COSMIN: Development and evaluation of a checklist to assess the methodological quality of studies on measurement properties

Mokkink, L.B.

2010

document version
Publisher's PDF, also known as Version of record

7

Inter-rater reliability of the COSMIN Checklist

LB Mokkink, CB Terwee, E Gibbons, PW Stratford, J Alonso, DL Patrick, DL Knol,
LM Bouter, HCW de Vet

Submitted for publication

COSMIN

Abstract

Introduction

The COSMIN checklist is a tool for evaluating the methodological quality of studies on measurement properties of health-related patient-reported outcomes. The aim of this study is to determine the inter-rater reliability of each COSMIN item (n=114).

Methods

75 articles evaluating measurement properties were randomly selected from the bibliographic database compiled by the Patient-Reported Outcome Measurement Group, Oxford, UK. Raters were asked to assess the methodological quality of three articles, using the COSMIN checklist. In a one-way design, intraclass kappa's or quadratic-weighted kappa's, and percentage agreement were calculated for each item.

Results

88 raters participated. Of the 75 selected articles, 26 articles were rated by four to six participants, and 49 by two or three participants. Overall, the kappa's for the COSMIN items were low (59% was below 0.40, 6% was above 0.75), while the percentage agreement was appropriate (64% was above 80% agreement). Reasons for low agreement were complexity of the task, the need for subjective judgement, and accustom to different standards, terminology and definitions.

Conclusions

The COSMIN items showed a large variation in agreement among raters. When using the checklist in a systematic review, we recommend getting some training and experience, completing it by two independent raters, and reaching consensus on one final rating. Instructions for using the checklist are improved.

Introduction

Recently, a checklist for the evaluation of the methodological quality of studies on measurement properties of health-related patient-reported outcomes (HR-PROs) – the COSMIN checklist – was developed in an international Delphi study [1]. COSMIN is an acronym for COnsensus-based Standards for the selection of health Measurement Instruments. This checklist can be used for the appraisal of the methodological quality of studies included in a systematic review of measurement properties of HR-PROs. It can also be used to design and report a study on measurement properties. Also, reviewers and editors could use it to identify shortcomings in studies on measurement properties, and to assess whether the methodological quality of such studies is high enough to justify publication.

The COSMIN checklist contains twelve boxes [1]. Ten boxes can be used to assess whether a study meets the standards for good methodological quality (ranging from 5-18 items). 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) and 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, one box (IRT box) contains requirements for articles in which IRT methods are applied (4 items), and one box (Generalisability box) is included in the checklist that contains requirements for the generalisability of the results (8 items).

To assess the quality of the COSMIN checklist itself, one of the relevant measurement properties is reliability. It is important that the researchers who use the COSMIN checklist give the same ratings on each item. Therefore, the aim of this study is to determine the inter-rater reliability of each item of the COSMIN checklist among potential users.

Methods

Because the COSMIN checklist will be applied in the future to a variety of studies on different topics and study populations, with low and high quality, it was our goal to generalise the results of this study to a broad range of articles on measurement properties. In addition, the COSMIN checklist will be used by many researchers, using the instructions in the COSMIN manual as guidance. We were interested in the reliability in this situation. Often, in an article only a selection of measurement properties are being evaluated. Consequently, only part of the COSMIN checklist can be completed. We decided in advanced that (1) we aimed for four ratings for each item of the COSMIN checklist on the same article; (2) we aimed for each measurement property to be evaluated in at least 20 different articles.

R1
R2
R3
R4
R5
R6
R7
R8
R9
R10
R11
R12
R13
R14
R15
R16
R17
R18
R19
R20
R21
R22
R23
R24
R25
R26
R27
R28
R29
R30
R31
R32
R33
R34
R35
R36
R37
R38
R39

Article selection

In this study we included articles that were representative of studies on measurement properties. We selected articles from the bibliographic database compiled by the Patient-Reported Outcome Measurement (PROM) Group, Oxford, UK (http//phi.uhce.ox.ac.uk). The selection of articles for this study was a two-step procedure. First, of the 30,000+ included articles it was determined, based on the title, whether it concerned an article of a study on the evaluation of measurement properties of a PRO. For example, the title included terms of a specific measurement property, such as reliability, validity, or responsiveness. A total of 5137 articles were eligible. Second, from these articles, we randomly selected studies that fulfilled our inclusion criteria.

Inclusion criteria were:

- Purpose of the study was to evaluate one or more measurement properties
- Instrument under study was a HR-PRO instrument
- English language publications

Articles from any setting and any population could be included, and articles could have used Classical Test Theory (CTT) or modern test theory (i.e, Item Response Theory (IRT)) or both.

Exclusion criteria:

- Systematic reviews, case reports, letters to editors
- Studies that evaluated construct validity of two or more instruments at the same time by correlating the scores of the instruments mutually, without indicating one of instruments as the instrument of interest. In these studies, it is unclear of which instrument the construct validity is being assessed.

One of the authors (LM) selected articles until each measurement property was assessed in at least 20 articles. It appeared that we needed to select 75 articles. For each included article LM determined the relative workload for a rater to evaluate the methodological quality of the article, i.e. high, moderate, or low workload. The relative workload was based on the number of measurement properties assessed in the study, the number of instruments that were studied, the number of pages, and whether IRT was used. For example, an article in which IRT is used is considered having a high workload, and an article in which three measurement properties were evaluated in a four page paper was considered as having a low workload. We decided to ask each rater to evaluate three articles. We provided each rater with one article with a low workload, one with a moderate workload and one with a high workload.

Selection of participants

Raters were professionals who had some experience with assessing measurement properties. This could range from having little experience to being an expert. We choose to select a heterogeneous group of raters, because this reflects best the raters who will potentially use the COSMIN checklist in the future. We invited the panel of the COSMIN Delphi study to participate in the inter-rater reliability study (n=91), attendees of two courses on clinimetrics given in 2009 by the department of Epidemiology and Biostatistics of the VU University Medical Center (n=72), researchers on the mailing list of the Dutch chapter of the International Society for Quality of Life Research (ISOQOL-NL) (n=295), members of the EMGO Clinimetrics working group (n=32), members of the PRO Methods Group of the Cochrane Collaboration (n=79), researchers who previously showed interest in the COSMIN checklist (n=15), colleagues of the authors, and other researchers who were likely to show interest. We also asked these people if they knew other researchers who were interested in participating.

Procedure

Those who agreed to participate received three selected articles, together with a manual of the COSMIN checklist [2] and a data collection form to enter their scores. For each article, they were firstly asked to indicate for each measurement property whether it was evaluated in the article ('yes/no') (step 1 of the COSMIN procedure). The participants had to determine themselves which boxes they should complete for each of the three papers. Secondly, they were asked whether IRT was used in the article, and if so, they were asked to complete the IRT box (step 2). Thirdly, they were asked to complete the relevant boxes of the COSMIN checklist (step 3). Fourthly, raters were asked to complete the Generalisability box (step 4) for each measurement property assessed in the article.

Instructions on how to complete the boxes were provided in the COSMIN manual [2]. Raters did not receive any additional training in completing the COSMIN checklist and were not familiar with the checklist. Items could be answered with "yes"/ "no", with "yes"/ "?"/ "no", or with "yes"/ "no"/ "not applicable" ("na"). One item had four response options, i.e., "yes"/ "?"/ "no"/ or "na".

Statistical analyses

Each rater scored three of the 75 selected articles, and in each article a selection of the measurement properties was evaluated. Therefore, inter-rater reliability of each COSMIN item was analyzed using a one-way design. Dichotomous items were analysed using intraclass kappa's [3, 4]; the scoring was yes=1 and no=0.

$$Intraclass\ Kappa_{COSMINitem} = \frac{\sigma^2_{article}}{P(1-P)},$$

where $\sigma^2_{article}$ denotes the variance due to systematic differences between the articles for which the item was scored, and P is the proportion of those articles that were scored 'yes'. Given a rating of an article on a COSMIN item, intraclass kappa can be interpreted as the difference of the conditional probability that a second rating of the same article is the same as the first rating minus the conditional probability that the second rating is different from the first [3, 4, 5]. Ordinal items were analyzed with weighted kappa's using quadratic weights; the scoring was 'yes'=1, '?'=2, and 'no'=3. These measures are numerically the same as intraclass correlation coefficients (ICCs) obtained from analysis of variance (ANOVA) [3, 6, 7].

Twenty-two items could be answered with "na", which makes the scale of these items a nominal scale. Kappa's nor ICCs are able to handle nominal items in a one-way design. Since we do not calculate overall scores per box, we did not calculate kappa's per box. We only calculated kappa's per COSMIN item. We considered a kappa for each item below 0.40 as poor, between 0.40 and 0.75 as moderate to good, and above 0.75 as excellent [7].

Reliability measures such as kappa are dependent on the distribution of the data ($\sigma^2_{article}$). Vach showed that reliability measures are low when data are skewed [8]. Therefore, we also calculated percentage agreement for each item. We considered a distribution of scores as skewed when more than 75% of the raters who responded to an item used the same response category. We considered the highest number of similar ratings per item per article as agreement, and the other ratings as non-agreement. For example, if five raters rated the same item for the same article, and three of the raters rated 'yes', and two rated 'no', we considered three ratings as agreement. Percentage agreement was calculated by the number of ratings with agreement on all articles, divided by the total number of ratings on all articles for which that measurement property was assessed. A percentage agreement > 80% was considered appropriate.

In our analysis we combined scores of the items on the Generalisability box for all measurement properties, so that we calculated kappa's and percentage agreement only once for each of the items from this box, and not separately for each measurement property.

Results

A total of 154 people agreed to participate in this study. We received the ratings from 88 (57%) of the participants. The responders came from the Netherlands (58%), Canada (10%), UK (7%), Australia or New Zealand (6%), Europe without Netherlands and UK (15%), and

other (5%). The mean number of years experience in research was 12 years (SD = 8.7), and 9 years (SD = 7.1) experience in research related to measurement properties.
Of the 75 selected articles, 8 articles were rated by 6 participants, 7 articles were rated by five participants, 11 by four participants, 38 by three participants, and 11 by two participants. The percentage missing items per box were 7% for box A internal consistency, 5% for box B reliability, 1% box D content validity, 11% box E structural validity, 7% box F hypotheses testing, 5% box G cross-cultural validity, 5% box H criterion validity, 18% box I responsiveness, 3% box J interpretability, and 1% for the Generalisability box.

Items of the IRT box had 26 ratings for 13 articles; for 6 articles this box was completed by one rater, for two articles by two raters, for four articles by three raters, and for one article by four raters. The box C measurement error had 17 ratings for 14 articles; for twelve articles this box was completed by one rater, for one article by two raters, and one article by three raters. The results of these items are not shown, because kappa's and percentage agreement based on such small numbers are unreliable. For the property measurement error, however, we have some information because 10 of the 11 items from this box (i.e. all items on design requirements) were exactly the same items as the items about design requirements from box B reliability (i.e. items B1 to B10, see Table 2).

Inter-rater reliability

Table 1 shows the agreement between raters about whether the property was evaluated in an article. Note that these scores are not summary scores of the overall methodological quality of the property. In this table we describe kappa's, the variance components (i.e., article variance and total variance), and percentages agreement. Two of the ten properties, i.e. reliability and responsiveness, had an excellent kappa's, i.e. above 0.75. Three properties had moderate to good kappa's and five had poor kappa's. All properties had high percentage agreement (range from 84% to 96%).

Table 1. Kappa's and percentage agreement on whether the property was evaluated in an article (step 1)

	Intraclass kappa[a]	Article variance	Total variance	% agreement
Internal consistency	0.66	0.13	0.19	**94**
Reliability	**0.77**	0.19	0.25	**94**
Measurement error	*0.02*	*0.00*	*0.06*	***94***
Content validity	0.29	0.07	0.23	**84**
Structural validity	0.48	0.12	0.25	**86**
Hypotheses testing	0.29	0.07	0.23	**87**
Cross-cultural validity	*0.66*	*0.07*	*0.11*	***95***
Criterion validity	*0.23*	*0.04*	*0.17*	***93***
Responsiveness	**0.81**	0.17	0.21	**96**
Interpretability	*0.02*	*0.00*	*0.13*	***86***

[a] number of ratings on the 75 articles = 263; printed in italic indicates items with low dispersal i.e. more than 75% of the raters who responded to an item rated the same response category; printed in bold indicates kappa>0.70 or % agreement >80%

In Table 2 we describe kappa's, the variance components, and percentages agreement for each item of the COSMIN checklist (step 3 and 4). Of the 104 items, kappa's of 22 items could not be calculated due to nominal response options of these items. Of the remaining 82 items five (6%) had an excellent kappa, twenty-nine (35%) had a moderate to good kappa, and 48 items (59%) had a poor kappa (including the 8 items of which we set negative variance component to 0). Sixty-seven items of the 104 items in Table 2 (64%) had a percentage agreement above 80%. Thirty items (29%) had a percentage agreement between 70% and 80%, and seven items (7%) between 60% and 70%. Sample sizes for kappa's and percentage agreement per item were slightly different, due to articles that were scored only once by one rater. When calculating percentage agreement, these articles were not taking into account.

We observed two issues. Firstly, thirty-two of the 114 items (Table 1 and 2; 28%) showed hardly any dispersal, i.e. more than 75% of the raters who responded to the item rated the same response category. These items were presented in italics in Tables 1 and 2. When data are skewed, the between article variance, i.e. $\sigma^2_{article,}$ is low, and thus the kappa will be low. Secondly, in Table 2 it can be seen that twenty-nine items (28%) had a sample size below 50 for the calculation of kappa's, of which four were below 30 (4%). For the calculation of percentage agreement thirty-five items (34%) had a sample size of below 50, of which twenty-nine was below 30 (28%). These kappa's and percentage agreement based on such small numbers should be interpreted with caution.

Table 2. Kappa's and percentage agreement of the items from the COSMIN checklist

Item nr	Item	N	Kappa	Article variance	Total variance	N (minus articles with 1 rating)[a]	% agree-ment
Step 3							
Box A Internal consistency (n=195)[b]							
A1	Does the scale consist of effect indicators, i.e. is it based on a reflective model?	193	0.06	0.02	0.37	185	**82**
Design requirements							
A2	Was the percentage of missing items given?	190	0.48	0.11	0.24	183	**87**
A3	Was there a description of how missing items were handled?	187	0.54	0.11	0.21	180	**90**
A4	*Was the sample size included in the internal consistency analysis adequate?*	*185*	*0.06*	*0.01*	*0.25*	*177*	***87***
A5	Was the unidimensionality of the scale checked? i.e. was factor analysis or IRT model applied?	187	0.69	0.18	0.25	180	**92**
A6	Was the sample size included in the unidimensionality analysis adequate?	178	0.27	0.11	0.42	166	79
A7	*Was an internal consistency statistic calculated for each (unidimensional) (sub)scale separately?*	*187*	*0.31*	*0.15*	*0.49*	*179*	***85***
A8	*Were there any important flaws in the design or methods of the study?*	*179*	*0.22*	*0.04*	*0.18*	*174*	***86***
Statistical methods							
A9	*for Classical Test Theory (CTT): Was Cronbach's alpha calculated?*	*187*	[c]			*179*	***93***
A10	*for dichotomous scores: Was Cronbach's alpha or KR-20 calculated?*	*165*	[c]			*151*	***91***
A11	*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)*	*167*	[c]			*154*	***93***
Box B. Reliability (n=141)[b]							
Design requirements							
B1	Was the percentage of missing items given?	140	0.39	0.08	0.19	129	**87**
B2	*Was there a description of how missing items were handled?*	*137*	*0.43*	*0.06*	*0.14*	*125*	***91***
B3	Was the sample size included in the analysis adequate?	139	0.35	0.27	0.77	127	77
B4	*Were at least two measurements available?*	*140*	***0.72***	*0.05*	*0.07*	*129*	***98***
B5	Were the administrations independent?	139	0.18	0.06	0.35	129	73
B6	*Was the time interval stated?*	*136*	*0.50*	*0.07*	*0.14*	*125*	***94***
B7	Were patients stable in the interim period on the construct to be measured?	138	0.24	0.09	0.38	126	75
B8	Was the time interval appropriate?	137	0.45	0.32	0.70	125	**84**

Item nr	Item	N	Kappa	Article variance	Total variance	N (minus articles with 1 rating)[a]	% agree-ment
B9	Were the test conditions similar for both measurements? e.g. type of administration, environment, instructions	138	0.30	0.11	0.36	127	**83**
B10	Were there any important flaws in the design or methods of the study?	129	0.08	0.02	0.24	117	77
Statistical methods							
B11	for continuous scores: Was an intraclass correlation coefficient (ICC) calculated?	133	[c]			119	**86**
B12	for dichotomous/nominal/ordinal scores: Was kappa calculated?	127	[c]			111	**81**
B13	for ordinal scores: Was a weighted kappa calculated?	127	[c]			111	**83**
B14	for ordinal scores: Was the weighting scheme described? e.g. linear, quadratic	124	[c]			108	**81**
Content validity (n=83) [b]							
Design requirements							
D1	Was there an assessment of whether all items refer to relevant aspects of the construct to be measured?	83	0.33	0.2	0.60	62	79
D2	Was there an assessment of whether all items are relevant for the study population? (e.g. age, gender, disease characteristics, country, setting)	83	0.46	0.32	0.70	62	76
D3	Was there an assessment of whether all items are relevant for the purpose of the measurement instrument? (discriminative, evaluative, and/or predictive)	83	0.21	0.15	0.74	62	66
D4	Was there an assessment of whether all items together comprehensively reflect the construct to be measured?	83	0.15	0.10	0.64	62	66
D5	Were there any important flaws in the design or methods of the study?	78	0.13	0.03	0.23	58	76
Structural validity (n=118) [b]							
E1	Does the scale consist of effect indicators, i.e. is it based on a reflective model?	116	0[d]	0	0.26	99	78
Design requirements							
E2	Was the percentage of missing items given?	110	0.41	0.09	0.21	95	**87**
E3	Was there a description of how missing items were handled?	109	0.55	0.11	0.19	93	**91**
E4	*Was the sample size included in the analysis adequate?*	*109*	*0.56*	*0.26*	*0.47*	*94*	**87**
E5	Were there any important flaws in the design or methods of the study?	103	0.27	0.06	0.21	89	**84**

Statistical methods							
E6	*for CTT: Was exploratory or confirmatory factor analysis performed?*	*106*	[c]			*92*	**90**
E7	for IRT: Were IRT tests for determining the (uni-) dimensionality of the items performed?	80	[c]			62	**87**
Hypotheses testing (n=170) [b]							
Design requirements							
F1	Was the percentage of missing items given?	168	0.41	0.09	0.21	158	**87**
F2	*Was there a description of how missing items were handled?*	*169*	*0.60*	*0.11*	*0.19*	*159*	**92**
F3	*Was the sample size included in the analysis adequate?*	*167*	*0.12*	*0.04*	*0.36*	*157*	***84***
F4	Were hypotheses regarding correlations or mean differences formulated a priori (i.e. before data collection)?	168	0.42	0.32	0.76	158	74
F5	Was the expected *direction* of correlations or mean differences included in the hypotheses?	169	[c]			159	75
F6	Was the expected absolute or relative *magnitude* of correlations or mean differences included in the hypotheses?	168	[c]			159	**82**
F7	for convergent validity: Was an adequate description provided of the comparator instrument(s)?	136	0.30	0.07	0.23	125	**83**
F8	for convergent validity: Were the measurement properties of the comparator instrument(s) adequately described?	135	0.35	0.09	0.25	124	**81**
F9	Were there any important flaws in the design or methods of the study?	145	0.17	0.04	0.21	131	**81**
Statistical methods							
F10	*Were design and statistical methods adequate for the hypotheses to be tested?*	*161*	[c]			*150*	*78*
Cross-cultural validity (n=33) [b]							
Design requirements							
G1	Was the percentage of missing items given?	32	0.52	0.14	0.26	25	**88**
G2	Was there a description of how missing items were handled?	30	0.32	0.08	0.24	22	**82**
G3	Was the sample size included in the analysis adequate?	33	0.23	0.16	0.70	26	**81**
G4	*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?*	*33*	*0.34*	*0.05*	*0.13*	*28*	***89***
G5	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	33	0.46	0.11	0.23	28	**86**
G6	Did the translators work independently from each other?	33	0.61	0.20	0.32	28	**89**

Item nr	Item	N	Kappa	Article variance	Total variance	N (minus articles with 1 rating)[a]	% agree-ment
G7	Were items translated forward and backward?	33	**1.00**	0.52	0.52	28	**100**
G8	Was there an adequate description of how differences between the original and translated versions were resolved?	33	0.50	0.13	0.26	28	**86**
G9	Was the translation reviewed by a committee (e.g. original developers)?	31	0.56	0.14	0.25	25	**88**
G10	Was the HR-PRO instrument pre-tested (e.g. cognitive interviews) to check interpretation, cultural relevance of the translation, and ease of comprehension?	29	0.61	0.14	0.23	21	**90**
G11	Was the sample used in the pre-test adequately described?	32	0[d]	0	0.15	28	79
G12	Were the samples similar for all characteristics except language and/or cultural background?	31	0.41	0.14	0.34	26	**81**
G13	Were there any important flaws in the design or methods of the study?	31	0.42	0.10	0.25	26	**85**
Statistical methods							
G14	for CTT: Was confirmatory factor analysis performed?	32	[c]			27	74
G15	for IRT: Was differential item function (DIF) between language groups assessed?	23	[c]			13	77
Criterion validity (n=57)[b]							
Design requirements							
H1	*Was the percentage of missing items given?*	*56*	*0.59*	*0.10*	*0.17*	*35*	***91***
H2	*Was there a description of how missing items were handled?*	*56*	***0.79***	*0.09*	*0.11*	*35*	***97***
H3	Was the sample size included in the analysis adequate?	54	0.06	0.05	0.70	35	69
H4	Can the criterion used or employed be considered as a reasonable 'gold standard'?	57	0[d]	0	0.13	37	62
H5	Were there any important flaws in the design or methods of the study?	54	0.10	0.03	0.25	33	79
Statistical methods							
H6	for continuous scores: Were correlations, or the area under the receiver operating curve calculated?	56	[c]			37	78
H7	for dichotomous scores: Were sensitivity and specificity determined?	47	[c]			29	**83**
Responsiviness (n=79)[b]							
Design requirements							
I1	*Was the percentage of missing items given?*	*76*	*0.14*	*0.02*	*0.17*	*71*	***82***

12	*Was there a description of how missing items were handled?*	*77*	*0.36*	*0.04*	*0.12*	*73*	**92**
13	Was the sample size included in the analysis adequate?	76	0.40	0.28	0.70	72	72
14	*Was a longitudinal design with at least two measurement used?*	*78*	***1.00***	*0.01*	*0.01*	*73*	***100***
15	*Was the time interval stated?*	*78*	*0.25*	*0.05*	*0.18*	*73*	**89**
16	If anything occurred in the interim period (e.g. intervention, other relevant events), was it adequately described?	75	0.17	0.05	0.28	72	78
17	*Was a proportion of the patients changed (i.e. improvement or deterioration)?*	*73*	*0.32*	*0.03*	*0.09*	*70*	**97**
Design requirements for hypotheses testing							
For constructs for which a gold standard was not available							
18	Were hypotheses about changes in scores formulated a priori (i.e. before data collection)?	72	0.35	0.30	0.87	65	69
19	Was the expected *direction* of correlations or mean differences of the change scores of HR-PRO instruments included in these hypotheses?	65	[c]			60	78
110	*Were the expected absolute or relative magnitude of correlations or mean differences of the change scores of HR-PRO instruments included in these hypotheses?*	*66*	[c]			*61*	**90**
111	Was an adequate description provided of the comparator instrument(s)?	63	0[d]	0	0.21	56	70
112	Were the measurement properties of the comparator instrument(s) adequately described?	63	0.06	0.01	0.22	56	**80**
113	Were there any important flaws in the design or methods of the study?	68	0.03	0.01	0.25	63	71
Statistical methods							
114	Were design and statistical methods adequate for the hypotheses to be tested?	67	[c]			63	73
Design requirements for comparison to a gold standard							
For constructs for which a gold standards was available:							
115	Can the criterion for change be considered as a reasonable 'gold standard'?	28	0[d]	0	0.76	21	67
116	Were there any important flaws in the design or methods of the study?	21	0[d]	0	0.	12	67
Statistical methods							
117	for continuous scores: Were correlations between change scores, or the area under the Receiver Operator Curve (ROC) curve calculated?	39	[c]			28	79
118	for dichotomous scales: Were sensitivity and specificity (changed versus not changed) determined?	37	[c]			28	79
Interpretability (n=42) [b]							
J1	Was the percentage of missing items given?	41	**0.80**	0.21	0.26	22	**95**
J2	Was there a description of how missing items were handled?	41	0.19	0.05	0.25	21	76
J3	Was the sample size included in the analysis adequate?	41	0[d]	0	0.32	23	74
J4	Was the distribution of the (total) scores in the study sample described?	41	0.08	0.02	0.23	23	74
J5	Was the percentage of the respondents who had the lowest possible (total) score described?	40	**0.84**	0.22	0.26	20	**95**

Item nr	Item	N	Kappa	Article variance	Total variance	N (minus articles with 1 rating)[a]	% agreement
J6	Was the percentage of the respondents who had the highest possible (total) score described?	41	**0.70**	0.18	0.26	21	**90**
J7	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	41	0.05	0.01	0.26	21	76
J8	*Was the minimal important change (MIC) or the minimal important difference (MID) determined?*	*40*	*0.26*	*0.03*	*0.13*	*19*	***89***
J9	Were there any important flaws in the design or methods of the study?	41	0[d]	0	0.20	21	71
Step 4							
Generalisability box (n=866)[b]		[e]					
Was the sample in which the HR-PRO instruments was evaluated adequately described? In terms of:							
1	median or mean age (with standard deviation or range)?	865	0.36	0.07	0.20	733	**86**
2	*distribution of sex?*	*863*	*0.38*	*0.07*	*0.18*	*735*	***88***
3	important disease characteristics (e.g. severity, status, duration) and description of treatment?	862	[c]			746	**80**
4	*setting(s) in which the study was conducted? e.g. general population, primary care or hospital/ rehabilitation care*	*863*	*0.30*	*0.04*	*0.12*	*735*	***89***
5	*countries in which the study was conducted?*	*861*	*0.40*	*0.05*	*0.13*	*733*	***90***
6	*language in which the HR-PRO instrument was evaluated?*	*861*	*0.41*	*0.07*	*0.18*	*733*	***86***
7	Was the method used to select patients adequately described? e.g. convenience, consecutive, or random	857	0.40	0.10	0.24	729	**81**
8	Was the percentage of missing responses (response rate) acceptable?	849	0.48	0.17	0.36	724	**82**

[a] When calculating percentage agreement, articles that were only scored once on the particular item were not taking into account; [b] number of times a box was evaluated; [c] Kappa's could not be calculated because of nominal response scale in a one-way design; [d] Kappa is zero due to negative variance component which was set at 0; [e] sample sizes of Generalisability box are much higher that other items, because scores of the items on the Generalisability box for all measurement properties were combined; printed in italic indicates items with low dispersal i.e. more than 75% of the raters who responded to an item rated the same response category; printed in bold indicates Kappa>0.70 or % agreement >80%.

Discussion

In this study we investigated the inter-rater reliability of the COSMIN checklist. Overall, more than half of the items had poor kappa's. We will start the discussion with a statistical reason for low kappa's, and reasons for low agreement due to the complexity of the task.

Reasons for low kappa's

While kappa's were low, the percentage agreement was appropriate (i.e. above 80%) for 64% of the items. For thirty-two of the items (28%) there is an obvious reason for this seemingly contradictory finding: the kappa's were low due to a lack of dispersal of the scores, i.e. more than 75% of the raters who responded to an item rated the same response category. For example, item 15 of the box Responsiveness (i.e. was the time interval stated) had a kappa of 0.25; 65 times raters scored "yes" (83%), and 13 times they scored "no" (17%). The percentage agreement was 89%. Low dispersal rates strongly influence the kappa, because if the variance between articles is low, the error variance is large in relation to the article variance.

Reasons for low agreement between raters

For many items of the COSMIN checklist a subjective judgement is needed. For example, in each box the item *'were there any important flaws in the design or the methods of the study'* was included (e.g., B10, 113, 116 and J9). To answer this question, the rater should judge this based on his own experience and knowledge. Therefore, some kind of subjective evaluation is involved. Some other items might be rather difficult to score, because the information needed to answer the item is not reported in the article. For example, information to be able to respond on the item *'were the administrations independent'* (B5) is often not reported. Although raters should score '?' in this case, raters are likely to guess, or skip these items. This influences the kappa and the percentage agreement.

Furthermore, the COSMIN checklist contains consensus-based standards that may deviate from how persons are used to evaluate measurement properties or a person may disagree on a particular item. Consequently, a rater may score an item differently than recommended in the COSMIN manual. For example, many people consider effect sizes as appropriate measures for responsiveness. Within the COSMIN Delphi study, we decided to consider this as inappropriate [9]. We believe that only when clear hypotheses are formulated about the expected magnitude of the effect sizes (ES) it is appropriate as an indicator of responsiveness (114). Another example is the issue about the gold standard. The COSMIN panel considered a commonly used measurement instrument, such as the SF-36, not as a reasonable gold standard. However, raters may disagree with this, and rate the item '*can the criterion (for change) be considered as a reasonable gold standards'* (H4 and 115) as 'yes' while according to

the COSMIN manual this item should be scored with 'no'. Consequently, the kappa and the percentage agreement will be low.

Last, the distinction between rating the methodological quality of the study and rating the quality of the instrument that is evaluated in the study may be difficult, especially for content validity. Therefore, the items on content validity are difficult to score. All items of box D of content validity had low kappa's and percentage agreement. They ask whether the article under study appropriately *investigated* whether the items were relevant and comprehensive. This refers to the methodological quality of a study. For example, an appropriate method to investigate the content validity of a HR-PRO is involving patients from the target population, by asking them about the relevance and comprehensiveness of the items. These COSMIN items do not ask whether the items of the PRO under study *are* relevant and comprehensive, which refers to the quality of an instrument. Raters may have been confused about this distinction.

Strength and weaknesses of the study

It was our aim to randomly select equal numbers of studies on each measurement property. However, studies on internal consistency and hypotheses testing are more common than studies on measurement error, and interpretability. Studies that are based on CTT are more common than studies that apply IRT methods. Consequently, these less common measurement properties were less often selected for this study. This prevented analysis of the items on measurement error and on IRT analysis.

In addition, it was our aim to include a representative sample of potential users of the COSMIN checklist. As expected, the years of experience of the participants in this study both in research in general and in research in measurement instruments differed widely. Although more than half of the raters came from the Netherlands, we do not expect that the country of origin will have a major influence on the results.

In this study it was not feasible to train the raters because we expected that this would dramatically decrease the response rate. However, we recommend getting some experience in completing the COSMIN checklist before conducting a systematic review. In the future, when more raters are trained in completing the checklist, a reliability study among trained raters could be performed.

Due to the incomplete study design (i.e. not all raters scored all articles, and in an article not all measurement properties are evaluated) we had a one-way design. Therefore, the variance due to raters could not be distinguished from the error variance. Other optional designs would be asking a few raters to evaluate many articles, or asking many raters to evaluate the same few articles. Both designs were considered poor. In the first case, it is likely that we would not find participants, due to the large amount of work each rater had to do. We

felt that we as authors of the COSMIN checklist should not be these raters, because of our involvement in the development of the checklist. The second design is considered poor because we would have to include a few articles in which all measurement properties were evaluated. It is very likely that these articles do not exist, and if such an article is published, it is very likely that it is not a good representation of studies on measurement properties.
Since the kappa's were in several cases difficult to interpret, due to skewness of data, low sample sizes, etc, we also calculated percentage agreement. However, this measure is a poor measure of reliability. Often, this measure leads to artificial high agreements, because it does not take 'agreement by chance' into account.

Recommendations for improvement of the inter-rater reliability of the COSMIN checklist

Firstly, based on the results of this study, and feedback we received from raters, we improved the wording and grammar of a few items and we adapted the instructions in the manual. This might improve the inter-rater reliability of COSMIN items. Secondly, the COSMIN checklist is not a ready-made checklist, in a sense that the user can instantly complete all items. We recommend that researchers who use the COSMIN checklist, for example in a systematic review, agree beforehand on how to handle items that need a subjective judgement, and how to deal with lack of reporting in the original article. For example, based on the topic of the review, they should agree on what they consider an appropriate time interval for reliability (B8), on an adequate description for the comparator instrument(s) (F7 and I11), or on an acceptable percentage of missing responses (item 8 of the Generalisability box). This may also increase the inter-rater reliability. Thirdly, some experience in completing the checklist before conducting a systematic review is also likely to increase the reliability of the COSMIN checklist. Therefore, we are developing a training set of articles (to be published on our website), explaining how these articles should be evaluated using the COSMIN checklist. Fourthly, we strongly recommend using the taxonomy and terminology of the COSMIN checklist. For example, if authors compare their PRO to a commonly used PRO such as the SF-36, and they refer to this as criterion validity, we recommend considering this an evaluation of hypotheses testing which is an aspect of construct validity, and complete box F. Fifthly, when using the checklist in a systematic review of HR-PROs, we recommend to complete the checklist by at least two independent raters, and to reach consensus on one final rating. In this study we used the ratings of single raters to determine the reliability of the checklist, because a design with consensus scores of two raters was not feasible. We recommend evaluating the reliability of the consensus scores of couples of raters in a future study, when more raters are trained.

R1
R2
R3
R4
R5
R6
R7
R8
R9
R10
R11
R12
R13
R14
R15
R16
R17
R18
R19
R20
R21
R22
R23
R24
R25
R26
R27
R28
R29
R30
R31
R32
R33
R34
R35
R36
R37
R38
R39

Other measurement properties of the COSMIN checklist

Content validity and hypothesis testing are also relevant measurement properties for evaluating the quality of the COSMIN checklist. To ensure good content validity, we choose to involve many experts from different fields in the development of the checklist. Therefore, it is highly likely that all relevant items of specific measurement properties are included, which contributes to the content validity of the checklist. However, since content validity is a subjective judgement, an unbiased judgment cannot be performed by the developers, and therefore other researchers should assess this.

Construct validity should be assessed by means of hypotheses testing. However, we consider it very difficult to formulate challenging and specific hypotheses for assessing construct validity of the COSMIN checklist. Firstly, because no overall scores per box are calculated, one can 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 of different checklists is often very similar. The difference with other checklists is the inclusion of different items, which is an aspect of content validity. Secondly, even when an overall score for each box can be calculated, it is difficult to formulate hypotheses. Related checklists, such as the criteria proposed by the SAC-MOS [10], and the Terwee criteria [11], focus on the quality of measurement instruments, which is a different construct than the methodological quality of studies on measurement properties. Moreover, the measurement properties of these checklists have not been investigated. So, if the COSMIN checklist is compared to these checklists, and results are poor, it is not clear if this is due to a poor quality of the COSMIN checklist, or to a poor quality of the comparison instruments.

Other measurement properties are not relevant. Internal consistency and structural validity are not relevant, because the COSMIN checklist is based on a formative model, i.e. the items together form the construct [12]. Measurement error cannot be assessed, because there is no parameter of measurement error for ordinal or nominal scales. The COSMIN checklist is only available in English, and we have no intention to translate the checklist into other languages. Therefore cross-cultural validity is currently not relevant. Criterion validity cannot be assessed, because there is no gold standard for assessing the methodological quality of studies on measurement properties. And last, responsiveness is not relevant because the studies that are being evaluated with the checklist do not change over time.

Conclusions

The items of the COSMIN checklist showed a large variation in agreement among raters. Some disagreements are likely to be influenced by a subjective judgement needed to answer an item. Therefore, we recommend making decisions in advanced about how to score these issues. The reliability of other items with large disagreements may have improved since we have tried to improve the instructions in the manual on some issues, based on the feedback of raters. When using the COSMIN checklist it is important to read the manual carefully, and get some training and experience in completing the checklist.

Reference List

(1) Mokkink LB, Terwee CB, Patrick DL, et al. (2010). 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, 19, 539-49.

(2) Mokkink LB, Terwee CB, Patrick DL, et al. (2009). The COSMIN checklist manual. www.cosmin.nl

(3) Kraemer HC, Periyakoil VS, Noda A. (2002). Tutorial in biostatistics. Kappa coefficients in medical research. Statistics in Medicine, 21, 2109-2129.

(4) Kraemer HC. (2006). Correlation coefficients in medical research: from product moment correlation to the odds ratio. Statistical Methods in Medical Research, 15, 525-545.

(5) Kraemer HC. (1992). Measurement of reliability for categorical data in medical research. Statistical Methods in Medical Research, 1, 183-199.

(6) Lin L, Hedayat AS, Wu W. (2007). A unified approach for assessing agreement for continuous and categorical data. Journal of Biopharmaceutical Statistics, 17, 629-652.

(7) Fleiss JL. (1981). Statistical methods for rates and proportions. New York: John Wiley & Sons.

(8) Vach W. (2005). The dependence of Cohen's kappa on the prevalence does not matter. Journal of Clinical Epidemiology, 58, 655-661.

(9) Mokkink LB, Terwee CB, Knol DL, et al. (2010). The COSMIN checklist for evaluating the methodological quality of studies on measurement properties: A clarification of its content. BMC Medical Research Methodology, 10, 22.

(10) Scientific Advisory Committee of the Medical Outcomes Trust. (2002). Assessing health status and quality-of-life instruments: Attributes and review criteria. Quality of Life Research, 11, 193-205.

(11) Terwee CB, Bot SD, De Boer MR, et al. (2007). Quality criteria were proposed for measurement properties of health status questionnaires. Journal of Clinical Epidemiology, 60, 34-42.

(12) Streiner DL. (2003). Being inconsistent about consistency: when coefficient alpha does and doesn't matter. Journal of Personality Assessment, 80, 217-222.