Reliability: on the reproducibility of assessment data

Steven M Downing

CONTEXT All assessment data, like other scientific experimental data, must be reproducible in order to be meaningfully interpreted.

PURPOSE The purpose of this paper is to discuss applications of reliability to the most common assessment methods in medical education. Typical methods of estimating reliability are discussed intuitively and non-mathematically.

SUMMARY Reliability refers to the consistency of assessment outcomes. The exact type of consistency of greatest interest depends on the type of assessment, its purpose and the consequential use of the data. Written tests of cognitive achievement look to internal test consistency, using estimation methods derived from the test-retest design. Rater-based assessment data, such as ratings of clinical performance on the wards, require interrater consistency or agreement. Objective structured clinical examinations, simulated patient examinations and other performance-type assessments generally require generalisability theory analysis to account for various sources of measurement error in complex designs and to estimate the consistency of the generalisations to a universe or domain of skills.

CONCLUSIONS Reliability is a major source of validity evidence for assessments. Low reliability indicates that large variations in scores can be expected upon retesting. Inconsistent assessment scores are difficult or impossible to interpret meaningfully and thus reduce validity evidence. Reliability coefficients allow the quantification and estimation of the random errors of measurement in assessments, such that overall assessment can be improved.

Department of Medical Education, College of Medicine, University of Illinois at Chicago, Chicago, Illinois, USA

Correspondence: Steven M Downing PhD, Associate Professor of Medical Education, University of Illinois at Chicago, College of Medicine, Department of Medical Education (MC 591), 808 South Wood Street, Chicago, Illinois 60612-7309, USA. Tel: 00 1 312 996 6428; Fax: 00 1 312 413 2048; E-mail: sdowning@uic.edu.

KEYWORDS education, medical, undergraduate/*standards; educational measurement/*standards; reproducibility of results.

Medical Education 2004; 38: 1006–1012
doi:10.1046/j.1365-2929.2004.01932.x

INTRODUCTION

This article discusses reliability of assessments in medical education and presents examples of various methods used to estimate reliability. It expands the brief discussion of reliability by Crossley *et al.*[1] in an earlier paper in this series and discusses uses of generalisability theory, which have been described in detail elsewhere.[2–4] The emphasis of this paper is applied and practical, rather than theoretical.

What is reliability? In its most straightforward definition, reliability refers to the reproducibility of assessment data or scores, over time or occasions. Notice that this definition refers to reproducing scores or data, so that, just like validity, reliability is a characteristic of the result or outcome of the assessment, not the measuring instrument itself. Feldt and Brennan[5] suggest that: 'Quantification of the consistency and inconsistency in examinee performance constitutes the essence of reliability analysis.' (p 105)

This paper explores the importance of reliability in assessments, some types of reliability that are commonly used in medical education and their methods of estimation, and the potential impact on students of using assessments with low reliability.

THE CONSISTENCY OF ASSESSMENT DATA

One fundamental principle of the scientific method is that experiments must be reproducible in order to be properly interpreted or taken seriously. If another

Overview

What is already known on this subject

Reliability is a major 'quality index' of assessment data.

Reliability refers to different types of 'consistencies', depending on the source of assessment data.

Reliability coefficients estimate random measurement error in assessment data.

What this study adds

High pass/fail *decision* reliability is essential for high stakes examinations.

Intraclass correlation appropriately estimates interrater reliability.

Small amounts of unreliability may cause misclassification errors and large score differences on retesting.

Suggestions for further research

Research and development is needed to make generalisability theory, with all its many versatile and useful applications, more understandable, accessible and user-friendly for medical educators.

researcher can not reproduce the results of an experiment, more or less, any conclusions drawn from the original experiment are suspect and generalisations are limited. This is also true for assessment data, which must have the property of reliability, such that the outcomes or scores can be meaningfully reproduced and interpreted. If the results of assessments are not consistent or not reproducible, what meaningful interpretation can be argued or defended? Thus, reliability is a necessary but not sufficient condition for validity[6] and reliability is a major source of validity evidence for all assessments.[7,8] In the absence of sufficient reliability, assessment data are uninterruptible, as the data resulting from low reliability assessments have a large component of random error.

There are many types of reliability discussed in the educational measurement literature. All reliability estimates quantify some consistency of measurement and indicate the amount of random error associated with the measurement data. The specific purpose of the assessment dictates what type of consistency or reliability is of greatest importance.

Theoretically, reliability is defined in classical measurement theory (CMT) as the ratio of true score variance to total score variance. (As reliability coefficients are interpreted like correlation coefficients, it is also accurate to think of reliability as the squared correlation of the true scores with the observed scores.[5]) Starting from the basic definitional formula, $X = T + e$ (the observed score is equal to the true score plus random errors of measurement), and making some statistical assumptions along the way, one can derive all the formulae commonly used to estimate reliability or reproducibility of assessments. In the ideal world there would be no error term in the formula and all observed scores would always be exactly equal to the true score (defined as the long-run mean score, much like μ, the population mean score). In reality, the measurement world contains much random error and reliability coefficients are used to estimate the amount of measurement error in assessments.

The next section overviews reliability and some methods of reliability estimation in the context of the 3 types of assessments most commonly used in medical education: written assessments of cognitive achievement, clinical performance assessments and oral examinations, and highly structured performance examinations, such as simulated patient examinations.

RELIABILITY OF ACHIEVEMENT EXAMINATIONS

The approach typically utilised to estimate the reproducibility of test scores in written examinations employs the concept of internal consistency, usually estimated by the Cronbach alpha[9] coefficient or Kuder-Richardson formula 20 (KR 20).[10] The logic of internal test consistency reliability is straightforward and intuitive. The statistical derivation of these formulae starts with the test-retest concept, such that a test is given on 1 occasion to a group of examinees and the same test (or an equivalent form of the same test) is re-administered to the same group of students at a later time (assuming that the students have not learned or forgotten anything between tests). If the test produces reliable scores, the students should obtain nearly the same scores at the second testing as at the first testing. While this test-retest concept is the foundation of most of the reliability estimates used in medical education, the test-retest design is rarely, if ever, used in actual practice, as it is logistically so difficult to carry out.

Happily, measurement statisticians sorted out ways to estimate the test-retest condition many years ago, from a single testing.[11] The logic is: the test-retest design divides a test into 2 random halves, perhaps scoring the even-numbered items as the first test and the odd-numbered questions as the second test. Assuming that a single construct is measured by the entire test, the 2 random half tests are a reasonable proxy for 2 complete tests administered to the same group of examinees. Further, the correlation of the scores from the 2 random half tests approximates the test-retest reproducibility of the examination scores. (Note that this is the reliability of only half of the test and a further calculation must be applied, using the Spearman–Brown prophecy formula, in order to determine the reliability of the complete examination.[12])

A further statistical derivation (making a few assumptions about the test and the statistical characteristics of the items) allows one to estimate internal consistency reliability from all possible ways to split the test into 2 halves: this is Cronbach's alpha coefficient, which can be used with polytomous data (0, 1, 2, 3, 4, …n) and is the more general form of the KR 20 coefficient, which can be used only with dichotomously scored items (0, 1), such as typically found on selected-response tests.

A high internal consistency reliability estimate for a written test indicates that the test scores would be about the same, if the test were to be repeated at a later time. Moreover, the random errors of measurement (e.g. examinee fatigue or inattention, intra-individual differences, blind guessing and so on) are reasonably low so that the test scores exemplify an important source of validity evidence: score reproducibility.

RELIABILITY OF RATER DATA

Ward evaluation of clinical performance is often used in medical education as a major assessment method in clinical education. In addition, oral examinations are used in some settings in which raters or judges evaluate the spoken or oral performance of medical students or residents. Both are examples of assessments that depend on the consistency of raters and ratings for their reproducibility or reliability.

For all assessments that depend on human raters or judges for their primary source of data, the reliability or consistency of greatest interest is that of the rater or judge. The largest threat to the reproducibility of such clinical or oral ratings is rater inconsistency or low interrater reproducibility. (Technically, in most designs, raters or judges are nested or confounded with the items they rate or the cases, or both, so that it is often impossible to directly estimate the error associated with raters except in the context of items and cases.)

The internal consistency (alpha) reliability of the rating scale (all items rated for each student) may be of some marginal interest to establish some communality for the construct assessed by the rating scale, but interrater reliability is surely the most important type of reliability to estimate for rater-type assessments.

There are many ways to estimate interrater reliability, depending on the statistical elegance desired by the investigator. The simplest type of interrater reliability is 'percent agreement', such that for each item rated, the agreement of the 2 (or more) independent raters is calculated. Percent-agreement statistics may be acceptable for in-house or everyday use, but would likely not be acceptable to manuscript reviewers and editors of high quality publications, as these statistics do not account for the chance occurrence of agreement. The kappa[13] statistic (a type of correlation coefficient) does account for the random-chance occurrence of rater agreement and is therefore sometimes used as an interrater reliability estimate, particularly for individual questions, rated by 2 independent raters. (The phi[14] coefficient is the same general type of correlation coefficient, but does not correct for chance occurrence of agreement and therefore tends to overestimate true rater agreement.)

The most elegant estimates of interrater agreement use generalisability theory (GT) analysis.[2–4] From a properly designed GT study, one can estimate variance components for all the variables of interest in the design: the persons, the raters and the items. An examination of the percentage of variance associated with each variable in the design is often instructive in understanding fully the measurement error due to each design variable. From these variance components, the generalisability coefficient can be calculated, indicating how well these particular raters and these specific items represent the whole universe of raters and items, how representative this particular assessment is with respect to all possible similar assessments, and, therefore, how much we can trust the consistency of the ratings. In addition, a direct assessment of the rater error can be estimated in such a study as an index of interrater consistency.

A slightly less elegant, but perhaps more accessible method of estimating interrater reliability is by use of the intraclass correlation coefficient.[15] Intraclass correlation uses analysis of variance (ANOVA), as does generalisability theory analysis, to estimate the variance associated with factors in the reliability design. The strength of intraclass correlation used for interrater reliability is that it is easily computed in commonly available statistical software and it permits the estimation of both the actual interrater reliability of the *n*-raters used in the study as well as the reliability of a single rater, which is often of greater interest. Additionally, missing ratings, which are common in these datasets, can be managed by the intraclass correlation.

RELIABILITY OF PERFORMANCE EXAMINATIONS: OSCES AND SPS

Some of the most important constructs assessed in medical education are concerned with behaviour and skills such as communication, history taking, diagnostic problem solving and patient management in the clinical setting. While ward-type evaluations attempt to assess some of these skills in the real setting (which tends to lower reliability due to the interference of many uncontrolled variables and a lack of standardisation), simulated patient (SP) examinations and objective structured clinical examinations (OSCEs) can be used to assess such skills in a more standardised, controlled fashion.

Performance examinations pose a special challenge for reliability analysis. Because the items rated in a performance examination are typically nested in a case, such as an OSCE, the unit of reliability analysis must necessarily be the case, not the item. One statistical assumption of all reliability analyses is that the items are locally independent, which means that all items must be reasonably independent of one another. Items nested in sets, such as an OSCE, an SP examination, a key features item set[16] or a testlet[17] of multiple choice questions (MCQs), generally violate this assumption of local independence. Thus, the case set must be used as the unit of reliability analysis. Practically, this means that if one administers a 20-station OSCE, with each station having 5 items, the reliability analysis must use the 20 OSCE scores, not the 100 individual item scores. The reliability estimate for 20 observations will almost certainly be lower than that for 100 observations.

The greatest threat to reliable measurement in performance examinations is case specificity, as is well documented.[18,19] Complex performance assessments require a complex reliability model, such as generalisability theory analysis, to properly estimate sources of measurement error variance in the design and to ultimately estimate how consistently a particular sample of cases, examinees and SPs can be generalised to the universe or domain.

HOW ARE RELIABILITY COEFFICIENTS USED IN ASSESSMENT?

There are many ways to use reliability estimates in assessments. One practical use of the reliability coefficient is in the calculation of the standard error of measurement (SEM). The SEM for the entire distribution of scores on an assessment is given by the formula:[12]

$$\text{SEM} = \text{Standard Deviation} \times \sqrt{(1 - \text{Reliability})}.$$

This SEM can be used to form confidence bands around the observed assessment score, indicating the precision of measurement, given the reliability of the assessment, for each score level.

HOW MUCH RELIABILITY IS ENOUGH?

The most frequently asked question about reliability may be: how high must the reliability coefficient be in order to use the assessment data? The answer depends, of course, on the purpose of the assessment, its ultimate use and the consequences resulting from the assessment. If the stakes are extremely high, the reliability must be high in order to defensibly support the validity evidence for the measure. Various authors, textbook writers and researchers offer a variety of opinions on this issue, but most educational measurement professionals suggest a reliability of at least 0.90 for very high stakes assessments, such as licensure or certification examinations in medicine, which have major consequences for examinees and society. For more moderate stakes assessments, such as major end-of-course or end-of-year summative examinations in medical school, one would expect reliability to be in the range of 0.80–0.89, at minimum. For assessments with lower consequences, such as formative or summative classroom-type assessments, created and administered by local faculty, one might expect reliability to be in the range of 0.70–0.79 or so. Shorter tests, of course, will tend to have lower reliability, so that 'check-up' type quizzes, given in the classroom or clinic, may have considerably lower reliability.

The consequences on examinees of false positive or false negative outcomes of the assessment are far more important than the 'absolute value' of the reliability coefficient. One excellent use of the reliability coefficient is the estimation of the degree of confidence one can have in the pass/fail decision made on the basis of the assessment scores. For example, in assessments with very high consequences, the degree of confidence one has in the accuracy of pass/fail classification is very important. Unreliability, of course, tends to reduce confidence in the status or outcome classification of examinees. One method of estimating this pass/fail decision reproducibility was presented by Subkoviak[20] and permits a calculation of a pass/fail reproducibility index, indicating the degree of confidence one can place on the pass/fail outcomes of the assessment. Pass/fail decision reliability, ranging from 0.0 to 1.0, is interpreted as the probability of an identical pass or fail decision being made upon retesting. Generalisability theory also permits a calculation of the precision of measurement at the cut score (a standard error of measurement at the passing score), which can be helpful in evaluating this all-important accuracy of classification.

What are some of the practical consequences of low reliability of the interpretation of assessment data? Wainer and Thissen[21] discuss the expected change in test scores, upon retesting, for various levels of score reliability (Table 1). Their analyses were conducted by simulating test score data, at various levels of score reliability, performing scatter plots on the test-retest data, and then calculating the percentage of scores (in standard deviation units) that changed between the 2 simulated test administrations.

Expected changes in test scores upon retesting can be quite large, especially for lower levels of reliability. Consider this example: a test score distribution has a mean of 500 and a standard deviation of 100. If the score reliability is 0.50, the standard error of measurement equals 71.

*Table 1 Expected proportion of examinees at 3 levels of score change by reliability**

Expected percentage of scores change by more than:

Reliability	**0.5 SD change**	**1.0 SD change**	**1.5 SD change**
0.00	72%	48%	29%
0.50	62%	32%	13%
0.70	52%	20%	5%
0.80	43%	11%	2%
0.90	26%	3%	0.1%

* Wainer and Thissen[21] (Table 1, p 24).

Thus, a 95% confidence interval for a student scoring of 575 on this test is 575 ± 139. Upon retesting this student, we could reasonably expect 95/100 retest scores to fall somewhere in the range of 436–714. This is a very wide score interval, at a reliability level that is not uncommon, especially for rater-based oral or performance examinations in medical education. Even at a more respectable reliability level of 0.75, using the same data example above, we would reasonably expect this student's scores to vary by up to 98 score points upon repeated retesting. The effect of reliability on reasonable and meaningful interpretation of assessment scores is indeed real.

IMPROVING RELIABILITY OF ASSESSMENTS

There are several ways to improve the reliability of assessments. Most important is the use of sufficiently large numbers of test questions, raters or performance cases. One frequent cause of low reliability is the use of far too few test items, performance cases or raters to adequately sample the domain of interest. Make sure the questions or performance prompts are clearly and unambiguously written and that they have been thoroughly reviewed by content experts. Use test questions or performance cases that are of medium difficulty for the students being assessed. If test questions or performance prompts are very easy or very hard, such that nearly all students get most questions correct or incorrect, very little information is gained about student achievement and the reliability of these assessments will be low. (In mastery-type testing, this will present different issues.)

If possible, obtain pretest or tryout data from assessments before they are used as live or scored questions. This may be nearly impossible for classroom-type assessments and even for some larger-scale medical school assessments. However, it is possible to bank effective test questions or performance cases in secure item pools for reuse later. Given the great cost and difficulty of creating effective, reliable test questions and performance prompts, securing effective questions and prompts which have solid psychometric characteristics can add greatly to reliable measurement.

CONCLUSION

Reliability estimates the amount of random measurement error in assessments. All reliability analyses are concerned with some type of consistency of measurement. For written tests, the internal test consistency is generally most important, estimated by reliability indices such as Cronbach's alpha or the Kuder-Richardson formula 20. The internal consistency coefficients are all derived from the test-retest design and approximate the results of such test-retest experiments.

Rater-based assessments, such as ratings of clinical performance on the wards or oral examinations at the bedside, look to interrater reliability or reproducibility as their major sources of consistency. Several methods of estimating interrater reliability have been reviewed here, with generalisability theory analysis suggested as the most statistically elegant. Use of the intraclass correlation to estimate intrarater reliability is perfectly legitimate and may be somewhat more accessible than GT for some medical educators.

Performance assessments, such as OSCEs and SP examinations, must use the case as the unit of reliability analysis and will benefit from the use of GT to estimate various sources of measurement error in the design.

Use of the standard error of measurement to create confidence bands around observed scores is suggested as one very practical use of reliability in practice. Calculation of the pass/fail decision reliability is noted as important for high stakes examinations.

In order to improve the reliability of assessments, one should maximise the number of questions or prompts, aim for middle difficulty questions, and make certain that all assessment questions are unambiguous and clearly written and are, if possible, critiqued by content-expert reviewers. Pretesting, item tryout and item banking are recommended as means of improving the reliability of assessments in medical education, wherever possible.

ACKNOWLEDGEMENTS

The author wishes to thank Brian Jolly PhD and an anonymous reviewer for their critical reviews of this manuscript.

FUNDING

There was no external funding for this project.

ETHICAL APPROVAL

This review-type paper had no requirement for University of Illinois at Chicago Institutional Review Board (IRB) approval, and, thus, was not submitted for review or approval.

REFERENCES

1 Crossley J, Humphris G, Jolly B. Assessing health professionals. *Med Educ* 2002;**36**:800–4.
2 Brennan RL. *Generalizability Theory*. New York: Springer-Verlag 2001.
3 Crossley J, Davies H, Humphris G, Jolly B. Generalisability: a key to unlock professional assessment. *Med Educ* 2002;**36**:972–8.
4 Streiner DL, Norman GT. *Health Measurement Scales: A Practical Guide to their Development and Use*. 3rd edn. New York: Oxford University Press 2003.
5 Feldt LS, Brennan RL. Reliability. In: Linn RL, ed. *Educational Measurement*. 3rd edn. New York: American Council on Education, Macmillan 1989: 105–46.
6 Mehrens WA, Lehmann IJ. *Measurement and Evaluation in Education and Psychology*. 4th edn. New York: Harcourt Brace College Publishers 1991.
7 American Educational Research Association, American Psychological Association, National Council on Measurement in Education. *Standards for Educational and Psychological Testing*. Washington DC: American Educational Research Association 1999.
8 Downing SM. Validity: on the meaningful interpretation of assessment data. *Med Educ* 2003;**37**:830–7.
9 Cronbach LJ. Coefficient alpha and the internal structure of tests. *Psychometrika* 1951;**16**:297–334.
10 Kuder GF, Richardson MW. The theory of the estimation of test reliability. *Psychometrika* 1937;**2**:151–60.

11 Stanley JC. Reliability. In: Thorndike RL, ed. *Educational Measurement.* 2nd edn. Washington: American Council on Education 1971: 356–442.

12 Crocker L, Algina J. *Introduction to Classical and Modern Test Theory.* Belmont, California: Wadsworth Group/Thomson Learning 1986.

13 Cohen J. A coefficient of agreement for nominal scales. *Educ Psych Measurement* 1960;**10**:37–47.

14 Howell DC. *Statistical Methods for Psychology.* 5th edn. Pacific Grove, California: Duxbury 2002.

15 Ebel RL. Estimation of the reliability of ratings. *Psychometrika* 1951;**16**:407–24.

16 Page G, Bordage G. The Medical Council of Canada's Key Features Project: a more valid written examination of clinical decision making skills. *Acad Med* 1995;**70**:104–10.

17 Thissen D, Wainer H, eds. *Test Scoring.* Mahwah, New Jersey: Lawrence Erlbaum 2001.

18 Elstein A, Shulman L, Sprafka S. *Medical Problem-Solving: An Analysis of Clinical Reasoning.* Cambridge, Massachusetts: Harvard University Press 1978.

19 van der Vleuten CPM, Swanson DB. Assessment of clinical skills with standardised patients: state of the art. *Teach Learn Med* 1990;**2**:58–76.

20 Subkoviak MJ. A practitioner's guide to computation and interpretation of reliability indices for mastery tests. *J Educ Measurement* 1988;**25**:47–55.

21 Wainer H, Thissen D. How is reliability related to the quality of test scores? What is the effect of local dependence on reliability? *Educational Measurement: Issues Pract* 1996;**15** (1):22–9.

Received 8 January 2004; editorial comments to author 5 February 2004; accepted for publication 8 March 2004