Updated criteria for good measurement properties; scoring form from the consensus-based standards for the selection of health measurement instruments (COSMIN) guidelines (1).

<table>
<thead>
<tr>
<th>Measurement property</th>
<th>Rating^1</th>
<th>Criteria</th>
</tr>
</thead>
</table>
| Structural validity        | +        | **CTT:**  
|                            |          | 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 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:  
|                            |          | $\chi^2 > 0.01$  
|                            |          | 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 evidence for sufficient structural validity not met  
|                            | –        | At least low evidence for sufficient structural validity AND Cronbach’s alpha(s) < 0.70 for each unidimensional scale or subscale  
| 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  
|                            | ?        | MIC not defined SDC or  
|                            | –        | LoA > MIC  
| Hypotheses testing for construct validity | +        | The result is in accordance with the hypothesis  
|                            | ?        | No hypothesis defined (by the review team)  
|                            | –        | The result is not in accordance with the hypothesis  

^1: Criteria for ‘+’ not met

^2: MIC not defined

^3: CFI or TLI or comparable measure >0.95 OR RMSEA <0.06 OR SRMR <0.08

^4: At least low evidence for sufficient structural validity

^5: Cronbach’s alpha(s) ≥ 0.70 for each unidimensional scale or subscale

^6: MIC not defined

^7: The result is in accordance with the hypothesis

**Hypotheses testing for construct validity**

<table>
<thead>
<tr>
<th>Hypotheses testing for construct validity</th>
<th>Rating</th>
<th>Criteria</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>+</td>
<td>The result is in accordance with the hypothesis</td>
</tr>
<tr>
<td></td>
<td>?</td>
<td>No hypothesis defined (by the review team)</td>
</tr>
<tr>
<td></td>
<td>–</td>
<td>The result is not in accordance with the hypothesis</td>
</tr>
<tr>
<td>Cross-cultural validity\measurement invariance</td>
<td>+</td>
<td>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 $R^2 &lt; 0.02$)</td>
</tr>
<tr>
<td>-----------------------------------------------</td>
<td>---</td>
<td>----------------------------------------------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td></td>
<td>?</td>
<td>No multiple group factor analysis OR DIF analysis performed</td>
</tr>
<tr>
<td></td>
<td>−</td>
<td>Important differences between group factors OR DIF was found</td>
</tr>
<tr>
<td>Criterion validity</td>
<td>+</td>
<td>Correlation with gold standard ≥ 0.70 OR AUC ≥ 0.70</td>
</tr>
<tr>
<td></td>
<td>?</td>
<td>Not all information for ‘+’ reported</td>
</tr>
<tr>
<td></td>
<td>−</td>
<td>Correlation with gold standard &lt; 0.70 OR AUC &lt; 0.70</td>
</tr>
<tr>
<td>Responsiveness</td>
<td>+</td>
<td>The result is in accordance with the hypothesis(^7) OR AUC ≥ 0.70</td>
</tr>
<tr>
<td></td>
<td>?</td>
<td>No hypothesis defined (by the review team)</td>
</tr>
<tr>
<td></td>
<td>−</td>
<td>The result is not in accordance with the hypothesis(^7) OR AUC &lt; 0.70</td>
</tr>
</tbody>
</table>

The criteria are based on e.g. Terwee et al.\(^2\) and Prinsen et al.\(^3\)

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: Standardized Root Mean Residuals, TLI = Tucker-Lewis index

\(^{1}\) “+” = sufficient, “−” = insufficient, “?” = indeterminate

\(^{2}\) To rate the quality of the summary score, the factor structures should be equal across studies \(^{3}\) unidimensionality refers to a factor analysis per subscale, while structural validity refers to a factor analysis of a (multidimensional) patient-reported outcome measure

\(^{4}\) As defined by grading the evidence according to the GRADE approach

\(^{5}\) This evidence may come from different studies

\(^{6}\) The criteria ‘Cronbach alpha < 0.95’ was deleted, as this is relevant in the development phase of a PROM and not when evaluating an existing PROM.

\(^{7}\) 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: