**Appendix A. Definition and criteria for adequacy for psychometric properties (Terwee et al., 2007)**

|  |  |  |
| --- | --- | --- |
| **Property** | **Definition** | **Criteria of adequacy a b** |
| 1. Content validity | The degree to which the content of an instrument is an adequate reflection of the construct to be measured | (+) A clear description is provided of the measurement aim, the target population, the concepts that are being measured, and the item selection AND target population and (investigators OR experts) were involved in item selection;(?) A clear description of above-mentioned aspects is lacking OR only target population involved OR doubtful design or method;(-) No target population involvement;(0) No information found on target population involvement. |
| 2. Internal consistency | The extent to which items are intercorrelated, thus measuring the same construct  | (+) Factor analyses performed (or results of past ones taken into consideration) on adequate sample size (7 \* # of items and ≥ 100) AND Cronbach’s alpha(s) between 0.70 and 0.95 for each scale; (?) Cronbach’s alphas presented without factor analysis OR doubtful design or method;(-) Cronbach’s alpha(s) < 0.70 or > 0.95;(0) No information found on internal consistency. |
| 3. Criterion validity | The extent to which scores on a particular questionnaire are an adequate reflection of a ‘gold standard’ | (+) Convincing arguments that gold standard is “gold” AND correlation with gold standard ≥ 0.70;(?) ≥ 0.70 correlation presented without convincing arguments that gold standard is “gold” OR doubtful design or method;(-) Correlation with gold standard < 0.70;(0) No information found on criterion validity. |
| 4. Construct validity | The extent to which scores on a particular questionnaire relate to other measures in a manner that is consistent with theoretically derived hypotheses concerning the concepts that are being measured | (+) Explicitly tested for AND at least 75% of the results are in expected direction and size;(?) Doubtful design or method (e.g., not explicitly tested); (-) Less than 75% of results as expected;(0) No information found on construct validity. |
| 5. Reproducibility 5.1 Agreement | The extent to which the scores on repeated measures are close to each other (absolute measurement error) | (+) SDC or LOA < MIC OR convincing arguments that agreement is acceptable; (?) MIC not defined AND no convincing arguments that agreement is acceptable OR doubtful design or method; (-) SDC > MIC OR MIC equals or inside LOA; (0) No information found on agreement.  |
|  5.2 Reliability | The extent to which patients can be distinguished from each other, despite measurement errors (relative measurement error) | (+) ICC or weighted Kappa > 0.70;(?) Doubtful design or method (e.g., time interval not mentioned or less valid measure then a Kappa used); (-) ICC or weighted Kappa < 0.70; (0) No information found on reliability.  |
| 6. Responsiveness | The ability of a questionnaire to detect clinically important changes over time  | (+) Treatment program outlined and longitudinal expected changes presented AND/OR > 75% of results are as expected OR SDC < MIC OR MIC outside the LOA OR RR > 1.96 OR AUC > 0.70;(?) Doubtful design or method;(-) SDC or SDC > MIC OR MIC equals or inside LOA OR RR < 1.96 OR AUC < 0.70;(0) No information found on responsiveness. |
| 7. Floor and ceiling effects | The number of respondents who achieved the lowest or highest possible score | (+) < 15% of the respondents achieved the highest or lowest possible scores; (?) Doubtful design or method;(-) > 15% of respondents achieved the highest or lowest possible scores, despite adequate design and methods;(0) No information found on interpretation.  |
| 8. Interpretability | The degree to which one can assign qualitative meaning to an instrument’s quantitative scores or change in scores | (+) Mean and SD scores presented of at least four relevant subgroups of patients and MIC defined;(?) Doubtful design or method OR less than four subgroups OR no MIC defined; 1. No information found on interpretation.
 |

MIC = minimal important change; SDC = smallest detectable change; LOA = limits of agreement; ICC = intraclass correlation; AUC = area under the receiver operating characteristics curve; RR = responsiveness ration; SD = standard deviation.

a + = positive rating; ? = intermediate rating; - = negative rating; 0 = no information available.

b Doubtful design or method = lacking a clear description of the design or methods of the study, sample size smaller than 50 subjects (should be at least 50 in every (subgroup) analysis), or any important methodological weakness in the design or execution of the study.