Qual Life Res (2010) 19:539–549
DOI 10.1007/s11136-010-9606-8

The COSMIN checklist for assessing the methodological quality of studies on measurement properties of health status measurement instruments: an international Delphi study

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

Accepted: 2 February 2010 / Published online: 19 February 2010

Abstract
Background Aim of the COSMIN study (COnsensus-based Standards for the selection of health status Measurement INstruments) was to develop a consensus-based checklist to evaluate the methodological quality of studies on measurement properties. We present the COSMIN checklist and the agreement of the panel on the items of the checklist.
Methods A four-round Delphi study was performed with international experts (psychologists, epidemiologists, statisticians and clinicians). Of the 91 invited experts, 57 agreed to participate (63%). Panel members were asked to rate their (dis)agreement with each proposal on a five-point scale. Consensus was considered to be reached when at least 67% of the panel members indicated 'agree' or 'strongly agree'.
Results Consensus was reached on the inclusion of the following measurement properties: internal consistency, reliability, measurement error, content validity (including face validity), construct validity (including structural validity, hypotheses testing and cross-cultural validity), criterion validity, responsiveness, and interpretability. The latter was not considered a measurement property. The panel also reached consensus on how these properties should be assessed.
Conclusions The resulting COSMIN checklist could be useful when selecting a measurement instrument, peer-reviewing a manuscript, designing or reporting a study on measurement properties, or for educational purposes.

L. B. Mokkink (✉) · C. B. Terwee · D. L. Knol · L. M. Bouter · H. C. W. de Vet
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
e-mail: w.mokkink@vumc.nl
URL: www.emgo.nl; www.cosmin.nl

C. B. Terwee
e-mail: cb.terwee@vumc.nl

D. L. Knol
e-mail: d.knol@vumc.nl

L. M. Bouter
e-mail: lm.bouter@dienst.vu.nl

H. C. W. de Vet
e-mail: hcw.devet@vumc.nl

D. L. Patrick
Department of Health Services, University of Washington, Thur Canal St Research Office, 146N Canal Suite 310, Seattle, WA 98103, USA
e-mail: donald@u.washington.edu

J. Alonso
Health Services Research Unit, Institut Municipal d'Investigacio Medica (IMIM-Hospital del Mar), Doctor Aiguader 88, 08003 Barcelona, Spain
e-mail: jalonso@imim.es

J. Alonso
CIBER en Epidemiología y Salud Pública (CIBERESP), Barcelona, Spain

P. W. Stratford
School of Rehabilitation Science and Department of Clinical Epidemiology and Biostatistics, McMaster University, 1400 Main St. West, Hamilton, ON, Canada
e-mail: stratford@mcmaster.ca

L. M. Bouter
Executive Board of VU University Amsterdam, De Boelelaan 1105, 1081 HV Amsterdam, The Netherlands

Keywords Delphi technique · Outcome assessment · Psychometrics · Quality of life · Questionnaire

Introduction

Measurement of health outcomes is essential in scientific research and in clinical practice. Based on the scores obtained with measurement instruments, decisions are made about the application of subsequent diagnostic tests and treatments. Health status measurement instruments should therefore be reliable and valid. Otherwise there is a serious risk of imprecise or biased results that might lead to wrong conclusions. Organisations such as the US Food and Drug Administration (FDA) and the European Medicines Agency (EMEA) require that measurement instruments must be well validated for its purpose [1, 2]. The need for reliable and valid measurement instruments of health outcomes was clearly demonstrated by Marshall, who showed in schizophrenia trials that authors were more likely to report that treatment was superior to control when an unpublished measurement instrument was used in the comparison, rather than a published instrument [3].

Before a health status measurement instrument can be used in research or clinical practice, its measurement properties, i.e. reliability, validity and responsiveness, should be assessed and considered adequate. Studies evaluating measurement properties should be of high methodological quality to guarantee appropriate conclusions about the measurement properties of an instrument. To evaluate the methodological quality of a study on measurement properties, standards are needed. Although many standards and criteria have been proposed, these have not been operationalised into user-friendly and easily applicable checklists (e.g. [4, 5]). Moreover, these standards do not pay attention to studies that apply Item Response Theory (IRT) models, or are not consensus based (e.g. [6, 7]). Such a checklist should contain a complete set of standards (which refers to the design requirements and preferred statistical methods) and criteria of adequacy for what constitutes good measurement properties. Broad consensus is necessary in order to achieve wide acceptance of a checklist.

Research on measurement properties is particularly important for health outcomes that are directly reported by patients, i.e. health-related patient-reported outcomes (HR-PROs). A HR-PRO is a measurement of any aspect of a patient's health status that is directly assessed by the patient, i.e. without the interpretation of the patient's responses by a physician or anyone else [2]. Modes of data-collection for HR-PRO instruments include interviewer-administered instruments, self-administered instruments, or computer-administered instrument [2]. Examples of HR-PROs are questionnaires assessing symptoms, functional status, and health-related quality of life. These are constructs which are not directly measurable. Because of the subjective nature of these constructs, it is very important to evaluate whether the measurement instruments measure these constructs in a valid and reliable way.

The COSMIN initiative (COnsensus-based Standards for the selection of health Measurement Instruments) aims to improve the selection of health measurement instruments. As part of this initiative, the aim of this study was to develop a checklist containing standards for evaluating the methodological quality of studies on measurement properties. The checklist was developed as a multidisciplinary, international collaboration with all relevant expertise involved. We performed a Delphi study to address two research questions:

1. Which measurement properties should be included in the checklist?
2. How should these measurement properties be evaluated in terms of study design and statistical analysis (i.e. standards)?

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

Methods

Focus of the COSMIN checklist

The COSMIN checklist is focused on evaluating the methodological quality of studies on measurement properties of HR-PROs. We choose to focus on HR-PROs, because of the complexity of these instruments. These instruments measure constructs that are both multidimensional and not directly measurable.

In addition, we focused on evaluative applications of HR-PRO instruments, i.e. longitudinal applications assessing treatment effects or changes in health over time. The specification of *evaluative* is necessary, because the requirements for measurement properties vary with the application of the instrument [8]. For example, instruments used for evaluation need to be responsive, while instruments used for discrimination do not.

The COSMIN Steering Committee (Appendix 1) searched the literature to determine how measurement properties are generally evaluated. Two searches were performed: (1) a systematic literature search was performed to identify all existing systematic reviews on measurement properties of health status measurement instruments [9]. From these reviews, information was extracted on which measurement properties were evaluated, and on standards that were used to evaluate the

measurement properties of the included studies. For each measurement property, we found several different standards, some of which were contradictory [9]. (2) The steering committee also performed another systematic literature search (available on request from the authors) to identify methodological articles and textbooks containing standards for the evaluation of measurement properties of health status measurement instruments. Articles were selected if the purpose of the article was to present a checklist or standards for measurement properties. Standards identified in the aforementioned literature were used as input in the Delphi rounds.

International Delphi study

Subsequently, a Delphi study was performed, which consisted of four written rounds. The first questionnaire was sent in March 2006, the last questionnaire in November 2007. We decided to invite at least 80 international experts to participate in our Delphi panel in order to ensure 30 responders in the last round. Based on previous experiences with Delphi studies [10, 11], we expected that 70% of the people invited would agree to participate, and of these people 65% would complete the first list. Once started, we expected that 75% would stay involved. We included experts in the field of psychology, epidemiology, statistics, and clinical medicine. Among those invited were authors of reviews, methodological articles, or textbooks. Experts had to have at least five publications on the (methods of) measurement of health status in PubMed. We invited people from different parts of the world.

In the first round, we asked questions about which measurement properties should be included in the checklist, and about their terms and definitions. For example, we asked for the measurement property internal consistency 'which term do you consider the best for this measurement property?', with the response options 'internal consistency', 'internal consistency reliability', 'homogeneity', 'internal scale consistency', 'split-half reliability', 'internal reliability', 'structural reliability', 'item consistency', 'intra-item reliability', or 'other' with some space to give an alternative term. Regarding the definitions, we asked 'Which definition do you consider the best for internal consistency?', and provided seven definitions that were found in the literature and the option 'other' where a panel member could provide an alternative definition. In round two, we introduced questions about preferred standards for each measurement property. We asked questions about design issues, i.e. 'Do you agree with the following requirements for the design of a study evaluating internal consistency of HR-PRO instruments in an evaluative application? (1) One administration should be available. (2) A check for uni-dimensionality per (sub) scale should be performed. (3) Internal consistency statistics should be calculated for each (sub) scale separately'. The panel could answer each item on a 5-point scale ranging from strongly disagree to strongly agree. Next, the panel was asked to rate which statistical methods they considered adequate for evaluating the measurement property concerned. A list of potential relevant statistical methods for each measurement property was provided. For example, for internal consistency the following often used methods were proposed: 'Cronbach's alpha', 'Kuder-Richardson formula-20', 'average item-total correlation', 'average inter-item correlation', 'split-half analysis', 'goodness of fit (IRT) at a global level, i.e. index of (subject) separation', 'goodness of fit (IRT) at a local level, i.e. specific item tests', or 'other'. Panel members could indicate more than one method. In the third round, we presented the most often chosen method, both the one based on CTT and the one based on IRT, and asked if the panel considered this method as the most preferred method to evaluate the measurement property. For internal consistency, these were 'Cronbach's alpha' and 'goodness of fit (IRT) at a global level, i.e. index of (subject) separation', respectively. In the third round, the panel members were asked whether the other methods (i.e. 'Kuder-Richardson formula-20', 'average item-total correlation', 'average inter-item correlation', 'split-half analysis', 'goodness of fit (IRT) at a local level, i.e. specific item tests') were also considered appropriate. Panel members could also have indicated 'other methods' in round 2. Indicated methods were 'eigen-values or percentage of variance explained of factor analysis,' 'Mokken Rho' or 'Loevinger H' for internal consistency. In round 3, the panel was also asked whether they considered these methods as appropriate for assessing internal consistency. In the final Delphi round, all measurement properties and standards that the panel agreed upon were integrated by the steering committee into a preliminary version of the checklist for evaluating the methodological quality of studies on measurement properties.

In each Delphi round, the results of the previous round were presented in a feedback report. Panel members were asked to rate their (dis)agreement with regard to proposals. Agreement was rated on a 5-point scale (strongly disagree—disagree—no opinion—agree—strongly agree). The panel members were encouraged to give arguments for their choices to convince other panel members, to suggest alternatives, or to add new issues. Consensus on an issue was considered to be reached when at least 67% of the panel members indicated 'agree' or 'strongly agree' on the 5-point scale. If less than 67% agreement was reached on a question, we asked it again in the next round, providing pro and contra arguments given by the panel members, or we proposed an alternative. When no consensus was reached, the Steering Committee took the final decision.

When necessary, we asked the panel members to indicate the preferred statistical methods separately for each measurement theory, i.e. Classical Test Theory (CTT) or Item Response Theory (IRT), or for each type of score, such as dichotomous, nominal, ordinal, or continuous scores.

Results

Panel members

We invited 91 experts to participate of whom 57 (63%) agreed to participate. The main reason for non-participation was lack of time. Nineteen experts (21%) did not respond. Of the 57 experts who agreed to participate, 43 (75%) experts participated in at least one round, and 20 (35%) participated in all four rounds. The average number (minimum–maximum) of years of experience in measuring health or comparable fields (e.g. in educational or psychological measurements) was 20 (6–40) years. Most of the panel members came from Northern America ($n = 25$) and Europe ($n = 29$), while two were from Australia and one was from Asia. The response rate of the rounds ranged from 48 to 74%. Six panel members (11%) dropped out during the process. The names of all panel members who completed at least one round are presented in the "Acknowledgements".

The COSMIN taxonomy

In the Delphi study, we also developed a taxonomy of the relationships of measurement properties that are relevant for evaluating HR-PRO instruments, and reached consensus on terminology and definitions of these measurement properties. The relationships between all properties are presented in a taxonomy (Fig. 1). The taxonomy comprises three domains (i.e. reliability, validity, and responsiveness), which contain the measurement properties. The measurement property construct validity contains three aspects, i.e. structural validity, hypotheses testing, and cross-cultural validity. Interpretability was also included in the taxonomy and checklist, although it was not considered a measurement property, but nevertheless an important characteristic. The percentages agreement on terminology and position in the taxonomy are described elsewhere [12].

Fig. 1 COSMIN taxonomy of relationships of measurement properties

The COSMIN checklist

The results of the consensus reached in the Delphi rounds were used to construct the COSMIN checklist (Appendix 2). The checklist contains twelve boxes. Ten boxes can be used to assess whether a study meets the standard 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 (box D), structural validity (box E), hypotheses testing (box F), cross-cultural validity (box G), criterion validity (box H) and responsiveness (box I), and one box contains standards for studies on interpretability (box J). In addition, two boxes are included in the checklist that contain general requirements for articles in which IRT methods are applied (IRT box), and general requirements for the generalizability of the results (Generalizability box), respectively.

To complete the COSMIN checklist, a 4-step procedure should be followed (Fig. 2) [13]. Step 1 is to determine which properties are evaluated in an article. Step 2 is to determine if the statistical methods used in the article are based on Classical Test Theory (CTT) or on Item Response Theory (IRT). For studies that apply IRT, the IRT box should be completed. Step 3 is to complete the boxes with standards accompanying the properties chosen in step 1. These boxes contain questions to rate whether a study meets the standards for good methodological quality. Items are included about design requirements and preferred statistical methods of each of the measurement properties (boxes A to I). In addition, a box with items on interpretability of the (change) score is included (box J). The number of items in these boxes range from 5 to 18. Step 4 of the procedure is to complete the box on general requirements for the generalizability of the results. This Generalizability box should be completed for each property identified in step 1. We developed a manual describing the rationale of each item, and suggestions for scoring [13].

Consensus among the panel

In Table 1, we present ranges of percentage agreement of the panel members for each box, both for the design requirements and the statistical methods. Most of these issues were discussed in rounds 2 and 3.

Percentage agreement among the panel members on the items 1–3 in the IRT box ranged from 81 to 96%. Item 4 (i.e. checking the assumptions for estimating parameters of the IRT model) was included based on a suggestion of a panel member in round 4. Therefore, no consensus was rated, and the Steering Committee decided on including this item.

Four items included in the checklist had less than 67% agreement of the panel: item 9 of box A internal consistency, item 11 for box C measurement error, and items 11 and 17 of box I responsiveness. All but one was about the statistical methods. For different reasons, which we will successively explain, the Steering Committee decided to include these four items in the checklist.

When asking about the preferred statistical method for internal consistency, we initially did not distinguish between types of scores, i.e. dichotomous or ordinal scores (item 9). Therefore, Cronbach alpha was preferred over Kuder-Richardson Formula 20 (KR-20). However, the Steering Committee decided afterward that KR-20 was considered appropriate for dichotomous scores as well.

Item 11 of box C measurement error contains three methods, i.e. standard error of measurement (SEM), smallest detectable change (SDC) and Limits of Agreement (LOA). In round 3, SEM was chosen as the preferred method for measuring measurement error (76% agreement). When asking about other appropriate methods, only 20% agreed with SDC, and 28% with LOA. Despite the low percentages agreement reached in round 3 on accepting SDC and LOA as appropriate methods, the Steering Committee decided afterward that both methods should be considered appropriate to measure measurement error and were included in the checklist. The SDC is a linear transformation of the SEM [14], i.e., $1.96 \times \sqrt{2} \times \text{SEM}$. Because the SEM is an appropriate method, SDC should also be considered appropriate. The LOA is a parameter indicating how much two measures differ [15]. When these two measures are repeated measures in stable patients, it can be used as a method for assessing measurement error. LOA is directly related to SEM [16], and we therefore decided to include this method in the checklist.

Item 11 of box I responsiveness (i.e. 'was an adequate description provided of the comparator instrument(s)') was approved by 64% of the panel. Although the percentage agreement was slightly too low, we decided to include this item because it was also included in box F hypotheses testing, reflecting the similarity between construct validity and responsiveness.

Item 17 of box I contains two methods, i.e. correlations between change scores and the area under the receiver operator curve (ROC). Seventy-six percent of the panel considered the first method as the preferred method. This method can be used when both the measurement instrument under study and its gold standard are continuous measures. Only 60% considered the ROC method as an appropriate method to measure responsiveness when a (dichotomous) gold standard is available. In analogy to

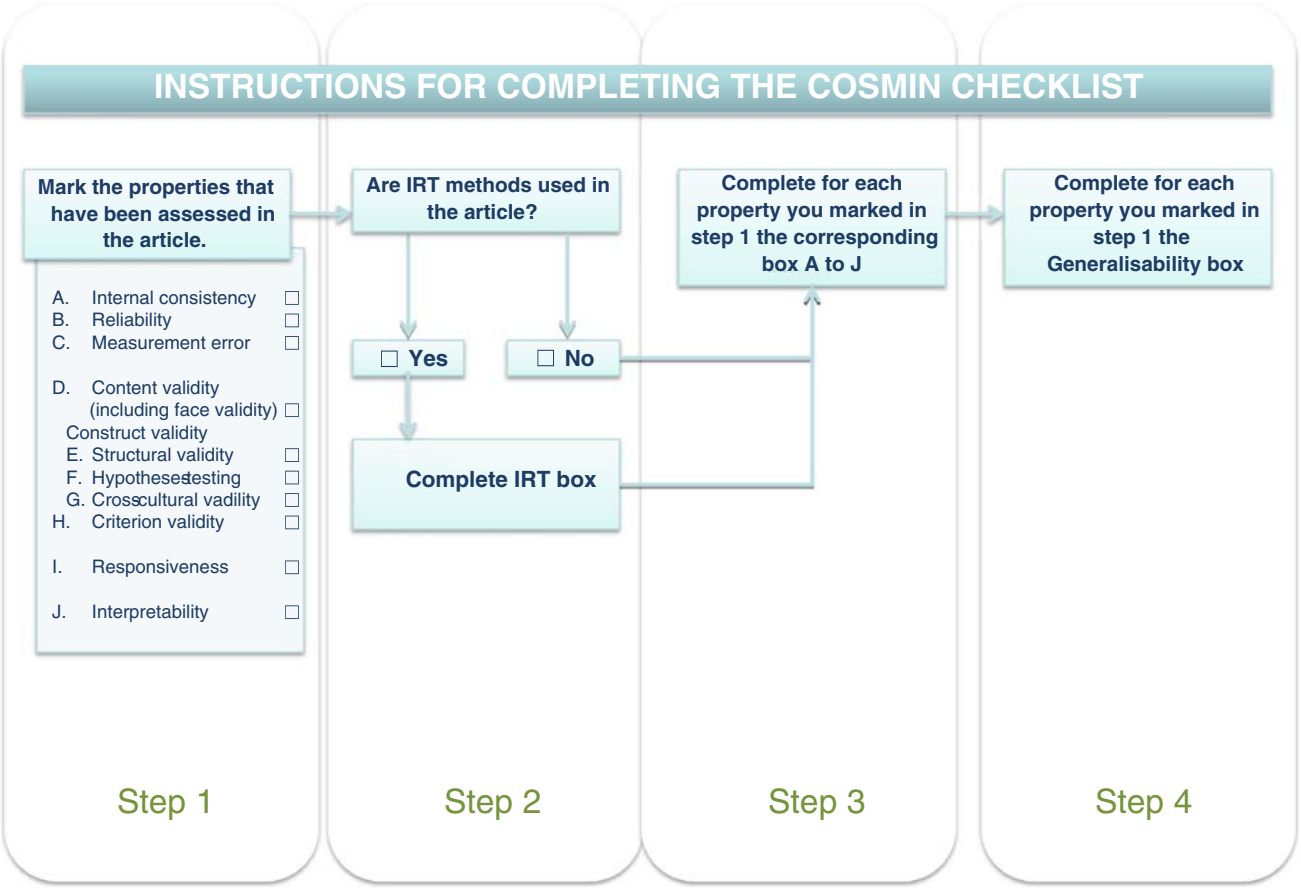

Fig. 2 The 4-step procedure to complete the COSMIN checklist

Table 1 Percentage agreement of panel members who (strongly) agreed with the items about design requirements and statistical methods for the COSMIN boxes A–J

	Design requirements (%)	Statistical methods (%)
Internal consistency	77–92 (R2)	40–88 (R2–4)
Reliability	77–97 (R2)	80–92 (R3)
Measurement error	Same items as for reliability	20–76 (R3)
Content validity	90–94 (R2)	na
Structural validity	72 (R3)	68–100 (R3)
Hypotheses testing	77–92 (R2, R4)	90 (R2)
Cross-cultural validity	70–79 (R3–4)	68–94 (R3)
Criterion validity	88 (R3)	88 (R3)
Responsiveness (general)	90–97 (R2)	na
Responsiveness (no gold standard available)	64–68 (R3)	88 (R3)
Responsiveness (gold standard available)	80 (R3)	60–76 (R3)
Interpretability	na	72–96 (R3)

R round in which consensus was reached, *na* not applicable

diagnostic research, the Steering Committee considered the ROC method an appropriate method to evaluate if a measurement instrument is as good as its gold standard. The Steering Committee therefore decided to include this method.

Discussion

In this Delphi study, we developed a checklist containing standards for evaluating the methodological quality of studies on measurement properties. We consider it useful to

separate the evaluation of the methodological quality of a study and the evaluation of its results, similar as is done for trials. The COSMIN checklist is meant for evaluating the methodological quality of a study on the measurement properties of a HR-PRO instrument, not for evaluating the quality of the HR-PRO instrument itself. To assess the quality of the instrument, criteria for what constitutes good measurement properties should be applied to the results of a study on measurement properties. Examples of such criteria were previously published by members of our group [6]. However, these criteria are not consensus based. Note that the COSMIN checklist does not include these criteria of adequacy.

Although we initially intended to develop these criteria [17], due to lack of time, and complexity of the issues, we have not developed criteria of adequacy of measurement properties yet. Consensus on such criteria should be obtained in the future. In addition, it might be useful to develop a rating system by which a study can be classified into different quality levels, e.g. excellent/good/fair/poor methodological quality.

The COSMIN checklist can be used to evaluate the methodological quality of studies on measurement properties of health status measurement instruments. For example, it can be used to assess the quality of a study on one measurement instrument or to compare the measurement properties of a number of measurement instruments in a systematic review (e.g. [18, 19]). In such a review, it is important to take the methodological quality of the selected studies into account. If the results of high quality studies differ from the results of low-quality studies, this can be an indication of bias. The COSMIN checklist can also be used as guidance for designing or reporting a study on measurement properties. Furthermore, students can use it when learning about measurement properties, and reviewers or editors of journals can use it to appraise the methodological quality of articles or grant applications of studies on measurement properties.

There are theoretical arguments that there is a need for an instrument to demonstrate good reliability, validity, and responsiveness. To our knowledge, Marshall [3] is the only one who empirically showed that the results of studies can differ when validated measurement instruments are used compared to studies in which non-validated instruments are used. However, more empirical research should be conducted to support the need. Studies could be conducted for this purpose, for example, in which the results of randomized controlled trials (RCTs) that uses well-responsive measurement instruments and RCTs that uses instruments with unknown responsiveness, are compared.

A Delphi approach is useful for situations in which there is a lack of empirical evidence, and there are strong differences of opinion. The answers of the research questions of the COSMIN study cannot empirically be investigated. Therefore, agreement among experts is useful. In the literature, cut-offs between 55 and 100% are used [20]. The cut-off of 67% for consensus was arbitrarily chosen.

It is impossible to draw a random sample from all experts. Therefore, the selection of experts was necessarily non-systematic. All first and last authors identified by any of the two systematic literature searches described in the method section were considered as potential experts. We added people who we considered experts and who were not yet on the list. As a check of being an expert, we searched PubMed to see whether an author had published at least five articles on measurement issues. We considered a total of 30 experts sufficient to have a spread over the variety of opinion, and not too large to keep it manageable.

In this study, we focused on HR-PRO instruments. However, the same measurement properties are likely to be relevant for other kind of health-related measurement instruments, such as performance-based instruments and clinical rating scales. Furthermore, we focused on evaluative instruments. However, for discriminative or predictive purposes, the design requirements and standards for the measurement properties are likely the same.

The COSMIN checklist gives general recommendations of HR-PRO measurements. Some of the standards in the COSMIN checklist need further refinement, e.g. by defining what an adequate sample size is or an adequate test–retest time interval or when something is adequately described. Since these issues are highly dependent on the construct to be measured, users should make these decisions for their own application.

To help future users of the COSMIN checklist, we described some of the discussions we have had in the Delphi rounds about the standards elsewhere [21]. In the manual [13], we described a rationale for each item and suggestions for scoring the items in the checklist.

The COSMIN initiative aims to improve the selection of measurement instruments. As a first step, we have reached consensus on which measurement properties are important and we have developed standards for how to evaluate these measurement properties. The COSMIN checklist was developed with the participation of many experts in the field. The COSMIN checklist will facilitate the selection of the most appropriate HR-PRO measure among competing instruments. By involvement of many experts in the development process of the COSMIN checklist, it is highly probable that all relevant items of all relevant measurement properties are included, contributing to its content validity. In addition, we are planning to evaluate the inter-rater reliability of the COSMIN checklist in a large international group of researchers.

Acknowledgments We are grateful to all the panel members who have participated in 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. And an additional thanks to Sharon Wood-Dauphinee for language corrections within the COSMIN checklist. This 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. These funding organizations did not play any role in the study design, data collection, data analysis, data interpretation, or publication.

Appendix 1: Members of the COSMIN Steering Committee

Wieneke Mokkink (epidemiologist): Department of Epidemiology and Biostatistics, EMGO Institute for Health and Care Research, VU University Medical Center, Amsterdam, The Netherlands.
Caroline Terwee (epidemiologist): Department of Epidemiology and Biostatistics, EMGO Institute for Health and Care Research, VU University Medical Center, Amsterdam, The Netherlands.
Donald Patrick (epidemiologist): Department of Health Services, University of Washington, Seattle, USA;
Jordi Alonso (clinician): Health Services Research Unit, Institut Municipal d'Investigacio Medica (IMIM-Hospital del Mar), Barcelona, Spain; CIBER en Epidemiología y Salud Pública (CIBERESP), Barcelona, Spain.
Paul Stratford (physiotherapist): School of Rehabilitation Science and Department of Clinical Epidemiology and Biostatistics, McMaster University, Hamilton, Canada.
Dirk Knol (psychometrician): Department of Epidemiology and Biostatistics, EMGO Institute for Health and Care Research, VU University Medical Center, Amsterdam, The Netherlands.
Lex Bouter (epidemiologist): Department of Epidemiology and Biostatistics, EMGO Institute for Health and Care Research, VU University Medical Center, Executive Board of VU University Amsterdam, Amsterdam, The Netherlands.
Riekie de Vet (epidemiologist): Department of Epidemiology and Biostatistics, EMGO Institute for Health and Care Research, VU University Medical Center, Amsterdam, The Netherlands.

Appendix 2: The COSMIN checklist

Step 1. Evaluated measurement properties in the article

	Internal consistency
	Reliability
	Measurement error
	Content validity
	Structural validity
	Hypotheses testing
	Cross-cultural validity
	Criterion validity
	Responsiveness
	Interpretability

Step 2. Determining if the statistical method used in the article are based on CTT or 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))	☐	☐	☐

Step 3. Determining if a study meets the standards for good methodological quality

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)	☐	☐	☐

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	☐	☐	☐	

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?	☐	☐	

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?	☐	☐	

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?	☐	☐	☐

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?	☐	☐	☐

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?	☐	☐	☐

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?	☐	☐	☐

Box I. Responsiveness

		yes	no	?
Design requirements				
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?	☐	☐	☐

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?	☐	☐	

Step 4: Determining the Generalisability of the results

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?	☐	☐	☐

References

1. Committee for Medicinal Products for Human Use (CHMP). (2005). *Reflection paper on the regulatory guidance for the use of health related quality of life (HRQL) measures in the evaluation of medicinal products*, EMEA, London, 2005. Available at: www.emea.europa.eu/pdfs/human/ewp/13939104en.pdf. Accessed November 10, 2008.
2. US Department of Health and Human Services FDA Center for Drug Evaluation and Research, US Department of Health and Human Services FDA Center for Biologics Evaluation and Research, & US Department of Health and Human Services FDA Center for Devices and Radiological Health. (2006). Guidance for industry: patient-reported outcome measures: Use in medical product development to support labeling claims: Draft guidance. *Health and Quality of Life Outcomes, 4*, 79.
3. Marshall, M., Lockwood, A., Bradley, C., Adams, C., Joy, C., & Fenton, M. (2000). Unpublished rating scales: A major source of bias in randomised controlled trials of treatments for schizophrenia. *British Journal of Psychiatry, 176*, 249–252.
4. Lohr, K. N., Aaronson, N. K., Alonso, J., Burnam, M. A., Patrick, D. L., Perrin, E. B., et al. (1996). Evaluating quality-of-life and health status instruments: Development of scientific review criteria. *Clinical Therapeutics, 18*, 979–992.
5. Nunnally, J. C. (1978). *Psychometric theory*. New York, London: McGraw-Hill.
6. Terwee, C. B., Bot, S. D., De Boer, M. R., Van der Windt, D. A., Knol, D. L., Dekker, J., et al. (2007). Quality criteria were proposed for measurement properties of health status questionnaires. *Journal of Clinical Epidemiology, 60*, 34–42.
7. Valderas, J. M., Ferrer, M., Mendivil, J., Garin, O., Rajmil, L., Herdman, M., et al. (2008). Development of EMPRO: A tool for the standardized assessment of patient-reported outcome measures. *Value in Health, 11*, 700–708.
8. Kirshner, B., & Guyatt, G. H. (1985). A methodological framework for assessing health indexes. *Journal of Chronic Diseases, 38*, 27–36.

9. Mokkink, L. B., Terwee, C. B., Stratford, P. W., Alonso, J., Patrick, D. L., Riphagen, I., et al. (2009). Evaluation of the methodological quality of systematic reviews of health status measurement instruments. *Quality of Life Research, 18*, 313–333.
10. Evers, S., Goossens, M., De Vet, H., Van Tulder, M., & Ament, A. (2005). Criteria list for assessment of methodological quality of economic evaluations: Consensus on Health Economic Criteria. *International Journal of Technology Assessment in Health Care, 21*, 240–245.
11. Verhagen, A. P., De Vet, H. C. W., De Bie, R. A., Kessels, A. G., Boers, M., Bouter, L. M., et al. (1998). 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, 51*, 1235–1241.
12. Mokkink, L. B., Terwee, C. B., Patrick, D. L., Alonso, J., Stratford, P. W., Knol, D. L., Bouter, L.M., De Vet H. C. W. 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* (accepted for publication).
13. Mokkink, L. B., Terwee, C. B., Patrick, D. L., Alonso, J., Stratford, P. W., Knol, D. L. et al. (2009). *The COSMIN checklist manual*. http://www.cosmin.nl. Accessed September 2009.
14. Pfennings, L. E., Van der Ploeg, H. M., Cohen, L., & Polman, C. H. (1999). A comparison of responsiveness indices in multiple sclerosis patients. *Quality of Life Research, 8*, 481–489.
15. Bland, J. M., & Altman, D. G. (1986). Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet, 1*, 307–310.
16. De Vet, H. C. W., Terwee, C. B., Knol, D. L., & Bouter, L. M. (2006). When to use agreement versus reliability measures. *Journal Clinical Epidemiology, 59*, 1033–1039.
17. Mokkink, L. B., Terwee, C. B., Knol, D. L., Stratford, P. W., Alonso, J., Patrick, D. L., et al. (2006). Protocol of the COSMIN study: COnsensus-based Standards for the selection of health Measurement INstruments. *BioMed Central Medical Research Methodology, 6*, 2.
18. De Boer, M. R., Moll, A. C., De Vet, H. C. W., Terwee, C. B., Volker-Dieben, H. J., & Van Rens, G. H. (2004). Psychometric properties of vision-related quality of life questionnaires: A systematic review. *Ophthalmic and Physiological Optics, 24*, 257–273.
19. Veenhof, C., Bijlsma, J. W., Van den Ende, C. H., Van Dijk, G. M., Pisters, M. F., & Dekker, J. (2006). Psychometric evaluation of osteoarthritis questionnaires: A systematic review of the literature. *Arthritis and Rheumatism, 55*, 480–492.
20. Powell, C. (2003). The Delphi technique: Myths and realities. *Journal of Advanced Nursing, 41*, 376–382.
21. Mokkink, L. B., Terwee, C. B., Patrick, D. L., Alonso, J., Stratford, P. W., Knol, D. L., et al. (2010). The COSMIN checklist for evaluating the methodological quality of studies on measurement properties: A clarification on its content. *BMC Medical Research Methodology*.