Criteria for good measurement properties

Measurement property	Rating*	Criteria
Content validity (including face validity)	+	All items refer to relevant aspects of the construct to be measured AND are relevant for the target population AND are relevant for the purpose of the measurement instrument AND together comprehensively reflect the construct to be measured
	?	Not all information for '+' reported
	–	Criteria for '+' not met
Structural validity	+	Unidimensionality: EFA: First factor accounts for at least 20% of the variability AND ratio of the variance explained by the first to the second factor greater than 4 OR Bi-factor model: Standardized loadings on a common factor >0.30 AND correlation between individual scores under a bi-factor and unidimensional model >0.90 Structural validity: CFI or TLI or comparable measure >0.95 AND (Root Mean Square Error of Approximation (RMSEA) <0.06 OR Standardized Root Mean Residuals (SRMR)<0.08)
	?	Not all information for '+' reported
	–	Criteria for '+' not met
Internal consistency	+	At least limited evidence for unidimensionality or positive structural validity AND Cronbach's alpha(s) ≥ 0.70 and ≤ 0.95
	?	Not all information for '+' reported OR conflicting evidence for unidimensionality or structural validity OR evidence for lack of unidimensionality or negative structural validity
	–	Criteria for '+' not met
IRT/Rasch analyses	+	At least limited evidence for unidimensionality or positive structural validity AND No evidence for violation of local independence: Rasch: standardized item-person fit residuals between -2.5 and 2.5; OR IRT: residual correlations among the items after controlling for the dominant factor < 0.20 OR Q3's < 0.37 AND no evidence for violation of monotonicity: adequate looking graphs OR item scalability >0.30 AND adequate model fit: Rasch: infit and outfit mean squares ≥ 0.5 and ≤ 1.5 OR Z-standardized values > -2 and <2 OR IRT: G^2 >0.01; Optional additional evidence: Adequate targeting; Rasch: adequate person-item threshold distribution; IRT: adequate threshold range No important DIF for relevant subject characteristics (such as age, gender, education), McFadden's R^2 < 0.02
	?	Model fit not reported
	–	Criteria for '+' not met
Reliability	+	ICC or weighted Kappa ≥ 0.70
	?	ICC or weighted Kappa not reported
	–	Criteria for '+' not met
Measurement error	+	SDC or LoA < MIC
	?	MIC not defined
	–	Criteria for '+' not met
Construct validity	+	At least 75% of the results are in accordance with the hypotheses
	?	No correlations with instrument(s) measuring related construct(s) AND no differences between relevant groups reported
	–	Criteria for '+' not met
Cross-cultural validity	+	No important differences found between language versions in multiple group factor analysis or DIF analysis
	?	Multiple group factor analysis AND DIF analysis not performed
	–	One or more criteria for '+' not met
Criterion validity	+	Convincing arguments that gold standard is "gold" AND correlation with gold standard ≥ 0.70
	?	Not all information for '+' reported
	–	Criteria for '+' not met
Responsiveness		
Responsiveness	+	At least 75% of the results are in accordance with the hypotheses
	?	No correlations with changes in instrument(s) measuring related construct(s) AND no differences between changes in relevant groups reported
	–	Criteria for '+' not met

Modified from Terwee et al. J Clin Epidemiol 2007;60:34-42.

AUC = area under the curve; 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; SEM = Standard Error of Measurement; SDC = smallest detectable change; TLI = Tucker-Lewis index

* + = positive rating, ? = indeterminate rating, – = negative rating