COSMIN content validity methodology

Instructions for completing the COSMIN boxes for content validity

1 CHECK the COSMIN website if the quality of the PROM development was already rated in another review. In that case, you can skip box 1A and use the quality
2 We recommend to score all PROMS with two raters, indepedently, and reach consensus afterwards. You can change "rater 1" and "rater 2" into the names of the raters
4 Add extra rows, columns or tables if needed
5 Tables 1, 2, and 3 will be filled automatically (you may need to add links to the other tabs). They can be included in a systematic review

COSMIN box 1. Standards for evaluating the quality of PROM development

Check the COSMIN website to see if the quality of the PROM development was already rated in another review

Ratings: V= very good; A = adequate; D = doubtful; I = inadequate; N= not applicable

1a. PROM design

	PROM ref			PROM ref			PROM ref		
General design requirements	Rater 1	Rater 2	Consensus	Rater 1	Rater 2	Consensus	Rater 1	Rater 2	Consensus
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									
4 Is a clear description provided of the context of use (i.e. discriminative, evaluative									
5 Was the PROM development study performed in a sample representing the target population for which the PROM was developed?									
Concept elicitation (relevance and comprehensiveness)	Rater 1	Rater 2	Consensus	Rater 1	Rater 2	Consensus	Rater 1	Rater 2	Consensus
6 Was an appropriate qualitative data collection method used to identify relevant items for									
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: was the sample size appropriate?									
SUBTOTAL QUALITY CONCEPT ELICITATION STUDY *Lowest score of items 6-13*									
TOTAL QUALITY OF THE PROM DESIGN *Lowest score of items 1-13*									

1b. Cognitive interview study or other pilot test

	Rater 1	Rater 2	Consensus	Rater 1	Rater 2	Consensus	Rater 1	Rater 2	Consensus
14 Was a cognitive interview study or other pilot test performed? *If NO skip items 15-35*									
General design requirements	Rater 1	Rater 2	Consensus	Rater 1	Rater 2	Consensus	Rater 1	Rater 2	Consensus
15 Was the cognitive interview study or other pilot test performed in a sample representing the target population?									
Comprehensibility	Rater 1	Rater 2	Consensus	Rater 1	Rater 2	Consensus	Rater 1	Rater 2	Consensus
16 Were patients asked about the comprehensibility of the PROM? *If NO or not clear, skip*									
	Rater 1	Rater 2	Consensus	Rater 1	Rater 2	Consensus	Rater 1	Rater 2	Consensus
17 Were all items tested in their final form?									
18 Was an appropriate qualitative method used to assess the comprehensibility of the PROM									
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?									
	SUBTOTAL QUALITY OF COMPREHENSIBILITY STUDY *Lowest score of items 15-25*									
	Comprehensiveness	Rater 1	Rater 2	Consensus	Rater 1	Rater 2	Consensus	Rater 1	Rater 2	Consensus
26	Were patients asked about the comprehensiveness of the PROM? *If NO or not clear, skip items 27-35*									
		Rater 1	Rater 2	Consensus	Rater 1	Rater 2	Consensus	Rater 1	Rater 2	Consensus
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?									
	SUBTOTAL QUALITY OF COMPREHENSIVENESS STUDY *Lowest score of items 15, 26-35*									
	TOTAL QUALITY OF THE PILOT STUDY *Lowest score of items 14-35*									
	TOTAL QUALITY OF THE PROM DEVELOPMENT STUDY *Lowest score of items 1-35*									

COSMIN box 2. Standards for evaluating the quality of content validity studies of PROMs

Only those parts of the box need to be completed for which information is available

Score: V= very good; A = adequate; D = doubtful; I = inadequate; N= not applicable

	PROM			PROM			PROM		
	ref			ref			ref		
2a. Asking patient about relevance	rater 1	rater 2	Consensus	rater 1	rater 2	Consensus	rater 1	rater 2	Consensus
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?									
SUBTOTAL QUALITY OF RELEVANCE STUDY *Lowest score of items 1-7*									
2b. Asking patients about comprehensiveness	rater 1	rater 2	Consensus	rater 1	rater 2	Consensus	rater 1	rater 2	Consensus
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?									
SUBTOTAL QUALITY OF COMPREHENSIVENESS STUDY *Lowest score of items 8-14*									
2c. Asking patients about comprehensibility	rater 1	rater 2	Consensus	rater 1	rater 2	Consensus	rater 1	rater 2	Consensus
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?									
SUBTOTAL QUALITY OF COMPREHENSIBILITY STUDY *Lowest score of items 15-21*									
2d. Asking professionals about relevance	rater 1	rater 2	Consensus	rater 1	rater 2	Consensus	rater 1	rater 2	Consensus
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?									
SUBTOTAL QUALITY OF RELEVANCE STUDY *Lowest score of items 22-26*									

2e. Asking professionals about comprehensiveness	rater 1	rater 2	Consensus	rater 1	rater 2	Consensus	rater 1	rater 2	Consensus
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?									
SUBTOTAL QUALITY OF COMPREHENSIVENESS STUDY *Lowest score of items 27-31*									

Rating the content validity of the PROM

Complete one tabel per PROM (subscale)

Criteria for content validity

To fill in ratings use apostrophe (') before the + / - / ± / ? signs

Score: + = sufficient; - = insufficient; ? = indeterminate; ± = inconsistent

PROM (subscale)	PROM development study	PROM development study	PROM development study	Content validity study 1	Content validity study 1	Content validity study 1	Content validity study 2[2]	Content validity study 2[2]	Content validity study 2[2]	Rating of reviewers	Rating of reviewers	Rating of reviewers	OVERALL RATINGS PER PROM[3]	OVERALL RATINGS PER PROM[3]	OVERALL RATINGS PER PROM[3]	QUALITY OF EVIDENCE	QUALITY OF EVIDENCE	QUALITY OF EVIDENCE
	(+ / - / ± / ?)	(+ / - / ± / ?)	(+ / - / ± / ?)	(+ / - / ± / ?)	(+ / - / ± / ?)	(+ / - / ± / ?)	(+ / - / ± / ?)	(+ / - / ± / ?)	(+ / - / ± / ?)	(+ / - / ± / ?)	(+ / - / ± / ?)	(+ / - / ± / ?)	+ / - / ±	+ / - / ±	+ / - / ±	High, moderate, low, very low	High, moderate, low, very low	High, moderate, low, very low
	rater 1	rater 2	consensus	rater 1	rater 2	consensus	rater 1	rater 2	consensus	rater 1	rater 2	consensus	rater 1	rater 2	consensus	rater 1	rater 2	consensus
Relevance																		
1 Are the included items relevant for the construct of interest?[1]																		
2 Are the included items relevant for the target population of interest?[1]																		
3 Are the included items relevant for the context of use of interest?[1]																		
4 Are the response options appropriate?																		
5 Is the recall period appropriate?																		
RELEVANCE RATING (+ / - / ± / ?)																		
Comprehensiveness																		
6 Are all key concepts included?																		
COMPREHENSIVENESS RATING (+ / - / ± / ?)																		
Comprehensibility																		
7 Are the PROM instructions understood by the population of interest as intended?																		
8 Are the PROM items and response options understood by the population of interest as intended?																		
9 Are the PROM items appropriately worded?																		
10 Do the response options match the question?																		
COMPREHENSIBILITY RATING (+ / - / ± / ?)																		
CONTENT VALIDITY RATING (+ / - / ± / ?)																		

[1] These criteria refer to the construct, population, and context of use of interest in the systematic review.

[2] Add more columns if more content validity studies are available

[3] If ratings are inconsistent between studies, consider using separate tables for subgroups of studies with consistent results.

COSMIN Risk of Bias checklist

Only those parts of the boxes need to be completed for which information is available

Score: V= very good; A = adequate; D = doubtful; I = inadequate; N= not applicable

	PROM			PROM			PROM		
Article reference:	ref			ref			ref		
3. Structural validity	rater 1	rater 2	Consensus	rater 1	rater 2	Consensus	rater 1	rater 2	Consensus
unidimensionality or 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?									
TOTAL *Lowest score of items 1-4*			V						
4. Internal consistency	rater 1	rater 2	Consensus	rater 1	rater 2	Consensus	rater 1	rater 2	Consensus
1 Was an internal consistency statistic calculated for each unidimensional (sub)scale 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									
5 Were there any other important flaws?									
TOTAL *Lowest score of items 1-5*			V						
5. Cross-cultural validity\measurement invariance	rater 1	rater 2	Consensus	rater 1	rater 2	Consensus	rater 1	rater 2	Consensus
1 Were the samples similar for relevant characteristics except for the group									
2 Was an adequate approach used to analyse the data?									
3 Was the sample size included in the analysis adequate?									
4 Were there any other important flaws?									
TOTAL *Lowest score of items 1-4*			V						
6. Reliability	rater 1	rater 2	Consensus	rater 1	rater 2	Consensus	rater 1	rater 2	Consensus
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?									
TOTAL *Lowest score of items 1-8*			V						
7. Measurement error	rater 1	rater 2	Consensus	rater 1	rater 2	Consensus	rater 1	rater 2	Consensus
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?								
	TOTAL *Lowest score of items 1-6*			v					

8. Criterion validity	rater 1	rater 2	Consensus	rater 1	rater 2	Consensus	rater 1	rater 2	Consensus
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?									
TOTAL *Lowest score of items 1-3*			v						

9. Hypotheses testing for construct validity									
9a. Comparison with other outcome measurement instruments (convergent validity)	rater 1	rater 2	Consensus	rater 1	rater 2	Consensus	rater 1	rater 2	Consensus
1 Is it clear what the comparator instrument(s) measure(s)?									
2 Were the measurement properties of the comparator instrument(s) adequate?									
3 Was the statistical method appropriate for the hypotheses to be tested?									
4 Were there any other important flaws?									
TOTAL *Lowest score of items 1-4*			v						
9b. Comparison between subgroups (discriminative or known-groups validity)	rater 1	rater 2	Consensus	rater 1	rater 2	Consensus	rater 1	rater 2	Consensus
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?									
TOTAL *Lowest score of items 5-7*			v						

10. Responsiveness									
10a. Criterion approach (i.e. comparison to a gold standard)	rater 1	rater 2	Consensus	rater 1	rater 2	Consensus	rater 1	rater 2	Consensus
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?									
TOTAL *Lowest score of items 1-3*			v						
10b. Construct approach (i.e. hypotheses testing; comparison with other outcome m	rater 1	rater 2	Consensus	rater 1	rater 2	Consensus	rater 1	rater 2	Consensus
4 Is it clear what the comparator instrument(s) measure(s)?									
5 Were the measurement properties of the comparator instrument(s) adequate?									
6 Was the statistical method appropriate for the hypotheses to be tested?									
7 Were there any other important flaws?									
TOTAL *Lowest score of items 4-7*			v						
10c. Construct approach: (i.e. hypotheses testing: comparison between subgroups)	rater 1	rater 2	Consensus	rater 1	rater 2	Consensus	rater 1	rater 2	Consensus
8 Was an adequate description provided of important characteristics of the									
9 Was the statistical method appropriate for the hypotheses to be tested?									
10 Were there any other important flaws?									

TOTAL *Lowest score of items 8-10*			v						
10d. Construct approach: (i.e. hypotheses testing: before and after intervention)	rater 1	rater 2	Consensus	rater 1	rater 2	Consensus	rater 1	rater 2	Consensus
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?									
TOTAL *Lowest score of items 11-13*			v						

Rating the measurement properties of the PROM

Use one Table per PROM
Add additional columns (studies) if necessary

PROM	Study 1			Study 2			Study 3			OVERALL					
	RATING	RATING	RATING	RATING	RATING	RATING	RATING	RATING	RATING	OVERALL RATING	OVERALL RATING	OVERALL RATING	QUALITY OF EVIDENCE	QUALITY OF EVIDENCE	QUALITY OF EVIDENCE
	+/-/?	+/-/?	+/-/?	+/-/?	+/-/?	+/-/?	+/-/?	+/-/?	+/-/?	+/-/±/?	+/-/±/?	+/-/±/?	High, moderate, low, very low	High, moderate, low, very low	High, moderate, low, very low
	rater 1	rater 2	consensus	rater 1	rater 2	consensus	rater 1	rater 2	consensus	rater 1	rater 2	consensus	rater 1	rater 2	consensus
Structural validity															
Internal consistency															
Cross-cultural validity															
Measurement invariance															
Reliability															
Measurement error															
Criterion validity															
Construct validity															
Responsiveness															

PROM	Study 1			Study 2			Study 3			OVERALL					
	RATING	RATING	RATING	RATING	RATING	RATING	RATING	RATING	RATING	OVERALL RATING	OVERALL RATING	OVERALL RATING	QUALITY OF EVIDENCE	QUALITY OF EVIDENCE	QUALITY OF EVIDENCE
	+/-/?	+/-/?	+/-/?	+/-/?	+/-/?	+/-/?	+/-/?	+/-/?	+/-/?	+/-/±/?	+/-/±/?	+/-/±/?	High, moderate, low, very low	High, moderate, low, very low	High, moderate, low, very low
	rater 1	rater 2	consensus	rater 1	rater 2	consensus	rater 1	rater 2	consensus	rater 1	rater 2	consensus	rater 1	rater 2	consensus
Structural validity															
Internal consistency															
Cross-cultural validity															
Measurement invariance															
Reliability															
Measurement error															
Criterion validity															
Construct validity															
Responsiveness															