Appendix 3. Criteria for good measurement properties

Measurement property	Rating	Criteria
Structural validity*	+	**CTT:** EFA/PCA: factor loadings of each item on its factor is at least 0.30 *AND* maximum 10% of the items load on more than one factor *AND* minimum explained variance is 50% and structure is in line with the theory about the construct to be measured *OR* results on scree plot or Kaiser criterion (Eigenvalues >1) are in line with the theory about the construct to be measured CFA: CFI or TLI or comparable measure >0.95 *OR* RMSEA <0.06 *OR* SRMR <0.08 **IRT/Rasch:** no violation of unidimensionality: 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 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: χ^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 eveidence for sufficient structural validity" not met
	-	At leats low evidence for sufficient structural validity *AND* Cronbach's alpha(s) <0.70 for each unidimensional scale or subscale
Reliability	+	ICC or (weighthed) kappa or Pearson/Spearman correlation ≥0.70
	?	ICC or (weighthed) kappa or Pearson/Spearman correlation not reported
	-	ICC or (weighthed) kappa or Pearson/Spearman correlation <0.70
Measurement error	+	SDC or LoA <MIC
	?	MIC not defined
	-	SDC or LoA > MIC
Hypotheses testing for construct validity	+	≥75% of the results is in accordance with predefined hypotheses
	?	No hypotheses defined (by the review team)
	-	≥75% of the results is not in accordance with predefined hypotheses
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* AUC ≥0.70
	?	Not all information for '+' reported
	-	Correlation with gold standard <0.70 *OR* AUC <0.70
Responsiveness	+	≥75% of the results is in accordance with predefined hypotheses *OR* AUC ≥0.70

	?	No hypotheses defined (by the review team)
	-	≥75% of the results is not in accordance with predefined hypotheses *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, EFA = exploratory factor analysis, ICC = intraclass correlation coefficient, IRT = item response theory, LoA = limits of agreement, MIC = minimal important change, PCA = principal component analyses, 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

*Standard 1 in Box 3 in the COSMIN Risk of Bias checklist[32] was rated very good if CFA was performed, adequate if EFA was performed, doubtful if PCA was performed and inadequate if none of the previous was performed.