A Method for Checking the Reliability of Cephalometric and Dental Morphologic Variables

JAN W. KUZMA, Ph.D.

TOM J. ZWEMER, D.D.S., M.S.

The use of cephalometric and dental morphologic characteristics is becoming increasingly popular in studying the variation and differences between various ethnic groups and tribal isolates. Frequently such investigations compare tribal data obtained by different investigators assuming that the investigators obtained the observation of each characteristic in a uniform and comparable way. Such an assumption does not seem unreasonable, particularly if standard indices are used for which the instructions for obtaining the measurements are clearly defined.

However, since measurement errors are usually involved in any study of cephalometric or dental morphologic characteristics, the examiners should check in advance the reliability of each measurement. Measurements that cannot be made with sufficient reliability should then be discarded as they would be useless in comparative studies. That is, it would be impossible to ascertain whether significant differences between tribal characteristics are real or are due to examiner errors.

For many years the method for checking reliability has utilized repeated observations by each examiner on the same set of records. Such a procedure permits a comparison of each examiner with himself and with each other examiner but it requires large numbers of observations. If, for example, five examiners are to be compared by making five repeated observations on each of ten records, a total of 250 examinations would be required for an adequate analysis of reliability for each of the dental indices used.

It is the purpose of this paper to suggest a considerably more efficient design, based on the use of modern, sophisticated statistical techniques, and to illustrate its use in the evaluation of reliability of thirty cephalometric and dental morphologic indices. Furthermore, it is shown that in addition to testing the significance of the inter- and intra-examiner errors, it is possible to test whether or not there is a short term learning or warming-up effect in obtaining repeated observations.

MATERIAL AND METHODS

The thirty characteristics used in this study were (1) the nineteen measurements of the cephalometric x-ray analyses of Downs and Riedel and (2) the following morphologic characteristics: mesiodistal crown diameter of maxillary central, mesiodistal crown diameter of maxillary lateral, mandibular arch length, mandibular arch width, maxillary arch length, maxillary arch width, height of palate, size of Carabelli cusp, depth of maxillary central shovel, buccolingual crown diameter of maxillary second premolar, and buccolingual crown diameter of maxillary first molar.

From the School of Medicine, University of California at Los Angeles and School of Dentistry, Loma Linda University, Loma Linda, Calif.

The computations were carried out at the Loma Linda Scientific Computation Facility supported by the USPHS grant FR-0027601.

EXAMINER	1	2	3	4	5
I	S_2	S_4	S_3	S_5	S_1
II	S_1	S_3	S_2	S_4	S_5
III	S_5	S_2	S_1	S_3	S_4
IV	S_4	S_1	S_5	S_2	S_3
V	S_3	S_5	S_4	S_1	S_2

Fig. 1 The Latin-square design.

Five graduate orthodontic students, experienced in obtaining cephalometric and dental morphologic measurements, were used as examiners in this study. To minimize the number of observations, a Latin-square design was used to answer the objectives of the study. Statistically, such a design is considered highly efficient as it requires only one observation by each of the five examiners on each of the five records. If the observations are taken in a prescribed order, only 25, instead of the expected 125, observations will be required. An example of the Latin-square used in this study is shown in Figure 1. Only five cases were required for this design. For example, examiner I observed subject two (S_2) first and subject four (S_4) next, while examiner II began with subject one (S_1) and followed with subject three (S_3), etc. Each case was selected for its moderate difficulty from each of the following tribes: Apache, Pima, Papago, Navajo and Eskimo. Although taking skulls from five different tribes is irrelevant to this study, it was done to assure the use of sufficiently difficult cases to illustrate potential measurement problems.

The plan of the study required that each of the five examiners measure in prescribed order the characteristics of each of the five cases. The same set of tracings was used by each of the five examiners to avoid introducing possible tracing errors. Three days after the measurements were completed the five examiners, using the same tracings, were asked to again obtain the same measurements on the same five cases in the same order without reference to the results of the previous trial. The examiners were not informed at the outset of the experiment that this would be done. All cephalometric measurements were taken from acetate tracings of lateral cephalometric films taken at a standard five-foot target-film distance using a headholder modified for field use. The dental morphologic characteristics were measured from plaster casts prepared from alginate impressions.

The specific hypotheses that were tested for each characteristic are:

1. There is no difference between the means of the five examiners.
2. There is no difference between the means of the order in which the observations were obtained.
3. There is no difference between the means of the five cases.
4. There is no difference between the means of the first and second attempts.

Results and Discussion

An analysis of variance of the Latin-squares was performed on each of the thirty variables for the first attempt, the second attempt, and for the differences of the two attempts. That is, a total of ninety analyses of variance were calculated. F tests with four and twelve degrees of freedom were used in testing the differences between the means. Table I lists the characteristics for which examiners' means were significantly different for the first and second attempts. On the first attempt there were five significant differences among the means of the examiners. Only four significant differences were observed on

TABLE I

List of characteristics for which examiners' means were significantly different for the two attempts.

FIRST ATTEMPT		SECOND ATTEMPT	
Characteristic	Significance Level	Characteristic	Significance Level
SNA-SNB Diff.	$p \leq .01$	SNA	$p \leq .01$
Facial Angle	$p \leq .01$	SNA-SNB Diff.	$p \leq .01$
Occlusal Plane	$p \leq .01$	1 - NP	$p \leq .01$
Max. A.W.	$p \leq .01$	Angle of Convexity	$p \leq .01$
Pal. Ht.	$p \leq .01$		

the second attempt; however, only for the SNA-SNB difference was a significant difference observed for both the first and second attempt.

The test for the differences in means of the first and second attempt showed that the only significant difference was for the occlusal plane. This is the only characteristic for which the examiner means on the first attempt were significantly different from those of the second attempt. Thus, in all but one of the characteristics observed, the examiners maintained internal consistency over the three-day period.

For seven of the eight characteristics shown in Table I, the difference in examiner means was significant in only one of the two attempts indicating that it is possible for these seven characteristics to be obtained uniformly. The SNA-SNB difference was significant in both attempts. Additional refinement in the technique and in landmark definition appears necessary before these characteristics can be used in a comparative study. Therefore, in a comparative study of characteristics of different groups twenty-nine of the thirty considered characteristics could be used if the same examiner collected data on all of the different groups. If these five examiners were to be used and their data compared, only twenty-two of the thirty characteristics would give meaningful comparisons.

The hypothesis of no difference in means of the order in which observations were obtained (learning or warming-up effect) was not rejected at the one percent level for any of the characteristics. This does not seem unreasonable as the measurement procedure is fairly standard and one would not expect an experienced examiner to change appreciably in one or the other direction in the measurement of five cases.

Significant differences were also observed among the five cases for all but two of the characteristics, namely, the A-B plane and maxillary arch length. This is as expected since the cases were from five different tribes.

Summary

A reliability study of thirty cephalometric and morphologic characteristics was evaluated using an efficient statistical design which required a total of only twenty-five observations in checking out five examiners and five cases. The results showed that:

1. There was a highly significant difference between examiner means in five characteristics on the first attempt and in four characteristics on the second attempt.
2. In only one of the thirty characteristics, occlusal plane, was there a difference between the first and the second attempt. The examiners

remained internally consistent in all but one of their measurements even though they did not agree among themselves on the measurements of these characteristics.

3. There were no significant differences in learning or warming-up effect.

Dept. of Biomathematics
Los Angeles, Cal. 90024

ACKNOWLEDGMENTS

We wish to acknowledge Dr. Garold Becker, Dr. Stephen Christensen, Dr. John Matis, Dr. Robert Packard, Dr. Wilmonte Penner for serving as the examiners in this study.

REFERENCES

1. Brothwell, D. R., *Dental Anthropology*, New York: The MacMillan Company, 1963.
2. Cochran, W. G. and Cox, G. M., *Experimental Designs*, New York: John Wiley and Sons, 1960 Second Edition.
3. Moorrees, Coenraad, F. A., *The Aleut Dentition*, Cambridge, Massachusetts: Harvard University Press, 1957.
4. Salzmann, J. A. *Roentgenographic Cephalometrics*, Philadelphia, Pennsylvania: J. B. Lippincott Company, 1961.