A comparative assessment of cephalometric errors

Joanna M. Battagel
Department of Child Dental Health, London Hospital Medical College Dental School, London, England

SUMMARY Measuring a cephalometric radiograph is an imprecise task, and the errors associated with this should be quantified and understood. The potential sources of error are outlined and the literature relating to the assessment of measurement errors is reviewed. Three techniques are examined: the error attributable to a single recording; its associated confidence limits; and the coefficient of reliability were the most commonly applied.

Data from two error studies were pooled and used to compare the accuracy of 12 skeletal, dental, and soft tissue measurements, using various assessments of each of the three techniques. Within each group, the use of different mathematical criteria led to variation in the results; one method was totally unsuitable. As with all statistical procedures, a knowledge of the technique facilitates interpretation of the results.

Some suggestions as to the most satisfactory approach to the estimation of measurement error in cephalometry are given.

Introduction

Since Broadbent introduced the cephalostat, (Broadbent, 1931), the lateral cephalometric radiograph has become an essential tool in orthodontic analysis. Radiographic appraisal is a prerequisite of both treatment planning and assessment of the efficacy of therapeutic techniques. To adequately assess treatment methods, the results of therapy must be quantified. Thus, the cephalogram must be measured and the measurements compared critically.

Statistically, in any study involving the comparison of two groups of individuals, there are three sources of variation (or error). These are 'between group' representing real differences between the two samples, 'between individual', representing biological variation and 'within individual', representing the measurement error. Although it is the first of these which is normally of interest, any measurement of it will be influenced by the other two. The best measure of between groups variation is the difference between means, with uncertainty (due to sampling) quantified by the standard error of this difference. The appropriate measure of within group variability is the standard deviation between individuals, whilst the appropriate measure of measurement error is the within patient standard deviation. A general framework encompassing all three, is the analysis of variance.

Because the cephalometric radiograph is not a precise tool, significant errors are involved in its measurement (Hatton and Grainger, 1958; Baumrind and Frantz, 1971a,b; Midtgård et al., 1974; Houston, 1983, Houston et al., 1986; Sandler, 1988a). Every effort should be made to minimize these errors, but since they are largely inherent in the film, this is not easy. Unlike most studies which involve measurement, precision cannot be improved by repeated recording, magnification of the object or using a more accurate recording device. Cephalometric studies thus differ considerably from those in other areas of dental research.

Nonetheless, measuring error is quantifiable and an error study, reporting the results of replicate measurements, should accompany every cephalometric investigation. An estimation of measuring error also indicates which points are reliable: a prospective error assessment will assist in the choice of variables ultimately recorded (Björk, 1947).

Depending on the technique chosen, associated errors may approach the therapeutic changes reported, raising doubts as to their validity (Broadway et al., 1962; Houston, 1983). This may occur despite statistical significance between the groups, since inter-group tests of significance do not allow for any measurement error: they assume that recordings are 100 per cent accurate, when, in fact, each includes a
variable measurement error. The contribution of this error adds to the variation in the data set and so to the difficulty in obtaining statistical significance. Thus, the results of clinically satisfactory techniques may be denied statistical significance. Following functional appliance therapy, the patient clearly looks better, but scientific proof of these changes is frequently lacking (Johnston, 1968; Isaacson et al., 1990). However, an ability to demonstrate statistical significance does not necessarily have any clinical meaning.

Statistical significance may also be absent if the number of subjects in the study is too small: it is not possible to gauge whether a clinically relevant finding is due to chance or not. Thus, in cephalometric studies an adequate sized sample is most important.

Prior to analysis, it is desirable that all radiographs are of high quality and are taken in standardized conditions. Thus, high speed film should be used with an appropriate soft tissue filter and a constant mid-sagittal plane to film distance maintained. Head position should be standardized and movement restricted.

Digital recording is more convenient than measuring with ruler and protractor. In terms of accuracy, there is little to choose between direct digitising and the use of an intervening tracing stage (Cohen, 1984; Sandler, 1988b) since the major component of recording error is that of point identification (Baumrind and Frantz, 1971a,b; Gravely and Murray Benzies, 1974; Sandler, 1988a). Bjork (1947), Chebib and Burdick (1973), and Houston (1983) described alternative calculations.

Confidence limits, describing the accuracy with which a given measurement could be made, were recommended by Gravely and Murray Benzies (1974), Kirkwood (1988), Bulman and Osborn (1989), and Snedecor and Cochran (1989). A similar procedure was reported by Chebib and Burdick (1973). A coefficient of variation was calculated by Savage et al. (1987), whilst estimates of reliability were described by Midtgård et al. (1974), Houston (1983), and Buschang et al. (1987). An analysis of variance was suggested by both Houston et al. (1986) and Buschang et al. (1987). More straightforward assessments involving the means and standard deviations were suggested by Broadway et al. (1962) and Richardson (1966). Because different kinds of error are being tested, the results of the various error evaluations will differ, depending on the calculation selected and the composition of the data under study.

The purpose of this study was, therefore, to review and compare critically the more commonly used methods of investigating cephalometric measurement error. To illustrate this, errors were calculated for a selection of hard and soft tissue measurements derived from a cross-sectional study.

**Review of the literature**

**Errors associated with recording a single measurement**

1. **Dahlberg’s (1940) statistic.** The concept of error and its relationship to biological measurements has been examined in depth by Dahlberg (1940). Replicate measurements for each of a series of patients are compared and the standard deviation of each of the paired measurements from its own pair mean calculated from the formula:

\[
S_e = \sqrt{\frac{\sum d^2}{2n}}
\]

where: \(S_e\) is the standard deviation of the differences of each of the replicates from its mean; \(n\) is the number of radiographs recorded; and \(d\) is the difference between the first and second recordings.
Whilst Dahlberg (1940) describes his error ($S_e$) as the standard error of the mean, this is, strictly speaking, the standard deviation of the difference of each of the paired measurements from its own pair mean as described above. Thus, $S_e$ as described here is identical to the $s$ described by Chebib and Burdick (1973) below.

2. Chebib and Burdick’s (1973) method. These authors suggest the following calculation:

$$
\hat{e} = \pm \frac{2}{\pi k} s
$$

where $\hat{e}$ is the expected error associated with each measurement; $s$ is the standard deviation of the measurement error; and $k$ is the number of readings of each measurement.

The Dahlberg (1940), and Chebib and Burdick (1973) formulae are mathematically related. Where films are digitized twice, Chebib and Burdick’s error is approximately half of that described by Dahlberg (1940). (The fraction on the right side of the equation simplifies to the square root of one divided by $\pi$ or 0.5642.) Both estimations describe the error associated with a single measurement based on a number of replicated readings. In theory, this would decrease as the number of replicates increased, but as Houston (1983) pointed out, repeated recording has little effect on the accuracy of cephalometric evaluation since the prime difficulty lies in point identification.

3. Houston’s (1983) estimate of random error. This is described as the variance of the difference between replicate determinations. On the basis that differences in both data sets may occur, the variance is divided by 2, thus producing a formula identical to that of Dahlberg, but by algebraic coincidence.

4. Björk’s (1947) quotient. Björk (1947) considered a different approach. Using replicate measurements, the standard error of the mean differences was compared with the mean of the differences themselves. For a measurement to be satisfactory, the mean of the differences should not exceed three times the standard error of replicate measurements. This assessment was based on the premise that for any experimental difference to be truly significant, the change must be three times greater than the standard error associated with that measurement.

5. Houston’s (1983) estimate of systematic error. Both Dahlberg’s (1940), and Chebib and Burdick’s (1973) estimations combine systematic and random components. Where the systematic error is zero, this will reflect the random error satisfactorily, but Houston (1983), stressed the need to separate the two. To test the systematic error, he recommended a one sample t-test between adequate numbers of replicate measurements, examined as a minimum, at the 10 per cent level of significance.

The coefficient of reliability

An alternative method of expressing the accuracy of a measurement is to consider the reliability of its assessment. Midtgår et al. (1974) emphasized that the error between double determinations was inadequate, unless it was also related to the variance of the landmarks in the material as a whole. They arbitrarily suggested that ideally, error variance should be less than 3 per cent of the study variance; with a coefficient greater than 10 per cent, the measurement would be inappropriate. It is more relevant to assess reliability in the context of a particular study than to apply selective thresholds.

Houston (1983) modified this approach, calculating the coefficient by subtracting the ratio of the random error variance divided by the total study variance, from unity. The random error variance was divided by two and the study variance of the first estimate was preferred to that of a mean of the two estimations. Thus, Houston’s (1983) calculation is related to that of Midtgår et al. (1974), but appears approximately twice as reliable.

In their assessment of reliability, Buschang et al. (1987) determined the ratio between the true and the total observed variances. The true variance was a hypothetical concept, being the variation which remained once random errors have been accounted for.

Confidence limits

The reliability of the recording of a single measurement may also be described in terms of its confidence limit (Gravely and Murray Benzies, 1974). The confidence limit within which a determination lies, is the product of the Dahlberg statistic and a value of $t$ dependent on the level of probability chosen and the number of cases under study. Frequently, the 95 per cent confidence limit is selected, with
the $t$ value being obtained from the 0.05 level of significance column in standard $t$ tables.

Whilst Chebib and Burdick (1973) describe
the maximum error associated with $p$ per cent of
recordings, this is based on a different mathe-
matical concept. It alludes to the value which
one is $p$ per cent confident that the true mean
does not exceed. The mean is calculated from
duplicate determinations and this would appear
a less satisfactory estimation. The two assess-
ments differ by a factor of one divided by root
2. Since the square root of two is 1.414, the
apparent accuracy has increased by 1/1.414 or
41 per cent, using this alternative.

Other confidence limits, described in text-
books of medical statistics (Kirkwood, 1988;
Bulman and Osborn, 1989; Snedecor and
Cochran, 1989) are based variously on the
standard deviation of the differences between
the two determinations (Bulman and Osborn,
1989; Snedecor and Cochran, 1989) or on the
measurements themselves (Kirkwood, 1988).

Materials and methods

The material used in this study comprised 246
lateral cephalometric radiographs: 105 were
films of a control group of children with accept-
able occlusions, whilst 141 comprised pretreat-
ment radiographs of patients with Class III
malocclusion. For 20 films in each group, rep-
licate recordings were made and all analyses of
error were based on these 40 films.

Each radiograph was traced onto acetate
paper using a fine 3H pencil. The tracing was
transferred to a drafting board, orientated
with the horizontal at 7 degrees to sella–nasion
(Burstone et al., 1978), and 19 skeletal, dental
and soft tissue points identified or constructed.
Points recorded conformed to British Standard
definitions (British Standards Institution, 1983).

Points were pricked with a stylus for better
identification and tracings digitized twice to a
tolerance of 0.1 mm. An average of the two
readings was used in subsequent calculations.
The radiograph was automatically re-orien-
tated at 7 degrees to S–N, and 12 measurements
determined. These were chosen to illustrate the
range of errors associated with a cross-section
of cephalometric variables. All measurements
were corrected for radiographic magnification
and converted to life size.

The entire procedure, from tracing to calcula-
tion was repeated several weeks later, when no
memory of the original films remained.

The following methods of describing error
were contrasted:

The errors associated with recording a single
measurement:

1. Dahlberg’s (1940) statistic.
2. Chebib and Burdick’s (1973) method.
4. Björk’s (1947) criteria.
5. Estimate of systematic error: Houston
(1983).

Confidence limits

2. Chebib and Burdick (1973)

Coefficients of reliability

1. Midergard et al. (1974)
2. Houston (1983)

The effects of the errors on the interpretation
of two cross-sectional samples of clinical data
(Table 1) were considered and their implications
discussed.

Results and discussion

The results are presented in Tables 1–4.

Errors associated with recording a single
measurement (Table 2)

Five methods of assessing the error associated
with the measurement of a single variable are
contrasted in Table 2. Column 1 shows the
Dahlberg (1940) statistic, column 2 depicts the
errors calculated by Chebib and Burdick’s
(1973) method, whilst column 3 gives Houston’s
(1983) assessment. Björk’s calculation is pre-
sented in column 4 and the $t$ value calculated
as recommended by Houston for the assessment
of systematic error, in column 5. The magnitude
of the errors varies slightly between methods of
assessment, but the greatest variation is seen
between variables.

1. Dahlberg (column 1).

Where the constituent points are well separated and relatively easy to
identify, the associated errors are half a unit. The values for both nasolabial angle
Table 1  Means and standard deviations of selected cephalometric variables for 105 Control and 141 Class III female children and the significance of any differences between them.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Control mean</th>
<th>Control SD</th>
<th>Class III mean</th>
<th>Class III SD</th>
<th>Difference Class III — control &amp; signif.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Maxilla</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SNA (degrees)</td>
<td>80.22</td>
<td>3.35</td>
<td>79.32</td>
<td>3.65</td>
<td>-0.90 NS</td>
</tr>
<tr>
<td>A to sella vertical (mm)</td>
<td>61.47</td>
<td>2.48</td>
<td>59.60</td>
<td>4.00</td>
<td>-1.88***</td>
</tr>
<tr>
<td>Articulare — ANS (mm)</td>
<td>83.65</td>
<td>3.73</td>
<td>80.05</td>
<td>4.33</td>
<td>-3.59***</td>
</tr>
<tr>
<td>Mandible</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SNB</td>
<td>77.25</td>
<td>3.21</td>
<td>80.74</td>
<td>3.97</td>
<td>3.49***</td>
</tr>
<tr>
<td>B to sella vertical (mm)</td>
<td>55.11</td>
<td>5.68</td>
<td>59.30</td>
<td>6.77</td>
<td>4.19***</td>
</tr>
<tr>
<td>Mandibular arc (degrees)</td>
<td>33.20</td>
<td>5.57</td>
<td>29.49</td>
<td>6.25</td>
<td>-3.72***</td>
</tr>
<tr>
<td>Articulare — pogonion (mm)</td>
<td>95.55</td>
<td>7.71</td>
<td>98.82</td>
<td>6.51</td>
<td>3.26***</td>
</tr>
<tr>
<td>Intermaxillary</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MM angle (degrees)</td>
<td>26.95</td>
<td>5.30</td>
<td>27.87</td>
<td>6.06</td>
<td>0.93 NS</td>
</tr>
<tr>
<td>Dental</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1/1 to sella vertical (mm)</td>
<td>63.53</td>
<td>5.43</td>
<td>62.51</td>
<td>5.78</td>
<td>-1.02 NS</td>
</tr>
<tr>
<td>Overjet (mm)</td>
<td>2.61</td>
<td>1.27</td>
<td>-0.44</td>
<td>1.63</td>
<td>-3.05***</td>
</tr>
<tr>
<td>Overbite (mm)</td>
<td>4.05</td>
<td>1.43</td>
<td>1.35</td>
<td>2.06</td>
<td>-2.70***</td>
</tr>
<tr>
<td>Soft tissue</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nasolabial angle (degrees)</td>
<td>94.43</td>
<td>5.13</td>
<td>89.52</td>
<td>6.85</td>
<td>-4.90***</td>
</tr>
</tbody>
</table>

Significance levels: * P≤0.05; ** P≤0.01, *** P≤0.001

Table 2  Five methods of assessing the error associated with recording a single measurement. Column 1 shows the Dahlberg (1940) statistic, column 2 the method of Chebib and Burdick (1973), column 3 illustrates Houston's (1983) assessment, column 4 gives Bjork's quotient and column 5 the result of Houston's (1983) test for systematic error.

All calculations are derived from the replicate recordings of 20 control and 20 Class III cases.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Maxilla</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SNA (degrees)</td>
<td>0.55</td>
<td>0.31</td>
<td>0.55</td>
<td>0.97</td>
<td>0.197 NS</td>
</tr>
<tr>
<td>A to sella vertical (mm)</td>
<td>0.53</td>
<td>0.30</td>
<td>0.53</td>
<td>1.25</td>
<td>0.239 NS</td>
</tr>
<tr>
<td>Articulare—ANS (mm)</td>
<td>0.97</td>
<td>0.55</td>
<td>0.96</td>
<td>1.46</td>
<td>0.315 NS</td>
</tr>
<tr>
<td>Mandible</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SNB</td>
<td>0.45</td>
<td>0.25</td>
<td>0.43</td>
<td>2.31</td>
<td>0.383 NS</td>
</tr>
<tr>
<td>B to sella vertical (mm)</td>
<td>0.66</td>
<td>0.37</td>
<td>0.53</td>
<td>2.10</td>
<td>0.289 NS</td>
</tr>
<tr>
<td>Mandibular arc (degrees)</td>
<td>2.15</td>
<td>1.21</td>
<td>2.12</td>
<td>1.41</td>
<td>0.697 NS</td>
</tr>
<tr>
<td>Articulare—pogonion (mm)</td>
<td>0.68</td>
<td>0.38</td>
<td>0.68</td>
<td>1.08</td>
<td>0.170 NS</td>
</tr>
<tr>
<td>Intermaxillary</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MM angle (degrees)</td>
<td>0.61</td>
<td>0.34</td>
<td>0.62</td>
<td>0.82</td>
<td>0.134 NS</td>
</tr>
<tr>
<td>Dental</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1/1 to sella vertical (mm)</td>
<td>0.51</td>
<td>0.29</td>
<td>0.51</td>
<td>1.24</td>
<td>0.159 NS</td>
</tr>
<tr>
<td>Overjet (mm)</td>
<td>0.40</td>
<td>0.23</td>
<td>0.40</td>
<td>1.10</td>
<td>0.262 NS</td>
</tr>
<tr>
<td>Overbite (mm)</td>
<td>0.35</td>
<td>0.20</td>
<td>0.33</td>
<td>1.98</td>
<td>0.414 NS</td>
</tr>
<tr>
<td>Soft tissue</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nasolabial angle (degrees)</td>
<td>2.60</td>
<td>1.47</td>
<td>2.63</td>
<td>0.55</td>
<td>0.299 NS</td>
</tr>
</tbody>
</table>
and Ricketts' mandibular arc exceed two units. These measurements involve points which are either close together (the nasolabial angle) or relatively difficult to locate, (XI point in the ascending ramus).

2. Chebib and Burdick (column 2). The errors calculated by Chebib and Burdick's (1973) method are consistently smaller than those of Dahlberg (1940). The two methods are mathematically related, with Chebib and Burdick's (1973) standard error approximately half that described by Dahlberg (1940).

3. Houston (column 3). The magnitude of the errors computed by Houston's (1983) equation is similar to that of Dahlberg (1940). There are two reasons for this. Both estimates examine the standard deviations of differences between replicate pairs, but they differ in their exact calculation of this statistic. Whilst Dahlberg (1940) subtracts each replicate from the mean of the pairs, Houston (1983) subtracts the mean of the summed differences from each difference between replicate pairs. Secondly, Houston divides his total by (n – 1), whereas Dahlberg divides by n. As the overall arithmetic mean is small in relation to the differences themselves and n approximates (n – 1), the results will be similar. Although Chebib and Burdick's (1973) calculation is identical to that of Dahlberg (1940), the rationale for including \( \pi \) in the denominator and, therefore, the proportionate reduction in their assessment, is not apparent from their presentation.

4. Björk (column 4). Björk (1947) takes two statistics from his paired data sets, the mean of the differences between replicate measurements and the standard error of the mean differences. The former is divided by the latter to produce the quotient given in column 4.

There was a wide range of variation between measurements, but none of the 12 variables breached Björk's acceptance criteria. Neither the nasolabial angle nor the mandibular arc showed sizable errors because of the wide range of cases under study. The larger the standard error, the better the result; thus, a drawback of Björk's technique is that homogeneity of data adversely affects the outcome.

5. Houston's estimation of systematic errors (column 5). No systematic errors were found. The critical value at the 90 per cent level of probability and (n – 1) degrees of freedom is 1.687. Houston (1983) stressed the need to include an adequate number of cases in the investigation of systematic error (at least 25), otherwise large standard deviations and associated standard errors could produce artificially small values of \( t \).

Because this test depends on the dispersion of the data, a non-homogeneous data set such as the combined control and Class III material presented here, will produce artificially good results and this has contributed to the absence of error reported. For the first three assessments, any pooling of data sets provides error values which are the arithmetic mean of the individual groups, but both Björk's and Houston's systematic assessments are sensitive to group composition. Thus, they are inappropriate for the data in this study.

**Confidence limits (Table 3)**

The traditional confidence limits described by Gravely and Murray Benzies (1974) are presented in Table 3, column 1. The analogous 'error which 95 per cent of measurements should not exceed', calculated by Chebib and Burdick (1973) is shown in column 2, whilst the assessment recommended by Kirkwood (1988) is given in column 3. Using Gravely and Murray Benzies' (1974) assessment, only three variables exhibited confidence limits which were less than one unit: overbite, overjet, and the angle SNB. Since these errors may occur in either direction, that is, the confidence interval is plus or minus the value given, the dispersion is twice the value presented. Thus, for most measurements, the range within which the actual value will lie, will be greater than two units.

The method described by Chebib and Burdick (1973) is similar and apparently more accurate, but like the authors' assessment of single measurement error, the rationale for this calculation is unclear. This further confuses the issue of the relevance of the different methods of error estimation and of the interpretation of subsequent study findings.

Kirkwood's (1988) calculations bear no relationship to either of the other methods, being based on the standard deviations of the measurements themselves. The number of individuals in the study is included in the calculation in its own right and, thus, the resultant confidence limit depends entirely on the dispersion of the
CEPHALOMETRIC ERRORS

Table 3 95 per cent confidence limits: Gravely and Murray Benzies (1974) column 1, the value within which 95 per cent of measurements will fall (Chebib and Burdick 1973) column 2, and the 95 per cent confidence limit described by Kirkwood (1988) column 3.

All calculations are derived from the replicate recordings of 20 control and 20 Class III cases.

<table>
<thead>
<tr>
<th>Variable</th>
<th>95 per cent confidence limit</th>
<th>95 per cent measures fall within this range</th>
<th>95 per cent confidence limit</th>
</tr>
</thead>
<tbody>
<tr>
<td>Maxilla</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SNA (degrees)</td>
<td>1.11</td>
<td>0.78</td>
<td>1.22</td>
</tr>
<tr>
<td>A to sella vertical (mm)</td>
<td>1.08</td>
<td>0.86</td>
<td>1.25</td>
</tr>
<tr>
<td>Articulare–ANS (mm)</td>
<td>197</td>
<td>1.39</td>
<td>1.35</td>
</tr>
<tr>
<td>Mandible</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SNB (degrees)</td>
<td>0.92</td>
<td>0.70</td>
<td>1.17</td>
</tr>
<tr>
<td>B to sella vertical (mm)</td>
<td>1.34</td>
<td>0.95</td>
<td>2.08</td>
</tr>
<tr>
<td>Mandibular arc (degrees)</td>
<td>4.34</td>
<td>3.07</td>
<td>1.93</td>
</tr>
<tr>
<td>Articulare–pogonion (mm)</td>
<td>1.38</td>
<td>0.98</td>
<td>2.03</td>
</tr>
<tr>
<td>Intermaxillary MM angle (degrees)</td>
<td>1.24</td>
<td>0.88</td>
<td>1.76</td>
</tr>
<tr>
<td>Dental</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>L/L to sella vertical (mm)</td>
<td>1.03</td>
<td>0.73</td>
<td>1.79</td>
</tr>
<tr>
<td>Overjet (mm)</td>
<td>0.81</td>
<td>0.57</td>
<td>0.75</td>
</tr>
<tr>
<td>Overbite (mm)</td>
<td>0.70</td>
<td>0.50</td>
<td>0.72</td>
</tr>
<tr>
<td>Soft tissue</td>
<td>Nasolabial angle (degrees)</td>
<td>5.27</td>
<td>3.72</td>
</tr>
</tbody>
</table>

data and the number of individuals surveyed. An increase in the number of persons automatically improves the confidence limit yet the precision of recording is unchanged. This is not a helpful attribute and the technique contributes nothing to the description of measurement error. The assessment described by Bulman and Osborn (1989), however, would be suitable.

The validity of the confidence limit as a measure of imprecision in cephalometric measurement should be considered carefully. At first sight, it would appear that all cephalometric studies will be unreliable. The confidence limit associated with SNB as calculated by Gravely and Murray Benzies (1974) approaches 1 degree. Thus, any individual measurement of this angle might have an associated error of nearly 2 degrees. Since the changes found in SNB in short-term serial studies are unlikely to exceed this value, the significance of such findings may seem questionable. For the cross-sectional data reported here, the significance between groups at the 0.01 per cent level calculated in Table 1 appears less suspect. It could be argued that the confidence limit is not an appropriate statistic to apply: it does not describe the specific error in a particular study, but a range of possibilities. Since it is the product of Dahlberg's statistic and a value which approximates 2 at the 95 per cent level, it will always be two to three times the Dahlberg value. Furthermore, the confidence limits quoted apply to 95 per cent of recordings; on the remaining 5 per cent of occasions, the error could be even worse.

Coefficients of reliability (Table 4)

Coefficients of reliability as calculated by Midtgård et al. (1974) are presented in Table 4, column 1, corresponding assessments by Houston's (1983) method are shown in column 2 and those of Buschang et al. (1987) are described in column 3.

Houston (1983) and Buschang et al. (1987) both subtract their coefficients from 1: therefore, these describe percentage reliability rather than error. The figures derived by Buschang et al.
Table 4  Coefficients of reliability as calculated by: Midtgård et al. (1974), column 1; by Houston (1983), column 2; and by Buschang et al. (1987), column 3. All calculations are derived from the replicate recordings of 20 control and 20 Class III cases. For the overjet, a separate calculation is included, based on the 20 control cases only.

<table>
<thead>
<tr>
<th>Variable</th>
<th>%</th>
<th>%</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Maxilla</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SNA</td>
<td>4.1</td>
<td>98.0</td>
<td>95.8</td>
</tr>
<tr>
<td>A to Sella vertical</td>
<td>3.4</td>
<td>98.2</td>
<td>96.3</td>
</tr>
<tr>
<td>Articulare-ANS</td>
<td>10.1</td>
<td>95.0</td>
<td>89.7</td>
</tr>
<tr>
<td>Mandible</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SNB</td>
<td>2.8</td>
<td>98.6</td>
<td>97.2</td>
</tr>
<tr>
<td>B to Sella Vert.</td>
<td>1.9</td>
<td>99.1</td>
<td>98.1</td>
</tr>
<tr>
<td>Mandibular arc</td>
<td>23.3</td>
<td>88.4</td>
<td>75.2</td>
</tr>
<tr>
<td>Art.-Pogion</td>
<td>2.5</td>
<td>98.8</td>
<td>97.3</td>
</tr>
<tr>
<td>Intermaxillary</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MM angle</td>
<td>2.7</td>
<td>98.7</td>
<td>97.5</td>
</tr>
<tr>
<td>Dental</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1/1 to Sella Vert.</td>
<td>1.6</td>
<td>99.2</td>
<td>98.4</td>
</tr>
<tr>
<td>Overjet</td>
<td>5.7</td>
<td>97.1</td>
<td>94.2</td>
</tr>
<tr>
<td>Overjet (20 control cases)</td>
<td>25.3</td>
<td>87.1</td>
<td>72.3</td>
</tr>
<tr>
<td>Overbite</td>
<td>4.4</td>
<td>97.8</td>
<td>95.6</td>
</tr>
<tr>
<td>Soft tissue</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nasolabial angle</td>
<td>41.0</td>
<td>79.5</td>
<td>54.3</td>
</tr>
</tbody>
</table>

(1987) are similar, but not identical to those deduced by Midtgård et al. (1974) subtracted from unity. In Houston’s (1983) calculation, the range is reduced because he considers that, since two series of measurements have been made, the error variance should be divided by 2 before this value is, in turn, divided by the total study variance.

The coefficients described by all three methods show a wide variation, with the nasolabial angle and mandibular arc consistently showing poor reliability. These variables show large associated errors by nearly all other methods examined. However, a poor coefficient of reliability is not invariably associated with a high Dahlberg error, as inspection of the errors associated with overjet will demonstrate.

Dahlberg’s (1940) method does not take into account the size of the error in relation to the magnitude of the variable itself. The coefficients of reliability seem to address this problem, but in so doing, create their own anomaly. The Dahlberg error associated with overjet (Table 2) is low: 0.40 mm. In proportion to the size of the mean overjets (Table 1), this is an appreciable component. The poor reliability in the homogeneous control group (25 per cent error using the Midtgård et al. (1974) method, 87 per cent reliability according to Houston (1983), and 72 per cent according to Buschang et al. (1987), reflects this satisfactorily.

When the Class III and control material are pooled and the reliability recalculated, it appears to have improved considerably: 6 per cent error according to Midtgård et al. (1974), 97 per cent reliable by Houston’s (1983) calculation and 94 per cent by the estimation of Buschang et al. (1987). The accuracy of recording has not altered, but a rather disparate subsample of data has been added. Since the calculation of reliability is designed to reflect the accuracy in relation to the total study material, widening the scope of the study, increases the variance and improves the coefficient. This is not always helpful and may be avoided in studies involving a broad spectrum of malocclusions by making a separate examination of the reliability of each subsample of data. As with all statistical tests, the results must not be interpreted blindly.
CEPHALOMETRIC ERRORS

Composition of the measurement error

The present study examined the errors associated with the measurement of cephalometric variables, as this was considered to be more realistic clinically than the analysis of errors associated with discrete points. The error of any measurement comprises the errors associated with all its component points. Thus, measurements are more accurate at right angles to the ellipses of the error envelopes of their constituent points (Baumrind and Frantz, 1971a, b). The distances between landmarks will also affect the outcome. Such errors will normally be greater than those associated with single points (Björk, 1947; Baumrind and Frantz, 1971b).

Conclusions

1. The results of a cephalometric study should be interpreted in relation to its associated measurement error.
2. No one method of error assessment can give a complete picture.
3. Dahlberg's estimation is the soundest method mathematically to evaluate measurement error.
4. Because the Dahlberg statistic does not take into account the proportionate size of the error in relation to the measurement itself, an alternative calculation may be indicated. This will depend on the form of the errors.
5. The coefficient of reliability as proposed by Midtgård et al. (1974) is suitable, provided its sensitivity to sample composition is recognized.
6. In relation to many orthodontic measurements, confidence limits based on Dahlberg's statistic are wider than the changes under study. They do not indicate the error associated with a single measurement, but rather the range of possible discrepancies over time.
7. Confidence limits calculated from the standard error of the data are unsuitable to describe the accuracy of measurement.
8. Any statistic which is influenced by the homogeneity of the data set should be interpreted with care.

It is suggested that a Dahlberg (1940) estimation is supplemented by the coefficient of reliability described by Midtgård et al. (1974). The limitation of the latter should be borne in mind.

Address for correspondence
Joanna M. Battagel
Department of Child Dental Health
London Hospital Medical College Dental School
Stepney Way
London E1 2AD
England

Acknowledgements

The author is indebted to Mr N. J. Battagel for designing de novo the computer software required to perform the analyses used in this study.

References

Björk A 1947 The face in profile. An anthropological X-ray investigation on Swedish children and conscripts. Svenska Tandläkar—Tidskrift 40: Suppl. 5B
British Standards Institution 1983 British standard glossary of dental terms. BS 4492, London
Broadbent B H 1931 A new X-ray technique and its application to orthodontia. Angle Orthodontist 1: 45-66
Dahlberg G 1940 Statistical methods for medical and biological students. InterScience Publications, New York


Richardson A 1966 An investigation into the reproducibility of some points, planes, and lines used in cephalometric analysis. American Journal of Orthodontics 52: 637–651


