J Forensic Sci, July 2002, Vol. 47, No. 6
Paper ID JFS2001404_476
Available online at: www.astm.org

Bradley J. Adams,[1] *Ph.D. and John E. Byrd,*[1] *Ph.D.*

Interobserver Variation of Selected Postcranial Skeletal Measurements*

ABSTRACT: Osteometric data are of great importance for analytical purposes in the field of forensic anthropology, and it is critical that interobserver concordance is high in order for the results of these analyses to be reliable. Significant interobserver variation of skeletal measurements is cause for concern since it may result in conflicting conclusions. The range of interobserver variation of selected postcranial measurements is addressed. Thirteen standard measurements familiar to most forensic anthropologists were examined, as well as nine nonstandard measurements that were unfamiliar to most participants in the study. Sixty-eight individuals participated in the study, the majority of whom considered themselves to be forensic anthropologists with various levels of experience. In general, interobserver variation of the tested standard postcranial measurements was found to be minimal, with the exception of pubis length and subtrochanteric femur measurements. The difficulties that can lead to measurement error are discussed and possible solutions are recommended.

KEYWORDS: forensic science, interobserver variation, osteometrics, skeletal measurements, validity, reliability, human osteology, forensic anthropology

Measurements can exist in numerous forms and, for purposes here, all forms are viewed as falling somewhere along a continuum of objectivity, from those that involve subjective judgments, such as “wide” or “narrow,” to those that are more specific observations, such as “open” or “closed,” to those that are observations taken with the aid of a measuring device, such as a length measurement taken with calipers. Values of the more subjective measurements can be idiosyncratic to the researcher, while those relying upon a specific device tend to be less affected by the observer.

The great philosopher of science Thomas Kuhn began his essay on the role of measurement in the physical sciences by referring to the quotation on the façade of the Social Sciences Research Building at the University of Chicago (1). The statement, made over a century ago by Lord Kelvin, is, “If you cannot measure, your knowledge is meager and unsatisfactory.” Lord Kelvin placed heavy emphasis on quantification in the study of nature. Kuhn determines that the most critical role of measurement (and mathematical analyses of measurement data) lies in the resolution of conflict between competing theories. Where scientists seek to choose the more desirable of two competing theories, they make a three-way comparison between theories and between each theory and the world. Here, measurement has its greatest advantage. For this reason, physical scientists have for centuries demonstrated a strong commitment to the objectivity and power offered by measurement data and mathematical analyses of the same.

Faith in measurement data is not unique to the physical sciences. The biologist D’Arcy Thompson, in the 1917 classic, *On Growth and Form,* expressed his view on measurement as follows:

> “ . . . numerical precision is the very soul of science, and its attainment affords the best, perhaps the only criterion of the truth of theories and the correctness of experiments” (2).

Thompson demonstrated that many aspects of biological form could be explained using the principles of geometry and served to showcase the power of mathematical treatment of biological measurement data. In keeping with developments in biology, physical anthropologists began to aggressively measure the human form during the infancy of the discipline in the 19th century (c.f., 3,4) and to base their interpretations concerning human types on statistical analyses of anthropometric data. It was understood from the very beginning that information such as average height, or the relationship between bone lengths and height, could be reliably determined only through measurement and statistical analysis.

Forensic science in general has not been oblivious to the advantages of measurement and appears to be rapidly expanding the role that measurement and quantitative analyses play in the investigative process (5). The Federal Rules of Evidence today place emphasis on reliability in the conclusions drawn by scientists. In many cases, reliability is determined by the measurement of error rates for the methods used, and it is understood that techniques incorporating statistical analyses of appropriate measurement data lend themselves to the accurate determination of error rates. Congress and the judiciary look to quantification in scientific analyses for the same reasons Kuhn outlined for the physical sciences: judges and juries must invariably select the most reasonable choice among competing theories. Measurement data are believed to be objective, and they allow the examiner to go beyond subjective assessments such as “similar” or “different.” With measurement data the examiner is able to quantify the de-

* Presented at the 53rd Annual Meeting of the American Academy of Forensic Sciences, Seattle, WA, February 2001.

[1] Forensic anthropologist, U.S. Army Central Identification Laboratory, Hawaii (CILHI).

Received 30 Nov. 2001; accepted 5 May 2002; published 18 Sept. 2002.

gree of difference or similarity and state how much confidence can be placed in this interpretation.

Statistical hypothesis testing is a formal expression of the quest for reliability and repeatability Forensic identification of skeletal remains provides a case in point. Konigsberg et al. (6) argue that the first goal of forensic anthropology—developing a biological profile of the remains under analysis—is actually a statistical problem and should be understood as probability based. The view that the results of scientific analyses must be treated as probabilities is becoming more important in forensic science as a whole (5). Thus, we can treat the respective candidates for an identification as competing theories and formally evaluate them against case-specific data and reference data by testing statistical hypotheses (7,8) and using likelihood ratios (6).

Forensic anthropologists have been particularly avid in pursuing quantitative methods in skeletal biology, as exemplified by the development of the Forensic Data Bank at the University of Tennessee in the 1980s and the many methods based upon that data (9,10). For example, forensic anthropologists routinely use cranial measurements to assist in the determination of race and sex of an unknown individual, as well as postcranial measurements to ascertain stature and sex. The software package FORDISC 2.0 (10) generates posterior probabilities that measure the similarity of skeletal elements to selected populations or sexes given the measurement data entered. Byrd and Adams (8) advocate conducting stature analysis with a statistical hypothesis testing approach. In addition, recent research has shown that metric techniques can be utilized for sorting commingled skeletal remains (7). Due to the significant implications that the results of these metric analyses hold (e.g., the identification or exclusion of an unknown individual), it is of utmost importance that the measurements utilized can be accurately and reliably taken and that they are replicable between observers. Incorrect measurements could result in potentially misleading results. Furthermore, significant interobserver measurement variation could potentially compromise pooled datasets compiled from multiple researchers and, in turn, bias research based on these data.

We must confront the reality that the methods of forensic anthropology do not provide absolute answers, but rather estimates with associated error rates. The variation inherent in the methods must be considered in the grand scheme. The error rate of a method utilizing measurements is the sum of at least two sources of error: error resulting from the natural variability in the trait being measured and error resulting from variability due to inconsistent measurement. An important example of the latter is interobserver variation. Though the overall error rates of methods relying upon subjective measurements can be lower than those relying upon the more objective, it is reasonable to assume that data collected with measuring devices will typically have lower interobserver error rates. But how low is this rate of error? Following the guidance of Konigsberg et al. (6) makes it necessary that we determine error rates for the methods of forensic anthropology. If interobserver variation is sufficiently high to affect results, it must be incorporated into interpretations.

While interobserver measurement error has been tested for anthropometric studies of the living (e.g., 11–13), measurement error of the postcranial skeleton has not been formally considered. In order to address this important issue, the validity and reliability of selected postcranial measurements were tested. The postcranial measurements selected for this study were chosen due to their anticipated difficulty and, as such, represent the "worst case" scenario in regard to interobserver variation.

Materials and Methods

In order to test the variability of selected postcranial measurements, it was necessary to collect measurement data from a sample of individuals with a wide range of osteometric experience. Participants for the study were primarily composed of individuals attending the 52nd Annual Meeting of the American Academy of Forensic Sciences in Reno, Nevada. While most studies that test interobserver error compare the results of only two individuals, the present research utilized a total sample of 68 participants. Names of participants were kept anonymous, but each individual was asked to provide information regarding the number of years experience with osteometrics, their field of study, and the average number of skeletons that they measure annually. The vast majority of participants were anthropologists, although odontologists and pathologists also contributed. The sample breakdown by years of experience with osteometrics is presented in Table 1.

A total of 22 postcranial measurements were selected for the study (Appendix 1). All participants measured the same skeletal samples. Thirteen of these measurements consisted of standard postcranial measurements as outlined by Moore-Jansen, Ousley, and Jantz (14), as well as nine nonstandard measurements described by Byrd and Adams (7). Digital calipers or an osteometric board were used for all measurements, and participants were instructed on which device to use. Participants were directed to record all measurements taken with the digital calipers to the tenth of a millimeter; measurements taken with the osteometric board were recorded to the nearest millimeter. As previously stated, many of the measurements were selected for this study because they were suspected of being problematic in terms of the difficulty in getting consistent results between observers. In addition, most of the participants had no previous experience with many of the nonstandard measurements.

The results of the study revealed several types of errors: (1) transposed numbers (e.g., recording 37 mm instead of 73 mm), (2) decimal place errors (e.g., osteometric board measurements recorded in centimeters despite instructions to use only millimeters), (3) failure to "zero out" the digital calipers, resulting in consistently skewed values, (4) recording the wrong measurement (e.g., recording maximum length of the femur when the desired measurement was epicondylar breadth), and (5) lack of understanding of the measurement definition or skeletal landmarks. In general, the errors of Types 1–4 were easily recognized and could be readily adjusted since the actual measurement was known. In other situations in which the specimen is not available for remeasurement, these errors would not be as easily resolved. The specific causes of error for other outliers (Type 5) in the dataset were not as easily recognized and likely stem from a general misunderstanding of the measurement definition or skeletal landmarks and cannot be attributed to carelessness.

TABLE 1—*Participants in study.*

Experience with Osteometrics	Number of Participants
0–1 year	7
1.1–5 years	19
5.1–10 years	17
10+ years	25
Total	68

In order to "clean" the data prior to statistical analysis, all measurements that were deemed to be so erroneous as to be beyond the range of possible human variation were removed from the dataset or were corrected if the cause of the error was apparent. This was accomplished through several steps. In most instances, the cause of the measurement error was the careless recording of the incorrect measurement (e.g., recording the maximum femur length instead of the epiphyseal breadth). These observations were placed with the appropriate measurement when possible. Next, all errors that were the result of data entry mistakes were corrected (e.g., transposed numbers or measurements recorded in centimeters instead of millimeters). At this point, the dataset was in a semi-clean state. All blatant errors stemming from carelessness had been either removed or corrected when the reason for the error was obvious. Still present within the data were numerous observations that were clearly erroneous, but the cause of the error was not readily apparent. In order to establish an arbitrary criterion for performing the final step in the cleaning process, the median and standard deviation were calculated for each of the 22 measurements. All measurements that fell outside of five standard deviations from the median were removed from the dataset. Although there was only a slight difference between the median and mean for each measurement, it was determined that since the majority of measurements tended to cluster around the median, this figure best represented the "gold standard." The exclusion of these data from consideration in the final analysis was justified in that they were widely divergent from a reasonable value (given the size of the bone) and should be easily recognized as such in any dataset. (Even if the measurement value went unnoticed, the use of the measurement in further analyses would produce outlandish results.)

As might be expected, Fig. 1 shows that the number of individuals that had to have observations edited (i.e., corrected) due to careless errors dropped based on the level of experience with osteometrics. The individuals with 0–1 year of experience were the worst, with 43% of the participants needing to have at least one measurement modified. The most experienced group (over ten years) did not have any entries that needed to be adjusted due to careless recording errors. Perhaps more surprising is that the number of individuals that had to have observations deleted (i.e., removed due to extreme and unexplainable error) is not as dependent on the level of experience. Individuals with over ten years osteometric experience committed a greater number of serious measurement errors than any other experience group except for the individuals with less than one year. Figure 1 shows that, again, the individuals with the least experience most frequently had unexplainable errors, with 57% of the participants needing to have at least one entry removed from the sample. The most experienced group had 24% of the participants that needed to have at least one value removed, while the intermediate experience groups fell between these frequencies. Overall, the number of measurements that needed to be "cleaned" was small and most measurements were taken accurately. Furthermore, as the measurements were not taken in a laboratory setting that was free of distractions (the majority of the data was collected in a hallway at the 52nd AAFS meeting), it is not surprising that some careless errors were committed.

The data were subsequently analyzed with the "clean" dataset and were sorted based on years of experience with osteometrics. In order to compare the results of the study in regard to the specific measurements, as well as by individuals, a Scaled Error Index (SEI) was calculated. This index permits comparison of measurements regardless of scale. Calculation of the index is provided in Eq 1. The absolute value of the difference between the raw measurement and the median is divided by the median. In order to convert this value to the percent error from the median, it is then multiplied by 100. All statistics regarding the Scaled Error Index were based on the clean dataset.

$$\text{Scaled Error Index} = \frac{|\text{Raw Measurement} - \text{Median}|}{\text{Median}} \times 100 \qquad (1)$$

Results

Validity is the degree to which a measurement measures what it is purported to measure and the extent to which it fulfills its purpose (15). *Reliability* is the degree to which a measurement yields the same results when taken on at least two different occasions by

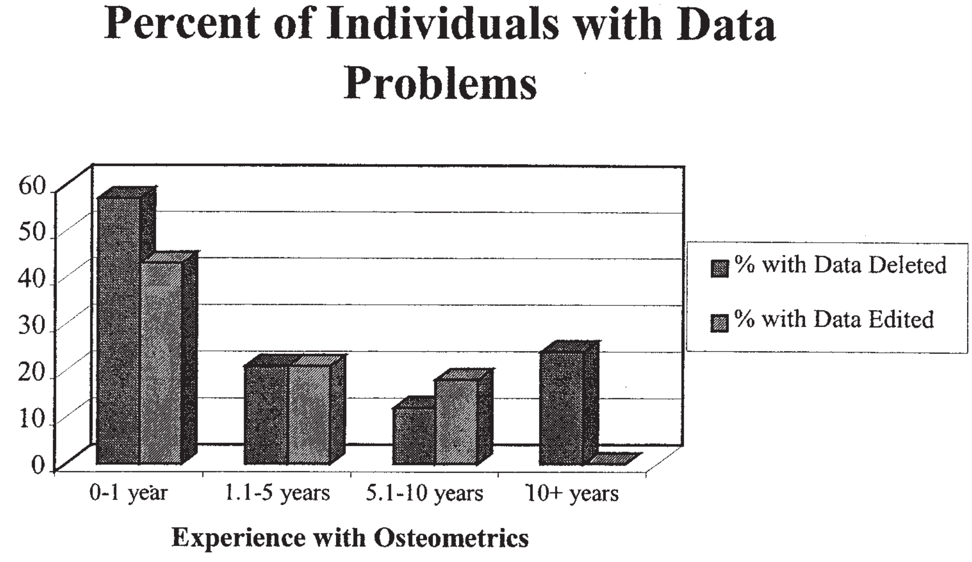


FIG. 1—*Results of data-cleaning process.*

a minimum of two different examiners (15). Thus, the reliability of a measurement can be evaluated by reference to the interobserver error rate, while the validity of a measurement determines whether it should be used at all. With high reliability (low interobserver error), measurements made by different observers may be used interchangeably without compromising the utility of the data.

A well-documented situation that exemplifies the potential implications of measurement error was elucidated by the research of Jantz, Hunt, and Meadows (16,17). They found compelling evidence that significant discrepancies existed between Trotter and Gleser's tibia measurement definition and the actual data used in the calculation of the stature estimation equations outlined in their 1952 article (18). Jantz et al. (16,17) found that, although the tibia definition explicitly stated that the medial malleolus should be included in the measurement, in actuality the malleolus had been excluded during data collection. The authors' research into primary documents from the former Central Identification Laboratories has determined that the tibia was measured inconsistently over time by Army personnel, which has possibly introduced significant noise into the Trotter and Gleser models. In turn, estimates of stature based on the guidelines of Trotter and Gleser's 1952 article will result in skewed estimates and possibly inflated standard errors. The variation between the defined measurement and the actual measurement used to calculate the regression equations will lead to overestimates of stature averaging 2.5 to 3.0 cm (17).

Due to the asymmetric morphology of the tibia, this bone has been notoriously problematic for osteometric analysis. The results of the present study show that there is still confusion among observers regarding the measurement of the tibia length (Figs. 2 and 3). While the variation seen in Fig. 2 is not extreme, it is greater than would be expected for a standard long bone length measure. Furthermore, the variation does not appear to be entirely dependent on the level of experience, as can be seen in Fig. 3 in which even the most experienced osteologists show variation in the measure. Figure 4, on the other hand, provides a good example of a measurement that has very little interobserver variation, regardless of the level of experience.

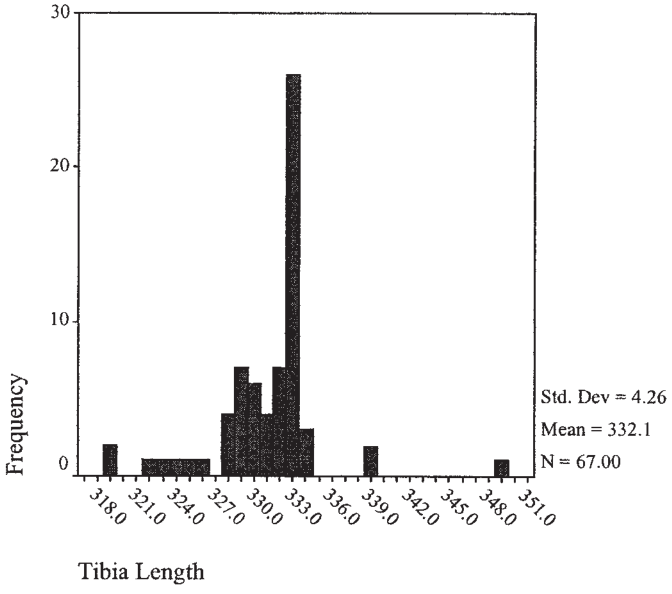


FIG. 2—*Variation observed in measurement of tibia length (based on "clean" dataset).*

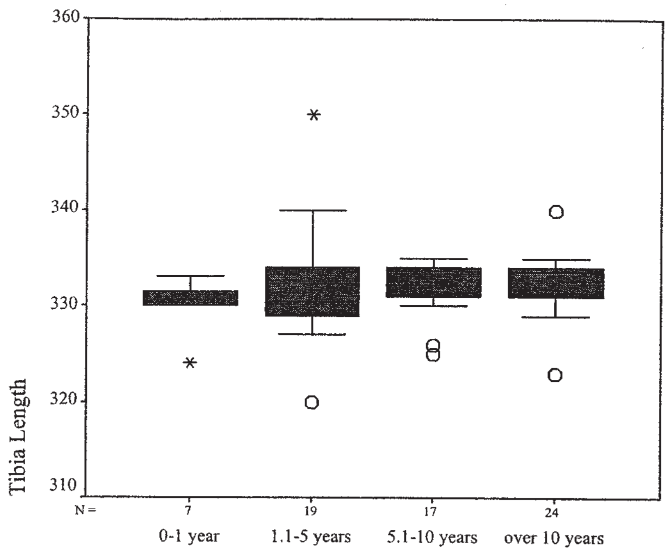


FIG. 3—*Variation observed in tibia length sorted by experience level (N = 67).*

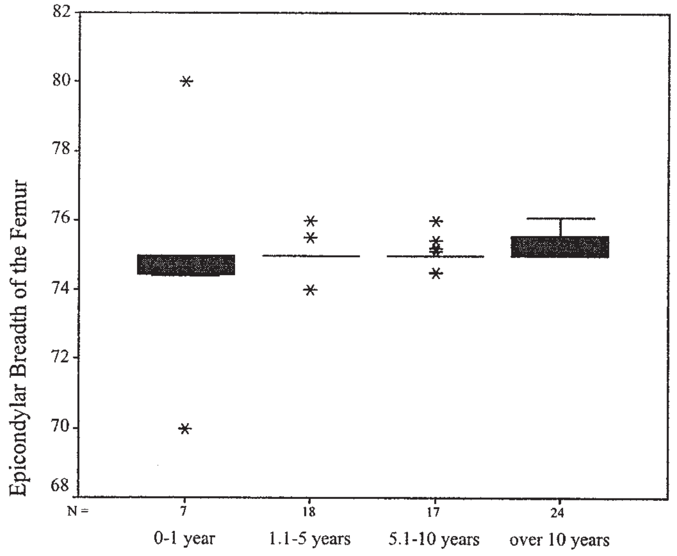


FIG. 4—*Example of a measurement with low interobserver variation.*

By far the most problematic standard measurement tested was found to be pubis length, a result that will be of no surprise to those familiar with osteometrics. Figure 5 demonstrates that the interobserver variation associated with this measurement is not specific to any one experience group. As can be seen in Fig. 6, the overall interobserver variation was high for this measurement, with an average SEI value of 8.32. Furthermore, the subtrochanteric femur measurements were found to be problematic, but not to the degree of the pubis length. Three of the nonstandard measurements were found to have large SEI values (Minimum A/P Diameter of the Femur Diaphysis, Maximum Diameter of the Femur along the Linea Aspera, and Ilium Thickness at the Sciatic Notch). Figure 6 shows that the average SEI values for these three measurements fall in between the subtrochanteric femur measurements and the pubis length. It is believed that the problems observed with these nonstandard measurements stemmed from an unfamiliarity with the landmarks and/or measurement description, and that accurate values could be derived with only minimal instruction. Table 2 provides SEI values for all tested measurements according to the experience level of the participants.

Analysis of the effect of experience on the degree of interobserver variation revealed that the overall SEI dropped as experience increased (Fig. 7). Those individuals having the least experience produced the most variation (SEI = 4.20), and those with the most experience displayed the least variation (SEI = 2.31). In order to observe whether the differences in the means were statistically significant, one-tailed t-test comparisons were performed, and the results are presented in Table 3. It was found that there is a statistically significant difference ($p < 0.05$ level) between those individuals with under five years experience and those with over five years experience. The data suggest that after five years of experience with osteometrics that there is no improvement in measuring ability. Those with less than five years experience show greater interobserver variation. In general, it was found that the same measurements were problematic regardless of the experience level (e.g., pubis length and subtrochanteric femur measurements); the main difference was the degree to which this variation was expressed.

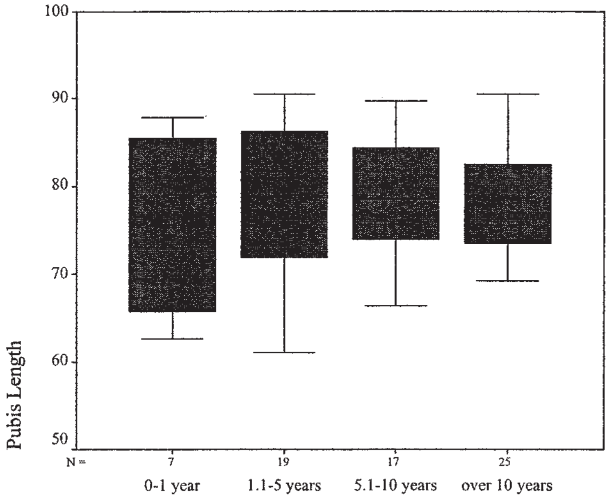


FIG. 5—*High interobserver variation exemplified by pubis length measurements.*

A comparison was performed to observe whether the measurement device had an effect on the interobserver variation, and it was found that measurements that utilize an osteometric board were generally found to be more consistent. As these measurements generally entail maximum length or breadth dimensions, the results are not surprising. The large difference in average SEI values according to the measurement device utilized is likely a reflection of the difficulty of the measurements and not an indication that one device is superior to another. Figure 8 shows the difference in the SEI for the two devices.

Conclusions and Recommendations

Anthropologists have followed the trend among scientists in general in placing increasing emphasis on measurement data as the fuel for their analytical fire. One of the reasons for this emphasis is the alleged reliability of measurements and the results of mathematical analyses of the same. We have formally evaluated the efficacy of this belief on the part of biological anthropologists by measuring the variability in selected skeletal measurements due to interobserver error. Where this variation is found to be inordinately high, then our faith is misplaced.

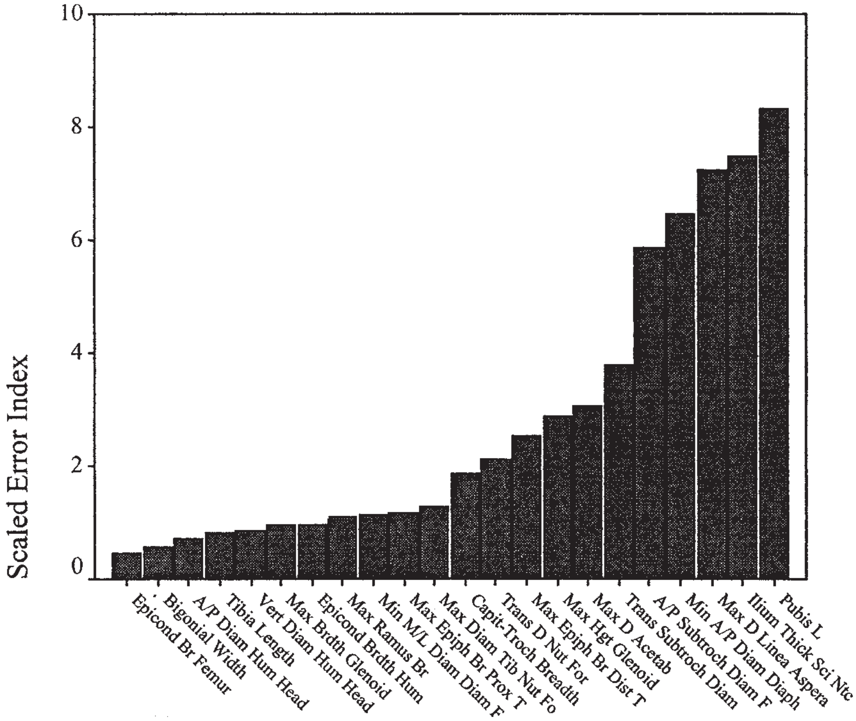
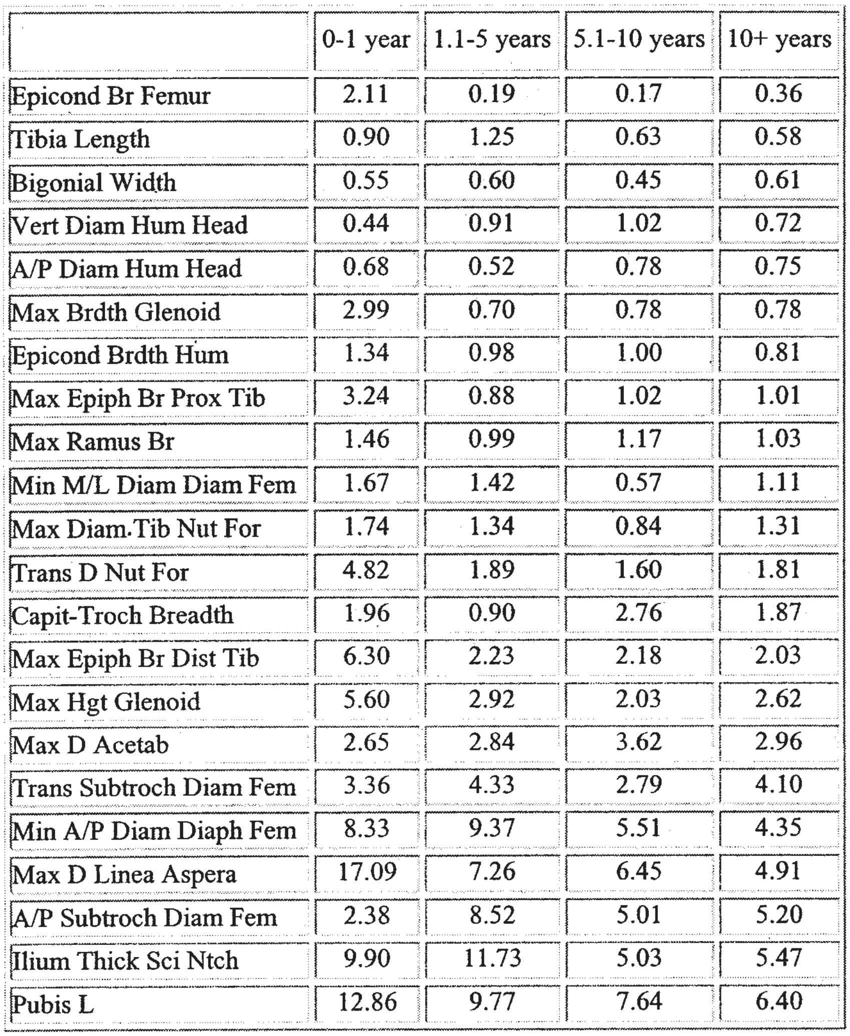


FIG. 6—*Ranked Scaled Error Index across all experience groups.*

TABLE 2—*Scaled error index by experience level.*



	0-1 year	1.1-5 years	5.1-10 years	10+ years
Epicond Br Femur	2.11	0.19	0.17	0.36
Tibia Length	0.90	1.25	0.63	0.58
Bigonial Width	0.55	0.60	0.45	0.61
Vert Diam Hum Head	0.44	0.91	1.02	0.72
A/P Diam Hum Head	0.68	0.52	0.78	0.75
Max Brdth Glenoid	2.99	0.70	0.78	0.78
Epicond Brdth Hum	1.34	0.98	1.00	0.81
Max Epiph Br Prox Tib	3.24	0.88	1.02	1.01
Max Ramus Br	1.46	0.99	1.17	1.03
Min M/L Diam Diam Fem	1.67	1.42	0.57	1.11
Max Diam.Tib Nut For	1.74	1.34	0.84	1.31
Trans D Nut For	4.82	1.89	1.60	1.81
Capit-Troch Breadth	1.96	0.90	2.76	1.87
Max Epiph Br Dist Tib	6.30	2.23	2.18	2.03
Max Hgt Glenoid	5.60	2.92	2.03	2.62
Max D Acetab	2.65	2.84	3.62	2.96
Trans Subtroch Diam Fem	3.36	4.33	2.79	4.10
Min A/P Diam Diaph Fem	8.33	9.37	5.51	4.35
Max D Linea Aspera	17.09	7.26	6.45	4.91
A/P Subtroch Diam Fem	2.38	8.52	5.01	5.20
Ilium Thick Sci Ntch	9.90	11.73	5.03	5.47
Pubis L	12.86	9.77	7.64	6.40

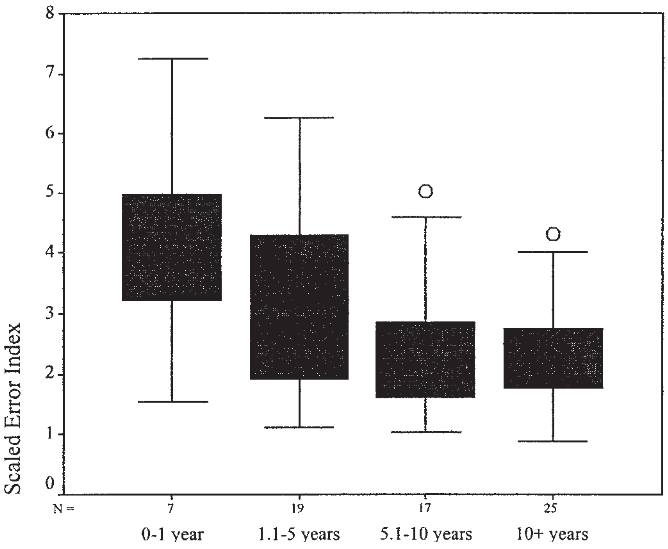


FIG. 7—*Box plot of Scaled Error Index across all variables and sorted by experience (N = 68).*

TABLE 3—*One-tailed T-test p-values comparing Scaled Error Index by experience.*

	1.1–5 Years	5.1–10 Years	Over 10 Years
0–1 year	0.095	0.004*	0.000*
1.1–5 years		0.038*	0.006*
5.1–10 years			0.371

* Indicates significance at 0.05 level.

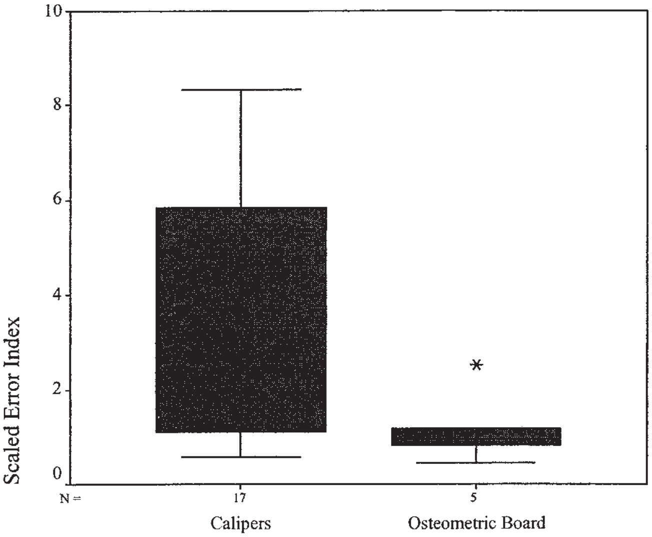


FIG. 8—*Scaled Error Index comparison between measurements derived with an osteometric board versus calipers.*

This study should be viewed in the perspective that the measurement sample represents some of the most difficult postcranial measurements and, as such, provides a "worst case" scenario for interobserver variation. The fact that there was agreement between most measurements recorded by most participants is satisfying, although the extreme variation observed in others is cause for concern. Overall, this study has shown that interobserver variation is a valid concern with osteometrics, and that it must be considered when we interpret the results of quantitative analyses. For example, the application of a stature estimation model to a forensic case is valid only if the case examiner takes the measurement(s) in the same manner as those who collected the data used to derive the model. Where this is uncertain, the conclusions should reflect the problem. Interobserver variation in the original reference data will tend to increase the standard error and could potentially bias the stature estimation model. Most of the measurements in this study have proven reliable, and we suspect that this will be true of the majority of skeletal measurements in use today. Although maximum length measurements of long bones were not tested (with the exception of tibia length), it is believed that these measurements can be accurately taken between observers due to the simplicity in their definitions and use of the osteometric board. As noted previously, special care needs to be taken with the tibia. The most problematic standard measurements tested were the pubis length and both subtrochanteric femur measurements. As the pubis length was found to be universally problematic, it can be determined that the ischium length would be equally as problematic since the same landmarks of the acetabulum are utilized.

With the exception of pubis length and the subtrochanteric femur measures, most of the observed errors can be explained as unfamiliarity with landmarks or a misunderstanding of the measurement description. We believe the pubis length to be an invalid measurement due to the problem of locating the landmark within the acetabulum and recommend that it not be used in analyses (the same is recommended for ischium length). The subtrochanteric measurements on the femur are valid, but require better measurement descriptions than currently used, and quite likely specific training from experienced osteologists to make them reliable. Two problems here are how far distal to the lesser trochanter should the measurement be taken, and how closely must the measurer maintain the anterior-posterior or transverse orientation (there is a tendency to rest the flattened portion of the anterior surface against the caliper jaw when taking the anterior-posterior measurement). In general, measurements requiring maximum or minimum dimensions are more reliable than measurements based on orientation (A/P or M/L) or landmarks. Nonstandard measurements that are based on recognizable landmarks and still show significant variation across all experience groups are evidence that training and practice are necessary. There appears to be a statistically significant improvement with osteometrics after five years of experience, but some of the same careless or unexplainable measurement errors are observed across all experience groups.

Quality control during the data-gathering process is essential and is an effective means of controlling errors. Measurements should be checked against reference data for obvious outliers. Our experience has shown that most mistakes are easy to spot because they produce numbers notably out of line with others on record. Those managing the Forensic Data Bank report similar findings, though they lament the amount of effort that is sometimes involved in quality control (Richard Jantz, personal communication). We stress the importance of university training in osteometrics to promote continuity in data collection. Beyond the university, forensic laboratories should include detailed measurement descriptions in their standard operating procedures and provide osteometric training to new staff members. The recent publication of laboratory manuals (14,19) has simplified this process. New procedures utilizing skeletal measurements should favor measurements that are relatively easy to take and provide clear definitions of the measurements in any publications. These steps will improve the overall precision and accuracy of anthropological findings derived from metric data.

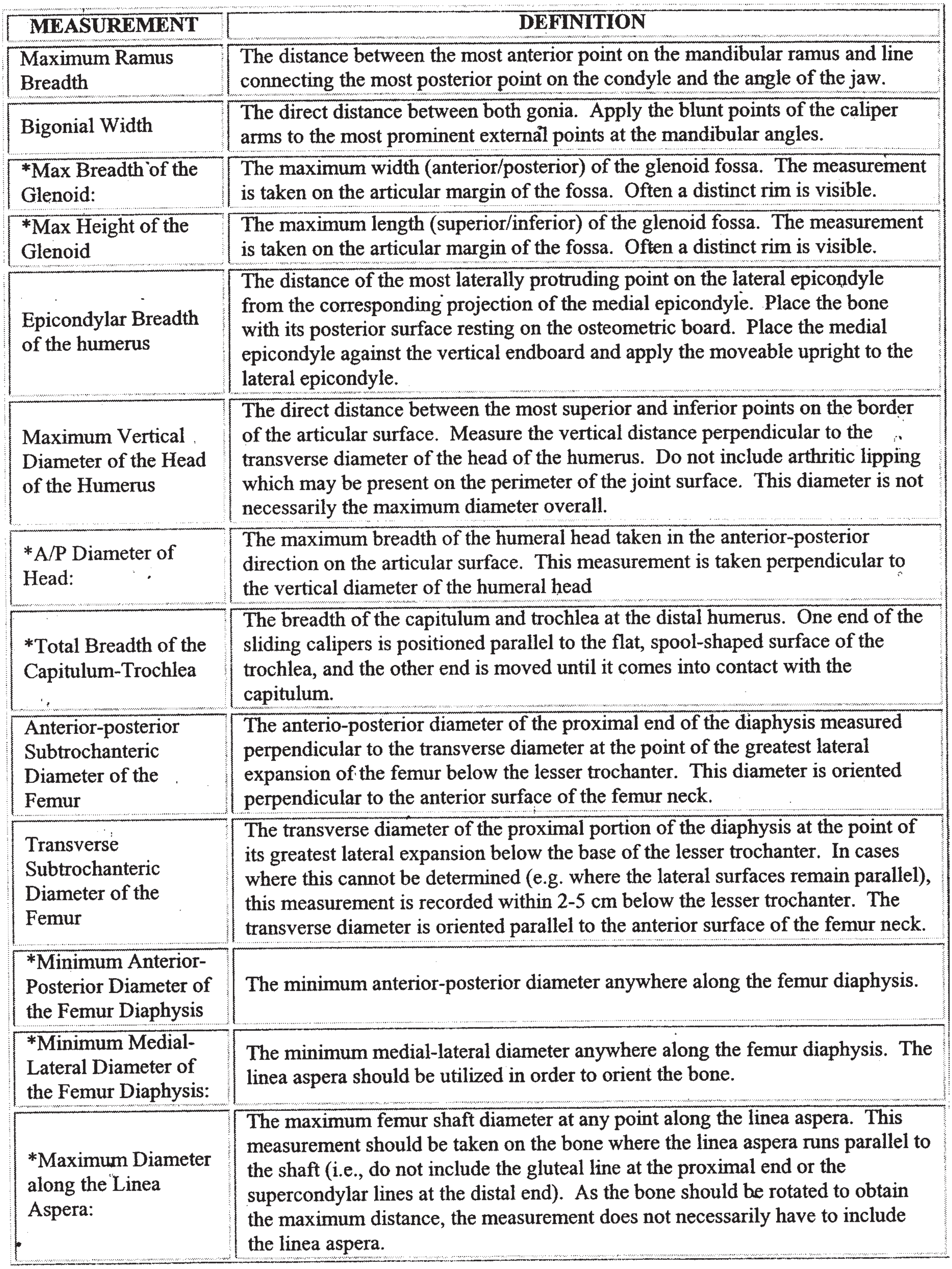
Overall, this study demonstrates that anthropologists have correctly placed a high degree of faith in the reliability of skeletal measurements. The selected standard measurements—some of the most difficult postcranial measurements—show only modest error rates (generally <3%) and, thus, high reliability. More important is that the distributions for most of the measurements have the majority of participants obtaining the same measurements value. It is believed that the interobserver variation for other postcranial measurements (e.g., most of the maximum long bone lengths) would be almost nonexistent. Interobserver variation of problematic skeletal measurements can be further reduced with the recommended steps of intensifying the training anthropology students receive in osteometrics and improving certain measurements descriptions.

Acknowledgments

We would like to thank those who participated in this study in addition to AAFS for allowing the data collection to occur during the 52nd annual meetings. Also, we would like to acknowledge Dr. Richard Jantz and Dr. Thomas Holland, who provided helpful comments during the preparation of this manuscript. Dr. Mark Leney provided many helpful suggestions regarding the statistical analysis.

APPENDIX 1

Measurements and Definitions Used in Study (denotes the "nonstandard' measurements, all othes as defined by Ref 14)*



MEASUREMENT	DEFINITION
Maximum Ramus Breadth	The distance between the most anterior point on the mandibular ramus and line connecting the most posterior point on the condyle and the angle of the jaw.
Bigonial Width	The direct distance between both gonia. Apply the blunt points of the caliper arms to the most prominent external points at the mandibular angles.
*Max Breadth of the Glenoid:	The maximum width (anterior/posterior) of the glenoid fossa. The measurement is taken on the articular margin of the fossa. Often a distinct rim is visible.
*Max Height of the Glenoid	The maximum length (superior/inferior) of the glenoid fossa. The measurement is taken on the articular margin of the fossa. Often a distinct rim is visible.
Epicondylar Breadth of the humerus	The distance of the most laterally protruding point on the lateral epicondyle from the corresponding projection of the medial epicondyle. Place the bone with its posterior surface resting on the osteometric board. Place the medial epicondyle against the vertical endboard and apply the moveable upright to the lateral epicondyle.
Maximum Vertical Diameter of the Head of the Humerus	The direct distance between the most superior and inferior points on the border of the articular surface. Measure the vertical distance perpendicular to the transverse diameter of the head of the humerus. Do not include arthritic lipping which may be present on the perimeter of the joint surface. This diameter is not necessarily the maximum diameter overall.
*A/P Diameter of Head:	The maximum breadth of the humeral head taken in the anterior-posterior direction on the articular surface. This measurement is taken perpendicular to the vertical diameter of the humeral head
*Total Breadth of the Capitulum-Trochlea	The breadth of the capitulum and trochlea at the distal humerus. One end of the sliding calipers is positioned parallel to the flat, spool-shaped surface of the trochlea, and the other end is moved until it comes into contact with the capitulum.
Anterior-posterior Subtrochanteric Diameter of the Femur	The anterio-posterior diameter of the proximal end of the diaphysis measured perpendicular to the transverse diameter at the point of the greatest lateral expansion of the femur below the lesser trochanter. This diameter is oriented perpendicular to the anterior surface of the femur neck.
Transverse Subtrochanteric Diameter of the Femur	The transverse diameter of the proximal portion of the diaphysis at the point of its greatest lateral expansion below the base of the lesser trochanter. In cases where this cannot be determined (e.g. where the lateral surfaces remain parallel), this measurement is recorded within 2-5 cm below the lesser trochanter. The transverse diameter is oriented parallel to the anterior surface of the femur neck.
*Minimum Anterior-Posterior Diameter of the Femur Diaphysis	The minimum anterior-posterior diameter anywhere along the femur diaphysis.
*Minimum Medial-Lateral Diameter of the Femur Diaphysis:	The minimum medial-lateral diameter anywhere along the femur diaphysis. The linea aspera should be utilized in order to orient the bone.
*Maximum Diameter along the Linea Aspera:	The maximum femur shaft diameter at any point along the linea aspera. This measurement should be taken on the bone where the linea aspera runs parallel to the shaft (i.e., do not include the gluteal line at the proximal end or the supercondylar lines at the distal end). As the bone should be rotated to obtain the maximum distance, the measurement does not necessarily have to include the linea aspera.

APPENDIX 1 *Continued*

Epicondylar Breadth of the Femur	The distance between the two most laterally projecting points on the epicondyles. Place the femur on the osteometric board so that it is resting on its posterior surface. Press one of the epicondyles against the vertical endboard while applying the movable upright to the other condyle. The measurement is parallel to the distal surfaces of the condyles.
Maximum Epiphyseal Breadth of the Proximal Tibia	The maximum distance between the two most laterally projecting points on the medial and lateral condyles of the proximal epiphysis. Place the tibia on the osteometric board resting on its posterior surface. Press the lateral condyle against the vertical endboard, and place the movable upright against the medial condyle. Tibiae exhibiting marked torsion may have to be rotated to obtain the maximum breadth, but do not include the occasionally prominent articular surface for the fibula.
Maximum Epiphyseal Breadth of the Distal Tibia	The distance between the most medial point on the medial malleolus and the lateral surface of the distal epiphysis. Place the two lateral protrusions of the distal epiphysis against the fixed side of the osteometric board and move the sliding board until it contacts the medial malleolus.
Maximum Diameter of the Tibia at the Nutrient Foramen	The distance between the anterior crest and the posterior surface at the level of the nutrient foramen. Rotate the caliper arms around the bone to get a maximum reading.
Transverse Diameter of the Tibia at the Nutrient Foramen:	The straight line distance of the medial margin from the interosseous crest. This is taken perpendicular to the maximum diameter at the nutrient foramen.
Length of the Tibia	The distance from the superior articular surface of the lateral condyle of the tibia to the tip of the medial malleolus. Place the tibia on the osteometric board resting on its posterior surface with the longitudinal axis of the bone parallel to the board. Place the lip of the medial malleolus on the vertical endboard and press the movable upright against the proximal articular surface of the lateral condyle.
Pubis Length	The distance from the point in the acetabulum where the three elements of the innominate meet to the upper end of the pubic symphysis. The measuring point in the acetabulum may be identified in the adult because: 1) frequently there is an irregularity there, both in the acetabulum and inside the pelvis; 2) there is a change in thickness which may be seen by holding the bone up to a light; 3) often there is a notch in the border of the articular surface in the acetabulum. In measuring the pubis care should be taken to hold the caliper parallel to the long axis of the bone.
*Thickness of the Ilium at the Sciatic Notch	Position one end of the calipers along the arcuate line, immediately next to the apex of the auricular surface (do not include the auricular surface). Slide the opposing end of the calipers to the lateral surface of the ilium to obtain the measurement.
*Maximum Diameter of the Acetabulum	The maximum distance across the acetabulum taken at any two points along the articular border of the lunate surface.

References

1. Kuhn TS. The function of measurement in modern physical science. Isis 1961;52:161–93.
2. Thompson DW. On growth and form. Cambridge: Cambridge University Press, 1971.
3. Pearson K. Mathematical contributions to the theory of evolution: on the reconstruction of the stature of prehistoric races. Phil Trans of the R Soc of Lond 1899;192A.
4. Boas F. Anthropometry of Shoshonean tribes. American Anthropologist 1899;1:751–8.
5. Kiely TF. Forensic evidence: science and the criminal law. Boca Raton: CRC Press, 2001.
6. Konigsberg LW, Herrmann NP, Wescott DJ. Commentary on McBride DG, Dietz MJ, Vennemeyer MT, Meadors SA, Benfer RA, and Furbee NL. Bootstrap methods for sex determination from the Os Coxae using the ID3 algorithm. J Forensic Sci 2002;47(2):424–6.
7. Byrd JE, Adams BJ. Sorting commingled human remains. Paper presented at the Advances in Personal Identification in Mass Disasters; Hickam AFB, Hawaii; 1999.
8. Byrd JE, Adams BJ. An alternative approach to the use of stature and long bone measurements in the identification process. Proceedings of the Meeting of the American Academy of Forensic Sciences; Reno, Nevada; 2000.
9. Jantz RL, Moore-Jansen PH. A database for forensic anthropology: structure, content, and analysis. Report of Investigations. Knoxville: The University of Tennessee, Department of Anthropology, 1988. Report No. 47.
10. Ousley S, Jantz R. FORDISC 2.0. Knoxville: University of Tennessee, 1996.
11. Himes JH. Reliability of anthropometric methods and replicate measurements. Am J Phys Anthropol 1989;79(1):77–80.
12. Jamison PL, Zegura SL. A univariate and multivariate examination of measurement error in anthropometry. Am J Phys Anthropol 1974;40(2): 197–203.
13. Jamison PL, Ward RE. Brief communication: measurement size, precision, and reliability in craniofacial anthropometry: bigger is better. Am J Phys Anthropol 1993;90(4):495–500.
14. Moore-Jansen PM, Ousley SD, Jantz RL. Data collection procedures for forensic skeletal material. Report of investigations. Knoxville: University of Tennessee, 1994. Report No. 48.
15. Nance JD. Reliability, validity, and quantitative methods in archaeology. In: Aldenderfer MS, editor. Quantitative research in archaeology. Newbury Park: Sage Publications, Inc., 1987;245–93.
16. Jantz RL, Hunt DR, Meadows L. Maximum length of the tibia: how did Trotter measure it? Am J Phys Anthropol 1994;93(4):525–8.
17. Jantz RL, Hunt DR, Meadows L. The measure and mismeasure of the tibia: implications for stature estimation. J Forensic Sci 1995;40(5):758–61.
18. Trotter M, Gleser GC. Estimation of stature from long bones of American whites and Negroes. Am J Phys Anthropol 1952;19:213–27.
19. Buikstra JE, Ubelaker DH, editors. Standards for data collection from human skeletal remains. Fayetteville: Arkansas Archaeological Survey, 1994.

Additional information—Reprints not available

Bradley Adams, Ph.D.
U.S. Army CILHI
310 Worchester Ave.
Hickam AFB, HI 96853
E-mail: adamsb@cilhi.army.mil