British Journal of Orthodontics/Vol. 24/1997/139–147

The Design and Analysis of Reliability Studies for the Use of Epidemiological and Audit Indices in Orthodontics

C. T. ROBERTS, B.SC., M.SC., PH.D.

The National Primary Research and Development Centre, The University of Manchester, Williamson Building, Oxford Road, Manchester, M13 9PL, U.K.

S. RICHMOND, B.D.S., F.D.S., R.C.S.(EDIN). D.ORTH, R.C.S., M.SC.D., PH.D.

Department of Child Dental Health, Dental School, University of Wales College of Medicine, Heath Park, Cardiff CF4 4XY, U.K.

Introduction

Epidemiological surveys have long demanded standardized methods of measurement, generally referred to as indices, to facilitate comparison. More recently formalized clinical audit has extended the need for such methods to assess average treatment outcome within and between clinical units. Indices are also being introduced to assess more objectively the need for treatment to maximize benefit from limited resources.

In order to assess the provision of orthodontic care in term of need and outcome, the Index of Orthodontic Treatment Need (IOTN) (Shaw *et al.*, 1991) and Peer Assessment Rating (PAR) (Richmond *et al.*, 1992) have been developed. Widespread use of these indices requires the training of examiners to establish a uniform standard of application. Once calibrated to a uniform standard it is important that consistency is maintained by an examiner and amongst groups of examiners working together.

Epidemiological or clinical indices may be developed and applied in one setting. Before such measures are used in different circumstances, it may be appropriate to reassess the value of the measure. For example, a measure may be more successful in a clinical setting, but fail when used in a field situation such as an epidemiological study using children. An index developed in the mixed or deciduous dentition may be unsuited to studying adult patients.

This paper addresses the statistical issues involved in the testing of indices and training and calibration of observers for their use in epidemiological and clinical work. Two basic terms are used in this context:

1. **Observer reliability**, is the extent to which a measurement is repeatable under identical conditions. The term intra-rater reliability referring to consistency of repeated observation by an observer with himself whilst inter-rater reliability relates to observations being consistent amongst a group of observers.
2. **Observer validity**, is the extent to which a measurement measures what it purports. In a clinical or epidemiological context the measurement of validity takes place against a validity or 'gold' standard.

Measurement Error

Errors in measurement are generally classified as either systematic or random. If a particular measurement is persistently under or over rated by an observer, then a systematic error is introduced. Measurement can be represented as an equation called the measurement model. If X is the observed value of the index for a specific case due to a particular observer, then,

$$X = T + r + e$$

where t is the true value for that case, r is the systematic error or bias due to the particular observer, and e is the random error varying from case to case. It is often assumed that the random error e does not depend on the general average for that case, in statistical parlance, T and e are independent. However, this may not be the case as a measurement may have greater random variation for larger true values.

In a training course the validity of the trainee's measurement is often compared against the validity standard defined by the instructor. It is assumed that the instructor's criterion is entirely reliable or that the possibility error by the instructor is small compared that of the trainee. The instructor score corresponds to T in the measurement model. Where no single individual can be considered entirely infallible this can be achieved by obtaining a consensus amongst several trained observers to obtain definitive scores.

Effects of measurement error

The relative importance of random and systematic error and their implications depend on the type of study planned. For example, if different observers examine separate subgroups in a study, comparison between the subgroups would be problematical unless is was clear that any systematic bias between observers was small relative to the magnitude of difference of clinical interest.

The effect of random error on the outcome measure of a study is to reduce efficiency. Random errors will not affect the mean of a sample, but will usually increase the variance and, hence, the standard deviations. This will not discredit the result, but will make a statistically significant difference more difficult to achieve. The sample size requirements will

0301-228X/97/002000+00$02.00

be greater for a less reliable measure than one that is more accurate. If several observers are involved in taking measurements in a study, systematic bias between observers also increases variance and reduces efficiency in the same way.

Increased random error will result in a reduction in the correlation observed between the variables, a process called attenuation. Although this bias towards the null value is generally acceptable, in certain situations it has been shown that spurious effects may be introduced. For example in study designs requiring the comparison of correlation or regression coefficients such as multiple regression or analysis of covariance, random error may lead to a statistical analysis showing a difference which would not be present if a more accurate measurement was employed (Healy, 1989). Sometimes a treatment effect is assessed by correlating the difference between pre- and post-treatment values with the pretreatment value. When the pretreatment measurement contains a random error an apparent negative correlation can occur where none would be present if the random error was removed from the calculation (Bloqvist, 1974). It might therefore be falsely concluded that large pretreatment values are associated with a greater change.

Assessment of measurement error

The remainder of this paper sets out to describe statistical aspects of the analysis of reliability or validity studies. Whether one wishes to compare a trainee with an instructor (validity) or a group of observers (reliability), each participant should independently observe the same sample of cases. The choice of sample depends on the population with which the index is to be used. For example, when training examiners for an epidemiological study in schools, the sample should be drawn from the relevant school population. If the measure is to be used in a new patient clinic, the sample should be typical of these.

One may wish to show how reliable or valid the index is when used by a particular type of examiner. In this situation it is important that more than two observers are used in order that the variation that exists between all possible observers is represented in the study.

In carrying out a reliability or validity study, interest may focus on the abilities of individual observers. Alternatively, it may be the overall quality of the measurement in a particular situation. One may be concerned with the ability of a group of users of an index rather than of an individual or the comparison of a pair. These are interlinked as poor reliability may be explained by specific individuals but the emphasis is different. Questions that might be addressed by a reliability or validity study include:

1. How close is the trainee's score likely to be to the standard score?
2. What is the range of likely differences between the two observers?
3. How large is the systematic bias of the trainee relative to the standard?
4. How large is the systematic differences between observers?
5. How large is the effect of random variation on a future study?

In the following discourse the appropriate statistical methods will be described to address each of these questions.

The statistical methods used to assess reliability can be applied with slight modification to the training/validity situation. In the assessment of reliability it is generally assumed that both measurements contain the same random errors relative to the general average or consensus value, whereas with validity it is generally assumed that the standard is made without error. Although this may not in practice be true, all of the random and systematic errors are attributed to the trainee.

Measurement scales

The statistical methods used to assess reliability and validity depend on the type of measurement scale being used. The simplest form of measure is dichotomous in that each subject can be allocated to only one or two categories (e.g. male or female). Where a scale has more than two categories these may be referred to as nominal categorical, involving a set of unordered categories that are qualitatively different (e.g. Class I, Class II, and Class III malocclusions). Alternatively, it may be ordered with categories having a natural sequence. An example, being the components of IOTN where each category represents an increased degree of severity.

Alternatively, measurement might be on continuous or interval scales with the distances between two values in one region of the scale meaning the same as an equal distance in another part of the scale (e.g. age, weight, or height). Ordered categorical scales stand between nominal or unordered categorical scales, and interval measurement, in that, there is generally some notion of distance between categories. Some ordered categorical scales have precise definitions of each category that reflect qualitative difference and hence are similar to nominal categorical scales. An example of this is the Dental Health Component of IOTN. Others may reflect an underlying continuum or latent measure, an example of this being the Aesthetic Component (AC) of IOTN. These may be considered to have greater similarity to interval scale measurement.

The PAR index score is derived by adding together a set of ordered categorical subcomponents. These have been weighted using a validation standard (Richmond *et al.*, 1992) to provide a summary weighted PAR score. Although based on ordered categorical scales, values range from 0 to above 50. It is expedient, but also reasonable to consider it as being an interval scale measurement due to the weighting of components. The process of validation has made a reduction in malocclusion from 30 to 20 on the weighted PAR scale equivalent to a change of say 20–10.

Reliability of Interval Scale Measurement

To illustrate the methods, data from a calibration exercise for the PAR index will be used. This involved the comparison of five trainees against an expert and the comparison of the five trainees amongst themselves. It includes both elements of validity and reliability studies. An example of such data for the first two trainees is given in Table 1.

TABLE 1 *Raw data and calculations*

Case	Rater 1 [1]	Rater 2 [2]	Mean $a = ([1] + [2])/2$	Diff $d_i = [1] - [2]$
1	19	2	10·5	−17
2	3	3	33·0	0
3	18	13	15·5	−5
4	2	2	2·0	0
5	30	8	19·0	−22
6	3	2	2·5	−1
7	31	30	30·5	−1
8	6	2	4·0	−4
9	34	31	32·5	−3
10	7	2	4·5	−5
11	18	10	14·0	−8
12	4	2	3·0	−2
13	43	31	37·0	−12
14	1	3	2·0	2
15	36	35	35·5	−1
16	16	8	12·0	−8
17	19	23	21·0	4
18	24	18	21·0	−6
19	37	33	35·0	−4
20	5	6	5·5	1
21	40	36	38·0	−4
22	0	3	1·5	3
23	14	14	14·0	0
24	6	4	5·0	−2
25	25	20	22·5	−5
26	16	19	17·5	3
27	39	40	39·5	1
28	12	10	11·0	−2
29	28	23	25·5	−5
30	10	5	7·5	−5

Mean	$x_a =$	16·4	$x_d =$	−3·6
Variance	$s^2_a =$	161·1	$s^2_d =$	32·0
S.D.	$s_a =$	12·70	$s_d =$	5·66
Root mean square error				4·69
Coefficient of reproducibility				13·26
Confidence limits of x_d				(−5·7 to −1·5)
Limits of agreement				(−4·9 to 7·7)
Reliability coefficient				0·89

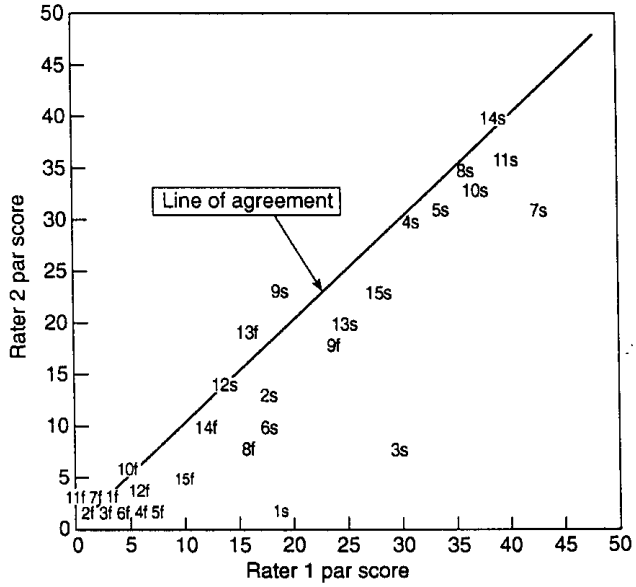

FIG. 1 Plot of pairs of PAR scores for each case of data in Table 1

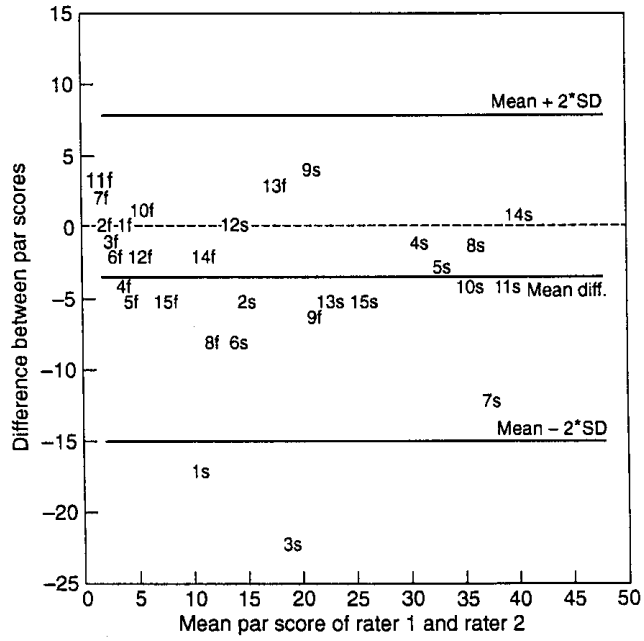

FIG. 2 Plot of difference between observers PAR scores against average of observers PAR scores.

Graphical presentation

In comparing the scores of two examiners (or a single examiner on two occasions) it is useful to look at the data graphically. A common practice is to construct a scatter-diagram by plotting the pairs of values for each case of patient (Fig. 1). If the two scores are in perfect agreement, then the values would be plotted on the dotted line of equality shown. It is insufficient that the two readings should be highly correlated, as correlation measures the strength of the linear relationship. Bland and Altman (1986) have noted that two scores being of perfect correlation could lie on a line with a positive gradient very different to the line of equality.

A better illustration suggested by Bland and Altman involves computing the difference and mean of the pairs of values being compared. In the example means a_i and differences d_i of the two observations have been computed (for each case). These are given in Table 1 together with their means ($\bar{x}_d$ and $\bar{x}_a$) and standard deviation (s_d and s_a). The difference for each case d_i is then plotted against the mean of the two raters a_i for each case (Fig. 2). The figure shows the magnitude of the difference between the two observers and will also reveal if the observers tend to differ more or less for cases of differing magnitude. The case numbers have been plotted so that any large disparity may be identified back to the specific case. To aid interpretation it is useful to add three horizontal lines $\bar{d}$, $\bar{d} + 2{,}s_d$ and $\bar{d} - 2.s_d$ to give the range of the difference between the two scores. The lines illustrates the systematic differences between observers. Assuming that the random errors have a normal distribution, 95 per cent of the points should lie between the lines $\bar{d} + 2.s_d$ and $\bar{d} - 2.s_d$. Bland and Altman refer to the two outer lines as the limits of agreement. In Fig. 2 it can be seen that the magnitude of the differences tends to

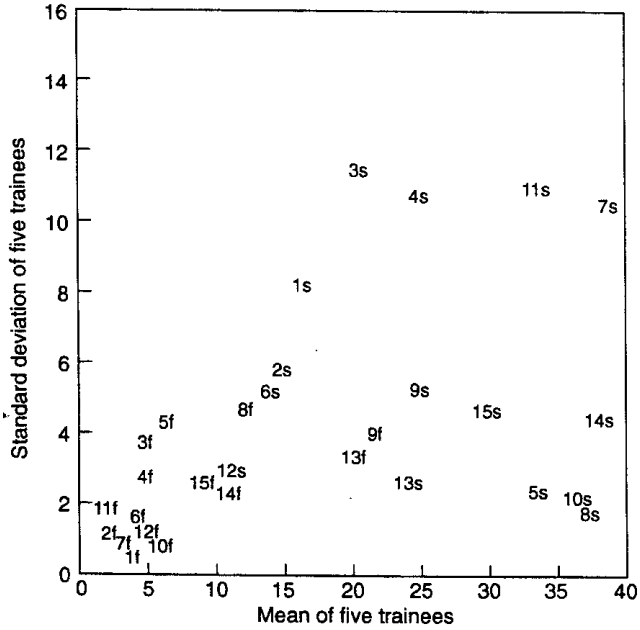

FIG. 3 Plot of standard deviation of 5 observers PAR scores case against the mean of 5 PAR scores for each case.

increase with the average of the score suggesting that the random error is not independent of the general average value. This is not surprising as the PAR index is the total of many subcomponents which can lead to errors due to omission.

Where a reliability study involves more than two observers, graphical representation can be provided by computing the standard deviation and the mean of the values for each case for all observers. The standard deviations are then plotted against the mean. Such a graph will illustrate the magnitude of random error and any change as the general average increases. It does not, however, give any impression of systematic bias between observers. This is illustrated for a sample of five trainee raters in Fig. 3. Again the increase in random error with magnitude is apparent.

When comparing a trainee with a standard or expert where it is assumed that only the trainee contributes to measurement error, the difference between scores should be plotted against a standard or expert score. This differs slightly from the plot used for reliability (Fig. 2) where both observations are considered to have measurement error relative to the general average value in the measurement model.

Summary statistic—systematic error

To assess how well on average the two observers agree, the mean of the differences $\bar{d}$ computed above gives a measure of any systematic error. This also applies when comparing a trainee with a standard.

As this mean difference $\bar{d}$ is calculated from a sample it will have sampling variation. The confidence interval needs to be computed to show the precision of $\bar{d}$. Provided the sample size is over 30, a close approximation to the limits of the 95 per cent confidence interval is given by $\bar{d} \pm 2.s_d/\sqrt{n}$ where n is the number of cases in the study, s_d is the standard deviation of the differences. In the example in Table 1 the values for the mean difference is −3·6 with 95 per cent confidence limits −5·7 to 1·5. The interpretation of this is that in any future series of similar cases the systematic error between these two observers will be between these two values with probability 0·95. It is important that the magnitude of limits is below what might constitute a clinically important difference. A probability of 0·95 may be too stringent in which case a confidence limit with say 0·9 or 0·8 probability might be used. In the example it can be seen that examiner 1 scores systematically more than examiner 2 confirming the impression of Figs 1 and 2.

TABLE 2 *Comparison of trainees with standard in PAR index calibration*

Trainee	Mean diff	Confidence interval	Paired *t*-test	RMS error	Limits agreement
1	−2·2	(−4·0 to −0·5)	0·0015	3·7	(−11·2 to 7·3)
2	−5·8	(−8·3 to 3·4)	0·0001	6·2	(−19·1 to 7·5)
3	−5·0	(−8·4 to −1·6)	0·007	7·4	(−23·1 to 13·7)
4	−1·1	(−3·7 to 1·5)	0·40	5·0	(−15·2 to 13·0)
5	−2·1	(−4·7 to 0·5)	0·11	5·2	(−16·3 to 12·1)

An alternative approach is to use a paired *t*-test to examine if the systematic bias is statistically significantly different from zero. The test statistic is computed as $T = \bar{d}.\sqrt{n}/s_d$. This is equivalent to examining whether the confidence interval does not include the zero difference. Such an approach is unfortunately flawed. Where random variation is greater the term s_d will be larger and the test statistic T smaller. This makes it less likely that a t-test will imply the bias is statistically significant. However, where the random variation is greater, the confidence limit will be wider suggesting the possibility of systematic bias of a greater magnitude. This illustrates one advantage of confidence limits compared to a significance test. This issue is illustrated in Table 2 which shows the comparison between a standard and five trainee raters. On the basis of the *t*-test it might be concluded that examiner 1 showed greater bias than examiner 5. Examination of the confidence limit suggests that the systematic bias of examiners 1 and 5 relative to the standard are little different. In a calibration study it is therefore suggested that systematic error is assessed by comparing confidence limits with a range based on clinical criteria.

Summary statistic—random error

To quantify random variation, a summary measure that is often recommended is the root mean square (RMS) error. This is given by the formula

$$\sqrt{\frac{\left(\sum_{i=1}^{n} d_i^2\right)}{2n}}$$

where d_i is the difference between the two raters. The RMS error is an estimate of the standard deviation representing the measurement error of a single measurement s_e provided there is no systematic bias. Where this is present

the RMS error will over-estimate the standard deviation by including both bias and random error.

A closely related quantity is the coefficient of reproducibility. It has been adopted by the British Standards Institute (1979) and is given by $2\sqrt{2} \times$ RMS error. In individual clinical measurement this has an interpretation when considering the difference between two measurements. In the absence of any underlying change 95 per cent of differences will generally be less than the coefficient of reproducibility. Hence, if two replicate readings are made on the same subject and the difference is less than the coefficient of reproducibility, then there is no evidence of change beyond what might be explained by measurement error. It is therefore desirable that the coefficient of reproducibility is less than what is deemed to be a clinically important difference if the measure is to be used for clinical decision making. Such a stringent requirement is not necessary for research measurement where we are concerned with samples rather than decision making on a single case. In our example RMS error is 4·69 which gives a coefficient of reproducibility of 13·26. This suggests that differences of less than 10 by these examiners could be explained by measurement error. Where more than two observers are involved the RMS error may be estimated using analysis of variance.

The above approach tends to obscure the systematic error which is often present in clinical measurement. An alternative approach is to consider the limits of agreement by $\bar{d} + 2.s_d$ and $\bar{d} - 2.s_d$. These were calculated above when constructing Fig. 2. Suppose the first rater observes a value X_1 then the second observer's value will be in the range from $X_1 + \bar{d} - 2.s_d$ up to $X_1 + \bar{d} + 2.s_d$ with probability 0.95. Likewise if the second rater observes a value X_2 then the value for the first observer should be in the range $X_2 - \bar{d} - 2.s_d$ up to $X_2 - \bar{d} + 2.s_d$. In the example if the first rater's value of PAR is X then the second will be in the range from $X - 12$ to $X + 7$. This approach can give meaningless negative values when applied to specific value if the data does not satisfy the assumption of being normally distributed.

The RMS error and coefficient of reproducibility provide a single figure for an overall assessment of a measure whilst the limits of agreement are more suited to comparison of specific pairs of raters. The latter method is better for addressing the comparison of a trainee with a standard.

The reliability coefficient

The methods described above assess measurement error in the units of the original measurement. Where variation between subject is comparatively large, random variation of particular magnitude will be less important than in a sample with small variation. For example, in a study of temperature measurement of human subjects the coefficient of reproducibility may be small in absolute terms, but the variation between subjects is also limited.

An alternative approach, applicable to research studies, is to relate the magnitude of the measurement error to the variability of the population being studied. Such a measure is the reliability coefficient (also confusingly called the intraclass correlation coefficient) that is defined as

$$R = 1 - \frac{S_e^2}{S_x^2}$$

where s^2x is the observed variation of the study sample and s_e^2 is the variation of the measurement error discussed above. Where there is no measurement error, that is s_e^2 equals zero, the coefficient of reliability equals 1 whilst if measurement error explains all the variation in the data, that is $s_e^2 = s^2x_1$ the reliability is zero. An estimate of the reliability coefficient is given by the formula:

$$R = \frac{2.s_a^2 - \frac{1}{2}s_d^2}{2.s_a^2 + (\text{RMSE})^2}$$

where s_a^2 is the variance of the average of the two value for each case, s_d^2 is the variance of the differences and RMSE is the root mean square error. Using this formula the estimate of R in our example is 0·89.

All analyses and graphics described above involving two observers can be easily obtained using any of the popular spread sheet software programs for IBM PC compatible or Apple Macintosh computers.

The interpretation of the reliability coefficient

The reliability coefficient has several useful properties in the assessment of some of the effects of random variation discussed above. The sample size required in a study with outcome measures having reliability R is increased by a factor $1/R$ compared to a study with an entirely reliable measure. The attenuation of a correlation coefficient can also be estimated from the reliability. For correlation r between two measurements X and Y, say each with reliability R_x and R_y the observed correlation will be under-estimated by a factor $\sqrt{R_x}.\sqrt{R_y}$ compared to the correlation between X and Y if there was no measurement error. These properties are useful as they give an estimate of the gain that may be achieved by improving the accuracy of measurement. For example, if limited numbers of patients are available for inclusion in a research project the reliability coefficient gives an estimate of the possible gain due to investment in improved measurement. It might be concluded that the measure is sufficiently reliable and therefore there is little benefit in attempting to make improvements. Alternatively, the decision may be made to use replicate measures of each case to improve reliability and save the expense of recruiting additional patients into the study.

It is apparent that the effect of random error depends on the context. As a consequence it is difficult to generalise as to what constitutes an acceptable level of reliability. Fleiss (1986) tentatively suggests that values of R below 0·4 or so might constitute poor reliability, between 0·4 and 0·75 fair to good, whilst above 0·75 represent excellent. These guidelines were suggested for use with the lower 95 per cent confidence limit of the reliability coefficient in order that account could be taken of the sampling variability. Calculation of this confidence limit for the reliability coefficient is beyond the scope of this article. Interested readers are referred to Fleiss (1986) who gives a comprehensive description for intra-observer and inter-observer reliability studies.

For more than two observers estimation of the reliability coefficient is generally based on an analysis of variance. Details of computation of the reliability coefficient are not given here and interested readers are referred to Fleiss (1986). For our example the estimate of the reliability coefficient for all five trainees obtained from analysis of variance is 0·89 with lower one-sided confidence limit of 0·77. This suggests that measurement error will cause only a slight loss of efficiency in a research study.

The reliability coefficient is a useful indication of the information contained in a measure. Whilst it has been suggested for making comparisons of the quality of pairs of raters (Kingman, 1986), calculation of limits of agreement is simpler and more informative, as systematic bias is also taken into consideration.

Reliability of measurement of change

In the assessment of treatment, interest sometimes focuses on the change of an index from start to finish. For example, it has been suggested that reduction and percentage reduction in PAR score provides a useful measures of outcome. It follows that reliability should be estimated for these measures particularly as the underlying variation in the sample of a measurement of chance tends to be much less.

Reliability of Categorical Measurement

The assessment of reliability for categorical measurement has many parallels with that for interval data. In Table 3, data from a reliability exercise for two examiners using the Dental Health Component of IOTN, a 5-point ordered categorical scale, is listed. To compare either a rater with the standard or a pair of raters a two-way table should be drawn up summarising the pairs of values (see Table 4). It is also worth calculating the proportion or percentage in each cell (these are given in brackets). If both observers were in perfect agreement all the frequencies would be on the main diagonal of category agreement that runs from top left of the table to bottom right. Examination of the table may also reveal categories that are being confused by observers for example in Table 4 the observers would appear to confuse categories 3 and 4. Identifying the specific cases involved may eliminate differing interpretation by observers and, hence, improve the category definition for future use.

TABLE 3 *Reliability data for two examiners using Dental Health Components of IOTN*

Case	Rater 1	Rater 2
1	4	4
2	2	2
3	4	3
4	2	2
5	4	5
6	3	2
7	3	4
8	2	2
9	4	4
10	2	2
11	4	4
12	4	2
13	4	2
14	2	1
15	4	4
16	4	3
17	4	4
18	4	4
19	5	5
20	2	2
21	4	4
22	2	1
23	4	5
24	2	1
25	4	3
26	3	3
27	5	5
28	3	2
29	5	5
30	2	2

TABLE 4 *Two-way table summarizing reliability data from Table 3*

		Rater 1 DHC Score					
		1	2	3	4	5	Row total
Rater 2 D HC Score	1	—	3 (0·10)	—	—	—	3 (0·10)
	2	—	6 (0·30)	2 (0·07)	2 (0·07)	—	10 (0·33)
	3	—	—	1 (0·30)	3 (0·10)	—	4 (0·13)
	4	—	—	1 (0·03)	7 (0·23)	0 (0·0)	8 (0·27)
	5	—	—	—	2 (0·7)	3 (0·10)	5 (0·17)
	Column Total	—	9 (0·3)	4 (0·13)	14 (0·47)	3 (0·10)	20

Proportions given in brackets

Unweighted kappa	0·42
Lower 95 per cent confidence limits	0·23
Weighted kappa with linear weights	0·62
Lower 95 per cent confidence limits	0·48

Assessment of bias

As well as examining the cell frequencies it is also worth comparing the column and row totals which give the frequency with which each examiner chose a specific category. If one rater uses a specific category more often than another, this is an example of systematic bias for a categorical variable. For the dichotomous case McNemar's test for comparing proportions for matched pairs (see Siegel and Castellan, 1988) may be applied to assess bias statistically. Formal statistical testing of this for nominal scale data is not straightforward. For ordinal scale measurement the Wilcoxon matched pairs test available in many statistical software packages can be used to test if the median difference in rater's scores differs from zero. In our example the Wilcoxon test gives $P = 0{\cdot}055$. Whilst this is not statistically significant at a 0·05 significance level, it is a basis for concern regarding bias. Examination of Table 4 reveals that on 10 occasions examiner 1 scored greater than examiner 2, but only on three occasions does the reverse occur.

As with interval scale measurement it could be argued

that a confidence interval for the median difference between raters should be used in place of a statistical test. Although a confidence interval can be calculated, this data is unlikely to be helpful as only whole or half units are available. Fortunately, it is difficult to conceive examples where a systematic difference might be statistically significant, but not clinical important unless the sample size is large. Conversely, a result having clinically important bias, but not being statistically significant might occur, but only where agreement between the two raters is poor.

Assessment of agreement

To assess reliability, the choice of summary statistic is often the percentage or proportion of agreement calculated by considering the percentage of frequencies on the main diagonal of the table. Some agreement between observers would occur by chance and this is substantial where one category has high prevalence for both observers. The agreement that is obtained by chance for each category is obtained by the formula.

$$\frac{\text{Row total} \bullet \text{Column total}}{\text{Grand total}}$$

in the same way as expected frequencies are calculated in the Chi-squared statistic. In the example in Table 5 the observed frequencies in each cell are no larger than what would be expected by chance. Although percentage agreement is high (75 per cent), the measurement has little value as agreement could be due to chance.

TABLE 5 *An example of high percentage agreement due to chance*

		Rater 1		Total
Rater 2		X+	X−	
	X+	10	50	60
	X−	50	250	300
	Total	60	300	360
Percentage agreement = 75%		Kappa = 0		

The Kappa statistic

Cohen (1960), devised the Kappa statistic to deal with this situation by subtracting the chance expected agreement from the observed agreement and then rescaling. Kappa is defined as

$$\kappa = \frac{p_0 - p_e}{1 - p_e}$$

where p_o = observed proportion of agreement and p_e = chance expected proportion of agreement. It has useful properties being scaled so that complete agreement between the two observers gives Kappa equal to one. Alternatively, where observed agreement is no better than that obtainable by chance, Kappa equals zero.

A more detailed description is given in the appendix for mathematically minded readers. A program that will run on an IBM compatible machine is available from the authors for computing Kappa and the related weighted Kappa statistic described below.

Weighted kappa

A scale which has ordered categories implies that disagreement between different pairs of categories have different seriousness depending on their positions in the sequence. Cohen (1968) proposed a modification of Kappa in which weights were assigned according to the degree of the agreement. On the main diagonal a weight of 1 is given and at extreme values a weight of 0. Intermediate values are given partial credit with a value between 0 and 1. Weighted kappa is defined by

$$\kappa_w = \frac{p_0(w) - p_e(w)}{1 - p_e(w)}$$

where $p_o(W)$ is the proportion of 'weighted agreement' observed and $p_e(W)$ the proportion of weighted agreement expected. As with kappa, a weighted kappa of 1 corresponds to perfect agreement, whilst a value of zero implies that observed agreement is no better than that obtained by chance.

Ideally, the choice of weights should be made prior to the study, based on the clinical or scientific implications of different forms of disagreement, although in practice it is difficult to obtain a consensus expressed in quantitative terms. Because different choices of weights give rise to different values of weighted kappa, tailoring weights to a specific situation may be perceived as deception as weights could be chosen to minimize disagreement. Therefore, it is generally advisable to use one or two standard weighting schemes referred to as quadratic and linear weights, unless there are strong reasons for doing otherwise. Details are given in the appendix. Both assume the distances between adjacent categories to be equal. Quadratic weights, first proposed by Cohen (1968), have been shown by Fleiss and Cohen (1973), to make the kappa statistic approximately equal to the reliability coefficient for large samples. This would suggest that these quadratic weights should be used where a scale has large numbers of categories or that it is realistic to think in terms of an underlying continuum as say with the AC component of IOTN. Linear weights proposed by Cicchetti and Allison (1971), would appear to be more suitable where numbers of categories are small and tend to differ qualitatively. An example of linear weights is illustrated for a five point scale such as DHC of IOTN in Table 6. The value of kappa obtained with these weights for the data in Table 3 is 0·62 whereas the unweighted kappa is 0·42. Whichever choice of weight is

TABLE 6 *Linear weights for weighted kappa for a 5-point scale*

		Rater 1				
		1	2	3	4	5
Rater 2	1	1	0·75	0·5	0·25	0
	2	0·75	1	0·75	0·5	0·25
	3	0·5	0·75	1	0·75	0·5
	4	0·25	0·5	0·75	1	0·75
	5	0	0·25	0·5	0·75	1

made, this should be made clear in any research report. It is important to note that unweighted, linear weighted and quadratic weighted kappa are not comparable. The value of unweighted kappa is less than that of weighted kappa with linear weights which is in turn less than that with quadratic weights.

Interpretation of kappa and weighted kappa

As with a reliability coefficient it is advisable to compute confidence limits. For kappa the concern is generally that a value of kappa is above a specific level and, hence, a lower 95 per cent confidence limit should be computed. This is available from the software described using the procedure described by Fleiss (1982). For our example, lower 95 per cent confidence limit is 0·48 showing that kappa is above this value with a probability of 0·95.

Whilst a value of kappa equal to 1 or zero has an easy interpretation, the intermediate values are more difficult. For nominal scale measurement it has been suggested (Landis and Koch, 1977) that a kappa of over 0·8 indicates good agreement, over 0·6 substantial, over 0·4 moderate, and that above 0·2 fair and below 0·2 poor. It is suggested that these criteria be applied to the lower 95 per cent confidence limit so in the example above the agreement can be considered to be moderate. As with those suggested for reliability these guidelines should not be used too rigidly, being based largely on experience rather than any precise justification. Where an index is to be used to make decisions regarding patient treatment a much higher level of reliability is likely to be required than for an index used for research or audit on groups of patients. They have also been suggested for use with weighted kappa (Fleiss, 1981) although here they may seem more arbitrary as the change of weighting scheme might move the value of weighted kappa into a different band.

Although interpretation of a single kappa value has problems, it can be used effectively in a comparative way. If more than two observers are involved in a reliability study, comparisons may be made using the kappa statistic for pairs of raters to identify groups of homogeneous or aberrant observers. Alternatively kappa may be used to assess intra-observer reliability to compare the consistency of each observer. This requires each observer to assess the same or equivalent samples.

As with reliability coefficients the value of kappa obtained relates to a particular population being adjusted for the underlying variability of the sample. For example, the value of kappa for a measure obtained when considering a clinical sample may be very different to that from an epidemiological survey. Although this may be in part due to the differing circumstances in which measurement is made, differing frequency distributions in either sample will also affect the value. With frequency distributions that differ greatly, values of kappa are not directly comparable.

Validity and categorical data

Although kappa is a measure of reliability, it can be used as a measure of agreement in validity studies. For dichotomous data validity is generally assessed using the statistics sensitivity and specificity (Nuttal and Davies, 1988). These terms are generally described in the context of a diagnostic test. Sensitivity is the proportion of positive diagnosis under the standard or definitive value that are detected by the trainee as being positive. Specificity is the proportion of the negative diagnosis under the standard that is detected by the trainee as being negative.

The ideas of sensitivity and specificity can be expanded to a scale with more than two values, although large number of values are difficult to interpret. To assess the validity of a trainee, applying the index in treatment decision making, categories might be grouped together. For example, for the DHC component of IOTN the values of 4 and 5 are considered to represent treatment need, whilst 1, 2 and 3 suggest no or only slight need (Richmond *et al.* 1994). The sensitivity and specificity for an individual operator could then be determined. For example, if the values of the first rater in Table 4 is the standard and the second the trainee, then sensitivity = 12/17 = 71 per cent, whilst specificity = 12/13 = 92 per cent. Collapsing the data in this way can provide a useful interpretation for decision making, but is no substitute to collecting and analysing ungrouped data. Dichotomizing the scale after the data are collected is in general felt to give a more reliable dichotomous measure, than permitting the observer to use two broad categories. For epidemiological work using ordered categorical scales it is better to encourage use of a complete scale as this provides more information for statistical analysis.

Conclusions

The adoption of standardized methods of measurement or indices is suggested as a means to provide comparable data in different epidemiological, clinical or audit studies. It must, however, be recognized that, even with well defined guidelines, examiners can be unreliable. Careful training and calibration provides no guarantee that results will be comparable due to differences in experience, personal biases regarding severity or individual aptitude. For this reason research studies should be designed to minimize the adverse effects of measurement error.

In general, if several observers are to be involved in a study it is advisable to avoid situations in which different observers assess different subgroups that are to be compared. For example, if two different observers examine different treatment groups, the suspicion that group difference reflect observer bias is likely to remain, making the study unconvincing, however good the calibration of the observers. Similar proportions of each subgroup should be examined by each examiner obtaining balance between examiners and study variables. By doing this the possibility that any examiner bias is confounded with group differences is removed. This might be obtained by random or systematic allocation of cases to observers. Unfortunately, situations can be envisaged when this is difficult to achieve, an example being studies comparing different regions or disparate health districts and clinical units. Similarly, where a single examiner observes all cases they should strive, where possible, for a balanced approach by avoiding examination of all cases of one treatment group prior to another otherwise any drift in measurement practice may be confounded with study variables. Where

this is not possible intra-rater repeatability exercises should be carried out at regular intervals to check and curtail any systematic drift over time.

References

Bland, J. M. Altman D. G. (1986)
Statistical methods for assessing agreement between two methods of clinical measurement,
Lancet, 307–310.

Bloqvist, N. (1974)
On the relation between change and initial value,
Journal American Statistical Association, **72,** 746–749.

British Standard Institute (1979)
Precision of test methods 1: Guide for the determination of repeatability and reproducibility for a standard method by inter-laboratory tests. BS 5497 part 1.
BSI, London.

Cicchetti, D. V. And Allison , T. (1971)
A new procedure for assessing reliability of scoring EEG sleep recording,
American Journal of EEG Technology, **11,** 101–109.

Cohen, J. (1960)
A coefficient of agreement for nominal scales,
Educational and Psychological Measurement, **20,** 37–46.

Cohen, J. (1988)
Nominal scale agreement with provision for rescaled disagreement and partial credit,
Psychological Bulletin, **70,** 213–220.

Fleiss, J. L. (1981)
Statistical Methods for Rates and Proportions,
Wiley, New York.

Fleiss, J. L. (1986)
Design and Analysis of Clinical Experiment,
Wiley, New York.

Fleiss, J. L. and Cohen, J. (1973)
The equivalence of weighted kappa and the intra-class correlation coefficient as measures of reliability,
Educational and Psychological Measurement, **33,** 613–619.

Healy, M. J. R. (1989)
Measuring error,
Statistics in Medicine, **8,** 893–906.

Kingman, A. (1986)
A procedure for evaluating reliability of a gingivitis index,
Journal of Clinical Periodontology, **13,** 385–391.

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

Nuttal, V. M. and Davies, J. A. (1988)
The capability of the 1983 Children's Dental Health Survey in Scotland to predict fillings and extractions subsequently undertaken,
Community Dental Health, **5,** 355–363.

Richmond, S. *et al. (1992)*
The development of the PAR Index (Peer Assessment Rating): reliability and validity,
European Journal of Orthodontics, **14,** 125–139.

Richmond, S., Shaw, W. G., O'Brien, K. D., Stephens, C. D., Roberts, C. T. and Andrews, M. (1994)
Reliability of Professional judgements of orthodontic treatment need,
British Dental Journal (in press).

Siegel, S. and Castellan, N. J. (1988)
Nonparametric Statistics for the Behavioral Sciences,
McGraw Hill, New York.

Shaw, W. C., *et al. (1991)*
Quality control in orthodontic indices of treatment need and treatment standards,
British Dental Journal, **170,** 107–112.

Appendix: kappa and weighted kappa

Using the notation in Table 7 kappa for dichotomous data is

$$\kappa = \frac{(p_{11} + p_{22}) - (p_{1+}p_{+1} + p_{2+}p_{+2})}{1 - (p_{1+}p_{+1}\,1\,p_{2+}p_{+2})}$$

TABLE 7 *Algebraic notation for two table as used in the appendix*

		Rater 1						
		1	2	..	j	..	*n*	Row total
Rater 2	1	p_{11}	p_{21}	..	p_{j}	..	p_{n1}	p_{+1}
	2	p_{12}	p_{22}	..	p_{j2}	..	p_{n2}	p_{+2}
	:	:	:	:	:	:		
	j	p_{1j}	p_{2j}	..	p_{jj}	..	p_{nj}	p_{+j}
	:	:	:	:	:	:	:	:
	n	p_{1n}	p_{2n}	..	p_{jn}	..	p_{nn}	p_{+n}
Column total		p_{1+}	p_{2+}		p_{j+}		p_{n+}	1

For nominal categorical data a kappa coefficient can be defined as a generalization from the dichotomous case

$$\kappa = \frac{\sum_{i=1}^{n} p_{ii} - \sum_{i=1}^{n} p_{i+}p_{+i}}{1 - \sum_{i=1}^{n} p_{i+}p_{+i}}$$

For weighted kappa the weights w_{ij}, for each pair of categories i and j, are restricted to the interval $0 \leqslant w_{ij} \leqslant 1$ with w_{ij} 5 w_{ji}. The observed proportion of agreement is then

$$p_0(w) = \sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}p_{ij}$$

$$p_e(w) = \sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}p_{i+}p_{+}$$

where p_{ij} are the observed cell proportions and p_{i+} and p_{+j} are the observed marginal proportions. Weighted kappa is given by

$$\kappa_w = \frac{p_0(w) - p_e(w)}{1 - p_e(w)}$$

If $w_{ij} = 0$ for $i \neq j$, then k_w becomes equal to unweighted kappa as given by above. Two standard weighting schemes for weighted kappa that assume that the sequence of categories equally spaced points on an integer scale are defined by the following formula. For each pair of categories i, j of a scale with n categories

$$w_{ij} = 1 - \frac{(i-j)^2}{(n-1)^2} \quad \text{—quadratic weights}$$

$$w_{ij} = 1 - \frac{|i-j|}{n-1} \quad \text{—linear weights.}$$