ELSEVIER

Journal of Clinical Epidemiology 60 (2007) 34−42

Journal of Clinical Epidemiology

Quality criteria were proposed for measurement properties of health status questionnaires

Caroline B. Terwee[a,*], Sandra D.M. Bot[a], Michael R. de Boer[a,b], Daniëlle A.W.M. van der Windt[a,c], Dirk L. Knol[a,d], Joost Dekker[a,e], Lex M. Bouter[a], Henrica C.W. de Vet[a]

[a]*EMGO Institute, VU University Medical Center, Van der Boechorststraat 7, 1081 BT Amsterdam, The Netherlands*
[b]*Department of Ophthalmology, VU University Medical Center, Amsterdam, The Netherlands*
[c]*Department of General Practice, VU University Medical Center, Amsterdam, The Netherlands*
[d]*Department of Clinical Epidemiology and Biostatistics, VU University Medical Center, Amsterdam, The Netherlands*
[e]*Department of Rehabilitation Medicine, VU University Medical Center, Amsterdam, The Netherlands*

Accepted 29 March 2006

Abstract

Objectives: Recently, an increasing number of systematic reviews have been published in which the measurement properties of health status questionnaires are compared. For a meaningful comparison, quality criteria for measurement properties are needed. Our aim was to develop quality criteria for design, methods, and outcomes of studies on the development and evaluation of health status questionnaires.

Study Design and Setting: Quality criteria for content validity, internal consistency, criterion validity, construct validity, reproducibility, longitudinal validity, responsiveness, floor and ceiling effects, and interpretability were derived from existing guidelines and consensus within our research group.

Results: For each measurement property a criterion was defined for a positive, negative, or indeterminate rating, depending on the design, methods, and outcomes of the validation study.

Conclusion: Our criteria make a substantial contribution toward defining explicit quality criteria for measurement properties of health status questionnaires. Our criteria can be used in systematic reviews of health status questionnaires, to detect shortcomings and gaps in knowledge of measurement properties, and to design validation studies. The future challenge will be to refine and complete the criteria and to reach broad consensus, especially on quality criteria for good measurement properties. © 2006 Elsevier Inc. All rights reserved.

Keywords: Reproducibility; Reliability; Validity; Responsiveness; Guidelines; Criteria

1. Introduction

The number of available health status questionnaires has increased dramatically over the past decades. Consequently, the choice of which questionnaire to use is becoming a major difficulty. Recently a large number of systematic reviews have been published of available questionnaires measuring a specific concept in a specific population, for example [1−11]. In these systematic reviews, typically, the content and measurement properties of the available questionnaires are compared. In analogy to systematic reviews of clinical trials, criteria are needed to determine the methodological quality of studies on the development and evaluation of health status questionnaires. In addition, criteria for good measurement properties are needed to legitimize what the best questionnaire is.

Several articles offer criteria for the evaluation of questionnaires. Probably the best-known and most comprehensive criteria are those from the Scientific Advisory Committee (SAC) of the Medical Outcomes Trust [12]. The SAC defined eight attributes of instrument properties that warrant consideration in evaluation. These include (1) conceptual and measurement model, (2) validity, (3) reliability, (4) responsiveness, (5) interpretability, (6) respondent and administrative burden, (7) alternative forms, and (8) cultural and language adaptations (translations). Within each of these attributes, specific criteria were defined by which instruments should be reviewed. Similar criteria have been defined, e.g., by Bombardier and Tugwell [13], Andresen [14], and McDowell and Jenkinson [15]. What is often lacking in these criteria, however, are explicit criteria for what constitutes good measurement properties. For example, for the

[*] Corresponding author. Tel.: +31-20-4448187; fax: +31-20-4446775.
E-mail address: cb.terwee@vumc.nl (C.B. Terwee).

0895-4356/06/$ − see front matter © 2006 Elsevier Inc. All rights reserved.
doi: 10.1016/j.jclinepi.2006.03.012

assessment of validity it is often recommended that hypotheses about expected results should be tested, but no criteria have been defined for how many hypotheses should be confirmed to justify that a questionnaire has good validity. No criteria have been defined for what constitutes good agreement (acceptable measurement error), good responsiveness, or good interpretability, and no criteria have been defined for the required sample size of studies assessing measurement properties.

As suggested by the SAC [12], we took on the challenge to further discuss and refine the available quality criteria for studies on the development and evaluation of health status questionnaires, including explicit criteria for the following measurement properties: (1) content validity, (2) internal consistency, (3) criterion validity, (4) construct validity, (5) reproducibility, (6) responsiveness, (7) floor and ceiling effects, and (8) interpretability. We used our criteria in two systematic reviews comparing the measurement properties of questionnaires for shoulder disability [1] and for visual functioning [4], and revised them based on our experiences in these reviews. Our criteria can also be used to detect shortcomings and gaps in knowledge of measurement properties, and to design validation studies.

In this article we define our quality criteria for measurement properties, discuss the difficult and sometimes arbitrary choices we made, and indicate future challenges. We emphasize that, just like the criteria offered by the SAC and others, our criteria are open to further discussion and refinement. Our aim is to contribute to the development of explicit quality criteria for the design, methods, and outcomes of studies on the development and evaluation of health status questionnaires.

2. Content validity

Content validity examines the extent to which the concepts of interest are comprehensively represented by the items in the questionnaire [16]. To be able to rate the quality of a questionnaire, authors should provide a clear description of the following aspects regarding the development of a questionnaire:

- *Measurement aim of the questionnaire*, i.e., discriminative, evaluative, or predictive [17]. The measurement aim is important, because different items may be valid for different aims. For example, a question on stiffness could be a valid item of a discriminative questionnaire used to measure the impact of osteoarthritis on quality of life (to distinguish between patients with different levels of quality of life), but would be considered invalid for an evaluative questionnaire used as an outcome measure in a pain medication trial, because it is unlikely to be changed by pain medication.
- *Target population*, i.e., the population for which the questionnaire was developed. This is important to judge the relevance and comprehensiveness of the items. For example, a questionnaire developed to measure functional status of patients with shoulder problems may be less valid to measure functional status of patients with wrist/hand problems, because some items may be less relevant for these patients (e.g., lifting above shoulder level), whereas important items for patients with wrist/hand problems may be missing (e.g., buttoning a shirt). The relevance of items may also depend on disease severity. An adequate description of the target population is therefore important for judging the comprehensiveness and the applicability of the questionnaire in (other) populations.
- *Concepts* that the questionnaire is intended to measure. To judge the suitability of a questionnaire for a specific purpose, it is important that authors provide a clear framework of what the overall concept to be measured is. Relevant concepts can be defined in terms of symptoms; functioning (physical, psychological, and social); general health perceptions; or overall quality of life [18]. These different outcome levels should clearly be distinguished and measured by separate subscales. For physical functioning it is important to distinguish between capacity (what a patient thinks he can do) and performance (what a patient actually does).
- *Item selection and item reduction*. The methods for item selection, item reduction, and the execution of a pilot study to examine the readability and comprehension should be justified and reported. Items in the questionnaire must reflect areas that are important to the target population that is being studied. Therefore, the target population should be involved during item selection. In some guidelines it is recommended that developers start with a large number of items and apply item reduction techniques to select a small number of final items. This strategy, however, does not guarantee a better content validity, because a comprehensive set of items can also be achieved without item reduction. Therefore, we do not consider this to be mandatory.
- *Interpretability of the items*. Completing the questionnaire should not require reading skills beyond that of a 12-year-old to avoid missing values and unreliable answers [19]. That means that items should be short and simple and should not contain difficult words or jargon terms. Moreover, items should not consist of two questions at the same time [19]. Furthermore, the time period to which the questions refer should be clearly stated and justified.

We give a positive rating for content validity if a clear description is provided of the measurement aim, the target population, the concepts that are being measured, and the item selection. Furthermore, the target population should have been involved during item selection, as well as either investigators or experts. If a clear description is lacking, content validity is rated as indeterminate.

3. Internal consistency

Internal consistency is a measure of the extent to which items in a questionnaire (sub)scale are correlated (homogeneous), thus measuring the same concept. Internal consistency is an important measurement property for questionnaires that intend to measure a single underlying concept (construct) by using multiple items. In contrast, for questionnaires in which the items are merely different aspects of a complex clinical phenomenon that do not have to be correlated, such as in the Apgar Scale [20], internal consistency is not relevant [21,22].

An internally consistent (homogeneous or unidimensional) scale is achieved through good construct definitions, good items, then principal component analysis or exploratory factor analysis, followed by confirmatory factor analysis. When internal consistency is relevant, principal component analysis or factor analysis should be applied to determine whether the items form only one overall scale (dimension) or more than one [23,24]. In case that there is no prior hypothesis regarding the dimensionality of a questionnaire, exploratory principal component analysis or factor analyses can be applied. But if there is a clear hypothesis regarding the factor structure, e.g., because of an existing theoretical model or because the factor structure has been determined previously, confirmatory factor analysis should be used [25,26]. The number of subjects included in a factor analysis is a matter of debate. Rules-of-thumb vary from four to 10 subjects per variable, with a minimum number of 100 subjects to ensure stability of the variance–covariance matrix [27].

After determining the number of (homogeneous) (sub)-scales, Cronbach's alpha should be calculated for each (sub)scale separately. Cronbach's alpha is considered an adequate measure of internal consistency. A low Cronbach's alpha indicates a lack of correlation between the items in a scale, which makes summarizing the items unjustified. A very high Cronbach's alpha indicates high correlations among the items in the scale, i.e., redundancy of one or more items. Furthermore, a very high Cronbach's alpha is usually found for scales with a large number of items, because Cronbach's alpha is dependent upon the number of items in a scale. Note that Cronbach's alpha gives no information on the number of subscales in a questionnaire, because alpha can be high when two or more subscales with high alphas are combined. Nunnally and Bernstein [28] proposed a criterion of 0.70–0.90 as a measure of good internal consistency. In our experience, however, many—in our view good—(subscales of) questionnaires have higher Cronbach's alphas. We give a positive rating for internal consistency when factor analysis was applied and Cronbach's alpha is between 0.70 and 0.95.

4. Criterion validity

Criterion validity refers to the extent to which scores on a particular instrument relate to a gold standard. We give a positive rating for criterion validity if convincing arguments are presented that the used standard really is "gold" and if the correlation with the gold standard is at least 0.70.

5. Construct validity

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 concepts that are being measured [17,19]. Construct validity should be assessed by testing predefined hypotheses (e.g., about expected correlations between measures or expected differences in scores between "known" groups). These hypotheses need to be as specific as possible. Without specific hypotheses, the risk of bias is high because retrospectively it is tempting to think up alternative explanations for low correlations instead of concluding that the questionnaire may not be valid. We therefore give a positive rating for construct validity if hypotheses are specified in advance and at least 75% of the results are in correspondence with these hypotheses, in (sub)groups of at least 50 patients.

6. Reproducibility

Reproducibility concerns the degree to which repeated measurements in stable persons (test–retest) provide similar answers. We believe that it is important to make a distinction between reliability and agreement [29,30]. Agreement concerns the absolute measurement error, i.e., how close the scores on repeated measures are, expressed in the unit of the measurement scale at issue. Small measurement error is required for evaluative purposes in which one wants to distinguish clinically important changes from measurement error. Reliability concerns the degree to which patients can be distinguished from each other, despite measurement error [19]. High reliability is important for discriminative purposes if one wants to distinguish among patients, e.g., with more or less severe disease (as in diagnostic applications). Reliability coefficients (intraclass correlation coefficients (ICC)) concern the variation in the population (interindividual variation) divided by the total variation, which is the interindividual variation plus the intraindividual variation (measurement error), expressed as a ratio between 0 and 1.

The time period between the repeated administrations should be long enough to prevent recall, though short enough to ensure that clinical change has not occurred. Often, 1 or 2 weeks will be appropriate, but there could be reasons to choose otherwise. Therefore, we do not rate the appropriateness of the time period, but only require that this time period is described and justified.

6.1. Agreement

The measurement error can be adequately expressed as the standard error of measurement (SEM) [30]. The

SEM equals the square root of the error variance of an ANOVA analysis, either including systematic differences ($\mathrm{SEM}_{\mathrm{agreement}}$) or excluding them ($\mathrm{SEM}_{\mathrm{consistency}}$). Many authors fail to describe how they calculated the SEM. We believe that systematic differences should be considered part of the measurement error, because we want to distinguish them from ‘‘real’’ changes, e.g., due to treatment. Therefore, we prefer $\mathrm{SEM}_{\mathrm{agreement}}$. The SEM can be converted into the smallest detectable change ($\mathrm{SDC} = 1.96 \times \sqrt{2} \times \mathrm{SEM}$), which reflects the smallest within-person change in score that, with $P < 0.05$, can be interpreted as a ‘‘real’’ change, above measurement error, in one individual ($\mathrm{SDC}_{\mathrm{ind}}$) [31,32]. The SDC measurable in a group of people ($\mathrm{SDC}_{\mathrm{group}}$) can be calculated by dividing the $\mathrm{SDC}_{\mathrm{ind}}$ by $\sqrt{n}$ [32,33].

Another adequate parameter of agreement is described by Bland and Altman [34]. Their limits of agreement equal the mean change in scores of repeated measurements ($\mathrm{mean}_{\mathrm{change}}$) ± 1.96 × standard deviation of these changes ($\mathrm{SD}_{\mathrm{change}}$). The limits of agreement are often reported because they are easily interpretable. Note that $\mathrm{SD}_{\mathrm{change}}$ equals $\sqrt{2} \times \mathrm{SEM}_{\mathrm{consistency}}$.

For evaluative purposes, the absolute measurement error should be smaller than the minimal amount of change in the (sub)scale that is considered to be important (minimal important change (MIC)). Therefore, the MIC of a (sub)scale should be defined (see under interpretability).

We give a positive rating for agreement if the SDC ($\mathrm{SDC}_{\mathrm{ind}}$ for application in individuals and $\mathrm{SDC}_{\mathrm{group}}$ for use in groups) or the limits of agreement (upper or lower limit, depending on whether the interest is in improvement or deterioration) are smaller than the MIC. Because this is a relatively new approach and not yet commonly presented, we also give a positive rating if authors provide convincing arguments (e.g., based on their experience with the interpretation of the questionnaire scores) that the agreement is acceptable. In both cases, we consider a sample size of at least 50 patients adequate for the assessment of the agreement parameter, based on a general guideline by Altman [35].

6.2. Reliability

The ICC is the most suitable and most commonly used reliability parameter for continuous measures. Many authors fail to describe which ICC they have used, e.g., an ICC for consistency ($\mathrm{ICC}_{\mathrm{consistency}}$) or an ICC for agreement ($\mathrm{ICC}_{\mathrm{agreement}}$) [19,36]. Because systematic differences are considered to be part of the measurement error, $\mathrm{ICC}_{\mathrm{agreement}}$ (two-way random effects model, or ICC (A,1) according to McGraw and Wong [36]) is preferred. The Pearson correlation coefficient is inadequate, because systematic differences are not taken into account [19]. For ordinal measures, the weighted Cohen's Kappa coefficient should be used. The absolute percentage of agreement is inadequate, because it does not adjust for the agreement attributable to chance. When quadratic weights are being used, the weighted Kappa coefficient is identical to the $\mathrm{ICC}_{\mathrm{agreement}}$ [19].

Often 0.70 is recommended as a minimum standard for reliability [28]. We give a positive rating for reliability when the ICC or weighted Kappa is at least 0.70 in a sample size of at least 50 patients.

7. Responsiveness

Responsiveness has been defined as the ability of a questionnaire to detect clinically important changes over time, even if these changes are small [37]. A large number of definitions and methods were proposed for assessing responsiveness [38]. We consider responsiveness to be a measure of longitudinal validity. In analogy to construct validity, longitudinal validity should be assessed by testing predefined hypotheses, e.g., about expected correlations between changes in measures, or expected differences in changes between ‘‘known’’ groups [38]. This shows the ability of a questionnaire to measure changes if they really have happened. Futhermore, the instrument should be able to distinguish clinically important change from measurement error. Responsiveness should therefore be tested by relating the SDC to the MIC, as described under agreement (see Section 6.1). This approach equals Guyatt's responsiveness ratio (RR), in which the clinically important change (MIC) is related to the between-subject variability in within-subject changes in stable subjects ($\mathrm{SD}_{\mathrm{change}}$; the same as in the limits of agreement) [39]. The RR should thus be at least 1.96 (at the value of 1.96 the MIC equals the $\mathrm{SDC}_{\mathrm{ind}}$, which is $1.96 \times \mathrm{SD}_{\mathrm{change}}$). Another adequate measure of responsiveness is the area under the receiver operating characteristics (ROC) curve (AUC) [40], which is a measure of the ability of a questionnaire to distinguish patients who have and have not changed, according to an external criterion. We consider an AUC of at least 0.70 to be adequate.

8. Floor or ceiling effects

Floor or ceiling effects are considered to be present if more than 15% of respondents achieved the lowest or highest possible score, respectively [41]. If floor or ceiling effects are present, it is likely that extreme items are missing in the lower or upper end of the scale, indicating limited content validity. As a consequence, patients with the lowest or highest possible score cannot be distinguished from each other, thus reliability is reduced. Furthermore, the responsiveness is limited because changes cannot be measured in these patients. We give a positive rating for (the absence of) floor and ceiling effects if no floor or ceiling effects are present in a sample size of at least 50 patients.

9. Interpretability

Interpretability is defined as the degree to which one can assign qualitative meaning to quantitative scores [42].

Investigators should provide information about what (change in) score would be clinically meaningful. Various types of information can aid in interpreting scores on a questionnaire: (1) means and SD of scores of (subgroups of) a reference population (norm values); (2) means and SD of scores of relevant subgroups of patients who are expected to differ in scores (e.g., groups with different clinical diagnoses, age groups, gender groups, primary vs. secondary care setting); (3) means and SD of scores of patients before and after treatment(s) of known efficacy [43]; and (4) means and SD of scores of subgroups of patients based on patients' global ratings of change. For example, a positive rating was given if mean scores and SD are presented of at least four subgroups. For example, if means and SD are presented for a general population (norm values), stratified by gender and age groups. In addition, an MIC should be defined to enable interpretation of change scores over time and sample size calculations. The MIC has been defined as "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" [44]. Various distribution-based and anchor-based methods have been proposed. Anchor-based approaches use an external criterion to operationalize an important change. Distribution-based approaches are based on statistical characteristics of the sample [45]. We recommend an anchor-based method, to determine the MIC because distribution-based methods do not provide a good indication of the importance of the observed change. We consider a sample size of at least 50 patients adequate to determine the MIC.

10. Population-specific ratings of measurement properties

A summary of the criteria for measurement properties of health status questionnaires is presented in Table 1. Each property is rated as positive, negative, or indeterminate, depending on the design, methods, and outcomes of the study. Measurement properties differ between populations and settings. Therefore, the evaluation of all measurement properties needs to be conducted in a population and setting that is representative for the population and setting in which the questionnaire is going to be used. The setting refers to the testing conditions, e.g., self-completed or interview, and language. A clear description of the design of each individual study has to be provided, including population (diagnosis and clinical features, age, and gender); design (e.g., language version, time between the measurements, completion before or after treatment); testing conditions (e.g., questionnaires completed at home or in a waiting room, self of in interview); and analyses of the data. If a clear description of the design of the study is lacking, the evaluated measurement properties are rated as indeterminate. In addition, if any important methodological weakness in the design or execution of the study is found, e.g., selection bias or an extremely heterogeneous study population, the evaluated measurement properties are also rated as indeterminate.

11. Overview table

In the final comparison of the measurement properties of different questionnaires, one has to consider all ratings together when choosing between different questionnaires. We recommend to compose a table that provides an overview of all ratings, such as the example given in Table 2. In Table 2 the results are presented from our systematic review of all questionnaires measuring disability in patients with shoulder complaints (because there is no gold standard for disability, criterion validity was not assessed) [1]. In Table 2 all ratings for each questionnaire are presented separately for each specific population or setting. For example, the Shoulder Pain and Disability Index (SPADI) was evaluated in several populations. Construct validity and responsiveness were rated positively for outpatients (c), but rated as indeterminate for primary care patients (b) and hospital patients (d). With this table one can make an evidence-based choice for the questionnaire with the best measurement properties, taking into account those measurement properties that are most important for a specific application (e.g., reliability when using a questionnaire for discrimination and responsiveness when using it for evaluation of a treatment effect) and the population and setting in which the questionnaire is going to be used.

12. Discussion

We developed quality criteria for the design, methods, and outcomes of studies on the development and evaluation of health status questionnaires. Nine measurement properties were distinguished: content validity, internal consistency, criterion validity, construct validity, reproducibility, longitudinal validity, responsiveness, floor and ceiling effects, and interpretability.

Our criteria are mostly opinion based because there is no empirical evidence in this field to support explicit quality criteria. They are useful rules of thumb, but other investigators may want to make their own choices.

We did not summarize the quality criteria into one overall quality score, as is often done in systematic reviews of randomized clinical trials [46]. An overall quality score assumes that all measurement properties are equally important, which is probably not the case. We consider content validity as one of the most important measurement properties. Only if the content validity of a questionnaire is adequate, one will consider using the questionnaire, and evaluation of the other measurement properties is useful. Furthermore, the aim of the questionnaire demands different qualities of the questionnaire with respect to reproducibility and responsiveness. Discriminative questionnaires require a high level of reliability to be able to distinguish between persons. Evaluative questionnaires require a high

Table 1
Quality criteria for measurement properties of health status questionnaires

Property	Definition	Quality criteria[a,b]
1. Content validity	The extent to which the domain of interest is comprehensively sampled by the items in the questionnaire	+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; 0No information found on target population involvement.
2. Internal consistency	The extent to which items in a (sub)scale are intercorrelated, thus measuring the same construct	+Factor analyses performed on adequate sample size (7 * # items and ⩾100) AND Cronbach's alpha(s) calculated per dimension AND Cronbach's alpha(s) between 0.70 and 0.95; ?No factor analysis OR doubtful design or method; −Cronbach's alpha(s) <0.70 or >0.95, despite adequate design and method; 0No information found on internal consistency.
3. Criterion validity	The extent to which scores on a particular questionnaire relate to a gold standard	+Convincing arguments that gold standard is "gold" AND correlation with gold standard ⩾0.70; ?No convincing arguments that gold standard is "gold" OR doubtful design or method; −Correlation with gold standard <0.70, despite adequate design and method; 0No 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	+Specific hypotheses were formulated AND at least 75% of the results are in accordance with these hypotheses; ?Doubtful design or method (e.g., no hypotheses); −Less than 75% of hypotheses were confirmed, despite adequate design and methods; 0No 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)	+MIC < SDC OR MIC outside the LOA OR convincing arguments that agreement is acceptable; ?Doubtful design or method OR (MIC not defined AND no convincing arguments that agreement is acceptable); −MIC ⩾ SDC OR MIC equals or inside LOA, despite adequate design and method; 0No 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); −ICC or weighted Kappa < 0.70, despite adequate design and method; 0No information found on reliability.
6. Responsiveness	The ability of a questionnaire to detect clinically important changes over time	+SDC 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, despite adequate design and methods; 0No 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 the respondents achieved the highest or lowest possible scores, despite adequate design and methods; 0No information found on interpretation.
8. Interpretatability	The degree to which one can assign qualitative meaning to quantitative 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; 0No information found on interpretation.

MIC = minimal important change; SDC = smallest detectable change; LOA = limits of agreement; ICC = Intraclass correlation; SD, standard deviation.

[a] + = positive rating; ? = indeterminate rating; − = negative rating; 0 = no information available.

[b] Doubtful design or method = lacking of 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.

Table 2
Summary of the assessment of the measurement properties of all questionnaires measuring disability in patients with shoulder complaints [1]

	Content	Internal	Construct	Reproducibility			Floor or	
Questionnaire	validity	consistency	validity	Agreement	Reliability	Responsiveness	ceiling effect	Interpretability
SDQ-UK	+	0	+	0	0	0	+ (b); − (a)	0
SIQ	+	?	+	+	0	+	+	+
OSQ	+	?	+	+	0	+	+	+
SDQ-NL	?	0	+	0	0	+ (b); ? (b)	− (b)	+
RC-QOL	+	0	+	?	0	0	+	?
DASH	+	0	+ (c,d); ? (c)	+ (c)	+ (c)	+ (c); ? (d)	+ (c)	+
WOSI	+	0	+	0	+	?	0	0
SSRS	−	0	? (c)	? (d)	? (d)	? (d)	+ (d)	+
SRQ	+	?	?	?	?	?	0	?
SST	+	?	+ (c); ? (d)	? (c,d)	? (d)	? (c,d)	+ (c)	+
WOOS	+	0	?	0	?	?	0	0
SSI	0	0	+ (c)	? (d)	? (d)	? (d)	+ (c)	0
UEFS	−	+	?	0	0	?	+	+
ASES	−	?	+ (c); ? (d)	? (d)	? (c,d)	? (c,d)	+ (d); ? (c)	+
SPADI	−	?	+ (c); ? (b,c)	+ (c); ? (d)	? (c,d)	+ (c); ? (b,d)	+ (b,c)	+
UEFL	−	0	+	0	0	0	− (a)	0

Rating: + = positive; 0 = intermediate; - = poor; ? = no information available.

Kinds of study population(s) used in the studies: (a) community, (b) primary care, (c) outpatients' clinic, and (d) hospital patients.

SDQ-UK = Shoulder Disability Questionnaire (English version); SIQ = Shoulder Instability Questionnaire; OSQ (Oxford) = Shoulder Questionnaire; SDQ-NL = Shoulder Disability Questionnaire (Dutch version); RC-QOL = Rotator Cuff Quality-of-Life Measure; DASH = Disabilities of the Arm, Shoulder and Hand Scale; WOSI = Western Ontario Shoulder Instability Index; SSRS = Subjective Shoulder Rating Scale; SST = Simply Shoulder Test; SSI = Shoulder Severity Index; UEFS = Upper Extremity Function Scale; ASES = American Shoulder and Elbow Surgeons Standardized Shoulder Assessment Form.

level of agreement to be able to measure important changes. Evaluative questionnaires should be responsive to change, whereas discriminative questionnaires do not necessarily need to be responsive to change.

We recommend to always compose a table that provides an overview of all ratings of all questionnaires, such as Table 2, which facilitates an assessment of all ratings together when choosing the most suitable questionnaire for a specific application. Two important issues should be kept in mind. Firstly, with our approach, poor quality questionnaires can be given positive ratings for some measurement properties. For example, the Upper Extremity Functional Limitation Scale (UEFL) received a positive rating for construct validity despite a negative rating for content validity. By considering all ratings together, one may decide for example to choose the Shoulder Rating Questionnaire (SRQ) or the Western Ontario Osteoarthritis of the Shoulder Index (WOOS), with positive ratings for content validity, over the UEFL, despite the fact that these questionnaires received indeterminate ratings for construct validity. Secondly, questionnaires with the highest number of positive ratings are not necessarily the best questionnaires. The ratings depend on the availability of information and the quality of reporting on the assessment of the measurement properties. For example, a questionnaire can be given many indeterminate ratings if measurement properties are not yet evaluated because the questionnaire is newly developed. Furthermore, poorly reported validation studies will lead to low ratings for questionnaires that are not necessarily poor in design or performance. The quality of reporting of studies on the development and evaluation of health status questionnaires therefore needs to be improved.

In applying our criteria in two systematic reviews [1,4], we found that several measurement properties are often not properly assessed or analyzed, nor clearly reported. Important information on content validity is often very limited. The aim of a questionnaire, i.e., whether the questionnaire was designed for discriminative or evaluative purposes, is often not described, and the concepts that the questionnaire is intended to measure are often ill defined. Furthermore, item selection is often poorly described. A statement like "A preliminary questionnaire was developed and completed by 30 patients.... A subset of these patients was interviewed and each question was assessed for clinical relevance, importance, and ease of completion" [47] does not justify that the items comprehensively represent all issues that are important to the target population. This hampers judgment about the applicability of the questionnaire in a given population.

One would assume that the number of scales corresponds with the number of dimensions identified in factor analysis, but this is often not the case. We found that several questionnaires claimed to cover more than one dimension, but consisted of one scale only, or vice versa [1].

Many authors fail to specify hypotheses for the assessment of construct validity or the hypotheses are not very specific or informative. For example, to validate a shoulder disability questionnaire, the authors tested the hypothesis that "restriction of shoulder movement on examination correlates with disability score" [48]. An informative hypothesis should evaluate, as directly as possible, specific claims made for a theory [49], and should thus include a justification of the expected direction and magnitude of the correlation. Without testing specific hypotheses, the risk of bias is high. Practically all authors of instrument evaluation studies conclude that their instrument is valid, whereas objective counting of the number of hypotheses that were confirmed frequently indicates otherwise [1,4].

Responsiveness is often ill defined and not well assessed. For example, many authors use multiple indices of responsiveness and conclude that their instrument is responsive when all indices point in the same direction [50–52]. Different authors, however, have pointed to the conceptual and algebraic differences between different indices of responsiveness, showing that different indices may lead to different conclusions [53,54]. Furthermore, many indices of responsiveness should actually be looked upon as measures of the magnitude of a treatment effect, which, in itself, tell us nothing about the quality of the instrument to serve its purpose [38].

We recommend using criteria like ours as an aid in the design and reporting of studies on the development and evaluation of health status questionnaires. This should lead to an improvement in the quality of (reporting of) such studies.

13. Future challenges

One might argue that our criteria are not discriminative enough to distinguish between good and very high-quality questionnaires. This would be important when many high-quality questionnaires are available, but in our experience, within the field of health status and health-related quality of life measurement, this is not (yet) the case. Therefore, we believe that our criteria work well to separate the wheat from the chaff. The next step would be to further refine and complete the criteria, e.g., by deciding how specific the hypotheses for testing construct validity or responsiveness should be, and by including criteria for the methods and results of studies using Item Response Theory (IRT) models. Furthermore, broad consensus is needed, especially on the criteria for good measurement properties.

Because the number of health status questionnaires is rapidly growing, choosing the right questionnaire for a specific purpose becomes a time-consuming and difficult task. An increasing number of systematic reviews are being published in which the measurement properties of health status questionnaires are being evaluated and compared. These systematic reviews are important tools for evidence-based instrument selection. Explicit quality criteria for studies on the development and evaluation of health status questionnaires are needed to legitimize what the best questionnaire is. Our criteria are a step in this direction. The future challenge will be to refine and complete the criteria and to reach broad consensus, especially on the quality criteria for good measurement properties.

References

[1] Bot SD, Terwee CB, van der Windt DA, Bouter LM, Dekker J, de Vet HC. Clinimetric evaluation of shoulder disability questionnaires: a systematic review of the literature. Ann Rheum Dis 2004;63:335–41.

[2] Jorstad EC, Hauer K, Becker C, Lamb SE. Measuring the psychological outcomes of falling: a systematic review. J Am Geriatr Soc 2005;53:501–10.

[3] Daker-White G. Reliable and valid self-report outcome measures in sexual (dys)function: a systematic review. Arch Sex Behav 2002;31: 197–209.

[4] de Boer MR, Moll AC, de Vet HC, Terwee CB, Volker-Dieben HJ, van Rens GH. Psychometric properties of vision-related quality of life questionnaires: a systematic review. Ophthalmic Physiol Opt 2004;24:257–73.

[5] Edwards B, Ung L. Quality of life instruments for caregivers of patients with cancer. Cancer Nurs 2002;25:342–9.

[6] Garratt AM, Brealey S, Gillespie WJ. Patient-assessed health instrument for the knee: a structured review. Rheumatology 2004;43: 1414–23.

[7] Hallin P, Sullivan M, Kreuter M. Spinal cord injury and quality of life measures: a review of instrument psychometric quality. Spinal Cord 2000;38:509–23.

[8] Haywood KL, Garratt AM, Fitzpatrick R. Quality of life in older people: a structured review of generic self-assessed health instruments. Qual Life Res 2005;14:1651–68.

[9] Haywood KL, Garratt AM, Dawes PT. Patient-assessed health in ankylosing spondylitis: a structured review. Rheumatology (Oxford) 2005;44:577–86.

[10] Dziedzic KS, Thomas E, Hay EM. A systematic search and critical review of measures of disability for use in a population survey of hand osteoarthritis (OA). Osteoarthritis Cartilage 2005;13:1–12.

[11] Ettema TP, Droes RM, de Lange J, Mellenbergh GJ, Ribbe MW. A review of quality of life instruments used in dementia. Qual Life Res 2005;14:675–86.

[12] Scientific Advisory Committee of the Medical Outcomes Trust. Assessing health status and quality-of-life instruments: attributes and review criteria. Qual Life Res 2002;11:193–205.

[13] Bombardier C, Tugwell P. Methodological considerations in functional assessment. J Rheumatol 1987;14(Suppl 15):6–10.

[14] Andresen EM. Criteria for assessing the tools of disability outcomes research. Arch Phys Med Rehabil 2000;81(Suppl 2):S15–20.

[15] McDowell I, Jenkinson C. Development standards for health measures. J Health Serv Res Policy 1996;1:238–46.

[16] Guyatt GH, Feeny DH, Patrick DL. Measuring health related quality of life. Ann Intern Med 1993;118:622–9.

[17] Kirschner B, Guyatt G. A methodological framework for assessing health indices. J Chronic Dis 1985;38:27–36.

[18] Wilson IB, Cleary PD. Linking clinical variables with health-related quality of life. JAMA 1995;273:59–65.

[19] Streiner DL, Norman GR. Health measurement scales. A practical guide to their development and use. New York: Oxford University Press; 2003.

[20] Apgar V. A proposal for new method of evaluation of the newborn infant. Curr Res Anesth Analg 1953;32:260–7.

[21] Fayers PM, Hand DJ. Causal variables, indicator variables and measurement scales: an example from quality of life. J R Stat Soc A 2002;165:233–61.

[22] Streiner DL. Being inconsistent about consistency: when coefficient alpha does and doesn't matter. J Pers Assess 2003;80:217–22.

[23] Floyd FJ, Widaman KF. Factor analysis in the development and refinement of clinical assessment instruments. Psychol Assess 1995;7:286–99.

[24] Streiner DL. Figuring out factors: the use and misuse of factor analysis. Can J Psychiatry 1994;39:135–40.

[25] Bollen KA. Structural equations with latent variables. New York: Wiley; 1989.

[26] de Vet HCW, Ader HJ, Terwee CB, Pouwer F. Are factor analytical techniques appropriately used in the validation of health status questionnaires? A systematic review on the quality of factor analyses of the SF-36. Qual Life Res 2005;14:1203−18.

[27] Kline P. The handbook of psychological testing. London: Routledge; 1993.

[28] Nunnally JC, Bernstein IH. Psychometric theory. 3rd ed. New York: McGraw-Hill; 1994.

[29] de Vet HCW. Observer reliability and agreement. In: Armitage P, Colton T, editors. Encyclopedia of biostatistics. Boston: John Wiley & Sons Ltd.; 1998. p. 3123−8.

[30] Stratford P. Reliability: consistency or differentiating among subjects? Phys Ther 1989;69:299−300.

[31] Beckerman H, Roebroeck ME, Lankhorst GJ, Becher JG, Bezemer PD, Verbeek AL. Smallest real difference, a link between reproducibility and responsiveness. Qual Life Res 2001;10:571−8.

[32] de Vet HCW, Bouter LM, Bezemer PD, Beurskens AJ. Reproducibility and responsiveness of evaluative outcome measures. Theoretical considerations illustrated by an empirical example. Int J Technol Assess Health Care 2001;17:479−87.

[33] de Boer MR, de Vet HCW, Terwee CB, Moll AC, Völker-Dieben HJM, van Rens GHMB. Change to the subscales of two vision-related quality of life questionnaires are proposed. J Clin Epidemiol 2005;58:1260−8.

[34] Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet 1986;307−10.

[35] Altman DG. Practical statistics for medical research. London: Chapman and Hall; 1991.

[36] McGraw KO, Wong SP. Forming inferences about some intraclass correlation coefficients. Psychol Methods 1996;1:30−46.

[37] Guyatt GH, Deyo RA, Charlson M, Levine MN, Mitchell A. Responsiveness and validity in health status measurement: a clarification. J Clin Epidemiol 1989;42:403−8.

[38] Terwee CB, Dekker FW, Wiersinga WM, Prummel MF, Bossuyt PM. On assessing responsiveness of health-related quality of life instruments: guidelines for instrument evaluation. Qual Life Res 2003;12: 349−62.

[39] Guyatt G, Walter S, Norman G. Measuring change over time: assessing the usefulness of evaluative instruments. J Chronic Dis 1987;40: 171−8.

[40] Deyo RA, Centor RM. Assessing the responsiveness of functional scales to clinical change: an analogy to diagnostic test performance. J Chronic Dis 1986;39:897−906.

[41] McHorney CA, Tarlov AR. Individual-patient monitoring in clinical practice: are available health status surveys adequate? Qual Life Res 1995;4:293−307.

[42] Lohr KN, Aaronson NK, Alonso J, Burnam MA, Patrick DL, Perrin EB, et al. Evaluating quality of life and health status instruments: development of scientific review criteria. Clin Ther 1996;18: 979−92.

[43] Guyatt GH, Bombardier C, Tugwell PX. Measuring disease-specific quality of life in clinical trials. CMAJ 1986;134:889−95.

[44] Jaeschke R, Singer J, Guyatt GH. Measurement of health status. Ascertaining the minimal clinically important difference. Control Clin Trials 1989;10:407−15.

[45] Crosby RD, Kolotkin RL, Williams GR. Defining clinically meaningful change in health-related quality of life. J Clin Epidemiol 2003;56: 395−407.

[46] Jadad AR, Moore RA, Carroll D, Jenkinson C, Reynolds DJ, Gavaghan DJ, et al. Assessing the quality of reports of randomized clinical trials: is blinding necessary? Control Clin Trials 1996;17: 1−12.

[47] L'Insalata JC, Warren RF, Cohen SB, Altchek DW, Peterson MG. A self-administered questionnaire for assessment of symptoms and function of the shoulder. J Bone Joint Surg Am 1997;79:738−48.

[48] Croft P, Pope D, Zonca M, O'Neill T, Silman A. Measurement of shoulder related disability: results of a validation study. Ann Rheum Dis 1994;53:525−8.

[49] Strauss ME. Introduction to the special section on construct validity of psychological tests: 50 years after Cronbach and Meehl (1955). Psychol Assess 2005;17:395.

[50] Beaton DE, Hogg-Johnson S, Bombardier C. Evaluating changes in health status: reliability and responsiveness of five generic health status measures in workers with musculoskeletal disorders. J Clin Epidemiol 1997;50:79−93.

[51] Stucki G, Liang MH, Fossel AH, Katz JN. Relative responsiveness of condition-specific and generic health status measures in degenerative lumbar spinal stenosis. J Clin Epidemiol 1995;48:1369−78.

[52] Wright JG, Young NL. A comparison of different indices of responsiveness. J Clin Epidemiol 1997;50:239−46.

[53] Stratford PW, Riddle DL. Assessing sensitivity to change: choosing the appropriate change coefficient. Health Qual Life Outcomes 2005;3:23.

[54] Zou GY. Quantifying responsiveness of quality of life measures without an external criterion. Qual Life Res 2005;14:1545−52.