COSMIN Risk of Bias checklist

Date: July, 2018

Contact
L.B. Mokkink, PhD
VU University Medical Center
Department of Epidemiology and Biostatistics
Amsterdam Public Health research institute
P.O. box 7057
1007 MB Amsterdam
The Netherlands
Website: www.cosmin.nl
E-mail: w.mokkink@vumc.nl

How to site the COSMIN Risk of Bias Checklist

Please refer to the following studies when using the COSMIN Risk of Bias Checklist:

Mokkink, L.B., De Vet, H.C.W., Prinsen, C.A.C, Patrick, D.L., Alonso, J., Bouter, L.M., et al. COSMIN Risk of Bias checklist for systematic reviews of Patient-Reported Outcome Measures. Accepted for publication in Quality of Life Research.

Prinsen, C. A., Mokkink, L. B., Bouter, L. M., Alonso, J., Patrick, D. L., Vet, H. C., et al. COSMIN guideline for systematic reviews of Patient-Reported Outcome Measures. Submitted.

Terwee, C. B., Prinsen, C. A., Chiarotto, A., Vet, H. C., Westerman, M. J., Patrick, D. L., et al. COSMIN methodology for evaluating the content validity of Patient-Reported Outcome Measures: a Delphi study. Submitted.

For details on how to use the COSMIN risk of Bias checklist see 'COSMIN methodology for systematic reviews of Patient-Reported Outcome Measures (PROMs) – user manual' and 'COSMIN methodology for assessing the content validity of Patient-Reported Outcome Measures (PROMs) - user manual' available from our website www.cosmin.nl.

Abbreviations used:

CTT – classical test theory

DIF – differential item functioning

IRT – Item response theory

MGCFA – multi-group confirmatory factor analysis

MI – measurement invariance

NA – not applicable

PROM – patient-reported outcome measure

1PL model – 1 parameter IRT model

2PL model – 2 parameter IRT model

Instructions

Tick the boxes that need to be completed for the article

	COSMIN Risk of Bias checklist
previously reported	Box 1. PROM development
previously reported ✓	Box 2. Content validity
✓	Box 3. Structural validity
✓	Box 4. Internal consistency
✓	Box 5. Cross-cultural validity\Measurement invariance
✓	Box 6. Reliability
✓	Box 7. Measurement error
	Box 8. Criterion validity
	Box 9. Hypotheses testing for construct validity
✓ partial - "construct approach" 10b and 10c	Box 10. Responsiveness

To assess the methodological quality of each study, i.e. assessing the risk of bias of the result of a study, the corresponding COSMIN Risk of Bias box should be completed. To determine the overall quality of a study the lowest rating of any standard in the box is taken (i.e. “the worst score counts” principle). For example, if for a reliability study one item in a box is rated as ‘inadequate’, the overall methodological quality of that reliability study is rated as ‘inadequate’. The response option ‘NA’ (not applicable) is at issue for some standards. For example, when a study on structural validity is based on CTT, the standard on IRT is not applicable and this standard should not be considered in the “worst score counts”-rating for that specific study. For standards where this option is not at issue, these cells are grey and shouldn’t be used.

Box 3. Structural validity

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

Does the study concern unidimensionality or structural validity? [2] (unidimensionality) / structural validity

Statistical methods	**very good**	**adequate**	**doubtful**	**inadequate**	**NA**
1 For CTT: Was exploratory or confirmatory factor analysis performed?	Confirmatory factor analysis performed	(Exploratory factor analysis performed)		No exploratory or confirmatory factor analysis performed	Not applicable
2 For IRT/Rasch: does the chosen model fit to the research question?	Chosen model fits well to the research question	Assumable that the chosen model fits well to the research question	Doubtful if the chosen model fits well to the research question	Chosen model does not fit to the research question	(Not applicable)
3 Was the sample size included in the analysis adequate?	(FA: 7 times the number of items and ≥100)	FA: at least 5 times the number of items and ≥100; OR at least 6 times number of items but <100	FA: 5 times the number of items but <100	FA: < 5 times the number of items	
	Rasch/1PL models: ≥ 200 subjects	Rasch/1PL models: 100-199 subjects	Rasch/1PL models: 50-99 subjects	Rasch/1PL models: < 50 subjects	
	2PL parametric IRT models OR Mokken scale analysis: ≥ 1000 subjects	2PL parametric IRT models OR Mokken scale analysis: 500-999 subjects	2PL parametric IRT models OR Mokken scale analysis: 250-499 subjects	2PL parametric IRT models OR Mokken scale analysis: < 250 subjects	

Other					
4 Were there any other important flaws in the design or statistical methods of the study?	No other important methodological flaws		Other minor methodological flaws (e.g. rotation method not described)	Other important methodological flaws (e.g. inappropriate rotation method)	

[1] If the scale is not based on a reflective model, unidimensionality or structural validity is not relevant.
[2] In a systematic review, it is helpful to make a distinction between studies where factor analysis is performed on each (sub)scale separately to evaluate whether the (sub)scales are unidimensional (unidimensionality studies) and studies where factor analysis is performed on all items of an instrument to evaluate the (expected) number of subscales in the instrument and the clustering of items within subscales (structural validity studies).

Box 4. Internal consistency

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

Design requirements	**very good**	**adequate**	**doubtful**	**inadequate**	**NA**
1 Was an internal consistency statistic calculated for each unidimensional scale or subscale separately?	Internal consistency statistic calculated for each unidimensional scale or subscale		Unclear whether scale or sub scale is unidimensional	Internal consistency statistic NOT calculated for each unidimensional scale or sub scale	
Statistical methods					
2 For continuous scores: Was Cronbach's alpha or omega calculated?	Cronbach's alpha or Omega calculated		Only item-total correlations calculated	No Cronbach's alpha and no item-total correlations calculated	Not applicable
3 For dichotomous scores: Was Cronbach's alpha or KR-20 calculated?	Cronbach's alpha or KR-20 calculated		Only item-total correlations calculated	No Cronbach's alpha or KR-20 and no item-total correlations calculated	Not applicable
4 For IRT-based scores: Was standard error of the theta (SE (θ)) or reliability coefficient of estimated latent trait value (index of (subject or item) separation) calculated?	SE(θ) or reliability coefficient calculated			SE(θ) or reliability coefficient NOT calculated	Not applicable
Other					
5 Were there any other important flaws in the design or statistical methods of the study?	No other important methodological flaws		Other minor methodological flaws	Other important methodological flaws	

[1] If the scale is not based on a reflective model, internal consistency is not relevant

Box 6. Reliability

Design requirements	very good	adequate	doubtful	inadequate	NA
1 Were patients stable in the interim period on the construct to be measured?	Evidence provided that patients were stable	Assumable that patients were stable	Unclear if patients were stable	Patients were NOT stable	
2 Was the time interval appropriate?	Time interval appropriate		Doubtful whether time interval was appropriate or time interval was not stated	Time interval NOT appropriate	
3 Were the test conditions similar for the measurements? e.g. type of administration, environment, instructions	Test conditions were similar (evidence provided)	Assumable that test conditions were similar	Unclear if test conditions were similar	Test conditions were NOT similar	
Statistical methods					
4 For continuous scores: Was an intraclass correlation coefficient (ICC) calculated?	ICC calculated and model or formula of the ICC is described	ICC calculated but model or formula of the ICC not described or not optimal. Pearson or Spearman correlation coefficient calculated with evidence provided that no systematic change has occurred	Pearson or Spearman correlation coefficient calculated WITHOUT evidence provided that no systematic change has occurred or WITH evidence that systematic change has occurred	No ICC or Pearson or Spearman correlations calculated	Not applicable
5 For dichotomous/nominal/ordinal scores: Was kappa calculated?	Kappa calculated			No kappa calculated	Not applicable

6	For ordinal scores: Was a weighted kappa calculated?	Weighted Kappa calculated		Unweighted Kappa calculated or not described		Not applicable	
7	For ordinal scores: Was the weighting scheme described? e.g. linear, quadratic	Weighting scheme described	Weighting scheme NOT described			Not applicable	
Other							
8	Were there any other important flaws in the design or statistical methods of the study?	No other important methodological flaws		Other minor methodological flaws	Other important methodological flaws		

Box 7. Measurement error

Design requirements	very good	adequate	doubtful	Inadequate	NA
1 Were patients stable in the interim period on the construct to be measured?	Patients were stable (evidence provided)	Assumable that patients were stable	Unclear if patients were stable	Patients were NOT stable	
2 Was the time interval appropriate?	Time interval appropriate		Doubtful whether time interval was appropriate or time interval was not stated	Time interval NOT appropriate	
3 Were the test conditions similar for the measurements? (e.g. type of administration, environment, instructions)	Test conditions were similar (evidence provided)	Assumable that test conditions were similar	Unclear if test conditions were similar	Test conditions were NOT similar	
Statistical methods					
4 For continuous scores: Was the Standard Error of Measurement (SEM), Smallest Detectable Change (SDC) or Limits of Agreement (LoA) calculated?	SEM, SDC, or LoA calculated	Possible to calculate LoA from the data presented		SEM calculated based on Cronbach's alpha, or on SD from another population	Not applicable
5 For dichotomous/nominal/ordinal scores: Was the percentage (positive and negative) agreement calculated?	% positive and negative agreement calculated	% agreement calculated		% agreement not calculated	Not applicable
Other					
6 Were there any other important flaws in the design or statistical methods of the study?	No other important methodological flaws		Other minor methodological flaws	Other important methodological flaws	

10b. Construct approach (i.e. hypotheses testing; comparison with other outcome measurement instruments)					
Design requirements	**very good**	**adequate**	**doubtful**	**inadequate**	**NA**
4 Is it clear what the comparator instrument(s) measure(s)?	Constructs measured by the comparator instrument(s) is clear			Constructs measured by the comparator instrument(s) is not clear	
5 Were the measurement properties of the comparator instrument(s) sufficient?	Sufficient measurement properties of the comparator instrument(s) in a population similar to the study population	Sufficient measurement properties of the comparator instrument(s) but not sure if these apply to the study population	Some information on measurement properties of the comparator instrument(s) in any study population	NO information on the measurement properties of the comparator instrument(s) OR evidence of poor quality of comparator instrument(s)	
Statistical methods					
6 Was the statistical method appropriate for the hypotheses to be tested?	Statistical method was appropriate	Assumable that statistical method were appropriate	Statistical method applied NOT optimal	Statistical method applied NOT appropriate	
Other					
7 Were there any other important flaws in the design or statistical methods of the study?	No other important methodological flaws		Other minor methodological flaws	Other important methodological flaws	

10c. Construct approach: (i.e. hypotheses testing: comparison between subgroups)					
Design requirements	**very good**	**adequate**	**doubtful**	**inadequate**	**NA**
8 Was an adequate description provided of important characteristics of the subgroups?	Adequate description of the important characteristics of the subgroups	Adequate description of most of the important characteristics of the subgroups	Poor or no description of the important characteristics of the subgroups		
Statistical methods					
9 Was the statistical method appropriate for the hypotheses to be tested?	Statistical method was appropriate	Assumable that statistical method was appropriate	Statistical method applied NOT optimal	Statistical method applied NOT appropriate	
Other					
10 Were there any other important flaws in the design or statistical methods of the study?	No other important methodological flaws		Other minor methodological flaws	Other important methodological flaws	