How to read and critically appraise a reliability article

Andrea Bialocerkowski, Nerida Klupp, Peter Bragge

Background: *This article is a guide to critically appraising reliability studies based upon fundamental reliability concepts. It focuses on reliability studies of physical measurement instruments used in the allied health professions.*

Content: *Eight critical appraisal questions specific to reliability studies are outlined, and their theoretical basis and importance described. The questions encompass key aspects of reliability theory; selection of clinically stable participants appropriate to the instrument and condition of interest; minimization of random error; appropriate periods of time between measurements; the interpretation of frequently used reliability statistics; and the generalizability of results. The importance of interpreting reliability studies in specific clinical contexts and settings is emphasized.*

Conclusions: *These questions will guide clinicians and researchers to make informed decisions regarding whether reliability evidence can be applied to their specific context.*

Key words: ■ critical appraisal ■ measurement studies ■ outcome measurement ■ reliability

Submitted 29 October 2009, sent back for revisions 25 November 2009; accepted for publication following double-blind peer review 15 December 2009

Andrea Bialocerkowski *is Associate Professor, School of Biomedical and Health Sciences, University of Western Sydney, Sydney;* **Nerida Klupp** *is Lecturer, School of Biomedical and Health Sciences, University of Western Sydney, Sydney; and* **Peter Bragge** *is Senior Research Fellow, National Trauma Research Institute, Monash University, Melbourne, Australia.*

Correspondence to: A Bialocerkowski Email: a.bialocerkowski@uws.edu.au

The term 'reliability' is used frequently in healthcare conversations, by clinicians, patients, and researchers. Often, it is used inappropriately, due to the lack of understanding of the theoretical concepts that underpin this measurement concept. In the article 'Measurement error and reliability testing: application to rehabilitation' (Bialocerkowski and Bragge, 2008), the fundamental reliability concepts were presented by discussing the relationship between reliability and measurement error, the sources of measurement error, methods used to minimize measurement error, and the interpretation of measurement results. Moreover, the various types of reliability studies and the relationship between reliability and validity were explored. Although these fundamental concepts are paramount in gaining a thorough understanding of reliability studies, there is a paucity of information available to guide clinicians and researchers on how to critically appraise reliability studies.

Critical appraisal is a process which is systematic in nature. It is used by clinicians and researchers to examine the methodological quality of research. Moreover, critical appraisal guides and tools are used to assist in determining whether research findings are relevant to a specific clinical or research context (Oxman et al, 1993; Straus et al, 2005). A plethora of critical appraisal tools currently exist and cover study designs such as the systematic reviews, and experimental, diagnostic, observational and qualitative studies (Katrak et al, 2004).

Although reliability studies can be considered as observational studies, observational study critical appraisal tools currently do not consider aspects that are unique to reliability studies. Similarly, when reliability is considered as a component of a validity study appraisal (Jerosch-Herold, 2005), the key reliability issues are unable to be thoroughly addressed. Therefore, in this article, a short list of questions that have been specifically designed to critically evaluate reliability studies are presented. These questions are based on the theory of reliability. It is envisaged that the responses to these questions will direct clinicians' and researchers' thinking towards whether the information presented in published reliability studies can be appropriately applied to their specific context.

The focus of this article is evaluation of reliability studies of physical measurement instruments used in the allied health professions. An example of such a study, that may be useful to read in parallel with this article, is a reliability study of a scapular dysfunction classification system by Kibler et al (2002). Throughout this article, the Kibler et al (2002) study is referred to, to present brief examples of key information for each appraisal question. Reliability studies of written questionnaires involve unique design and statistical approaches and evaluation of these is beyond the scope of this article.

EVALUATING RELIABILITY STUDIES

Question 1: Is the aim of the study clear and appropriate?

All research studies, regardless of their design or purpose, should provide an introduction to the topic under investigation (Greenhalgh, 1997). In the context of a reliability study, the introduction should therefore describe the measurement instrument or procedure under investigation, its relevance to clinical practice (either generally or in relation to a specific condition), and briefly outline what is known from previous studies about the reliability of this instrument. In doing so, the introduction should justify the importance of the study. The study aim and hypothesis, which are usually found towards the end of the introduction, should be clearly outlined so that readers have an understanding of the information that is contained in the remainder of the article.

In the recommended parallel article by Kibler et al (2002), who investigated the reliability of a scapular dysfunction classification system, the aim and hypothesis are clearly outlined at the end of the introduction section. Further, the appropriateness of this research aim has been reasonably justified by the preceding description of current evidence and clinical challenges about scapular dysfunction.

Question 2: Was the study sample appropriate?

This question addresses three issues pertaining to the sample: patient characteristics, assessor characteristics and sample size. It is suggested that at least two of these issues should be adequately addressed in order to answer ‘yes’ to this question during critical appraisal.

In studies of diagnostic accuracy, it is important to sample a spectrum of patients that are representative of the patients on which the test will be used in clinical practice or research (Whiting et al, 2003). This concept is applicable to reliability studies. For example, a reliability study of a mobility assessment scale for elderly, arthritic patients should not be conducted on a young, healthy cohort. Therefore, readers of reliability evidence need to identify the characteristics of the sample used in the reliability study and determine whether it is well matched to the population on which the measurement tool was developed for use. For example, the participants of the Kibler et al (2002) study have an average age (29.5 ± 9 years), weight (81.2 ± 15.95 kg), height (178 ± 11.9cm) and breadth of diagnostic categories, which are likely to represent some adult patient groups who present to physiotherapy for the treatment of common shoulder injuries.

Consideration should also be given to the qualifications and experience of the assessors. Assessors vary in their competence to use measurement tools and their ability to follow standardized protocols (Bartlett and Frost, 2008). These aspects directly affect measurement error (Bialocerkowski and Bragge, 2008). Furthermore, where there is more than one assessor, large differences in assessor qualifications and experience may adversely affect reliability estimates. Therefore, assessor qualifications and experience should be appropriate to those required to undertake the measurement procedure under study and this information should be documented in reliability studies. Kibler et al (2002) provide some details of the qualifications and experience of the raters and outline the training in the use of the evaluation protocol provided to them.

As in all clinical research studies, sample size should be justified. Various statistical techniques and formulas can be employed for estimating sample size in reliability studies. Such techniques follow the principles of power calculations in comparative clinical studies. For example, Walter et al (1998) provide tables and formulas to estimate the required number of subjects in a reliability study using the intraclass correlation coefficient (described below), given an acceptable range of intraclass correlation coefficient values, the number of observations per subject, and α and β. Similarly, Cicchetti (1981) demonstrated that when using the weighted Kappa statistic (also described below), the minimum number of subjects required to yield valid results is $2k^2$ (where k is the number of classification categories, ranging from 3–10). In the study by Kibler et al (2002), in which the Kappa statistic was used, the outcome (scapular dyskinesis pattern) has four categories. This implies a required sample size of $2 \times 16 = 32$ subjects; however, it is not known whether the Kappa was weighted in this study, and no statistical justification of the actual sample size of 26 was provided.

Question 3: Were a broad range of values generated for the target measurement?

To be able to adequately evaluate the reliability of a measurement tool, participants in reliability studies must be heterogenous with respect to the magnitude of the variable that is being measured. This is because between-participant variability affects the evaluation of reliability. Homogenous samples (i.e. samples with little inter-participant variation) tend to produce lower estimates of reliability compared with heterogenous samples, in the presence of the same magnitude of measurement error (Bartlett and Frost, 2008). Therefore, the lack of variability within a sample may lead to the incorrect conclusion that a measurement tool lacks reliability. Thus, in addition to being broadly representative of the condition under study, participants in a reliability study should also be selected to maximize variation in the measurement under investigation.

In addition to baseline participant characteristics, information regarding heterogeneity of participants can also be found in the results section of reliability studies. Authors should provide the range of values gained on each measurement occasion and by each assessor. These values should span the range of the measurement tool that is being evaluated.

Participants studied by Kibler et al (2002), included individuals with a range of shoulder conditions as well as those without shoulder pathology, to maximize the range of measured values. However, these values, which in this study are scapular pattern types, have not been presented in the article.

TABLE 1.
Sources of random error and methods to reduce random error for physical tests.

Sources of random error	Methods to reduce random error
Equipment	
Damage: e.g. bent plastic goniometer, hydraulic fluid leaking out of a Jamar Dynamometer	Regular checking and maintenance of equipment
Lack of calibration of equipment: e.g. KinCom, EMG	Regular calibration of equipment
Lack of adequate warm-up of equipment: e.g KinCom, Vicon	Standardized warm-up procedure
Patient	
Lack of comprehension of the task: e.g. confusion	Provide simple instructions and check comprehension
Lack of confidence or nervousness	Provide familiarization of task
Distraction: e.g. not concentrating on the measurement task	Evaluate in a quiet environment
Fatigue or practice effect	Standardized number of repetitions
Variable pain intensity	Evaluate the same patient at the same time of the day Evaluate pain intensity prior to and during measurement, and analgesic use, and use this information to interpret the measured value
Assessor (clinician or researcher)	
Lack of competence in measurement	Competence in measurement
Fatigue	Evaluate the same patient at the same time of the day
Distraction: e.g. not concentrating on the measurement task	Evaluate in a quiet environment
Lack of use of standardized measurement protocol: including starting position, handling techniques, method of stabilization, patient instructions, method of recording measurements (where appropriate)	Use a standardized measurement protocol Document where deviations occur from the standardized protocol

Sources: Rothstein, 1985; Wright and Feinstein, 1992; Bindra et al, 2003; Domholdt, 2005

Question 4: Did the researchers minimize random error in their methodology?

Random error is error that occurs due to chance. This type of error varies in magnitude and direction between patients and assessment occasions (Carmines and Zellar, 1979). Consequently, it affects reliability – 'the extent to which a measurement is consistent and free from error' (Portney and Watkins, 2000). Minimization of random error is important both in clinical practice as well as research, to produce measurements that are as close as possible to the true value (Rothstein, 1985).

Random error can originate from the measurement tool, the patient, or the assessor, for physical tests (Rothstein, 1985). *Table 1* provides examples of the specific sources of random error that may occur during measurement, and methods that can be used by clinicians and researchers to minimize random error.

When reading reliability studies, it is important to identify the methods used to minimize random error. The contents of *Table 1* could be used to guide this process. Specifically, readers should identify whether the methods used to minimize random error were applied consistently across all measurement occasions and by all assessors. In addition, readers should evaluate whether the methods used by the researchers adequately minimized random error. The contents of *Table 1* could be compared with what one would reasonably expect would occur in a reliability study given the context of the study. This is an important process in the critical appraisal of a reliability study, as the lack of minimization of random error may lead to the incorrect conclusion that an outcome instrument lacks reliability.

Although not all potential sources of random error were addressed by Kibler et al (2002), the researchers satisfied a level of reasonable requirement for this appraisal item. This was achieved by attention to, and discussion about, the provision of basic training for assessors, a testing protocol including repetitions and standardized movement, and creation of a controlled environment, across all measurement occasions. Kibler et al (2002) acknowledge that the pragmatic use of video assessments, while minimizing error due to fatigue, pain and patient position, is likely to underestimate reliability.

Question 5: Were clinically stable participants used in the study?

In studies evaluating intra-rater reliability, patients are measured on two or more occasions by one or more assessors. In research literature, the terms intra-rater reliability and test-retest reliability are often used interchangeably. Test-retest reliability is the ability of an instrument to obtain the same result with repeated administration. This term is often used when evaluating the reliability of questionnaires, and therefore is beyond the scope of this article. Intra-rater reliability, which is a focus of this article, refers to a similar concept, where one assessor gains the same results with repeated use of an instrument (Portney and Watkins, 2000).

Intra-rater reliability can be calculated for each assessor. In contrast, for inter-rater reliability studies, two or more assessors evaluate a group of patients on one occasion (Portney and Watkins, 2000). Inter-rater reliability therefore refers to the consistency of measurements between pairs of assessors, who may not have the same measurement competence or experience (Domholdt, 2005). A common feature of both intra-rater and inter-rater reliability studies is that assessments take place over a period of time.

Patients have the capacity to change in their 'health' status during an assessment (Portney and Watkins, 2000). For example, a patient's pain intensity may increase during the single assessment session in an inter-rater reliability study, as multiple assessors are measuring the same variables over a period of time. In intra-rater reliability studies, a patient's condition may improve or deteriorate between assessments. The above mentioned aspects may impact on the evaluation of the reliability of a measurement tool, if they are not identified and controlled for by the researchers. For example, if a patient's condition improves over time, there may be more variability (i.e. less consistency) in the measurements over time, compared with if the patient's condition remained stable. This may lower the reliability of the measurement tool.

When reading reliability studies it is important to identify if reliability was determined using data from clinically stable participants. Specifically, readers should identify whether there is evidence to suggest that variables which potentially influence the reliability of a measurement tool (e.g. pain intensity, analgesic use) have not significantly changed during the reliability study. This information may be provided in the form of bivariate or multivariate analyses of potentially influencing variables. More frequently, researchers identify and only use data from stable patient participants i.e. patients who report that their health condition has not changed during conduct of the reliability study (Portney and Watkins, 2000). An example of this methodology can be found in a publication by Bialocerkowski (2008).

In the study by Kibler et al (2002), this source of reliability error has been controlled for by the use of video assessment. That is, although scapular dysfunction might be a stable clinical presentation, the need for repeat testing on patient participants has been removed, and as such there can be no change over time in their fatigue, pain or function.

Question 6: Was the period of time between measurements appropriate?

All reliability studies involve patients undergoing repeated measurements. In addition to changes in health status as discussed above, patients may inadvertently influence measurement by recalling their initial performance on physical tests upon subsequent testing. Therefore, the time period between measurements needs to be sufficiently long to reduce the possibility of recall (and to reduce the likelihood of the increase in pain intensity), but short enough to reduce the risk of a change in health status (DeVon et al, 2007).

The optimal period of time between measurements is dependent on a variety of factors, such as the intended purpose of the reliability evaluation, the characteristics of the participants (i.e. the degree of symptomology), the complexity of the measurement process (and the probability of participant fatigue and learning), and the likelihood of recall affecting measurement (Portney and Watkins, 2000). Therefore, when reading a reliability study, the time period that was used between measurements should be described, and readers should

TABLE 2.
Statistical tests used to calculate the reliability

Type of variable	
Continuous	**Categorical**
Paired t test and Pearson correlation coefficient	Percentage agreement
Intraclass Correlation Coefficient and 95% confidence interval	Kappa statistic and 95% confidence interval
Standard error of measurement	
Coefficient of variation	
Bland and Altman's 95% limits of agreement	

Sources: Atkins and Nevill, 1998; Bartlett and Frost, 2008

determine whether this period of time is likely to impact on the measurement process, based upon their knowledge of the instrument and condition under study. Again considering Kibler et al (2002), the use of video recordings means there is no time effect to be considered for the patient participants, who underwent assessment only once. For the assessors, there were 17 days between their first and second video assessments of scapular dysfunction, and it is unlikely that recall of prior results would affect the second measurement.

Question 7: Are the results meaningful?

As stated in Bialocerkowski and Bragge (2008), measurement error is associated with all measurements, and it is the responsibility of the researcher, or the clinician, to minimize these errors as much as possible. However, readers need to interpret whether a reported reliability estimate is adequate for their clinical or research needs. This concept is related to the magnitude of the reliability statistic.

There are many statistical tests that may be used to evaluate reliability (*Table 2*). Considerable controversy exists in the calculation of reliability for continuous variables due to inherent limitations of the statistical tests (listed in *Table 2*). Consequently, various statistical tests are used by researchers to establish reliability (Atkinson and Nevill, 1998). Readers of reliability studies, therefore, require grounding in basic statistics to interpret reliability estimates in context with their intended use of the measurement tool. Below is a discussion on how to interpret two of the most frequently-used reliability statistical tests for continuous-scale rehabilitation measurement tools: the intraclass correlation coefficient and standard error of measurement.

Intraclass correlation coefficients (ICC) are frequently calculated to evaluate the reliability of continuous variables. They yield a number which ranges from 0.00 (no reliability) to 1.00 (perfect agreement) (Portney and Watkins, 2000). Researchers often define, a priori, acceptable reliability in terms of the magnitude of an ICC. Portney and Watkins (2000) state that as a general rule, ICCs below 0.5 can be thought of as representing 'poor reliability'. ICCs between 0.51 and 0.75 indicate 'moderate' reliability and values above 0.75 represent 'good' reliability.

However, decisions regarding acceptable reliability need to be matched to the purpose of the measurement tool (Streiner and Norman, 2003). For example, measurement tools which yield information that is used to make life-threatening health care decisions tend to require reliability estimates well above 0.75. In contrast, ICCs between 0.51 and 0.75 may be deemed acceptable for measures that involve subjective assessor judgements not resulting in critical healthcare decisions. Readers need to consider the threshold ICC value for acceptable reliability in their specific context when appraising reliability articles. However, this is not a simple process, as it is difficult to conceptualize the magnitude of measurement error by interpreting a number which ranges between 0.00 and 1.00.

As a consequence, researchers often calculate the standard error of measurement (SEM) which quantifies measurement error in the same units as the original measurement (Streiner and Norman, 2003). One SEM plus or minus the measured value provides a range that can be interpreted such that the researcher is 68% confident that the true value of the measurement (i.e. the value that is free from measurement error) is contained within this range. If the SEM is multiplied by 1.96, this increases the confidence to 95%. The SEM is calculated by the formula:

$$\text{SEM} = \text{SD}_{\text{test}} \times \sqrt{(1 - r_{\text{test}})}$$

In this formula, SD_{test} is the standard deviation of the measurements taken and r_{test} is the reliability coefficient, such as the ICC. Having a high ICC produces a small SEM. Small SEMs are desirable as they equate to less measurement error (Berg and Latin, 2004).

Readers need to consider the magnitude of the SEM in reliability studies, and determine whether it is appropriate to their clinical or research context. Like the interpretation of the magnitude of an ICC, this is a subjective process. Streiner and Norman (2003) suggest that the scale of the measurement tool should be considered when making this judgement. For example, if a scale was measured in 5 degree increments, a SEM of 5 degrees may be acceptable. However, if the scale was in increments of 0.5 degrees, then a SEM of 5 degrees may be unsatisfactory.

Where measurements are made on a categorical scale, different statistical tests are required to measure agreement. One of the most frequently used statistical tests is the Kappa statistic, which can be used when there are multiple raters, participants and measurement categories (Sim and Wright 2005). The Kappa statistic is a more robust measure of agreement than a simple percentage agreement because it corrects for agreement that would be expected by chance alone. A Kappa value of 0 represents agreement by chance only, and the maximum value of 1.0 represents perfect chance-corrected agreement (Sim and Wright, 2005). Landis and Koch (1977) propose the following scale for evaluating strength of agreement from Kappa values:

- 0.0–0.20: slight
- 0.21–0.40: fair
- 0.41–0.60: moderate
- 0.61–0.80: substantial
- 0.81–1.00: almost perfect.

A weighted Kappa makes further correction for the magnitude of disagreement on a multi-level categorical scale. For example when two raters use a 5-point categorical scale, a weighted Kappa will account for the fact that if Rater 1 rates '3' and Rater 2 rates '4', this is closer in agreement than if Rater 1 rates '1' and Rater 2 rates '5' (Tooth and Ottenbacher, 2004). As with other statistical measures of agreement, readers should interpret Kappa values in the context of their clinical or research setting.

In Kibler et al (2002), the Kappa statistic was used to determine reliability of scapular dysfunction classifications because categorical data, rather than continuous data, were obtained. The results section provides two person comparisons for intra-tester and inter-tester reliability, with results ranging from $\kappa = 0.31$ to $\kappa = 0.59$. Using the Landis and Koch (1977) scale, this represents fair to moderate reliability with intra-tester reliability slightly stronger than inter-tester reliability. Further meaning could have been gained had the authors provided 95% confidence intervals to reflect the sampling error (Sim and Wright, 2005). These results are not sufficiently strong to support the widespread clinical use of this test. The results suggest that this test should be further refined and subsequently evaluated.

Question 8: Can the results be generalized to my clinical/research context?

Reliability is not an inherent characteristic of a measurement tool. Rather, reliability exists within the context in which it was evaluated (Portney and Watkins, 2000). This means that reliability is population-specific (Rothstein, 1985); that is, reliability estimates provide information on the magnitude of measurement error associated with a specific measurement tool, which is used in a specific manner, by a specific group of assessors, on a specific group of individuals. As consumers of research evidence, readers need to be cognisant that:

- Different tools or instruments can be used to evaluate the same variable. However, there is great potential for the measurement error to differ between these tools. Even when using identical tools, variability in measurements may stem from using different administration or scoring protocols. As such, applying reliability evidence to clinical or research settings is dependent on using the same measurement tools and identical administration and scoring protocols
- As described earlier, assessor skills and experience can vary and influence measurement error. Readers, therefore, should identify whether the skill-level of assessors is comparative to their clinical/research counterparts and only apply reliability evidence if comparability exists
- The participant demographic characteristics in a reliability study may be sufficiently different to those of the patients on which you wish to use the measurement tool. Characteristics, such as pain, deformity, spasticity, level of comprehension, weakness and anxiety, may potentially impact on the measurement process (Wright and Feinstein, 1992). Readers should determine whether the demographic characteristics of the study participants are comparative to the intended use of the measurement tool. Reliability evidence should only be applied if comparability exists.

In the study by Kibler et al (2002), the information for determining comparability is found in the methods section. Similar reliability could be anticipated in relevant clinical settings as the study participants and assessors reasonably represent the populations of patients and clinicians for whom the test is intended to be used. Moreover, the assessment procedure has well defined training, administration and scoring protocols. Further research is required to determine the reliability for physical, rather than two-dimensional videotaped, assessments. In addition, Kibler et al (2002) suggested an increase in assessor training which may increase the reliability of this measurement.

TABLE 3.
Critical appraisal questions for reliability studies

Question	Yes	No
1. Is the aim of the study clear and appropriate?		
2. Was the study sample appropriate?		
3. Were a broad range of values generated for the target measurement?		
4. Did the researchers minimise random error in their methodology?		
5. Were clinically stable participants used in the study?		
6. Was the period of time between measurements appropriate?		
7. Are the results meaningful?		
8. Can the results be generalized to my clinical / research context?		

CONCLUSIONS

This article described eight questions that will assist readers to critically appraise reliability studies. These questions are unique to reliability studies and are based on the theory that underpins reliability. These questions are summarized in *Table 3*. It is hoped that these questions will guide clinicians and researchers to make informed decisions regarding whether reliability evidence can be applied to their specific context. IJTR

Conflict of interest: none

Atkinson G, Nevill AM (1998) Statistical methods for assessing measurement error (reliability) in variables relevant to sports medicine. *Sports Med* **26**(4): 217–38

Bartlett JW, Frost C (2008) Reliability, repeatability and reproducibility: analysis of measurement error in continuous variables. *Ultrasound Obstet Gynecol* **31**(4): 466–75

Berg KF, Latin RW (2004) *Essentials of Research Methods in Health, Physical Education, Exercise Science, and Recreation.* Lippincott, Williams and Wilkins, Philadelphia

Bialocerkowski AE (2008) Activity limitations and compensatory mechanism use following limited wrist fusion. *Arthritis Rheum* **59**(10): 1504–11

KEY POINTS

- Consumers of research evidence should critically evaluate reliability evidence before it is applied to their specific context.
- Currently, there is a paucity of information on how to critically appraise reliability studies.
- Based on the theory of reliability, eight critical appraisal questions were developed specifically for reliability studies.
- These questions cover the concepts of: minimization of random error, the use of heterogenous but clinically stable participants, the use of appropriate periods of time between measurements, the presentation of meaningful results, and the generalizability of results.

Bialocerkowski A, Bragge P (2008) Measurement error and reliability testing: application to rehabilitation. *International Journal of Therapy and Rehabilitation* **15**(10): 422-427

Bindra RR, Dias JJ, Heras-Palau C, Amadio PC, Chung KC, Burke FD (2003) Assessing outcome after hand surgery: the current state. *J Hand Surg Br* **28**(4): 289–9

Carmines E, Zeller R (1979) *Reliability and Validity Assessment.* Sage Publications, Beverley Hills

Cicchetti DV (1981) Testing the Normal Approximation and Minimal Sample Size Requirements of Weighted Kappa When the Number of Categories is Large. *Applied Psychological Measurement* **5**(1): 101–04

DeVon HA, Block ME, Moyle-Wright P et al (2007) A psychometric toolbox for testing validity and reliability. *J Nurs Scholarsh* **39**(2): 155–64

Domholdt E (2005) *Physical Ttherapy Research. Principles and Applications.* WB Saunders Company, Philadelphia

Greenhalgh T (1997) How to read a paper: Getting your bearings (deciding what the paper is about). *BMJ* **315**(7102): 243-246

Jerosch-Herold C (2005) An evidence-based approach to choosing outcome measures: a checklist for the critical appraisal of validity, reliability and responsiveness studies. *British Journal of Occupational Therapy* **68**(8): 347–53

Katrak P, Bialocerkowski AE, Massy-Westropp N, Kumar S, Grimmer KA (2004) A systematic review of the content of critical appraisal tools. *BMC Med Research Methodol* **4**: 22 http://tinyurl.com/ydnjnrf (accessed 9 February 2010)

Kibler WB, Uhl TL, Maddux JW, Brooks PV, Zeller B, McMullen J (2002) Qualitative clinical evaluation of scapular dysfunction: a reliability study. *J Shoulder Elbow Surg* **11**(6): 550–56

Landis JR, Koch GG (1977) The measurement of observer agreement for categorical data. *Biometrics* **33**(1): 159–74

Oxman AD, Sackett DL, Guyatt G (1993) Users' guides to the medical literature. I. How to get started. The Evidence-Based Medicine Working Group. *JAMA* **270**(17): 2093–5

Portney LG, Watkins MP (2000) *Foundations of Clinical Research. Applications to Practice.* Prentice Hall Health, Upper Saddle River, New Jersey

Rothstein J (1985) *Measurement and Clinical Practice: Theory and Application*. Churchill Livingstone, New York

Sim J, Wright CC (2005) The kappa statistic in reliability studies: use, interpretation, and sample size requirements. *Phys Ther* **85**(3): 257–68

Straus SE, Richardson WS, Glasziou P, Haynes RB (2005) *Evidence-Based Medicine: How to Practice and Teach EBM (3rd edition).* Churchill Livingstone, Edinburgh

Streiner DL, Norman GR (2003) *Health Measurement Scales: A Practical Guide to their Development and Use (3rd edition)*. Oxford Medical Publications, Oxford

Tooth LR, Ottenbacher KJ (2004) The kappa statistic in rehabilitation research: an examination. *Arch Phys Med Rehabil* **85**(8): 1371–6

Walter SD, Eliasziw M, Donner A (1998). Sample size and optimal designs for reliability studies. *Stat Med* **17**(1): 101–10

Whiting P, Rutjes AW, Reitsma JB, Bossuyt PM, Kleijnen J (2003) The development of QUADAS: a tool for the quality assessment of studies of diagnostic accuracy included in systematic reviews. *BMC Med Res Methodol* **10**(3): 25 http://tinyurl.com/ygczeee (accessed 9 February 2010)

Wright JG, Feinstein AR (1992) Improving the reliability of orthopaedic measurements. *J Bone Joint Surg Br* **74**(2): 287–91