COSMIN checklist manual

Lidwine B. Mokkink
Caroline B Terwee
Donald L Patrick
Jordi Alonso
Paul W Stratford
Dirk L Knol
Lex M Bouter
Henrica CW de Vet

Contact
CB Terwee, PhD
VU University Medical Center
Department of Epidemiology and Biostatistics
EMGO Institute for Health and Care Research
1081 BT Amsterdam
The Netherlands
Website: www.cosmin.nl, www.emgo.nl
E-mail: cb.terwee@vumc.nl

Table of contents

Foreword

Studies evaluating the measurement properties of an instrument should be of high methodological quality to guarantee appropriate conclusions about the measurement properties of the instrument. To evaluate the methodological quality of a study on measurement properties, standards are needed for design requirements and preferred statistical analyses.

The COSMIN group developed a checklist containing such standards. This checklist can be used to evaluate the methodological quality of studies on measurement properties. The COSMIN checklist was developed in a multidisciplinary, international consensus-study in which 43 experts in health status measurement from all over the world participated.

This manual contains user-friendly data extraction forms and detailed instructions for how to complete the COSMIN checklist. Possible applications of the COSMIN checklist are described. In addition, background information is provided on the development and validation of the checklist and the rationale behind all items.

The COSMIN study was financially supported by the EMGO Institute for Health and Care Research, VU University Medical Center, Amsterdam, and the Anna Foundation, Leiden, the Netherlands.

1. Background information

1.1 The COSMIN initiative

COSMIN stands for COnsensus-based Standards for the selection of health Measurement INstruments. The COSMIN initiative aims to improve the selection of health measurement instruments.

Measurement in medicine is hampered by a lack of evidence and consensus on which are the best instruments. Numerous measurement instruments are used for measuring a given construct. These measurement instruments vary in content, purpose (i.e. discrimination or evaluation), and quality (i.e. the measurement properties). This leads to non-comparable study results, risk of incorrect conclusions, and non-evidence-based practice.

Instrument selection can be facilitated by standardized assessment of the content and measurement properties of measurement instruments. The aim of the COSMIN initiative is to provide tools for evidence-based instrument selection. To select the best instrument, e.g. in a systematic review of measurement properties, several steps should be taken. One of those steps is to evaluate the methodological quality of studies on measurement properties. The COSMIN checklist is one such tool. A next step is to assess the quality of a measurement instrument. For more information on how to select the best measurement instrument, we refer to De Vet et al, 2011. [1].

COSMIN steering committee

Lidwine B. Mokkink[1]
Caroline B Terwee[1]
Donald L Patrick[2]
Jordi Alonso[3]
Paul W Stratford[4]
Dirk L Knol[1]
Lex M Bouter[5]
Henrica CW de Vet[1]

[1] Department of Epidemiology and Biostatistics and the EMGO Institute for Health and Care Research, VU University Medical Center, Van der Boechorststraat 7, 1081 BT, Amsterdam, The Netherlands;

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

[3] Health Services Research Unit, Institut Municipal d'Investigacio Medica (IMIM-Hospital del Mar), Doctor Aiguader 88, 08003, and CIBER en Epidemiología y Salud Pública (CIBERESP), Barcelona, Spain;

[4] School of Rehabilitation Science and Department of Clinical Epidemiology and Biostatistics, McMaster University, 1400 Main St. West, Hamilton, Canada;

[5] Executive Board of VU University Amsterdam, De Boelelaan 1105, 1081 HV, Amsterdam, The Netherlands.

1.2 Development of the COSMIN checklist

The COSMIN checklist was developed in an international Delphi study. The aims of this study were:
(1) To reach consensus on which measurement properties should be evaluated of Health-Related Patient-Reported Outcomes (HR-PROs) and how they should be defined
(2) To develop standards for how these measurement properties should be evaluated in terms of study design and statistical analysis

Initially, a third aim of the study was to reach consensus on criteria for what constitutes adequate measurement properties. These criteria could be used to evaluate the quality of the instrument. However, due to lack of time, such criteria have not (yet) been developed.

Focus of the COSMIN checklist
In the development of the COSMIN checklist the focus was on evaluating the methodological quality of studies on measurement properties of HR-PROs. We chose to focus on HR-PROs, because of the complexity of these instruments. These instruments are aimed to measure not directly measurable, multidimensional constructs. Therefore most examples in this manual are from studies on HR-PRO instruments. However, the checklist can also be used for evaluating the methodological quality of studies on other measurement instruments because the same measurement properties are likely to be relevant for other kind of health-related measurement instruments, such as performance-based instruments or clinical rating scales.

In addition, we initially focussed on evaluative applications of HR-PRO instruments, i.e. longitudinal applications assessing treatment effects or changes in health over time. However, for instruments used for discriminative or predictive purposes, the design requirements and standards for the measurement properties are likely to be the same.

International Delphi study
An international Delphi study was performed consisting of four written rounds in 2006-2007 among a panel of 43 experts in the field of psychology, epidemiology, statistics and clinical medicine (Appendix 1).

As a preparation for the Delphi study a literature search was performed to determine how measurement properties are generally defined and evaluated [2]. A systematic search was performed to identify all systematic reviews of measurement properties of health status measurement instruments. Additional searches were performed to identify relevant methodological articles and textbooks that presented standards for evaluating measurement properties.

Each Delphi round consisted of a series of questions. Questions were asked about which measurement properties should be included when evaluating HR-PROs, how they should be called and defined, and how they relate to each other in a taxonomy. In addition, questions were asked about design requirements and preferred statistical methods for assessing the measurement properties. Preferred statistical methods were asked separately for studies using Classical Test Theory (CTT) and Item Response Theory (IRT). The results of previous rounds were presented in a feedback report, containing all results of the previous round, including arguments provided by the panel members. Consensus was considered to be reached when at least 67% of the panel members agreed with a proposal.

Results of the Delphi study

Consensus was reached on terminology (74% to 88%, except for structural validity (56%)), and definitions of measurement properties (68% to 88%) and on the position of each measurement property in the taxonomy (68% to 84%) [3]. Consensus was also reached on design requirements (68-94%) and preferred statistical methods (68-100%) [4].

The results of the consensus reached in the Delphi rounds were used to construct the COSMIN checklist. The COSMIN checklist consists of 12 boxes (see chapter 2). Two boxes are used to evaluate whether general requirements of a study on measurement properties are met. Nine boxes are used to evaluate the quality of the assessment of different measurement properties: (1) internal consistency, (2) reliability, (3) measurement error, (4) content validity (including face validity), (5-7) construct validity (subdivided into three boxes, about structural validity, hypotheses testing, and cross-cultural validity), (8) criterion validity, and (9) responsiveness. Finally, one box is used to evaluate the quality of a study on interpretability of a HR-PRO. Interpretability is not considered a measurement property, but it is an important requirement for the suitability of an instrument in research or clinical practice.

The COSMIN taxonomy, showing the relationships among the measurement properties, and their definitions, is presented in chapter 1.3.

1.3 Taxonomy and definitions

Lack of consensus on taxonomy, terminology, and definitions has let to confusion about which measurement properties are relevant, which concepts they represent, and how they should be evaluated. Before one could get consensus on the appropriate methods for evaluating a measurement property, one needs to have consensus on the terms, relevance, and definitions of the measurement properties. Therefore, the COSMIN initiative developed a taxonomy of measurement properties relevant for evaluating health instruments. This taxonomy formed the foundation on which the COSMIN checklist was based. In the COSMIN Delphi study consensus was reached on terminology and definitions of all included measurement properties in the COSMIN checklist.

Taxonomy of measurement properties

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

Figure 1. The COSMIN taxonomy

In assessing the quality of a HR-PRO instrument we distinguish three quality domains, i.e. reliability, validity, and responsiveness. Each domain contains one or more measurement properties. The domain reliability contains three measurement properties: internal consistency, reliability, and measurement error. The domain validity also contains three measurement properties: content validity, construct validity and criterion validity. The domain responsiveness contains only one measurement property, which is also called responsiveness. The term and definition of the domain and measurement property responsiveness are actually the same, but they are distinguished in the taxonomy for reasons of clarity. Some measurement properties contain one or more aspects, that were defined separately: Content validity includes face validity, and construct validity includes structural validity, hypotheses testing and cross-cultural validity.

Definitions of measurement properties

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

Issues that were discussed in the COSMIN Delphi study regarding terminology, definitions of measurement properties and the positions of the measurement properties in the taxonomy are described elsewhere [3].

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

Term			Definition
Domain	**Measurement property**	**Aspect of a measurement property**	
Reliability			**The degree to which the measurement is free from measurement error**
Reliability (extended definition)			**The extent to which scores for patients who have not changed are the same for repeated measurement under several conditions: e.g. using different sets of items from the same health related-patient reported outcomes (HR-PRO) (internal consistency); over time (test-retest); by different persons on the same occasion (inter-rater); or by the same persons (i.e. raters or responders) on different occasions (intra-rater)**
	Internal consistency		**The degree of the interrelatedness among the items**
	Reliability		**The proportion of the total variance in the measurements which is due to 'true'[†] differences between patients**
	Measurement error		**The systematic and random error of a patient's score that is not attributed to true changes in the construct to be measured**
Validity			**The degree to which an HR-PRO instrument measures the construct(s) it purports to measure**
	Content validity		**The degree to which the content of an HR-PRO instrument is an adequate reflection of the construct to be measured**
		Face validity	**The degree to which (the items of) an HR-PRO instrument indeed looks as though they are an adequate reflection of the construct to be measured**
	Construct validity		**The degree to which the scores of an HR-PRO instrument are consistent with hypotheses *(for instance with regard to internal relationships, relationships to scores of other instruments, or differences between relevant groups)* based on the assumption that the HR-PRO instrument validly measures the construct to be measured**
		Structural validity	**The degree to which the scores of an HR-PRO instrument are an adequate reflection of the dimensionality of the construct to be measured**
		Hypotheses testing	**Idem construct validity**
		Cross-cultural validity	**The degree to which the performance of the items on a translated or culturally adapted HR-PRO instrument are an adequate reflection of the performance of the items of the original version of the HR-PRO instrument**
	Criterion validity		**The degree to which the scores of an HR-PRO instrument are an adequate reflection of a 'gold standard'**
Responsiveness			**The ability of an HR-PRO instrument to detect change over time in the construct to be measured**
	Responsiveness		**Idem responsiveness**
Interpretability*			**Interpretability is the degree to which one can assign qualitative meaning - that is, clinical or commonly understood connotations – to an instrument's quantitative scores or change in scores.**

[†] *The word 'true' must be seen in the context of the CTT, which states that any observation is composed of two components – a true score and error associated with the observation. 'True' is the average score that would be obtained if the scale were given an infinite number of times. It refers only to the consistency of the score, and not to its accuracy (ref Streiner & Norman)*

** Interpretability is not considered a measurement property, but an important characteristic of a measurement instrument*

1.4 Validation of the COSMIN checklist

The COSMIN checklist could be considered a measurement instrument itself, measuring the methodological quality of a study on measurement properties. Therefore, the measurement properties of the COSMIN checklist itself should be thoroughly investigated.
Only three measurement properties are relevant for the validation of the COSMIN checklist: content validity, construct validity (hypotheses testing), and reliability. The other measurement properties are not relevant or cannot be assessed. Internal consistency and structural validity are not relevant because the items in the COSMIN boxes are not summarized into total scores (see also chapter 3 for this perspective). Measurement error cannot be assessed because there is no parameter of measurement error for ordinal or nominal scales. Cross-cultural validity is currently not relevant because the COSMIN checklist is only available in English. We have no intentions to translate the checklist into other languages. Criterion validity cannot be assessed because there is no gold standard for assessing the methodological quality of studies on measurement properties. And finally, responsiveness is not relevant because the studies that are being evaluated with the checklist do not change over time.

Content validity
Many experts in the field of measurement from all over the world with different backgrounds (e.g. psychometricians, epidemiologists, statisticians, clinicians) participated in the development of the COSMIN checklist. By using this approach, it is highly likely that all relevant items of all relevant measurement properties are included, contributing to the content validity of the checklist. However, since content validity is a subjective judgment, an unbiased judgment cannot be performed by the developers, and therefore other researchers should assess this. A thorough evaluation of content validity would contribute to the confidence in the validity of the checklist. This evaluation could consist of a survey among experts in the field (who were not involved in the development of the checklist), asking them to evaluate the relevance and comprehensiveness of the items in the checklist.

Construct validity (hypotheses testing)
No formal assessment of construct validity have been performed yet. Construct validity could be assessed by comparing the COSMIN standards to other standards that have been developed for assessing the methodological quality of studies on measurement properties, such as the scientific review criteria of the Medical Outcomes Trust [5], the checklist of Bombardier et al. [6], or the EMPRO tool [7]. The COSMIN standards are expected to correlate highly with each of these other standards because there is a high amount of overlap among these standards. Such a comparison, however, would be difficult because in the other instruments no total scores are used. Therefore it would be difficult to formulate suitable hypotheses for assessing construct validity. One could only compare individual items of the COSMIN checklist with individual items from different checklists. This may however, lead to obviously high correlations because on item level, the formulation and content is often very similar The difference with other checklists is the inclusion of different items, which is an aspect of content validity. An alternative approach could be to investigate known groups validity. This could be done, for example, by examining whether the COSMIN checklist can discriminate between high and low quality studies on measurement properties, as determined by some external

criterion. As an external criterion, the opinion of an expert panel (of people who were not involved in the development of the checklist) could be used. Such a study has not yet been performed.

Reliability

It is important to evaluate the inter-rater reliability of the COSMIN items to assess whether different users score articles in the same way. The inter-rater reliability of the COSMIN checklist has been assessed in an international study [8]. A total of 88 researchers (raters) from different countries participated in this study. Each rater evaluated the quality of three studies on measurement properties. Inter-rater reliability was analyzed by calculating Intraclass Correlation Coefficients (ICCs) per item of the checklist. Also percentages agreement were calculated for each item because many items had a skewed distribution of scores (i.e. more than 75% of the raters used the same response category). The ICCs were generally low, but the percentage agreement was appropriate (i.e. above 80%) for two thirds of the items. Low ICCs were due to skewed distributions, the subjective nature of some items, confusion about terminology, and lack of reporting. As a result of this study, some modifications to the checklist and this manual were made.

Note that the reliability results as described above apply to ratings of individual raters on item level. When using the COSMIN checklist in a systematic review of measurement properties, we recommend to complete the checklist by at least two independent raters, and to reach consensus on one final rating. We also recommend to reach agreement among the raters beforehand on how to handle items that need a subjective judgement, and how to deal with lack of reporting in the original article. Finally, for systematic review of measurement properties we recommend to use the COSMIN checklist with 4-point rating scale, and to apply total quality scores per measurement property. This approach is described in Chapter 3.

1.5 Applications of the COSMIN checklist

The COSMIN checklist can be applied in several different situations by different users:

Systematic reviews of measurement properties
Authors of systematic reviews of measurement properties can use the COSMIN checklist to evaluate the methodological quality of the included studies on measurement properties. The assessment of the methodological quality of the included studies is an important step in any kind of systematic review because low quality studies have a high risk of biased results. The COSMIN checklist is comparable to similar checklists that have been developed for use in systematic reviews of other types of studies, such as the Delphi list for assessing the methodological quality of randomized clinical trials [9], the QUADAS checklist for assessing the methodological quality of diagnostic studies [10], and the QUIPS checklist for assessing the methodological quality of prognostic studies (available from JA Hayden, Toronto, Canada: jhayden@dal.ca).

Measurement instrument selection
Researchers who are selecting a measurement instrument for their study can use the COSMIN checklist to assess the quality of the available evidence on the measurement properties of the selected instrument(s) or of different available measurement instruments. They can use the COSMIN checklist in combination with criteria for good measurement properties (e.g. those developed by Terwee et al. [11]) to selected the best measurement instrument for a given purpose.

Identification of the need for further research on measurement properties
Application of the COSMIN checklist and taxonomy can also identify the need for further research on the measurement properties of a measurement instrument. The COSMIN taxonomy can be used to see whether all measurement properties have been evaluated. The taxonomy includes all measurement properties that should be evaluated when an instrument is used for evaluative purposes. The COSMIN checklist can be used to assess whether the available evidence on the measurement properties is of high quality.

Designing a study on measurement properties
Researchers who are designing a study on the measurement properties of a particular measurement instrument can use the COSMIN checklist to make sure that their study meets the standards for excellent quality. For example, a researcher using the COSMIN checklist to design a study on the construct validity of a PRO instrument may decide, based on the items in the COSMIN checklist box on hypothesis testing, to formulate and test specific hypotheses about expected mean differences between groups or expected correlations between the scores on the instrument of interest and other, related instruments. This will ascertain the quality of the validation study.

Reporting a study on measurement properties
Researchers who are reporting a study on measurement properties can use the COSMIN checklist to make sure that they report all information that is needed to enable an appropriate evaluation of the quality of their study. We recommend to use the COSMIN terminology and definitions of measurement properties to facilitate uniform reporting and avoid confusion in the literature on terms and definitions.

Reviewing the quality of a submitted manuscript on measurement properties
Editors or reviewers of submitted manuscripts can use the COSMIN checklist to assess whether the quality of a study on measurement properties is high enough to justify publication of the study. In addition, they can use the COSMIN checklist to identify issues that have not (yet) been (properly) reported. In the review process, the COSMIN checklist can be a useful tool to increase the quality of reporting of studies on measurement properties.

2. Instructions for completing the COSMIN checklist

Throughout the manual we provide some examples to explain the COSMIN items, for example, about adequate sample sizes or about hypotheses. We would like to emphasize that these are used as examples, arbitrarily chosen and are not based on consensus within the COSMIN panel. The definition of criteria for good measurement properties was beyond the scope of the COSMIN study.

2.1 General instructions for completing the COSMIN checklist

In this chapter we describe how to use the COSMIN checklist. We first describe the general structure of the checklist. Next, we provide a four step procedure to be followed when using the checklist.
To illustrate the use of the checklist, we focus in this chapter on the application of the checklist to evaluate the methodological quality of a published article on measurement properties (e.g. when performing a systematic review of measurement properties). Other possible applications of the checklist are discussed in chapter 1.5.

General structure of the checklist
The checklist contains twelve boxes. Ten boxes can be used to assess whether a study meets the standards for good methodological quality. Nine of these boxes contain standards for the included measurement properties: internal consistency (box A), reliability (box B), measurement error (box C), content validity (including face validity)(box D), construct validity (i.e. structural validity (box E), hypotheses testing (box F), and cross-cultural validity (box G)), criterion validity (box H), and responsiveness (box I). One box contains standards for studies on interpretability (box J). In addition, two boxes are included in the checklist that contain general requirements. One box for articles in which IRT methods are applied (IRT box), and one box containing general requirements for the generalisability of the results of a study on one or more measurement properties (Generalisability box).

Four step procedure for completing the checklist
When completing the COSMIN checklist, four steps should be taken (Figure 2), which will be further explained in the next paragraphs.

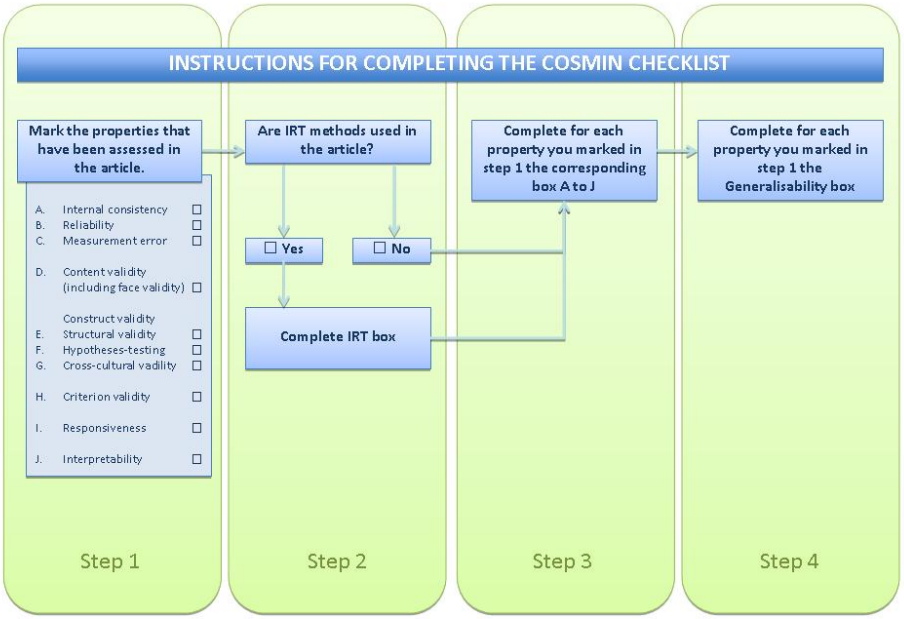


Figure 2. Four-step procedure for completing the COSMIN checklist.

Step 1. Determine which boxes need to be completed

The COSMIN checklist should be used as a modular tool. This means that it may not be necessary to complete the whole checklist when evaluating the quality of a particular study. The measurement properties evaluated in the study determine which boxes are relevant. For example, if in a study the internal consistency and reliability of an instrument were assessed, two boxes need to be completed. For evaluating the quality of the assessment of internal consistency box A should be completed. For evaluating the quality of the assessment of reliability box B should be completed. If in the same study measurement error of the instrument was not assessed, then box C does not need to be completed. Etcetera. This modular system was developed because not all measurement properties are assessed in all studies.

Sometimes the same measurement property is assessed in multiple (sub)groups in one study. For example when an instrument is validated in two different language groups from different countries in one study. In that case, the same box may need to be completed multiple times if the design of the study was different among countries. The user of the checklist should decide which boxes should be completed (and how often). This should be marked in step 1.

Step 1 sometimes requires a subjective judgement because the terms for measurement properties used in an article may not be similar to the terms used in COSMIN. Also definitions used in an article for a certain measurement property may be different from the COSMIN definitions. And finally, methods used in an article to evaluate a certain measurement property may be considered parameters of another measurement property according to the COSMIN taxonomy. Some examples and recommendations can be found on the COSMIN website under Frequently Asked Questions (www.cosmin.nl).

Step 2. Complete the IRT box if IRT methods were applied in the article
IRT methods are increasingly used for developing a PRO instrument and assessing its measurement properties. When articles use IRT methods to evaluate the measurement properties, the IRT box should be completed. This box needs to be completed only once for an article, even if multiple measurement properties were evaluated with IRT methods in the article (e.g. internal consistency and structural validity). This is because the questions in the IRT box refer to general issues about the IRT analyses, such as the software package and IRT model that were used, which are assumed to be similar for all measurement properties that were evaluated. If this is not the case, one can decide to complete the IRT box multiple times, for each assessment of a measurement property separately.

Step 3: Complete the corresponding boxes marked in step 1
In step 3, the corresponding boxes should be completed for each measurement property that was marked in step 1, to determine if the measurement properties were assessed according to the standards for good methodological quality.
In addition, box J (interpretability) should be completed if this was marked in step 1. We recommend to complete box J for quality assessment only for studies that explicitly aim to assess the interpretability of an instrument. Furthermore, in systematic reviews of measurement properties, we recommend to use the items in box J in a data extraction form, to extract data on the distribution of scores in the study population and in relevant subgroups, and data on floor-and ceiling effects and MIC from all included studies. This gives you an overview of all available information on the interpretability of scores of the instruments of interest.

Step 4. Complete the Generalisability box for each property marked in step 1.
The measurement properties of an instrument are likely to be different for different (patient) populations. For example, when evaluating reliability, the intraclass correlation coefficient (ICC) depends very much on the variation in scores in the study population. The value of the ICC is usually much higher in a heterogeneous population than in a homogeneous population. Therefore, it should be clear to which population the results of a study can be generalized. It is important that users of PRO instruments can judge whether the results of published studies on measurement properties can be assumed to apply to their study population. This should be decided based on the characteristics of the patient sample (e.g. age, gender, disease characteristics) that was used in the analyses of the measurement properties. The Generalisability box was developed to assess whether the sample in which the PRO instrument was evaluated was adequately enough described to decide to which population the results of the study can be generalized.
Alternatively, instead of using this box to rate the generalisability of the findings, one can also use items 1-6 of the Generalisability box in a data extraction form, to extract information about the characteristics of the study sample in which the measurement properties were assesed. We recommend this approach in systematic reviews of measurement properties because the information is required to decide in the data synthesis whether the results from different studies can be pooled. See for example the review of Schellingerhout et al. [12].

The Generalisability box should be completed several times, for each property that was marked in step 1. We recommend this approach because within one article different (patient) samples may have been used for the evaluation of different measurement properties. These samples may have different characteristics. It could also be possible that the characteristics of one sample were more extensively described than the characteristics of another sample. Some examples and recommendations can be found on the COSMIN website under Frequently Asked Questions (www.cosmin.nl).

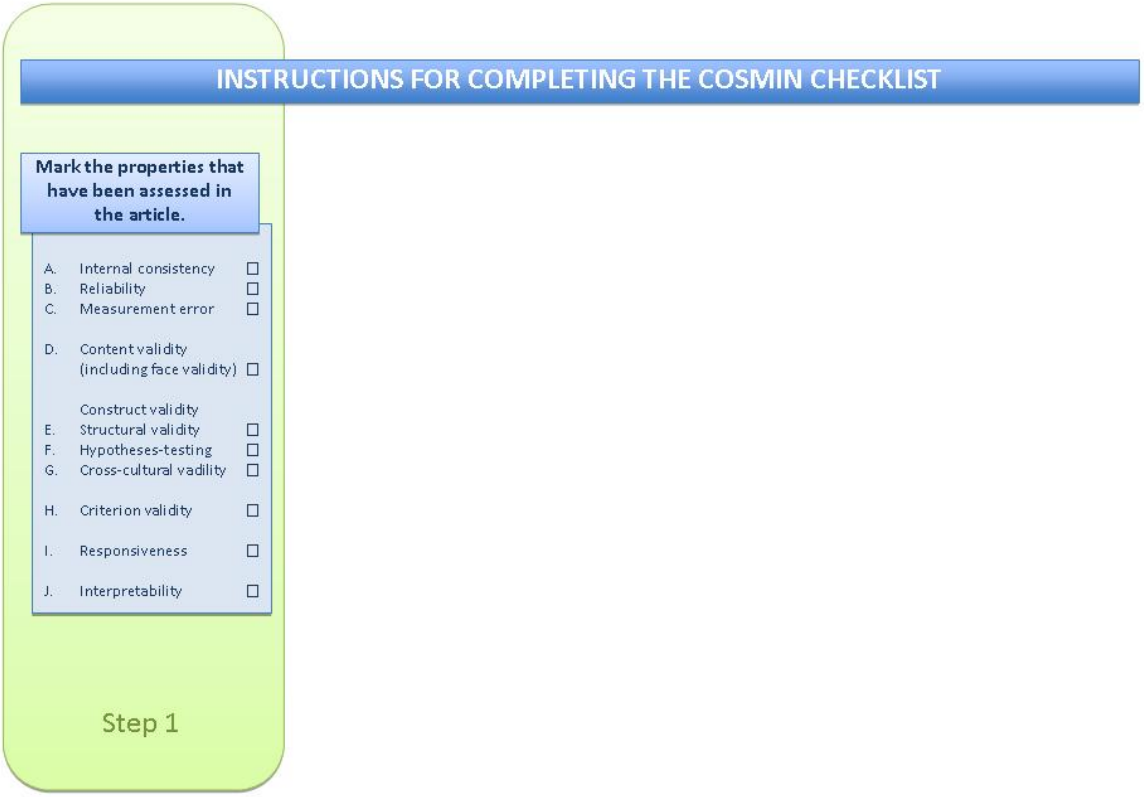
2.2 Instructions for completing the COSMIN boxes

In this chapter specific explanations and instructions for completing the COSMIN checklist are provided for with each box. There are a number of items that are included in all boxes A through J. These items are included in all boxes because they refer to general design issues, such as sample size and the number of missing values, that are relevant for the assessment of all measurement properties. These questions need to be answered for each assessment of a measurement property (thus again in each box) because different populations or designs may have been used for the evaluation of different measurement properties. For example, in a study internal consistency was evaluated in the whole study population, while reliability was evaluated in only a subgroup of the study population. The reliability study suffered from a larger number of missing values because it involved two measurements instead of just one.
Instructions for completing these general items are presented on a separate page (page 13)

Electronic data extraction
The COSMIN checklist is a modular tool. This means that it may not be necessary to complete the whole checklist when evaluating the quality of a particular study. We intend to develop an electronic data extraction form with the possibility to skip boxes. In the electronic data extraction form users can indicate which boxes they want to complete (step 1), and next only those boxes will be shown. Entered data will be stored in a database that can be opened in e.g. Excell or SPSS. The electronic data extraction form will be made available through the COSMIN website.

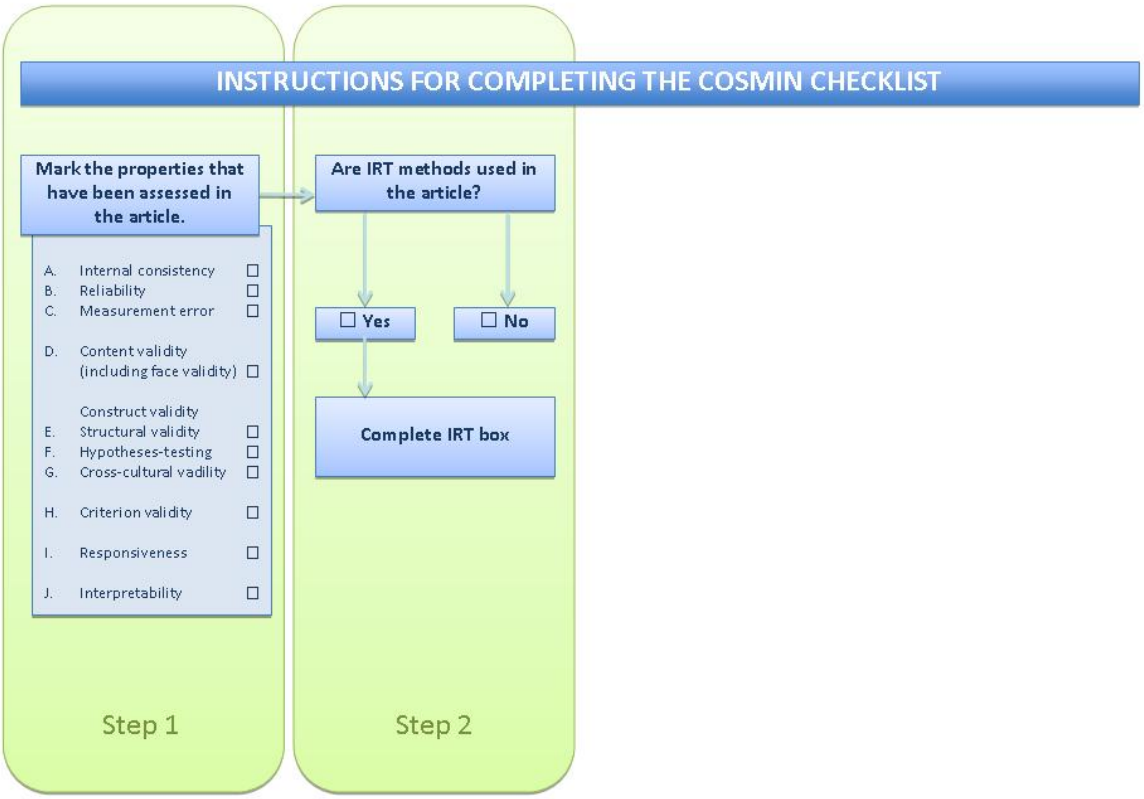
Step 1. Determine which boxes need to be completed

Mark the boxes of the (measurement) properties that have been evaluated in the article. This indicates the boxes that should be completed. If a box needs to be completed more than once (e.g. for intra-rater reliability and inter-rater reliability) you can mark this in the margin of the figure.



Step 2. Determine if the IRT box needs to be completed

When articles use IRT methods, mark here that the IRT box should be completed. This box needs to be completed only once for an article.
Then complete the IRT box (page 13).



Box IRT

Box General requirements for studies that applied Item Response Theory (IRT) models				
		yes	**no**	**?**
1	Was the IRT model used adequately described? e.g. One Parameter Logistic Model (OPLM), Partial Credit Model (PCM), Graded Response Model (GRM)	☐	☐	
2	Was the computer software package used adequately described? e.g. RUMM2020, WINSTEPS, OPLM, MULTILOG, PARSCALE, BILOG, NLMIXED	☐	☐	
3	Was the method of estimation used adequately described? e.g. conditional maximum likelihood (CML), marginal maximum likelihood (MML)	☐	☐	
4	Were the assumptions for estimating parameters of the IRT model checked? e.g. unidimensionality, local independence, and item fit (e.g. differential item functioning (DIF))	☐	☐	☐

Explanation and instructions

Item 2: Different software packages use slightly different methods to estimate IRT parameters, and therefore, the used software package must be described.

Item 4: Often assumptions must be checked before IRT methods can be applied. For example, many IRT models require unidimensional scales, and items must be locally independent.

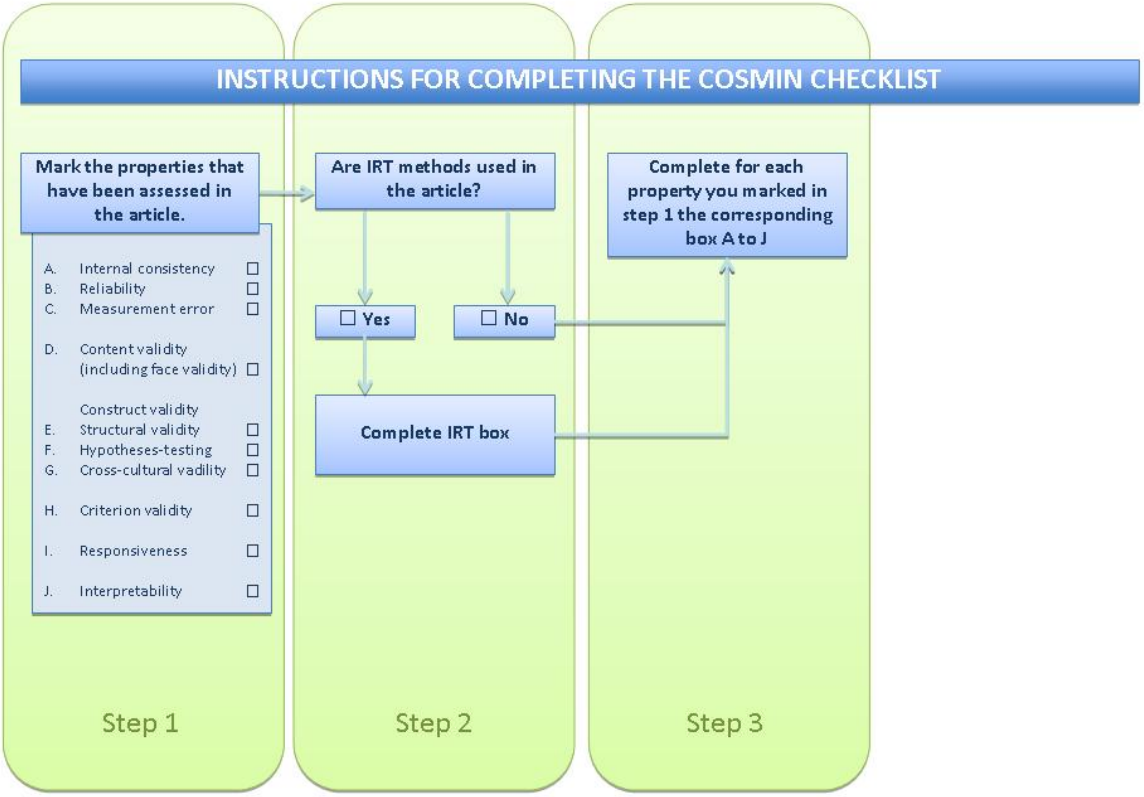
Step 3. Complete the corresponding boxes that were marked in step 1

Complete the corresponding boxes for each measurement property that was marked in step 1. In addition, complete box J (interpretability) if this was box marked in step 1.

Before completing the boxes, first read the instructions for completing the questions on general design issues (e.g. on missing items and sample size) that are included in all boxes. These instructions can be found on page 22.

Each box is accompanied with specific explanations and instructions for completing the specific items of each box.

For systematic reviews on measurement properties, we recommend to use the COSMIN checklist with 4-point rating scale. This version is described in Chapter 3. However, we still recommend to read the instructions on the next pages, because the rationale of scoring the items is similar in both versions.



Instructions for completing the questions on general design issues

There are a number of items that are included in all boxes A through J, shown in Table 2. These items are included in all these boxes because they refer to general design issues that are relevant for the assessment of all measurement properties. A rationale and instructions for completing these items are given below.

Table 2. Questions on general design issues that will appear in all boxes A through J.

	yes	no	?
Was the percentage missing items described?	☐	☐	
Was described how missing items were handled?	☐	☐	
Was the sample size included in the internal consistency analysis adequate?	☐	☐	☐
Were there any important flaws in the design or methods of the study?	☐	☐	

Firstly, each box (except box D on content validity) contains one item asking if the percentage of missing items was described. A high number of missing items can introduce bias in the results of the study if the missings were not random. Missing items could refer to the average number of missing items per instrument or the percentage of missing responses per item. We recommend to score "yes" if any or both of these two kinds of missings was described. A second item asks if it was adequately described how missing items were handled. It is important that this information is known because it may have a large influence on the scores on the instrument.

Secondly, each box (except box D) contains an item asking if the sample size of the study was adequate. This should be judged by the user of the checklist and may differ between methods. For example, factor analyses and IRT analyses require a large sample size. For factor analyses rules of thumb vary between a subject-to-variables ratio of 4:1 to 10:1, with a minimum of 100 subjects [13]. For IRT analyses the sample size depends on factors such as the IRT model chosen, the discriminative ability of the items, and the number of item parameters that are being estimated [14,15]. Recommendations vary from 100 subjects for Rasch models, to 500 subjects for models with more parameters [14]. For other measurement properties a smaller sample size may suffice. We have previously suggested a minimum sample size of 50 for studies using CTT [11], although 100 would be even better. For some measurement properties a sample size calculation can be performed. For example, for reliability studies one can estimate the number of subjects required to obtain a certain confidence interval around an ICC [16]. For example, a sample size of 50 patients is needed to obtain a confidence interval from 0.70-0.90 around an ICC of 0.80 [16]. It is also possible to perform sample size calculations for expected correlations among measures in validity studies. For this we refer to general handbooks in statistics. Note that the sample size requirements in the COSMIN checklist refer to the final sample size included in the analyses (this can often be found in the Tables of an article), which may be lower than the total sample size included in the study due to missing values or drop outs.

Thirdly, each box contains an item asking if there were any important other methodological flaws that are not covered by the checklist, but that may lead to biased results or conclusions. For example, if in a study patients were only included in the analyses if their data were complete, this could be considered a methodological flaw because selection bias might have occurred. Bias may also occur, for example, when a long version of a questionnaire is compared to a short version, while the scores of the short version were computed using the responses obtained with the longer version.

Box A – internal consistency

Box A. Internal consistency				
		yes	**no**	**?**
1	Does the scale consist of effect indicators, i.e. is it based on a reflective model?	☐	☐	☐
Design requirements		**yes**	**no**	**?**
2	Was the percentage of missing items given?	☐	☐	
3	Was there a description of how missing items were handled?	☐	☐	
4	Was the sample size included in the internal consistency analysis adequate?	☐	☐	☐
5	Was the unidimensionality of the scale checked? i.e. was factor analysis or IRT model applied?	☐	☐	
6	Was the sample size included in the unidimensionality analysis adequate?	☐	☐	☐
7	Was an internal consistency statistic calculated for each (unidimensional) (sub)scale separately?	☐	☐	☐
8	Were there any important flaws in the design or methods of the study?	☐	☐	
Statistical methods		**yes**	**no**	**NA**
9	for Classical Test Theory (CTT): Was Cronbach's alpha calculated?	☐	☐	☐
10	for dichotomous scores: Was Cronbach's alpha or KR-20 calculated?	☐	☐	☐
11	for IRT: Was a goodness of fit statistic at a global level calculated? e.g. χ^2, reliability coefficient of estimated latent trait value (index of (subject or item) separation)	☐	☐	☐

Explanation and instructions

Item 1: This item concerns the relevance of the assessment of the measurement property internal consistency for the PRO instrument under study. The internal consistency statistic only gets an interpretable meaning, when the interrelatedness among the items is determined of a set of items that together form a reflective model [17,18]. This means the (sub)scale for which internal consistency is assessed should be based on a reflective model. A reflective model is a model in which all items are a manifestation of the same underlying construct. These kind of items are called effect indicators. These items are expected to be highly correlated and interchangeable. Its counterpart is a formative model, in which the items together form the construct. These items do not need to be correlated. Therefore, internal consistency is not relevant for items that form a formative model [19-21]. For example, stress could be measured by asking about the occurrence of different situations and events that might lead to stress, such as job loss, death in a family, [22]divorce etc. These events do not need to be correlated, thus internal consistency is not relevant for such an instrument. More examples can be found on the COSMIN website under Frequently Asked Questions (www.cosmin.nl).

When the HR-PRO instrument is based on a formative model, item 1 should be scored with "no", and the other items in box A can be skipped.

Often, authors do not explicitly describe whether their instrument is based on a reflective or formative model. To decide afterwards which model is used, one can do a simple "thought test". With this test one should consider whether all item scores are expected to change when the construct changes. If yes, the construct can be considered a reflective model. If not, the HR-PRO instrument is probably based on a formative model [19,20].
It is not always possible to decide afterwards if the instrument is based on a reflective or formative model and thus whether internal consistency is relevant. In this case, we recommend to score item 1 with "?", and complete the other items in the box to assess the quality of the analyses performed.

Item 5. A second requirement for an internal consistency statistic to get an interpretable meaning is that the scale needs to be unidimensional. Internal consistency and unidimensionality are not the same. Unidimensionality is a prerequisite for a clear interpretation of the internal consistency statistics [17,18]. Unidimensionality of a scale can be investigated for example by factor analysis [23] or IRT methods, such as item factor analysis [22]. Several computer programs are available to check for unidimensionality in IRT, such as The Mokken Scale Analysis for Polytomous Items computer program (MSP), DETECT, HCA/CCPROX, and DIMTEST [24].

Item 4 and item 6. Sample size requirements for factor analysis and IRT analysis are higher than for assessing Cronbach's alpha (see the instructions for completing the questions on general design issues on page 13). Therefore, the adequateness of the sample size needs to be considered separately for the assessment of unidimensionality and the assessment of the internal consistency coefficient.

Item 7. The internal consistency coefficient should be calculated for each unidimensional subscale separately. If unidimensionality was not checked, but the authors relied on factor analyses reported in another article and internal consistency coefficients are reported for each previously identified subscale, we recommend to rate item 5 with "no" and rate item 7 with "yes".

Box B – reliability

Box B. Reliability: relative measures (including test-retest reliability, inter-rater reliability and intra-rater reliability)					
Design requirements			**yes**	**no**	**?**
1	Was the percentage of missing items given?		☐	☐	
2	Was there a description of how missing items were handled?		☐	☐	
3	Was the sample size included in the analysis adequate?		☐	☐	☐
4	Were at least two measurements available?		☐	☐	
5	Were the administrations independent?		☐	☐	☐
6	Was the time interval stated?		☐	☐	
7	Were patients stable in the interim period on the construct to be measured?		☐	☐	☐
8	Was the time interval appropriate?		☐	☐	☐
9	Were the test conditions similar for both measurements? e.g. type of administration, environment, instructions		☐	☐	☐
10	Were there any important flaws in the design or methods of the study?		☐	☐	
Statistical methods		**yes**	**no**	**NA**	**?**
11	for continuous scores: Was an intraclass correlation coefficient (ICC) calculated?	☐	☐	☐	
12	for dichotomous/nominal/ordinal scores: Was kappa calculated?	☐	☐	☐	
13	for ordinal scores: Was a weighted kappa calculated?	☐	☐	☐	☐
14	for ordinal scores: Was the weighting scheme described? e.g. linear, quadratic	☐	☐	☐	

Explanation and instructions

Item 4 and item 5. To evaluate reliability the instrument should have been administered twice. The administrations should have been independent. This implies that the first administration has not influenced the second administration. At the second administration the patient or rater should not have been aware of the scores on the first administration.

Item 6 and item 8. The time interval between the administrations must be appropriate. The time interval should be long enough to prevent recall bias, and short enough to ensure that patients have not been changed on the construct to be measured. What an appropriate time interval is, depends on the construct to be measured and the target population. A time interval of about 2 weeks is often considered appropriate for the evaluation of PRO instruments [25].

Item 7. Patients should be stable with regard to the construct to be measured between the administrations. What "stable patients" are depends on the construct to be measured and the target population. Evidence that patients were stable could be, for example, an assessment of a global rating of change, completed by patients or physicians. When an intervention is given in the interim period, one can assume that (many of) the patients have changed on the construct to be measured. In that case, we recommend to score item 7 with "no".

Item 9. A last requirement is that the test conditions should be similar. Test conditions refer to the type of administration (e.g. a self-administered questionnaire, interview, performance-test), the setting in which the instrument was administered (e.g. at the hospital, or at home), and the instructions given. These test conditions may influence the responses of a patient. The reliability may be underestimated if the test conditions are not similar.

Item 11. The preferred reliability statistic depends on the type of response options. For continuous scores the intraclass correlation coefficient (ICC) is preferred [25,26]. The use of the Pearson's and Spearman's correlation coefficient is considered not adequate, because they do not take systematic error into account.

Item 12, item 13, and item 14. For dichotomous scores or nominal scores the Cohen's kappa is the preferred statistical method [25]. For ordinal scales partial chance agreement should be considered, and therefore a weighted kappa [25,27] is preferred. A description of the weights (e.g., linear or quadratic weights [28]) should be given. Proportion agreement is considered not adequate, because it does not correct for chance agreement.

No IRT methods for assessing reliability were found in the literature or suggested by any of the panel members.

Box C – measurement error

Box C. Measurement error: absolute measures				
Design requirements		**yes**	**no**	**?**
1	Was the percentage of missing items given?	☐	☐	
2	Was there a description of how missing items were handled?	☐	☐	
3	Was the sample size included in the analysis adequate?	☐	☐	☐
4	Were at least two measurements available?	☐	☐	
5	Were the administrations independent?	☐	☐	☐
6	Was the time interval stated?	☐	☐	
7	Were patients stable in the interim period on the construct to be measured?	☐	☐	☐
8	Was the time interval appropriate?	☐	☐	☐
9	Were the test conditions similar for both measurements? e.g. type of administration, environment, instructions	☐	☐	☐
10	Were there any important flaws in the design or methods of the study?	☐	☐	
Statistical methods		**yes**	**no**	**?**
11	for CTT: Was the Standard Error of Measurement (SEM), Smallest Detectable Change (SDC) or Limits of Agreement (LoA) calculated?	☐	☐	

Explanation and instructions

Item 4 and item 5. To evaluate measurement error the instrument should have been administered twice. The administrations should have been independent. This implies that the first administration has not influenced the second administration. At the second administration the patient or rater should not have been aware of the scores on the first administration.

Item 6 and item 8. The time interval between the administrations must be appropriate. The time interval should be long enough to prevent recall bias, and short enough to ensure that patients have not been changed on the construct to be measured. What an appropriate time interval is, depends on the construct to be measured and the target population. A time interval of about 2 weeks is often considered appropriate for the evaluation of PRO instruments [25].

Item 7. Patients should be stable with regard to the construct to be measured between the administrations. What "stable patients" are depends on the construct to be measured and the target population. Evidence that patients were stable could be, for example, an assessment of a global rating of change, completed by patients or physicians. When an intervention is given in the interim period, one can assume that (many of) the patients

have changed on the construct to be measured. In that case, we recommend to score item 7 with "no".

Item 9. A last requirement is that the test conditions should be similar. Test conditions refer to the type of administration (e.g. a self-administered questionnaire, interview, performance-test), the setting in which the instrument was administered (e.g. at the hospital, or at home), and the instructions given. These test conditions may influence the responses of a patient. The measurement error may be overestimated if the test conditions are not similar.

Item 11. The preferred statistic for measurement error in studies based on CTT is the standard error of measurement (SEM). Note that the requirement of two administrations for evaluating measurement error implies that the calculation of the SEM based on Cronbach's alpha is considered not appropriate, because it does not take the variance between time points into account [29]. Other appropriate statistics for assessing measurement error are the limits of agreement (LoA) and the smallest detectable change (SDC) [29]. Both parameters are directly related to the SEM [29]. Changes within the LoA or smaller than the SDC are likely to be due to measurement error and changes outside the LoA or larger than the SDC should be considered as real change. Note that this does not indicate that these changes are also meaningful to patients. This depends on what change is considered important, which is an issue of interpretability (Box J).

Box D – content validity

Box D. Content validity (including face validity)				
General requirements		**yes**	**no**	**?**
1	Was there an assessment of whether all items refer to relevant aspects of the construct to be measured?	☐	☐	☐
2	Was there an assessment of whether all items are relevant for the study population? (e.g. age, gender, disease characteristics, country, setting)	☐	☐	☐
3	Was there an assessment of whether all items are relevant for the purpose of the measurement instrument? (discriminative, evaluative, and/or predictive)	☐	☐	☐
4	Was there an assessment of whether all items together comprehensively reflect the construct to be measured?	☐	☐	☐
5	Were there any important flaws in the design or methods of the study?	☐	☐	

Explanation and instructions

Content validity should be assessed by making a judgment about the *relevance* and the *comprehensiveness* of the items. The user of the checklist should assess whether the authors of an article on content validity have adequately judged the relevance and the comprehensiveness of the items. An appropriate method might be to let experts judge the relevance and comprehensiveness of the items.

Item 2. When a new instrument is developed, the focus and detail of the content of the instrument should match the target population. This could have been assessed by letting the target population judge this. If the instrument concerns a PRO instrument, then patients should judge the relevance of the items for the patient population. In addition, many missing observations on an item can be an indication that the item is not relevant for the population.
Sometimes an instrument is used in a different population than the original target population for which it was developed. In that case it should be assessed whether all items are relevant for this new study population. For example, a questionnaire measuring shoulder disability (i.e., the Shoulder Disability Questionnaire) may include the item "my shoulder hurts when I bring my hand towards the back of my head" [30]. When one decides to use this questionnaire in a population of patients with wrist problems to measure wrist disability, one could not simply change the word "shoulder" into "wrist" because this item might not be relevant for patients with wrist problems. Moreover, an item like "Do you have difficulty with the grasping and use of small objects such as keys or pens?" [31] will probably not be included in a questionnaire for shoulder disability, while it is relevant to ask patients with wrist problems.

Item 4. To assess the comprehensiveness of the items three aspects should have been taken into account: the content coverage of the items, the description of the domains, and the theoretical foundation. The first two refer to the question if *all* relevant aspects of the construct are covered by the items and the domains. The theoretical foundation refers to a clear description of the construct, and the theory on which it is based. A part of this theoretical foundation could be a description of how different constructs within a concept are interrelated, like for instance described in the model of health status of Wilson and Cleary [32] or the International Classification of Functioning, Disability and Health (ICF) model [33]. When patients or experts were asked whether they missed items this could be considered as an indication that the comprehensiveness of the items was assessed. A large number of patients with the highest or lowest possible score on a scale may be an indication that items are missing.

Face validity requires a subjective judgement. Therefore, there were no standards developed for assessing face validity.

Box E – structural validity

Box E. Structural validity				
		yes	**no**	**?**
1	Does the scale consist of effect indicators, i.e. is it based on a reflective model?	☐	☐	☐
Design requirements		**yes**	**no**	**?**
2	Was the percentage of missing items given?	☐	☐	
3	Was there a description of how missing items were handled?	☐	☐	
4	Was the sample size included in the analysis adequate?	☐	☐	☐
5	Were there any important flaws in the design or methods of the study?	☐	☐	
Statistical methods		**yes**	**no**	**NA**
6	for CTT: Was exploratory or confirmatory factor analysis performed?	☐	☐	☐
7	for IRT: Were IRT tests for determining the (uni-) dimensionality of the items performed?	☐	☐	☐

Explanation and instructions

Item 1. Structural validity is only relevant for instruments that are based on a reflective model (see box A for an explanation). Therefore, item 1 is included to check the relevance of evaluating structural validity. When item 1 is scored with "no", the other items in box E are not relevant and can be skipped.

Item 6. To determine the structure of the instrument, a factor analysis is the preferred statistic when using CTT. Although confirmatory factor analysis is preferred over explorative factor analysis, both could be useful for the evaluation of structural validity. Confirmative factor analysis tests whether the data fit a premeditated factor structure [34]. Based on theory or previous analyses a priori hypotheses are formulated and tested. Explorative factor analysis can be used when no clear hypotheses exist about the underlying dimensions, or to reduce the number of items [34].
In the COSMIN study we did not discuss specific requirements for factor analyses, such as the choice of the explorative factor analysis (principal component analysis or common factor analysis), the choice and justification of the rotation method (e.g. orthogonal or oblique rotation), or the decision about the number of relevant factors. Such specific requirements are described by e.g. Floyd & Widaman [34] and de Vet et al. [35]. When there are serious flaws in the quality of the factor analysis, we recommend to score item 5 with "yes".

Item 7. Some IRT software programs also include analyses to check the dimensionality of the items such as The Mokken Scale Analysis for Polytomous Items computer program (MSP), DETECT, HCA/CCPROX, and DIMTEST [24].

Box F – hypotheses testing

Box F. Hypotheses testing				
Design requirements		**yes**	**no**	**?**
1	Was the percentage of missing items given?	☐	☐	
2	Was there a description of how missing items were handled?	☐	☐	
3	Was the sample size included in the analysis adequate?	☐	☐	☐
4	Were hypotheses regarding correlations or mean differences formulated a priori (i.e. before data collection)?	☐	☐	☐*
		yes	**no**	**NA**
5	Was the expected *direction* of correlations or mean differences included in the hypotheses?	☐	☐	☐
6	Was the expected absolute or relative *magnitude* of correlations or mean differences included in the hypotheses?	☐	☐	☐
7	for convergent validity: Was an adequate description provided of the comparator instrument(s)?	☐	☐	
8	for convergent validity: Were the measurement properties of the comparator instrument(s) adequately described?	☐	☐	
9	Were there any important flaws in the design or methods of the study?	☐	☐	
Statistical methods		**yes**	**no**	**NA**
10	Were design and statistical methods adequate for the hypotheses to be tested?	☐	☐	☐

Explanation and instructions

Item 4. Specific hypotheses to be tested should have been formulated a priori, and preferably stated in the methods section of an article. Without specific hypotheses, the risk of bias is high because retrospectively it is tempting to think up alternative explanations for low correlations instead of concluding that an instrument may not be valid. The hypotheses should concern expected mean differences between groups or expected correlations between the scores on the instrument and other variables, such as scores on other instruments, or demographic or clinical variables. The hypotheses may also concern the relative magnitude of correlations, for example a statement that instrument A is expected to correlate higher with instrument B than with instrument C.

Hypotheses testing is an ongoing, iterative process [36]. The more specific the hypotheses are and the more hypotheses are being tested, the more evidence is gathered for construct validity. However, the COSMIN panel considered it not possible to formulate standards for the amount of hypotheses that need to be tested in a construct

validity study. This depends on the construct to be measured and the content and measurement properties of the comparator instruments.

Item 5 and item 6. The expected direction (positive or negative) and magnitude (absolute or relative) of the correlations or differences should have been included in the hypotheses (e.g. [36-39]). Without this specification it is difficult to decide afterwards whether the hypothesis is confirmed or not. For example, authors could have stated that they expected a correlation of at least 0.60 between two instruments that intend to measure the same construct. Or authors could have stated that they expected a mean difference of 10 points in score on the instrument between two patient groups who are expected to differ in the construct to be measured.
The hypotheses may also concern the relative magnitude of correlations. For example, authors could have stated that they expected that the score on measure A correlates at least 0.10 points higher with the score on measure B than with the score on measure C.

Item 7. When hypotheses were formulated about expected relations with other instruments, these comparator instruments should have been appropriately described in terms of the construct they intend to measure. For example, if the comparator instrument is a VAS measuring pain, it should have been described if the pain refers to the average or worst pain, and to which time period.

Item 8. The measurement properties of the comparator instruments should be adequate. Otherwise it is difficult to decide afterwards whether negative results are due to lack of validity of the instrument under study or poor quality of the comparator instrument. The measurement properties of the comparator instruments should have been described or references should have been provided to studies in which these properties are described. Ideally, the measurement properties of the comparator instruments should have been assessed in the same language version, and the same patient population as is used in the study.

Item 10. Many different hypotheses can be formulated and tested. The users of the COSMIN checklist have to decide whether or not the statistical methods used in the article are adequate for testing the stated hypotheses. P-values should be avoided in testing hypotheses, because it is not relevant to examine whether correlations statistically differ from zero [40]. The validity issue is about whether the direction and magnitude of a correlation is similar to what could be expected based on the construct(s) that are being measured. When assessing differences between groups, it is also less relevant whether these differences are statistically significant (which depends on the sample size) than whether these differences are as large as could be expected.

Box G – cross-cultural validity

Box G. Cross-cultural validity				
Design requirements		**yes**	**no**	**?**
1	Was the percentage of missing items given?	☐	☐	
2	Was there a description of how missing items were handled?	☐	☐	
3	Was the sample size included in the analysis adequate?	☐	☐	☐
4	Were both the original language in which the HR-PRO instrument was developed, and the language in which the HR-PRO instrument was translated described?	☐	☐	
5	Was the expertise of the people involved in the translation process adequately described? e.g. expertise in the disease(s) involved, expertise in the construct to be measured, expertise in both languages	☐	☐	
6	Did the translators work independently from each other?	☐	☐	☐
7	Were items translated forward and backward?	☐	☐	☐
8	Was there an adequate description of how differences between the original and translated versions were resolved?	☐	☐	
9	Was the translation reviewed by a committee (e.g. original developers)?	☐	☐	
10	Was the HR-PRO instrument pre-tested (e.g. cognitive interviews) to check interpretation, cultural relevance of the translation, and ease of comprehension?	☐	☐	
11	Was the sample used in the pre-test adequately described?	☐	☐	
12	Were the samples similar for all characteristics except language and/or cultural background?	☐	☐	☐
13	Were there any important flaws in the design or methods of the study?	☐	☐	
Statistical methods		**yes**	**no**	**NA**
14	for CTT: Was confirmatory factor analysis performed?	☐	☐	☐
15	for IRT: Was differential item function (DIF) between language groups assessed?	☐	☐	☐

<u>Explanation and instructions</u>

When evaluating cross-cultural validity, all items are applicable. When an instrument is only translated, but cross-cultural validity was not assessed, the items 4 though 11 can be used to evaluate the quality of the translation procedure.

Because of language and cultural differences, a simple translation is not sufficient. An adequate procedure contains multiple forward and backward translations with at least two translators per step. The standards in this box are based on existing guidelines for

translation and adaptation of measurement instruments, such as guidelines developed by International Quality of Life Assessment (IQOLA) [41], the MAPI Research Institute [42], or the European Organisation of Research and Treatment of Cancer (EORTC) [43].

Item 5. The characteristics and qualifications of each of the members should have been described, in terms of expertise in the languages, in the disease of the target population, and in the construct to be measured.
The specific qualifications of the translators have not been discussed in the COSMIN study. However, such specific requirements can be found in many translation guidelines. It is generally recommended that the forward translators should have the target language as their mother tongue. It is recommended that one translator has expertise on the construct to be measured, the second one being a language expert, but naive on the topic. The back translators should have the original language as their mother tongue. They should be blind for the original version of the questionnaire. It is recommended that the back translators are both language experts and naïve to the constructs to be measured. If users of the COSMIN checklist consider the qualifications of the translators inadequate, we recommend to score item 13 with "yes".

Item 6. To allow detection of errors, divergent interpretation or ambiguous items in the original version [44], the translators should have worked independently from each other. If only one translator was involved, item 6 should be scored "no".

Item 7 and item 8. To further uncover mistakes in the new version, the items should have been translated forward (into the new language) and backward (back to the original language). If differences occurred between the original version, and the backward translated version, it should have been described how these differences were resolved.

Item 9. A committee should have reviewed the final translation. Preferably including the developers of the original instrument, as they know best what the items were aimed to measure. This team should be multidisciplinary, with expertise in the disease involved, and the construct to be measured, and with the involvement of members of the target population who speak the language in which the instrument was translated. These latter persons are well able to judge whether or not culturally relevant idioms are used [44].

Item 10 and item 11. A pre-test should have been performed to check the interpretation and cultural relevance of the items, and the ease of comprehension. The sample in which the translation was pre-tested should have been described in terms of age, gender, disease characteristics, and setting.

Item 12. When cross-cultural validity is assessed, the samples should be similar (e.g. in terms of age, gender, disease characteristics) except for their language.

Item 14. The preferred statistical method for assessing cross-cultural validity using CTT is confirmatory factor analysis (CFA). Based on the theoretical foundation and the factor structure of the original instrument the hypothesized factor structure can be tested using CFA.

Item 15. The preferred statistical method for assessing cross-cultural validity using IRT methods is differential item functioning (DIF) analyses [45]. DIF examines the equivalence between two versions of the same instrument. It examines whether respondents with the same level of the scale score do respond similar to a particular item. DIF can also be examined by using regression analyses (see for example [46]).

Box H – criterion validity

Box H. Criterion validity				
Design requirements		**yes**	**no**	**?**
1	Was the percentage of missing items given?	☐	☐	
2	Was there a description of how missing items were handled?	☐	☐	
3	Was the sample size included in the analysis adequate?	☐	☐	☐
4	Can the criterion used or employed be considered as a reasonable 'gold standard'?	☐	☐	☐
5	Were there any important flaws in the design or methods of the study?	☐	☐	
Statistical methods		**yes**	**no**	**NA**
6	for continuous scores: Were correlations, or the area under the receiver operating curve calculated?	☐	☐	☐
7	for dichotomous scores: Were sensitivity and specificity determined?	☐	☐	☐

Explanation and instructions

Item 4. The criterion used should be considered as a reasonable "gold standard". The COSMIN panel reached consensus that no gold standard exist for HR-PRO instruments. The only exception is when a shortened instrument is compared to the original long version. In that case, the original long version can be considered the gold standard. Often, authors consider their comparator instrument wrongly as a gold standard, for example when they compare the scores of a new instrument to a widely used instrument.

Item 6 and item 7. When both the HR-PRO instrument and the gold standard have continuous scores, correlation is the preferred statistical method. When the instrument scores are continuous and scores on the gold standard are dichotomous the area under the receiver operating characteristic (ROC) is the preferred method, and when both scores are dichotomous sensitivity and specificity are the preferred methods to use.

Box I – responsiveness

Box I. Responsiveness				
Design requirements		**yes**	**no**	**?**
1	Was the percentage of missing items given?	☐	☐	
2	Was there a description of how missing items were handled?	☐	☐	
3	Was the sample size included in the analysis adequate?	☐	☐	☐
4	Was a longitudinal design with at least two measurement used?	☐	☐	
5	Was the time interval stated?	☐	☐	
6	If anything occurred in the interim period (e.g. intervention, other relevant events), was it adequately described?	☐	☐	
7	Was a proportion of the patients changed (i.e. improvement or deterioration)?	☐	☐	
Design requirements for hypotheses testing		**yes**	**no**	**?**
For constructs for which a gold standard was not available:				
8	Were hypotheses about changes in scores formulated a priori (i.e. before data collection)?	☐	☐	☐*
		yes	**no**	**NA**
9	Was the expected *direction* of correlations or mean differences of the change scores of HR-PRO instruments included in these hypotheses?	☐	☐	☐
10	Were the expected absolute or relative *magnitude* of correlations or mean differences of the change scores of HR-PRO instruments included in these hypotheses?	☐	☐	☐
11	Was an adequate description provided of the comparator instrument(s)?	☐	☐	
12	Were the measurement properties of the comparator instrument(s) adequately described?	☐	☐	
13	Were there any important flaws in the design or methods of the study?	☐	☐	
Statistical methods		**yes**	**no**	**NA**
14	Were design and statistical methods adequate for the hypotheses to be tested?	☐	☐	☐
Design requirement for comparison to a gold standard		**yes**	**no**	**?**
For constructs for which a gold standard was available:				
15	Can the criterion for change be considered as a reasonable gold standard?	☐	☐	☐
16	Were there any important flaws in the design or methods of the study?	☐	☐	

Statistical methods		yes	no	NA
17	for continuous scores: Were correlations between change scores, or the area under the Receiver Operator Curve (ROC) curve calculated?	☐	☐	☐
18	for dichotomous scales: Were sensitivity and specificity (changed versus not changed) determined?	☐	☐	☐

Explanation and instructions

Although responsiveness is considered to be a separate measurement property from validity, the only difference between cross-sectional (construct and criterion) validity and responsiveness is that validity refers to the validity of a single score, and responsiveness refers to the validity of a change score [3]. Therefore, the standards for responsiveness are similar to the standards of construct and criterion validity. Similarly as evaluating construct and criterion validity, the design requirements for assessing responsiveness are divided in situations in which a gold standard is available (items 1-7 and 15 to 18 should be completed), and situations in which a gold standard is not available (items 1-14 should be completed). Although the approach is similar to construct and criterion validity, the COSMIN panel decided not to use the terms construct and criterion responsiveness, because these terms are unfamiliar in the literature.

Item 4. Because responsiveness is about detecting changes in scores on an instrument, at least two measurements should have been administered in a longitudinal design.

Item 5 and item 6. To examine if changes in scores can be expected, the time interval between the measurements should have been stated and it should have been described what happened in the interim period.

Item 7. Some evidence should have been provided to assume that at least a proportion of the patients have been improved or deteriorated on the construct to be measured. Otherwise it is difficult to decide afterwards whether the patients did not truly change or whether the measurement instrument was not responsive when no change on the instrument was observed. If the sample existed of patients with chronic progressive diseases, or if an intervention has been given, or another relevant event has happened, it is likely that part of the patients have changed. Sometimes a global rating scale of change was used to ask patients if they considered themselves as changed on the construct. Some authors assume that patients have changed because an intervention 'with proven effectiveness' was applied to the patients. Users of the checklist should be cautious with interpreting such a statement, because often the evidence of the effectiveness of the intervention comes from the same study in which the responsiveness of the measurement instrument is being evaluated. This is considered circular reasoning.

Item 8. Specific hypotheses to be tested should have been formulated a priori, and preferably stated in the methods section of an article. Without specific hypotheses, the risk of bias is high because retrospectively it is tempting to think up alternative

explanations for low correlations instead of concluding that an instrument may not be responsive.
One of the most difficult tasks when testing hypothesis, is formulating challenging hypotheses. By testing hypotheses we aim to show that the instrument truly measures changes in the construct(s) it purports to measure. In practice, this means that the instrument should measure changes in the right construct(s) and not changes in something else, but also that it should measure the right amount of change, i.e. it should not under- or overestimate the real change in the construct that has occurred (see items 9 and 10). This latter aspect is often overlooked in assessing responsiveness.
The hypotheses should concern expected mean differences between changes in groups or expected correlations between changes in the scores on the instrument and changes in other variables, such as scores on other instruments, or demographic or clinical variables. Hypotheses about expected effect size (ES) or similar measures such as standardized response mean (SRM) can also be used, but only when an explicit hypothesis (and rationale) for the expected magnitude of the effect size is given. The hypotheses may also concern the relative magnitude of correlations, for example a statement that change in instrument A is expected to correlate higher with change in instrument B than with change in instrument C.

Hypotheses testing is an ongoing, iterative process [36]. The more specific the hypotheses are and the more hypotheses are being tested, the more evidence is gathered for responsiveness. However, the COSMIN panel considered it not possible to formulate standards for the amount of hypotheses that need to be tested in a responsiveness study. This depends on the construct to be measured and the content and measurement properties of the comparator instruments.

Item 9 and item 10. The expected direction (positive or negative) and magnitude (absolute or relative) of the correlations or differences should have been included in the hypotheses (e.g. [36,38,39,47]). Without this specification it is difficult to decide afterwards whether the hypothesis is confirmed or not. For example, authors could have stated that they expected a correlation of at least 0.60 between changes in two instruments that intend to measure the same construct. Or authors could have stated that they expected a mean difference of 10 points in changes in scores on the instrument between two patient groups who are expected to differ in the construct to be measured. The hypotheses may also concern the relative magnitude of correlations. For example, authors could have stated that they expected that the change in score on measure A correlates at least 0.10 points higher with the change in score on measure B than with the change in score on measure C.

Item 11. When hypotheses were formulated about expected relations with other instruments, these comparator instruments should have been appropriately described in terms of the construct they intend to measure. For example, if the comparator instrument is a global rating scale assessing perceived change, it should have been described if the change refers to overall change or change in a specific construct, e.g. change in pain or functioning. A global rating scale is not considered to be a gold standard, because its validity and reliability is not perfect [48,49]. However, this scale can be used as a construct approach of responsiveness. In that case it is recommended to define and test hypotheses e.g. about the expected correlation between changes on the instrument under study and the global rating scale.

Item 12. The measurement properties of the comparator instruments should be adequate. Otherwise it is difficult to decide afterwards whether negative results are due to lack of responsiveness of the instrument under study or poor quality of the comparator instrument. The measurement properties of the comparator instruments should have been described or references should have been provided to studies in which these properties are described. Ideally, the measurement properties of the comparator instruments should have been assessed in the same language version, and the same patient population as is used in the study.

Item 14. Since many different hypotheses can be formulated and tested, the users of the COSMIN checklist have to decide whether or not the statistical methods used in the article are adequate for testing the stated hypotheses. P-values should be avoided in testing hypotheses, because it is not relevant to examine whether correlations statistically differ from zero [50]. The responsiveness issue is about whether the direction and magnitude of a correlation is similar to what could be expected based on the construct(s) that are being measured. When assessing differences between groups, it is also less relevant whether these differences are statistically significant (which depends on the sample size) than whether these differences are as large as could be expected.

Item 15. No gold standards exists for HR-PRO instruments, only the original longer version of a HR-PRO can be considered a gold standard, when it is compared to its shorter version. In studies on PRO instruments often a global rating of change is used as a comparator instrument. This measure has a high face validity (provided that the rating scale asks about the same construct as the instrument under study). However, some authors have questioned the reliability and validity of such retrospective measures of change [51]. Therefore, this rating scale was not considered an appropriate gold standard for assessing responsiveness. It could, however, be considered a useful comparator instrument in a construct approach (see also item 11).

Item 17 and Item 18. A correlation between change scores is the preferred method for comparing changes in the instrument with changes in a gold standard. If the scores on the gold standard are dichotomous, the area under the Receiver Operator Curve (ROC) curve is the preferred method. If the scores of the instrument under study are also dichotomous, sensitivity and specificity are the preferred parameters.

<u>Inappropriate measures of responsiveness</u>
There are a number of parameters proposed in the literature to assess responsiveness that we consider inappropriate. The use of effect sizes (mean change score / SD baseline) [52], and related measures, such as standardised response mean (mean change score / SD change score) [53], Norman's responsiveness coefficient (σ^2 change/ σ^2 change + σ^2 error) [54], and relative efficacy statistic ((t-statistic$_1$ / t-statistic$_2$)^2) [55] are inappropriate measures of responsiveness.
These measures are considered measures to interpret changes in health status, or to interpret the magnitude of an intervention or other event, rather than measures of the quality of the measurement instrument [56,57].

It is impossible to assess in one study both the treatment effect and the responsiveness of measurement instrument based on the same effect size. If the effect size is zero, either the intervention has no effect or the outcome measure is not responsive. If the effect size is moderate, multiple conclusions are possible: either the effect is moderate and the outcome measure is responsive, or the effect is large or small and the outcome measure has poor responsiveness because the true effect is over- of underestimated by the instrument. So the argument of the COSMIN panel is that the effect size only has meaning as a measure of responsiveness if we know (or assume) beforehand what the magnitude of the effect of the intervention is. If, for example, we expect a large effect of the intervention we can test the hypothesis that the measurement instrument shows an effect size of 0.8 or higher. But if we expect a small effect of the intervention, we would not expect such a high effect size. This example shows that a high effect size does not necessarily indicates a good responsiveness.

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

Guyatt's responsiveness ratio (MIC/SD change score of stable patients) [58] was also considered to be inappropriate, because it takes the minimal important change into account. Minimal important change is about the interpretation of the change score, not about the validity of the change score. The paired t-test was considered to be inappropriate because it is a measure of significant change instead of valid change, and it is dependent on the sample size of the study [50].

Box J – interpretability

Box J. Interpretability				
		yes	**no**	**?**
1	Was the percentage of missing items given?	☐	☐	
2	Was there a description of how missing items were handled?	☐	☐	
3	Was the sample size included in the analysis adequate?	☐	☐	☐
4	Was the distribution of the (total) scores in the study sample described?	☐	☐	
5	Was the percentage of the respondents who had the lowest possible (total) score described?	☐	☐	
6	Was the percentage of the respondents who had the highest possible (total) score described?	☐	☐	
7	Were scores and change scores (i.e. means and SD) presented for relevant (sub) groups? e.g. for normative groups, subgroups of patients, or the general population	☐	☐	
8	Was the minimal important change (MIC) or the minimal important difference (MID) determined?	☐	☐	
9	Were there any important flaws in the design or methods of the study?	☐	☐	

Explanation and instructions

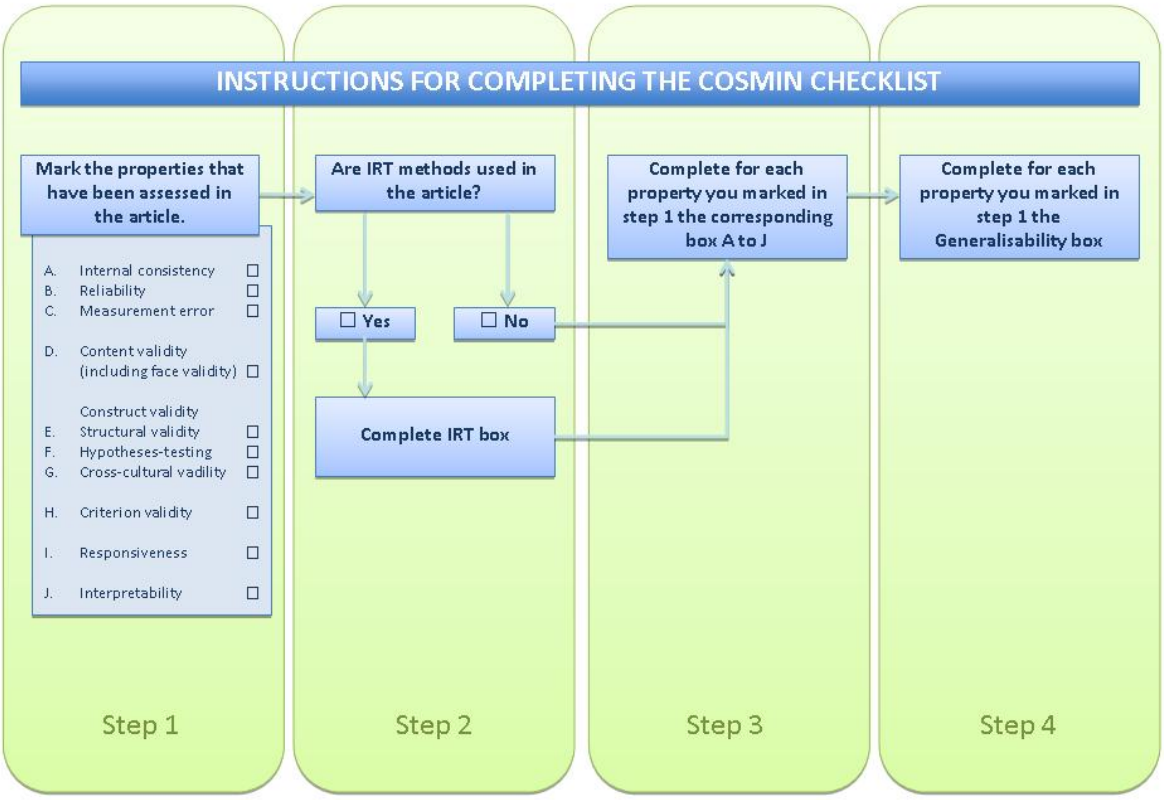
Item 4 and item 7. To facilitate interpretation of scores, the distribution of the scores in the study population should be described, preferably by showing the entire distribution (e.g. in a histogram), in addition to the mean and SD. Also mean and SD of scores and change scores in relevant (sub) groups should be presented, for example in different age and gender groups or in groups with different disease characteristics.

Item 5 and item 6. The percentage patients who have the lowest and highest possible scores should be described. If many patients have the same score this may influences reliability because patients who have the same score cannot be distinguished from each other. It may also influence responsiveness because patients who already have the highest or lowest possible score cannot change anymore in that direction.

Item 8. The minimal important change (MIC) or minimal important difference (MID) should be determined. The MIC is the smallest change in score in the construct to be measured which patient perceive as important. The MID is the smallest differences in the construct to be measured between patients that is considered important [59]. There is an ongoing discussion in the literature about which methods should be used to determine the MIC or MID of a HR-PRO instrument. Therefore in the COSMIN study no standards for assessing MIC were defined. If users of the COSMIN checklist think there were major flaws in the methods for assessing MIC or MID, we recommend to score item 9 with “yes”.

Step 4. Complete the Generalisability box for each property marked in step 1

The Generalisability box should be completed several times, for each property that was marked in step 1.



Box generaliability

Box Generalisability box					
			yes	**no**	**NA**
	Was the sample in which the HR-PRO instrument was evaluated adequately described? In terms of:				
	1	median or mean age (with standard deviation or range)?	☐	☐	
	2	distribution of sex?	☐	☐	
	3	important disease characteristics (e.g. severity, status, duration) and description of treatment?	☐	☐	☐
	4	setting(s) in which the study was conducted? e.g. general population, primary care or hospital/rehabilitation care	☐	☐	
	5	countries in which the study was conducted?	☐	☐	
	6	language in which the HR-PRO instrument was evaluated?	☐	☐	
7	Was the method used to select patients adequately described? e.g. convenience, consecutive, or random		☐	☐	
			yes	**no**	**?**
8	Was the percentage of missing responses (response rate) acceptable?		☐	☐	☐

Explanation and instructions

Item 1, item 2, and item 3. To know to which patient population the results of a study on a measurement property can be generalized, the study sample must be adequately described in terms of their (mean) age, gender, and important disease characteristics, such as severity or duration of disease. What the most important disease characteristics are, should be determined by the users of the checklist. Item 3 is not relevant when the underlying population is a healthy population. In that case we recommend to score item 3 with "not applicable".

Item 4, item 5, and item 6. The setting in which the patients were recruited is important because measurement properties can be different for patients recruited for example in primary care versus patients recruited in secondary care. It is also important to know the country in which the study was performed because the measurement properties may be different in different countries due to cultural differences. For the same reason, it is important for questionnaires to know which language version of a questionnaire was studied. In our experience often the country in which the study is performed and the language version of the measurement instrument that was used are not mentioned explicitly, but can be deduced from the affiliation of the authors.

Item 7. It is more clear to which patient population the results of a study can be generalized when a randomly selected sample, or a sample of consecutive patients was used, than when a convenience sample was used. Especially when patients are recruited

by newspaper advertisements, or patient-societies, it may be unclear to which patient population the results can be generalized.

Item 8. Finally, the percentage of missing responses should be acceptable. A high percentage of missing responses on the instrument can be an indication of selection bias [60]. The percentage of missing responses can be interpreted as the percentage of patients who did not want to participate in the study (non-response rate), or as the percentage of patients who received the instrument but did not complete it. When the study has a longitudinal design, the non-response on the second administration should be considered as well. It is up to the user of the checklist what is considered acceptable.

3. Criteria for adequate methodological quality of a study on measurement properties - COSMIN checklist with 4-point scale

The COSMIN checklist was developed to rate the methodological quality of a study on one of more measurement properties. The COSMIN checklist consists of 12 boxes, containing 4-18 items per box (119 items in total). The methodological quality of a study is considered adequate if all items in a box are considered adequate.

It is, however, often the case that not all items in a box are scored adequate. In the COSMIN Delphi study, it was not discussed how the methodological quality of a study should be rated if not all items are scored adequate. Later, however, a scoring system for the COSMIN checklist was developed to calculate quality scores per measurement property when using the checklist in systematic reviews of measurement properties [61].

Part of the COSMIN group developed a 4-point rating scale to classify each assessment of a measurement property as excellent, good, fair, or poor, based on the scores of the items in the corresponding COSMIN box. We developed this rating scale based on discussions in our Clinimetrics working group (www.clinimetrics.nl), as well as on the application of this scale to rate the quality of all studies on measurement properties described in 46 articles on neck disability questionnaires. Based on this application, the scale was further discussed and refined.

In general, a rating will be assigned as follows: We argued that meeting all COSMIN standards represents the ideal situation, which might be considered more than 'good'. Therefore, in general, an item is scored as excellent when there is evidence that the methodological quality aspect of the study to which the item is referring is adequate (this equals the original response option ''yes''). For example, if evidence is provided (e.g., from a global rating scale) that patients remained stable between the test and retest (item 7, box B), this item is scored as excellent. An item is scored as good when relevant information is not reported in an article, but it can be assumed that the quality aspect is adequate. For example, if it can be assumed that patients were stable between the test and retest (e.g., based on the clinical characteristics of the patients and the time interval between the test and retest), the item is scored as good. An item is scored as fair if it is doubtful whether the methodological quality aspect is adequate. For example, when it is unclear whether the patients were stable in a reliability study, the item is scored as fair. Finally, an item is scored as poor when evidence is provided that the methodological quality aspect is not adequate, for example, if patients were treated between the test and retest.
For each item in the COSMIN checklist, specific criteria were developed for 'excellent', 'good', 'fair', and 'poor' quality. Each item is rated individually on the 4-point rating scale. Subsequently, an overall score for the assessment of a given measurement property is obtained by taking the lowest score for any of the items in the box ('worst score counts` method). Thus if one item in a box is scored as 'poor', the overall score for the study on that measurement property will be 'poor'.
The specific criteria for each item can be found in Terwee et al. 2011 [61]. We recommend to use the COSMIN checklist with 4-point rating score in systematic reviews of measurement properties (to download from www.cosmin.nl).

4. How to cite the COSMIN checklist

This manual was based on four articles, published in peer-reviewed scientific journals. If you use the COSMIN checklist, please refer to one of the COSMIN articles instead of this manual:

The protocol of the Delphi study is described in:

> Mokkink LB, Terwee CB, Knol DL, Stratford PW, Alonso J, Patrick DL, Bouter LM, de Vet HCW. Protocol of the COSMIN study: COnsensus-based Standards for the selection of health Measurement INstruments. BMC Med Res Methodol 2006;6:2.

In this paper we present the COSMIN checklist and describe the agreement of the panel concerning the items included in the checklist.:

> Mokkink LB, Terwee CB, Patrick DL, Alonso J, Stratford PW, Knol DL, Bouter LM, de Vet HCW. The COSMIN checklist for assessing the methodological quality of studies on measurement properties of health status measurement instruments: an international Delphi study. Quality of Life Research 2010;19:539-549.

In this article we explain our choices for some of the included design requirements and preferred statistical methods, for which no evidence is available in the literature or about which we have had major discussion among the members of the Delphi panel:

> Mokkink LB, Terwee CB, Knol DL, Stratford PW, Alonso J, Patrick DL, Bouter LM, de Vet HCW. The COSMIN checklist for evaluating the methodological quality of studies on measurement properties: A clarification of its content. BMC Medical Research Methodology 2010;10:22.

In this article we described the COSMIN taxonomy and definitions:

> Mokkink LB, Terwee CB, Patrick DL, Alonso J, Stratford PW, Knol DL, Bouter LM, de Vet HCW. International consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes: results of the COSMIN study. Journal of Clinical Epidemiology 2010;63:737-745.

The results of the inter-rater reliability of the COSMIN checklist are described in:

> Mokkink LB, Terwee CB, Gibbons E, Stratford PW, Alonso J, Patrick DL, Knol DL, Bouter LM, de Vet HCW. Inter-rater reliability of the COSMIN (COnsensus-based Standards for the selection of health status Measurement Instruments) checklist. BMC Research Methodology 2010;10:82.

In this article the development of the COSMIN checklist with 4-point rating scale is described:

> Terwee CB, Mokkink LB, Knol DL, Ostelo RWJG, Bouter LM, de Vet HCW. Rating the methodological quality in systematic reviews of studies on measurement properties: a scoring system for the COSMIN checklist. Qual Life Res 2011; Jul 6. [Epub ahead of print]

5. Publications

Development and validation of the COSMIN checklist

Mokkink LB, Terwee CB, Knol DL, Stratford PW, Alonso J, Patrick DL, Bouter LM, de Vet HCW. Protocol of the COSMIN study: COnsensus-based Standards for the selection of health Measurement INstruments. BMC Med Res Methodol 2006;6:2.

Mokkink LB, Terwee CB, Patrick DL, Alonso J, Stratford PW, Knol DL, Bouter LM, de Vet HCW. The COSMIN checklist for assessing the methodological quality of studies on measurement properties of health status measurement instruments: an international Delphi study. Quality of Life Research. 2010;19(4):539-49.

Mokkink LB, Terwee CB, Knol DL, Stratford PW, Alonso J, Patrick DL, Bouter LM, de Vet HCW. The COSMIN checklist for evaluating the methodological quality of studies on measurement properties: A clarification of its content. BMC Medical Research Methodology. 2010;10:22.

Mokkink LB, Terwee CB, Patrick DL, Alonso J, Stratford PW, Knol DL, Bouter LM, de Vet HCW. The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes. Journal of Clinical Epidemiology. 2010;63(7):737-45.

Mokkink LB, Terwee CB, Gibbons E, Stratford PW, Alonso J, Patrick DL, Knol DL, Bouter LM, de Vet HCW. Inter-rater agreement and reliability of the COSMIN (COnsensus-based Standards for the selection of health status Measurement Instruments) checklist. BMC Med Res Methodol. 2010 Sep 22;10:82.

Terwee CB, Mokkink LB, Knol DL, Ostelo RWJG, Bouter LM, de Vet HCW. Rating the methodological quality in systematic reviews of studies on measurement properties: a scoring system for the COSMIN checklist. Qual Life Res 2011; Jul 6. [Epub ahead of print]

Studies in which the COSMIN checklist was used can be found on the COSMIN website.

6. References

[1] De Vet HCW, Terwee CB, Mokkink LB, Knol DL. Measurements in Medicine. A practical guide. 1 ed. Cambridge: Cambridge University Press; 2011.

[2] Mokkink LB, Terwee CB, Stratford PW, Alonso J, Patrick DL, Riphagen I, et al. Evaluation of the methodological quality of systematic reviews of health status measurement instruments. Qual Life Res 2009;18:313-33.

[3] Mokkink LB, Terwee CB, Patrick DL, Alonso J, Stratford PW, Knol DL, et al. International consensus on taxonomy, terminology, and definitions of measurement properties: results of the COSMIN study. Journal of Clinical Epidemiology 2010;63:737-45.

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

[5] Lohr KN, Aaronson NK, Alonso J, Burnam MA, Patrick DL, Perrin EB, et al. Evaluating quality of life and health status instruments: Development of scientific review criteria. Clinical Therapeutics 1996;18:979-92.

[6] Bombardier C, Tugwell P. Methodological considerations in functional assessment. Journal of Rheumatology 1987;14:(suppl 15) 6-10.

[7] Valderas JM, Ferrer M, Mendivil J, Garin O, Rajmil L, Herdman M, et al. Development of EMPRO: a tool for the standardized assessment of patient-reported outcome measures. Value Health 2008;11:700-8.

[8] Mokkink LB, Terwee CB, Gibbons E, Stratford PW, Alonso J, Patrick DL, et al. Inter-rater agreement and reliability of the COSMIN (COnsensus-based Standards for the selection of health status Measurement Instruments) checklist. BMC Med Res Methodol 2010;10:82.

[9] Verhagen AP, de Vet HCW, de Bie RA, Kessels AGH, Boers M, Bouter LM, et al. The Delphi list: A criteria list for quality assessment of randomized clinical trials for conducting systematic reviews developed by Delphi consensus. Journal of Clinical Epidemiology 1998;51:1235-41.

[10] Whiting P, Rutjes AW, Reitsma JB, Bossuyt PM, Kleijnen J. The development of QUADAS: a tool for the quality assessment of studies of diagnostic accuracy included in systematic reviews. BMC Med Res Methodol 2003;3:25.

[11] Terwee CB, Bot SDM, de Boer MR, van der Windt DAWM, Knol DL, Dekker J, et al. Quality criteria were proposed for measurement properties of health status questionnaires. J Clin Epidemiol 2007;60:34-42.

[12] Schellingerhout JM, Verhagen AP, Heymans MW, Koes BW, de Vet HCW, Terwee CB. Measurement properties of disease-specific questionnaires in patients with neck pain: a systematic review. Qual Life Res 2011;July 7 {epub ahead of print].

[13] Kline P. Handbook of psychological testing. London: Routledge; 1993.

[14] Edelen MO, Reeve BB. Applying item response theory (IRT) modeling to questionnaire development, evaluation, and refinement. Qual Life Res 2007;16 Suppl 1:5-18.

[15] Embretson SE, Reise SP. Item response theory for psychologists. Mahwah, New Jersey: Lawrence Erlbaum Associates, Inc.; 2000.

[16] Giraudeau B, Mary JY. Planning a reproducibility study: How many subjects and how many replicates per subject for an expected width of the 95 per cent confidence interval of the intraclass correlation coefficient. Statistics in Medicine 2001;20:3205-14.

[17] Cortina JM. What is coefficient alpha? An examination of theory and applications. Journal of Applied Psychology 1993;78:98-104.

[18] CRONBACH LJ. Coefficient Alpha and the Internal Structure of Tests. Psychometrika 1951;16:297-334.

[19] Fayers PM, Hand DJ. Factor analysis, causal indicators and quality of life. Quality of Life Research 1997;6:139-50.

[20] Fayers PM, Hand DJ. Causal variables, indicator variables and measurement scales: An example from quality of life. J R Statist Soc A 2002;165:233-61.

[21] Streiner DL. Being inconsistent about consistency: when coefficient alpha does and doesn't matter. J Pers Assess 2003;80:217-22.

[22] Wirth RJ, Edwards MC. Item factor analysis: Current approaches and future directions. Psychological Methods 2007;12:58-79.

[23] Streiner DL. Figuring out factors: The use and misuse of factor analysis. Can J Psychiatry 1994;39:135-40.

[24] Van Abswoude AAH, Van der Ark LA, Sijtsma K. A Comparative Study of Test Data Dimensionality Assessment Procedures Under Nonparametric IRT Models. Applied Psychology Measurement 2004;28:3-24.

[25] Streiner DL, Norman GR. Health measurement scales. A practical guide to their development and use. 4 ed. New York: Oxford University Press; 2008.

[26] Shrout PE, Fleiss JL. Intraclass Correlations: Uses in assessing rater reliability. Psychological Bulletin 1979;86:420-8.

[27] Cohen J. Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. Psychological Bulletin 1968;70:213-20.

[28] Fleiss JL, Cohen J. The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. Educational and Psychological Measurement 1973;33:613-9.

[29] de Vet HCW, Terwee CB, Knol DL, Bouter LM. When to use agreement versus reliability measures. J Clin Epidemiol 2006;59:1033-9.

[30] van der Heijden GJMG, Leffers P, Bouter LM. Shoulder disability questionnaire design and responsiveness of a functional status measure. Journal of Clinical Epidemiology 2000;53:29-38.

[31] Levine DW, Simmons BP, Koris MJ, Daltroy LH, Hohl GG, Fossel AH, et al. A self-administered questionnaire for the assessment of severity of symptoms and functional status in carpal tunnel syndrom. J Bone Joint Surg Am 1993;75:1585-92.

[32] Wilson IB, Cleary PD. Linking clinical variables with health-related quality of life. JAMA 1995;273:59-65.

[33] World Health Organization. ICF: international classification of functioning, disability and health. Geneva : World **Health** Organization; 2001.

[34] Floyd FJ, Widaman KF. Factor analysis in the development and refinement of clinical assessment instruments. Psychological Assessment 1995;7:286-99.

[35] de Vet HCW, Ader HJ, Terwee CB, Pouwer F. Are factor analytical techniques appropriately used in the validation of health status questionnaires? A systematic review on the quality of factor analyses of the SF-36. Quality of Life Research 2005;14:1203-18.

[36] Strauss ME, Smith GT. Construct Validity: Advances in Theory and Methodology. Annu Rev Clin Psychol 2008.

[37] CRONBACH LJ, MEEHL PE. Construct validity in psychological tests. Psychol Bull 1955;52:281-302.

[38] McDowell I, Jenkinson C. Development standards for health measures. J Health Serv Res Policy 1996;1:238-46.

[39] Messick S. The standard problem. Meaning and values in measurement and evaluation. American Psychologist 1975;oct:955-66.

[40] Altman DG. Practical statictics for medical research. London: Chapman and Hall; 1991.

[41] Bullinger M AJAGLASMW-DSGBWAANBPFSKSWJftIPG. Translating health status questionnaires and evaluating their quality: The IQOLA project approach. Journal of Clinical Epidemiology 1998;51:913-23.

[42] Acqaudro C, Conway K, Wolf B, Hareendran A, Mear I, Anfray C, et al. Development of a standardized classification system for the translation of Patient-Reported Outcome (PRO) measures. PRO newsletter 2008;39:5-7.

[43] Koller M, Aaronson NK, Blazeby J, Bottomley A, Dewolf L, Fayers P, et al. Translation procedures for standardised quality of life questionnaires: The

European Organisation for Research and Treatment of Cancer (EORTC) approach. Eur J Cancer 2007;43:1810-20.

[44] Guillemin F, Bombardier C, Beaton D. Cross-cultural adaptation of health-related quality of life measures: Literature review and proposed guidelines. Journal of Clinical Epidemiology 1993;46:1417-32.

[45] Teresi JA, Ocepek-Welikson K, Kleinman M, Cook KF, Crane PK, Gibbons LE, et al. Evaluating measurement equivalence using the item response theory log-likelihood ratio (IRTLR) method to assess differential item functioning (DIF): applications (with illustrations) to measures of physical functioning ability and general distress. Qual Life Res 2007;16 Suppl 1:43-68.

[46] Petersen MA, Groenvold M, Bjorner JB, Aaronson NK, Conroy T, Cull A, et al. Use of differential item functioning analysis to assess the equivalence of translations of a questionnaire. Qual Life Res 2003;12:373-85.

[47] CRONBACH LJ, MEEHL PE. Construct validity in psychological tests. Psychol Bull 1955;52:281-302.

[48] Norman GR, Stratford PW, Regehr G. Methodological problems in the retrospective computation of responsiveness to change: the lesson of Cronbach. J Clin Epidemiol 1997;50:869-79.

[49] Guyatt GH, Norman GR, Juniper EF, Griffith LE. A critical look at transition ratings. J Clin Epidemiol 2002;55:900-8.

[50] Altman DG. Practical statistics for medical research. London: Chapman & Hall/CRC; 1991.

[51] Norman GR, Stratford P, Regehr G. Methodological problems in the retrospective computation of responsiveness to change: The lesson of Cronbach. Journal of Clinical Epidemiology 1997;50:869-79.

[52] Cohen J. Statistical power analysis for the behavioural sciences. 2nd ed. Hillsdale, NJ: Lawrence Erlbaum Associates; 1988.

[53] McHorney CA, Tarlov AR. Individual-patient monitoring in clinical practice: are available health status surveys adequate? Qual Life Res 1995;4:293-307.

[54] Norman GR. Issues in the use of change scores in randomized trials. J Clin Epidemiol 1989;42:1097-105.

[55] Stockler MR, Osoba D, Goodwin P, Corey P, Tannock IF. Responsiveness to change in health-related quality of life in a randomized clinical trial: a comparison of the Prostate Cancer Specific Quality of Life Instrument (PROSQOLI) with analogous scales from the EORTC QLQ-C30 and a trial specific module. European Organization for Research and Treatment of Cancer. J Clin Epidemiol 1998;51:137-45.

[56] Streiner DL, Norman GR. Health measurement scales. A practical guide to their development and use. 4th ed ed. Oxford: University Press; 2008.

[57] Terwee CB, Dekker FW, Wiersinga WM, Prummel MF, Bossuyt PM. On assessing responsiveness of health-related quality of life instruments: guidelines for instrument evaluation. Qual Life Res 2003;12:349-62.

[58] Guyatt GH, Walter S, Norman GR. Measuring change over time: assessing the usefulness of evaluative instruments. J Chronic Dis 1987;40:171-8.

[59] De Vet H, Beckerman H, Terwee CB, Terluin B, Bouter LM, for the clinimetrics working group. Definition of clinical differences. Letter to the Editor. Journal of rheumatology 2006;33:434.

[60] Fayers PM, Curran D, Machin D. Incomplete quality of life data in randomized trials: missing items. Stat Med 1998;17:679-96.

[61] Terwee CB, Mokkink LB, Knol DL, Ostelo RW, Bouter LM, de Vet HC. Rating the methodological quality in systematic reviews of studies on measurement properties: a scoring system for the COSMIN checklist. Qual Life Res 2011; July 7 [epub ahead of print].

Appendix 1. The COSMIN panel members

Listed are the panel members who have participated in at least one round of the COSMIN study:

Neil Aaronson, Linda Abetz, Elena Andresen, Dorcas Beaton, Martijn Berger, Giorgio Bertolotti, Monika Bullinger, David Cella, Joost Dekker, Dominique Dubois, Arne Evers, Diane Fairclough, David Feeny, Raymond Fitzpatrick, Andrew Garratt, Francis Guillemin, Dennis Hart, Graeme Hawthorne, Ron Hays, Elizabeth Juniper, Robert Kane, Donna Lamping, Marissa Lassere, Matthew Liang, Kathleen Lohr, Patrick Marquis, Chris McCarthy, Elaine McColl, Ian McDowell, Don Mellenbergh, Mauro Niero, Geoffrey Norman, Manoj Pandey, Luis Rajmil, Bryce Reeve, Dennis Revicki, Margaret Rothman, Mirjam Sprangers, David Streiner, Gerold Stucki, Giulio Vidotto, Sharon Wood-Dauphinee, Albert Wu.