Appendix II COSMIN Checklists

eTable 1 COSMIN risk of bias checklist

PROM Development	Results
1. Is a clear description provided of the construct to be measured?	
2. Is the origin of the construct clear: was a theory, conceptual framework or disease model used or clear rationale provided to define the construct to be measured?	
3. Is a clear description provided of the target population for which the PROM was developed?	
4. Is a clear description provided of the context of use	
5. Was the PROM development study performed in a sample representing the target population for which the PROM was developed?	
6. Was an appropriate qualitative data collection method used to identify relevant items for a new PROM?	
7. Were skilled group moderators/interviewers used?	
8. Were the group meetings or interviews based on an appropriate topic or interview guide?	
9. Were the group meetings or interviews recorded and transcribed verbatim?	
10. Was an appropriate approach used to analyse the data?	
11. Was at least part of the data coded independently?	
12. Was data collection continued until saturation was reached?	
13. For quantitative studies (surveys): was the sample size appropriate?	
14. Was a cognitive interview study or other pilot test conducted?	
15. Was the cognitive interview study or other pilot test performed in a sample representing the target population?	
16. Were patients asked about the comprehensibility of the PROM?	
17. Were all items tested in their final form?	
18. Was an appropriate qualitative method used to assess the comprehensibility of the PROM instructions, items, response options, and recall period?	
19. Was each item tested in an appropriate number of patients?	
20. Were skilled interviewers used?	
21. Were the interviews based on an appropriate interview guide?	
22. Were the interviews recorded and transcribed verbatim?	
23. Was an appropriate approach used to analyse the data?	
24. Were at least two researchers involved in the analysis?	
25. Were problems regarding the comprehensibility of the PROM instructions, items, response options, and recall period appropriately addressed by adapting the PROM?	
26. Were patients asked about the comprehensiveness of the PROM?	
27. Was the final set of items tested?	
28. Was an appropriate method used for assessing the comprehensiveness of the PROM?	
29. Was each item tested in an appropriate number of patients?	
30. Were skilled interviewers used?	
31. Were the interviews based on an appropriate interview guide?	
32. Were the interviews recorded and transcribed verbatim?	

33. Was an appropriate approach used to analyse the data?	
34. Were at least two researchers involved in the analysis?	
35. Were problems regarding the comprehensiveness of the PROM appropriately addressed by adapting the PROM?	
Content validity	
1. Was an appropriate method used to ask patients whether each item is relevant for their experience with the condition?	
2. Was each item tested in an appropriate number of patients?	
3. Were skilled group moderators/interviewers used?	
4. Were the group meetings or interviews based on an appropriate topic or interview guide?	
5. Were the group meetings or interviews recorded and transcribed verbatim?	
6. Was an appropriate approach used to analyse the data?	
7. Were at least two researchers involved in the analysis?	
8. Was an appropriate method used for assessing the comprehensiveness of the PROM?	
9. Was each item tested in an appropriate number of patients?	
10. Were skilled group moderators/interviewers used?	
11. Were the group meetings or interviews based on an appropriate topic or interview guide?	
12. Were the group meetings or interviews recorded and transcribed verbatim?	
13. Was an appropriate approach used to analyse the data?	
14. Were at least two researchers involved in the analysis?	
15. Was an appropriate qualitative method used for assessing the comprehensibility of the PROM instructions, items, response options, and recall period?	
16. Was each item tested in an appropriate number of patients?	
17. Were skilled group moderators/interviewers used?	
18. Were the group meetings or interviews based on an appropriate topic or interview guide?	
19. Were the group meetings or interviews recorded and transcribed verbatim?	
20. Was an appropriate approach used to analyse the data?	
21. Were at least two researchers involved in the analysis?	
22. Was an appropriate method used to ask professionals whether each item is relevant for the construct of interest?	
23. Were professionals from all relevant disciplines included?	
24. Was each item tested in an appropriate number of professionals?	
25. Was an appropriate approach used to analyse the data?	
26. Were at least two researchers involved in the analysis?	
27. Was an appropriate method used for assessing the comprehensiveness of the PROM?	
28. Were professionals from all relevant disciplines included?	
29. Was each item tested in an appropriate number of professionals?	
30. Was an appropriate approach used to analyse the data?	
31. Were at least two researchers involved in the analysis?	
Structural validity	

1. For CTT: Was exploratory or confirmatory factor analysis performed?	
2. For IRT/Rasch: does the chosen model fit to the research question?	
3. Was the sample size included in the analysis adequate?	
4. Were there any other important flaws in the design or statistical methods of the study?	
Internal consistency	
1. Was an internal consistency statistic calculated for each unidimensional scale or subscale separately?	
2. For continuous scores: Was Cronbach's alpha or omega calculated?	
3. For dichotomous scores: Was Cronbach's alpha or KR-20 calculated?	
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?	
5. Were there any other important flaws in the design or statistical methods of the study?	
Cross-cultural validity	
1. Were the samples similar for relevant characteristics except for the group variable?	
2. Was an appropriate approach used to analyse the data?	
3. Was the sample size included in the analysis adequate?	
4. Were there any other important flaws in the design or statistical methods of the study?	
Reliability	
1. Were patients stable in the interim period on the construct to be measured?	
2. Was the time interval appropriate?	
3. Were the test conditions similar for the measurements? e.g. type of administration, environment, instructions	
4. For continuous scores: Was an intraclass correlation coefficient (ICC) calculated?	
5. For dichotomous/nominal/ordinal scores: Was kappa calculated?	
6. For ordinal scores: Was a weighted kappa calculated?	
7. For ordinal scores: Was the weighting scheme described? e.g. linear, quadratic	
8. Were there any other important flaws in the design or statistical methods of the study?	
Measurement error	
1. Were patients stable in the interim period on the construct to be measured?	
2. Was the time interval appropriate?	
3. Were the test conditions similar for the measurements? (e.g. type of administration, environment, instructions)	
4. For continuous scores: Was the Standard Error of Measurement (SEM), Smallest Detectable Change (SDC) or Limits of Agreement (LoA) calculated?	
5. For dichotomous/nominal/ordinal scores: Was the percentage (positive and negative) agreement calculated?	
6. Were there any other important flaws in the design or statistical methods of the study?	
Criterion validity	
1. For continuous scores: Were correlations, or the area under the receiver operating curve calculated?	
2. For dichotomous scores: Were sensitivity and specificity determined?	
3. Were there any other important flaws in the design or statistical methods of the study?	
Hypotheses testing for construct validity	

1. Is it clear what the comparator instrument(s) measure(s)?	
2. Were the measurement properties of the comparator instrument(s) sufficient?	
3. Was the statistical method appropriate for the hypotheses to be tested?	
4. Were there any other important flaws in the design or statistical methods of the study?	
5. Was an adequate description provided of important characteristics of the subgroups?	
6. Was the statistical method appropriate for the hypotheses to be tested?	
7. Were there any other important flaws in the design or statistical methods of the study?	
Responsiveness	
1. For continuous scores: Were correlations between change scores, or the area under the Receiver Operator Curve (ROC) curve calculated?	
2. For dichotomous scales: Were sensitivity and specificity (changed versus not changed) determined?	
3. Were there any other important flaws in the design or statistical methods of the study?	
4. Is it clear what the comparator instrument(s) measure(s)?	
5. Were the measurement properties of the comparator instrument(s) sufficient?	
6. Was the statistical method appropriate for the hypotheses to be tested?	
7. Were there any other important flaws in the design or statistical methods of the study?	
8. Was an adequate description provided of important characteristics of the subgroups?	
9. Was the statistical method appropriate for the hypotheses to be tested?	
10. Were there any other important flaws in the design or statistical methods of the study?	
11. Was an adequate description provided of the intervention given?	
12. Was the statistical method appropriate for the hypotheses to be tested?	
13. Were there any other important flaws in the design or statistical methods of the study?	

eTable 2 Criteria for good measurement properties

Measurement property	Rating	Criteria
Structural validity	+	**CTT:** CFA: CFI or TLI or comparable measure >0.95 OR RMSEA <0.06 OR SRMR <0.082 **IRT/Rasch:** No violation of unidimensionality3: CFI or TLI or comparable measure >0.95 OR RMSEA <0.06 OR SRMR <0.08 AND no violation of local independence: residual correlations among the items after controlling for the dominant factor <0.20 OR Q3's < 0.37 AND no violation of monotonicity: adequate looking graphs OR item scalability >0.30 AND adequate model fit: IRT: $\chi 2$ >0.01 Rasch: infit and outfit mean squares ≥ 0.5 and ≤ 1.5 OR Z standardized values >-2 and <2
	?	CTT: Not all information for '+' reported IRT/Rasch: Model fit not reported
	-	Criteria for '+' not met
Internal consistency	+	At least low evidence for sufficient structural validity AND Cronbach's alpha(s) ≥ 0.70 for each unidimensional scale or subscale
	?	Criteria for "At least low evidence4 for sufficient structural validity" not met
	-	At least low evidence for sufficient structural validity AND Cronbach's alpha(s) < 0.70 for each unidimensional scale or subscale
Reliability	+	ICC or weighted Kappa ≥ 0.70
	?	ICC or weighted Kappa not reported
	-	ICC or weighted Kappa < 0.70
Measurement error	+	SDC or LoA < MIC
	?	MIC not defined
	-	SDC or LoA > MIC
Hypotheses testing for construct validity	+	The result is in accordance with the hypothesis
	?	No hypothesis defined (by the review team)
	-	The result is not in accordance with the hypothesis
Cross-cultural	+	No important differences found between group factors (such

validity\measurement invariance		as age, gender, language) in multiple group factor analysis OR no important DIF for group factors (McFadden's R2 < 0.02)
	?	No multiple group factor analysis OR DIF analysis performed
	-	Important differences between group factors OR DIF was found
Criterion validity	+	Correlation with gold standard ≥ 0.70 OR AUC ≥ 0.70
	?	Not all information for '+' reported
	-	Correlation with gold standard < 0.70 OR AUC < 0.70
Responsiveness	+	The result is in accordance with the hypothesis7 OR AUC ≥ 0.70
	?	No hypothesis defined (by the review team)
	-	The result is not in accordance with the hypothesis7 OR AUC < 0.70

AUC: area under the curve; CFA: confirmatory factor analysis; CFI: comparative fit index; CTT: classical test theory; DIF: differential item functioning; ICC: intraclass correlation coefficient; IRT: item response theory; LoA: limits of agreement; MIC: minimal important change; RMSEA: Root Mean Square Error of Approximation; SEM: Standard Error of Measurement; SDC: smallest detectable change; SRMR: Standardized Root Mean Residuals; TLI: Tucker-Lewis index; "+": sufficient; "-": insufficient; "?": indeterminate.

eTable 3 Modified GRADE approach for assessing certainty of evidence*

Domain	Grade	Reason
Risk of bias	-0 level: No	There are multiple studies of at least adequate quality, or there is one study of very good quality available
	-1 level: Serious	There are multiple studies of doubtful quality available, or there is only one study of adequate quality
	-2 level: Very serious	There are multiple studies of inadequate quality, or there is only one study of doubtful quality available
	-3 level: Extremely serious	There is only one study of inadequate quality available
Inconsistency	-0 level: No	There is no inconsistency among pooled studies or there is only one study in subgroups
	-1 level: Serious	There are severe inconsistencies among pooled studies
	-2 level: Very serious	There are very severe inconsistencies among pooled studies.
Imprecision	-0 level: No	Total sample size>50-100
	-1 level: Serious	Total sample size=50-100
	-2 level: Very serious	Total sample size <50
Indirectness	-0 level: No	There is no indirectness between results and conclusion
	-1 level: Serious	There is severe indirectness between results and conclusion
	-2 level: Very serious	There is very severe indirectness between results and conclusion

*The starting point of quality level is high evidence. The quality of evidence is subsequently downgraded to moderate, low, or very low evidence.