Supplementary Table 1: Quality criteria for rating of measurement properties concerning validity, reliability and responsiveness

Psychometric properties	Definition	Rating	Quality criteria
Reliability	The degree to which the measurement is free from measurement error (Prinsen et al., 2016)	+	ICC OR weighted Kappa *r* >0.70
		-	ICC OR weighted Kappa *r* <0.70
		?	ICC OR weighted kappa not reported
Internal consistency	The extent to which items in a (sub)scale are intercorrelated, thus measuring the same construct (Terwee et al., 2007)	+	(Sub)scale unidimensional AND Cronbach alpha >0.70
		-	(Sub)scale not unidimensional OR Cronbach alpha <0.70
		?	Dimensionality not known OR Cronbach alpha not determined
Content validity	The extent to which the domain of interest is comprehensively sampled by the items in the measurement instrument (Terwee et al., 2007).	+	The target population considers all items in the measurement instrument to be relevant AND considers the tool to be complete
		-	The target population considers all items in the measurement instrument to be irrelevant OR considers the tool to be incomplete
		?	No target population involvement
Discriminant validity	The extent to which results from a test relate to results of another test which measures a different construct (i.e., the ability to discriminate between dissimilar constructs) (Robertson et al., 2017)	+	Area under ROC curve is >0.5
		-	Area under ROC curve is <0.5
		?	Area under ROC curve not determined
Construct / structural validity	The degree to which the scores of a measurement instrument are an adequate reflection of the dimensionality of the construct to be measured (Prinsen et al., 2016)	+	Factors should explain at least 50% of the variance
		-	Factors explain <50% of the variance
		?	Explained variance not mentioned

Hypothesis testing for construct validity	The extent to which scores on a particular measurement instrument relate to other measures in a manner that is consistent with theoretically derived hypotheses concerning the concepts that are being measured (Terwee et al., 2007)	+ - ?	At least 75% of the result is in accordance with the hypothesis <75% of the result is not in accordance with the hypothesis No hypothesis defined (by the review team)
Cross-cultural validity/ measurement invariance	The degree to which the performance of the items on a translated or culturally adapted measurement instrument is an adequate reflection of the performance of the items of the original version of the measurement instrument (Prinsen et al., 2016)	+ - ?	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) Important differences between group factors OR DIF was found No multiple group factor analysis OR DIF analysis performed
Criterion/Concurrent validity	The extent to which scores on a particular measurement instrument relate to an alternative, previously validated measure of the same construct (Robertson et al., 2017) or a gold standard (Terwee et al., 2007)	+ - ?	Convincing arguments that gold standard is ''gold'' OR alternative measure has been previously validated AND correlation with gold standard OR alternative measure >0.70 Correlation with gold standard OR alternative measure <0.70 despite adequate design and method No convincing arguments that gold standard is ''gold'' OR alternative measure has been validated OR doubtful design or method
Responsiveness	The ability of a measurement instrument to detect important changes over time (Terwee et al., 2007; Prinsen et al., 2016)	+ - ?	SDC OR SDC < MIC OR MIC outside the LOA OR RR > 1.96 OR AUC> 0.70 SDC OR SDC> MIC OR MIC equals or inside LOA OR RR <1.96 OR AUC <0.70, despite adequate design and methods Doubtful design or method
Floor and ceiling effects	The number of respondents who achieved the lowest or highest possible score (Terwee et al., 2007)	+ -	<15% of the respondents achieved the highest or lowest possible scores; >15% of the respondents achieved the highest or lowest possible scores, despite adequate design an methods

		?	Doubtful design or method

The criteria are based on Robertson et al. (2017), Prinsen et al. (2016), Terwee et al. (2007) and Mokkink et al. (2010)

ICC = Intraclass correlation; ROC = receiver operating characteristic; DIF = differential item functioning; SDC = smallest detectable change; MIC = minimum important change; LOA = limits of agreement; RR = relative risk; ACU = area under the curve.

+ = positive rating; - = negative rating; ? = indeterminate rating.