Evaluating Agreement: Conducting a Reliability Study

By Paul J. Karanicolas, MD, PhD, Mohit Bhandari, MD, MSc, FRCSC, Hans Kreder, MD, FRCSC, Antonio Moroni, MD, Martin Richardson, MD, Stephen D. Walter, PhD, Geoff R. Norman, PhD, and Gordon H. Guyatt, MD, MSc, on Behalf of the Collaboration for Outcome Assessment in Surgical Trials (COAST) Musculoskeletal Group*

Instruments that are useful in clinical or research practice will, when the object of measurement is stable, yield similar results when applied at different times, in different situations, or by different users. Studies that measure the relation of differences between patients or subjects and measurement error (reliability studies) are becoming increasingly common in the orthopaedic literature. In this paper, we identify common aspects of reliability studies and suggest features that improve the reader's confidence in the results. One concept serves as the foundation for all further consideration: in order for a reliability study to be relevant, the patients, raters, and test administration in the study must be similar to the clinical or research context in which the instrument will be used. We introduce the statistical measures that readers will most commonly encounter in reliability studies, and we suggest an approach to sample-size estimation. Readers interested in critically appraising reliability studies or in developing their own reliability studies may find this review helpful.

Introduction

With every clinical encounter, surgeons make and interpret measurements. When they assess a patient prior to surgery, they inquire about the patient's age, height, and weight and make an assessment of the patient's pain, range of motion, and physical deformities. Subsequently, they monitor the heart rate, blood pressure, and urine output. All of these measurements are associated with some degree of error. Subconsciously, surgeons are aware of this error and, for every measurement that is taken, they decide how much error they are willing to accept. For example, surgeons would be content knowing the height of a patient within a margin of several centimeters, but a measurement of fracture displacement would need to be much more precise, in the magnitude of millimeters.

The expected range of measurements is the main factor that determines the amount of measurement error that is acceptable. The real worth of a measurement is in how effectively it can be compared to one or more other measurements, either between patients or from the same patient at different times. If the error of a measurement is as large as the expected difference between measurements, the instrument will be useless. In the measurement of height, the expected range of measurements might be 50 to 60 cm, so even with an error of 4 to 5 cm it is still possible to differentiate patients according to the categories of short, average, or tall height. When considering the extent of fracture displacement, the difference between anatomic alignment and severe malalignment may only be 10 to 20 mm, so the measurement error must be much smaller than that.

Reliability refers to the relationship between measurement error and the expected distribution of measurements over time and across observers and situations[1,2]. Reliability is not the same as agreement. The fundamental difference is that reliability is measured relative to the distribution of measurements. A new test that always yields a result of 100.00 regardless of the rater, patient, or any other circumstances would have perfect agreement but would provide no more information to the clinician. This important distinction makes reliability a much more powerful estimate of the usefulness of an instrument than simple measures of agreement are.

Statistically, reliability is the ratio of between-subject variability (in other words, the "true" differences between subjects) to the total variability (the "true" differences plus measurement error), and ranges from 0 (indicating that all of the variation in the sample is due to error) to 1 (indicating perfect reliability; i.e., all variation is due to "true" differences between subjects) (Fig. 1).

An instrument must be reliable in order to be useful in measuring differences between patients. Once investigators have established that an instrument is reliable, they must determine if it measures what it is intended to measure (validity),

*Collaboration for Outcome Assessment in Surgical Trials (COAST) Musculoskeletal Group Steering Committee: Antonio Barquet, Uruguay; Ole Brink, Denmark; Hans Kreder, Canada; Takashi Matsushita, Japan; Antonio Moroni, Italy; Martin Richardson, Australia; Parag Sancheti, India; Paul Tornetta III, USA

Disclosure: In support of their research for or preparation of this work, one or more of the authors received, in any one year, outside funding or grants in excess of $10,000 from the Canadian Institutes of Health Research. Neither they nor a member of their immediate families received payments or other benefits or a commitment or agreement to provide such benefits from a commercial entity.

J Bone Joint Surg Am. 2009;91 Suppl 3:99-106 • doi:10.2106/JBJS.H.01624

$$\text{Reliability} = \frac{\text{True Subject Variability}}{\text{True Subject Variability} + \text{Measurement Error}}$$

Fig. 1

The definition of reliability.

whether the time and expense is practical (feasibility), and whether clinicians will actually use it in practice (acceptability).

In this paper, we identify common aspects of reliability studies and suggest features that improve readers' confidence in the results. One concept serves as the foundation for all further consideration: in order for a reliability study to be relevant, the patients, raters, and test administration in the study must be similar to the clinical or research context in which the instrument will be used.

Features of a Reliability Study

Few guidelines exist to assist readers in appraising reliability studies or to assist researchers in designing them[3]. In this section, we suggest seven questions surgeons can ask themselves as they read a report of a reliability study or write a protocol to conduct one (Table I).

Was the Research Question Appropriate?

Investigators undertaking a reliability study must precisely define which instrument(s) are being tested and how the instrument(s) will be used in clinical or research practice. Furthermore, investigators must determine what type(s) of reliability they will measure in the study. The most common measures are the internal consistency, intraobserver, test-retest, and interobserver reliabilities.

Internal consistency reflects the correlation between an individual's responses within an instrument and suggests whether or not the items seem to be measuring the same thing. For example, Leggin et al. measured the internal consistency of the Penn Shoulder Score, which includes three items regarding pain: pain at rest, pain with normal activities, and pain with strenuous activities[4]. It might be expected that individuals with a high level of pain at rest would also have substantial pain with normal and strenuous activities and, conversely, that patients with no pain at rest might have no or low levels of pain with normal and strenuous activities. In this study, the internal consistency (measured with use of the Cronbach alpha test for inter-item correlation[5]) was 0.93, which indicates a very high correlation between the items.

Perhaps because calculation of internal consistency requires only a single administration of an instrument, internal consistency appears commonly in the literature and authors may refer to the measurement as "reliability." There are, however, many potential sources of measurement error that this calculation does not incorporate, such as differences between times, observers, and settings. Therefore, internal consistency represents the weakest form of reliability, and readers should interpret the results with caution[2].

The other three types of reliability share an important characteristic: they measure the agreement between two or more test administrations.

TABLE I Key Questions to Ask About a Reliability Study

1. Was the research question appropriate?
2. Were the raters representative of the individuals who will apply the instrument in practice?
3. Were the patients or subjects representative of the population that will be rated in practice?
4. Did raters assign the ratings in a clinically relevant manner?
5. Were the data analyzed with use of appropriate reliability statistics?
6. How was the sample size (of raters and subjects) determined?
7. How can the results be interpreted?

Test-retest reliability measures the extent to which one observer who is rating a subject on multiple occasions achieves similar results. Since time elapses between ratings, the characteristics being rated may also change. For example, the range of motion of a knee may change substantially over the course of a two-week period (a common interval used for test-retest reliability measurements).

Intraobserver reliability is similar to test-retest reliability, except that the characteristics being rated are fixed. Of course, this type of measurement is only possible in certain circumstances, such as during the rating of radiographs or videos. Since time is the only factor that varies between administrations, this form of study design will typically yield a higher reliability estimate than that obtained with test-retest or interobserver reliability studies.

Interobserver reliability measures the extent to which two or more observers obtain similar scores when rating the same subject. Interobserver reliability is the broadest and—when error related to observers is highly relevant—the most clinically useful measure of reliability. Since the intraobserver reliability will usually be higher than or equal to the interobserver reliability, if researchers document an acceptable level of interobserver reliability in the appropriate context, no further reliability testing is necessary. However, if the interobserver reliability is poor, knowledge of the test-retest or intraobserver reliability might assist researchers in identifying sources of error and in making appropriate modifications. Furthermore, measuring interobserver reliability is inappropriate if only one individual will apply the test (e.g., self-reported quality-of-life questionnaires); in this situation, the test-retest reliability is more appropriate.

Were the Raters Representative of the Individuals Who Will Apply the Instrument in Practice?

The individuals who make the ratings are an obvious potential source of variation. For instruments that are self-administered (such as quality-of-life questionnaires) the rater is also the subject; we will discuss the principles of selecting these individuals in the next section. Here we outline some important points to consider for situations in which one or more raters apply an instrument to multiple subjects (such as a fracture

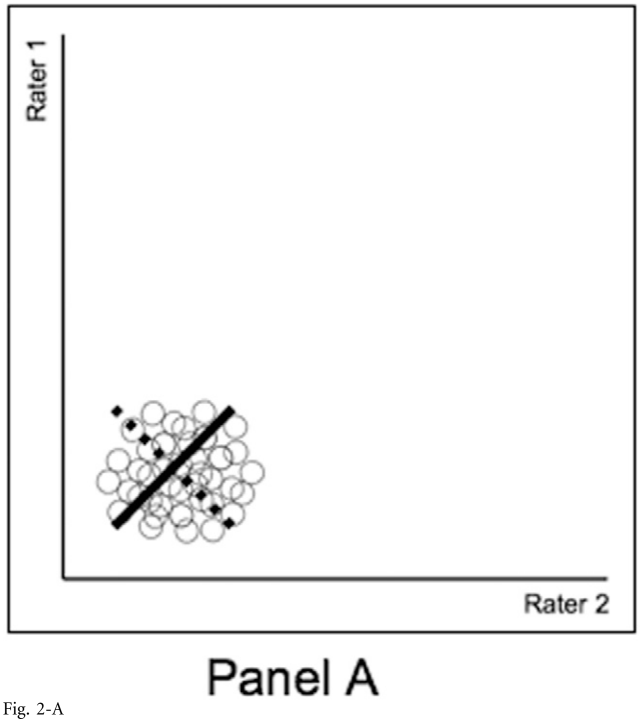

Fig. 2-A

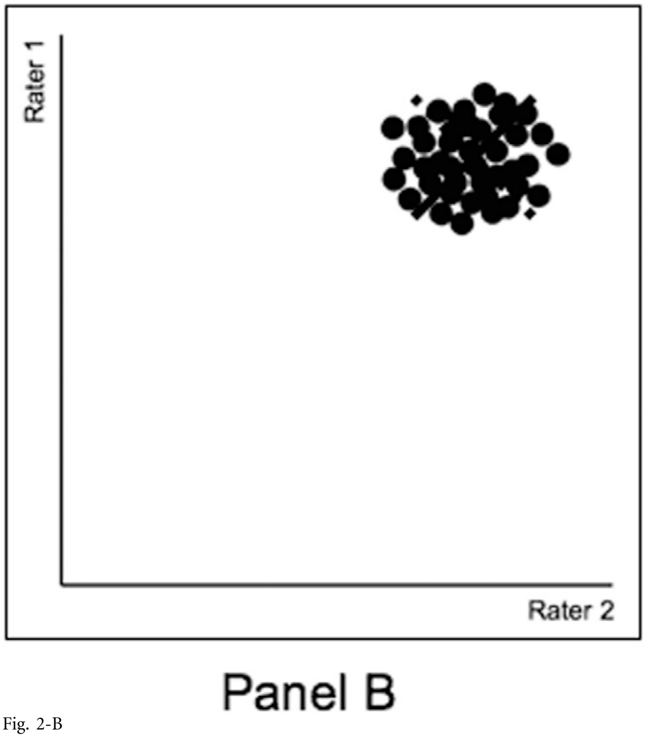

Fig. 2-B

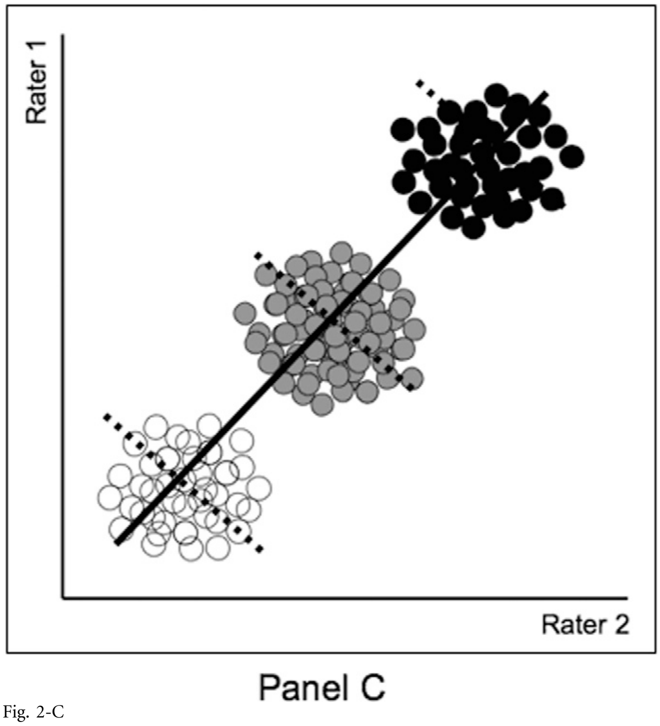

Fig. 2-C

Figs. 2-A, 2-B, and 2-C The influence of subject heterogeneity on reliability. Panel A (Fig. 2-A) and B (Fig. 2-B) depict two studies with homogeneous groups of subjects. Panel C (Fig. 2-C) depicts a study with a heterogeneous group of subjects. The solid lines represent the true between-subject variability, and the dashed lines represent error, or between-rater variability. The reliability in Panel C will be higher than the reliability in Panels A and B.

classification system). Two factors may contribute to the variability between raters: the expertise level of each rater, and the raters' practice settings.

With respect to the level of expertise, a reliable rating is more likely to be assigned by a rater with more training and experience than by a rater with minimal or no training and experience. If raters with varying levels of expertise use the tool in practice (as is usually the case), then including raters with all potential levels of expertise will provide more informative results.

The same principle applies to the raters' practice settings. Of course, it is usually not practical to conduct a study that

TABLE II Advantages and Disadvantages of Common Statistical Measures of Reliability

Statistic	Advantages	Disadvantages
Proportion agreement	Simple	Potentially misleading as it does not account for chance
	Easy to calculate and interpret	Not appropriate for continuous data
(Weighted) Kappa	Accounts for chance	Less accurate if responses are skewed
		Only appropriate for categorical data
Phi	Accounts for chance	Only useful for dichotomous data
	Resistant to skewed responses	Not commonly used
Pearson correlation	Common, easy to calculate	Measures relationship between two variables, not agreement or reliability
Intraclass correlation coefficient	Appropriate for continuous (or categorical) data	Different options may introduce calculation errors
	Different options depending on context	Less transparent

incorporates every level of expertise and practice setting to which surgeons may wish to extrapolate the results; however, researchers should include as diverse and representative a group of raters as possible. When reporting the results of a reliability study, researchers should state who the raters were and provide information regarding the expertise of the raters in the particular rating process.

Were the Patients or Subjects Representative of the Population That Will Be Rated in Practice?

The principles of selecting the patients or subjects for the study are very similar to those discussed for the raters. The patients in the study should represent the actual population that the clinicians will evaluate in clinical practice. For example, in a study assessing knee laxity, investigators measured the intra-rater and inter-rater reliability in a group of twenty healthy volunteers[6]. Unfortunately, this study is of little relevance to clinicians, who are not interested in measuring knee laxity in healthy individuals. This study would have been strengthened if the investigators included patients with very stable knees, very unstable knees, and knees for which stability fell somewhere between the two extremes.

Including patients who represent a broad range of pathology, disability, or whatever the measurement focus of the study happens to be also provides a statistical advantage. Intuitively, it might appear that clinicians would be more likely to agree on ankle stability in a group of healthy volunteers. This highlights the difference between agreement and reliability: although the raw agreement may be higher in a homogeneous (similar) sample, the reliability will be lower.

Figures 2-A, 2-B, and 2-C depict this principle graphically. The panels in Figures 2-A and 2-B represent two studies with homogeneous groups of subjects at both extremes of a scale. In each of these cases, the true between-subject variability (the solid lines) is small relative to the error, or between-rater variability (the dashed lines). The panel in Figure 2-C represents the results from a study involving subjects from each of the extremes, plus a group in the middle. Here the between-subject variability is much larger relative to the error, so the reliability that is measured in this study will be substantially higher than that measured in the other two studies.

Thus, reliability can be manipulated. If you wish to make your instrument appear highly reliable, include normal subjects and those with extreme pathology or dysfunction. If you wish to make a competitor's instrument appear unreliable, choose a homogeneous population. Researchers should resist these temptations and recruit patients who are representative of the spectrum of disease that clinicians will see in practice.

Did Raters Assign the Ratings in a Clinically Relevant Manner?

The administration of the ratings will vary depending on the nature of the raters and the subjects. Nevertheless, the objective of the rating sessions should be the same: to mimic, as closely as possible, the clinical practice environment. For example, when assessing the reliability of a fracture classification system that involves the estimation of lengths and angles, raters should only use tools such as rulers or protractors if they would use them in the real-life clinical practice. Furthermore, researchers must consider what additional information that they will make available to raters, such as patient history or other physical findings. The most pragmatic approach is to provide raters with as much clinical information as they would normally have access to in clinical practice[7]. If, however, researchers wish to determine the impact of different instruments in isolation, they should only provide subject information that is directly relevant to the instrument being tested. Returning to the example of knee laxity, knowledge of the clinical history of a participant could easily influence a rater: clinicians would expect healthy volunteers to have stable knees, while injured patients would be much more likely to demonstrate instability.

Irrespective of the context, all raters should independently complete the evaluations in similar test settings. For example, in a study of classification systems for fractures of the distal part of the radius[8], each rater might view digital radiographs on a personal computer or hard copies from a light box. Either method would

TABLE III Summary of Results from an Intraobserver Reliability Study of Resistance to Shoulder Internal Rotation[16]

		Test 2		
		Strong	Weak	Total
Test 1	Strong	24	1	25
	Weak	1	2	3
	Total	25	3	28

be acceptable (the ideal method would be whichever was used most commonly in clinical practice), but it would not be appropriate for some individuals to view the images on a computer and others to view hard copies unless this variability represented the regular practice of the raters.

A web-based approach is an innovative method of administering reliability studies for radiographic images, such as fracture classification systems. Current web-based technology allows researchers in North America to send images to Asia faster than they can walk into an adjoining office. Researchers in a variety of medical fields have reported that web-based technology has improved efficiency and collaboration in clinical research and practice[9-11]. The Collaboration for Outcome Assessment in Surgical Trials (COAST) has developed a web-based methodology for conducting reliability studies of radiographic images[12].

Were the Data Analyzed with Use of Appropriate Reliability Statistics?

Statisticians have described a wide variety of techniques to measure agreement or reliability (Table II). Given the broad analytical options, investigators should consider calculating and reporting more than one statistical estimate[13]. We will briefly discuss some common forms of reliability analyses for categorical and continuous data; readers interested in learning more about reliability analyses should refer to a statistical text or focused review[2,14,15].

Categorical Data

The simplest measure of agreement, the *proportion* or percentage *agreement,* fails to address the agreement that one would expect due to chance. Consider, for example, the data in Table III, summarized from an intraobserver study of resistance testing in subjects with shoulder pain[16]. Adding the "agreement" cells and dividing by the total yields the proportion agreement; 93% (26 of 28) in this case, which seems extremely good. Table IV displays the data that would result if raters guessed at random, but with the same overall distribution of "strong" to "weak." Here the raw agreement is 79% (22 of 28), which is also quite good. Clearly, the value of 93% does not accurately reflect the reliability of this measure, because it does not account for the agreement that may be due to chance alone. Fortunately, there are several statistical approaches that do address chance agreement.

The *kappa coefficient*, the most commonly reported statistic in orthopaedic fracture reliability studies[1], accounts for chance agreement in categorical responses by comparing the observed agreement with the possible agreement beyond chance[17]. This statistic yields a maximum value of 1.0 (indicating perfect agreement), with 0.0 indicating no agreement beyond chance, and negative values indicating agreement worse than chance. Examining the shoulder stability data once more (Table III), the kappa is 0.63, substantially lower than the raw agreement of 0.93.

Researchers can use kappa to calculate agreement for two or more observers, and with two or more categories of response. In the latter context, if some responses are closer than others (i.e., most commonly ordered responses, such as a severity score of 1 to 4) they can employ a *"weighted" kappa* that incorporates partial agreement[18]. One disadvantage of kappa occurs when the distribution of responses is very skewed: in this case there is little room for agreement above chance, so kappa may be deceivingly small[19].

The *phi* statistic is a measure of "chance-independent" agreement[20]. The biggest advantage of phi is its resistance to skewed distributions. The phi statistic from the shoulder stability data is 0.75, a reliability estimate between the values calculated with kappa and the raw agreement. Since the distribution of responses from this study is skewed, phi is probably the best measure of the reliability. Despite this attractive feature, phi is uncommonly reported in medical statistics.

Continuous Data

The *Pearson correlation* represents a familiar approach to continuous data, but it is limited in that two sets of measurements may be perfectly correlated but have poor agreement. Figure 3 demonstrates this point with data from two hypothetical reliability studies: in both studies, the ratings from reviewer 1 are perfectly correlated with the ratings from reviewer 2. In one study, however, the agreement between reviewers is perfect (red line), while in the other study the reviewers do not actually agree on any measurements (green line). Therefore, the Pearson correlation insufficiently describes the relationship between two variables for the purposes of a reliability study.

Intraclass correlation coefficients are a set of related measures of reliability, derived from a repeated measures analysis of variance[21], that yield a value that is closest to the formal definition of reliability. One intraclass correlation coefficient measures the proportion of total variability that is due

TABLE IV Expected Results from an Intraobserver Reliability Study of Resistance to Shoulder Internal Rotation When Rater Guessed at Random, but with the Same Total Distribution of Responses

		Test 2		
		Strong	Weak	Total
Test 1	Strong	22	3	25
	Weak	3	0	3
	Total	25	3	28

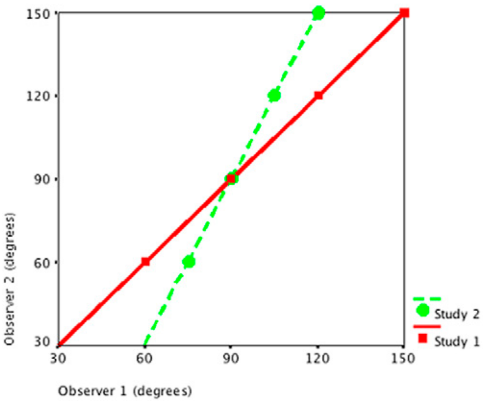

Fig. 3

Scatter-plot of results from two hypothetical reliability studies. The agreement between observers is better in study 1 than it is in study 2, but the Pearson correlation is 1.0 in both.

to true between-subject variability[22]. Although analysts most commonly calculate an intraclass correlation coefficient for continuous outcomes, when applied to categorical data it is equivalent to the weighted kappa with quadratic weighting. Several variations of the intraclass correlation coefficient facilitate its use in addressing a variety of reliability issues[2].

In summary, data analysts have many statistical options available to estimate the reliability of two or more sets of measurements. The following are the most commonly reported statistics: kappa for dichotomous responses, weighted kappa for polytomous (more than two categories) responses, and intraclass correlation coefficient for continuous data. Investigators who encounter more complex analytical situations, such as a comparison of two or more reliability estimates[23,24] or separation of the error into multiple sources (such as observers, times, and locations) in a single analysis[23,25], should involve a statistician familiar with these techniques.

How Was the Sample Size (of Raters and Subjects) Determined?

Researchers control the size of two samples in reliability studies: the number of raters and the number of subjects. Although increasing the number in either group will yield a more precise reliability estimate (a narrower confidence interval), the number of subjects has a much greater impact on the precision than the number of raters does (especially when there are more than four or five raters)[2]. Therefore, we recommend determining the number of raters based on generalizability and feasibility, then estimating the number of subjects required to achieve the desired precision.

The number of raters that are needed to satisfy the generalizability requirement depends on the characteristics of the raters. The feasibility of performing multiple ratings also depends on the nature of the subjects: radiographs can easily be rated several times by different individuals, but living patients are unlikely to be as accommodating. Thus, the ultimate decision about the number of raters to include involves balancing the theoretical benefits of increased generalizability with feasibility considerations.

When the number of raters has been determined, investigators can perform a sample-size calculation to estimate the required number of subjects. As with any sample-size estimation, the calculation is dependent on the analysis plan. We will describe the approach to sample-size estimation for studies that use an intraclass correlation coefficient; interested readers can find estimates for other reliability statistics in the cited material[15].

Researchers may use two approaches to estimate the appropriate number of subjects. In the first method, investigators choose the minimum acceptable reliability and estimate the sample size needed to prove that the actual reliability is higher[26]. In most reliability studies, the minimum acceptable reliability is not intuitively obvious. The second approach is based on the desired precision of the reliability estimate. The calculation incorporates the number of raters, the expected intraclass correlation coefficient (estimated from past studies or simply a "best guess"), the confidence interval (usually

TABLE V Sample-Size Estimation with Use of the Intraclass Correlation Coefficient as Based on Giraudeau and Mary[27]

Number of Raters	Expected Intraclass Correlation Coefficient	Number of Subjects Required for 95% Confidence Interval at Three Different Confidence Levels		
		±0.05	±0.10	±0.20
2	0.9	56	14	4
	0.8	200	50	13
	0.7	400	100	25
	0.6	630	158	40
	0.5	865	217	55
4	0.9	36	9	3
	0.8	119	30	8
	0.7	222	56	14
	0.6	322	81	21
	0.5	401	101	26
6	0.9	31	8	2
	0.8	103	26	7
	0.7	187	47	12
	0.6	263	66	17
	0.5	314	79	20
10+	0.9	29	8	2
	0.8	92	23	6
	0.7	164	41	11
	0.6	224	56	14
	0.5	259	65	17

95%), and the width of the confidence interval. Table V displays sample-size estimates for selected parameters; readers may find a description of the full calculation elsewhere[27].

How Can the Results Be Interpreted?

Fortunately, most of the statistics that we have discussed yield values on the same scale: 0.0 indicates that all of the variability is due to error, and 1.0 indicates that all of the variability is due to true between-subject differences. Unfortunately, reliability studies rarely yield estimates close to either of these values; actual results are more likely to be somewhere between 0.3 and 0.7[1]. So what is an "acceptable" level of reliability?

Researchers have proposed guidelines to assist readers in interpreting reliability estimates[28-32]. All of these are variations of the same theme and not surprising: values close to 0 (or negative) represent poor reliability, values close to 1.0 represent excellent reliability, and values around 0.5 represent moderate reliability. Ultimately, whether or not a given level of reliability is acceptable will depend on the context of the measurement and the other instruments available. If the instrument being studied is the only tool available to measure an important quality, then it will have to suffice until investigators develop a more reliable tool.

Because interpretation of a reliability study is context-specific, readers must determine if the raters, subjects, and instrument administration in the study reflect their clinical or research setting. If the contexts are similar, readers may comfortably expect similar reliability in their setting. However, if the settings are sufficiently different, readers must apply the results cautiously or repeat the reliability testing in more applicable circumstances.

Summary

When considering implementing a new instrument into research or clinical practice, potential users should first ensure that the reliability of the instrument has been measured appropriately in a similar setting. The features discussed in this article will help readers to critically appraise reliability studies and to design their own rigorous reliability studies. ■

Paul J. Karanicolas, MD, PhD
Mohit Bhandari, MD, MSc, FRCSC
Stephen D. Walter, PhD
Geoff R. Norman, PhD
Gordon H. Guyatt, MD, MSc,
Department of Clinical Epidemiology and Biostatistics, McMaster University, 293 Wellington Street North, Suite 110, Hamilton, ON L8L 2X2, Canada. E-mail address for M. Bhandari: bhandam@mcmaster.ca

Hans Kreder, MD, FRCSC
Department of Surgery, University of Toronto,
Sunnybrook Campus, MG365, 2075 Bayview Avenue,
Toronto, ON M4N 3M5, Canada

Antonio Moroni, MD
Rizzoli Orthopaedic Institute, University of Bologna,
Via G.C. Pupilli 1, Bologna 40136, Italy

Martin Richardson, MBBS, FRACS (Orth)
Department of Orthopaedic Surgery, Royal Melbourne Hospital,
University of Melbourne, Grattan Street, Parkville 3050,
Melbourne, Victoria, Australia

References

1. Audigé L, Bhandari M, Kellam J. How reliable are reliability studies of fracture classifications? A systematic review of their methodologies. Acta Orthop Scand. 2004;75:184-94.

2. Streiner DL, Norman GR. Health measurement scales: a practical guide to their development and use. 3rd ed. Oxford: Oxford University Press; 2003.

3. Karanicolas PJ. The reliability study. In: Bhandari M, Joensson A, editors. Clinical research for surgeons. Stuttgart, Germany: Thieme Medical; 2009. p 91-9.

4. Leggin BG, Michener LA, Shaffer MA, Brenneman SK, Iannotti JP, Williams GR Jr. The Penn shoulder score: reliability and validity. J Orthop Sports Phys Ther. 2006;36:138-51.

5. Cronbach LJ. Coefficient alpha and the internal structure of tests. Psychometrika. 1951;16:297-334.

6. van der Esch M, Steultjens M, Ostelo RW, Harlaar J, Dekker J. Reproducibility of instrumented knee joint laxity measurement in healthy subjects. Rheumatology (Oxford). 2006;45:595-9.

7. Irwig L, Macaskill P, Walter SD, Houssami N. New methods give better estimates of changes in diagnostic accuracy when prior information is provided. J Clin Epidemiol. 2006;59:299-307.

8. Flikkilä T, Nikkola-Shito A, Kaarela O, Pääkkö E, Raatikainen T. Poor interobserver reliability of AO classification of fractures of the distal radius. Additional computed tomography is of minor value. J Bone Joint Surg Br. 1998;80:670-2.

9. Boulos MN, Maramba I, Wheeler S. Wikis, blogs and podcasts: a new generation of Web-based tools for virtual collaborative clinical practice and education. BMC Med Educ. 2006;6:41.

10. Castel JM, Figueras A, Vigo JM. The internet as a tool in clinical pharmacology. Br J Clin Pharmacol. 2006;61:787-90.

11. Sachs NA, Nulud PL, Loeb, GE. Virtual Visit: improving communication for those who need it most. Stud Health Technol Inform. 2003;94: 302-8.

12. Collaboration for Outcome Assessment in Surgical Trials. http://www.coastresearch.ca. 2007 Apr 16.

13. Luiz RR, Szklo M. More than one statistical strategy to assess agreement of quantitative measurements may usefully be reported. J Clin Epidemiol. 2005;58:215-6.

14. Ludbrook J. Statistical techniques for comparing measurers and methods of measurement: a critical review. Clin Exp Pharmacol Physiol. 2002;29: 527-36.

15. Sim J, Wright CC. The kappa statistic in reliability studies: use, interpretation, and sample size requirements. Phys Ther. 2005;85: 257-68.

16. Hayes KW, Petersen CM. Reliability of classifications derived from Cyriax's resisted testing in subjects with painful shoulders and knees. J Orthop Sports Phys Ther. 2003;33:235-46.

17. Cohen J. A coefficient of agreement for nominal scales. Educ Psychol Meas. 1960;20:37-46.

18. Cohen J. Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. Psychol Bull. 1968;70: 213-20.

19. Kraemer HC. Ramifications of a population model for kappa as a coefficient of reliability. Psychometrika. 1979;44:461-72.

20. McGinn T, Guyatt G, Cook R, Meade M. Measuring agreement beyond chance. In: Guyatt G, Rennie D, editors. Users' guides to the medical literature: a manual for evidence-based clinical practice. Chicago: AMA Press; 2001. p 461-70.

21. Fisher RA. Statistical methods for research workers. Edinburgh: Oliver and Boyd; 1925.

22. Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. Psychol Bull. 1979;86:420-8.

23. Donner A, Shoukri MM, Klar N, Bartfay E. Testing the equality of two dependent kappa statistics. Stat Med. 2000;19:373-87.

24. Donner A, Zou G. Testing the equality of dependent intraclass correlation coefficients. Statistician. 2002;51:367-79.

25. Shavelson RJ, Webb NM, Rowley GL. Generalizability theory. Am Psychol. 1989;44:922-32.

26. Walter SD, Eliasziw M, Donner A. Sample size and optimal designs for reliability studies. Stat Med. 1998;17:101-10.

27. Giraudeau B, Mary JY. Planning a reproducibility study: how many subjects and how many replicates per subject for an expected width of the 95 per cent confidence interval of the intraclass correlation coefficient. Stat Med. 2001; 20:3205-14.

28. Altman DG. Practical statistics for medical research. London: Chapman & Hall/CRC; 1990. Inter-rater agreement; p 403-9.

29. Brage ME, Rockett M, Vraney R, Anderson R, Toledano A. Ankle fracture classification: a comparison of reliability of three X-ray views versus two. Foot Ankle Int. 1998;19:555-62.

30. Fleiss JL. Statistical methods for rates and proportions. 2nd ed. New York: Wiley-Interscience; 1981.

31. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics. 1977;33:159-74.

32. Svanholm H, Starklint H, Gundersen HJ, Fabricius J, Barlebo H, Olsen S. Reproducibility of histomorphologic diagnoses with special reference to the kappa statistic. APMIS. 1989;97:689-98.