Psychometric Evaluation of Self-Report Questionnaires - the development of a checklist

Sandra DM Bot*    Caroline B Terwee†
Daniëlle AWM van der Windt‡    Lex M Bouter†
Joost Dekker§    Henrica CW de Vet†

In the last decade a large number of self-report questionnaires have been developed, which are designed to assess disease-specific physical functioning (i.e. the performance of daily activities). The choice of which questionnaire to use may be based on the study population, on the purpose of the questionnaire, on practical considerations (e.g. ease of scoring, and time to complete), and on its psychometric performances. Unfortunately, there are only a few publications about the comparison of content and psychometric quality of these questionnaires. Consequently, little evidence is available to guide the clinician and researcher during questionnaire selection. To evaluate the psychometric properties of self-report questionnaires standardized criteria are needed. We developed a checklist to evaluate the psychometric quality of self-report questionnaires in terms of validity, reproducibility, responsiveness, interpretability and practical burden. This checklist was applied in a systematic review of the psychometric properties of shoulder disability questionnaires. For each questionnaire all papers reporting on its psychometric properties were retrieved and submitted to quality assessment using this checklist.

The checklist

A checklist was composed to evaluate and compare the psychometric properties of the questionnaire according to the following steps: (i) evaluation of methods used for development of a self-report questionnaire; (ii) evaluation of the design and conduct of studies reporting on the psychometric properties of the

*Address of correspondence: Ms. S. D. M. Bot, Institute for Research in Extramural Medicine (EMGO Institute), VU University Medical Center, Amsterdam, The Netherlands. E-mail: S.BOT.EMGO@MED.VU.NL
†Institute for Research in Extramural Medicine, VU University Medical Center, Amsterdam, The Netherlands
‡Department of General Practice, VU University Medical Center, Amsterdam, The Netherlands
§Department of Rehabilitation Medicine, VU University Medical Center, Amsterdam, The Netherlands
questionnaire; (iii) rating of the results regarding psychometric properties of the questionnaire; and (iv) summarizing the results of different studies that assess the same questionnaire. Our checklist is partly based on the criteria developed by the Scientific Advisory Committee of the Medical Outcome Trust (Lohr et al., 1996) and the checklist developed by Bombardier and Tugwell (1987). The checklist contains items on validity, reproducibility, responsiveness, interpretability and practical burden.

**Validity**

Establishing validity is essential in the evaluation of a self-report questionnaire. An instrument is of no use when it does not measure what it is supposed to measure. Testing of validity can be done by several methods (L'Insalata et al., 1997). We decided to rate content validity and construct validity. A gold standard for measuring physical functioning is often not available, which means that criterion validity cannot be assessed.

Content validity examines the extent to which the domains of interest are comprehensively sampled by the items in the questionnaire (Guyatt et al., 1993). In order to rate content validity the methods for item selection, item-reduction, and the execution of a pilot study to examine the level of reading and comprehension were evaluated. Item selection can be done by interviewing patients or experts in the field, by reviewing the literature, or by using the investigators' expertise. Items on the questionnaire must reflect areas that are important to patients suffering from the disorder that is being studied. Therefore, studies achieved a positive rating for content validity when at least patients had been involved during item selection.

The internal consistency of a questionnaire can be considered to be another aspect of content validity. Internal consistency was rated by checking if a factor analysis had been performed to explore the dimensional structure of the questionnaire. If an instrument had more than one subscale, a Cronbach's alpha had to be presented for each subscale. A Cronbach's alpha in the range of 0.70-0.90 was considered to be acceptable (Nunnally, 1978; Streiner & Norman, 1995). An alpha exceeding 0.90 may indicate redundancy (Fitzpatrick et al., 1998).

Construct validity refers to the extent to which scores on a particular instrument relate to other measures in a manner that is consistent with theoretically derived hypotheses concerning the constructs that are measured (Kirshner & Guyatt, 1985). We considered construct validity to be adequately tested if hypotheses were specified and the results were in correspondence with these hypotheses. Finally, the presence of floor or ceiling effects was evaluated. Floor and ceiling effects were considered present if more than 15% of respondents achieved the highest or lowest possible score, respectively (McHorney & Tarlov, 1995). Therefore, authors had to provide descriptive statistics for the distribution of scores, which included information on the presence of floor or ceiling effects.
Reproducibility

Reproducibility is the extent to which an instrument is free of measurement error. It was assessed by rating test-retest reliability and agreement. Test-retest reliability refers to the extent to which the same results are obtained on repeated administrations of the same questionnaire when no change in physical functioning is expected (Lohr et al., 1996; De Vet et al., 2001). The period between administrations should be long enough that memory effects can be ignored, though short enough to ensure that clinical change has not occurred. We considered calculation of the intraclass correlation coefficient (ICC) per domain an adequate method for test-retest reliability (Bland & Altman, 1996). An ICC above 0.70 for group comparisons was rated as positive test-retest reliability (Nunnally, 1978; Lohr et al., 1996; Fitzpatrick et al., 1998), accompanied by confidence intervals. Application of Pearson correlation coefficients to estimate test-retest reliability was rated as doubtful, as it neglects systematic errors if present (Deyo et al., 1991; Bland & Altman, 1996). For discriminative instruments, which are used to distinguish between individuals or groups, systematic changes may not be important (Kirshner & Guyatt, 1985). However, for evaluative instruments (i.e. instruments that are developed to measure clinical change over time) agreement should be established. Calculation of the 95 % limits of agreement (Bland & Altman, 1986), the Kappa coefficient (Cohen, 1960) or the Standard Error of Measurement (SEM) were regarded as adequate measures of agreement. Calculation of the percentage of agreement was considered to be inadequate, as it does not adjust for the agreement attributable to chance. It was not possible to define sensible cut-off points for the results regarding agreement. Therefore, a positive rating was given when an adequate method for studying agreement had been used.

Responsiveness

Responsiveness refers to an instrument's ability to detect important change over time in the concept being measured (Testa & Simonson, 1996; De Bruin et al., 1997; Terwee et al., 2003). There is no single agreed method to assess responsiveness. Terwee et al. retrieved 25 definitions and 31 different measures of responsiveness from the literature. All available measures could be subdivided into two groups: (i) measures of treatment effect (e.g. effect size, paired t-test); and (ii) correlation between change scores of different measures, also referred to as longitudinal validity. The limitation of the first alternative is that it gives information about the magnitude of the treatment effect, rather than about the quality of the questionnaire. Furthermore, this method depends not only on the magnitude of the observed change, but also on the size of the study population, and the variability of the questionnaire at issue (Husted et al., 2000). The second method may be better suited to assess responsiveness. This method requires predictions about how the results of the questionnaire should correlate with other related measures. Responsiveness was considered adequate if specific hypotheses had been formulated and when the results were in correspondence with these hypotheses.
Validity, reproducibility and responsiveness depend on the setting and the population in which it is assessed. Therefore, a clear description of the design of each individual psychometric study had to be provided. We recorded if authors presented a clear description of the study population (including diagnosis and clinical features), measurements, testing conditions and analysis of the data. Methodological weaknesses in the design or execution of a study that might have influenced the results or conclusions of the study were recorded.

**Interpretability**

Interpretability may be defined as the degree to which one can assign qualitative meaning to quantitative scores (Lohr et al., 1996). The investigators had to provide information about what (difference in) score would be clinically meaningful. The minimal clinically important difference (MCID) is “the smallest difference in score in the domain of interest which patients perceive as beneficial and would mandate, in the absence of troublesome side effects and excessive cost, a change in the patient’s management” (Jaeschke et al., 1989). Besides a MCID, various types of information can aid in interpreting scores on a questionnaire: (i) presentation of means and standard deviations (SD) of scores before and after treatment; (ii) comparative data on the distribution of scores in relevant subgroups; (iii) information on the relationship of scores to well-known functional measures or to clinical diagnosis; and (iv) information on the association between changes in score and patients’ global ratings of the magnitude of change they have experienced. Investigators had to provide at least two of these types of information for a positive rating of interpretability.

**Practical burden**

Time required for administration was included in the checklist to determine respondent burden. Information should be provided on the average time needed to complete the questionnaire. A positive rating was given when a questionnaire can be completed within ten minutes. Assessment of administrative burden involved the presence of information about scoring and the ease of scoring. The scoring method was rated as easy when the items were simply summed up, moderate when a VAS or simple formula was used, and difficult when either a VAS in combination with a formula, or a complex formula was used.

**Application of the checklist**

Systematic searches resulted in the identification of 28 studies referring to the psychometric characteristics of 16 shoulder disability questionnaires. When applying the checklist to review the psychometric quality of these questionnaires we encountered several dilemmas. For instance, when studies do not give an adequate description of the study design and population characteristics, rating
of the performances of the questionnaire is hampered. The psychometric properties of a questionnaire may vary among different settings and populations (Streiner & Norman, 1995), therefore it is indispensable that a proper description is given. Furthermore, the methods used in the studies should be sound before one can evaluate the results. If this was not the case, we decided to rate the results of these studies as doubtful.

When constructing a questionnaire, one should specify beforehand which dimensions (constructs) it is supposed to measure (i.e. whether the questionnaire will be a unidimensional or multidimensional instrument) and subsequently test this hypothetical dimensional structure by using factor analysis. We found that this was often not properly done, or not done at all, which made it difficult to rate internal consistency. Factor analysis was rarely applied, and when applied the results of the factor analysis were not always used correctly, e.g. some authors used one total score for a questionnaire, even when factor analysis clearly showed multidimensionality. The internal consistency (Cronbach’s alpha) cannot be interpreted properly when the dimensionality of the questionnaire is not analyzed.

Several aspects of construct validity are subject to discussion. To examine construct validity the correlation of the questionnaire with other instruments is usually assessed. There are no agreed standards as to how strong such a correlation should be in order to establish adequate construct validity. In addition, it is indispensable to formulate hypotheses before validity testing. These hypotheses should specify both the magnitude and direction of the expected correlation. The more hypotheses are specified, the better, as confirmation of these hypotheses gives support to the validity of the questionnaire. However, a hypothesis may be rejected because of a wrong presumption, while the questionnaire is valid. Taking these aspects into account, it is not feasible to establish clear-cut guidelines for rating construct validity. It is evident that the same applies to responsiveness.

The presence of floor and ceiling effects may influence the responsiveness of an instrument to detect clinically relevant change. If floor effects are present the effect of an intervention will be missed for individuals that occupy the lower levels of the scale even before the intervention (i.e. the instrument is not capable of demonstrating an improvement in score when the patient clinically improves). The cut-off value we used (15% of respondents achieving the highest or lowest score) is arbitrary and may be questionable. Authors should at least provide descriptive statistics about the distribution of scores. In addition, as floor and ceiling effects are dependent upon the population being studied, a description of demographics and clinical characteristics of the study population is required.

We considered calculation of the intraclass correlation coefficient (ICC) per domain as an adequate method for test-retest reliability. Several studies have stated than an ICC should be above 0.70 for an adequate test re-test validity at group comparisons (Nunnally, 1978; Lohr et al., 1996; Fitzpatrick et al., 1998). Authors should provide confidence intervals around the ICC to indicate
the degree of uncertainty in the estimation of the reliability coefficient. This is important because often small sample sizes (< 40) are used for studying test-retest reliability. When statistical estimates are derived from very small populations, they may be inaccurate, and confidence intervals will be wide. Therefore it might be better to consider the lower limit of the confidence interval around the ICC to be above 0.70. When questionnaires are used for clinical assessment of individual patients higher measurement standards are demanded than when questionnaires are used for research purposes in groups of patients. For individual comparisons an ICC of at least 0.90 should be required (Hays et al., 1993; McHorney & Tarlov, 1995).

For evaluative instruments, reliability should be demonstrated with a measure of agreement. We could not establish sensible cut-off points for the results of the agreement studies, as these depend on the definition of a clinically relevant difference. Therefore, studies should present measures of agreement as well as the MCID.

In our review only five out of 28 studies paid attention to interpretation of the results of a questionnaire, and a MCID was stated for only three questionnaires. When investigators do not provide an indication of how to interpret the (changes in) score, the findings are of limited use to clinicians (Guyatt, 2000). Information about the meaning of scores can be gathered by correlating the scores with those of other well-known measures or by comparing scores between known groups. Among others, Lydick and Epstein have described different approaches for interpretation of HRQOL-changes (Lydick & Epstein, 1993). It should be recognized that the interpretation of the results is questionable when the psychometric quality of an instrument is unknown or has not been adequately tested. The more frequently an instrument is tested, and the more situations in which it performs as expected, the greater our confidence in its psychometric properties. The use of various methods and various populations helps 'building' these properties. However, this complicates the rating, as it is not easy to summarize the results when more than one study is found on a single questionnaire, especially when the results are contradictory.

The cut-off value for practical burden (i.e. time needed to complete the questionnaire, ease of scoring) is arbitrary and can be changed in another value dependent on the importance one attaches on it.

Conclusions

To assess the quality of self-report questionnaires we developed a checklist that contained five dimensions (i.e. validity, reproducibility, responsiveness, interpretability and practical burden). We have the impression that these are indeed the essential concepts (see Guyatt et al., 1993; Streiner & Norman, 1995; Fitzpatrick et al., 1998). For each dimension, we developed criteria to evaluate the methodological quality of the instrument development/evaluation. The criteria we used may be disputed. However, it was not our intention to create a definite checklist, but to take the first step in creating a checklist for evaluating the
psychometric properties of self-report questionnaires and to stimulate further development of it. Much more discussion is needed about which criteria to use for such a checklist, and how each step in the process of quality appraisal can be taken. Guidelines are needed to set standards and define the criteria by which these instruments should be assessed (Fletcher, 1995). In addition, a relatively new method to develop and evaluate health status questionnaires is item response theory (IRT). In the future, the checklist should also accommodate an evaluation of IRT methods and results.

Finally, we found that a lot of authors do not provide enough information about the development and psychometric performance of the questionnaire to make an adequate quality assessment possible. The evaluation of psychometric studies depend on complete and accurate reporting, therefore it is necessary to develop a standard on the reporting of psychometric studies analogous to for instance the CONSORT statement on the reporting of randomized controlled trials (Begg et al., 1996). Such a statement can help authors to improve reporting by use of a checklist and flow diagram.

References


