International Journal of Nursing Studies 48 (2011) 661–671

Contents lists available at ScienceDirect

International Journal of Nursing Studies

journal homepage: www.elsevier.com/ijns

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

Jan Kottner [a,*], Laurent Audige [b], Stig Brorson [c], Allan Donner [d], Byron J. Gajewski [e], Asbjørn Hróbjartsson [f], Chris Roberts [g], Mohamed Shoukri [h], David L. Streiner [i]

[a] *Fanningerstr. 61, 10365 Berlin, Germany*
[b] *AO Clinical Investigation and Documentation, Dübendorf, Switzerland*
[c] *Department of Orthopaedic Surgery, Herlev University Hospital, Herlev, Denmark*
[d] *Department of Epidemiology and Biostatistics, Schulich School of Medicine and Dentistry, The University of Western Ontario, London, Ontario, Canada*
[e] *Department of Biostatistics, University of Kansas Schools of Medicine & Nursing, Kansas City, KS, USA*
[f] *The Nordic Cochrane Centre, Rigshospitalet, Copenhagen, Denmark*
[g] *School of Community Based Medicine, The University of Manchester, Manchester, UK*
[h] *Department of Biostatistics, Epidemiology and Scientific Computing, King Faisal Specialist Hospital and Research Center, The University of Western Ontario, London, Ontario, Canada*
[i] *Department of Psychiatry, University of Toronto, Toronto, Ontario, Canada*

ARTICLE INFO

Keywords:
Agreement
Guideline
Interrater
Intrarater
Reliability
Reproducibility

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.

What is already known about this topic?

- Reporting of interrater/intrarater reliability and agreement is often incomplete and inadequate.
- Widely accepted criteria, standards, or guidelines for reliability and agreement reporting are lacking.

What this paper adds

- Fifteen issues that should be addressed when reporting reliability and agreement studies are proposed.
- The proposed guidelines help to improve reporting.

1. Background

Reliability and agreement are important issues in classification, scale and instrument development, quality

☆ This article was originally published in the *Journal of Clinical Epidemiology* 2011: 64 (96–106). The article is republished with permission from the *Journal of Clinical Epidemiology*.

* Corresponding author. Tel.: +49 30 48624860;
fax: +49 30 450 529 900.
E-mail address: jan.kottner@email.de (J. Kottner).

0020-7489/$ – see front matter © 2011 Published by Elsevier Ltd.
doi:10.1016/j.ijnurstu.2011.01.016

assurance, and in the conduct of clinical studies (Dunn, 2004; Mulsant et al., 2002; Szklo and Nieto, 2007). 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 (Dunn, 2004; Szklo and Nieto, 2007; Polit and Beck, 2008; Shoukri, 2004; Streiner and Norman, 2008).

The terms 'reliability' and 'agreement' are often used interchangeably. However, the two concepts are conceptually distinct (Bland and Altman, 1990; De Vet et al., 2006; Feinstein and Cicchetti, 1990; Gwet, 2008; Kirshner and Guyatt, 1985; Terwee et al., 2007). 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 (Dunn, 2004; Streiner and Norman, 2008). Therefore, reliability can be defined as the ability of a measurement to differentiate among 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, classification, instrument or procedure, assesses the same subjects or object 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 (Mulsant et al., 2002; Terwee et al., 2007; Bot et al., 2004; Kottner and Dassen, 2008a; Kottner et al., 2009a,b; Nwosu et al., 1998; Ratanawongsa et al., 2008; Stochkendahl et al., 2006; Sun et al., 1997; Swingler, 2001). Because of inadequate reporting, interpretation and synthesis of the results is often difficult. Moreover, widely accepted formal criteria, standards, or guidelines for reliability and agreement reporting in the health care and medical fields are lacking (Mulsant et al., 2002; Audigé et al., 2004; Brorson and Hróbjartsson, 2008). Authors of reviews have often established their own criteria for data extraction and quality assessment (Terwee et al., 2007; Kottner et al., 2009a; Nwosu et al., 1998; Stochkendahl et al., 2006; Swingler, 2001; Audigé et al., 2004; Brorson and Hróbjartsson, 2008; Hestbaek and Leboeuf-Yde, 2000; Innes and Straker, 1999). The American Educational Research Association (AERA), American Psychological Association (APA) and the National Council on Measurement in Education (NCME) have attempted to improve reporting of reliability and agreement, but they focus on psychological testing (1999).

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 which should be addressed when reliability and agreement studies are reported.

The development process contained elements of Glaser's state of the art method (Glaser, 1980) and of the nominal group technique (Fink et al., 1984). 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 which are related to reliability and agreement studies. Therefore, definitions of key concepts were discussed among the team members, and definitions were established (Appendix A). 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.

3. Guidelines

The Guidelines for Reporting Reliability and Agreement Studies (GRRAS) are shown in Table 1. They contain issues

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)
Methods	6.	Explain how the sample size was chosen. State the determined number of raters, subjects/objects, and replicate observations
	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 that were included and the number of replicate observations that 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)

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 below in order 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 that 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 characterise a wide range of situations, in which persons make judgements 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 (Streiner, 2003): interobserver variation (Buntinx et al., 1996), observer variation (Thomsen et al., 2002), interreader reliability (Colle et al., 2002), intercoder reliability (Vikström et al., 2007), intercoder agreement (Hwang et al., 2006), inter examiner reliability (Nanda et al., 2008), intertester reliability (Hall et al., 2008), repeatability (Peat and Barton, 2005), 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', 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 like 'reliability' or 'agreement' were not used (Kottner et al., 2009b). 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, classifications or scales that are used (Dunn, 2004; Shoukri, 2004; Cicchetti et al., 2006; Kraemer, 1979; Suen, 1988). 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 (Vikström et al., 2007; Hwang et al., 2006; Slongo et al., 2006). Additionally, there may be several definitions for the same measured construct (Barone and Madlinger, 2006; Bhat and Rockwood, 2005). 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 (Kraemer et al., 2002; Maclure and Willett, 1987). 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, 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) (Kottner et al., 2009b; Cicchetti et al., 2006).

In the case of continuous measurements, the value of interrater reliability/agreement coefficients depends on their range (Müller and Büttner, 1994; Rousson et al., 2002). 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 cut-off 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 (Dunn, 2004; Szklo and Nieto, 2007; Kraemer, 1979; Müller and Büttner, 1994; Gjørup, 1988).

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 characterised by rater qualifications, clinical background, knowledge, degree of expertise, and training, as these characteristics may have a major impact on reliability and agreement estimates (Szklo and Nieto, 2007; Nanda et al., 2008; Kraemer, 1979; Suen, 1988; Kobak et al., 2004; Shrout, 1998).

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 (Szklo and Nieto, 2007). 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 (Cicchetti et al., 2006; Kobak et al., 2005) and such information is needed in later reliability generalisation studies (Streiner and Norman, 2008).

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 generalisation studies should be the preferred sources. It should be explained why this new study is 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 (Altaye et al., 2001; Cantor, 1996; Cicchetti, 1999; Donner and Eliasziw, 1987, 1992; Giraudeau and Mary, 2001; Shoukri et al., 2004; Walter et al., 1998). 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 to organisational, financial, or ethical constraints (Cicchetti, 2001). 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 in order to make the study transparent and credible.

3.3.2. Item 7: describe the sampling method

Rationale: Enrolment of subjects in interrater reliability and agreement studies is often not clearly stated (Nwosu et al., 1998). 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 (Dunn, 2004; Shoukri, 2004; Shrout and Fleiss, 1979) and guides the reader in generalising the reported results (D'Olhaberriague et al., 1996). 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 (Zegers et al., 2010). 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 (Dunn, 2004; De Vet et al., 2006; Hestbaek and Leboeuf-Yde, 2000; Kobak et al., 2004; Bours et al., 1999; Gould et al., 2004; Hart et al., 2006; Scinto et al., 2001). Standardisation of the measurement/rating process helps to prevent bias (Nanda et al., 2008), but when instruments or classifications are to be utilised 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 (Stochkendahl et al., 2006; Audigé et al., 2005; House et al., 1981; McAlister et al., 1999).

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 (Nwosu et al., 1998; Audigé et al., 2004; Brorson and Hróbjartsson, 2008). However, describing the availability of clinical information does not mean that these data must be described or analysed in detail. It should also be stated whether raters were blinded in the sense that they were not aware that their judgement will be compared to those of other raters, removing the possibility of a Hawthorne effect (i.e., ensuring that the rater's behaviour is not altered due to an awareness being observed) (Polit and Beck, 2008; Wickström and Bendix, 2000). A sufficient description of the measurement process, including information regarding blinding is missing in many research reports (Nwosu et al., 1998; Hestbaek and Leboeuf-Yde, 2000).

In addition, it should be stated if the final measure or rating to be used result from single or repeated measures. In order 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 (European Pressure Ulcer Advisory Panel, 2005; Perkins et al., 2000; Richardson, 1972). 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 reliability (Streiner and Norman, 2008; Shrout and Fleiss, 1979). 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 (Shrout and Fleiss, 1979). 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 (Kobak et al., 2008; Topf, 1988).

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 are simultaneously conducting 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 assessment (Swingler, 2001; Defloor and Schoonhoven, 2004). 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 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 (Kottner et al., 2009a,b; Nwosu et al., 1998; Audigé et al., 2004).

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 (Dunn, 2004; Shoukri, 2004; Müller and Büttner, 1994). Therefore, it is not possible to be too prescriptive regarding the 'best' statistical method, with the choice depending on the purpose as well as design of the study.

Table 2
Statistical methods for analysing 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 (ICC)	Proportions of agreement (ranges) Proportions of specific agreement (ranges) Standard errors of measurement (SEM) Coefficients of variation (c.v.) Bland–Altman-plots and limits of agreement

Table 2 lists frequently applied statistical approaches as arranged by the level of measurement (Stevens, 1946) and by the use of reliability vs. agreement statistics. Kappa-like statistics provide useful information about reliability for categorical data (Dunn, 2004; Kraemer et al., 2002). 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 (Donner and Eliasziw, 1992). Kappa coefficients have been frequently criticized for their dependence on the rater prevalence, but, as with other measures of reliability or diagnostic accuracy this behaviour 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 (Kraemer et al., 2002; Vach, 2005). In addition, it might reflect the inability of raters to distinguish between adjacent categories (Darroch and McCloud, 1986).

Ordinal measurements are common in research and practice. Reliability calculations for such data have been proposed by Whitfield (1949), Rothery (1979), Müller and Büttner (1994), and Roberts and McNamee (2005).

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 (Shrout and Fleiss, 1979; McGraw and Wong, 1996). Moreover, although the ICC is reported in many research reports, it is often not clear which ICC was used (Kottner and Dassen, 2008a; Bhat and Rockwood, 2005). 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 (Colle et al., 2002; Shoukri et al., 2004; Donner and Eliasziw, 1994).

Proposed measures of agreement include proportions of exact agreement (De Vet et al., 2006; Fleiss et al., 2003; Uebersax, 2002), proportions of specific agreement (Fleiss et al., 2003; Uebersax, 2002), repeatability coefficients, and the graphical method proposed by Bland and Altman (Bland and Altman, 1999). For continuous measurements standard errors of measurement (SEM) (De Vet et al., 2006; Stratford and Goldsmith, 1997) and proportions of agreement within specified limits provide useful information as well (De Vet et al., 2006).

When reliability or agreement data is 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 summarising reliability coefficients from different groups or samples have been made (Shoukri, 2004; Barlow et al., 1991; Donner and Klar, 1996; Gajewski et al., 2007; Charter, 2003; Vacha-Haase, 1998). Although not frequently done (e.g., Bååth et al., 2008), the heterogeneity between multiple centres should be reported, because empirical evidence suggests that it is almost always present (Kottner et al., 2009b; Gajewski et al., 2007; Bååth et al., 2008).

There are other approaches which might also be taken (e.g., coefficients of variation, item response theory, or the 'signal to noise ratio' (Elkum and Shoukri, 2008)). 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 (Sun et al., 1997; Thomsen et al., 2002; Shoukri et al., 2004; Walter et al., 1998; Bonnet, 2002; Saito et al., 2006). 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 generalisability 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 (Dunn, 2004; Shrout and Fleiss, 1979). However, in the case of self-administered questionnaires, only the number of respondents needs to be provided.

Recently, Bates-Jensen et al. (2008) 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 (Stochkendahl et al., 2006; Bhat and Rockwood, 2005; Audigé et al., 2005) and whether results may be extrapolated to other populations (Innes and Straker, 1999). 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 characterising the sample (Sainsbury et al., 2005).

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 (Kottner and Dassen, 2008a; Kottner et al., 2009b). Statements like 'The percentage agreement ... was $r = 0.83$ to $r = 0.96$' (Oot-Giromini, 1993) or 'A data collection team ... were trained ... to achieve an interrater reliability of 0.90.' (Lewicki et al., 2000) are insufficient.

Single summary measures of reliability and agreement provide only limited information (Bland and Altman, 1990; Kottner and Dassen, 2008a; Stochkendahl et al., 2006; Thomsen et al., 2002). 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 (Donner and Eliasziw, 1992; Gardener and Altman, 1986). Where investigators wish to demonstrate a satisfactory level of reliability and agreement, particular attention should be given in the interpretation of the lower limit (Roberts and McNamee, 1998).

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 labelled as 'poor' or 'perfect', 'low' or 'high,' or whether the reliability/agreement is 'high enough' (Audigé et al., 2004; Cicchetti, 2001; Fleiss et al., 2003; Landis and Koch, 1977; Lee et al., 1989). Although these guidelines are clearly arbitrary (Cicchetti et al., 2006; Landis and Koch, 1977) 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". Nevertheless, these 'labels' do not indicate the practical or clinical relevance of results (Stochkendahl et al., 2006; Innes and Straker, 1999). 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 (Dunn, 2004; Bland and Altman, 1990; Kraemer et al., 2002; Mok et al., 2008).

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 (Terwee et al., 2007; Shoukri et al., 2004; Kottner and Dassen, 2008b). For example, ICC values for a scale measuring pressure ulcer risk should be at least 0.90 or higher when applied in clinical practice (Kottner and Dassen, 2008b). If individual and important decisions are made on the basis of reliability estimates, values should be at least 0.90 (Polit and Beck, 2008) or 0.95 (Nunnally and Bernstein, 1994).

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 which provide little help as to what should be done next (Audigé et al., 2004).

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 (Gajewski et al., 2007; Charter, 2003). For instance, Stochkendahl et al. (2006), 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 4-fold 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 (Kottner et al., 2009a,b; Ratanawongsa et al., 2008; Sainsbury et al., 2005; Beckman et al., 2004; Gouttebarge et al., 2004; Steenbeek et al., 2007). 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 (Kottner et al., 2009b; Innes and Straker, 1999; Hall et al., 2008; Bhat and Rockwood, 2005). 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) (Bossuyt et al., 2003) and with the Standards for Educational and Psychological Testing (American Educational Research Association, American Psychological Association, National Council on Measurement in Education, 1999).

The STARD (Bossuyt et al., 2003) 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 (Fleiss et al., 2003). No rater or method is considered as a reference standard (Kraemer, 1992). 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 (American Educational Research Association, American Psychological Association, National Council on Measurement in Education, 1999) 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 behaviour is evaluated and scored using a standardised 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 (De Villiers et al., 2005). 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 summarising and presenting the group response, and on how disagreements are resolved (Hutchings et al., 2006). It has also been shown that Delphi methods do not result in different outcomes when compared to the Nominal group method and that groups of up to 12 participants can achieve agreement (Kadam et al., 2006; Vella et al., 2000). In our multidisciplinary group, discussions among group members were allowed. An understanding of reasons for disagreement was possible, 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 A. 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 judgement about a person or object
Reliability	Reliability is the ability of scores of a measuring device to differentiate among subjects or objects

Appendix A (*Continued*)

Concepts	Definitions
Repeatability	Repeatability is the degree 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 subject or objects under repeated similar assessment conditions. Synonym with intrarater reliability

Conflict of interest

The authors declare that there is no conflicht of interest.

Funding

None.

Ethical approval

The issue Ethical Approval is not applicable here.

References

Altaye, M., Donner, A., Eliasziw, M., 2001. A general goodness-of-fit approach for inference procedures concerning the kappa statistic. Statistics in Medicine 20 (16), 2479–2488.

American Educational Research Association, American Psychological Association, National Council on Measurement in Education, 1999. Standards for Educational and Psychological Testing. American Educational Research Association, Washington.

Audigé, L., Bhandari, M., Hanson, B., Kellam, J., 2005. A concept for the validation of fracture classifications. Journal of Orthopaedic Trauma 19 (6), 404–409.

Audigé, L., Bhandari, M., Kellam, J., 2004. How reliable are reliability studies of fracture classifications? Acta Orthopaedica Scandinavica 75 (2), 184–194.

Bååth, C., Hall-Lord, M.L., Idvall, E., Wiberg-Hedman, K., Larsson, B.W., 2008. Interrater reliability using Modified Norton Scale, Pressure Ulcer Card, Short Form-Mini Nutritional Assessment by registered and enrolled nurses in clinical practice. Journal of Clinical Nursing 17 (5), 618–626.

Barlow, W., Lai, M.Y., Azen, S.P., 1991. A comparison of methods for calculating a stratified kappa. Statistics in Medicine 10 (9), 1465–1472.

Barone, J.E., Madlinger, R.V., 2006. Should an Allen test be performed before radial artery cannulation? The Journal of Trauma 61 (12), 468–470.

Bates-Jensen, B.M., McCreath, H.E., Pongquan, V., Apeles, N.C.R., 2008. Subepidermal moisture differentiates erythema and stage I pressure ulcers in nursing home residents. Wound Repair and Regeneration 16 (2), 189–197.

Beckman, T.J., Ghosh, A.K., Cook, D.A., Erwin, P.J., Mandrekar, J.N., 2004. How reliable are assessments of clinical teaching? Journal of General Internal Medicine 19 (9), 971–977.

Bhat, R., Rockwood, K., 2005. Inter-rater reliability of delirium rating scales. Neuroepidemiology 25 (1), 48–52.

Bland, J.M., Altman, D.G., 1990. A note on the use of the intraclass correlation coefficient in the evaluation of agreement between two methods of measurement. Computers in Biology and Medicine 20 (5), 337–340.

Bland, J.M., Altman, D.G., 1999. Measuring agreement in method comparison studies. Statistical Methods in Medical Research 8 (2), 135–160.

Bonnet, D.G., 2002. Sample size requirements for estimating intraclass correlations with desired precision. Statistics in Medicine 21 (9), 1331–1335.

Bossuyt, P.M., Reitsma, J.B., Bruns, D.E., Gatsonis, C.A., Glasziou, P.P., Irwig, L.M., et al., 2003. The STARD statement for reporting studies of diagnostic accuracy: explanation and elaboration. Annals of Internal Medicine 138 (1), W1–W12.

Bot, S.D., Terwee, C.B., van der Windt, D.A., Bouter, L.M., Dekker, J., De Vet, H.C., 2004. Clinimetric evaluation of shoulder disability questionnaires: a systematic review. Annals of the Rheumatic Diseases 63 (4), 335–341.

Bours, G.J., Halfens, R.J., Lubbers, M., Haalboom, J.R., 1999. The development of a National Registration Form to measure the prevalence of pressure ulcers in the Netherlands. Ostomy Wound Management 45 (11), 28–40.

Brorson, S., Hróbjartsson, A., 2008. Training improves agreement among doctors using the Neer system for proximal humeral fractures in a systematic review. Journal of Clinical Epidemiology 61 (1), 7–16.

Buntinx, F., Beckers, H., De Keyser, G., Flour, M., Nissen, G., Raskin, T., et al., 1996. Inter-observer variation in the assessment of skin ulceration. Journal of Wound Care 5 (4), 166–170.

Cantor, A.B., 1996. Sample-size calculations for Cohen's kappa. Psychological Methods 1 (2), 150–153.

Charter, R.A., 2003. Combining reliability coefficients: possible application to meta-analysis and reliability generalization. Psychological Reports 93, 643–647.

Cicchetti, D., 1999. Sample size requirements for increasing the precision of reliability estimates: problems and proposed solutions. Journal of Clinical and Experimental Neuropsychology 21 (4), 567–570.

Cicchetti, D., 2001. The precision of reliability and validity estimates revisited: distinguishing between clinical and statistical significance of sample size requirements. Journal of Clinical and Experimental Neuropsychology 23 (5), 695–700.

Cicchetti, D., Bronen, R., Spencer, S., Haut, S., Berg, A., Oliver, P., et al., 2006. Rating scales, scales of measurement, issues of reliability. The Journal of Nervous and Mental Disease 194 (8), 557–564.

Colle, F., Rannon, F., Revel, M., Fermanian, J., Poiraudeau, S., 2002. Impact of quality scales on levels of evidence inferred from a systematic review of exercise therapy and low back pain. Archives of Physical Medicine and Rehabilitation 83 (12), 1745–1752.

Darroch, J.N., McCloud, P.I., 1986. Category distinguishability and observer agreement. Australian & New Zealand Journal of Statistics 28 (3), 371–388.

De Vet, H.C.W., Terwee, C.B., Knol, D.L., Bouter, L.M., 2006. When to use agreement versus reliability measures. Journal of Clinical Epidemiology 59 (10), 1033–1039.

De Villiers, M.R., de Villiers, P.J.T., Kent, A.P., 2005. The Delphi technique in health sciences education research. Medical Teacher 27 (7), 639–643.

Defloor, T., Schoonhoven, L., 2004. Inter-rater reliability of the EPUAP pressure ulcer classification system using photographs. Journal of Clinical Nursing 13 (8), 952–959.

D'Olhaberriague, L., Litvan, I., Mitsias, P., Mansbach, H.H., 1996. A reappraisal of reliability and validity studies in stroke. Stroke 27 (12), 2331–2336.

Donner, A., Eliasziw, M., 1987. Sample size requirements for reliability studies. Statistics in Medicine 6 (4), 441–448.

Donner, A., Eliasziw, M., 1992. A goodness-of-fit approach to inference procedures for the kappa statistic: confidence interval construction, significance testing and sample size estimation. Statistics in Medicine 11 (11), 1511–1519.

Donner, A., Eliasziw, M., 1994. Statistical implications of the choice between a dichotomous or continuous trait in studies of interobserver agreement. Biometrics 50 (2), 550–555.

Donner, A., Klar, N., 1996. The statistical analysis of kappa statistics in multiple samples. Journal of Clinical Epidemiology 49 (9), 1053–1058.

Dunn, G., 2004. Statistical Evaluation of Measurement Errors: Design and Analysis of Reliability Studies, 2nd ed. Arnold, London.

Elkum, N., Shoukri, M.M., 2008. Signal-to-noise ratio (SNR) as a measure of reproducibility: design, estimation, and application. Health Services and Outcomes Research Methodology 8, 119–133.

European Pressure Ulcer Advisory Panel, 2005. EPUAP Statement on Prevalence and Incidence Monitoring of Pressure Ulcer Occurrence 2005. Retrieved March 8, 2009, from http://www.epuap.org/review6_3/page5.html.

Feinstein, A.R., Cicchetti, D.V., 1990. High agreement but low kappa: I. The problems of two paradoxes. Journal of Clinical Epidemiology 43 (6), 543–549.

Fink, A., Kosecoff, J., Chassin, M., Brook, R.H., 1984. Consensus methods: characteristics and guidelines for use. American Journal of Public Health 74 (9), 979–983.

Fleiss, J.L., Levin, B., Paik, M.C., 2003. Statistical Methods for Rates and Proportions, 3rd ed. Wiley, New Jersey.

Gajewski, B.J., Hart, S., Bergquist, S., Dunton, N., 2007. Inter-rater reliability of pressure ulcer staging: probit Bayesian Hierarchical Model that allows for uncertain rater response. Statistics in Medicine 26 (25), 4602–4618.

Gardener, M.J., Altman, D.G., 1986. Confidence intervals rather than P values: estimation rather than hypothesis testing. British Medical Journal 292 (6522), 746–750.

Giraudeau, B., Mary, J.Y., 2001. 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. Statistics in Medicine 20 (21), 3205–3214.

Gjørup, T., 1988. The kappa coefficient and the prevalence of a diagnosis. Methods of Information in Medicine 27 (4), 184–186.

Glaser, E.M., 1980. Using behavioral science strategies for defining the state-of-the-art. Journal of Applied Behavioral Science 16, 79–92.

Gould, D., Goldstone, L., Kelly, D., Gammon, J., 2004. Examining the validity of pressure ulcer risk assessment scales: a replication study. International Journal of Nursing Studies 41 (3), 331–339.

Gouttebarge, V., Wind, H., Kuijer, P.P., Frings-Dresen, M.H.W., 2004. 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. International Archives of Occupational and Environmental Health 77 (8), 527–537.

Gwet, K.L., 2008. Computing inter-rater reliability and its variance in the presence of high agreement. The British Journal of Mathematical and Statistical Psychology 61 (Pt 1), 29–48.

Hall, T.M., Robinson, K.W., Fujinawa, O., Akasaka, K., Pyne, E.A., 2008. Intertester reliability and diagnostic validity of the cervical flexion-rotation test. Journal of Manipulative and Physiological Therapy 31 (4), 293–300.

Hart, S., Bergquist, S., Gajewski, B., Dunton, N., 2006. Reliability testing of the national database of nursing quality indicators pressure ulcer indicator. Journal of Nursing Care Quality 21 (3), 256–265.

Hestbaek, L., Leboeuf-Yde, C., 2000. Are chiropractic tests for the lumbo-pelvic spine reliable and valid? A systematic review. Journal of Manipulative and Physiological Therapeutics 23 (4), 258–275.

House, A.E., House, B.J., Campbell, M.B., 1981. Measures of interobserver agreement: calculation formulas and distribution effects. Journal of Behavioral Assessment 3 (1), 37–57.

Hutchings, A., Raine, R., Sanderson, C., Black, N., 2006. A comparison of formal consensus methods used for developing clinical guidelines. Journal of Health Services Research & Policy 11 (4), 218–224.

Hwang, J.C., Yu, A.C., Casper, D.S., Starren, J., Cimino, J.J., Chiang, M.F., 2006. Representation of ophthalmology concepts by electronic systems: intercoder agreement among physicians using controlled terminologies. Ophtalmology 113 (4), 511–519.

Innes, E., Straker, L., 1999. Reliability of work-related assessments. Work 13 (2), 107–124.

Kadam, U.T., Jordan, K., Croft, P.R., 2006. A comparison of two consensus methods for classifying morbidities in a single professional group showed the same outcomes. Journal of Clinical Epidemiology 59 (11), 1169–1173.

Kirshner, B., Guyatt, G., 1985. A methodological framework for assessing health indices. Journal of Chronic Diseases 38 (1), 27–36.

Kobak, K.A., Engelhardt, N., Williams, J., Lipsitz, J.D., 2004. Rater training in multicenter clinical trials: issues and recommendations. Journal of Clinical Psychopharmacology 24 (2), 113–117.

Kobak, K.A., Lipsitz, J.D., Williams, J.B.W., Engelhardt, N., Bellew, K.M., 2005. A new approach to rater training and certification in a multicenter clinical trial. Journal of Clinical Psychopharmacology 25 (5), 407–411.

Kobak, K.A., Williams, J.B.W., Engelhardt, N., 2008. A comparison of face-to-face and remote assessment of inter-rater reliability on the Hamilton Depression Rating Scale via videoconferencing. Psychiatry Research 158 (1), 99–103.

Kottner, J., Dassen, T., 2008a. Interpreting interrater reliability coefficients of the Braden scale: a discussion paper. International Journal of Nursing Studies 45 (8), 1239–1246.

Kottner, J., Dassen, T., 2008b. An interrater reliability study of the Braden scale in two nursing homes. International Journal of Nursing Studies 45 (10), 1501–1511.

Kottner, J., Dassen, T., Tannen, A., 2009a. Inter- and intrarater reliability of the Waterlow pressure sore risk scale: a systematic review. International Journal of Nursing Studies 46 (3), 369–379.

Kottner, J., Raeder, K., Halfens, R., Dassen, T., 2009b. A systematic review of interrater reliability of pressure ulcer classification systems. Journal of Clinical Nursing 18 (3), 315–336.

Kraemer, H.C., 1979. Ramifications of a population model for κ as a coefficient of reliability. Psychometrika 44 (4), 461–472.

Kraemer, H.C., 1992. Measurement of reliability for categorical data in medical research. Statistical Methods in Medical Research 1 (2), 183–199.

Kraemer, H.C., Periyakoil, V.S., Noda, A., 2002. Kappa coefficients in medical research. Statistics in Medicine 21 (14), 2109–2129.

Landis, J.R., Koch, G.G., 1977. The measurement of observer agreement for categorical data. Biometrics 33 (1), 159–174.

Lee, J., Koh, D., Ong, C.N., 1989. Statistical evaluation of agreement between two methods for measuring a quantitative variable. Computers in Biology and Medicine 19 (1), 61–70.

Lewicki, L.J., Mion, L.C., Secic, M., 2000. Sensitivity and specificity of the Braden scale in the cardiac surgical population. Journal of Wound Ostomy and Continence Nursing 27 (1), 36–41.

Maclure, M., Willett, W.C., 1987. Misinterpretation and misuse of the kappa statistic. American Journal of Epidemiology 126 (2), 161–169.

McAlister, F.A., Straus, S.E., Sackett, D.L., 1999. Why we need large, simple studies of clinical examination: the problem and a proposed solution. Lancet 354 (13), 1721–1724.

McGraw, K.O., Wong, S.P., 1996. Forming inferences about some intraclass correlation coefficients. Psychological Methods 1 (1), 30–46.

Mok, J.M., Berven, S.H., Diab, M., Hackbarth, M., Hu, S.S., Deviren, V., 2008. Comparison of observer variation in conventional and three digital radiographic methods used in the evaluation of patients with adolescent idiopathic scoliosis. Spine 33 (6), 681–686.

Müller, R., Büttner, P., 1994. A critical discussion of intraclass correlation coefficients. Statistics in Medicine 13 (23–24), 2465–2476.

Mulsant, B.H., Kastango, K.B., Rosen, J., Stone, R.A., Mazumdar, S., Pollock, B.G., 2002. Interrater reliability in clinical trials of depressive disorders. The American Journal of Psychiatry 159 (9), 1598–1600.

Nanda, R., Gupta, S., Kanapathipillai, P., Liow, R.Y.L., Rangan, A., 2008. An assessment of the inter examiner reliability of clinical tests for subacromial impingement and rotator cuff integrity. European Journal of Orthopaedic Surgery and Traumatology 18, 495–500.

Nunnally, J.C., Bernstein, I.H., 1994. Psychometric Theory. McGraw-Hill, New York.

Nwosu, C.R., Khan, K.S., Chien, P.F., Honest, M.R., 1998. Is real-time ultrasonic bladder volume estimation reliable and valid? A systematic overview. Scandinavian Journal of Urology and Nephrology 32 (5), 325–330.

Oot-Giromini, B., 1993. Pressure ulcer prevalence, incidence and associated risk factors in the community. Decubitus 6 (5), 24–32.

Peat, J., Barton, B., 2005. Medical Statistics: A Guide to Data Analysis and Critical Appraisal. Blackwell, Oxford.

Perkins, D.O., Wyatt, R.J., Bartko, J.J., 2000. Penny-wise and pound-foolish: the impact of measurement error on sample size requirements in clinical trails. Biological Psychiatry 47 (8), 762–766.

Polit, D.F., Beck, C.T., 2008. Nursing Research: Generating and Assessing Evidence for Nursing Practice, 8th ed. Lippincott Williams & Wilkins, Philadelphia.

Ratanawongsa, N., Thomas, P.A., Marinopoulos, S.S., Dorman, T., Wilson, L.M., Ashar, B.H., et al., 2008. The reported validity and reliability of methods for evaluating continuing medical education: a systematic review. Academic Medicine 83 (3), 274–283.

Richardson, F.M., 1972. Peer review of medical care. Medical Care 10 (1), 29–39.

Roberts, C., McNamee, R., 1998. A matrix of kappa-type coefficients to assess the reliability of nominal scales. Statistics in Medicine 17 (4), 471–488.

Roberts, C., McNamee, R., 2005. Assessing the reliability of ordered categorical scales using kappa-type statistics. Statistical Methods in Medical Research 14 (5), 493–514.

Rothery, P., 1979. A nonparametric measure of intraclass correlation. Biometrika 66 (3), 629–639.

Rousson, V., Gasser, T., Seifert, B., 2002. Assessing intrarater, interrater and test–retest reliability of continuous measurements. Statistics in Medicine 21 (22), 3431–3446.

Sainsbury, A., Seebass, G., Bansal, A., Young, J.B., 2005. Reliability of the Barthel Index when used with older people. Age Ageing 34 (3), 228–232.

Saito, Y., Sozu, T., Hamada, C., Yoshimura, I., 2006. Effective number of subjects and number of raters for inter-rater reliability studies. Statistics in Medicine 25 (9), 1547–1560.

Scinto, J.D., Galusha, D.H., Krumholz, H.M., Meehan, T.P., 2001. The case for comprehensive quality indicator reliability assessment. Journal of Clinical Epidemiology 54 (11), 1103–1111.

Shoukri, M.M., 2004. Measures of Interobserver Agreement. Chapman & Hall/CRC, Boca Raton.

Shoukri, M.M., Asyali, M.H., Donner, A., 2004. Sample size requirements for the design of reliability study: review and new results. Statistical Methods in Medical Research 13, 251–271.

Shrout, P.E., 1998. Measurement reliability and agreement in psychiatry. Statistical Methods in Medical Research 7 (3), 301–317.

Shrout, P.E., Fleiss, J.L., 1979. Intraclass correlations: uses in assessing rater reliability. Psychological Bulletin 86 (2), 420–428.

Slongo, T., Audigé, L., Schlickewei, W., Clavert, J.-M., Hunter, J., 2006. 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. Journal of Pediatric Orthopedics 26 (1), 43–49.

Steenbeek, D., Ketelaar, M., Galama, K., Gortner, J.W., 2007. Goal attainment in paediatric rehabilitation: a critical review of the literature. Developmental Medicine & Child Neurology 49 (7), 550–556.

Stevens, S.S., 1946. On the theory of scales of measurement. Science 103 (2684), 677–680.

Stochkendahl, M.J., Christensen, H.W., Hartvigsen, J., Vach, W., Haas, M., Hestbaek, L., et al., 2006. Manual examination of the spine: a systematic critical literature review of reproducibility. Journal of Manipulative and Physiological Therapeutics 29 (6), 475–485.

Stratford, P.W., Goldsmith, C.H., 1997. Use of the standard error as a reliability index of interest: an applied example using elbow flexor strength data. Physical Therapy 77 (7), 745–750.

Streiner, D.L., 2003. Clinimetrics versus psychometrics: an unnecessary distinction. Journal of Clinical Epidemiology 56 (12), 1142–1145.

Streiner, D.L., Norman, G.R., 2008. Health Measurement Scales: A Practical Guide to Their Development and Use, 4th ed. Oxford University Press, Oxford.

Suen, H.K., 1988. Agreement, reliability, accuracy, and validity: toward a clarification. Behavioral Assessment 10 (4), 343–366.

Sun, Y., Stürmer, T., Günther, K.P., Brenner, H., 1997. Reliability and validity of clinical outcome measurements of osteoarthritis of the hip and knee – a review of the literature. Clinical Rheumatology 16 (2), 185–198.

Swingler, G.H., 2001. Observer variation in chest radiography of acute lower respiratory infections in children: a systematic review. BMC Medical Imaging 1 (1), 1.

Szklo, M., Nieto, F.J., 2007. Epidemiology Beyond the Basics, 2nd ed. Jones and Bartlett, Sudbury.

Terwee, C.B., Bot, S.D., de Boer, M.R., van der Windt, D.A., Knol, D.L., Dekker, J., et al., 2007. Quality criteria were proposed for measurement properties of health status questionnaires. Journal of Clinical Epidemiology 60 (1), 34–42.

Thomsen, N.O., Olsen, L.H., Nielsen, S.T., 2002. Kappa statistics in the assessment of observer variation: the significance of multiple observers classifying ankle fractures. Journal of Orthopaedic Science 7 (2), 163–166.

Topf, M., 1988. Interrater reliability decline under covert assessment. Nursing Research 37 (1), 47–49.

Uebersax, J., 2002. Raw Agreement Indices. Retrieved April 1, 2008, from http://ourworld.compuserve.com.homepages/jsuebersax/raw.htm.

Vach, W., 2005. The dependence of Cohen's kappa on the prevalence does not matter. Journal of Clinical Epidemiology 58 (7), 655–661.

Vacha-Haase, T., 1998. Reliability generalisation: exploring variance in measurement error affecting score reliability across studies. Educational and Psychological Measurement 58 (1), 6–20.

Vella, K., Goldfrad, C., Rowan, K., Bion, J., Black, N., 2000. Use of consensus development to establish national priorities in critical care. BMJ 320 (7240), 976–980.

Vikström, A., Skånér, Y., Strender, L.E., Nilsson, G.H., 2007. 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 Medical Informatics and Decision Making 7, 9.

Walter, S.D., Eliasziw, M., Donner, A., 1998. Sample size and optimal designs for reliability studies. Statistics in Medicine 17 (1), 101–110.

Whitfield, J.W., 1949. Intra-class rank correlation. Biometrika 36 (3–4), 463–467.

Wickström, G., Bendix, T., 2000. The "Hawthorne effect": what did the original Hawthorne studies actually show? Scandinavian Journal of Work, Environment & Health 26 (4), 363–367.

Zegers, M., de Bruijne, M.C., Wagner, C., Groenewegen, P.P., van der Wal, G., de Vet, H.C., 2010. The inter-rater agreement of retrospective assessments of adverse events does not improve with two reviewers per patient record. Journal of Clinical Epidemiology 63 (1), 94–102.