gross, F Guillemin, G Guyatt, A Hall, M Hanskamp-Sebregts, K Haywood, J Hendrikx, B Hill, R Holman, R Ismail, M Johnston, U Kaiser, R Kane, N Kaur, T Kendzerska, C Kerr, V Khullar, N Kline Leidy, L Klokker, C Kopkow, SL Kroman, J Lambert, J Lane, CM Larsen, K Lasch, HH Lauridsen, EH Lee, J Lee, KN Lohr, M Lundberg, MR Lynn, JM Maaskant, S Magasi, ML Martin, L Maxwell, E MCColl, H McConachie, CM McDonough, I McDowell, D Mijnarends, D Miller, L Mitchell, VP Moen, M Monticone, AN Naegeli, S Nolte, R Osborne, R Ostelo, M Oude Voshaar, S Parry, AM Peeters, E Petitclerc, S Polinder, DF Polit, J Pool, K Potter, E Proud, L Rajmil, N Ramisetty, BB Reeve, AK Reimers, D Revicki, W Riley, KA Robinson, J Rodgers, EM Roos, N Rothrock, N Roussel, K Royal, I Scholl, VAB Scholtes, J Singh, R Speyer, M Sprangers, P Spuls, D Stevanović , J Stinson, LI Strand, S Svekington, SS Tavernier, K Thorborg, H Tigerstrand Grevnerts, K Toupin April, C Treanor, P Tugwell, YY Tung, S Tyson, C Vrijman, K Wales, S Weldam, S Wheelwright, B Wiitavaara, M Wilberforce, H Wittink, CKH Wong, JG Wright, G Zangger.