European Journal of Orthodontics 24 (2002) 457–470

Natural head posture—considerations of reproducibility

D. Bister*, R. J. Edler*, B. D. M. Tom** and A. T. Prevost**

*Department of Orthodontics, Norman Rowe Maxillofacial Unit, Queen Mary's University Hospital, Roehampton, London and **Department of Public Health and Primary Care, Institute of Public Health, Robinson Way, Cambridge, UK

SUMMARY This three-part study investigated the reproducibility of natural head posture (NHP) using radiographs and photographs. In part 1, reproducibility of cephalograms 1 year after the introduction of NHP was investigated and found to be less favourable (Dahlberg coefficient of 2.99 degrees) than most other previous investigations. In order to minimize radiation exposure of patients, reproducibility of photographs and method agreement between photographs and cephalograms were investigated in part 2. Reproducibility of the two photographs was poor (2.71 degrees). However, method agreement between cephalograms and the photographs taken at the same time was good (1.39 degrees). Replacement of the radiographic method with the photographic method for assessing NHP reproducibility appeared justified. Changing the protocol for achieving NHP in part 3 of the study improved reproducibility substantially (1.41 degrees).

Various statistical methods were used to assess reproducibility and method agreement. Bland and Altman's graphical representation was found to be the most appropriate for method agreement. The Dahlberg coefficient, commonly used to assess NHP repeatability/reproducibility, does not provide an extreme enough interval to allow a sufficient clinical assessment of a method to be undertaken, compared with the reproducibility coefficient. That is, the latter provides a 95 per cent range, compared with 52 per cent with Dahlberg.

Introduction

Although it has been known for some time that use of intra-cranial reference lines for assessment of anterior–posterior skeletal relationships is inherently unreliable (Downs, 1956), they are still widely used for diagnosis and treatment planning. The variability of planes such as sella–nasion and Frankfort Horizontal to each other as well as to the true horizontal is such that measurements based on these planes are likely to give misleading information (Houston, 1991; Moorrees, 1995). As pointed out by Proffit and White (1991), such measurements when used on orthognathic patients are likely to be even more misleading; so use of the true horizontal and/or vertical planes as alternatives would appear to be essential.

Whilst the overwhelming majority of publications (e.g. Bjerin, 1957; Moorrees and Kean, 1958; Solow and Tallgren, 1971; Siersbæk-Nielsen and Solow, 1982) on repeatibility/reproducibility of natural head posture (NHP) show good results, even after 15 years (Peng and Cooke, 1999), its use has not been more widespread. Indeed critical assessments of NHP reproducibility are rare (Luyk *et al*., 1986). There might be several reasons why NHP has not found common acclaim: confusion over both terminology and methodology in achieving NHP, lack of reliable reference data, and the fact that taking radiographs in NHP may be more time-consuming than simply positioning Frankfort Horizontal parallel to the horizontal.

However, there is certainly no contra-indication to ensuring that patients' heads are orientated in NHP before lateral cephalograms are taken. Accordingly, when in 1997 a dedicated radiographer was appointed to the Maxillofacial Unit at Queen Mary's Hospital, Roehampton, London and new cephalometry equipment installed, it seemed appropriate to try and ensure that films should be taken in NHP and thus to train the radiographer appropriately and then test the reproducibility of the films. In the orthodontic

literature it has become common practice to quote Dahlberg's (1940) coefficient when assessing the reproducibility of a single method or the agreement between two methods. For assessment of reproducibility of NHP, a coefficient that takes a value below a cut-off point of approximately 1.5–2 degrees has been used to indicate good reproducibility or agreement (e.g. Cooke and Wei, 1988a). However, it would appear that neither the appropriateness of this coefficient nor the validity of the cut-off point has been formally addressed in the orthodontic literature. This study presents an attempt at rectifying this by giving a clinical interpretation to Dahlberg's coefficient and comparing it with other statistical methods, namely the International Standards Organization, 1994 (ISO; 5725) definition of reproducibility, Fleiss and Cohen's (1973) intraclass correlation coefficient (ICC), and Bland and Altman's (1986) 'limits of agreement' approach.

The aims of the study were to:

1. assess the reproducibility of NHP after the introduction of new X-ray equipment in 1997;
2. determine which statistical method best represents reproducibility of a method and agreement between methods in a clinical situation;
3. investigate the effect of altering the NHP protocol on reproducibility;
4. devise a time-efficient method for auditing reproducibility without the use of ionizing radiation, for the purpose of training radiographers in NHP.

Materials and method

Terminology

According to the ISO (1994), repeatability refers to test conditions that are as constant as possible, where independent test results are obtained with the same method on identical test items in the same laboratory by the same operator, using the same equipment within short intervals of time. Reproducibility refers to test conditions under which results are obtained with the same method on identical test items, but in different laboratories with different operators and using different equipment. Because the time interval plays an important role in the definition of reproducibility, this term was used in this study rather than repeatability. Method agreement was used instead of reproducibility when comparing the results obtained with different methods on identical test items.

Part 1 retrospectively assessed the reproducibility of NHP involving all patients who had two lateral cephalograms taken after the introduction of new X-ray equipment in 1997 and March 1998. Patients were not exposed to unnecessary radiation for the purpose of this study; all radiographs were clinically indicated either for diagnostic purposes (start of treatment) or for assessment of treatment progress. The time interval between the two radiographs was 3–8 months.

Part 2 was prospective and aimed to assess both method agreement between radiography and photography, and reproducibility of NHP comparing two photographs.

Part 3 was also prospective and assessed reproducibility of NHP using the photographic method, but with a modified protocol. For parts 2 and 3, an initial photograph (photo 1) was taken immediately before the cephalogram. After development of the radiograph the patient was re-positioned in the cephalostat and a second photograph was taken approximately 10 minutes after photograph 1 (photo 2).

Protocol for head posture

The protocol recommended by Solow (1994) was followed in parts 1 and 2. The patients were first asked to walk around and relax before standing in the cephalostat. Once positioned, the patients were then asked to walk on the spot and tilt the head forwards and backwards with decreasing amplitude until a natural head balance was reached and to look straight into her/his own eyes in a mirror, which was mounted on a foldaway door. The head-holder was then adjusted until the ear rods could be inserted into the ears and the radiograph was taken. The above procedure was repeated if the patient's head position changed during the adjustment of the ear rods. No occipital support was used.

The protocol for part 3 was simplified by omitting walking and head-tilting exercises, and the radiographer was allowed to interfere and repeat the procedure if the patient's head was clearly in an over- or under-extended position, i.e. not in NHP.

Subjects

All patients in need of cephalometric assessment regardless of gender, age, or severity of malocclusion were included in the study. All three groups subsequently included patients before, during, and after orthodontic treatment, as well as orthognathic patients (Orthogn) and those with cleft lip and palate (CLP; Table 1). There was no obvious difference in patient mix between the three groups. Prior approval for taking photographs was obtained.

Radiography

A Proline 2002 CC (Planmeca, Helsinki, Finland) was used to take all lateral head films. The film distance to the X-ray tube was fixed at 160 cm. The film distance to mid-sagittal plane of the patient's head was also fixed at 18 cm. The resulting magnification was 10 per cent. Films were exposed at 68–70 kV, 12 mAs, and a filter of 2.5 mm aluminium equivalent was used. For definition of the true vertical, a 0.5-mm lead wire suspending a weight and ear rods was used for identification of the transverse plane.

Photography

A Contax RTS Camera with a T*2.8/60 mm macro Lens (Carl Zeiss, Germany), was positioned on top of the X-ray machine. To avoid optical distortion, the actual facial profile of the patient covered only the central portion of the film. The aperture was fixed at 8 and the resulting depth of focus was 1.2 and 5 m. To achieve exposure times ranging from 1/60 to 1/125 of a second a fast black and white film (ASA 1600, Neopan Professional; Fuji Photo Film, Japan) was used. Exposure was triggered by cable release. After development of the negatives the area of interest was identified and was magnified to 20.3×25.4 cm (10×8 inches) so eventually the magnification was approximately 1:1 compared with the cephalogram.

Table 1 Patient mix for Parts 1, 2, and 3.

	Group 1	Group 2	Group 3
Female/male	10/6	8/4	24/13
Age/SD	13.9/2.54	14.3/2.35	13.5/3.73
CLP/Orthogn.	1/1	0/0	2/1

Measurements

The line connecting soft tissue nasion and subnasale (V-line), as well as the E-line of Ricketts (1957), was traced by one person for all photographs and cephalograms on two occasions (see Figure 1). Their respective angle to the true vertical was also measured twice and the mean of the two measurements was used. The difference of the two measurements was also used to assess intra-examiner error.

Statistical methods

The bias (or systematic error) between the replicate measurements for a method or between single measurements on two methods was assessed using either the paired *t*-test with a 5 per cent significance level or equivalently by constructing a 95 per cent confidence interval (CI) around the mean difference. Tests for either

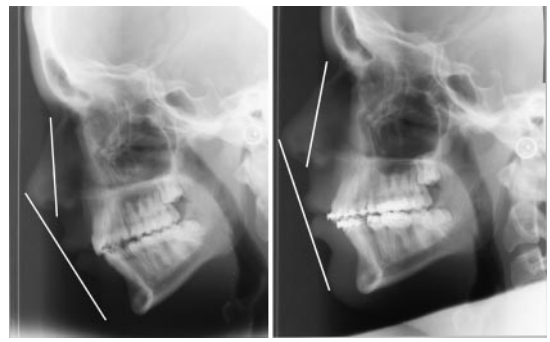

Figure 1 Two cephalograms of a patient before and during orthodontic treatment in natural head posture. This was the 'worst' case in terms of reproducibility. Demonstration of the two lines measured: soft tissue N/subnasale (V-line), and E-line.

association between the random error and the size of the measurement or differential method variability were performed using Pearson's correlation coefficient.

To assess the reproducibility of the tracings of the lateral cephalograms and the photographs and the agreement between photographs and cephalograms for the three series, the Dahlberg coefficient, the coefficients of reproducibility and reliability (i.e. Intraclass correlation coefficient, ICC), and the limits of agreement approach of Bland and Altman were used (see the Appendix for a detailed description of the statistical methods).

Results

Intra-examiner error

Dahlberg's coefficient for assessment of intra-examiner error for repeat tracings of all photographs and cephalographs was used. The results (Table 2) were similar to those found in other studies and no systematic errors in measurement of the angles were observed.

Table 2 Intra-examiner error (Dahlberg's coefficient) for the two reference lines in degrees.

	V-line	E-line
Part 1 X-ray 1	0.353	0.418
Part 1 X-ray 2	0.239	0.382
Part 2 photo 1	0.550	0.243
Part 2 X-ray	0.310	0.263
Part 2 photo 2	0.323	0.318
Part 3 photo 1	0.359	0.432
Part 3 X-ray	0.389	0.372
Part 3 photo 2	0.253	0.421

Part 1

This involved a retrospective assessment of the reproducibility of NHP after tracing the films of all patients who had at least two lateral cephalograms taken after the introduction of the new X-ray equipment in 1997 and March 1998. Sixteen patients' radiographs were assessed. The results for V- and E-lines are presented in Table 3, and the reproducibility plots of the difference between repeated angle measurements and their means are shown in Figure 2. There was no obvious tendency for the differences to widen or narrow as the angle increased (Figure 2 and Table 3).

The 95 per cent CI for the systematic bias contains zero for both the V- and E-lines (95 per cent CIs –3.92 and 0.29 degrees, and –3.88 and 0.88 degrees). Thus there was no statistical evidence of a 'drift' having occurred. However, these CIs were sufficiently wide to indicate an uncertainty in the level of bias. The Dahlberg coefficients for the V- and E-lines were unexpectedly high at 2.99 and 3.24 degrees. The reproducibility coefficients were 8.47 and 9.17 degrees, respectively.

Part 2

Method agreement was assessed between cephalograms and photographs, as well as reproducibility of the photographs. The results

Table 3 Statistical comparisons for V- and E-lines of repeated cephalographs for part 1 of the study.

Methods	Systematic bias with 95% confidence interval	Intra-class correlation	Variance components between subjects and components	Dahlberg's coefficient (residual SD)	*Limits of reproducibility/ **Limits of agreement	Test for differential method variability
Ceph1 versus Ceph2 (V-line)	–1.81 (–3.92, 0.29)	0.38	5.61 and 8.97	2.99	±8.47*	$P = 0.57$
Ceph1 versus Ceph2 (E-line)	–1.5 (–3.88, 0.88)	0.57	14.11 and 10.5	3.24	±9.17*	$P = 0.37$

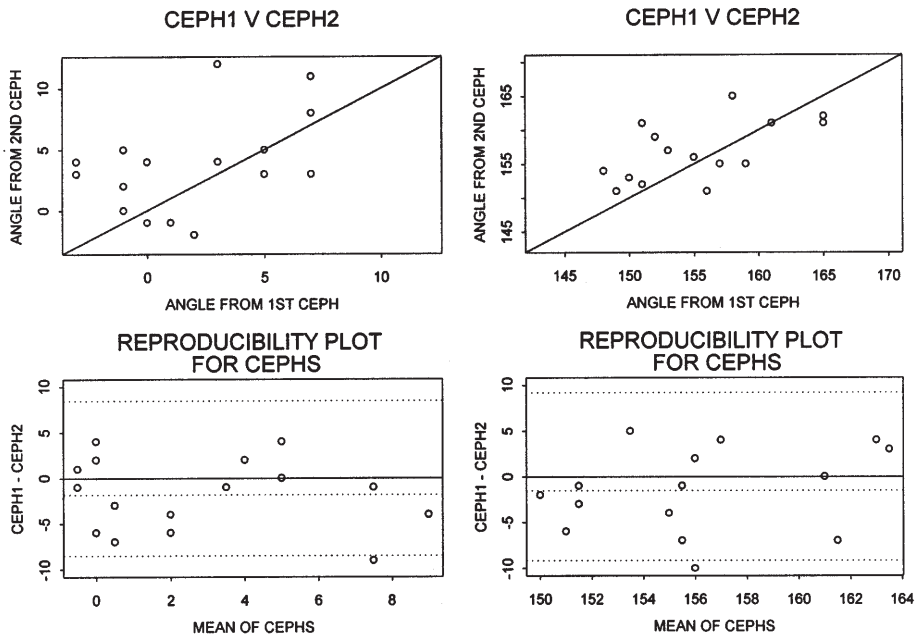

Figure 2 The Bland–Altman reproducibility plots for the V- (left) and E-lines (right) for part 1 of the study. Both lines for the two lateral head radiographs show poor reproducibility. There was no significant bias.

for the 12 patients in part 2 are shown in Tables 4 and 5. The tests for systematic bias were found to be statistically non-significant (95 per cent CIs for the V-line: –3.21 and 1.79 degrees; and the E-line: –2.58 and 2.08 degrees). The Dahlberg coefficients for two photographs for the V- and E-lines were 2.71 and 2.49 degrees, respectively. The corresponding ICCs for the photographic method were 0.82 and 0.89. The reproducibility coefficients for the photographic method were 7.66 and 7.05 degrees. The reproducibility plots are shown in Figures 3 and 4. Method agreement between photograph 1 and the cephalograph showed the best result of this

Table 4 Statistical comparisons for V-line for two photographs, photograph 1–cephalograph and cephalograph–photograph 2 for the second part of the study.

Methods	Systematic bias with 95% confidence interval	Intra-class correlation	Variance components between subjects and components	Dahlberg's coefficient (residual SD)	*Limits of reproducibility/ **Limits of agreement	Test for differential method variability
Photo1 versus Photo2	–0.71 (–3.21, 1.79)	0.82	30.59 and 6.82	2.71	±7.66*	$P = 0.23$
Photo1 versus Ceph	–0.54 (–1.79, 0.71)	0.94	30.02 and 1.80	1.39	(–4.48, 3.40)**	$P = 0.47$
Photo2 versus Ceph	0.17 (–1.80, 2.13)	0.89	36.17 and 4.41	2.1	(–6.02, 6.35)**	$P = 0.31$

Table 5 Statistical comparisons for E-line for two photographs, photograph 1–cephalograph and cephalograph–photograph 2 for the second part of the study.

Methods	Systematic bias with 95% confidence interval	Intra-class correlation	Variance components between subjects and components	Dahlberg's coefficient (residual SD)	*Limits of reproducibility/ **Limits of agreement	Test for differential method variability
Photo1 versus Photo2	–0.25 (–2.58, 2.08)	0.89	44.79 and 5.74	2.49	±7.05*	$P = 0.15$
Photo1 versus Ceph	–0.67 (–1.69, 0.36)	0.97	40.25 and 1.19	1.19	(–3.90, 2.56)**	$P = 0.51$
Photo2 versus Ceph	–0.42 (–2.69, 1.86)	0.89	47.07 and 5.88	2.44	(–7.58, 6.74)**	$P = 0.25$

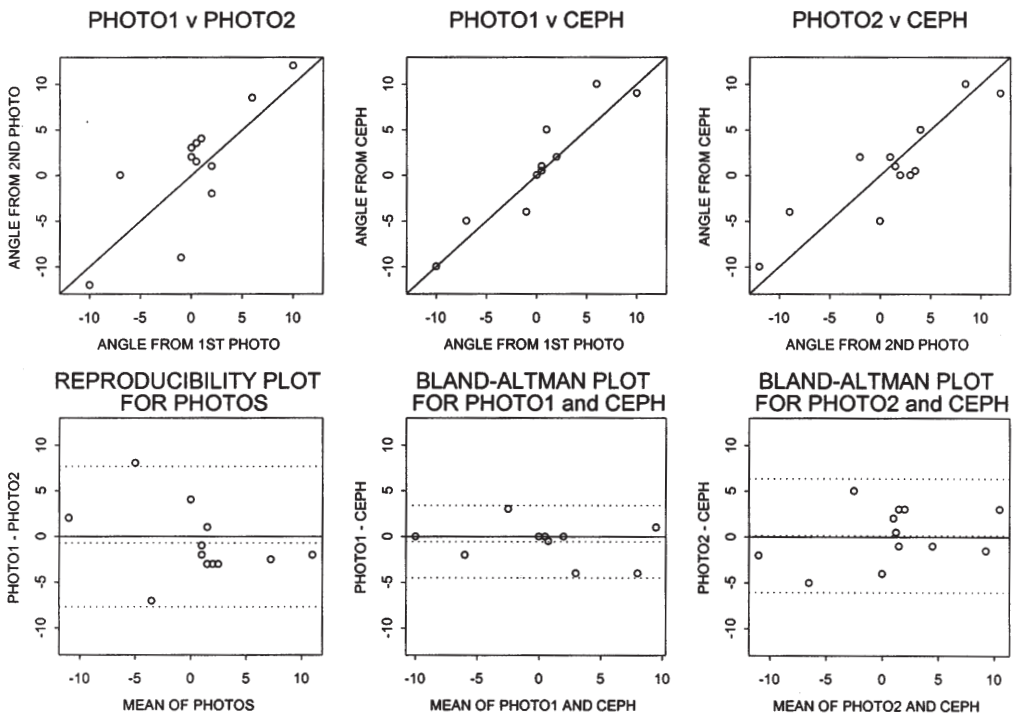

Figure 3 V-line second series; data deviates more from mean between photograph 1 and 2 and cephalograph and photo 2, than between photo 1 and cephalograph. From this it was concluded that the photographic and cephalometric methods are compatible and that reproducibility between two photographs was poor probably due to implementation problems with the protocol for NHP.

series in terms of Dahlberg coefficient [1.39 (V-line) and 1.19 degrees (E-line)] and limits of agreement [–4.48/3.4 (V-line) and –3.90/2.56 (E-line)]. This was probably the result of the short time lapse between the two recordings.

Method agreement between the cephalograph and photograph taken at time 2 was not as good [–6.02/6.35 (V-line) and –7.58/6.74 (E-line)]; Dahlberg coefficient 2.10 (V-line) and 2.44 (E-line) degrees. This corresponded with the

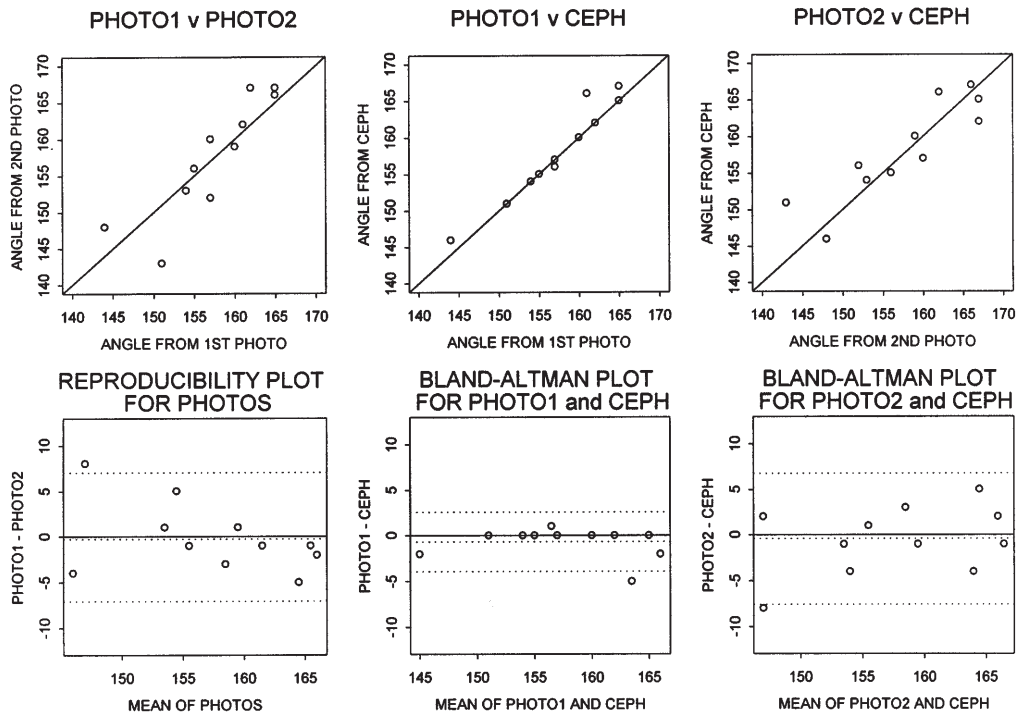

Figure 4 E-line second series; data shown here substantiated findings of data in Figure 3.

poor reproducibility between the two photographs and was thought to be due to the time lapse between the recordings (approximately 10 minutes).

Part 3

This investigated the reproducibility of NHP using the photographic method with a modified (essentially simplified) protocol. The results for the 37 patients are shown in Tables 6 and 7. The Dahlberg coefficients comparing the two photographs were 1.41 and 1.47 degrees for the V- and E-lines, respectively. The corresponding coefficients of reliability were 0.93 and 0.84. The estimated reproducibility coefficients were 4.00 and 4.17 degrees, respectively. The paired *t*-tests for assessing systematic bias were all statistically non-significant. With a change in protocol from part 2 to part 3, reproducibility for the photographs was much improved; the agreement between cephalograms and the initial photographs (time 1), and that between cephalograms and photographs taken after the cephalographs (time 2) were determined. Table 6 shows the results after assessing the agreement between cephalograms and photographs at times 1 and 2 when applied to the V-line. The Dahlberg coefficients were found to be 1.33 and 1.88 degrees for the comparisons between the cephalograms and the initial photographs, and between the cephalograms and the subsequent photographs. The corresponding coefficients of reliability were 0.94 and 0.88. The limits of agreement for the two comparisons were –4.19 and 3.19 degrees and –5.39 and 5.36 degrees, respectively. The Bland–Altman plots are shown in Figures 5 and 6. Table 7 shows the corresponding results for the E-line.

When applied both to the V- and E-lines, the 95 per cent limits of agreement for the cephalographs and the initial photographs were narrower than, and lie inside, the limits for the cephalographs and the subsequent photographs.

Table 6 Statistical comparisons for V-line for two photographs, photograph 1–cephalograph and cephalograph–photograph 2 for the third part of the study.

Methods	Systematic bias with 95% confidence interval	Intra-class correlation	Variance components between subjects and components	Dahlberg's coefficient (residual SD)	*Limits of reproducibility/ **Limits of agreement	Test for differential method variability
Photo1 versus Photo2	–0.49 (–1.14, 0.17)	0.93	27.02 and 2.00	1.41	±4.00*	$P = 0.09$
Photo1 versus Ceph	–0.5 (–1.12, 0.12)	0.94	24.84 and 1.70	1.33	(–4.19, 3.19)**	$P = 0.77$
Photo2 versus Ceph	–0.01 (–0.91, 0.88)	0.88	25.84 and 3.61	1.88	(–5.39, 5.36)**	$P = 0.30$

Table 7 Statistical comparisons for E-line for two photographs, photograph 1–cephalograph and cephalograph–photograph 2 for the third part of the study.

Methods	Systematic bias with 95% confidence interval	Intra-class correlation	Variance components between subjects and components	Dahlberg's coefficient (residual SD)	*Limits of reproducibility/ **Limits of agreement	Test for differential method variability
Photo1 versus Photo2	–0.28 (–0.98, 0.41)	0.84	31.89 and 5.95	1.47	±4.17*	$P = 0.13$
Photo1 versus Ceph	–0.01 (–0.54, 0.51)	0.9	32.25 and 3.52	1.1	(–3.16, 3.14)**	$P = 0.53$
Photo2 versus Ceph	0.27 (–0.45, 0.99)	0.85	32.65 and 5.96	1.53	(–4.07, 4.61)**	$P = 0.32$

Discussion

The Dahlberg coefficient of NHP, 1 year after its introduction, was not as reliable as those reported by other investigators. These were 2.99 and 3.24 degrees for the V- and E-lines, respectively. However there also appeared to be a discrepancy between the values of the Dahlberg coefficient and the clinical impression of the reproducibility of NHP; the latter appearing even more variable clinically. It thus became appropriate to investigate the suitability of Dahlberg's coefficient in assessing reproducibility of NHP. It was concluded that reproducibility of NHP in part 1 of this study was poor, either because the protocol for achieving head posture itself was faulty, or its implementation was inadequate. Although the variability of facial photographs for orthodontics has been investigated previously (Farkas *et al*., 1980; Bengal, 1985; Gordon and Wander, 1987; Claman *et al*., 1990; Bishara *et al*., 1995; Strauss *et al*., 1997), there are only a few studies investigating method agreement between cephalography and photography (Tsang and Cooke, 1999). The study by Benson and Richmond (1997) concluded that the photographic technique showed clinically acceptable validity and reproducibility and that such methods would be appropriate for multi-centre trials.

In part 2, method agreement between cephalographs and photographs taken at the same time was found to be better than the reproducibility of two photographs taken with a time interval of approximately 10 minutes. Therefore, there was a greater impact on the reliability of the results from positioning the patient twice (after a 10-minute time delay) than the comparing

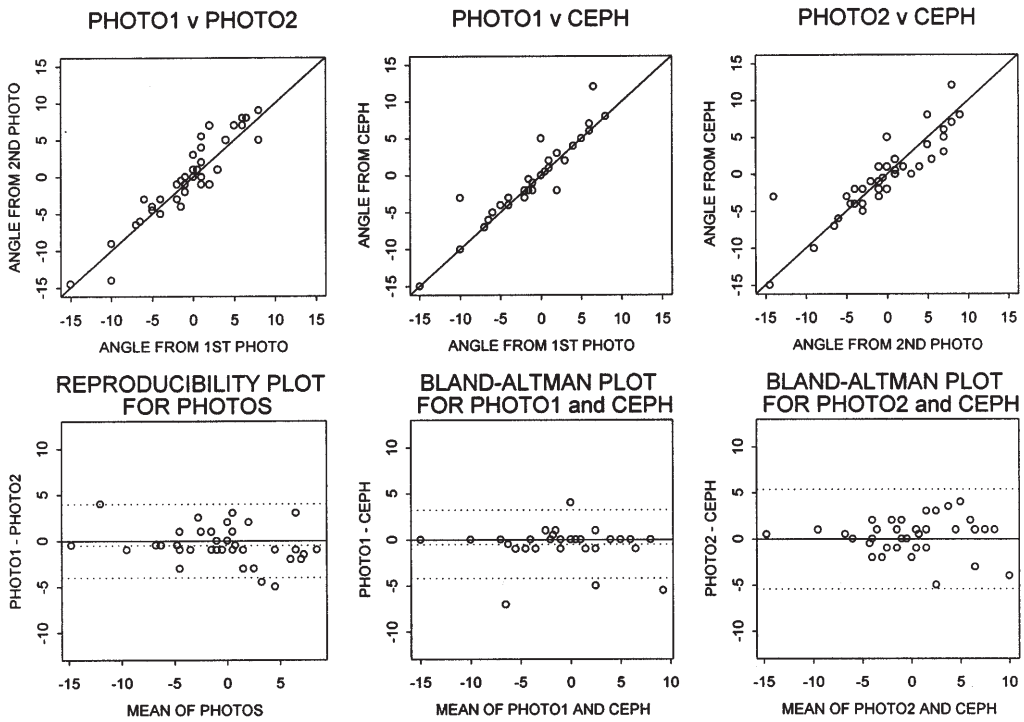

Figure 5 V-line third series using a different protocol for NHP; improved reproducibility between two photographs taken at a time interval of approximately 10 minutes; data show less deviation from mean than in the second series.

agreement between radiographs and photographs. Thus the protocol was changed.

In part 3, the results for reproducibility of NHP improved considerably. Although Siersbæk–Nielsen and Solow (1982) did 'not accept any operator unless they had recorded at least 20 patients', no study has yet addressed the problem of staff training. This is dependent on the protocol used and no protocol has yet been universally adopted (Cooke and Wei, 1998a). In earlier literature, comparative studies regarding the reproducibility of self-balance and mirror position showed that the mirror position seemed to be superior (Solow and Tallgren, 1971; Cooke and Wei, 1988b). However, Proffit and White (1991) considered the use of a window leading on to the horizon as an acceptable alternative. The difference of reproducibility for sitting and standing positions has also been investigated previously, but the results remain controversial and Moorrees (1994) considered either position appropriate. There is also controversy about the use of a nasal bridge-holder (Luyk *et al*., 1986; Viazis, 1991) and ear rods. The use of ear rods seems to have a greater influence on the quality of the films rather than the reproducibility of NHP (Cooke and Wei, 1988b). However Moorrees (1995) preferred the use of the midline-ruler to keep the head truly vertical rather than ear rods. However, no previous study has further investigated the various types of protocols for the mirror position and their influence on the outcome of repeatability/reproducibility, and fairly minor changes in protocol may have significant effects on reproducibility.

Reproducibility of NHP for two photographs taken with a time interval of approximately 10 minutes, and cephalographs and photographs for part 3, after protocol simplification, were similar to those reported by others (Bjerin, 1957; Moorrees and Kean, 1958; Solow and Tallgren, 1971; Siersbæk-Nielsen and Solow, 1982).

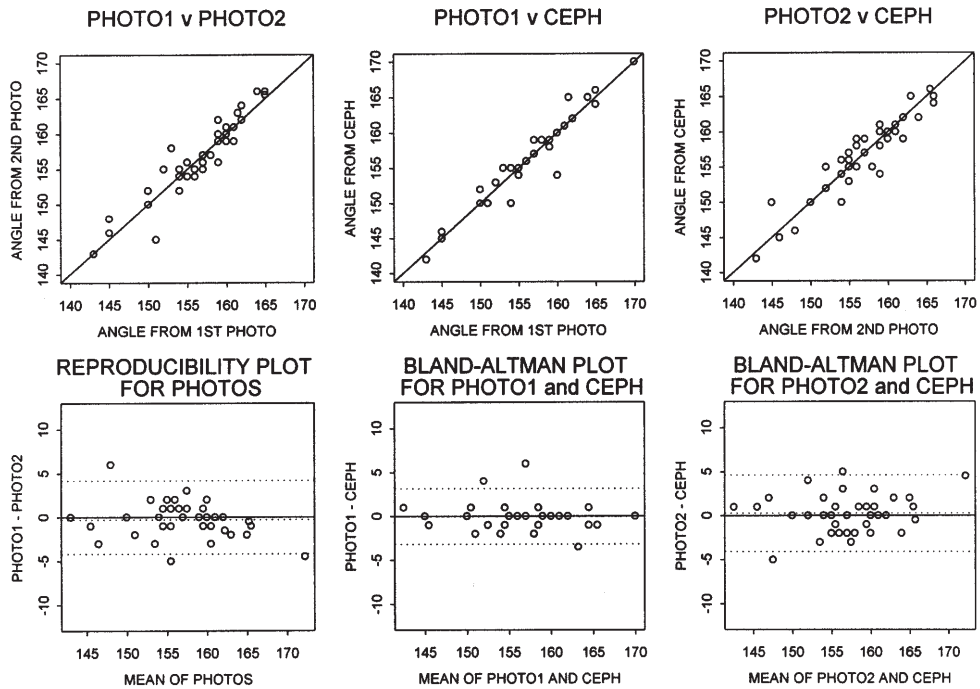

Figure 6 E-line third series; standard deviation for E-line also substantially improved.

Figure 7 Initial photograph, cephalogram, and second photograph. The differences between the angular measurements for this patient were within the 95 per cent CI. The difference in head posture is clearly visible.

Figure 8 Initial photograph, cephalogram, and second photograph of another patient. The differences between the angular measurements for this patient were also within the 95 per cent CI.

However, although Dahlberg's coefficient was approximately 1.5 degrees, this camouflaged the true variability of the results. Figures 7 and 8 show two photographs and the cephalograph of two patients; although patient data were within the 95 per cent reproducibility limits, the difference in head posture between the two photographs is clearly visible. It appears that such differences in reproducibility of NHP are clinically significant and their usefulness in the decision-making process of clinical orthodontics remains to be established. Figure 1 shows the two radiographs of the 'worst offender' for NHP in series 1. Bjerin (1957) as well as Solow and Tallgren (1971) argued that the variability of intra-cranial reference lines between subjects is greater than the variability of NHP over time.

However, comparison of variability of intra-cranial reference lines of different patients to the 'true horizontal' uses cross-sectional data. Reproducibility of NHP, however, makes use of longitudinal data. To our knowledge, no study has yet compared the influence of the two measurements on the variability of intra-cranial reference lines. The only way to judge whether the use of one method (NHP versus use of intra-cranial reference lines) is advantageous would be to show that the use of one particular method makes a clinically significant difference in the decision-making process.

Assessment of reproducibility for NHP

The Dahlberg coefficient and the reproducibility coefficient measure the variability of the random measurement. Dahlberg's coefficient is the standard deviation of the random measurement error, whilst the reproducibility coefficient is approximately 2.83 times the standard deviation of this random error.

By definition one would expect 95 per cent of the absolute differences between replicates to lie below the reproducibility coefficient, and only 52 per cent of the absolute differences between replicates to lie below the Dahlberg coefficient. This follows from assuming a normal distribution for the differences between replicates, together with the property that ±2 standard deviations of a standard normal distribution contain approximately 95 per cent of the probability density. By definition, 2.83 standard deviations of the measurement error (i.e. the reproducibility coefficient) are equivalent to 2 standard deviations of a standard normal distribution. Thus, by division, 1 standard deviation of the measurement error (i.e. the Dahlberg coefficient) is equivalent to 0.71 standard deviations of a standard normal distribution; and 52 per cent of the probability density is contained within ±0.71 standard deviations. Note also that the interval $(-D, D)$ can be interpreted, approximately, as the inter-quartile range of the differences between replicate measurements. Thus the Dahlberg coefficient does not give a wide enough interval to allow a sufficient clinical assessment of the method to be undertaken. This is generally not known and has obvious implications for the measurement of other orthodontic parameters.

The ICC (or coefficient of reliability) relates the variability of the measurement error to the total variability of the population being studied. It is defined as the proportion of the total variance attributable to the subjects. A value 'close' to unity has been taken to indicate a 'good' level of reproducibility. The ICC will be close to unity (i.e. indicating good reproducibility) in samples where a small value for the variance of the measurement error is estimated. However the ICC may also be close to unity in samples where the total variance is large relative to the variance of the random measurement error, irrespective of the actual magnitude of the measurement error. Therefore an ICC that is close to unity does not necessarily indicate high precision or low level of measurement error. In fact, the ICC is not a fixed characteristic of the method being employed, as it is dependent not only on the nature of the method (i.e. measurement error), but also on the variability of the sampled population of subjects being measured. For example, the inclusion of CLP patients with 'normal' patients may increase the variability between subjects and therefore improve the ICC relative to what might have been obtained separately for either group. However, the present sample was too small to assess the subgroups separately in a meaningful way. Thus the ICC has no absolute meaning and can be interpreted only as a relative measure of reproducibility, and thus this makes it difficult to say what are acceptable levels of reproducibility for different situations. However, this does not necessarily mean that it is not a valid measure of reproducibility, as it may still be useful in assessing the performance of a particular method in a given clinical setting. The conclusions obtained from using the coefficient of reliability (or ICC) may be different from those obtained from using the Dahlberg or reproducibility coefficients. For example, in the first series, the coefficients of reliability for the cephalogram method, in their application to the V- and E-lines, were 0.38 and 0.57, respectively. This would indicate that the reproducibility of NHP for the cephalogram method is better in its application

to the E- than V-line. However, this conclusion is contradicted by the other measures of reproducibility that are based purely on measurement error, namely the reproducibility coefficient and the Dahlberg coefficient. For instance, the Dahlberg coefficients for the first series were 2.99 and 3.24 degrees for the V- and E-lines, respectively. Therefore it would be misleading to report the ICC because, compared with that for the V-line, the larger measurement error for the E-line would be hidden by its substantially larger between-subject variability. Also, the coefficients of reliability (ICC) in the second series indicate a good level of reproducibility, whereas the reproducibility coefficients indicate the converse.

Conclusions

1. The initial reproducibility of NHP after the introduction of new X-ray equipment in 1997 was poor (Dahlberg's coefficient of 2.99 degrees for cephalograms and 2.71 degrees for photographs for the V-line); however, photographic reproducibility subsequently improved by simplifying the protocol (Dahlberg's coefficient of 1.41 degrees for the V-line).
2. For assessing reproducibility, the reproducibility coefficient and its corresponding graphical representation were most appropriate. This gives a wide enough interval (95 per cent) to allow sufficient clinical assessment of the method as compared with the Dahlberg coefficient, which represents a 52 per cent interval. For assessing the agreement between two methods, the limits of agreement approach of Bland and Altman was most appropriate because, unlike the Dahlberg and reproducibility coefficients, it correctly incorporates the systematic bias between the methods.
3. The protocol for NHP itself appears to have an influence on reproducibility. There is some evidence to suggest that the success of a certain protocol is operator dependent. However, in a clinical situation it seems to be difficult to improve reproducibility beyond 1.4 degrees, even when only one dedicated radiographer takes the cephalographs/photographs. This means that the limit of reproducibility is ±4.0 degrees. Overall there was a small but noticeable change in head posture within the time interval in which the records were taken; this trend differed for parts 2 and 3 of the study.
4. Photography is useful for the training of radiographers during the introduction of NHP in cephalometry. Reproducibility can be audited after a short time interval without unnecessary radiation exposure of patients.

Address for correspondence

Dr D. Bister
Department of Orthodontics and Paediatric Dentistry
Floor 22 Guy's Tower
Guy's Hospital
London SE1 9RT, UK

Acknowledgements

The authors would like to thank Ranji Uthayashanker, Department of Orthodontics, Queen Mary's Hospital, Roehampton, for helping with this project by taking the photographs and radiographs.

References

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

Bengal W 1985 Standardization in dental photography. International Dental Journal 35: 210–217

Benson P E, Richmond S 1997 A critical appraisal of measurement of the soft tissue outline using photographs and video. European Journal of Orthodontics 19: 397–409

Bishara S E, Cummins D M, Jorgensen G J, Jakobsen J R 1995 A computer assisted photogrammetric analysis of soft tissue changes after orthodontic treatment. Part I: methodology and reliability. American Journal of Orthodontics and Dentofacial Orthopedics 107: 633–639

Bjerin R 1957 A comparison between the Frankfort horizontal and the sella turcica-nasion line as reference planes in cephalometric analysis. Acta Odontologica Scandinavica 15: 1–12

Bland J M, Altman D G 1986 Statistical methods for assessing agreement between two methods of clinical measurement. The Lancet 2: 307–310

Bland J M, Altman D G 1990 A note on the use of the intraclass correlation coefficient in the evaluation of

agreement between two methods of measurement. Computers in Biology and Medicine 20: 337–340

Claman L, Patton D, Rashid R 1990 Standardized portrait photography for dental patients. American Journal of Orthodontics and Dentofacial Orthopedics 98: 197–204

Cooke M S, Wei S H Y 1988a A summary five factor cephalometric analysis based on natural head posture and the true horizontal. American Journal of Orthodontics and Dentofacial Orthopedics 93: 213–223

Cooke M S, Wei S H Y 1988b The reproducibility of natural head posture: a methodological study. American Journal of Orthodontics and Dentofacial Orthopedics 93: 280–288

Dahlberg G 1940 Statistical methods for medical and biological students. George Allen & Unwin Ltd, London, p. 124

Downs W B 1956 Analysis of the dento-facial profile. Angle Orthodontist 4: 191–212

Farkas L, Bryson W, Klotz J 1980 Is photogrammetry of the face reliable? Plastic and Reconstructive Surgery 66: 346–355

Fleiss J L, 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

Gordon P, Wander P 1987 Techniques for dental photography. British Dental Journal 162: 307–316

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

Houston W J B 1991 Bases for the analysis of cephalometric radiographs; intra-cranial reference structures or natural head position. Proceedings of the Finnish Dental Society 87: 43–49

International Standards Organization 1994 (ISO 5725) Trueness and precision of test methods. Part 1: British Standards Institution 13:24:18 2000, p. 3

Luyk N H, Whitfield P H, Ward-Booth R P, Williams E D 1986 The reproducibility of the natural head position in lateral cephalometric radiographs. British Journal of Oral and Maxillofacial Surgery 24: 357–366

Moorrees C F A 1994 Natural head position—a revival. American Journal of Orthodontics and Dentofacial Orthopedics 105: 512–513

Moorrees C F A 1995 Natural head position: The key to cephalometry. In: Jacobson A (ed.) Radiographic cephalometry. Quintessence, Chicago, pp. 175–184

Moorrees C F A, Kean M R 1958 Natural head position, a basic consideration in the interpretation of cephalometric radiographs. American Journal of Physiology and Anthropology 16: 213–234

Peng L, Cooke M S 1999 Fifteen-year reproducibility of natural head posture: a longitudinal study. American Journal of Orthodontics and Dentofacial Orthopedics 116: 82–85

Proffit W R, White R P (eds) 1991 In: Surgical orthodontic treatment. Mosby, St Louis, pp. 96–141

Ricketts R M 1957 Planning treatment on the basis of the facial pattern and an estimate of its growth. Angle Orthodontist 27: 14–37

Siersbæk-Nielsen, Solow B 1982 Intra- and interexaminer variability in head posture recorded by dental auxiliaries. American Journal of Orthodontics 82: 50–57

Solow B 1994 Cervical and cranio-cervical posture in realation to craniofacial growth. Acta Medica Romania 32: 232–249

Solow B, Tallgren A 1971 Natural head position in standing subjects. Acta Odontologica Scandinavica 29: 591–607

Strauss R A, Weis B D, Lindauer S J, Rebellato J, Isaacson R J 1997 Variability of facial photographs for use in treatment planning for orthodontics and orthognathic surgery. International Journal of Adult Orthodontics and Orthognathic Surgery 12: 197–203

Student M 1908 The probable error of the mean. Biometrica 6: 1–25

Tsang K H S, Cooke M S 1999 Comparison of cephalometric analysis using a non-radiographic sonic digitizer (DigiGraph™ Workstation) with conventional radiography. European Journal of Orthodontics 21: 1–13

Viazis A D 1991 A cephalometric analysis based on natural head position. Journal of Clinical Orthodontics 25: 172–181

Appendix

Dahlberg's (1940) coefficient

In the orthodontic literature, where typically two measurements are compared per subject, the Dahlberg's coefficient has found widespread use and is defined as D,

$$D = \sqrt{\frac{\sum_i d_i^2}{2n}}$$

where n is the number of subjects and d_i is the difference between repeated measurements on the ith subject.

The Dahlberg coefficient has typically been used in two situations:

1. when two measurements are compared using the same method (i.e. repeatability/reproducibility);
2. when measurements from two methods have been compared (method agreement).

In the case of repeatability/reproducibility studies, the Dahlberg coefficient is equivalent to s_e, the standard deviation of the measurement error of one method. That is,

$$D = s_e$$

Thus for more than two readings per subject using the same method, the Dahlberg coefficient should be taken to be s_e, where s_e is calculated as described below. The relationship between the Dahlberg coefficient and the reproducibility coefficient (see ISO definition) is easily obtained. It is:

$$i_r = 2\sqrt{2}D \qquad \text{or} \qquad i_r = 2.83D$$

Therefore, using the normal distribution assumption for the differences, the Dahlberg's coefficient has an interpretation in terms of the absolute difference between two replicates, namely, that just 52 per cent of the absolute differences are expected to lie below D.

International Standards Organization (1994) definition of reproducibility

The ISO definition of reproducibility is expressed in terms of the reproducibility coefficient i_r, where i_r is the value below which 95 per cent of the absolute differences would be expected to lie. This is derived from the same assumption that the differences between pairs of replicates are observed from a normal distribution. Here

$$i_r = 2\sqrt{2}s_e$$

Intra-class correlation coefficient (ICC)

The ICC, also known as the coefficient of reliability (Houston, 1983), is defined as the proportion of the total variance attributable to the subjects. Denoted by r, it is given by

$$r = \frac{s_b^2}{(s_b^2 + s_e^2)} = 1 - \frac{s_e^2}{(s_b^2 + s_e^2)}$$

where s_b^2 and s_e^2 are, respectively, the between-subjects and within-subject (i.e. measurement error) variance components, obtained either from an analysis of variance model for balanced data (i.e. equal number of replicates per subject), or from restricted maximum likelihood for unbalanced data. Equivalently, it is defined as the average Pearson's correlation coefficient amongst all possible orderings resulting from interchanging the replicate measurements of any of the subjects (Bland and Altman, 1990). A value of r 'close' to unity has been taken to indicate a 'good' level of reproducibility.

Graphical presentation of reproducibility

A graphical approach to presenting reproducibility consists of plotting the difference between the two repeated readings against their mean for each subject. The difference gives a measure of the within-subject variability, whilst the mean gives a measure of the size or magnitude of the measurement. Limits within which 95 per cent (or 52 per cent) of the differences are expected to lie appear on the graph as horizontal lines. These are positioned at i_r and at $-i_r$ (or D and $-D$). These limits are centred about the zero difference, as no systematic bias is expected between repeated readings.

Graphical presentation of method agreement

A similar approach to presenting method agreement has been proposed by Altman and Bland (1983) and Bland and Altman (1986). Here each subject's difference in readings between methods is plotted against their mean reading. Limits within which 95 per cent of the differences are expected to lie appear on the graph as horizontal lines. These are positioned at $\bar{d} + 2sd$ and at $\bar{d} - 2sd$, where $\bar{d}$ and sd are the mean and standard deviation of the differences, respectively. The mean bias, $\bar{d}$, also appears on the graph as a horizontal line. Note that these limits are centred about $\bar{d}$ not 0 as in the approach for reproducibility.