

Guidelines for Reporting Reliability and Agreement Studies (GRRAS) were proposed

Jan Kottner^{a,*}, Laurent Audigé^b, Stig Brorson^c, Allan Donner^d, Byron J. Gajewski^e,
Asbjørn Hróbjartsson^f, Chris Roberts^g, Mohamed Shoukri^h, David L. Streinerⁱ

^aDepartment of Nursing Science, Centre for Humanities and Health Sciences, Charité-Universitätsmedizin Berlin, Berlin, Germany

^bAO Clinical Investigation and Documentation, Dübendorf, Switzerland

^cDepartment of Orthopaedic Surgery, Herlev University Hospital, Herlev, Denmark

^dDepartment of Epidemiology and Biostatistics, Schulich School of Medicine and Dentistry, The University of Western Ontario, London, Ontario, Canada

^eDepartment of Biostatistics, University of Kansas Schools of Medicine & Nursing, Kansas City, KS, USA

^fThe Nordic Cochrane Centre, Rigshospitalet, Copenhagen, Denmark

^gSchool of Community Based Medicine, The University of Manchester, Manchester, UK

^hDepartment of Biostatistics, Epidemiology and Scientific Computing, King Faisal Specialist Hospital and Research Center,

The University of Western Ontario, London, Ontario, Canada

ⁱDepartment of Psychiatry, University of Toronto, Toronto, Ontario, Canada

Accepted 2 March 2010

Abstract

Objective: Results of reliability and agreement studies are intended to provide information about the amount of error inherent in any diagnosis, score, or measurement. The level of reliability and agreement among users of scales, instruments, or classifications is widely unknown. Therefore, there is a need for rigorously conducted interrater and intrarater reliability and agreement studies. Information about sample selection, study design, and statistical analysis is often incomplete. Because of inadequate reporting, interpretation and synthesis of study results are often difficult. Widely accepted criteria, standards, or guidelines for reporting reliability and agreement in the health care and medical field are lacking. The objective was to develop guidelines for reporting reliability and agreement studies.

Study Design and Setting: Eight experts in reliability and agreement investigation developed guidelines for reporting.

Results: Fifteen issues that should be addressed when reliability and agreement are reported are proposed. The issues correspond to the headings usually used in publications.

Conclusion: The proposed guidelines intend to improve the quality of reporting. © 2011 Elsevier Inc. All rights reserved.

Keywords: Agreement; Guideline; Interrater; Intrarater; Reliability; Reproducibility

1. Background

Reliability and agreement are important issues in classification, scale and instrument development, quality assurance, and in the conduct of clinical studies [1–3]. Results of reliability and agreement studies provide information about the amount of error inherent in any diagnosis, score, or measurement, where the amount of measurement error determines the validity of the study results or scores [1,3–6].

The terms “reliability” and “agreement” are often used interchangeably. However, the two concepts are

conceptually distinct [7–12]. Reliability may be defined as the ratio of variability between subjects (e.g., patients) or objects (e.g., computed tomography scans) to the total variability of all measurements in the sample [1,6]. Therefore, reliability can be defined as the ability of a measurement to differentiate between subjects or objects. On the other hand, agreement is the degree to which scores or ratings are identical. Both concepts are important, because they provide information about the quality of measurements. Furthermore, the study designs for examining the two concepts are similar. We focus on two aspects of these concepts:

- Interrater agreement/reliability (different raters, using the same scale, classification, instrument, or procedure, assess the same subjects or objects)
- Intrarater agreement/reliability (also referred to as test–retest) (the same rater, using the same scale,

* Corresponding author. Department of Nursing Science, Centre for Humanities and Health Sciences, Charité-Universitätsmedizin Berlin, Augustenburger Platz 1, 13353 Berlin, Germany. Tel.: +49-30-450-529-054; fax: +49-30-450-529-900.

E-mail address: jan.kottner@charite.de (J. Kottner).

What is new?**Key finding**

- Reporting of interater/intrarater reliability and agreement is often incomplete and inadequate.
- Widely accepted criteria, standards, or guidelines for reliability and agreement reporting in the health care and medical fields are lacking.
- Fifteen issues that should be addressed when reliability and agreement are reported are proposed.

What this adds to what is known

- There is a need for rigorous interrater and intrarater reliability and agreement studies to be performed in the future.
- Systematic reviews and meta-analysis of reliability and agreement studies will become more frequent.

What is the implication and what should change now?

- Reporting of reliability and agreement studies should be improved.
- The proposed guidelines help to improve reporting.

classification, instrument, or procedure, assesses the same subjects or objects at different times).

- Issues regarding internal consistency are not addressed here.

When reporting the results of reliability and agreement studies, it is necessary to provide information sufficient to understand how the study was designed and conducted and how the results were obtained. Reliability and agreement are not fixed properties of measurement tools but, rather, are the product of interactions between the tools, the subjects/objects, and the context of the assessment. Reliability and agreement estimates are affected by various sources of variability in the measurement setting (e.g., rater and sample characteristics, type of instrument, administration process) and the statistical approach (e.g., assumptions concerning measurement level, statistical model). Therefore, study results are only interpretable when the measurement setting is sufficiently described and the method of calculation or graphical presentation is fully explained.

After reviewing many reliability and agreement studies, it becomes apparent that important information about the study design and statistical analysis is often incomplete [2,12–21]. Because of inadequate reporting, interpretation and synthesis of the results are often difficult. Moreover,

widely accepted formal criteria, standards, or guidelines for reliability and agreement reporting in the health care and medical fields are lacking [2,22,23]. Authors of reviews have often established their own criteria for data extraction and quality assessment [12,15,17,19,21–25]. The American Educational Research Association, American Psychological Association (APA), and the National Council on Measurement in Education have attempted to improve the reporting of reliability and agreement, but they focus on psychological testing [26].

2. Project

In the absence of standards for reporting reliability and agreement studies in the medical field, we evolved the idea that formal guidelines might be useful for researchers, authors, reviewers, and journal editors. The lead author initially contacted 13 experts in reliability and agreement investigation and asked whether they saw a need for such guidelines and whether they wished to take part in this project. The experts were informally identified based on substantial contributions and publications in the field of agreement and reliability. Each expert was also asked whether he knew an additional expert who would be interested in participating. Finally, from all these individuals, an international group of eight researchers, who were experienced in instrument development and evaluation (L.A., D.L.S., S.B., and A.H.), reliability and agreement estimation (A.D., B.J.G., C.R., and M.S.), or in systematic reviews (L.A., S.B., and A.H.) of reliability studies, was formed.

The objective was to develop guidelines for reporting reliability and agreement studies. The specific aim was to establish items that should be addressed when reliability and agreement studies are reported.

The development process contained elements of Glaser's state-of-the-art method [27] and of the nominal group technique [28]. Based on an extensive literature review, the project coordinator (J.K.) produced a first draft of guidelines, including an initial list of possible items. It was sent to all team members, and they were asked to review, comment, change the document and wording, and to indicate whether they agree or disagree. Feedback indicated that there was a need for clarification of key concepts that are related to reliability and agreement studies. Therefore, definitions of key concepts were discussed among the team members, and definitions were established (Appendix). Based on the comments from the first review, a second draft was created and again sent to team members, accompanied by a summary report of all criticisms, discussions, and suggestions that had been made. Based on the critiques of the second draft, a third draft was created, reviewed, and discussed again. After the review of the fourth draft, consensus was achieved, and team members approved these guidelines in their current form.

Table 1
Guidelines for Reporting Reliability and Agreement Studies (GRRAS).

TITLE AND ABSTRACT	1. Identify in title or abstract that interrater/intrarater reliability or agreement was investigated.
INTRODUCTION	2. Name and describe the diagnostic or measurement device of interest explicitly.
	3. Specify the subject population of interest.
	4. Specify the rater population of interest (if applicable).
	5. Describe what is already known about reliability and agreement and provide a rationale for the study (if applicable).
	6. Explain how the sample size was chosen. State the determined number of raters, subjects/objects, and replicate observations.
METHODS	7. Describe the sampling method.
	8. Describe the measurement/rating process (e.g. time interval between repeated measurements, availability of clinical information, blinding).
	9. State whether measurements/ratings were conducted independently.
	10. Describe the statistical analysis.
RESULTS	11. State the actual number of raters and subjects/objects which were included and the number of replicate observations which were conducted.
	12. Describe the sample characteristics of raters and subjects (e.g. training, experience).
	13. Report estimates of reliability and agreement including measures of statistical uncertainty.
DISCUSSION	14. Discuss the practical relevance of results.
AUXILIARY MATERIAL	15. Provide detailed results if possible (e.g. online)

3. Guidelines

The Guidelines for Reporting Reliability and Agreement Studies (GRRAS) are shown in Table 1. They contain issues that should be addressed when reliability and agreement are investigated. The underlying rationale, arguments, or empirical data to support each item are given later. The proposed issues correspond to the headings and order usually used in publications. The items aim to cover a broad range of clinical test scores, classifications or

diagnosis. However, some items are only partly applicable to self-completed questionnaires (e.g., items 4, 6, 11).

Studies may be conducted with the primary focus on reliability and agreement estimation itself or they may be a part of larger diagnostic accuracy studies, clinical trials, or epidemiological surveys. In the latter case, researchers report agreement and reliability as a quality control, either before the main study or by using data of the main study. Typically, results are reported in just

a few sentences, and there is usually only limited space for reporting. Nevertheless, it seems desirable to address all issues listed in the following sections to allow data to be as useful as possible. Therefore, reliability and agreement estimates should be reported in another publication or reported as part of the main study. Our guidelines focus on reporting in detail, but we provide some minimum requirements for reporting results as a minor part of larger studies as well.

3.1. Title and abstract

3.1.1. Item 1: Identify in title or abstract whether interrater/intrarater reliability or agreement was investigated.

Rationale: Bibliographic databases and the Internet have now become the primary resources for searching for evidence. To use this evidence, rapid and clear identification of reliability and agreement studies is necessary. We recommend using the terms “interrater/intrarater reliability/agreement” in the title or abstract. The term “rater” seems suitable to characterize a wide range of situations, in which persons make judgments about other persons or objects. The terms “reliability” and “agreement” refer, in particular, to these kinds of studies.

Today, specific searches for these types of studies are limited, because a variety of different terms for reliability and agreement studies are used [29]: interobserver variation [30], observer variation [31], interreader reliability [32], intercoder reliability [33], intercoder agreement [34], inter-examiner reliability [35], intertester reliability [36], repeatability [37], and others. In the hierarchical classification of Medical Subject Heading Terms (MeSH) of the U.S. National Library of Medicine, the entry term “reliability” refers to the MeSH “reproducibility of results,” and the entry terms “interobserver” and “intraobserver” refer to the MeSH “observer variation.” The entry term “agreement” does not refer to agreement studies, and there are no entries at all for the terms “interrater” and “intrarater.” In other databases, such as the Educational Resource Information Center, the terms “interrater” or “intrarater” are not indexed. On the other hand, “interrater,” “reliability,” and “interrater reliability” are subject headings in the Excerpta Medica database (EMBASE) and in the Thesaurus of Psychological Index Terms used in the database provided by the APA (PsycINFO). Neither in EMBASE nor in PsycINFO is the term “agreement” used for agreement studies. To overcome the diversity of applied terms and to enhance communication and research, we suggest that MeSH terms for reliability and agreement studies need to be fixed in the future.

Reliability and agreement studies that take place and are reported as part of a larger study should be identified in the abstract or keywords as well, because they provide empirical evidence regarding measurement error. Systematic reviews on interrater reliability studies have reported that many reliability and agreement estimates are obtained during the course of larger cross-sectional or prospective study designs.

Most of these investigations would have been missed when specific search terms, such as “reliability” or “agreement,” were not used [16]. If indexing is absent, these reliability and agreement investigations are hardly detectable. It should be noted that original keywords provided by authors or publishers that do not correspond to the terminology or taxonomy used by indexing databases will get lost but they are valuable when using other search strategies (e.g., hand searching) or databases (e.g., Science direct, Springerlink).

3.2. Introduction

3.2.1. Item 2: Name and describe the diagnostic or measurement device of interest explicitly.

Rationale: The degree of reliability/agreement is related to the properties of instruments and classifications or scales that are used [1,5,38–40]. However, measurement devices exist in various versions and languages and many have been adapted several times. Therefore, definitions of items or categories should be made explicit, because raters may have different understandings of the wording used, thus, creating difficulties in the interpretation of concepts [33,34,41]. Additionally, there may be several definitions for the same measured construct [42,43]. Readers must know exactly which instrument or scale, and which version, was applied. A standard reference is insufficient for that.

In the case of categorical classifications, the total number of categories used must be stated, because the value and interpretation of interrater reliability/agreement coefficients are related to this number [44,45]. Furthermore, many classification systems were designed to measure the severity of a disease or conditions (e.g., pressure ulcer classifications: grades 1–4; intensity of phobic anxiety symptoms: mild, moderate, and severe). When such classifications are applied to nonaffected persons, it is common to use additional categories, such as “not affected.” Under these circumstances, readers must know exactly how many and which categories were actually applied (e.g., pressure ulcer classifications: grades 0–4) [16,38].

In the case of continuous measurements, the value of interrater reliability/agreement coefficients depends on their range [46,47]. When continuous measurements are converted into categorical data and split into distinct categories (e.g., normal blood pressure or hypertension), authors should state explicitly the chosen cutoff scores.

The aforementioned statements hold good in situations where the investigated measurement device exists and has already been published. If a new instrument or scale is being developed, the Introduction section should contain the rationale and justification for this. The detailed description of the new tool should be part of the Methods section.

3.2.2. Item 3: Specify the subject population of interest.

Rationale: Measurement or diagnostic devices were designed for special populations (e.g., care settings, age groups, stages of disease). Moreover, characteristics of

subjects affect the interpretation of reliability and agreement, because the resulting coefficients are closely linked to this population. Reliability and agreement coefficients are population specific and depend on the prevalence of the studied characteristic or trait [1,3,39,46,48].

3.2.3. Item 4: Specify the rater population of interest (if applicable).

Rationale: Classifications and instruments or scales that are not self-completed are designed for persons working in various fields, having varying levels of training, and under specific conditions. Usually, the rater population of interest will be all persons working in these areas and possibly using the instrument in question. This population should be characterized by rater qualifications, clinical background, knowledge, degree of expertise, and training, as these characteristics may have a major impact on reliability and agreement estimates [3,35,39,40,49,50].

Usually, the focus of reliability and agreement studies as part of larger studies involves the measurement of reliability or agreement among researchers, research assistants, or all other raters, who are responsible for the data collection [3]. In these cases, the raters involved are the only raters of interest. Nevertheless, rater characteristics (e.g., clinical skills, training) should be described on the grounds that they potentially can influence results [38,51] and such information is needed in later reliability generalization studies [6].

3.2.4. Item 5: Describe what is already known about reliability and agreement and provide a rationale for the study (if applicable).

Rationale: For studies carried out with existing scales or instruments, readers should be provided with an overview of existing evidence about reliability and agreement. This should be accomplished by a review of the literature. Systematic reviews and reliability generalization studies should be the preferred sources. It should be explained why this new study was necessary and why it was important to investigate agreement and reliability in this situation. What will be added to existing knowledge?

3.3. Methods

3.3.1. Item 6: Explain how the sample size was chosen. State the determined number of raters, subjects/objects, and replicate observations.

Rationale: Although investigations into sample size determination for reliability and agreement studies are small in number, some suggestions have appeared in the literature [52–59]. Note that, in studies investigating scores of self-administered questionnaires, sample size determination refers to the subjects only.

Situations may arise where the predetermined number of replicate observations, subjects, or raters cannot be achieved because of organizational, financial, or ethical constraints [60]. Furthermore, in smaller reliability and

agreement studies, the maximum possible number of raters may be determined by the design of the main study. This information should be made explicit to make the study transparent and credible.

3.3.2. Item 7: Describe the sampling method.

Rationale: Enrollment of subjects in interrater reliability and agreement studies is often not clearly stated [17]. The sampling method (e.g., at random, consecutive, convenient) for both rated subjects and raters should be stated, because it has implications for the statistical analysis [1,5,61] and guides the reader in generalizing the reported results [62]. We further suggest that authors should explain in detail what “random,” “consecutive,” and “convenient” mean in their study. Recently, Zegers et al. conducted an interrater agreement study of the results of patient-record reviews [63]. They stated “In this study, 55 trained physicians reviewed in several different hospitals (average 5.2 hospitals per physician)” (p. 96). Even though the authors provide detailed eligibility criteria, the sampling method is not clear.

3.3.3. Item 8: Describe the measurement/rating process (e.g., time interval between repeated measurements, availability of clinical information, blinding).

Rationale: It is important to provide readers with sufficient information regarding the measurement/rating process, because reliability and agreement estimates may vary according to the time interval between repeated measurements; the general conditions underlying the measurement situation (e.g., atmosphere, location); the specific measurement setting (e.g., imaging modalities, light); or the complexity of the measurement/rating process or characteristics of the rated subjects themselves [1,8,24,49,64–67]. Standardization of the measurement/rating process helps to prevent bias [35], but when instruments or classifications are to be used in broader clinical contexts and in daily practice, reliability and agreement should also be investigated under conditions as close as possible to the clinical daily routine or other natural setting [19,68–70].

The completeness of clinical information about a person’s health status, identity, gender, age, or history can also influence the diagnostic process and, consequently, the assessment result. It should be stated what information was given in which way, whether raters were blinded to subject/object information; and how was the blinding organized [17,22,23]. However, describing the availability of clinical information does not mean that these data must be described or analyzed in detail. It should also be stated whether raters were blinded in the sense that they were not aware that their judgments will be compared with those of other raters, removing the possibility of a Hawthorne effect (i.e., ensuring that the rater’s behavior is not altered because of an awareness being observed) [4,71]. A sufficient description of the measurement process, including

Table 2
Statistical methods for analyzing interrater/intrarater reliability and agreement studies

Level of measurement	Reliability measures	Agreement measures
Nominal	Kappa statistics	Proportions of agreement Proportions of specific agreement
Ordinal	Ranked intraclass correlation Matrix of kappa coefficients Weighted kappa	Proportions of agreement Proportions of specific agreement
Continuous	Intraclass correlation coefficients	Proportions of agreement (ranges) Proportions of specific agreement (ranges) Standard errors of measurement Coefficients of variation Bland–Altman plots and limits of agreement

information regarding blinding, is missing in many research reports [17,24].

In addition, it should be stated if the final measure or rating to be used result from single or repeated measures. To increase reliability, it is often recommended to use the mean of two or more raters rating the same persons or objects in the study [72–74]. Such mean ratings can be assessed for reliability in a similar way as single ratings. Consensus ratings or the mean of repeated measures usually show higher reliabilities [6,61]. In those cases, it must be stated whether reliability coefficients are based on the average of ratings and scores or whether reliability measures refer to the score of a single rater [61]. This information is an integral part of the measurement/rating process.

The measurement/rating process in reliability and agreement investigations as part of larger studies should be as similar as possible to that of the main study. Otherwise, results may not adequately reflect the measurement error of the main study [75,76].

3.3.4. Item 9: State whether measurements/ratings were conducted independently.

Rationale: When diagnostic information, scores, or other test results in clinical practice are obtained by individual raters alone, the measurements or ratings in the study should be conducted similarly as well. This is important, because the magnitude of reliability and agreement coefficients may be directly influenced by the study setting. For example, in studies where raters simultaneously conduct an assessment or scoring exercise, no communication should be allowed. However, some form of communication under these study conditions may still have taken place. Thus, two raters may agree with each other more often when they are aware of each other's assessments [21,77]. On the other hand, repeated scorings or ratings should not be conducted independently in every situation, because many decisions are made by groups. If one is interested in comparing decisions or scorings between such groups, then the authors should, instead, describe the degree of independence among the groups involved. Systematic reviews have revealed that information regarding the independence of classifications or scorings is frequently missing [15–17,22].

3.3.5. Item 10: Describe the statistical analysis.

Rationale: There are several statistical approaches that may be used in the measurement of reliability and agreement. Because they were often developed within different disciplines, no single approach can be regarded as standard. Every method is also based on assumptions concerning the type of data (nominal, ordinal, continuous), the sampling (at random, consecutive, convenience), and on the treatment of random and systematic error [1,5,46]. Therefore, it is not possible to be too prescriptive regarding the “best” statistical method, with the choice depending on the purpose and the design of the study.

Table 2 lists frequently applied statistical approaches as arranged by the level of measurement [78] and by the use of reliability vs. agreement statistics. Kappa-like statistics provide useful information about reliability for categorical data [1,44]. However, there are several types of kappa statistics, including Cohen's kappa, Cohen's weighted kappa, and the intraclass kappa statistic. Inference procedures also vary depending on the particular kappa statistic adopted, for example, the goodness-of-fit approach for the intraclass kappa statistic [56]. Kappa coefficients have been frequently criticized for their dependence on the rater prevalence, but, as with other measures of reliability or diagnostic accuracy, this behavior exactly reflects the population specificity. Low kappa values indicate the inability of the investigated measure or classification to make clear distinctions between subjects of a population in which those distinctions are very rare or difficult to achieve [44,79]. In addition, it might reflect the inability of raters to distinguish between adjacent categories [80].

Ordinal measurements are common in research and practice. Reliability calculations for such data have been proposed by Whitfield [81], Rothery [82], Müller and Büttner [46], and Roberts and McNamee [83].

The intraclass correlation coefficient (ICC) based on analysis of variance (ANOVA) models and kappa statistics using quadratic weights may be adopted for measuring the reliability of continuous scales. ANOVA models are typically justified by assuming normally distributed errors. The treatment of sampling errors because of different raters is crucial for the appropriate selection of an ICC [61,84]. Moreover, although the ICC is reported in many research

reports, it is often not clear which ICC was used [14,43]. When continuous measurements are split into distinct categories (see item 2), it is recommended that results be calculated for the continuous measurement as well, because the transformation of continuous values into categories may cause difficulties in interpretation and lead to a reduction in statistical power [32,58,85].

The proposed measures of agreement include proportions of exact agreement [8,86,87], proportions of specific agreement [86,87], repeatability coefficients, and the graphical method proposed by Bland and Altman [88]. For continuous measurements, standard errors of measurement [8,89] and proportions of agreement within specified limits provide useful information as well [8].

When reliability or agreement data are collected in a clustered fashion, for example, in multicenter studies, it should be reported whether the data have been pooled and, if so, which pooling procedure was used. Proposals for summarizing reliability coefficients from different groups or samples have been made [5,90–94]. Although not frequently done (e.g., Bååth et al. [95]), the heterogeneity between multiple centers should be reported, because empirical evidence suggests that it is almost always present [16,92,95].

There are other approaches that might also be used (e.g., coefficients of variation, item response theory, or the “signal to noise ratio” [96]). Researchers should clearly state a priori their assumptions, why a certain approach was chosen, and what was intended to be demonstrated. Finally, the statistical software used should be reported.

3.4. Results

3.4.1. Item 11: State the actual number of raters and subjects/objects that were included and the number of replicate observations that were conducted.

Rationale: These numbers are necessary to evaluate the precision of the study and to make further calculations (e.g., in meta-analysis) possible [20,31,58,59,97,98]. A flow diagram allows readers to follow the inclusion and exclusion process from the intended sample of raters and subjects to the actual sample. This information also provides some information about the generalizability of results. Finally, features of the data collection dealing with crossings of raters and subjects/objects help readers to decide whether the statistical analysis was appropriate [1,61]. However, in the case of self-administered questionnaires, only the number of respondents needs to be provided.

Recently, Bates-Jensen et al. [99] investigated whether subepidermal moisture measures can be used to predict skin damage using, among others, a four-stage pressure ulcer classification: “Interrater agreement was assessed on 98 pairs of observations. For erythema presence kappas ranged from 0.70 to 1.00 across anatomic sites and for erythema severity (blanchable vs. nonblanchable) kappas ranged from 0.59 to 1.00. Interrater agreement on stage was 1.00 on 10 pressure ulcers” (p. 191). In addition to the fact that

it remains unclear which kappa statistic was applied, it is impossible to understand how many raters, patients, skin sites, and types of ulcers were involved.

3.4.2. Item 12: Describe the sample characteristics of raters and subjects (e.g., training, experience).

Rationale: Sample characteristics should be described. This information helps to evaluate whether a representative sample of raters and subjects was included [19,43,68] and whether results may be extrapolated to other populations [25]. Participating raters should be similar to the intended users. Additionally, information about the “true” prevalence, the severity of the rated characteristic, or the actual number of categories is helpful in characterizing the sample [100].

3.4.3. Item 13: Report estimates of reliability and agreement, including measures of statistical uncertainty.

Rationale: As there are various statistical approaches that can be adopted, it must be made clear what the calculated numeric expressions mean. Recent reviews revealed that the type of coefficient obtained is sometimes reported ambiguously [14,16]. Statements, such as “The percentage agreement ... was $r = 0.83$ to $r = 0.96$ ” [101] or “A data collection team ... were trained ... to achieve an interrater reliability of 0.90” [102], are insufficient.

Single summary measures of reliability and agreement provide only limited information [7,14,19,31]. We recommend reporting a combination of coefficients (e.g., kappa statistics and percentage of agreement), which allow the reader to get a detailed impression of the degree of the reliability and agreement. Graphical methods (e.g., Bland–Altman) also provide useful information about the distribution of scores.

Confidence intervals as measures of statistical uncertainty should be reported, because the ranges of values that are considered to be plausible for the population of interest are useful for interpreting results [56,103]. Where investigators wish to demonstrate a satisfactory level of reliability and agreement, particular attention should be given to the interpretation of the lower limit [104].

3.5. Discussion

3.5.1. Item 14: Discuss the practical relevance of results.

Rationale: There are various suggestions in the literature regarding the degree to which reliability or agreement coefficients can be labeled as “poor” or “perfect,” “low” or “high,” or whether the reliability/agreement is “high enough” [22,60,86,105,106]. Although these guidelines are clearly arbitrary [38,105], they have been widely adopted in the reporting of results. As an example, Zegers et al. (2010) stated “A κ -value between 0.00 and 0.20 was classified as ‘slight’; between 0.21 and 0.40 as ‘fair’; between 0.41 and 0.60 as ‘moderate’; between 0.61 and 0.80 as ‘substantial’; and between 0.81 and 1.00 as ‘almost perfect’” [63]. Nevertheless, these “labels” do not indicate the practical or clinical relevance of results [19,25]. In other words, even if one obtains

high reliability or agreement coefficients, disagreements might have occurred, which are clinically unacceptable. The magnitude of acceptable differences between scorings or ratings is not solely a statistical decision but also a clinical one. In clinical practice, it depends on the purpose and consequences of test results, scores, or diagnostic results regarding how much error will be allowed to be introduced into the clinical decision making [1,7,44,107].

Values of 0.60, 0.70, or 0.80 are often used as the minimum standards for reliability coefficients, but this may be only sufficient for group-level comparisons or research purposes [12,58,108]. For example, ICC values for a scale measuring pressure ulcer risk should be at least 0.90 or higher when applied in clinical practice [108]. If individual and important decisions are made on the basis of reliability estimates, values should be at least 0.90 [4] or 0.95 [109].

Finally, results should be interpreted in terms of influencing factors. Authors should state what could and should be done to improve results. There are various studies concluding that reliability and agreement are poor but provide little help as to what should be done next [22].

3.6. Auxiliary material

3.6.1. Item 15: Provide detailed results if possible (e.g., online).

Rationale: Considering the variety of factors influencing reliability and agreement estimates (rater and sample characteristics, instruments, statistical methods), it is evident that single summary measures provide only limited information. Thus, systematic reviews and the meta-analysis of reliability and agreement studies will likely become more frequent. Detailed results or even raw data are valuable resources for recalculations and meta-analysis [92,93]. For instance, Stockendahl et al. [19], in conducting a meta-analysis of reliability coefficients of binary data, were unable to decide whether kappa values were influenced by differences in observed prevalence between raters or by lack of agreement. Presentation of relevant fourfold tables would have solved this problem, perhaps, presented as auxiliary material. Otherwise, authors should carefully consider the way in which results are presented in the article.

4. Discussion

The level of reliability and agreement among users of scales, instruments, or classifications in many different areas is largely unknown [15,16,18,100,110–112]. Therefore, there is a clear need for rigorous interrater and intrarater reliability and agreement studies to be performed in the future. Studies are also needed for investigating reliability in clinical practice [16,25,36,43]. We hope that the guidelines will help to improve the quality of reporting.

To our knowledge, no document focusing on reporting of reliability and agreement studies in the medical field

has yet been published. However, there is some overlap of the present guidelines with the Standards for Reporting of Diagnostic Accuracy (STARD) [113] and with the Standards for Educational and Psychological Testing [26].

The STARD [113] contains a checklist of essential elements of diagnostic accuracy studies that must be reported. Diagnostic accuracy studies differ from reliability or agreement studies in comparing one or more test results with results obtained with a reference value obtained on the same subject. In interrater and intrarater reliability/agreement studies, results are compared from the same or from different raters rating the same subjects or objects and using the *same* scale, method, or classification. In these kinds of studies, raters or methods are treated symmetrically [86]. No rater or method is considered as a reference standard [114]. Reliability/agreement estimates provide information about the degree of measurement error in the results, not of the validity of the results. Additionally, STARD gives only limited guidance on the reporting reliability and agreement coefficients.

The purpose of the Standards for Educational and Psychological Testing [26] is to provide criteria for the evaluation of tests, testing practices, and the effects of test use. Reliability and measurement error issues are addressed in 20 of these standards, but emphasis here is placed on psychological tests in which an examinee's behavior is evaluated and scored using a standardized process. Measurement and instrument issues (e.g., scale units, reporting of different standard errors) are discussed in great detail, whereby other issues (e.g., indexing, sampling) are not considered.

Our recommendations aim to cover the reporting of reliability and agreement studies over a wide range of disciplines, especially in health care. Today, there are no established standards in this field.

5. Limitations

We chose a pragmatic approach in developing the guidelines. Eight experts participated, and they were blinded to each other in the first round only. It is commonly assumed that Delphi methods are more reliable, because the group interaction is indirect and more people can be involved [115]. Furthermore, no single expert with a strong opinion and ego can override the opinion of the other experts. However, consensus achieved by Delphi methods also heavily depends on the participating experts, techniques of summarizing and presenting the group response, and on how disagreements are resolved [116]. It has also been shown that Delphi methods do not result in different outcomes when compared with the Nominal group method and that groups of up to 12 participants can achieve agreement [117,118]. In our multidisciplinary group, discussions among group members were allowed. An understanding of reasons for disagreement was possible, and new ideas were developed, discussed, and incorporated in the guidelines.

Because we provide only guidelines for reporting studies, formal validation approaches are not applicable. The

only possible validation of our guidelines would be to investigate whether another group of experts with comparable levels of expertise in a comparable situation would have produced the same results, which would be very impractical. However, we do strongly encourage users of these guidelines to comment on and criticize our work so as to improve it accordingly.

6. Conclusions

Interrater and intrarater reliability and agreement examinations are needed to estimate the amount of error in the rating or scoring of tests and classification procedures. We have proposed a set of general guidelines for reporting reliability and agreement studies. The guidelines are broadly useful and applicable to the vast majority of diagnostic issues. We believe that this first draft may be improved upon and updated in the future. We appreciate any comments or suggestions by readers and users.

Appendix

Concepts related to reliability and agreement studies

Concepts	Definitions
Agreement	Agreement is the degree to which scores or ratings are identical.
Interrater agreement	Interrater agreement is the degree to which two or more raters achieve identical results under similar assessment conditions.
Interrater reliability	Interrater reliability is the degree to which two or more raters are able to differentiate among subjects or objects under similar assessment conditions.
Rater	Every person who makes a judgment about a person or object.
Reliability	Reliability is the ability of scores of a measuring device to differentiate among subjects or objects.
Repeatability	Repeatability is the degree of how close scores or ratings obtained under similar conditions are.
Test–retest reliability (intrarater reliability)	Test–retest reliability is the degree to which a measurement device is able to differentiate among subjects or objects under repeated similar assessment conditions. Synonymous with intrarater reliability.

References

- [1] Dunn G. Statistical evaluation of measurement errors: design and analysis of reliability studies. 2nd edition. London, UK: Arnold; 2004.
- [2] Mulsant BH, Kastango KB, Rosen J, Stone RA, Mazumdar S, Pollock BG. Interrater reliability in clinical trials of depressive disorders. *Am J Psychiatry* 2002;159:1598–600.
- [3] Szklo M, Nieto FJ. Epidemiology beyond the basics. 2nd edition. Sudbury, MA: Jones and Bartlett; 2007.
- [4] Polit DF, Beck CT. Nursing research: generating and assessing evidence for nursing practice. 8th edition. Philadelphia, PA: Lippincott Williams & Wilkins; 2008.
- [5] Shoukri MM. Measures of interobserver agreement. Boca Raton, FL: Chapman & Hall/CRC; 2004.
- [6] Streiner DL, Norman GR. Health measurement scales: a practical guide to their development and use. 4th edition. Oxford, UK: Oxford University Press; 2008.
- [7] Bland JM, Altman DG. A note on the use of the intraclass correlation coefficient in the evaluation of agreement between two methods of measurement. *Comput Biol Med* 1990;20:337–40.
- [8] De Vet HCW, Terwee CB, Knol DL, Bouter LM. When to use agreement versus reliability measures. *J Clin Epidemiol* 2006;59:1033–9.
- [9] Feinstein AR, Cicchetti DV. High agreement but low kappa: I. the problems of two paradoxes. *J Clin Epidemiol* 1990;43:543–9.
- [10] Gwet KL. Computing inter-rater reliability and its variance in the presence of high agreement. *Br J Math Stat Psychol* 2008;61:29–48.
- [11] Kirshner B, Guyatt G. A methodological framework for assessing health indices. *J Chron Dis* 1985;38:27–36.
- [12] Terwee CB, Bot SDM, de Boer MR, van der Windt DAWM, Knol DL, Dekker J, et al. Quality criteria were proposed for measurement properties of health status questionnaires. *J Clin Epidemiol* 2007;60:34–42.
- [13] Bot SDM, Terwee CB, Van der Windt DAWM, Bouter LM, Dekker J, De Vet HCW. Clinimetric evaluation of shoulder disability questionnaires: a systematic review. *Ann Rheum Dis* 2004;63:335–41.
- [14] Kottner J, Dassen T. Interpreting interrater reliability coefficients of the Braden scale: A discussion paper. *Int J Nurs Stud* 2008;45:1239–46.
- [15] Kottner J, Dassen T, Tannen A. Inter- and intrarater reliability of the Waterlow pressure sore risk scale: a systematic review. *Int J Nurs Stud* 2009;46:369–79.
- [16] Kottner J, Raeder K, Halfens R, Dassen T. A systematic review of interrater reliability of pressure ulcer classification systems. *J Clin Nurs* 2009;18:315–36.
- [17] Nwosu CR, Khan KS, Chien PF, Honest MR. Is real-time ultrasonic bladder volume estimation reliable and valid? A systematic overview. *Scand J Urol Nephrol* 1998;32:325–30.
- [18] Ratanawongsa N, Thomas PA, Marinopoulos SS, Dorman T, Wilson LM, Ashar BH, et al. The reported validity and reliability of methods for evaluating continuing medical education: a systematic review. *Acad Med* 2008;83:274–83.
- [19] Stochkendahl MJ, Christensen HW, Hartvigsen J, Vach W, Haas M, Hestbaek L, et al. Manual examination of the spine: a systematic critical literature review of reproducibility. *J Manipulative Physiol Ther* 2006;29:475–85.
- [20] Sun Y, Stürmer T, Günther KP, Brenner H. Reliability and validity of clinical outcome measurements of osteoarthritis of the hip and knee— a review of the literature. *Clin Rheumatol* 1997;16:185–98.
- [21] Swingler GH. Observer variation in chest radiography of acute lower respiratory infections in children: a systematic review. *BMC Med Imaging* 2001;1:1.
- [22] Audigé L, Bhandari M, Kellam J. How reliable are reliability studies of fracture classifications? *Acta Orthop Scand* 2004;75:184–94.
- [23] Brorson S, Hróbjartsson A. Training improves agreement among doctors using the Neer system for proximal humeral fractures in a systematic review. *J Clin Epidemiol* 2008;61:7–16.
- [24] Hestbaek L, Leboeuf-Yde C. Are chiropractic tests for the lumbo-pelvic spine reliable and valid? A systematic review. *J Manipulative Physiol Ther* 2000;23:258–75.
- [25] Innes E, Straker L. Reliability of work-related assessments. *Work* 1999;13:107–24.
- [26] American Educational Research Association, American Psychological Association, National Council on Measurement in Education.

- Standards for educational and psychological testing. Washington, DC: American Educational Research Association; 1999.
- [27] Glaser EM. Using behavioral science strategies for defining the state-of-the-art. *J Appl Behav Sci* 1980;16:79–92.
- [28] Fink A, Kosecoff J, Chassin M, Brook RH. Consensus methods: characteristics and guidelines for use. *Am J Public Health* 1984;74:979–83.
- [29] Streiner DL. Clinimetrics versus psychometrics: an unnecessary distinction. *J Clin Epidemiol* 2003;56:1142–5.
- [30] Buntinx F, Beckers H, De Keyser G, Flour M, Nissen G, Raskin T, et al. Inter-observer variation in the assessment of skin ulceration. *J Wound Care* 1996;5:166–70.
- [31] Thomsen NO, Olsen LH, Nielsen ST. Kappa statistics in the assessment of observer variation: the significance of multiple observers classifying ankle fractures. *J Orthop Sci* 2002;7:163–6.
- [32] Colle F, Rannon F, Revel M, Fermandian J, Poiradeau S. Impact of quality scales on levels of evidence inferred from a systematic review of exercise therapy and low back pain. *Arch Phys Med Rehabil* 2002;83:1745–52.
- [33] Vikström A, Skånér Y, Strender LE, Nilsson GH. Mapping the categories of the Swedish primary health care version of ICD-10 to SNOMED CT concepts: rule development and intercoder reliability in a mapping trial. *BMC Med Inform Decis Making* 2007;2:9.
- [34] Hwang JC, Yu AC, Casper DS, Starren J, Cimino JJ, Chiang MF. Representation of ophthalmology concepts by electronic systems: intercoder agreement among physicians using controlled terminologies. *Ophthalmology* 2006;113:511–9.
- [35] Nanda R, Gupta S, Kanapathipillai P, Liow RYL, Rangan A. An assessment of the inter examiner reliability of clinical tests for subacromial impingement and rotator cuff integrity. *Eur J Orthop Surg Traumatol* 2008;18:495–500.
- [36] Hall TM, Robinson KW, Fujinawa O, Akasaka K, Pyne EA. Intertester reliability and diagnostic validity of the cervical flexion-rotation test. *J Manipulative Physiol Ther* 2008;31:293–300.
- [37] Peat J, Barton B. Medical statistics: a guide to data analysis and critical appraisal. Oxford, UK: Blackwell; 2005.
- [38] Cicchetti D, Bronen R, Spencer S, Haut S, Berg A, Oliver P, et al. Rating scales, scales of measurement, issues of reliability. *J Nerv Ment Dis* 2006;194:557–64.
- [39] Kraemer HC. Ramifications of a population model for κ as a coefficient of reliability. *Psychometrika* 1979;44:461–72.
- [40] Suen HK. Agreement, reliability, accuracy, and validity: toward a clarification. *Behav Assess* 1988;10:343–66.
- [41] Slongo T, Audigé L, Schlickewei W, Clavert J-M, Hunter J. International Association for Pediatric Traumatology Development and validation of the AO Pediatric Comprehensive Classification of Long Bone Fractures by the pediatric expert group of the AO Foundation in collaboration with AO Clinical Investigation and Documentation and the International Association for Pediatric Traumatology. *J Pediatr Orthop* 2006;26:43–9.
- [42] Barone JE, Madlinger RV. Should an Allen test be performed before radial artery cannulation? *J Trauma* 2006;61:468–70.
- [43] Bhat R, Rockwood K. Inter-rater reliability of delirium rating scales. *Neuroepidemiology* 2005;25:48–52.
- [44] Kraemer HC, Periyakoil VS, Noda A. Kappa coefficients in medical research. *Stat Med* 2002;21:2109–29.
- [45] Maclure M, Willett WC. Misinterpretation and misuse of the kappa statistic. *Am J Epidemiol* 1987;126:161–9.
- [46] Müller R, Büttner P. A critical discussion of intraclass correlation coefficients. *Stat Med* 1994;13:2465–76.
- [47] Rousson V, Gasser T, Seifert B. Assessing intrarater, interrater and test-retest reliability of continuous measurements. *Stat Med* 2002;21:3431–46.
- [48] Gjørup T. The kappa coefficient and the prevalence of a diagnosis. *Methods Inf Med* 1988;27:184–6.
- [49] Kobak KA, Engelhardt N, Williams J, Lipsitz JD. Rater training in multicenter clinical trials: issues and recommendations. *J Clin Psychopharmacol* 2004;24:113–7.
- [50] Shrout PE. Measurement reliability and agreement in psychiatry. *Stat Methods Med Res* 1998;7:301–17.
- [51] Kobak KA, Lipsitz JD, Williams JBW, Engelhardt N, Bellew KM. A new approach to rater training and certification in a multicenter clinical trial. *J Clin Psychopharmacol* 2005;25:407–11.
- [52] Altaye M, Donner A, Eliasziw M. A general goodness-of-fit approach for inference procedures concerning the kappa statistic. *Stat Med* 2001;20:2479–88.
- [53] Cantor AB. Sample-size calculations for Cohen's kappa. *Psychol Methods* 1996;1:150–3.
- [54] Cicchetti D. Sample size requirements for increasing the precision of reliability estimates: problems and proposed solutions. *J Clin Exp Neuropsychol* 1999;21:567–70.
- [55] Donner A, Eliasziw M. Sample size requirements for reliability studies. *Stat Med* 1987;6:441–8.
- [56] Donner A, Eliasziw M. A goodness-of-fit approach to inference procedures for the kappa statistic: confidence interval construction, significance testing and sample size estimation. *Stat Med* 1992;11:1511–9.
- [57] Giraudeau B, Mary JY. Planning a reproducibility study: how many subjects and how many replicates per subject for an expected width of the 95 percent confidence interval of the intraclass correlation coefficient. *Stat Med* 2001;15:3205–14.
- [58] Shoukri MM, Asyali MH, Donner A. Sample size requirements for the design of reliability study: review and new results. *Stat Methods Med Res* 2004;13:251–71.
- [59] Walter SD, Eliasziw M, Donner A. Sample size and optimal designs for reliability studies. *Stat Med* 1998;17:101–10.
- [60] Cicchetti D. The precision of reliability and validity estimates re-visited: distinguishing between clinical and statistical significance of sample size requirements. *J Clin Exp Neuropsychol* 2001;23:695–700.
- [61] Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. *Psychol Bull* 1979;86:420–8.
- [62] D'Olhaberriague L, Litvan I, Mitsias P, Mansbach HH. A reappraisal of reliability and validity studies in stroke. *Stroke* 1996;27:2331–6.
- [63] Zegers M, de Bruijne MC, Wagner C, Groenewegen PP, van der Wal G, de Vet HC. The inter-rater agreement of retrospective assessments of adverse events does not improve with two reviewers per patient record. *J Clin Epidemiol* 2010;63:94–102.
- [64] Bours GJJW, Halfens RJG, Lubbers M, Haalboom JRE. The development of a National Registration Form to measure the prevalence of pressure ulcers in the Netherlands. *Ostomy Wound Manage* 1999;45:28–40.
- [65] Gould D, Goldstone L, Kelly D, Gammon J. Examining the validity of pressure ulcer risk assessment scales: a replication study. *Int J Nurs Stud* 2004;41:331–9.
- [66] Hart S, Bergquist S, Gajewski B, Dunton N. Reliability testing of the national database of nursing quality indicators pressure ulcer indicator. *J Nurs Care Qual* 2006;21:256–65.
- [67] Scinto JD, Galusha DH, Krumholz HM, Meehan TP. The case for comprehensive quality indicator reliability assessment. *J Clin Epidemiol* 2001;54:1103–11.
- [68] Audigé L, Bhandari M, Hanson B, Kellam J. A concept for the validation of fracture classifications. *J Orthop Trauma* 2005;19:401–6.
- [69] House AE, House BJ, Campbell MB. Measures of interobserver agreement: calculation formulas and distribution effects. *J Behav Assess* 1981;3:37–57.
- [70] McAlister FA, Straus SE, Sackett DL. Why we need large, simple studies of clinical examination: the problem and a proposed solution. *Lancet* 1999;354:1721–4.

- [71] Wickström G, Bendix T. The “Hawthorne effect”: what did the original Hawthorne studies actually show? *Scand J Work Environ Health* 2000;26:363–7.
- [72] European Pressure Ulcer Advisory Panel. EPUAP statement on prevalence and incidence monitoring of pressure ulcer occurrence. 2005. Available at http://www.epuap.org/review6_3/page5.html. Accessed March 8, 2009.
- [73] Perkins DO, Wyatt RJ, Bartko JJ. Penny-wise and pound-foolish: the impact of measurement error on sample size requirements in clinical trials. *Biol Psychiatry* 2000;47:762–6.
- [74] Richardson FM. Peer review of medical care. *Med Care* 1972;10:29–39.
- [75] Kobak KA, Williams JBW, Engelhardt N. A comparison of face-to-face and remote assessment of inter-rater reliability on the Hamilton Depression Rating Scale via videoconferencing. *Psychiatry Res* 2008;28:99–103.
- [76] Topf M. Interrater reliability decline under covert assessment. *Nurs Res* 1988;37:47–9.
- [77] Defloor T, Schoonhoven L. Inter-rater reliability of the EPUAP pressure ulcer classification system using photographs. *J Clin Nurs* 2004;13:952–9.
- [78] Stevens SS. On the theory of scales of measurement. *Science* 1946;103:677–80.
- [79] Vach W. The dependence of Cohen’s kappa on the prevalence does not matter. *J Clin Epidemiol* 2005;58:655–61.
- [80] Darroch JN, McCloud PI. Category distinguishability and observer agreement. *Aust N Z J Stat* 1986;28:371–88.
- [81] Whitfield JW. Intra-class rank correlation. *Biometrika* 1949;36:463–7.
- [82] Rothery P. A nonparametric measure of intraclass correlation. *Biometrika* 1979;66:629–39.
- [83] Roberts C, McNamee R. Assessing the reliability of ordered categorical scales using kappa-type statistics. *Stat Methods Med Res* 2005;14:493–514.
- [84] McGraw KO, Wong SP. Forming inferences about some intraclass correlation coefficients. *Psychol Meth* 1996;1:30–46.
- [85] Donner A, Eliasziw M. Statistical implications of the choice between a dichotomous or continuous trait in studies of interobserver agreement. *Biometrics* 1994;50:550–5.
- [86] Fleiss JL, Levin B, Paik MC. Statistical methods for rates and proportions. 3rd edition. Hoboken, NJ: Wiley; 2003.
- [87] Uebersax J. Raw agreement indices, 2002. Available at <http://ourworld.compuserve.com/homepages/juebersax/raw.htm>. Accessed April 1, 2008.
- [88] Bland JM, Altman DG. Measuring agreement in method comparison studies. *Stat Methods Med Res* 1999;8:135–60.
- [89] Stratford PW, Goldsmith CH. Use of the standard error as a reliability index of interest: an applied example using elbow flexor strength data. *Phys Ther* 1997;77:745–50.
- [90] Barlow W, Lai MY, Azen SP. A comparison of methods for calculating a stratified kappa. *Stat Med* 1991;10:1465–72.
- [91] Donner A, Klar N. The statistical analysis of kappa statistics in multiple samples. *J Clin Epidemiol* 1996;49:1053–8.
- [92] Gajewski BJ, Hart S, Bergquist S, Duntun N. Inter-rater reliability of pressure ulcer staging: probit Bayesian Hierarchical Model that allows for uncertain rater response. *Stat Med* 2007;26:4602–18.
- [93] Charter RA. Combining reliability coefficients: possible application to meta-analysis and reliability generalisation. *Psychol Rep* 2003;93:643–7.
- [94] Vacha-Haase T. Reliability generalisation: exploring variance in measurement error affecting score reliability across studies. *Educ Psychol Meas* 1998;58:6–20.
- [95] Bååth C, Hall-Lord ML, Idvall E, Wiberg-Hedman K, Larsson BW. Interrater reliability using Modified Norton Scale, Pressure Ulcer Card, Short Form-Mini Nutritional Assessment by registered and enrolled nurses in clinical practice. *J Clin Nurs* 2008;17:618–26.
- [96] Elkum N, Shoukri MM. Signal-to-noise ratio (SNR) as a measure of reproducibility: design, estimation, and application. *Health Serv Outcomes Res Methodol* 2008;8:119–33.
- [97] Bonnet DG. Sample size requirements for estimating intraclass correlations with desired precision. *Stat Med* 2002;21:1331–5.
- [98] Saito Y, Sozu T, Hamada C, Yoshimura I. Effective number of subjects and number of raters for inter-rater reliability studies. *Stat Med* 2006;25:1547–60.
- [99] Bates-Jensen BM, McCreath HE, Pongquan V, Apeles NCR. Subepidermal moisture differentiates erythema and stage I pressure ulcers in nursing home residents. *Wound Repair Regen* 2008;16:189–97.
- [100] Sainsbury A, Seebass G, Bansal A, Young JB. Reliability of the Barthel Index when used with older people. *Age Ageing* 2005;34:228–32.
- [101] Oot-Giromini B. Pressure ulcer prevalence, incidence and associated risk factors in the community. *Decubitus* 1993;6:24–32.
- [102] Lewicki LJ, Mion LC, Secic M. Sensitivity and specificity of the Braden scale in the cardiac surgical population. *J Wound Ostomy Continence Nurs* 2000;27:36–41.
- [103] Gardner MJ, Altman DG. Confidence intervals rather than P values: estimation rather than hypothesis testing. *Br Med J* 1986;292:746–50.
- [104] Roberts C, McNamee R. A matrix of kappa-type coefficients to assess the reliability of nominal scales. *Stat Med* 1998;17:471–88.
- [105] Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics* 1977;33:159–74.
- [106] Lee J, Koh D, Ong CN. Statistical evaluation of agreement between two methods for measuring a quantitative variable. *Comput Biol Med* 1989;19:61–70.
- [107] Mok JM, Berven SH, Diab M, Hackbarth M, Hu SS, Deviren V. Comparison of observer variation in conventional and three digital radiographic methods used in the evaluation of patients with adolescent idiopathic scoliosis. *Spine* 2008;33:681–6.
- [108] Kottner J, Dassen T. An interrater reliability study of the Braden scale in two nursing homes. *Int J Nurs Stud* 2008;45:1501–11.
- [109] Nunnally JC, Bernstein IH. Psychometric theory. New York, NY: McGraw-Hill; 1994.
- [110] Beckman TJ, Ghosh AK, Cook DA, Erwin PJ, Mandrekar JN. How reliable are assessments of clinical teaching? *J Gen Intern Med* 2004;19:971–7.
- [111] Goutteborge V, Wind H, Kuijter PP, Frings-Dresen MHW. Reliability and validity of functional capacity evaluation methods: a systematic review with reference to Blankenship system, Ergos work simulator, Ergo-Kit and Isernhagen work system. *Int Arch Occup Environ Health* 2004;77:527–37.
- [112] Steenbeek D, Ketelaar M, Galama K, Gortner JW. Goal attainment in paediatric rehabilitation: a critical review of the literature. *Dev Med Child Neurol* 2007;49:550–6.
- [113] Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig LM, et al. The STARD statement for reporting studies of diagnostic accuracy: explanation and elaboration. *Ann Intern Med* 2003;138:W1–W12.
- [114] Kraemer HC. Measurement of reliability for categorical data in medical research. *Stat Methods Med Res* 1992;1:183–99.
- [115] De Villiers MR, de Villiers PJT, Kent AP. The Delphi technique in health sciences education research. *Med Teach* 2005;27:639–43.
- [116] Hutchings A, Raine R, Sanderson C, Black N. A comparison of formal consensus methods used for developing clinical guidelines. *J Health Serv Res Policy* 2006;11:218–24.
- [117] Kadam UT, Jordan K, Croft PR. A comparison of two consensus methods for classifying morbidities in a single professional group showed the same outcomes. *J Clin Epidemiol* 2006;59:1169–73.
- [118] Vella K, Goldfrad C, Rowan K, Bion J, Black N. Use of consensus development to establish national priorities in critical care. *BMJ* 2000;320:976–80.