AMERICAN JOURNAL OF PHYSICAL ANTHROPOLOGY 125:132–137 (2004)

Reliability Test of the Visual Assessment of Cranial Traits for Sex Determination

Dana E. Walrath,[1]* Paul Turner,[1] and Jaroslav Bruzek[2]

[1]*Department of Medicine, University of Vermont, Burlington, Vermont 05401*
[2]*URA 376 du CNRS, Université Bordeaux I, Talence, France 33405*

KEY WORDS interobserver variation; sex identification; dimorphism; cranium; Inuit

ABSTRACT Interobserver variation in visual evaluation was analyzed for 10 cranial traits in a homogeneous archaeological series. Two observers independently scored cranial traits commonly used for determination of sex. Though determination of sex did not differ significantly for the two observers, individual traits had different levels of interobserver reliability. In addition, indices of relative "maleness" and "femaleness" derived by the two observers differed at statistically significant levels. Because such indices are used in cross-population comparisons of relative gracility and robusticity of diverse samples, these comparisons should be interpreted with caution when more than one investigator has performed an assessment. Most of our instances of interobserver discordance derived from character traits described in subjective terms without accompanying diagrams. Clarity of definition, rather than number of character traits, was found to be critical for effective determination of sex by the visual assessment method. Use of fewer, more precisely defined character traits can improve interobserver reliability. Am J Phys Anthropol 125:132–137, 2004.

Reconstructing the biology and behavior of past populations relies on accurate determination of sex from bony elements. However, considerable overlap in the size and degree of robusticity in male and female skeletons complicates the determination of sex. In addition, higher levels of robusticity, particularly in the mandible and skull, in past populations led to a bias toward identification of skeletal and fossil specimens as male (Weiss, 1972). Environmental and behavioral influences on robusticity also resulted in incorrect determination of sex in more recent populations (Krogman and Isçan, 1986; Novotny, 1986; Novotny et al., 1993; St. Hoyme and Isçan, 1989; Bruzek, 1992, 1995, 2002).

Although the accuracy of sex determination is greater from the pelvis and other postcranial elements, the skull is frequently used for determination of sex in archaeological contexts because it is better preserved (Novotny et al., 1993). However, determination of sex based on the skull is incorrect in as many as 20% of specimens (Masset, 1987). Meindl et al. (1985) suggested that this sex determination error stems from inherent variability in the expression of dimorphism in the skull, compared to the canalization of pelvic dimorphism by the selective effects of childbirth. These authors showed that cranial and pelvic expressions of sexual dimorphism are poorly correlated, and that interobserver discrepancy is more pronounced in the case of the skull.

While metric and nonmetric features can be assessed for sex identification of the skull, morphometric analyses can be confounded by sex differences in shape. To overcome this limitation, Giles and Elliot (1963) combined nine standard cranial measures into a single discriminant function for sex identification from the skull. Presently, any combination of standard measurements can be used for sex identification using FORDISC 2.0, a computer program capable of performing multiple discriminant function analysis (Ousley and Jantz, 1996). This program optimizes the capabilities of discriminant functions, minimizing the biasing effects of size variation on discriminant functions documented in previous studies (VanGerven, 1974; Burr et al., 1977). However, population differences in size and expression of dimorphism remain problematic. Because of population differences in the degree of dimorphism, discriminant functions are population-specific (Meindl et al., 1985; MacLaughlin and Bruce, 1986, 1990; vanVark and Schaafoma 1992). While the overlap in size of male and female ranges is still the biggest problem for sexing skeletal material, the accuracy of sex determination depends on the degree of dimorphism expressed in the individual specimen.

With greater emphasis on *shape* rather than *size*, visual assessment methods provide a valuable tool

*Correspondence to: Dana Walrath, Ph.D., Department of Medicine, University of Vermont, 371 Pearl St., Burlington, VT 05401. E-mail: dana.walrath@uvm.edu

Received 26 August 2002; accepted 21 July 2003.

DOI 10.1002/ajpa.10373
Published online 12 January 2004 in Wiley InterScience (www.interscience.wiley.com).

for determination of sex and comparisons across populations (Ferembach et al., 1980; Meindl et al., 1985; Bruzek et al., 1992; Buikstra et al., 1994). Meindl et al. (1985) showed that subjective assessment methods compare favorably with discriminant function analyses of the skull.

Visual assessment of character traits is commonly used in determination of sex from the skull for two reasons. First, the method is rapid. Second, male and female ranges of cranial morphometric characteristics are population-specific (vanVark and Schaafoma, 1992). However, because the visual assessment method is subjective, interobserver variation can result in differences in indices of population sexual dimorphism and sex identification.

Reliability tests, common to diverse branches of biological anthropology, can be used to determine the effectiveness of a particular methodology. These studies aim to distinguish true biological variation from interobserver variation, and so play a particularly important role in the examination of nonmetric features. Reliability studies of visual assessment criteria were published for maturity evaluation via Tanner staging (Espeland et al., 1990), dental age determination (Liversidge, 1994), paleopatholoical diagnosis (Waldron and Rogers, 1991), age at death (Meindl et al., 1990), and forensic identification (Hogge et al., 1994). While it might be assumed that an increased number of criteria would improve reliability in the analysis of dimorphic traits, the optimal set of criteria for reliable determination of sex is unknown. The present work tests the reliability of visual assessment of sexually dimorphic cranial character traits in an archaeological population of unknown sex. The implications of differences in determination of sex by two observers are considered.

MATERIALS AND METHODS

The study sample consisted of 42 well-preserved, complete, or nearly complete Inuit crania, curated at the University of Pennsylvania Museum. The skeletal material was excavated from Point Barrow, Alaska, in 1898, in 1917–1919, and again in 1928 (McIlhenny, 1900; Mason, 1930; Van Valin, 1941). The majority of the collection dates from the Birnirk period (AD 500–900), representing some of the earliest Arctic human remains found in association with artifacts. This series represents a homogeneous group of individuals adapted to extreme Arctic conditions. Because the sample consisted entirely of robust, cold-adapted individuals, the reliability of visual assessment for determination of sex could be tested without confounding by inherent population heterogeneity.

The approach employed in the present study is based on methodologies or standards developed during symposia on consistency of techniques for skeletal biologists. The first method was developed, during the 1970s, by the "Workshop of European Anthropologists" to define uniform methods for tasks such as determination of sex (Ferembach et al., 1980). The second method was developed at the 1991 seminar held at the Field Museum of Natural History in Chicago in response to federal law 101-601, the "Graves Protection and Repatriation Act." The resulting standards (Buikstra et al., 1994) were designed to optimize data collection from Native American skeletal material and facilitate the repatriation process. Each of these symposia contributed to the methods employed in the present study. The method of Ferembach et al. (1980) has the advantages of precision and manipulation of data for determination of sex, while character trait descriptions are brief. Numerical values are assigned to each of the diagnostic features according to a five-point scale ranging from -2 to $+2$, corresponding to hyperfeminine and hypermasculine, respectively. This technique effectively ranks the importance of certain traits. Individual features are multiplied by one, two, or three, based on their significance for determination of sex. Thus, final determination of sex is the result of weighted averages. Buikstra et al. (1994) contributed detailed character trait descriptions, as well as comparative figures to evaluate them, to the sex determination methodology of Ferembach et al. (1980). These two complementary approaches can easily be used in tandem, as in the present study.

Figure 1 presents the definitions of the characters used in the study. The first and third authors, both trained and experienced osteologists, scored features independently for each cranium in the series. Determination of sex was based on the "index of sexualization" (IS) (Ascadi and Nemeskeri, 1970; Ferembach et al., 1980), calculated according to the formula:

$$\mathrm{IS} = \frac{\Sigma(\text{score} \times \text{weight})}{\Sigma\text{weight}}$$

Positive IS values identify a specimen as male, while negative IS values identify a specimen as female. If the IS score is zero or approaching zero, the sex of the specimen must be regarded as uncertain (Ascadi and Nemeskeri, 1970). Accordingly, this study defines the interval of ± 0.2 as the criterion for indeterminate sex identification.

Because females are generally more gracile and males more robust, the IS score is also used to reflect the relative gracility and robusticity of an individual specimen. In this regard, the IS score can also be used to assess an entire sample in two ways. First, the mean sample IS reflects the gracility or robusticity of a sample. Second, the IS scores of males and females can be compared, allowing assessment of the degree of sexual dimorphism in the sample.

Choice of analytical methods was driven by two factors: 1) the categorical nature of data in the sex determination methodology being tested, and 2) the non-normal distribution of the data. The nonparametric gamma statistic was used to assess interobserver reliability for the two observers on the 10

Trait	Weight	Sexualization				
		Hyperfeminine (-2)	Feminine (-1)	Indeterminate 0	Masculine (-1)	Hypermasculine (+2)
Glabella	3	smooth	slightly delimited	delimited	marked	massive, prominent
Mastoid process	3	very small	small	medium	large	very large
Nuchal plane	3	smooth	slightly arched traces of nuchal lines	nuchal lines and occipital crest evident	nuchal lines and occipital crest marked	nuchal lines and occipital crest with rough surface
Zygomatic process of the temporal	3	very thin and low	thin and low	medium	thick and high	very thick and high
Superciliary arches	2	smooth	slighly delimited	delimited, marked	marked	very marked
Frontal and parietal eminences	2	marked	medium	moderate	indistinct	missing
External occipital protuberance	2	smooth	hardly	medium	marked	very marked
Zygomatics	2	very low, smooth surface	low, smooth surface	medium, irregular surface	high, irregular surface	very high, irregular surface
Frontal profile	1	vertical	almost vertical	little inclined	medium inclined	strongly inclined
Orbital form	1	very round, sharp border	round, sharp border	transitory form, medium border	quadrangular, rounded border	very quadrangular, rounded border

Fig. 1. Character traits for visual assessment of sex determination of the cranium. Modified from Ferembach et al. (1980) and Buikstra et al. (1994).

cranial traits individually. Rather than the more traditional nonparametric Spearman's rho or Kendall's tau. Gamma was used due to the number of tied scores inherent in the rating scheme. In such cases, the more traditional nonparametric statistics yield biased estimates, whereas the gamma statistic

TABLE 1. Interobserver reliability for 10 cranial traits[1]

Trait	Gamma
Glabella	0.85
Mastoid	0.915
Nuchal plane	0.67
Zygomatic process of temporal	0.725
Supercilliary arches	0.92
Frontal and parietal eminences	0.21
External occipital protuberances	0.89
Zytomatics	0.85
Frontal profile	0.58
Orbital form	0.73

[1] All traits excluding frontal and parietal eminences demonstrated significant interobserver reliability at $P < 0.001$. Interobserver reliability for frontal and parietal eminences was nonsignificant ($P = 0.27$).

adequately controls for such biases (Chen and Popovich, 2002; Siegel and Castellan, 1988). The gamma statistic is reported on a scale of 0–1.0, with 0 indicative of no association between observer's scores, and 1.0 indicative of perfect association.

Cohen's kappa (Cohen, 1960) was used to assess interobserver reliability for the two observers on the determination of sex based on the weighted index of sexualization (IS) score. Kappa assesses the proportion of agreement between observers corrected for chance and the standard measure of interobserver reliability with nominal data (e.g., male vs. female). Kappa is scaled on a range from −1 to +1. A negative value indicates a poorer than chance agreement, zero indicates agreement totally by chance alone, a positive value indicates a better than chance agreement, and +1 indicates perfect agreement (Shrout et al., 1987). Five separate kappa scores were generated to assess interobserver reliability: 1) the overall IS reliablity based on all 10 cranial traits, 2) the IS score for the 4 cranial traits providing visual cues, 3) the IS score for the remaining 6 cranial traits without visual cues, 4) the IS score for the 5 traits with a gamma value of 0.85 and greater, and 5) the IS for the remaining 5 traits with a gamma of less than 0.85. Based on the assessment criteria of Fleiss (1981) for the adequacy of kappa, an a priori kappa score of 0.60–0.69 was uses as the indicator of "good agreement/interobserver reliability" for all five coefficients.

The Wilcoxon sign ranks (Siegel and Castellan, 1988), the nonparametric equivalent of the dependent *t*-test, was used to assess significant difference in the IS scores of the two observers. Statistical significance was assessed at $P < 0.05$ for the kappa, gamma, and Wilcoxon sign ranks statistics.

RESULTS AND DISCUSSION

Table 1 presents the interobserver reliability for each of the 10 character traits of the Inuit cranial sample, as assessed using the gamma statistic. Though all traits except the frontal and parietal eminences demonstrate statistically significant interobserver reliability at $P < 0.001$, the gamma scores range from 0.58–0.92. The characters with the highest gamma scores (0.85 and above) possess clear definitions, often accompanied by illustrations (glabella, mastoid process, superciliary arches, external occipital protuberance, and zygomatics). Though a diagram accompanies the orbital form, scoring this character trait is complicated by the incorporation of two distinct morphological characteristics: the form border, and the shape of the entire orbit in the character's definition. Other traits with weaker gamma scores possess definitions distinguishing the extreme score from the masculine or feminine score only by the addition of the adjective "very." The definitions for the frontal and parietal eminences, the traits with poor interobserver reliability, describe the character only in terms of a series of ungrounded descriptors, ranging from "marked" to "missing."

Interobserver variation in determination of sex can be accounted for by the variation in the scoring of character traits as described above. Traits with lower gammas and high weights are the most problematic, as they contribute proportionately more to the determination of sex. In contrast, the characters with lower gammas and low weight do not exert such extreme effects when it comes to the overall determination of sex.

Table 2 presents the simulation of the determination of sex as a function of the IS, using various combinations of cranial character traits. The specific kappa for each of the five interobserver reliability coefficients is summarized, along with the mean and median IS scores and the analysis of interobserver difference, using the Wilcoxon sign rank.

When all 10 traits are used with an interval of uncertainty for the indeterminant category, the presumed sex ratio between the two observers is not significantly different, and Cohen's kappa is 0.61, indicative of good interrater reliability. When the number of traits used for determination of sex is reduced, interobserver reliability varies according to the quality of the character traits used. For these tests, we followed the same procedure described in our methodology, but limited IS scoring to: 1) the 4 characters accompanied by diagrams as compared to the 6 without diagrams, and 2) the characters with gamma values of 0.85 and above as compared to those with gammas below 0.85. Selective reduction of the number of traits improves interobserver reliability and concordance, minimizing discordance. The quality, rather than quantity, of character traits minimizes interobserver discordance. These results are similar to the findings for visual assessment with pelvic characters (Bruzek and Ferembach, 1992).

Though osteologists concur that "the most accurate and precise sexing and aging are obtained when it is possible to arrange many skeletal specimens within a series (to seriate) and to compare within a single biological population" (White, 1991, p. 306), in practice this ideal cannot always be attained. Frequently, individual specimens are found in isolation,

TABLE 2. Interobserver reliability and difference scores for measures of index of sexuality

IS measure	Number of male, female, and indeterminate specimens		Kappa[1]	Wilcoxon signed ranks (*P*-value)	Mean/median (per Rater)
	Rater 1	Rater 2			
All 10 traits	M: 18	23			
	F: 16	16	0.61	3.35 ($P = 0.001$)	Rater 1: 0.19/0.41
	I: 8	3			Rater 2: −0.11/−0.045
Visual items	M: 21	22			
	F: 18	16	0.62	2.21 ($P = 0.027$)	Rater 1: 0.07/0.28
	I: 3	4			Rater 2: −0.11/0.17
Nonvisual items	M: 14	3			
	F: 15	24	0.495	3.31 ($P = 0.001$)	Rater 1: 0.28/0.46
	I: 13	15			Rater 2: −0.11/−0.04
Items with gamma ≥0.85	M: 18	23			
	F: 16	16	0.785	2.73 ($P = 0.006$)	Rater 1: 0.13/0.38
	I: 8	3			Rater 2: −0.09/0.25
Items with gamma <0.85	M: 18	23			
	F: 16	16	0.30	3.14 ($P = 0.002$)	Rater 1: 0.27/0.40
	I: 8	3			Rater 2: −0.15/−0.10

[1] Kappa: 0.50–0.59 = fair interrater reliability: 0.60–0.69 = good interrater reliability; ≥0.70 = excellent interrater reliability (Fleiss, 1981).

with determination of sex occuring without the benefit of seriation. Therefore, the present assessment of interobserver reliability of methodologies for the determination of sex without seriation is appropriate.

Similarly, results derived from the visual assessment of traits for determination of sex are taken beyond the population-specific context inherent in the ideal approach described above. Mean IS values for a given sample can be used to describe the average population gracility or robusticity, and can be applied to comparisons of diverse samples (Ascadi and Nemeskeri, 1970). The present study shows that the observer's mean population IS scores derived for this sample differed at statistically significant levels. Table 2 presents mean and median IS scores for each observer, as determined using 10 traits and each of the truncated simulations described above. Interobserver differences were analyzed using Wilcoxon rank signs. Observer 1 tended to "masculinize" the sample in all five simulations, with sample mean scores ranging from 0.07–0.28. Observer 2 tended to "feminize" the sample, with sample mean scores ranging from −0.09 to −0.15. The subjective nature of IS values limits interobserver comparisons of the results. However, this index can be used for intraobserver comparison of diverse samples.

CONCLUSIONS

This study demonstrates how clarity of definition, rather than number of character traits, was critical for effective determination of sex by the visual assessment method. Although visual assessment of cranial traits did not result in significant interobserver discordance for determination of sex, sample mean IS scores differed significantly between the two observers. Most interobserver discordance derived from the specific character traits with subjective definitions and no accompanying diagrams. Use of fewer, more precisely defined character traits can improve interobserver concordance.

These findings have several implications. First, skeletal and particularly fossil remains are often fragmentary. This study demonstrates that all osteological fragments are not equal when it comes to the determination of sex, state of preservation notwithstanding. Second, because accurate determination of sex is critical for the reconstruction of biology and behavior of past populations, the reliability of sexing methodologies must be tested thoroughly. Finally, cross-population comparisons of the relative robusticity or degree of sexual dimorphism are most reliable when a single individual performs the analysis. Future investigators will rely on the ongoing documentation of Native American skeletal material by diverse investigators. Given the level of interobserver variation shown in this study, population comparisons of mean IS scores should be interpreted with caution when more than one investigator has performed the assessment.

ACKNOWLEDGMENTS

The authors thank Jeremy Sabloff, Director of the University of Pennsylvania Museum, for his kind invitation to J.B. to work on these materials. Many thanks also go to Alan Mann and Janet Monge for their generous support and comments on the manuscript, to Alex Pezzati, Archivist of the University of Pennsylvania Museum, for providing details about the Museum's Point Barrow excavations, and to Marilee Jones for her careful attention to details.

LITERATURE CITED

Ascadi G, Nemeskeri J. 1970. History of human lifespan and mortality. Budapest: Akademiai Kiado.

Bruzek J. 2002. A method for visual determination of sex, using the human hip bone. Am J Phys Anthropol 117:157–168.

Bruzek J, Ferembach D. 1992. Fiabilité de la méthode visuelle de détermination du sexe à partir du bassin, proposée par le

"Groupe de travail d'Anthropologues Européanes": application à l'os coxal. Estratto Arch Antropol Etnol 72:146–161.
Buikstra J, Ubelaker D, Aftandilian DA. 1994. Standards for data collection from human skeletal remains. Fayetteville, AK: Arkansas Archaeological Survey.
Burr D, Gerven DV, Gustav B. 1977. Sexual dimorphism and mechanics of the human hip: a multivariate assessment. Am J Phys Anthropol 47:273–78.
Chen PY, Popovich PM. 2002. Correlation: parametric and nonparametric measures. Thousand Oaks, CA: Sage.
Cohen J. 1960. A coefficient of agreement for nominal scales. Educ Psychol Meas 20:37–46.
Espeland M, Gallagher D, Tell G, Davison L, Platt O. 1990. Reliability of Tanner stage assessment in a multi-center study. Am J Hum Biol 2:503–510.
Ferembach D, Schwidetzky I, Stloukal M. 1980. Recommendations for age and sex diagnoses of skeletons. J Hum Evol 9:517–549.
Fleiss JL. 1981. Statistical methods for rates and proportions. 2nd ed. New York: John Wiley & Sons, Inc.
Giles E, Elliot O. 1963. Sex determination by discriminant function analysis of the crania. Am J Phys Anthropol 21:53–68.
Hogge J, Messmer J, Doan Q. 1994. Radiographic identification of unknown human remains and interpreter experience level. J Forensic Sci 39:373–377.
Krogman WM, Isçan MY. 1986. The Human Skeleton in Forensic Medicine. Springfield, IL: Charles C. Thomas.
Liversidge H. 1994. Accuracy of age estimation from developing teeth of a population of known age (0–5.4 years). Int J Osteoarchaeol 4:37–45.
MacLaughlin S, Wood B. 1995. Population variation in sexual dimorphism in the human innominate. Hum Evol 1:221–31.
Mason JA. 1930. Archaeological work in Alaska. Univ Mus Bull 1:10–13.
Masset C. 1987. Le recrutement d'un ensemble funéraire. In: Duday H, Masset C, editors. Anthropologie physique et archéologie: méthodes d'études des sépultures. Paris: Editions du CNRS. p 111–134.
McIlhenny EA. 1900. Collections and publications. Bull Free Mus Sci Art 2:172.
Meindl R, Lovejoy C, Mensforth R, DonCarlos L. 1985. Accuracy and direction of error in the sexing of the skeleton: implications for paleodemography. Am J Phys Anthropol 68:79–85.
Meindl RS, Russell KF, Lovejoy CO. 1990. Reliability of age at death in the Hamann-Todd collection: validity of subselection procedures used in blind tests of the summary age technique. Am J Phys Anthropol 83:349–57.
Novotny V. 1986. Sex determination of the pelvic bone: a systems approach. Anthropologie 24:197–206.
Novotny V, Isçan M, Loth S. 1993. Morphologic and osteometric assessment of age, sex, and race from the skull. In: Isçan MY, Helmer RP, editors. Forensic analysis of the skull. New York: Wiley-Liss. p. 71–88.
Ousley S, Jantz R. 1996. FORDISC 2.0. Personal computer forensic discriminant functions. Knoxville: University of Tennessee Press.
Shrout PE, Spitzer RL, Fleiss JL. 1987. Quantification of agreement in psychiatric diagnosis revisited. Arch Gen Psychiatry 44:172–177.
Siegel S, Castellan NJ. 1988. Nonparametric statistics for the behavioral sciences. 2nd ed. New York: McGraw-Hill.
St. Hoyme L, Isçan MY. 1989. Determination of sex and race: accuracy and assumptions. In Isçan MY, Kennedy K, editors. Reconstruction of life from the skeleton. New York: Alan R. Liss. p 53–93.
VanGerven D. 1974. The contribution of size and shape variation to patterns of sexual dimorphism of the human femur. Am J Phys Anthropol 37:49–60.
Van Valin W. 1941. Midnight sun and nocturnal moons: Eskimoland speaks. Caldwell, ID: Caxton Printers.
vanVark G, Schaafoma W. 1992. Advances in the quantitative analysis of skeletal morphology. In: Sanders S, Katzenberg M, editors. Skeletal biology of past peoples: research methods. New York: Wiley and Liss. p 225–257.
Waldron T, Rogers J. 1991. Inter-observer variation in coding osteoarthritis in human skeletal remains. Int J Osteoarchaeol 1:49–56.
Weiss KM. 1972. On the systematic bias in skeletal sexing. Am J Phys Anthropol 37:239–250.
White TD. 1991. Human Osteology. San Diego: California: Academic Press, Inc.