Cephalometric Reliability

A Full ANOVA Model for the Estimation of True and Error Variance

Peter H. Buschang
Richard Tanguay
Arto Demirjian

A detailed description of sampling designs for assessing the reliability of cephalometric measurements, emphasing distinctions between 1) true and observed variance, 2) random and systemic components of variance, and 3) complete and minimal models for evaluating measurement error.

KEY WORDS: • Cephalometrics • Error • Reliability • • Statistics • Variance

Since its introduction by Broadbent in 1931, cephalometrics has become an integral part of orthodontics, including research, teaching, and clinical practice. Considerable progress has been achieved over the years in standardizing equipment and techniques to minimize the effects of distortion and magnification, the *errors of projection.* Much more progress is needed in controlling the *errors of identification.*

Carlsson (1967) has shown that the greatest source of error lies in locating cephalometric landmarks. Technical errors exist both within observers (Björk 1947, Solow 1966,Richardson 1966; Starbrun and Danielsen, 1982) and between observers (Mattila and Haataja, 1968, Vinkka and Koski, Baumrind and Frantz, 1971). In lieu of their elimination, errors must be quantified in order to substantiate the validity of cephalometric research and clinical applications.

To this end, determining the reliability of landmark identification remains a prerequisite for the meaningful manipulation and interpretation of cephalometric data. However, few studies adequately document the reliability of their methodology. The traditional approach to evaluate error variance has been *method error.*

Dr. Buschang is a research associate, Section d'orthodontie, département de santé buccale, and Centre de recherche sur la croissance humaine, Université de Montréal. He holds a Ph.D. from the University of Texas and was a postdoctoral research fellow with the Department of Orthodontics, School of Dental Medicine, University of Connecticut.

Author Address:

Dr. P. H. Buschang
Faculté de médicine dentaire
Département de santé buccale
Case postale 6128, Succursale "A"
Montréal, P.Q. H3C 3J7
CANADA

Mr. Tanguay is a biostatistician at the Centre de recherche sur la croissance humaine and holds an M.S. in biostatistics from the Université de Montréal.

Dr. Demirjian is director, Centre de recherche sur la croissance humaine and Professor of Anatomy, Université de Montréal. He holds D.D.S. degrees from the Universities of Istanbul and Montréal, and a M.S. in anatomy from the University of Toronto.

Depending on the design of the analysis, method error alone could produce inaccurate results (Buschang et al. 1984). Moreover, comparisons of error variance are difficult to interpret due to the lack of standardization. In contrast, the coefficient of reliability that is presented here, because it is a *relative measure of error,* is immediately interpretable and comparable.

— Reliability Defined —

Two kinds of errors affect cephalometric measurements — random and nonrandom. Random error refers to the various chance factors included in the description of a measurement. The measurement process itself inevitably introduces the unsystematic effects of random error. Nonrandom error introduces a systematic biasing effect which pertains directly to the validity of the measurement.

Each observed cephalometric measurement can be broken down to its true value and an error component. The true value is a theoretical construct, an average that would be obtained if a landmark were remeasured an infinite number of times.

Reliability is ascertained from repeated measurements across individuals, and depends on the variance components associated with the observed measure. Since the observed value is equal to the true value plus error variance,

$$\text{VAR(observed)} = \text{VAR(true)} + \text{VAR(error)} \qquad (1)$$

it follows that —

Reliability of the measurement technique expresses the *ratio of true variance on true variance plus error variance,* or —

$$\text{Reliability} = \frac{\text{Var(true)}}{\text{Var(true)} + \text{Var(error)}} \qquad (2)$$

Reliability estimates provide the proportion of observed variance of a measure which can be treated directly as true variance. For example, a reliability of 0.90 indicates that 90% of the observed variance is true, while 10% is error.

The residual variance, error which is attributable to the methodology, may include both random and systematic components.

The estimate of true variance provides a basis for comparison; it represents the actual within-individual variance. True variance is central to reliability estimation.

The variance components necessary to calculate the coefficient of reliability may be obtained by *analysis of variance* (ANOVA), using a model with the appropriate adjustments for the random and fixed factors present in the design (Healy 1958, Kirk 1968, Weiner 1971). Assuming a minimal design, as with one set of replicate measurements from a homogeneous sample, evaluated by a single operator, ANOVA provides accurate estimates of the true and error variance with —

$$\text{Y} = \text{Mean} + \text{T} + \text{X} + \text{TX} \qquad (3)$$

where T = systematic tracing variance
X = true variance (within), and
TX = random tracing variance (residual).

Most research and clinical applications include additional sources of variation which must be evaluated to obtain accurate estimates of variance. If age, for example, explains some of the variance for a particular set of cephalometric data, its inclusion into the model will affect the estimation of true variance, which could alter a measurement's reliability. Interaction between the age and tracing effects could change the error variance.

Including age (A) as a component of variation (nested in subjects), the ANOVA model should be —

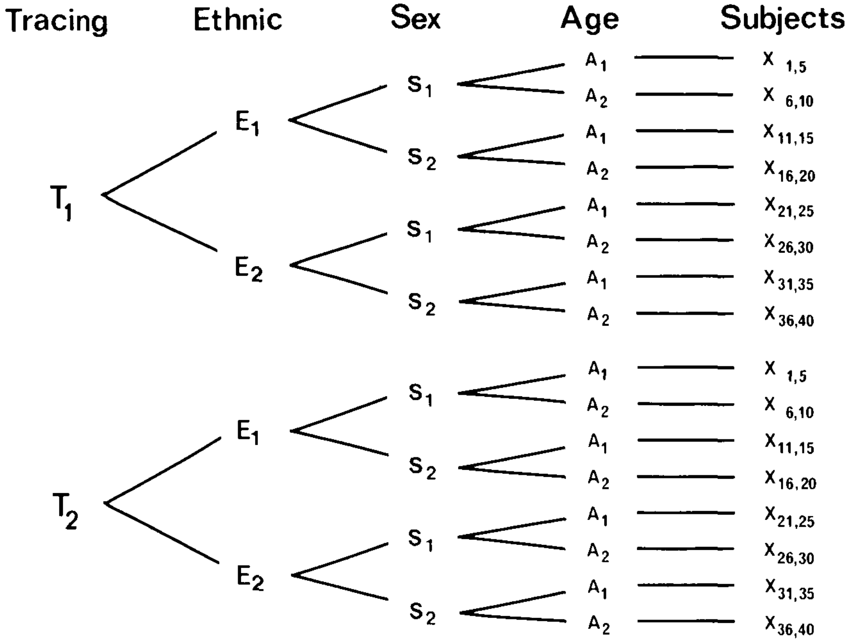

Fig. 1 Sampling design (full model) for the evaluation of duplicate tracings, with 40 subjects nested in two ethnic, two sex, and two age categories.

$$Y = \text{Mean} + A + T + AT + X(A) + TX(A) \qquad (4)$$

The minimal model (3) includes age (A) and the interaction between age and tracing (AT) in the true and random variance components, respectively. The design becomes more complex with each additional source of variation. Although complete models are complex, they are essential to accurate estimation of true and error variance.

A full model also allows assessment of global reliability, which combines the error variance for multiple tracings and/or operators. Global reliability is the best means for evaluating the repeatability of cephalometric landmark identification.

— Materials and Methods —

To demonstrate the application of a full ANOVA model, this report evaluates the intraobserver technical reliability of eight standard landmarks derived from cephalometric tracings of lateral head films. The landmarks chosen to illustrate the methodology include the horizontal and vertical perspectives of Bolton point (Bo), gonion (Go), pogonion (Po), supramentale (SM), subspinale (SS), and orbitale (Or), lower central incisor tip (LI), upper central incisor tip (UI).

The full model, which applies to a variety of research and clinical situations, is compared to a minimal model (3), which includes only tracing and random effects.

The design (Fig. 1) includes duplicate tracings (T) for the two age (A) categories 10 and 14 years, of both sexes (S), and two ethnic groups (E) (French-Canadian children from Montreal and Anglo-Saxon children from Burlington, Ontario), for each of the five subjects (X).

Subjects are nested in ethnicity, sex and age X(ESA). Moreover, the model is mixed; T and X are random, while E, S, A, are fixed. The sample includes a total of 80 tracings.

Due to the potential interactions of the main effects, the complete model includes a total of 18 terms for which the variance components have been calculated (Table 1).

— Results —

Table 2 presents the tracing reliability for the horizontal and vertical perspectives of the eight landmarks.

Global reliability reflects the combined effects of systematic and random tracing errors. It estimates the proportion of observed variance which is true variance.

Systematic reliability is consistently high. For Bolton point, the difference between the minimal and full models may be attributed to rounding errors. With the exception of Bolton point (horizontal), the reliability associated with random errors is also high.

Significantly, the minimal and full models may yield different estimates. For

Table 1

Full Analysis of Variance (ANOVA) Model
Showing Sources, Degrees of Freedom, and Expected Mean Squares

Term	Source	Degrees of Freedom	Expected Mean Square
1	Mean	1	80(1) + 40(3) + 2(16) + (18)
2	Ethnicity	1	40(2) + 20(6) + 2(16) + (18)
3	Tracings	1	40(3) + (18)
4	Sex	1	40(4) + 20(8) + 2(16) + (18)
5	Age	1	40(5) + 20(10) + 2(16) + (18)
6	Eth Trac	1	20(6) + (18)
7	Eth Sex	1	20(7) + 10(12) + 2(16) + (18)
8	Trac Sex	1	20(8) + (18)
9	Eth Age	1	20(9) + 10(13) + 2(16) + (18)
10	Trac Age	1	20(10) + (18)
11	Sex Age	1	20(11) + 10(15) + 2(16) + (18)
12	Eth Trac Sex	1	10(12) + (18)
13	Eth Trac Age	1	10(13) + (18)
14	Eth Sex Age	1	10(14) + 2(16) + 5(17) + (18)
15	Trac Sex Age	1	10(15) + (18)
16	X (Eth Sex Age)	32	2(16) + (18)
17	Eth Trac Sex Age	1	5(17) + (18)
18	Trac X (Eth Sex Age)	32	(18)
		80	

example, the minimal model's estimates for Bolton point (horizontal), subspinale (vertical), and orbitale (horizontal), are substantially higher than those derived from the full model.

Finally, reliability estimates using the full model may also be higher than those based on the minimal model (see vertical Bolton point and horizontal upper central incisor tip).

The different reliability estimates obtained with the full and minimal models can be explained by considering the true and residual variance components associated with the respective measurements. As shown in Table 3, the minimal and full models differ in their estimates of true variance. Differences are particularly marked for the landmarks' vertical perspectives.

This shows that the minimal model's estimate of true variance includes the main effects (i.e., ethnic, sex, and age). The random (residual) variance components are in most instances comparable. Table 4 substantiates that a considerable portion of the total variation may be attributed to age. Consequently, some of the true variation estimated by the minimal model is attributable to age changes.

Variation associated with sex is consistently low. Little or no ethnic variation is indicated.

Different estimates of error variance reflect interactions between tracing and the main effects.

Table 2

Comparisons of
Systematic (S), Random (R), and Global (G)
Reliability of Cephalometric Landmarks
Using Minimal (M) and Full (F) Analysis of Variance Models

		Horizontal			Vertical		
Landmark	Sex	S	R	G	S	R	G
Bolton Point	♂	0.997	0.791	0.789	0.992	0.947	0.940
	♀	0.996	0.751	0.757	0.992	0.955	0.948
Gonion	♂	1.000	0.990	0.990	1.000	0.990	0.990
	♀	1.000	0.990	0.990	1.000	0.987	0.987
Pogonion	♂	1.000	0.991	0.991	1.000	0.997	0.997
	♀	1.000	0.989	0.989	1.000	0.996	0.996
Supramentale	♂	0.999	0.991	0.990	0.998	0.977	0.975
	♀	0.999	0.990	0.990	0.998	0.972	0.970
Subspinale	♂	1.000	0.978	0.978	0.999	0.955	0.954
	♀	1.000	0.975	0.975	0.999	0.919	0.918
Orbitale	♂	0.999	0.924	0.923	0.999	0.893	0.891
	♀	0.999	0.903	0.903	0.999	0.890	0.889
Lower Incisor Tip	♂	1.000	0.989	0.989	1.000	0.994	0.994
	♀	1.000	0.988	0.988	1.000	0.993	0.993
Upper Incisor Tip	♂	1.000	0.988	0.988	1.000	0.993	0.993
	♀	1.000	0.991	0.991	1.000	0.987	0.987

Table 3

Estimates of the True and Residual Variance Components
by the Minimal (M) and Full (F) ANOVA models

		Horizontal Variance		Vertical Variance	
Landmark	Sex	True	Residual	True	Residual
Bolton Point	♂	1.553	0.410	2.327	0.129
	♀	1.246	0.395	2.469	0.117
Gonion	♂	1.904	0.020	3.516	0.028
	♀	2.024	0.021	2.479	0.032
Pogonion	♂	5.631	0.053	6.698	0.017
	♀	5.580	0.060	4.539	0.019
Supramentale	♂	4.091	0.038	3.749	0.090
	♀	4.095	0.040	2.971	0.014
Subspinale	♂	2.008	0.045	1.696	0.080
	♀	1.918	0.050	1.004	0.089
Orbitale	♂	1.396	0.115	0.731	0.088
	♀	0.963	0.103	0.707	0.087
Lower Incisor Tip	♂	3.299	0.038	2.548	0.016
	♀	3.528	0.042	2.054	0.014
Upper Incisor Tip	♂	3.432	0.040	2.922	0.022
	♀	3.783	0.036	1.729	0.023

Variance Contributed by
Ethnic, Sex and Age effects
estimated by the full ANOVA model

	Horizontal			Vertical		
Landmark	Ethnic	Sex	Age	Ethnic	Sex	Age
Bolton Point	—	0.171	0.727	—	0.063	—
Gonion	—	—	0.139	0.148	0.962	1.268
Pogonion	0.336	0.030	—	—	0.698	3.951
Supramentale	0.208	0.084	—	0.004	0.360	1.460
Subspinale	—	0.182	0.136	—	0.234	0.998
Orbitale	—	0.245	0.391	—	0.045	—
Lower Incisor Tip	—	0.174	0.032	—	0.223	1.080
Upper Incisor Tip	—	0.176	—	—	0.281	2.091

— Discussion —

This paper is intended to demonstrate the application of a full (mixed) model for the evaluation of cephalometric reliability. The fact that the full and minimal models differ is sufficient cause for due consideration of the sources of variation in methodological design and testing. Tests of significance for systematic effects alone provide no indication of reliability, which depends on their relative contribution to true variance.

Reliability estimates obtained from a minimal model are generally, though not always, higher than those derived from a full model for the design presented.

Differences between the models reflect changes in both the true and random components of variance. Age in particular, as well as sex, contribute to the variance of cephalometric data, which in turn affects the estimates of true variance and sometimes reliability.

Since reliability is defined by a ratio of variances, it is possible that measures with higher error variance are more reliable than those with lower error variance. For example, Bolton point (vertical) is more reliable than orbitale (vertical), even though the error variance for the former is almost twice that of the latter. Such results demonstrate the importance of correctly estimating true variance in order to evaluate error variance.

The sources of variation that should be included in a model will depend on the particular characteristics of each data set, as well as the eventual application of the results. If the design includes several operators, inter- and intraobserver reliability should be evaluated by the same model; their combined systematic and random components provide better estimates of global reliability. Traditionally, inter- and intraobserver errors have been analyzed separately, but it should be remembered that any single tracing made by any one individual incorporates both types of errors.

This methodology is applicable in both the clinical and research environments. For the clinician, the coefficient of reliability allows comparison of the accuracy of identifying different cephalometric landmarks. Error variances, as obtained by the "error method", are not directly comparable. Considering the extent of variation in individuals with normal occlusion (Moorrees 1953), clinicians are well advised to evaluate their technical reliability before appraising the abnormality for individual patients.

Most important, the methodology provides a flexible tool for investigating the many factors that influence cephalometric reliability. Since variance estimates are available for all of the terms in the model, it is possible to evaluate any combination of systematic and nonsystematic components present in complex designs. Reliability standards could be developed to establish the range of error for cephalometric data.

Any limit of acceptability must be arbitrary; it varies with the measure and objectives of the analysis. Figures greater than 0.90 are desirable, while reliability estimates below 0.80 render a measure doubtful. Moreover, reliabilities allow the estimation of a measure's *unobserved* true score variance, which is essential for computing the true correlations between different measures. For instance, if a correlation of 0.80 is observed between two measures, each of which has a reliability of 0.80, then their *true* correlation in 1.00, a perfect relationship.

The true correlation between measures may be estimated by —

$$\mathrm{TC} = \frac{\mathrm{OC}}{\sqrt{\mathrm{r1} \times \mathrm{r2}}} \tag{5}$$

where TC = True correlation, OC = Observed correlation, and r1 and r2 refer to the reliability of the two measures being correlated.

Finally, the methodology offers a formal means of monitoring the progress of students and technicians undergoing training in cephalometrics.

— Summary —

Cephalometrics requires accurate estimates of technical reliability if it is to serve the orthodontist in confirming diagnosis, treatment planning, and research. Estimates must be based on the ratio of true to observed variance; the more true variance relative to observed variance, the greater the reliability of the measure.

Method error, the common approach for evaluating cephalometric error, does not provide for reliability estimates and, with complex designs, its estimates of error variance may be inaccurate.

Models which evaluate variance must be defined to take all potential sources of variation into consideration. Restricting an analysis to the variation associated with the measurement process, such as inter- or intraobserver error, could result in inaccurate estimates of true and/or method error, which might alter assessments of reliability.

This research was supported by MRC grant #MA-8917.

A/O

References

Baumrind, S. and Frantz, R. C. 1971. The reliability of head film measurements. 1. Landmark identification. *Am. J. Orthod.* 60: 111-127.

Björk, A. 1947. The face in profile. *Berlingska Boktryckereit.* Lund.

Broadbent, B. Holly. 1931, 1981. A new x-ray technique and its application to orthodontia. *Angle Orthod.* 1:April,1931. Reprinted 1981. 51:93-114.

Buschang, P. H., Tanguay, R., Patterson, D. K., and Demirjian, A. 1984. Cephalometric reliability: a comparison of two assessment methods. *Am. J. Phys. Anthrop.* 63: 142-143.

Carlsson, G. E. 1967. Error in x-ray cephalometry. *Odont. T.* 75: 99-129.

Healy, M. J., 1958. Variations within individuals in human biology. *Human Biology* 30: 210-218.

Kirk, R. E., 1968. *Experimental Design: Procedures for the Behavioral Sciences.* Cole, Belmont.

Mattila, K., and Haataja, J., 1968. On the accuracy of determining certain reference point in cephalometric radiography. *Odont. T.* 76: 249-295.

Moorrees, C. F. A. 1953. Normal variation and its bearing on the use of cephalometric radiographs in orthodontic diagnosis. *Am. J. Orthod.* 39: 942-950.

Richardson, A., 1966. An investigation into the reproducibility of some point, planes and lines used in cephalometric analysis. *Amer. J. Orthod.* 52: 637-651.

Solow, B., 1966. The pattern of craniofacial associations. *Acta. odont. Scand.* suppl. 46.

Starbrun, A. E., and Danielsen, K. 1982. Precision in cephalometric landmark identification. *Eur. J. Orthod.* 4: 185-196.

Vinkka, H., and Koski, K. 1974. Inter- and intraobserver variability in an x-ray craniometric analysis method. *Proc. Finn. Dent. Soc.* 70: 156-160.

Weiner, B. J. 1971. *Statistical Principles in Experimental Design,* 2nd Ed. McGraw-Hill, New York.