Supplementary File 2: COSMIN Definitions and Criteria

COSMIN Taxonomy Measurement property definitions [1]

Measurement Property			**Definition**
Domain	Measurement Property	Aspect of a measurement property	
Reliability			The degree to which the measurement is free from measurement error
Reliability (extended definition)			The extent to which scores for patients who have not changed are the same for repeated measurement under several conditions e.g. using different sets of items from the same outcome measure (internal consistency); over time (test-retest); by different persons on the same occasion (inter-rater); or by the same persons (i.e. raters or responders) on different occasions (intra-rater)
	Internal consistency		The degree of the interrelatedness among the items
	Reliability		The proportion of the total variance in the measurements which is due to 'true'[#] differences between patients
	Measurement Error		The systematic and random error of a patients score that is not attributed to true changes in the construct to be measured.
Validity			The degree to which an outcome measure measures the construct(s) it purports to measure
	Content Validity		The degree to which the content of an outcome measure is an adequate reflection of the construct to be measured
		Face validity	The degree to which (the items of) an outcome measure indeed looks as though they are adequate reflection of the construct to the measured
	Construct validity		The degree to which the scores of an outcome measure are consistent with hypotheses (for instance with regard to internal relationships, relationships to scores of other instruments, or differences between relevant groups) based on the assumption that the

			outcome measure validly measures the construct to be measured
		Structural validity	The degree to which the scores of an outcome measure are an adequate reflection of the dimensionality of the construct to be measured
		Hypotheses testing	Idem construct validity
		Cross-cultural validity	The degree to which the performance of the items on a translated or culturally adapted outcome measure are an adequate reflection of the performance of the items of the original version of the outcome measure.
	Criterion validity		The degree to which the scores of a outcome measure are an adequate reflection of a 'gold standard'
Responsiveness			The ability of an outcome measure to detect change over time in the construct to measured
	Responsiveness		Idem responsiveness
Interpretability*			Interpretability is the degree to which one can assign qualitative meaning – that is, clinical or commonly understood connotations – to an outcome measure's quantitative scores or change in scores.

[#]The word 'true' must be seen in the context of the CTT, which states that any observation is composed 9of two components – a true score and error associated with the observation. 'True' is the average score that would be obtained if the scale were given a infinite number of times. It refers only to the consistency of the score, and not to its accuracy.
*Interpretability is not considered a measurement property, but an important characteristic of a measurement instrument.

<u>Criteria for Good Measurement Properties</u> [2]

Measurement Property	Rating	Criteria
Structural Validity	+	**CTT**: CFA or TLI or comparable measure >0.95 or RMSEA <0.06 or SRMR ,0.08[a] **IRT/Rasch:** No violation or unidemnsionality[b]: CFI or TLR 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 x^2 >0.001. Rasch: infit and

	? -	outfit means squares ≥ and ≤1.5 OR Z-standardised 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[c] for sufficient structural validity[d] AND Cronbach's alpha(s) ≥0.70 for each unidimensional scale or subscale[e] Criteria for "at least low evidence[c] for sufficient structural validity[d]" not met at least low evidence[c] for sufficient structural validity[d] AND Cronbach's alpha(s) <0.70 for each unidimensional scale or subscale[e].
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[d] MIC not defined SDC or LoA >MIC[d]
Hypothesis testing for construct validity	+ ? -	The result is in accordance with the hypothesis[f] No hypothesis defined (by the review team) The result is not in accordance with the hypothesis[f]
Cross-cultural validity/measurement invariance	+ ? -	No important differences found between group factors (such as age, gender, language) in multiple group factor analysis OR no important DIF for group factors (McFadden's R^2 <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 AYC <0.70 Not all information for + was reported Correlation with gold standard <0.70 or AUC <0.70
Responsiveness	+	The result is in accordance with the hypothesis[f] or AUC ≥ 0.70 No hypothesis defined (by review team) The result is not in accordance with the hypothesis[f] 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 standardised root mean residuals, TLR Tucker-Lewis Index

+ = sufficient

? = indeterminate

- = insufficient

[a]To rate the quality of the summary score, the factor structures should be equal across the studies

[b]Unidimensionality refers to a factor analysis per subscale, while structural validity refers to a factor analysis of a (multidimensional) patient reported outcome measure
[c] As defined by the grading the evidence according to the GRADE approach
[d]This evidence may come from different studies
[e]the criteria Cronbach alpha <0.95 was deleted as this is relevant in the development phase of a PROM and ot when evaluating an existing PROM
[f]The results of all studies should be taken together and it should then be decided if 75% of the results are in accordance with the hypotheses.

References

1. Mokkink LB, Terwee CB, Patrick DL, et al. The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes. *Journal of Clinical Epidemiology* 2010;63(7):737-45. doi: 10.1016/j.jclinepi.2010.02.006
2. Prinsen CAC, Mokkink LB, Bouter LM, et al. COSMIN guideline for systematic reviews of patient-reported outcome measures. *Quality of Life Research* 2018 doi: 10.1007/s11136-018-1798-3