European Journal of Orthodontics 34 (2012) 158–163
doi:10.1093/ejo/cjr010
Advance Access Publication 29 March 2011

The effect of sample size and bias on the reliability of estimates of error: a comparative study of Dahlberg's formula

S. D. Springate
Orthodontic Department, Eastman Dental Institute, London, UK

Correspondence to: Dr S. D. Springate, 111 Harley Street, London W1G 6AW, UK. E-mail: orthodontics@london.com

SUMMARY This study examined the effects of different sample sizes and different levels of bias (systematic error) between replicated measurements on the accuracy of estimates of random error calculated using two common formulae: Dahlberg's and the 'method of moments' estimator (MME). Computer-based numerical simulations were used to generate clinically realistic measurements involving random errors with a known distribution. For each simulation, two sets of 'measured values' were generated to provide the replicated data necessary for the estimation of the random error. Dahlberg's and the MME formula were applied to these paired data sets and the resulting estimates of error compared with the 'true' error. Nine different sample sizes ($n = 2$, 5, 10, 15, 20, 25, 30, 50, and 100) and two different types of bias (additive and multiplicative) were examined for their effect on the estimated error. In each case, the estimates of the random error were based on the distribution of 5000 separate simulations.

The results indicate that with a sample of less than 25–30 replicated measurements, the resulting estimates of error are potentially unreliable and may under or overestimate the true error, irrespective of the formula used in the calculation. Where, however, a bias exists between the replicate measurements, Dahlberg's formula can be expected to overestimate the true value of the random error even where the biases are small and difficult to detect by standard statistical tests. No such distorting effect was found for the MME formula, which provided estimates of error that were not meaningfully different from the true value even where relatively large biases existed between the replicates.

These results suggest the following: 1. A sample of at least 25–30 cases should be replicated to provide an estimate of the random error. 2. Where the original study contains fewer than 20 cases, the estimate of error will be unreliable. In these circumstances, it would be helpful if a confidence interval for the true error was also quoted. 3. Unless one can be absolutely sure that no bias exists between the replicate measurements, Dahlberg's formula should be avoided and the MME formula used instead.

Introduction

All physical measurements are approached with some degree of error. This is particularly so for anthropometric measurements of the type that commonly occur in clinical orthodontic research. If the errors are significant in relation to the measurements being made, they reduce the usefulness of those measurements. In comparative studies, measurement errors complicate interpretation of the results by potentially concealing important differences between groups or by indicating differences, which, in reality, do not exist. For this reason, it has become standard practice to include a statistical estimate of the measurement error in published reports of laboratory and clinical studies.

For reasons of mathematical and conceptual convenience, the total measurement error is generally partitioned into two separate classes of error: systematic and random. Systematic errors (also known as 'bias') are reproducible inaccuracies that lead to a measured value that is consistently larger or smaller than the true value. Random errors lead to variable differences from the true value and give rise, unpredictably, to measurements that are greater or smaller than the true value.

Without knowing the true value of the quantity being measured, it is not possible to determine the magnitude of any systematic error that may exist. Systematic errors can, however, be controlled by careful (and repeated) calibration of the observer and measuring apparatus against a known standard. This is not the case for random errors, but these can be reduced by averaging over a number of observations. Nevertheless, random errors set a limit on the ultimate accuracy that can be achieved no matter how many observations are averaged, and it is random error that is generally meant when speaking of 'experimental' or 'method' error.

The standard approach to estimating the random error is to replicate the measurements used in the study and then calculate some measure of the spread of the differences between them. One popular method of calculation uses the following equation originally described by Dahlberg (1940):

$$S_D = \sqrt{\frac{\sum_{i=1}^{n} d_i^2}{2n}}$$

where d is the difference between the pairs of replicate measurements, n is the number of cases, and S_D is the statistical estimate of the 'true' error (standard deviation), σ_r.

However, Houston (1983) cautioned that Dahlberg's formula will only provide a reliable estimate of the error where no bias (systematic error) exists between the two sets of replicated measurements. Unfortunately, as Houston (1983) pointed out, it is very difficult to exclude even quite large biases with certainty particularly where the sample is small.

An alternative that provides a reliable estimate of the error even when a constant bias exists between the replicates is the square root of the 'method of moments' variance estimator (MME) given by the equation:

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

where d is the difference between the pairs of replicate measurements, n is the number of cases, $\bar{d}$ is the mean of the ds, and S_M is the statistical estimate of the 'true' error (standard deviation), σ_r.

Although the first reference in the dental literature to the advantages of the MME formula over Dahlberg's formula was made over 40 years ago (Ingervall, 1964), there seems to have been no attempt to calculate the effect of using Dahlberg's formula in the presence of bias between replicate measurements. Even the study by Houston (1983), which highlighted the danger of using Dahlberg's formula, did not give any clear indication of the effect of using it in the presence of an (unidentified) bias. It seems reasonable, therefore, to ask if Dahlberg's formula produces a noticeably incorrect value for the true random error or is this simply a theoretical problem with no practical consequences?

In addition, in those studies where measurements are excessively time consuming or tedious, the researcher will usually wish to make replicate measurements of the smallest number of cases. A question that naturally arises is 'how many replicate measurements should be made and is there a price to be paid for replicating too few?'. Houston (1983) suggested that at least 25 cases should be replicated, but otherwise the existing orthodontic literature gives little guidance on this point.

The present study was undertaken to examine the performance of the two estimators (Dahlberg's and MME) using different sample sizes and in the presence of various biases in the replicated data. It was hoped that the findings would allow some simple guidelines to be formulated to assist researchers in making informed decisions regarding the answers to these questions in the different circumstances of each investigation.

Materials and methods

Numerical simulations of the process of measurement involving error were carried out by computer using the 'common precision model' (Jaech, 1985) as the basis for generating the data (Figure 1). That is, a series of measured values were generated by adding a random error and a known bias to a series of predetermined 'true' values.

For each simulation, two sets of 'measured values' were generated to provide the replicated data necessary for the estimation of the error. Statistical estimates of the random error were made by applying Dahlberg's and the MME formula to these paired data sets.

To ensure that the simulations were clinically realistic, the true values were derived from the frequency distribution of anterior cranial base length (S–N distance) for 12-year-old females reported in the Kings' College cephalometric growth study (Bhatia and Leighton, 1993). A set of true values (mean 64.9 mm; standard deviation 1.8 mm) was generated by randomly sampling from this distribution. Similarly, realistic levels of error were defined by reference to the root mean square error (error standard deviation) for S–N distance also reported by Bhatia and Leighton (1993). A Gaussian (normal) distribution of the random error was generated with a mean of zero and standard deviation of 0.5 mm. Samples were drawn at random from this distribution and added to the true values.

Two different types of bias were applied to these data: an additive offset (constant additive bias) as might occur with instrumental drift or a change in how an observer reads the measuring instrument and a multiplicative bias (non-constant bias) of the type that might occur where the radiographic object-film distance had altered over time giving a (small) difference in magnification between the two sets of measurements.

Three experiments were conducted to determine the effect of different sample sizes and different biases on the estimates of error.

Experiment 1: the effect of sample size on the estimates of error

Nine different sample sizes were used to calculate the estimates of random error: $n = 2, 5, 10, 15, 20, 25, 30, 50$, and

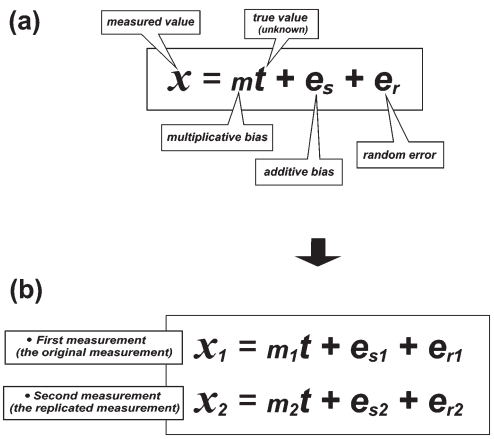

Figure 1 The mathematical models of measurement involving error employed in the study. a) The standard 'base model' for measurement error. b) The 'common precision model', which in its simplest form consists of two instances of the base model representing the two occasions of measurement. The term 'common precision' derives from the requirement that the variance of the random error (the precision) should be the same for both measurement occasions.

100. No bias was applied to any of the measurements in this experiment. For each distribution, the mean and (empirical) 95 per cent confidence limits were calculated. The 95 per cent confidence intervals (CI) were determined directly from the distributions and centred on the mean value.

Experiment 2: the effect of different levels of additive bias on the estimates of error

Using a sample size of n = 50, different additive biases were applied and the estimates of the random error were calculated for each bias. Four different magnitudes of additive bias were applied: 0 (no bias), 0.25, 0.5, and 1 mm.

Experiment 3: the effect of different levels of multiplicative bias on the estimates of error

Using a sample size of n = 50, the effect of four different magnitudes of multiplicative bias was examined by increasing the true value for one of each pair of replicates by: 0 (no bias), 1, 2, and 5 per cent. Estimates of the random error were calculated for each bias.

For each experiment, the estimates of the random error were based on the distribution of 5000 separate simulations. The mean of each distribution was taken as the most probable estimate of the random error. The estimates derived in this way were then compared with the true random error (0.5 mm). However, the differences were not tested for statistical significance because with the large samples sizes used in the simulations (n = 5000), differences of even a few micrometres in the mean values would be highly statistically significant but otherwise clinically meaningless.

The simulations were carried out on a desktop personal computer using the software program, Resampling Stats Version 3 (Resampling Stats Inc., Arlington, Virginia, USA; Chernick and Friis, 2003).

Results

Experiment 1: the effect of sample size

The results are shown in Figure 2. Although the mean of each distribution was close to the true value of 0.5 mm, for the smallest sample sizes, the range of estimates was very large with correspondingly wide CIs. The CI narrowed sharply with increasing sample size until a sample size of between 25 and 30 was reached. From then on, there was only a small reduction in the width of the CI with increasing sample size up to n = 100. No meaningful differences were found between the two estimators for any of the sample sizes examined (Table 1). As a consequence of this experiment, a sample size of n = 50 was chosen for the remaining simulations.

Experiment 2: the effect of different levels of additive bias

As can be seen from Figure 3, although there was no detectable effect on the estimates made using the MME formula, even the smallest bias of 0.25 mm had a significant influence on the estimate derived using Dahlberg's formula. For biases greater than 0.5 mm, the Dahlberg estimates of the error were grossly misleading with a magnitude of more than twice the true value.

Experiment 3: the effect of different levels of multiplicative bias

The results are shown graphically in Figure 4. The effect of a multiplicative bias was markedly different for the two methods of estimating the random error. For Dahlberg's formula, even the smallest bias (1 per cent) produced an estimate of error that was 34 per cent greater than the true value, while a 5 per cent bias led to an estimate of error that was over four times as great as the true value. For the MME formula, the same levels of bias led to estimates of error that were greater than the true value by only 4 and 6 per cent, respectively.

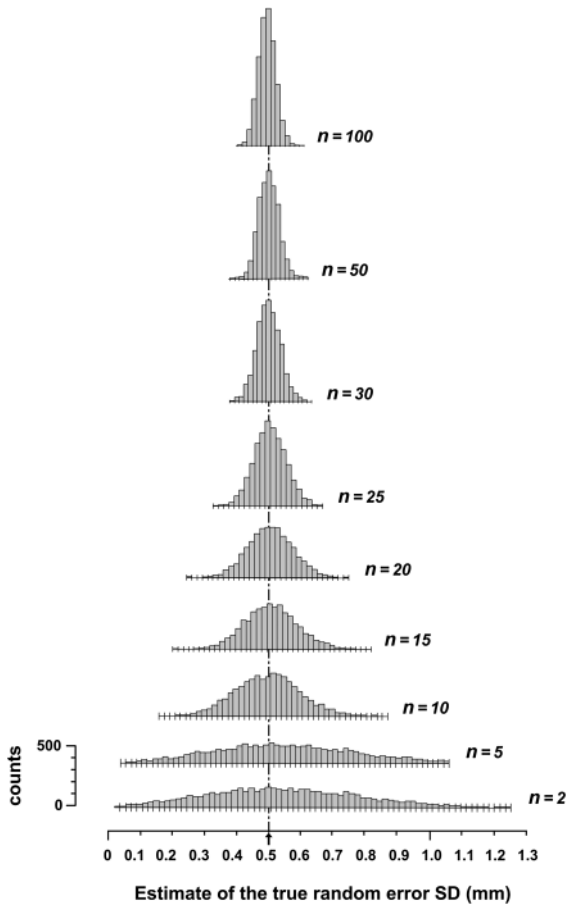

Figure 2 The effect of sample size on the distributions of the estimates of error in the absence of bias between the replicate measurements. The distributions of the estimates of random error based on 5000 simulations for each of the 9 different sample sizes examined in the first simulation experiment (experiment 1). The distributions shown are those for the Dahlberg formula. However, the distributions found using the method of moments estimator formula were similar in all respects as shown in Table 1.

Discussion

The use of replicated measurements to establish the likely error associated with a particular measurement process relies on several assumptions. One of the more important of these is that the errors of all replicate pairs can be legitimately pooled (Utermohle *et al.*, 1983). In the strictest sense, this requires the 'expected' error to be the same for all cases in a study, which is unlikely to be met in practice. However, because the error statistic is calculated using squared differences, it will not be a simple average of the individual errors but will instead be weighted towards the higher magnitude errors (Utermohle *et al.*, 1983). Consequently, unless the sample selected for replication is as representative as possible of the cases in the original study, the pooling of the individual errors will not provide an accurate estimate of the true random error.

In practice, this means that the sample must not only be selected at random but also be large enough to avoid major sampling errors that would lead to the inclusion of disproportionate numbers of cases with high or low magnitude errors. So how large should the sample be? The first simulation experiment was undertaken in an attempt to answer this question.

As can be seen in Figure 2, small samples ($n < 15$) have very wide distributions with a high proportion of the values in the tails. Consequently, such small samples run a serious risk of including unrepresentative numbers of extreme values (extremely high or extremely low). As the sample size increases towards 30, the distribution narrows rapidly towards the mean value. Above 30 cases, further narrowing of the distribution occurs only slowly with increasing sample sizes even up to $n = 100$.

The suggestion by Houston (1983) that a minimum of 25 cases must be replicated is therefore broadly supported by the results of this study. Greater numbers of cases will provide a more reliable result but replicating more than 50 cases provides almost no meaningful advantage. This result was true regardless of which formula is used to estimate the error.

For studies with fewer than 25–30 cases, measurements from all the cases should therefore be replicated but where there are fewer than approximately 20 cases, the resulting error statistic cannot be relied upon to provide an accurate representation of the true error. In these cases, it would be helpful to include the upper and lower bounds of the 95 per cent CI when quoting the error statistic. The formulae for the 95 per cent CIs and a simplified method for their calculation are given in the Appendix.

Unfortunately, where there are only a small number of cases in the original study, it is almost impossible to exclude the presence of even a relatively large bias between pairs of replicate measurements as pointed out by Houston (1983). In such circumstances, the Dahlberg statistic should be treated with suspicion because of its sensitivity to such biases. This sensitivity can be seen in the results of the remaining simulation studies where in every case, regardless of whether the bias was additive or multiplicative, Dahlberg's formula overestimated the magnitude of the true random error. On the other hand, the MME formula provided an error statistic that was very close to the true random error.

This effect of overestimation (rather than underestimation) with even quite small biases (0.25 mm additive bias or 1 per cent multiplicative bias) arises directly from the way the statistic is calculated (by squaring the differences between the replicates). That is, on average, Dahlberg's formula will

Table 1 Means and 95 per cent confidence intervals (CIs) of the distributions of error for the two estimators.

Sample size (n)	Dahlberg's estimator		Methods of moments estimator	
	Mean	95% CI	Mean	95% CI
2	0.42	0.00–0.92	0.39	0.00–0.99
5	0.48	0.08–0.88	0.49	0.07–0.92
10	0.50	0.19–0.81	0.51	0.19–0.83
15	0.51	0.25–0.76	0.52	0.26–0.77
20	0.48	0.29–0.67	0.49	0.30–0.68
25	0.51	0.31–0.71	0.52	0.32–0.71
30	0.51	0.33–0.69	0.51	0.33–0.70
50	0.49	0.37–0.62	0.49	0.37–0.62
100	0.50	0.43–0.57	0.50	0.44–0.57

For each sample size, the mean and 95% CIs are based on 5000 separate simulations.

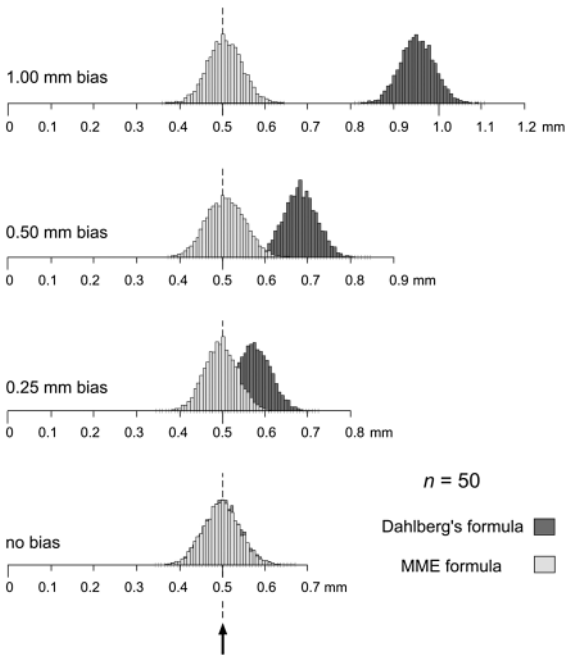

Figure 3 The effect of different levels of additive bias on the estimates of error. The distributions of the estimates of random error (each based on 5000 simulations) for the four different levels of additive bias (including zero bias) are shown for the two methods (Dahlberg's estimator and the methods of moments estimator) for estimating the error. The dashed vertical line and arrow at the base of the figure indicate the 'true' value of the random error (0.5 mm).

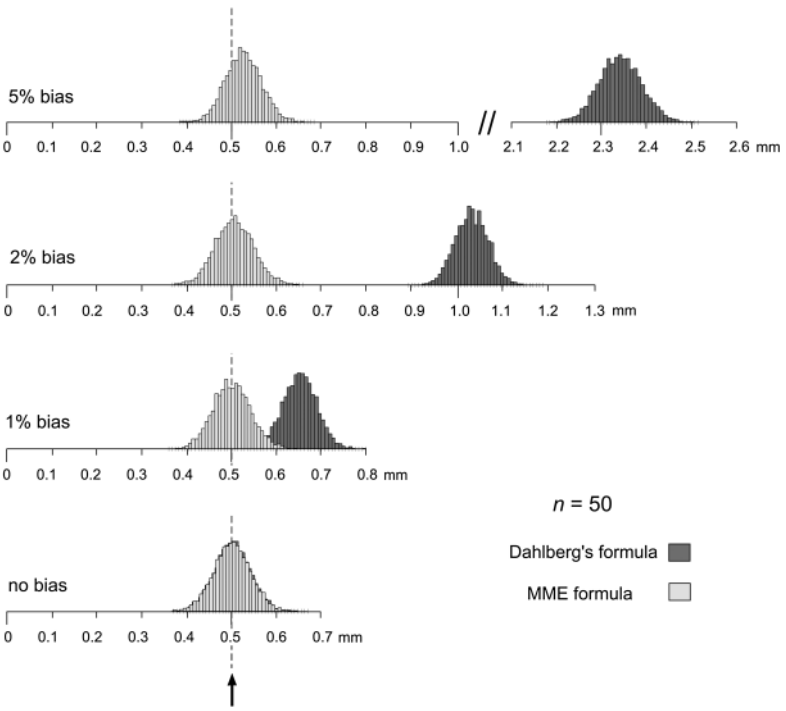

Figure 4 The effect of different levels of multiplicative bias on the estimates of error. The distributions of the estimates of random error (each based on 5000 simulations) for the four different levels of multiplicative bias (including zero bias) are shown for the two methods (Dahlberg's estimator and the methods of moments estimators) for estimating the error. The dashed vertical line and arrow at the base of the figure indicate the 'true' value of the random error (0.5 mm).

always overestimate and never underestimate the true error where a bias exists.

The distorting effect of bias between replicate measurements is likely to have its greatest impact where the true random error is very small in relation to the average size of the original measurements. That is, for those measurements that are, in reality, most precise. In such cases, where a particular measurement is believed to be very precise, the MME formula should be used in preference to Dahlberg's formula even where the sample size is considered adequate ($25 < n > 30$) because of the difficulty of excluding the presence of small biases with certainty.

Where a bias between the replicate measurements is known or believed to exist, it is clear that the MME formula should be used in preference to Dahlberg's formula. However, this should not be taken to imply that any pre-existing bias can be safely ignored or that it is unnecessary to test for bias if the MME formula is employed. The existence of bias between the replicate measurements not only indicates that an inaccuracy has arisen in the process of estimating the error, it should also alert the researcher to a possible problem in the measurement process of the original study. It should, therefore, be standard practice to examine the replicate data for the presence of bias and to test this using a single-sample Student's *t*-test.

The results of this study appear to indicate that Dahlberg's formula offers no advantage over the MME formula. There is, however, a theoretical advantage in using Dahlberg's formula where the sample size is very small and where the researcher can be certain that no bias exists. Under these circumstances, it should provide a more reliable estimate of the error than the MME formula because there are fewer degrees of freedom in the denominator—$2n$ as opposed to $2(n-1)$. This advantage is only likely to occur where fewer than 10 cases existed in the original study, but, as indicated above, with such small numbers, it becomes almost impossible to rule out the presence of bias.

Some evidence of this theoretical advantage was found in the present simulation study. In the absence of bias, the empirical 95 per cent CIs for $n = 2$, 5, and 10 were slightly narrower for Dahlberg's statistic than for the MME statistic by 0.07, 0.05, and 0.02 mm, respectively (Table 1). Nevertheless, such a small advantage is unlikely to confer any clinical or practical benefit.

A further advantage that is sometimes claimed for Dahlberg's formula is that it provides a measure of the total effect of both systematic and random errors (Krarup *et al.*, 2005). This view is incorrect and appears to result from a misunderstanding of Houston's use of the term 'systematic error' (Houston, 1983), which he used to indicate that part of the total error that was not random and also the bias between the replicate measurements used in estimating the random error. It is simply not possible to determine the true level of any systematic error that might have existed in the original study.

Conclusions

The results of this study indicate the following:

1. A sample of at least 25–30 cases should be replicated to provide an estimate of the random error. Where the original study contains fewer than 20 cases, the estimate of error will be unreliable regardless of which formula is used in the calculation. In such cases, it would be helpful if an estimated CI for the true error was also quoted.
2. Unless one can be certain that no bias exists between the replicate measurements, it is preferable to use the standard MME formula rather than Dahlberg's formula to estimate the random error.

Acknowledgement

I would like to thank Professor Nigel Hunt for his help and support during the original study from which the work arose.

References

Bhatia S N, Leighton B C 1993 A manual of facial growth. A computer analysis of longitudinal cephalometric growth data. Oxford University Press, Oxford

Chernick M R, Friis R H 2003 Introductory biostatistics for the health sciences. Modern applications including the bootstrap. John Wiley and Sons, Hoboken

Dahlberg G 1940 Statistical methods for medical and biological students. Interscience Publications, New York

Houston W J B 1983 The analysis of errors in orthodontic measurements. American Journal of Orthodontics 83: 382–390

Ingervall B 1964 Retruded contact position of mandible. A comparison between children and adults. Odontologisk Revy 15: 130–149

Jaech J L 1985 Statistical analysis of measurement errors. John Wiley and Sons, New York

Krarup S, Darvann T A, Larsen P, Marsh J L, Kreiborg S 2005 Three-dimensional analysis of mandibular growth and tooth eruption. Journal of Anatomy 207: 669–682

Utermohle C J, Zegura S L, Heathcote G M 1983 Multiple observer, humidity and choice of statistics: factors influencing craniometric data quality. American Journal of Physical Anthropology 61: 85–95

Appendix

A CI estimate for the true random error, σ_r, can be calculated using the following formulae to determine the upper and lower limits of the 95 per cent CI:

Table 1A Simplified calculation of the 95 per cent confidence limits for the 'true' error

Sample size (n)	Lower 95% confidence limit multiplication factor	Upper 95% confidence limit multiplication factor
2	0.260	4.446
3	0.327	3.043
4	0.367	2.490
5	0.395	2.194
6	0.416	2.008
7	0.433	1.884
8	0.447	1.792
9	0.459	1.721
10	0.469	1.664
11	0.478	1.618
12	0.485	1.581
13	0.492	1.548
14	0.499	1.520
15	0.505	1.495
16	0.510	1.473
17	0.515	1.454
18	0.519	1.437
19	0.523	1.421
20	0.527	1.408

1. Lower limit of the 95 per cent CI: $\sqrt{\frac{(n-1)S_d^{\,2}}{2\chi^2_{0.975(\nu)}}}$.
2. Upper limit of the 95 per cent CI: $\sqrt{\frac{(n-1)S_d^{\,2}}{2\chi^2_{0.025(\nu)}}}$,

where n is the number of cases used to estimate the error, $S_d^{\,2}$ is the mean squared difference between the replicate measurements for the n cases [using $(n-1)$ as the denominator], $\chi^2_{0.025(\nu)}$ is the 2.5 percentile of the χ^2 distribution with ν degrees of freedom, and $\chi^2_{0.975(\nu)}$ is the 97.5 percentile of the χ^2 distribution with ν degrees of freedom (where $\nu = n-1$).

For any given sample size, the expression $(n-1)/\chi^2_{0.975(\nu)}$ will be a constant, as will the expression $(n-1)/\chi^2_{0.025(\nu)}$. In addition, $S_d^{\,2}/2$ is simply $S_M^{\,2}$ (the MME of the random error variance). Consequently, the confidence limits can be calculated more simply from S_M (the MME estimate of the random error) using Table 1A.

To calculate the 95 per cent CI for a sample of less than 20 pairs of replicated measurements: first, find the row corresponding to the sample size and then multiply the calculated random error by the numbers in the lower and upper 95 per cent CI limit columns to establish the lower and upper limits of the CI.

For example, if the random error is calculated as 0.50 mm (using the MME formula) for a sample of 17 replicated measurements: the lower 95 per cent confidence limit will be 0.26 mm (=0.5 × 0.515) and the upper 95 per cent confidence limit will be 0.73 mm (=0.5 × 1.454). The random error should then be quoted thus: 0.50 mm (95% CI 0.26–0.73 mm).