Rating the methodological quality in systematic reviews of studies on measurement properties: a scoring system for the COSMIN checklist

Caroline B. Terwee · Lidwine B. Mokkink · Dirk L. Knol · Raymond W. J. G. Ostelo · Lex M. Bouter · Henrica C. W. de Vet

Accepted: 25 June 2011 / Published online: 6 July 2011

Abstract

Background The COSMIN checklist is a standardized tool for assessing the methodological quality of studies on measurement properties. It contains 9 boxes, each dealing with one measurement property, with 5–18 items per box about design aspects and statistical methods. Our aim was to develop a scoring system for the COSMIN checklist to calculate quality scores per measurement property when using the checklist in systematic reviews of measurement properties.

Methods The scoring system was developed based on discussions among experts and testing of the scoring system on 46 articles from a systematic review. Four response options were defined for each COSMIN item (excellent, good, fair, and poor). A quality score per measurement property is obtained by taking the lowest rating of any item in a box ("worst score counts").

Results Specific criteria for excellent, good, fair, and poor quality for each COSMIN item are described. In defining the criteria, the "worst score counts" algorithm was taken into consideration. This means that only fatal flaws were defined as poor quality. The scores of the 46 articles show how the scoring system can be used to provide an overview of the methodological quality of studies included in a systematic review of measurement properties.

Conclusions Based on experience in testing this scoring system on 46 articles, the COSMIN checklist with the proposed scoring system seems to be a useful tool for assessing the methodological quality of studies included in systematic reviews of measurement properties.

Keywords Reproducibility of results · Validation studies · Outcome assessment · Psychometrics · Systematic review · Questionnaire

Introduction

Systematic reviews of measurement properties are useful for selecting the best measurement instrument for a specific purpose [1]. These reviews are becoming increasingly important because the number of measurement instruments for assessing one particular construct is still rising, especially in the field of health-related patient-reported outcomes. The number of systematic reviews of measurement properties of health status measurement instruments has increased from less than 5 per year before 1996 to 45 in 2008 (www.cosmin.nl). However, the methodology of systematic reviews of measurement properties is still under development.

In a systematic review, not only the results of the included studies but also their methodological quality should be taken into account. The assessment of the methodological quality of a study and the assessment of the

C. B. Terwee (✉) · L. B. Mokkink · D. L. Knol · H. C. W. de Vet
Department of Epidemiology and Biostatistics and the EMGO Institute for Health and Care Research, VU University Medical Center, Van der Boechorststraat 7, 1081 BT Amsterdam, The Netherlands
e-mail: cb.terwee@vumc.nl

R. W. J. G. Ostelo
Department of Health Sciences and the EMGO Institute for Health and Care Research, Faculty of Earth and Life Sciences, VU University Amsterdam, Amsterdam, The Netherlands

L. M. Bouter
Executive Board of VU University Amsterdam, Amsterdam, The Netherlands

quality of the instrument at issue are two different things and should be performed separately in systematic reviews. If the methodological quality of a study on the measurement properties of a specific instrument is appropriate, the results can be used to assess the quality of the instrument at issue. However, when the methodological quality of a study is inadequate, the results cannot be trusted and the quality of the instrument under study remains unclear [2]. Some authors of systematic reviews of measurement properties have evaluated the methodological quality of the included studies [3–6]. However, different methods are used to evaluate methodological quality. For example, Haywood et al. [3] considered the number and kind of tests and studies performed; Alla et al. [4] used criteria designed for the quality assessment of trials; Marinus et al. [5] considered the appropriateness of the analyses and sample size; and Wind et al. [6] considered the objective, population, assessment method, and analyses and presentation of the statistical outcomes.

Recently, an international Delphi study was carried out to develop the COnsensus-based Standards for the selection of health Measurement INstruments (COSMIN) checklist for assessing the methodological quality of studies on measurement properties [7, 8]. The COSMIN checklist is increasingly used in systematic reviews of measurement properties. It contains nine boxes, each dealing with one measurement property. Each box contains 5–18 items that can be used to assess whether a study on a specific measurement property meets the standard for good methodological quality. The boxes concern internal consistency, reliability, measurement error, content validity, construct validity (i.e., structural validity, hypotheses testing, and cross-cultural validity), criterion validity, and responsiveness. An additional box is included to assess the methodological quality of a study on interpretability. Finally, the checklist contains one box including additional methodological standards for studies that use item response theory (IRT) models and one box to assess the generalizability of the results of a study on measurement properties (12 boxes in total) [7, 8]. Each measurement property is evaluated separately. This means that if multiple measurement properties are assessed in one study, several COSMIN boxes need to be completed. Instructions for use of the COSMIN checklist are available in a manual on the COSMIN web site (www.cosmin.nl).

The number of questions per COSMIN box varies from 5 to 18. The response options are “yes,” “no,” and either “unknown (?)” or “not applicable (NA).” So far, it has not been defined how an overall methodological quality score per measurement property (per box) can be obtained. It is, however, highly desirable in systematic reviews to obtain a methodological quality score for each study on a given measurement property. The aim of this study was therefore to develop a scoring system that can be used to calculate methodological quality scores per measurement property for the COSMIN checklist.

Methods

The scoring system was developed based on discussions in the Clinimetrics working group of the EMGO Institute of Health and Care Research of the VU University Medical Center in Amsterdam (www.clinimetrics.nl), alternated with testing the scoring system on a set of 46 articles that were included in a systematic review on the measurement properties of 8 neck disability questionnaires [9]. The Clinimetrics working group consists of about 20 investigators, including PhD students, post docs, senior researchers, and professors. The group convenes once a month to discuss research proposals, progress of ongoing projects, manuscripts in preparation, or methodological papers from the literature.

The aim of the scoring system is to obtain an overall methodological quality score per measurement property for a given study. Note that methodological quality scores for different studies are not combined. For example, if three studies on the reliability of the same measurement instrument are included in a systematic review, the methodological quality of each study is rated separately.

It was decided to change the dichotomous response options (yes, no) of the COSMIN items into four response options (excellent, good, fair, and poor) in order to increase the discriminative ability of the items. Four response options for each item of the COSMIN checklist were defined, representing excellent, good, fair, and poor methodological quality. Subsequently, a methodological quality score per box is obtained by taking the lowest rating of any item in a box (“worst score counts”). For example, if one item in the box “Reliability” is scored poor, the methodological quality of the assessment of reliability in that study is rated as poor. A poor score on any item is thus considered to represent a fatal flaw.

A first draft of the scoring system was made by one of the authors [CBT] and was discussed in the Clinimetrics working group. Based on the discussion, adaptations were made. A second draft of the scoring system was applied to a set of 46 articles from a systematic review on the measurement properties of 8 neck disability questionnaires [9]. Each article was scored by one rater [CBT] using the COSMIN checklist with the 4-point response options. Methodological quality scores per measurement property were obtained by taking the lowest rating of any item in the relevant box. The results were compared with the rater’s overall judgement of the quality of the study, and discrepancies were noted. These discrepancies were discussed in the Clinimetrics working group, and adaptations to the scoring system were made. Subsequently, a third draft was again applied to the 46 articles by the same rater.

Below we describe the scoring system in more detail and present the quality scores of the 46 articles as an example of how this scoring system can be applied in a systematic review of measurement properties. Alternative scoring systems that were considered will be discussed in the discussion section.

The Interpretability box and the Generalizability box are mainly used as data extraction forms. We recommend to use the Interpretability box to extract all information on the interpretability issues described in this box (e.g., norm scores, floor-ceiling effects, and minimal important change) of the instruments under study from the included articles. Similar, we recommend to use the Generalizability box to extract data on the characteristics of the study population and sampling procedure. Therefore, no scoring system was developed for these boxes.

Results

General description of the scoring system

The specific criteria for excellent, good, fair, and poor quality per item of each COSMIN box are described in the COSMIN checklist with a 4-point scale (available from the web site www.cosmin.nl). As an example, the box "Reliability" (box B) with a 4-point scale is presented in Table 1. In general, an item is scored as excellent when there is evidence that the methodological quality aspect of the study to which the item is referring is adequate (this equals the original response option "yes"). For example, if evidence is provided (e.g., from a global rating scale) that patients remained stable between the test and retest (item 7, box B), this item is scored as excellent. An item is scored as good when relevant information is not reported in an article, but it can be assumed that the quality aspect is adequate. For example, if it can be assumed that patients were stable between the test and retest (e.g., based on the clinical characteristics of the patients and the time interval between the test and retest), the item is scored as good. An item is scored as fair if it is doubtful whether the methodological quality aspect is adequate. For example, when it is unclear whether the patients were stable in a reliability study, the item is scored as fair. Finally, an item is scored as poor when evidence is provided that the methodological quality aspect is not adequate, for example, if patients were treated between the test and retest.

In defining the response options, the "worst score counts" algorithm was taken into consideration. Only fatal flaws in the design or statistical analyses were regarded as poor methodological quality. For example, when in a construct validity study no hypotheses were formulated a priori regarding the relation of the instrument under study with other measures, and it was unclear what was expected, this is considered poor methodological quality. For some items, the worst possible response option was defined as good or fair instead of poor because we did not want these items to have too much impact on the methodological quality score per box. For example, item 1 in most boxes refers to whether the percentage of missing items is given. The only two possible answers are yes or no, which were rated as excellent and good, respectively. This does not mean, however, that we consider it good practice if this information is not reported. It rather means that, in our opinion, a study that did not report the number of missing items can still obtain an overall score of good methodological quality for a measurement property, if all other items are scored good or excellent. Item 2 in most boxes refers to whether it was described how missing items were handled. If this is not described, this is not necessarily a fatal flaw in the study. Therefore, it was decided to score this item as fair instead of poor if it was not described how missing items were handled. Finally, for some items, it was not possible to define four different response options. For these items, only two or three response options were defined. For example, item 9 in box E (structural validity) refers to whether exploratory or confirmatory factor analysis was performed. The only possible answers are (1) yes (excellent), (2) yes but exploratory factor analysis was performed while confirmatory would have been more appropriate (good), or (3) no (poor).

In all boxes, a small sample size was considered poor methodological quality. As a rule of thumb, a sample size of 100 is considered as excellent, 50 as good, 30 as fair, and less than 30 as poor [10]. For the assessment of some measurement properties, larger sample sizes are required, e.g., for factor analysis, the sample size should be at least five to seven times the number of items with a minimum of 100 (item 6, box A and item 4 box E) [11].

Scoring the quality of IRT studies

If studies use IRT models, the COSMIN IRT box should be completed in addition to the specific boxes for the measurement properties that were evaluated in the IRT study. IRT models are most often used for assessing internal consistency and cross-cultural validity (Differential Item Functioning). If the IRT model, the computer software package, or the method of estimation was not adequately described (IRT box items 1–3), these items are scored good instead of excellent. If the assumptions for estimating parameters of the IRT model were not checked or this is unknown (item 4), this item is scored fair. To obtain a total score for the methodological quality of studies that use IRT methods, the "worst score counts" algorithm should be

Table 1 Example of one COSMIN box with 4-point scale

Box B. Reliability: relative measures (including test–retest reliability, inter-rater reliability, and intra-rater reliability)

	Excellent	Good	Fair	Poor
Design requirements				
1. Was the percentage of missing items given?	Percentage of missing items described	Percentage of missing items NOT described		
2. Was there a description of how missing items were handled?	Described how missing items were handled	Not described but it can be deduced how missing items were handled	Not clear how missing items were handled	
3. Was the sample size included in the analysis adequate?	Adequate sample size ($\geq$100)	Good sample size (50–99)	Moderate sample size (30–49)	Small sample size (<30)
4. Were at least two measurements available?	At least two measurements			Only one measurement
5. Were the administrations independent?	Independent measurements	Assumable that the measurements were independent	Doubtful whether the measurements were independent	Measurements NOT independent
6. Was the time interval stated?	Time interval stated		Time interval NOT stated	
7. Were patients stable in the interim period on the construct to be measured?	Patients were stable (evidence provided)	Assumable that patients were stable	Unclear whether patients were stable	Patients were NOT stable
8. Was the time interval appropriate?	Time interval appropriate		Doubtful whether time interval was appropriate	Time interval NOT appropriate
9. Were the test conditions similar for both measurements? e.g., type of administration, environment, and instructions	Test conditions were similar (evidence provided)	Assumable that test conditions were similar	Unclear whether test conditions were similar	Test conditions were NOT similar
10. Were there any important flaws in the design or methods of the study?	No other important methodological flaws in the design or execution of the study		Other minor methodological flaws in the design or execution of the study	Other important methodological flaws in the design or execution of the study
Statistical methods				
11. For continuous scores: Was an intraclass correlation coefficient (ICC) calculated?	ICC calculated and model or formula of the ICC is described	ICC calculated but model or formula of the ICC not described. Pearson or Spearman correlation coefficient calculated with evidence provided that no systematic change has occurred	Pearson or Spearman correlation coefficient calculated WITHOUT evidence provided that no systematic change has occurred or WITH evidence that systematic change has occurred	No ICC or Pearson or Spearman correlations calculated
12. For dichotomous/nominal/ordinal scores: Was kappa calculated?	Kappa calculated			Only percentage agreement calculated
13. For ordinal scores: Was a weighted kappa calculated?	Weighted Kappa calculated		Unweighted Kappa calculated	Only percentage agreement calculated
14. For ordinal scores: Was the weighting scheme described? e.g. linear, quadratic	Weighting scheme described	Weighting scheme NOT described		

applied to the combination of the IRT box and the box of the measurement property that was evaluated in the IRT study. For example, if IRT methods are used to study internal consistency and item 4 in the IRT box is scored fair, while the items in the internal consistency box (box A) are all scored as good or excellent, the methodological quality score for the internal consistency study will be fair. However, if any of the items in box A is scored poor, the methodological quality score for the internal consistency study will be poor.

Adaptations made during testing

In comparing the initial COSMIN scoring with the rater's overall judgement of the methodological quality of the study, we found a few discrepancies. In most cases, the rater's overall judgement was more positive than the rating obtained with the COSMIN scoring system. For example, when rating the methodological quality of a study on construct validity and the expected direction of correlations or mean differences were not included in the hypotheses for testing construct validity, this was originally rated as fair quality. However, the rater argued that it was often possible to deduce what was expected. We therefore changed the scoring of this response option into good.

Example of the application of the scoring system in systematic reviews of measurement properties

The scoring system was applied on a set of 46 articles from a systematic review on the measurement properties of 8 neck disability questionnaires [9]. The results are presented in Fig. 1. This figure shows how the scoring system can be used to provide an overview of the methodological quality of the included studies on measurement properties in a systematic review. For example, in 41 of the 46 articles, construct validity was evaluated. 5 (11%) of these studies were rated as excellent, 8 (19%) as good, 16 (40%) as fair, and 12 (30%) as poor.

Subsequently, the methodological quality of the studies should be taken into account in the evaluation of the results of the included studies. In this phase of the review, the results from different studies are combined [12]. In this systematic review, levels of evidence were used to rate the quality of the instruments, like is done in reviews of randomized clinical trials [13, 14]. In applying levels of evidence, the methodological quality of the studies is taken into account, as well the number of studies and their results. As the results of studies with poor methodological quality cannot be trusted, they do not contribute any evidence, while excellent studies provide strong evidence. The highest level of evidence was applied to the results of studies of excellent methodological quality, and the lowest level of evidence was applied to the results of studies of fair methodological quality [9].

Discussion

In this article, we presented a scoring system for the COSMIN checklist that can be used to obtain an overall score for the methodological quality of a study on a specific measurement property. Four response options for each item of the COSMIN checklist were defined, representing excellent, good, fair, and poor quality. Subsequently, a methodological quality score per box can be obtained by taking the lowest rating of any item in a box ("worst score counts"). A methodological quality score for each study on a measurement property is highly desirable in systematic reviews, as it allows to present conclusions on the quality of the instruments under study accompanied by various levels of evidence.

In the scoring system, items 1 and 2 in most boxes (on the number of missing items and how missing items are handled) are scored less strict than the other items. This information is often not reported in articles. If this lack of information is rated as fair or poor, most studies would get a methodological quality score of fair or poor when using the "worst score counts" principle. We hope that with increasing use of the COSMIN checklist, the reporting of studies on measurement properties will improve. Then, these response options could be reconsidered.

For obtaining a methodological quality score per measurement property, it was decided to take the lowest rating of any item in the corresponding box ("worst score counts"). Three alternative methods were considered but regarded to be less optimal. First, it was considered to rate the methodological quality of a study as good when most, but not all items are adequate and as poor when more than a defined number of items are inadequate. This option, however, seemed against the consensus reached in the COSMIN Delphi study because the COSMIN Delphi panel considered all included items important. Therefore, this option was considered undesirable.

Second, a more simple "worst score counts" algorithm with three response options (good, fair, and poor) instead of four was considered. However, the additional response option "excellent" was considered useful to discriminate between studies in which all items are adequate (which is considered as excellent) and studies in which almost all items are adequate (which is considered as good).

A third alternative method that was considered less optimal was to calculate a "mean score" per box. With this method, each response option is scored (e.g., poor = 0, fair = 1, good = 2, and excellent = 3), and a total score is calculated by summarizing the scores of the completed items and dividing it by the number of completed items. An

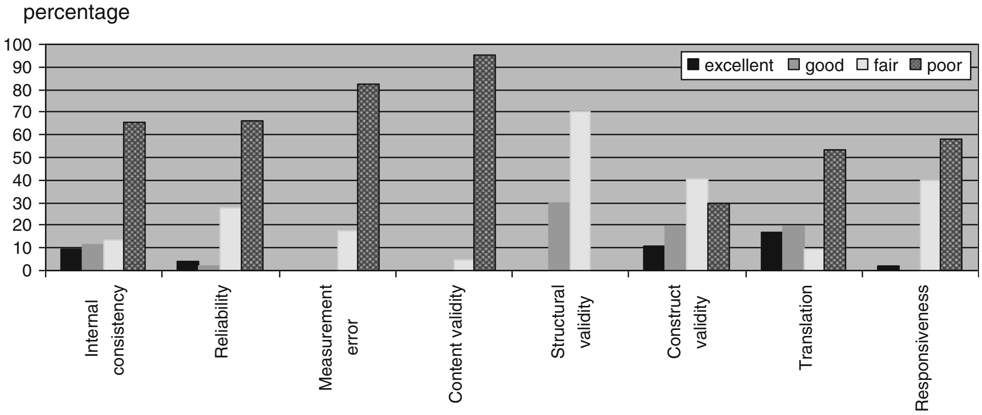

Fig. 1 Percentage of studies with excellent, good, fair, or poor quality. Included number of studies: Internal consistency 35; reliability 36; measurement error 21; content validity 16; structural validity 11; construct validity (hypotheses testing) 41; translation 25; responsiveness 37 (criterion validity was not assessed in any of the studies)

advantage of this method is that the total score is not dependent on the number of items in the box. A disadvantage is that fatal flaws in the design or analyses can be compensated by other good design aspects, which was considered undesirable.

Because the "worst score counts" algorithm was taken into account when developing the response options, we recommend users of the COSMIN scoring system to present no quality ratings on item level using the 4-point scale. For some items, the worst possible response option was defined as good or fair instead of poor because we did not want these items to have too much impact on the methodological quality score per box. This means that no studies will score poor on these items. If scores are presented on item level, this may give the wrong impression that studies score good (or fair) on these items, while in fact, the quality of the study on this aspect is low. If authors want to report on the quality of the studies on item level, we recommend to present the scores dichotomously (as in the original COSMIN checklist). It is not difficult to transform the 4-point scale back to the original dichotomous response options. The response option excellent equals the original response option "yes." The response options good, fair, and poor can be transformed into the response options "no" or "unknown." Exception is the item on sample size where we consider the response option good (a sample size of at least 50) also "yes." This provides the possibility to report on item level which quality criteria were met and which were not met for each study. We have used this approach in an article in which we report on the methodological quality of studies on the measurement properties of neck disability questionnaires, based on data from a systematic review [15].

It should be noted that not all scoring decisions need to be used exactly as defined in this article. For example, each box contains an item referring to whether the sample size included in the analysis was adequate. We presented criteria (e.g., 100 is excellent) as a rule of thumb. We consider this useful, especially for less experienced users of the checklist. However, as we stated in the articles on the development of the COSMIN checklist [7, 8], what is adequate may depend on a number of issues. We therefore recommend that users should make such scoring decisions for their own application. For some items, we therefore did not present any rules of thumb at all, for example, on whether the time interval in a test–retest study is adequate. This is highly dependent on the construct to be measured and should therefore be decided on by the users of the checklist.

The decision for the currently presented scoring system was based on arguments rather than evidence. The validity and reliability of the current scoring system has not yet been assessed. In this study, we compared the scoring system with the opinion of just one rater. Moreover, this rater was the same person that applied the scoring system, which means that the ratings were not independent. However, since the scoring system was developed, it has been or is currently used in (at least) eight reviews (manuscripts in preparation), in which more than 10 different reviewers were involved. The scoring system as presented in this manuscript was considered useful and seemed to have a good face validity. An empirical study comparing the validity of different kind of scoring systems (e.g., by comparing them to an independent overall expert opinion) is currently being performed.

The COSMIN checklist with the presented scoring system is the only tool available at this moment to evaluate the methodological quality of studies on measurement properties in a standardized way. Standards for studies on measurement properties have been published before, such as the criteria proposed by the Scientific Advisory Committee of the Medical Outcomes Trust [16], but they are not presented in a user-friendly checklist and were not developed as a methodological quality assessment tool for systematic reviews. Other checklists that have been published, such as EMPRO [17], and a checklist we previously designed [18] were developed for rating the quality of an instrument rather than the methodological quality of a study. Although in these checklists methodological quality aspects of the study are taken into account in the criteria for the quality of an instrument, the assessment of the methodological quality of a study and the assessment of the quality of an instrument are fundamentally different things and should be performed separately in systematic reviews. This distinction is also increasingly being made in systematic reviews of randomized trials and diagnostic studies [13, 19].

Based on our experience in testing this scoring system in 46 articles and using it in several additional systematic reviews of measurement properties (manuscripts in preparation or submitted for publication), we firmly believe that the COSMIN checklist with the proposed scoring system is a useful tool for assessing the methodological quality of studies included in systematic reviews of measurement properties.

References

1. Mokkink, L. B., Terwee, C. B., Stratford, P. W., Alonso, J., Patrick, D. L., Riphagen, I., et al. (2009). Evaluation of the methodological quality of systematic reviews of health status measurement instruments. *Quality of Life Research, 18*, 313–333.
2. Higgins, J. P. T. & Green, S. (Eds.). (2009). *Cochrane Handbook for Systematic Reviews of Interventions Version 5.0.2 [updated September 2009].* (Available from www.cochrane-handbook.org: Cochrane Collaboration).
3. Haywood, K. L., Garratt, A. M., & Fitzpatrick, R. (2006). Quality of life in older people: A structured review of self-assessed health instruments. *Expert Review of Pharmacoeconomics and Outcomes Research, 6*, 181–194.
4. Alla, S., Sullivan, S. J., Hale, L., & McCrory, P. (2009). Self-report scales/checklists for the measurement of concussion symptoms: a systematic review. *British Journal of Sports Medicine, 43*(Suppl 1), i3–i12.
5. Marinus, J., Ramaker, C., van Hilten, J. J., & Stiggelbout, A. M. (2002). Health related quality of life in Parkinson's disease: A systematic review of disease specific instruments. *Journal of Neurology, Neurosurgery and Psychiatry, 72*, 241–248.
6. Wind, H., Gouttebarge, V., Kuijer, P. P. F. M., & Frings-Dresen, M. H. W. (2005). Assessment of functional capacity of the musculoskeletal system in the context of work, daily living, and sport: A systematic review. *Journal of Occupational Rehabilitation, 15*, 253–272.
7. Mokkink, L. B., Terwee, C. B., Patrick, D. L., Alonso, J., Stratford, P. W., Knol, D. L., et al. (2010). The COSMIN checklist for assessing the methodological quality of studies on measurement properties of health status measurement instruments: An international Delphi study. *Quality of Life Research, 19*, 539–549.
8. Mokkink, L. B., Terwee, C. B., Knol, D. L., Stratford, P. W., Alonso, J., Patrick, D. L., et al. (2010). The COSMIN checklist for evaluating the methodological quality of studies on measurement properties: A clarification of its content. *BMC Medical Research Methodeology, 10*, 22.
9. Schellingerhout, J. M., Verhagen, A. P., Heymans, M. W., Koes, B. W., de Vet, H. C. W., & Terwee, C. B. (2011). Measurement properties of disease-specific questionnaires in patients with neck pain: a systematic review. *Quality of Life Research*. doi: 10.1007/s11136-011-9965-9.
10. Stevens, J. (1996). *Applied multivariate statistics for the social sciences*. Mahway, NJ: Lawrence Erlbaum.
11. de Vet, H. C. W., Ader, H. J., Terwee, C. B., & Pouwer, F. (2005). 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. *Quality of Life Research, 14*, 1203–1218.
12. de Vet, H. C. W., Terwee, C. B., Mokkink, L. B., & Knol, D. J. (2011). *Measurement in medicine. A practical guide*. Cambridge: Cambridge University Press.
13. Guyatt, G. H., Oxman, A. D., Vist, G. E., Kunz, R., Falck-Ytter, Y., Onso-Coello, P., et al. (2008). GRADE: An emerging consensus on rating quality of evidence and strength of recommendations. *British Medical Journal, 336*, 924–926.
14. Furlan, A. D., Pennick, V., Bombardier, C., & van Tulder, M. (2009). 2009 updated method guidelines for systematic reviews in the Cochrane Back Review Group. *Spine (Phila Pa 1976), 34*, 1929–1941.
15. Terwee, C. B., Schellingerhout, J. M., Verhagen, A. P., Koes, B. W., & de Vet, H. C. W. (2011). Methodological quality of studies on the measurement properties of neck pain and disability questionnaires: a systematic review. *Journal of Manipulative and Physiological Therapeutics, 34*, 261–272.
16. Scientific Advisory Committee of the Medical Outcomes Trust. (2002). Assessing health status and quality-of-life instruments: Attributes and review criteria. *Quality of Life Research, 11*, 193–205.
17. Valderas, J. M., Ferrer, M., Mendivil, J., Garin, O., Rajmil, L., Herdman, M., et al. (2008). Development of EMPRO: A tool for the standardized assessment of patient-reported outcome measures. *Value in Health, 11*, 700–708.
18. Terwee, C. B., Bot, S. D. M., de Boer, M. R., van der Windt, D. A. W. M., Knol, D. L., Dekker, J., et al. (2007). Quality criteria were proposed for measurement properties of health status questionnaires. *Journal of Clinical Epidemiology, 60*, 34–42.
19. Brozek, J. L., Akl, E. A., Jaeschke, R., Lang, D. M., Bossuyt, P., Glasziou, P., et al. (2009). Grading quality of evidence and strength of recommendations in clinical practice guidelines: Part 2 of 3. The GRADE approach to grading quality of evidence about diagnostic tests and strategies. *Allergy, 64*, 1109–1116.