Appendix 2. Criteria for grading PROs based on the guidelines proposed by Khadka et al.[13]

Content Development	
Item identification	A: Comprehensive consultation with the patients, experts (interviews, FGD) and literature review B: Minimal consultation with appropriate patients, experts and literature review C: No consultation with patients
Item selection	A: Pilot instrument developed and tested with Rasch or factor analysis, Statistical justification for reducing items, Items with floor and ceiling effects removed; missing data considered B: Only some of these techniques C: No pilot instrument, or no statistical justification of the items in the final instrument.
CTT based psychometric properties	
Acceptability[170]	A: The percentage of missing data for majority of items : ≤5% B: The percentage of missing data for majority of items : > 5%; ≤40% C: The percentage of missing data for majority of items: > 40%
Targeting[170, 171]	A: End-point responses ≤5% for majority of items B: End-point responses > 5%; ≤40% for majority of items C: End-point responses > 40% for majority of items
Internal Consistency[174] [172]	A: 0.95 >= Cronbach α ≥ 0.7 B: 0.7 >Cronbach α ≥ 0.6, Or Cronbach α >0.95 C: α < 0.6
Item dependency[170]	A: Inter-item correlations < 0.3 B: Inter-item correlations >= 0.3; < 0.6 C: Inter-item correlations >= 0.6
Dimensionality[12, 173]	A: 1st factor loading >0.4 for all items; Principal component analysis (PCA) - variance explained by the measure > 60% and eigenvalue of the first contrast <2.0 B: Cronbach α : 0.7 < α > 0.9; PCA - variance explained by the measure ≥50% to < 60%; and eigenvalue <2.0 C: Cronbach α : 0.7 > α or α >0.9; PCA- variance explained by the measure <50%; and eigenvalue > 2.0 (indicates multidimensionality)
Rasch based psychometric properties	
Response categories	A: All the categories were ordered or ordering of the categories were obtained after repairing disordered categories and evenly spaced categories B: All the categories were ordered or ordering of the categories were obtained after repairing disordered categories and categories were not evenly spaced. C: Unrepairable disordered categories
Dimensionality	A: PCA of residuals: variance explained by the measure ≥ 60%; and eigenvalue of the first contrast <2.0 B: PCA of residuals : variance explained by the measure ≥50% to < 60%; and eigenvalue <2.0 C. PCA of residuals : variance explained by the measure <50%; and eigenvalue >2.0 (indicates multidimensionality)
Measurement Precision	A: Person separation index (PSI) ≥2.5; Reliability (α)>0.85

	B: PSI 2.0 < 2.50; 0.8=< α < 0.85 C:PSI <2.0; α <0.80
Item fit statistics	A: All items with infit and outfit mean square between 0.70 and 1.30 B: One or two items within the 0.50 and 1.50 limit C: More than two items outside the 0.50 and 1.50 limit
Differential Item functioning (DIF)	A: All items with DIF < 0.5 logits B: Some items 0.5 to 1 logits, and one at the most > 1 logits C: More than one item > 1.0 logit
Targeting	A: Difference between item and person means =< 1 logit B: >1 to =<2 logits C: > 2 logits
Validity	
Convergent validity	A: Tested against appropriate measure, and correlation: (0.3 – 0.9) B: Debatable choice of measure and correlation: (0.3 – 0.9) C: Correlation < 0.3 or > 0.9)
Discriminant validity	A: Tested with an appropriate measure, and correlation: <0.3 B: Debatable choice of measure, and correlation: <0.3 C: Correlation > 0.3
Concurrent validity	A: Tested with an appropriate measure and correlation: 0.3 – 0.9 B: Debatable choice of measure, and correlation: 0.3 – 0.9 C: Correlation < 0.3 or > 0.9
Known group validity	A: Tested between appropriate clinical groups; significant difference between groups. B: Tested between debatable groups; significant difference between groups C: Not tested, or insignificant difference between groups
Repeatability or reproducibility/Responsiveness	
Test retest agreement	A: Intra-class correlation (ICC) ≥ 0.8, B: 0.6 ≤ ICC < 0.8 C: ICC < 0.6
Interobserver/intermode agreement[12]	A: Limits of agreement (LOA) < Minimally important difference (MID), Weighted kappa> 0.8, Intermodal correlation > 0.7 B: LOA broader but still close to MID, Kappa: 0.6 – 0.79, intermodal correlation 0.5-0.7 C: LOA >> MID, kappa <0.6, intermodal correlation < 0.5, or incorrect statistical test or inadequate sample (<30)
Responsiveness	A: Score changes over time > MID, or changes with intervention; Effect size (>=1) or responsiveness statistics given B: Changes over time but relationship to MID nor reported, effect size >=0.5 to <1; small sample, and inadequate time frame C: Score changes ≤ MID; effect size < 0.5

Note: A = Excellent; B = Fair/OK, C = Unsatisfactory