European Journal of Orthodontics 27 (2005) 286–291
doi:10.1093/ejo/cjh078

An update on the analysis of agreement for orthodontic indices

Rebecca Brown* and Stephen Richmond*
*Department of Dental Health and Biological Sciences, Cardiff University Dental School, Cardiff, UK

SUMMARY The training of clinicians in the correct use of commonly used orthodontic indices involves calibration. The level of agreement between the trainee and a standard is assessed both as a measure of reproducibility and the success of training programmes. For the Peer Assessment Rating (PAR) index and the Index of Complexity, Outcome and Need (ICON), the recommended level of acceptable inter-rater agreement is no more than ±12 and ±18, respectively. Many commonly used methods of analysing this type of agreement are inappropriate. The method used in this investigation allows the calculation of limits of agreement, which easily demonstrate any major departures in agreement between trainee scores and standard scores. The basic method assumes that the differences between trainee and standard scores are normally distributed and that there is no relationship between these differences and the magnitude of the index. An extension to this approach is required when the assumptions of the basic method are not upheld. This extension provides a regression-based approach to calculating limits of agreement.

The results of this study demonstrate that the assumptions of the basic approach need to be checked for each comparison of trainee versus standard. In addition, regression-based methods are a more accurate means of calculating limits of agreement when these assumptions are not upheld. They also provide more information about bias and the range of disagreement between raters.

Introduction

Epidemiological and clinical orthodontic indices were developed in order to standardize the assessment of orthodontic care (Richmond *et al.*, 1993). These indices have widespread use and considerable numbers of clinicians use indices in daily practice. It is important to ensure that the clinicians who use the indices are appropriately trained and calibrated to facilitate comparisons between studies.

Previous investigations have examined the development and application of two such indices, namely the Peer Assessment Rating (PAR) index (Richmond *et al.*, 1992) and the Index of Complexity, Outcome and Need (ICON) (Daniels and Richmond, 2000). For PAR and ICON, the recommended level of acceptable inter-rater agreement is no more than ±12 and ±18, respectively. These levels are set at perceived levels of clinical significance. The statistical issues involved in the testing of these types of indices have been well documented (Roberts and Richmond, 1997). However, further analysis may be required when the assumptions of the basic approach are not upheld.

Bland and Altman (1986) outlined a method for measuring agreement that gives limits of agreement for continuous measures. Their more recent paper (Bland and Altman, 1999) outlined an extension to the basic approach where data exhibit a relationship between difference and magnitude. This may apply to the analysis of agreement in orthodontic indices and this extension has been applied in the present study. In this way, a more accurate assessment of agreement may be carried out.

Methods

A calibration study was undertaken to compare the scores of trainees with the standard score for PAR and ICON. Thirty models were arranged in sequence in a circle. The trainees were positioned equidistant with a gap of three apart and the cases were scored in turn in a clockwise direction. The trainees were chosen to illustrate the statistical methods; their age, gender, and experience were not recorded.

Both PAR and ICON scores are weighted summary measures on an interval scale, which range from 0 to 60 and 0 to 120, respectively. In order to illustrate the statistical methods outlined in this study, the ICON scores from one trainee and the PAR scores from a second trainee were used.

In method comparison studies, the measurements made by one observer are compared with the measurements from a particular standard. Often the true values can be extremely difficult to measure, so they remain unknown and even the best measure or 'gold standard' is rarely without error. Thus, some lack of agreement is inevitable. The following methods were used to attempt to quantify the degree of disagreement between scorers.

Statistical analysis

Regression analysis was used to investigate the relationships between differences and magnitude. The method of Bland and Altman (1999) for measuring agreement was followed and regression-based limits of agreement calculated. These limits were compared with the constant limits previously calculated for the data (Roberts and Richmond, 1997).

Results

Table 1 gives the results of the calibration study for ICON and PAR scores for two trainees (Rater 1 and Rater 2). The mean (standard deviation; SD) for the ICON standard score was 61.3 (28.6) and for Rater 1 57.6 (22.6). For the standard PAR and Rater 2, the mean (SD) scores were 25.6 (15.0) and 23.4 (13.0), respectively.

In order to investigate agreement between pairs of ICON and PAR scores, the rater scores were plotted against the standard scores. Figure 1a, b shows the line of exact agreement superimposed on both scatter plots. Deviation away from this line is evident and occurs towards the upper end of the scales. A Bland–Altman plot of difference between the scores versus the average of the scores makes these deviations clearer. Figure 2a shows this plot for the ICON scores and Figure 2b for the PAR scores.

The differences should not be plotted against either measure separately even when one is a standard. This is because the difference will be related to each individual measure, a well-known statistical phenomenon (Bland and Altman, 1995). Figure 2a, b demonstrates any consistent bias in the scores (away from zero difference) which can be adjusted for if appropriate. The figure also shows horizontal lines for the mean difference and 95 per cent limits of agreement. These limits are calculated as the mean difference ± 2SD of the differences. Provided differences within these observed limits of agreement are acceptable clinically, there is agreement between the rater and standard scores. Satisfactory agreement and how far apart scores can be without leading to problems is a question of clinical judgement, which should be defined in advance.

Table 1 Calibration study results.

Case	Standard ICON Score	Rater 1 ICON Score	Standard PAR Score	Rater 2 PAR Score
1	79	81	37	31
2	9	26	5	4
3	36	63	23	25
4	21	27	5	11
5	52	61	18	16
6	58	70	24	14
7	61	61	31	29
8	9	20	3	1
9	83	72	36	26
10	27	20	2	2
11	92	62	37	36
12	87	87	38	41
13	73	80	38	42
14	31	25	4	5
15	57	52	16	22
16	67	53	17	15
17	74	56	23	22
18	61	51	33	24
19	92	83	37	39
20	10	24	5	5
21	71	43	42	34
22	87	54	34	36
23	75	75	35	30
24	113	89	45	30
25	93	77	22	18
26	62	66	18	17
27	83	73	42	36
28	21	20	5	9
29	90	89	37	39
30	64	68	55	42

ICON, Index of Complexity, Outcome and Need; PAR, Peer Assessment Rating Index.

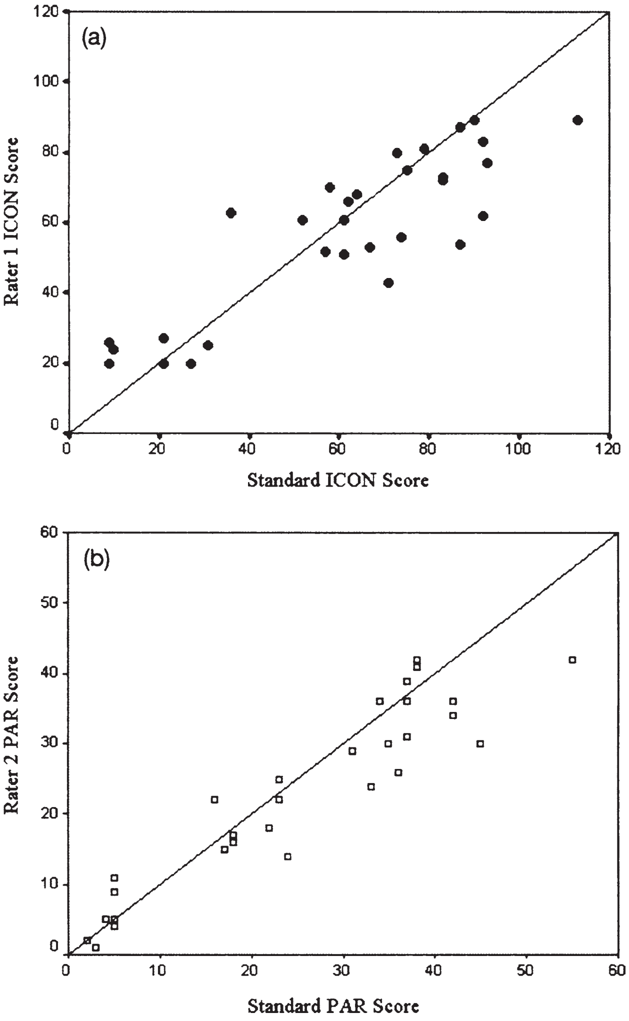

Figure 1 Scatter plots of standard versus trainee for (a) Index of Complexity, Outcome and Need (ICON) and (b) Peer Assessment Rating (PAR) scores.

Ninety-five per cent of the differences should lie between these limits if the errors are normally distributed. In order to check the distribution of the errors, histograms of the differences between standard and rater scores were examined. Figure 3a shows that the ICON errors were normally distributed, whereas the PAR score errors were skewed

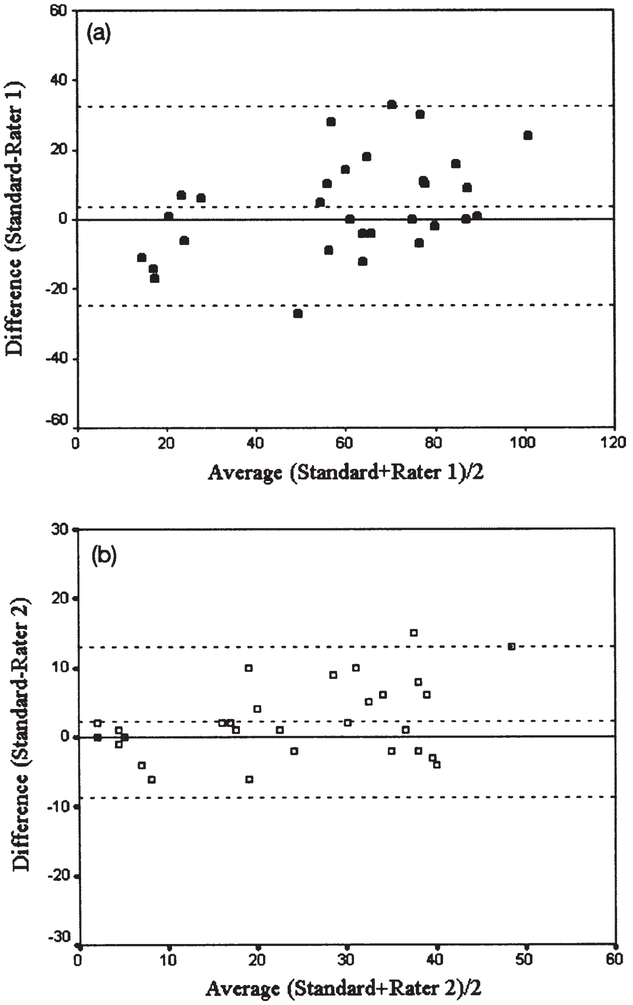

Figure 2 Bland–Altman plot of difference versus average of standard and trainee for (a) Index of Complexity, Outcome and Need (ICON) and (b) Peer Assessment Rating (PAR) scores.

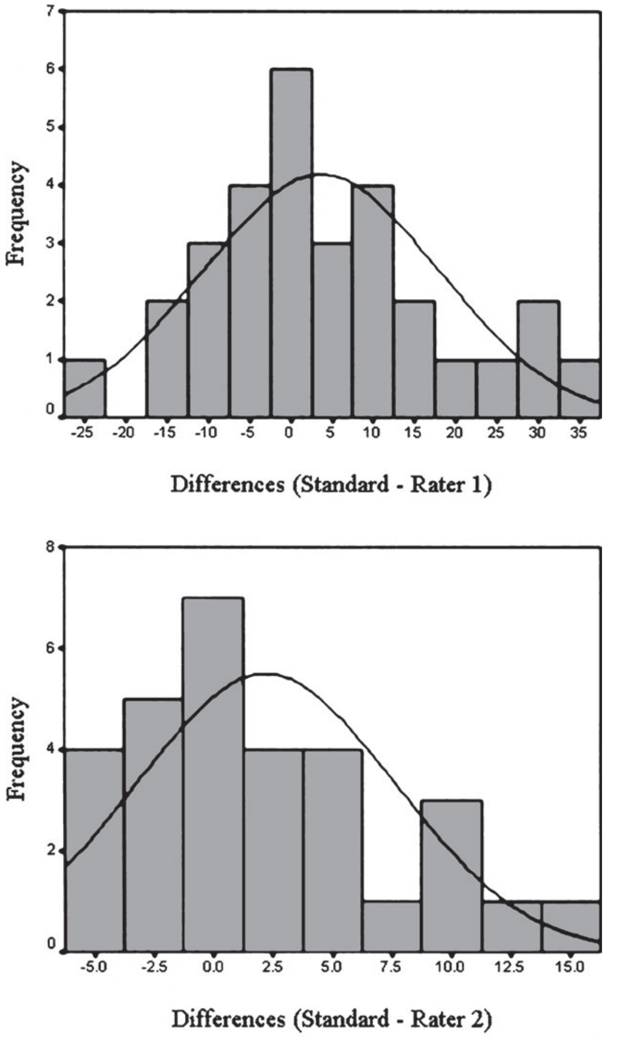

Figure 3 Histogram of differences between standard and trainee scores for (a) Index of Complexity, Outcome and Need (ICON) showing a normal distribution and (b) Peer Assessment Rating (PAR) scores showing a skewed distribution.

(Figure 3b). This is the first assumption of the method, which must be satisfied before the constant limits can be relied upon. The second assumption of the method is that there is no dependence of the differences on the average and this should also be checked before relying on the constant limits of agreement. It can be seen in the plot of differences versus the average that there is a spreading out of the differences with increasing magnitude in average ICON score (Figure 2a). This means there is an increase in variability of the differences for larger scores. This relationship can be ignored and the constant 95 per cent limits of agreement used. However, they will be wider apart than necessary for low ICON scores and narrower than they should be for larger ICON scores.

Regression analysis was used to assess the relationship between the differences and the average scores. The results suggest that there was a significant dependence of the differences on average ICON scores ($P = 0.015$). Spearman's rank correlation also clearly demonstrated this relationship (r = 0.441). In this case, Bland and Altman (1999) recommended removing this relationship either by transformation, or if that failed, the use of regression-based limits of agreement. Logarithmic transformation of both measurements before analysis will usually enable the standard approach to be used. Limits of agreement derived from log-transformed data can be back transformed to give limits for the ratio of the actual measurements (Bland and Altman, 1986). The ICON score data were log transformed and the analysis repeated. However, this failed to remove the relationship. This is because the differences tended to be in one direction for low values of ICON score and in the other direction for higher ICON scores (Figure 1a).

A better approach to deal with such data is the use of regression-based limits. Here the variability of the SD of the differences is modelled directly as a function of the level of measurement. The method uses absolute residuals from a fitted regression line and is based on the approach used to derive age-related reference intervals (Altman, 1993; Playle *et al.*, 1998).

The variation around the line of best fit from the regression of the differences on the average is modelled by an examination of the regression residuals. These residuals (saved quite easily in most statistical software packages) typically have a normal distribution and their absolute values are regressed on the average ICON scores.

The regression equation of the differences in ICON score ($\hat{D}_I$) on the average ICON score (A_I) was

Equation 1 $\hat{D}_I = -11.423 + 0.254A_I$ ($P = 0.015$)

The regression equation of the absolute values of the regression residuals ($\hat{R}_I$) on the average ICON score (A_I) was

Equation 2 $\hat{R}_I = 8.811 + 0.027A_I$ ($P = 0.627$)

The absolute residuals follow a half normal distribution. The SD of these residuals ($SD_{ABSRESID}$) is obtained by multiplying the fitted values of the regression of the absolute residuals on the average score by $\sqrt{(\pi/2)}$. If the regression of the absolute regression residuals against the average score is significant then the regression-based limits of agreement are obtained by combining the two regression equations (Altman, 1993). If the regression of the absolute values of the regression residuals on the average ICON score is not significant, the SD of the absolute residuals is estimated by multiplying the mean of the absolute residuals by $\sqrt{(\pi/2)}$. So, to obtain regression-based limits of agreement the following was used:

Equation 3 $\hat{D}_I \pm 1.96 \times SD_{ABSRESID}$

where $SD_{ABSRESID} = \sqrt{(\pi/2)} \times$ mean absolute residuals.

Upper 95 per cent limit of agreement =
$-11.423 + 1.96(10.415 \times \sqrt{(\pi/2)}) + 0.254A_I$
Lower 95 per cent limit of agreement =
$-11.423 - 1.96(10.415 \times \sqrt{(\pi/2)}) + 0.254A_I$

These regression-based limits are illustrated in Figure 4a.
As the differences in PAR score data were not normally distributed, the data were transformed using natural logs and re-examined. The relationship between the differences and the average still remained after transformation. Therefore, regression-based limits were also calculated for the PAR score data.
The regression equation of the differences in PAR score ($\hat{D}_P$) on the average PAR score (A_P) was

Equation 1 $\hat{D}_P = -1.303 + 0.143A_P$ ($P = 0.049$)

The regression equation of the absolute values of the regression residuals ($\hat{R}_P$) on the average ICON score (AP) was

Equation 2 $\hat{R}_P = 1.39 + 0.107A_P$($P = 0.006$)

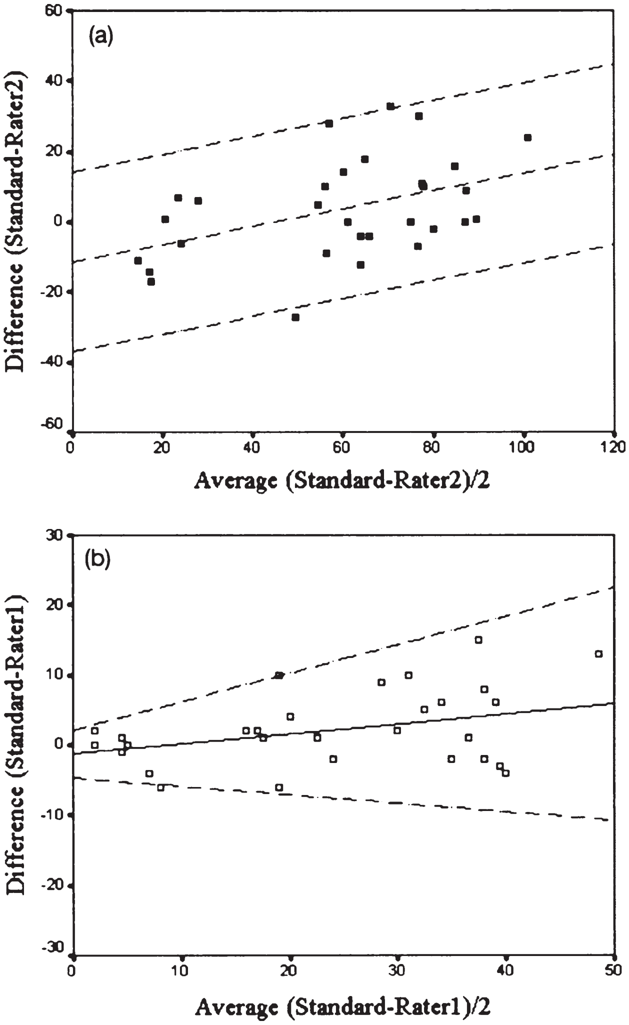

Figure 4 Regression-based 95 per cent limits of agreement for (a) Index of Complexity, Outcome and Need (ICON) and (b) Peer Assessment Rating (PAR) scores.

As in this case equation 2 was also significant, the equations were combined to obtain regression-based limits of agreement

Equation 3 $\hat{R}_P \pm 1.96 \times SD_{ABSRESID}$

where $SD_{ABSRESID} = \sqrt{(\pi/2)} \times \hat{R}_P$

The PAR score data $SD_{ABSRESID}$ was estimated by multiplying the fitted values of the regression in equation 2 by $\sqrt{(\pi/2)}$

Upper 95 per cent limit = $-1.303 + 1.96((1.39+0.107A_P) \times \sqrt{(\pi/2)}) + 0.143A_P$
$= 2.112 + 0.407A_P$
Lower 95 per cent limit = $-11.423 - 1.96(10.415 \times \sqrt{(\pi/2)}) + 0.254A_P$
$= -4.718 - 0.121A_P$

These regression-based limits are illustrated in Figure 4b.

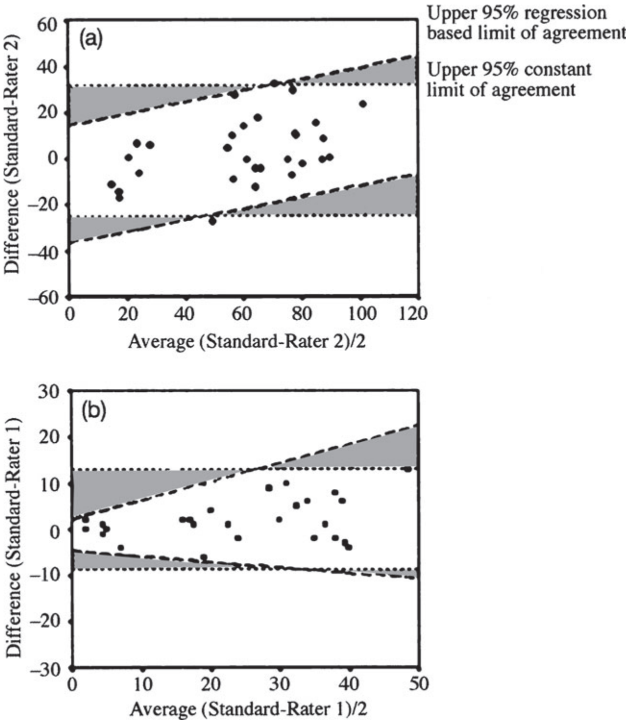

Figure 5 Constant and regression-based 95 per cent limits of agreement for (a) Index of Complexity, Outcome and Need (ICON) and (b) Peer Assessment Rating (PAR) scores.

Differences between constant and regression-based limits of agreement

It is clear that using constant limits of agreement when there is a relationship between the differences and the average can be misleading. This can happen in two ways depending on the relationship. For the ICON scores there was an increasing trend of differences with increasing magnitude of the average score. This means that for scores at either end of the scale, the limits of agreement may be either under- or overestimated using the basic approach rather than the regression-based approach. Figure 5a shows a comparison of the two sets of limits, demonstrating these important differences.

For the PAR scores it was evident that not only was there an increasing trend of differences versus average, but in addition there was a spreading out of the scatter around this trend line. In effect, the regression-based agreement limits diverge towards upper levels of average PAR score, meaning that constant limits would be extremely misleading. The differences in this case are illustrated in Figure 5b. Shaded areas indicate discrepancies between the constant limits and the regression-based limits. The limits are overestimated at the lower end of the scale of PAR score and underestimated at higher magnitudes using the basic approach compared with the regression-based approach.

Discussion

The aim of this investigation was to describe the application of a method for quantifying agreement between test raters and a standard measure for ICON and PAR scores. It is an extension to the commonly used limits of agreement originally proposed by Altman and Bland (1983). In order to reach a wider audience in the clinical community, the method was republished (Bland and Altman, 1986). Uptake of the method has been slow and incorrect analyses for agreement studies still persist (Bland and Altman, 1999).

The methods used here, advocated by Bland and Altman, focus on quantifying disagreement. For many clinical measures, agreement is a quality that is quantifiable. Hence, methods that use hypothesis testing are not applicable in this setting. Correlation is a commonly misused analysis of such data, as are regression and comparison of means (Schoolman, 1968; Westgard and Hunt, 1973; Feinstein, 1976; Altman and Bland, 1983). Correlation indices, such as the intraclass correlation coefficient, are also not suitable as they were designed for the quantification of measurement validity rather than agreement (Bland and Altman, 1990).

It was found that the use of constant 95 per cent limits of agreement for PAR and ICON scores was not suitable without extension of the approach. Regression-based limits take into account the nature of the data and inherent relationships. These limits quantify agreement between raters and standard measurements more accurately than the simple constant limits and additional analysis of the data is justified.

Once these limits of agreement are calculated it is a question of clinical judgement whether there is acceptable agreement in the particular situation. For example, it may be recommended in advance that an acceptable level of agreement for ICON scores is a difference of no more than ±18. An examination of the differences in ICON score data from Figure 4a shows not only bias at either end of the scale of measurement (in opposite directions), but also more disagreement than the recommended level. For the PAR score data, a suggested level of agreement may be a difference of no more than ±12. Examination of Figure 4b demonstrates increasing disagreement with increasing magnitude of PAR score, which is also greater than that recommended over more than half of the scale.

Bland and Altman (1999) recommended that repeated measures are also included in agreement studies as this would add additional useful information to the analysis. However, this is rarely done. It is important that these methods are demonstrated in as many fields of medicine and dentistry as possible, so that correct analyses are performed in subsequent studies. This investigation has shown that regression-based limits of agreement are more accurate for assessing the range and nature of the disagreement between raters. They therefore provide additional information that is essential for improving training programmes and the standard of clinical practice in the use of orthodontic indices.

Address for correspondence

R. A. Brown
Department of Dental Health and Biological Sciences
Cardiff University Dental School
Heath Park
Cardiff CF14 4XY
UK
E-mail: BrownRA2@Cardiff.ac.uk

References

Altman D G 1993 Calculating age-related reference centiles using absolute residuals. Statistics in Medicine 12: 917–924

Altman D G, Bland M J 1983 Measurement in medicine: the analysis of method comparison studies. Statistician 32: 307–317

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

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

Bland J M, Altman D G 1995 Comparing methods of measurement: why plotting difference against standard method is misleading. Lancet 346: 1085–1087

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

Daniels C P, Richmond S 2000 The development of the Index of Complexity, Outcome and Need (ICON). Journal of Orthodontics 27: 149–162

Feinstein A R 1976 Clinical biostatistics. XXXVII. Demeaned errors, confidence games, nonplussed minuses, inefficient coefficients, and other statistical disruptions of scientific communication. Clinical Pharmacology and Therapeutics 20: 617–631

Playle R *et al.* 1998 Determining true glomerular filtration status in newly presenting type 2 diabetic subjects using age and sex adjustment. Diabetes Care 21: 1893–1896

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 C, Stephens C D, Webb W G, Roberts C T, Andrews M 1993 Orthodontics in the general dental service of England and Wales: a critical assessment of standards. British Dental Journal 174: 315–329

Roberts C T, Richmond S 1997 The design and analysis of reliability studies for the use of epidemiological and audit indices in orthodontics. British Journal of Orthodontics 24: 139–147

Schoolman H M 1968 Statistics in medical research: principles versus practices. Journal of Laboratory and Clinical Medicine 71: 357–367

Westgard J O, Hunt M R 1973 Use and interpretation of common statistical tests in method-comparison studies. Clinical Chemistry 19: 49–57