JOURNAL OF FORENSIC SCIENCES

J Forensic Sci, 2014
doi: 10.1111/1556-4029.12563
Available online at: onlinelibrary.wiley.com

PAPER

ANTHROPOLOGY

Alexandra R. Klales,[1] *M.S.; and Michael W. Kenyhercz,*[2] *M.S.*

Morphological Assessment of Ancestry using Cranial Macromorphoscopics*,†

ABSTRACT: Ancestry estimation is essential for biological profile estimation in forensic anthropology. Hefner (2009) and Osteoware (Smithsonian Institution, 2011) presented 16 macromorphoscopic traits that can be scored for standardized data collection and can also be used within a statistical framework to estimate ancestry. The primary purpose of this research was to examine the utility of these traits for assessing ancestry. Tests of observer agreement and the range of variation in trait expression were evaluated. A sample of 208 American whites and blacks from the Hamann–Todd Collection were scored, and several classification methods were utilized in accordance with Hefner (2009). Correct classifications for the pooled sex analyses ranged from 73.3% to 86.6% and from 46.7% to 64.3% when the sexes were analyzed independently. Interobserver agreement was variable and was found to be lower than that presented in Hefner (2009). Trait expression was variable in both groups and was generally consistent with Hefner's findings.

KEYWORDS: forensic science, forensic anthropology, biological profile, ancestry estimation, cranial nonmetrics, validation study, standardization

Biological profile estimation is fundamental for forensic anthropologists attempting to identify an unknown individual. Forensic anthropologists are tasked with estimating the parameters of an individual's biological profile to provide law enforcement with a description that can then narrow the list of potential victims and that can also be matched to missing person reports. Of the parameters for biological profile estimation, ancestry is the most controversial and, arguably, one of the most researched. Currently, there is a dichotomy in the field of anthropology concerning the use of "race" or ancestry. A full discussion of this debate is beyond the scope of this paper (for further discussion of this debate, see [1–3]). However, despite the "race debate," and for the sake of biological profile estimation, forensic anthropologists estimate ancestry and are frequently able to differentiate between groups with a reasonable degree of certainty based on morphological traits and skeletal measurements (3).

Recently, there has been a concentrated shift in forensic anthropology to develop biological profile methods that are both scientifically valid and reliable and that also meet the criteria set forth by the *Daubert* ruling (4). As a result, previously and traditionally used methods for biological profile estimation are being revisited for the tests of classification accuracy in independent samples and are also being revised to include statistical measures of accuracy and associated error rates (5). To date, multiple quantitative and qualitative skeletal methods have been developed to estimate ancestry. Despite being inherently more subjective than metric methods, morphological techniques continue to be taught and employed in ancestry estimation. However, research by Hefner and Ousley (6–8) noted that the skeletal nonmetric traits historically used in ancestry estimation are not unique to specific groups, but rather occur in all groups with varying frequencies. It then becomes the job of the forensic anthropologist to "identify degrees of phenotypic traits. . .observable in the skeleton that occur with high frequencies in certain populations" for use in ancestry estimation (9:118).

Hefner (8) responded to the call for methodological improvements with a newly proposed data collection module for the standardized scoring of morphological skull traits. Hefner (8) created ordinal scores with descriptions and corresponding illustrations for 11 morphological traits, termed macromorphoscopics, that have been commonly, or historically, applied to ancestry estimation in the skull. The traits utilized were compiled from a number of sources but were largely based on Hooton's Harvard List (8). The traits included by the author are displayed in Table 1 (see Hefner [8] for illustrations and complete descriptions of each trait). Hefner first analyzed the frequency distribution of these traits and then also used these traits within a statistical framework for the classification of group membership. A total of 747 individuals were scored by Hefner (8) for each of these traits from four broad geographical descent groups: African, Asian, European, and Native American. Significant differences were found between ancestral groups in the frequency distribution of all traits with the exception of the malar tubercle protrusion, which is located at the inferior junction of the zygomaticomaxillary suture. Additionally, eight of the 11 traits had moderate to high correlations with the exception of the malar tubercle, nasal overgrowth, and the zygomaticomaxillary suture.

[1]Department of Anthropology, University of Manitoba, 432 Fletcher Argue Building, 15 Chancellor Circle, Manitoba, MB R3T 2N2, Canada.
[2]Department of Anthropology, University of Alaska, Bunnell Building Room 405, PO Box 757720, Fairbanks, AK 99775.

*Presented in part at the 64th Meeting of the American Academy of Forensic Sciences, February 20–25, 2012, in Atlanta, GA.
†Finanaical Support provided by the University of Manitoba Graduate Fellowship & Manitoba Graduate Scholarship.

Received 26 June 2013; and in revised form 2 Nov. 2013; accepted 17 Nov. 2013.

TABLE 1—*Trait names and abbreviations.*

Trait Name	Abbreviation
Anterior Nasal Spine	ANS
Inferior Nasal Aperture	INA
Interorbital Breadth	IOB
Malar Tubercle	MT
Nasal Aperture Shape*	NAS
Nasal Aperture Width	NAW
Nasal Bone Contour	NBC
Nasal Bone Shape*	NBS
Nasal Overgrowth	NO
Nasofrontal Suture*	NFS
Orbital Shape*	OS
Post-Bregmatic Depression	PBD
Posterior Zygomatic Tubercle*	PZT
Supranasal Suture	SS
Transverse Palatine Suture	TPS
Zygomaticomaxillary Suture Closure	ZS

*Traits added to Osteoware (13).

Hefner (8) obtained no significant differences between males and females, so the two were pooled for each ancestral group. Using logistic regression, naïve Bayesian and *k*-nearest neighbor, Hefner reported high classification accuracies that ranged from 84% to 93% depending on the combination of traits used in the analysis and on the statistical classification method employed. Additionally, the collection procedures and trait scoring had relatively low observer error. Using Cohen and Fleiss's kappa statistics, eight of the 11 traits tested for intraobserver error and all 11 traits tested for interobserver had moderate to perfect levels of agreement according to parameters outlined by Landis and Koch (10). Results indicated that there was considerable variation present in the frequency distribution of the expected macromorphoscopic traits in specific ancestral groups; therefore, Hefner noted that ancestry estimation based only on the "experience-based" assessment of morphoscopic traits alone is "an art that is unscientific" (8:985). More recent research by L'Abbé et al. (11) and Vitek (12) have further assessed the utility of the Hefner traits for ancestry estimation.

Use of these above-mentioned traits within a statistical framework for classification can provide a method of ancestry estimation that is scientific, reliable, valid, and upholds the spirit and intent of the *Daubert* criteria. The results of Hefner (8) highlighted the potential utility of using standardized traits and data collection procedures for morphological ancestry estimation. Furthermore, high classification rates are encouraging and indicate the potential applicability of these traits for ancestry estimation in forensic anthropology. The Hefner (8) data collection procedure is included as one of the modules in the Osteoware Standardized Skeletal Documentation Software (13). The Osteoware package includes the 11 traits described in Hefner (8), in addition to: nasal aperture shape (NAS), nasal bone shape (NBS), nasofrontal suture (NFS), orbital shape (OS), and posterior zygomatic tubercle (PZT) (Table 1). Similarly, subsets of the aforementioned traits traits from Hefner (8) and Osteoware (13) have been combined with an analytical technique termed the Optimized Summed Scoring Attributes (OSSA) (14). With the OSSA method, users enter ordinal scores for six of the original 16 traits (8,13) to classify an unknown crania as white or black. The Osteoware data collection module (13) and the OSSA (14) classification method using a combination of these traits are currently both being used for ancestry estimation in active forensic cases. To be accepted as a valid and reliable method of data collection and classification, it remains imperative to test the consistency in scoring of these traits by multiple observers and to test how well the collected traits perform using different analytical tools and independent samples. The purpose of the current research was (i) to evaluate the degree of variation in traits between populations, (ii) to test the utility of the 16 macromorphoscopic traits for statistical classification of ancestry and differentiation between populations in an independent skeletal collection, and (iii) to test the reliability between observers for trait scoring.

Materials and Methods

Sample and Scoring

A sample of 208 crania was scored from the Hamann–Todd Osteological Collection (HTH) from two ancestral groups: American whites and American blacks (Table 2). Individuals were selected from the collection using a stratified random sample ensuring that males and females from both ancestral groups were evenly represented. Individuals in the HTH collection were of documented sex and ancestry from the late nineteenth and early twentieth centuries. All individuals had no apparent pathological or traumatic conditions present and were complete enough to score at least 14 of the 16 traits: the 11 from Hefner (8) and the five additional traits included in the Osteoware package (13) (Fig. 1). One observer (ARK) with previous familiarity and experience using the traits described by Hefner (8) and Osteoware (13) blindly scored each individual using the descriptions and illustrations from the Macromorphoscopic software program provided by Dr. Hefner. A second experienced observer (MWK) scored a subset 84 individuals for tests of interobserver error. A third observer with basic skeletal knowledge, but no experience using the aforementioned traits for ancestry estimation also scored a subsample of 10 male individuals.

Statistical Methods

Frequencies & Trait Correlations—First, trait frequency distributions were calculated for each of the four ancestry/sex groups. A Fisher–Freeman–Halton test was then run to determine whether significant differences in trait frequencies existed between the two ancestral groups. The Freeman–Halton extension of the Fisher's exact test is used to test the significance of categorical expressions in a contingency table; in this instance, the trait frequencies are used to test the contingency of each trait against each ancestry group. The tables are submitted to a chi-squared test to develop probability estimates for the interaction of each specific trait and ancestry, thus allowing for the assessment of which traits occur more significantly in certain populations, if at all. To determine which specific scores produced the significant differences between populations for traits, the residual for each trait score (difference between the observed and expected frequency) was converted into a *z*-score and compared with the critical value of ± 1.96 (level appropriate for an alpha value of 0.05). Differences in trait frequencies between ancestral

TABLE 2—*Sample included in this study.*

Ancestry/Sex Group	*n*
Black Females (BF)	52
White Females (WF)	54
Black Males (BM)	50
White Males (WM)	52

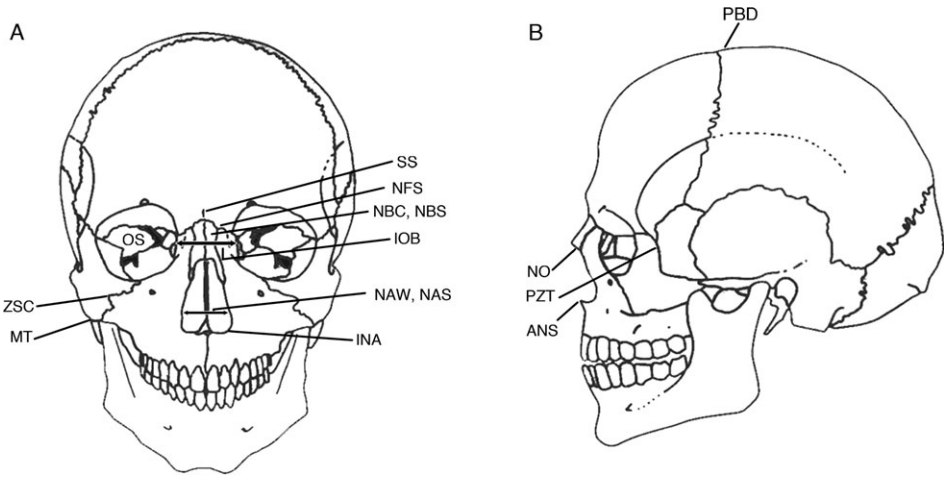

FIG. 1—*Anatomical location of the cranial macromorphoscopic traits. A) Anterior view, B) Lateral view Transverse palatine suture not pictured. See (8) for additional definitions and illustrations.*

groups provides insight into whether or not each group deviates from the traits commonly believed to be indicative of that particular population. Furthermore, the correlation results from this study can be compared with those reported by Hefner (8) for the measures of congruency between different samples. Polychoric correlations were calculated for the ordinal variables, while tetrachoric correlations were calculated for the binary variables for each group as a measure of association between traits. Polychoric correlations are best utilized on ordinal level variables wherein there is an underlying assumption of normality. Further, tetrachoric correlations are essentially polychoric correlations that have been adjusted for use when both traits under analysis are binary.

Classification—Classification accuracies were calculated using the multiple methods included in Hefner's original article, ordinal logistic regression (OLR), *k*-nearest neighbor (kNN), and naïve Bayesian (NB) with the inclusion of linear discriminant function analysis (LDFA) and random forest models (RFM). For each statistical method, both two-way (pooled sexes) and four-way (male and female separate) analyses were conducted for ancestry assessment. While some of the analyses are not normally used for ordinal data, as Walker (15) has noted, the goals of statistical analysis are more important than a rigid typology of data types; this is especially true of statistical classification methods, in which the correct classification rates of the reference groups is the overarching practical criterion (cf. [16] and citations within for more on the measurement-statistics debate). The LDFA utilized forward Wilks stepwise selection to select the variables that contributed most significantly to the discriminant functions and also leave-one-out-cross-validation to avoid unrealistic classification rates (17). Missing values were replaced with the variable mean in LDFA. In kNN analysis, classification of an individual is based on *k* most similar individuals in the reference sample and on the group identities of those individuals. Correct classification rates can vary with the selection of *k*, or nearest neighbors; therefore, group membership was set as the target so that the number of neighbors could automatically be selected (from three to seven) based on the model (18). Missing variable values were treated as invalid and were excluded from the analysis. The NB utilized a 20% culled training set from the total sample. For the NB, individuals with missing data were pairwise deleted prior to analysis. Ordinal logistic regression is designed for use with data that includes ordinal dependent variables and is an extension of logistic regression which can be used with continuous or dichotomous variables (19). Predicted values were calculated for the missing dependent variables and a 95% confidence interval was selected for analysis. Of the methods used for classification in this research, OLR is the most appropriate given the structure of the data. Random forest models were introduced by Breiman (20) as an improvement upon classification trees. The RFM used 2500 decision trees with each of the 16 variables being randomly assessed at each node. The present version of the "randomForest" package in R (21) does not handle missing data, so any specimens that did not contain 14 of the 16 variables under analysis were pairwise deleted prior to being subjected to the RFM. All analyses were conducted using SPSS (18) and R (21).

Interobserver Error—Cohen's kappa (*K*) was utilized to test the degree of agreement for trait scoring between the original observer and the additional two and to serve as a proxy for assessing the role of experience in trait scoring. Agreement levels used for this study are defined by Landis and Koch (10) and were chosen to compare the results of this research to Hefner's (8), which also utilized this scale.

Results

Frequencies & Trait Correlations

Frequency distributions of trait scores for each group are presented in Table 3. Results for pooled sexes were also included for comparision to Hefner's (8) results. Significant differences between the two ancestral groups were obtained at the $p < 0.05$ level for ANS, INA, IOB, NAW, NBC, NFS, and SS using a chi-square test of independence (Table 4). For INA, the residual values above the critical value revealed for score 1 (sloped) (1.9 in blacks and −1.9 in whites) and score 4 (right angle) (−3.0 in blacks and 2.9 in whites) indicating that the black population had higher then expected sloping or guttering of the nasal aperture, while whites had a nasal floor at a right angle to the maxilla with a superior rise of the anterior floor that was higher than expected. For IOB, the residual values above the critical value were obtained for score 1 (narrow) (−2.6 in blacks and 2.6 in whites) and score 3 (broad) (2.9 in blacks and −2.8 in whites) indicating that the black population had higher than expected

TABLE 3—*Trait frequencies in the four ancestry/sex groups and for pooled sexes.*

	Ancestry/Sex Group								Pooled Sexes		Hefner (8)	
	BF		WF		BM		WM					
Traits & Scores	*n*	%	*n*	%	*n*	%	*n*	%	B %	W %	B %	W %
ANS												
1	13	25.5	5	9.3	5	10.0	3	5.7	17.8	7.5	69.7	36.3
2	25	49.0	28	51.9	30	60.0	25	47.2	54.5	49.5	20.2	26.0
3	13	25.5	21	38.9	15	30.0	25	47.2	27.7	43.0	10.1	37.7
INA												
1	5	9.8	0	0.0	6	12.0	2	3.8	10.9	1.9	29.4	0.7
2	16	31.4	2	3.7	12	24.0	4	7.5	27.7	5.6	28.9	3.4
3	19	37.3	15	27.8	22	44.0	14	26.4	40.6	27.1	21.6	24.0
4	8	15.7	32	59.3	8	16.0	20	37.7	15.8	48.6	13.3	41.1
5	3	5.9	5	9.3	2	4.0	13	24.5	5.0	16.8	6.9	30.8
IOB												
1	10	19.6	28	51.9	6	11.1	19	35.8	15.8	43.9	9.6	30.8
2	24	47.1	18	33.3	12	22.2	24	45.3	35.6	39.3	34.4	63.0
3	17	33.3	8	14.8	32	59.3	10	18.9	48.5	16.8	56.0	6.2
NAW												
1	1	2.0	17	31.5	1	2.0	12	22.6	2.0	27.1	3.7	54.1
2	24	47.1	26	48.1	28	56.0	37	69.8	51.5	58.9	40.8	32.9
3	26	51.0	11	20.4	21	42.0	4	7.5	46.5	14.0	55.5	13.1
MT												
0	12	23.5	22	40.7	18	36.0	13	24.5	29.7	32.7	50.5	51.4
1	25	49.0	22	40.7	12	24.0	21	39.6	36.6	40.2	27.5	32.2
2	12	23.5	9	16.7	15	30.0	16	30.2	26.7	23.4	14.7	12.3
3	2	3.9	1	1.9	5	10.0	3	5.7	6.9	3.7	7.3	4.1
ZS												
0	5	9.8	14	25.9	12	24.0	7	13.2	16.8	19.6	5.1	1.5
1	25	49.0	17	31.5	22	44.0	20	37.7	46.5	34.6	31.6	37.0
2	21	41.2	22	40.7	14	28.0	25	47.2	34.7	43.9	49.7	42.2
3	0	0.0	1	1.9	2	4.0	1	1.9	2.0	1.9	13.6	19.3
NO												
0	20	57.1	24	58.8	27	69.2	19	45.2	63.5	51.8	68.1	52.7
1	15	42.9	17	41.5	12	30.8	23	54.8	36.5	48.2	31.9	49.2
PBD												
0	24	47.1	29	53.7	26	52.0	38	71.7	49.5	62.6	52.8	82.9
1	27	52.9	25	46.3	24	48.0	15	28.3	50.5	37.4	47.2	17.1
SS												
0	36	70.6	20	37.7	11	22.0	7	13.2	46.5	25.5	42.8	39.0
1	3	5.9	9	17.0	8	16.0	9	17.0	10.9	17.0	31.2	39.0
2	12	23.5	24	45.3	31	62.0	37	69.8	42.6	57.5	26.0	22.0
TPS												
1	12	26.1	12	22.2	10	23.8	16	34.8	25.0	29.5	18.3	29.0
2	16	34.8	16	29.6	12	28.6	12	26.1	31.8	29.5	47.2	27.6
3	12	26.1	12	22.2	17	40.5	12	26.1	33.0	25.3	25.0	33.8
4	6	13.0	9	16.7	3	7.1	6	13.0	10.2	15.8	9.4	9.7
NBC												
0	7	13.7	1	1.9	1	2.0	0	0.0	7.9	0.9	52.3	7.5
1	14	27.5	6	11.1	4	8.0	8	15.1	17.8	13.1	22.9	15.8
2	2	3.9	0	0.0	4	8.0	0	0.0	5.9	0.0	10.1	18.5
3	11	21.6	28	51.9	16	32.0	28	52.8	26.7	52.3	10.6	25.3
4	17	33.3	19	35.2	25	50.0	17	32.1	41.6	33.6	4.1	32.9
OS												
1	10	19.6	11	20.4	11	22.0	11	20.8	20.8	20.6		
2	31	60.8	23	42.6	23	46.0	31	58.5	53.5	50.5		
3	10	19.6	20	37.0	16	32.0	11	20.8	25.7	29.0		
PZT												
0	4	7.8	2	3.7	1	2.0	1	1.9	5.0	2.8		
1	20	39.2	27	50.0	13	26.0	21	39.6	32.7	44.9		
2	17	33.3	18	33.3	23	46.0	22	41.5	39.6	37.4		
3	10	19.6	7	13.0	13	26.0	9	17.0	22.8	15.0		
NBS												
1	4	7.8	5	9.3	7	14.0	5	9.4	10.9	10.3		
2	26	51.0	36	66.7	25	50.0	21	58.5	50.5	58.8		
3	13	25.5	10	18.5	9	18.0	15	28.3	21.8	25.8		
4	8	15.7	3	5.6	9	18.0	2	3.8	16.8	5.2		
NFS												
1	12	23.5	16	29.6	16	29.6	33	62.3	27.7	46.2		
2	8	15.7	11	20.4	13	24.1	6	11.3	20.8	16.0		
3	4	7.8	1	1.9	6	11.1	3	5.7	9.9	3.8		

TABLE 3—*Continued.*

	Ancestry/Sex Group								Pooled Sexes		Hefner (8)	
	BF		WF		BM		WM					
Traits & Scores	*n*	%	*n*	%	*n*	%	*n*	%	B %	W %	B %	W %
4	27	52.9	26	48.1	15	27.8	10	18.9	41.9	34.0		
NAS												
1	41	80.4	44	81.5	40	74.1	44	83.0	80.2	82.2		
2	9	17.6	3	5.6	4	7.4	3	5.7	12.9	5.6		
3	1	2.0	7	13.0	6	11.1	6	11.3	6.9	12.1		

TABLE 4—*Significance of trait frequencies by group using a Chi-squared test.*

Trait	χ^2
ANS	**0.02**
INA	**<0.001**
IOB	**<0.001**
MT	0.68
NAS	0.11
NAW	**<0.001**
NBC	**<0.001**
NBS	0.07
NFS	**0.03**
NO	0.09
OS	0.87
PBD	0.06
PZT	0.24
SS	**0.01**
TPS	0.44
ZS	0.39

Values significant at $p < 0.05$ are in bold.

broad IOB and conversely whites had higher than expected narrow IOB. A similar trend was noticed for NAW with score 1 (narrow) (−3.4 in blacks and 3.3 in whites) and score 3 (broad) (3.1 in blacks and −3.0 in whites). Finally, score 3 for NBC (−2.1 in blacks and 2.0 in whites) indicates that whites had a higher propensity for a steep contour of the nasals rather than a low, rounded or Quonset-hut appearance. Polychoric and tetrachoric correlations for each of the variables are presented in Table 5. The only combinations to demonstrate at least low-to-moderate levels of correlation (0.3–0.5) were between INA and ANS (0.41), NAW and ANS (−0.41), IOB and INA (−0.37), NAW and INA (0.38), and NAW and IOB (0.46).

Classification

Classification accuracies in the two-way analyses ranged from 73.3% to 88.6% depending on the method; OLR performed best for pooled sex ancestry estimation (Table 6). In the four-way analyses in which sex and ancestry groups were separated, NB performed best (64.3% accuracy). Accuracy ranged from 46.7% to 60.4% for the remaining methods, with OLR classifying best after NB (Table 7). In LDFA, three traits were forward Wilk's stepwise selected in the two-way analysis (Table 8), while in the four-way analysis, five traits were selected (Table 9). In both, inferior nasal aperture accounted for the greatest separation and was the first variable selected. In the four-way, white females classified much lower than the other three groups. In the two-way kNN analysis, the model selected three traits (IOB, NAW, PZT) and $k = 5$, while in the four-way, three different traits were selected (NFS, ANS, INA) and $k = 5$. As with the LDFA, white females classified the poorest. Because missing values are excluded in kNN, the sample size was reduced to 63 black and 71 white. In the NB, the four-way analysis utilized all traits and yielded a total correct classification of 64.3%. The two-way ancestry analysis had a slightly higher total correct classification at 76.2%, wherein the white sample classified higher compared to the black sample. The RFM four-way analysis yielded a total correct classification of 46.7%. Each of the available variables was tested at each node, but there was a preference in each decision node for either NAW or IOB. In the two-way RFM, total correct classification was 73.3%, with both whites and blacks demonstrating approximately equal classification accuracies. The sample sizes for both NB and RFM were reduced to 131 individuals (63 black, 68 white) due to the pairwise deletion of missing values.

Observer Error

Using Cohen's Kappa, nine of the 11 traits had values lower than those reported by Hefner (8). Trait agreement was also generally lower than presented by L'Abbé et al. (11). Based on parameters outlined in Landis and Koch (10), four traits had slight agreement (ANS, NBC, NBS, ZS), six traits showed fair agreement (INA, MT, NAS, NFS, NO, PZT) and six exhibited moderate levels of agreement (IOB, NAW, OS, PBD, SS, TPS) when the two experienced observers' scores were compared (Table 10). Two traits, PBD ($K = 0.411$) and IOB ($K = 0.412$), had higher kappa values than Hefner reported, but still only showed moderate levels of agreement. The five additional traits not included in Hefner (8) but added to Osteoware (13) were also assessed. Each of the five additional traits fell into the slight to moderate level of agreement categories (Table 10). Agreement between the experienced (Exp) and the inexperienced (Inexp) observer varied considerably. Four of the traits displayed agreement values less than chance (ANS, INA, MT, PBD), four showed slight agreement (NAW, NBC, NBS, NFS), four exhibited fair agreement (IOB, IOS, PZT, ZS), three had moderate agreement (NAS, NO, SS), and one showed substantial agreement (TPS) (Table 10). Five of the traits had higher levels of agreement between the inexperienced/experienced observer pairing than between the experienced observers pairing: NAS, NO, PZT, TPS, and ZS (Table 10).

Discussion

Consistent with Hefner's original study (8:991), a wide range of variation was revealed in trait expression within and between groups and supports the author's notion that "compiled trait lists for ancestry ignore a substantial amount of variation within

TABLE 5—*Polychoric and Tetrachoric Correlation matrix.*

Traits	ANS	INA	IOB	NAW	MT	ZS	NO	PBD	SS	TPS	NBC	OS	PZT	NBS	NFS	NAS
ANS	–	0.41	−0.1	−0.41	0.0	0.15	0.13	−0.0	0.14	−0.1	0.28	−0	−0.1	−0.1	0.02	−0.2
INA		–	−0.4	−0.38	0.0	0.12	0.05	−0.1	0.01	−0.1	0.16	0.05	−0.1	−0.1	0.01	−0.1
IOB			–	0.46	0.1	−0.1	−0.0	0.02	−0.2	0.0	−0.16	0.04	0.08	−0.1	−0.1	−0.1
NAW				–	0.1	−0.1	0.11	0.04	−0.3	0.05	−0.2	−0.1	−0.1	0.08	0.08	−0.0
MT					–	0.0	−0.1	0.12	0.03	0.09	−0.27	0.05	0.14	−0.2	−0.1	−0.2
ZS						–	−0.1	−0.1	0.02	−0.1	−0.01	0.17	0.16	0.0	0.06	0.08
NO							–	−0.3	0.14	−0.1	0.13	−0.1	0.13	−0.1	−0.2	−0.2
PBD								–	0.01	0.11	−0.13	−0.0	−0.1	0.1	0.08	0.05
SS									–	−0.2	0.04	−0.0	0.15	−0.0	−0.1	−0.2
TPS										–	−0.03	−0.1	−0.1	0.07	0.22	−0.1
NBC											–	−0.1	0.09	0.03	−0.0	−0.1
OS												–	−0.1	0.01	−0.1	0.13
PZT													–	−0.1	−0.1	−0.1
NBS														–	0.07	0.0
NFS															–	−0.1
NAS																–

TABLE 6—*Classification (%) accuracies by method in the two-way analysis for ancestry with pooled sexes.*

Method	Black	White	Total
OLR	85.7	91.5	88.6
NB	71.4	80.9	76.2
kNN (3)	68.3	78.9	73.6
LDFA	71.3	76.6	74.0
RFM	74.6	72.1	73.3

TABLE 7—*Classification (%) accuracies by method in the four-way analysis for separate ancestry/sex groups.*

Method	BF	WF	BM	WM	Total
NB	74.2	62.5	61.8	58.8	64.3
OLR	71.0	52.8	43.8	74.3	60.4
kNN (3)	41.9	38.9	46.9	73.5	50.3
LDFA	56.9	35.2	56.0	52.8	50.2
RFM	51.6	50.0	38.2	47.1	46.7

TABLE 8—*Variables retained in two-way LDFA.*

Variable	Tolerance	F to Remove	Wilks' Lambda
INA	0.95	20.72	0.75
NAW	0.92	21.11	0.75
NO	0.95	4.68	0.67

TABLE 9—*Variables retained in four-way LDFA.*

Variable	Tolerance	F to Remove	Wilks' Lambda
INA	0.92	9.78	0.50
SS	0.95	9.60	0.49
IOB	0.82	9.09	0.49
NBC	0.91	5.26	0.45
NFS	0.94	5.21	0.45

groups." For many of the traits, more intermediate scores were obtained in the current study versus extreme scores found in Hefner (8). For example, the absence of a pronounced anterior nasal spine (score 1) and the presence of a guttered or sloping inferior nasal aperture (score 1 & 2) are traits typically associated with black populations. Hefner's sample had a higher percentage of ANS score 1 (69.7% vs. 17.8% in the present study) and INA score 1 (29.4% vs. 10.9% in the present study) in the black populations (Table 3). In contrast, the frequencies for both

TABLE 10—*Comparison of interobserver agreements using Cohen's Kappa by trait.*

Trait	Exp Observer* $n = 84$	Inexp Observer† $n = 10$	Hefner (8) $n = 7$	L'Abbe et al. (11) $n = 30$
NBC	0.141	0.032	0.231	0.54
ANS	0.165	−0.250	0.506	0.55
ZS	0.166	0.357	0.541	0.11
NBS	0.198	0.155	–	–
PZT	0.251	0.365	–	–
NFS	0.281	0.032	–	–
INA	0.284	−0.522	0.376	0.65
NAS	0.324	0.412	–	–
NO	0.374	0.500	1.000	0.73
MT	0.382	−0.538	0.470	0.44
PBD	0.411	−0.250	0.232	–
IOB	0.412	0.242	0.325	0.44
OS	0.453	0.375	–	–
TPS	0.485	0.767	0.700	0.38
NAW	0.507	0.167	0.732	0.56
SS	0.586	0.412	0.650	–

*Experienced observer.
†Inexperienced observer.

of these traits in the current research are centered among the intermediate scores (ANS score 2 and INA scores 3 & 4) for both the black and white populations suggesting a greater degree of overlap in expression between the two population groups. The same trend was true for NAW and PBD. Conversely, in the present research, a higher percentage of extreme scores were obtained in IOB and TPS than in the original research by Hefner (8), while similar score distributions were revealed for MT and NO. The differences in observed trait frequencies are not unusual given the different study samples utilized and the compositional differences of the two samples: in the original research, contemporary populations were combined with historical populations from two regions, while the sample utilized in this study consists solely of African Americans from a historical sample (discussed in more detail below). However, and most importantly, these results corroborate Hefner's notion that phenotypic variation is present between blacks and whites in trait expression and this likely contributes to the continued application and utility of morphological traits for ancestry estimation. Given the nature of trait frequency overlaps, adequate use of the traits will largely depend on trait selection and the multivariate classification techniques employed by the researcher.

Significant differences were found in the frequency distribution of nearly half of the traits between ancestral groups (Table 4). These aforementioned findings are generally inconsistent with Hefner (8) and L'Abbé (11) and are likely due to the overall intermediacy displayed in the trait distribution scores described previously. All significantly different traits centered on the nasal region, following conventional wisdom of nonmetric trait assessment between the two groups. Specifically the residuals for INA, IOB, NAW, and NBC clearly adhere to the traditional knowledge or expectations for differences between white and black populations. Not surprisingly, the traits that demonstrated significant differences were also likely to be more highly correlated (Table 5); in particular, ANS and INA, IOB and INA, NAW and ANS, NAW and INA, and NAW to IOB (Table 5). Both the frequency distribution and correlation of midface variables were also noted by Hefner (8): however, the lower correlations between traits in the present study may, again, be an effect of the more intermediate trait expressions in the study sample, a product of the additional traits scored that were available in Osteoware but not included in Hefner's original study (8), or because wherein Hefner (8) evaluated four ancestral groups, and L'Abbe et al. (11) evaluated three, the present study only evaluated two.

Classification accuracies varied considerably by statistical method utilized. Ordinal logistic regression performed best in the two-way analysis, while naïve Bayesian performed best in the four-way analysis (Tables 6 and 7). For the two-way analysis, all methods with the exception of ordinal logistic regression performed worse than accuracy rates reported by Hefner (8). Poorer classification accuracies may be an artifact of the sample differences. The black sample included by Hefner is comprised of individuals native to Africa ($n = 32$), as well as African Americans from an early twentieth-century population ($n = 150$) and a modern population ($n = 38$). The sample utilized in the current study consists solely of African Americans from a late nineteenth and early twentieth century sample. Similarly, the European or white sample employed by Hefner included native Europeans ($n = 15$) and American whites ($n = 170$) from an early twentieth century population, while the current study only employed European Americans from a late nineteenth and early twentieth century sample. Secular change in trait expression has been documented in both European and African Americans for many of the traits included by Hefner and may be a contributing factor to the differences obtained in this study as compared to Hefner's original work (12). Presumably, using a sample from a single temporal period should have produced greater classification accuracy than when combining two temporal periods given secular change; however, the opposite proved true in this study. Furthermore, for some of the analyses (e.g., kNN, NB, and RFM), the original sample size of 208 was reduced due to missing variables and this may have impacted classification accuracy. Two traits involving the course and pattern of sutures (SS and TPS) were impossible to score due to obliteration in many cases. Nasal overgrowth was the trait most frequently omitted as the bony area is extremely fragile and was often damaged postmortem. Both the NO and SS were utilized in the two and four-way LDFA using stepwise selection of variables. Given their propensity for damage or unscorability this is problematic. Revisions to the Osteoware (13) data collection module and the Macromorphoscopic (8) program could include an ordinal score for obliteration of these two variables to resolve this problem. While not designed for ordinal data, LDFA outperformed some of the methods more appropriate for ordinal data in both the two- and four-way analyses, though as Hand (23) pointed out, the leave-one-out-cross-validation procedure does not utilize the assumption of normality, thereby not explicitly violating any assumptions of the linear discriminant function. Random forest modeling performed the poorest in both the two- and four-way analyses, though its utility in combining ordinal and continuous data was presented by Hefner et al. (22).

The Hefner collection procedures discussed previously, in conjunction with the OSSA method (14), are currently being used in active forensic casework primarily because they seem to work. In on-going research by the authors, the OSSA classification method, using the Hefner collection procedures, achieved 100% accuracy at assigning the ancestry of an unknown individual in a small sample of active forensic cases with positive identification (24). The varying classification accuracy indicates that the appropriate statistical treatment of the collected data is paramount in the successful application of the collection procedures for assigning ancestry to an unknown individual. Furthermore, the OSSA method of classification using a subset of Hefner's (8) traits may provide better classification than trait combinations in the other statistical methods tested in the current study.

Finally, results from the current study suggest that inexperienced observers should become intimately familiar with the trait definitions and illustrations provided by Hefner prior to utilizing the data collection protocols. Previous research by the authors (24) demonstrated that higher classification accuracy was achieved by the observer with more experience. Interobserver agreement was lower than in previous studies assessing these traits (8,11) and below substantial agreement between observers implies that this method may require extensive practical experience with the traits, the scores, and the illustrations and definitions before it can be confidently utilized as a data collection technique. Most importantly, the collection procedures outlined by Hefner (8) offers a means of standardization in data recording for traits that were historically ill-defined and often ambiguous. However, standardization through ordinal scores can be problematic if an observed trait does not fit into one of the available scores/definitions. The "atypical" trait occurrence was observed by the authors in a limited number of cases, specifically in regard to the palatine suture and frontonasal suture; the prior proved difficult because that suture closes and obliterates with senescence. However, when a trace TPS was visible, it was scored, which may have added to the lower agreement between observers. Additionally, the frontonasal suture sometimes presented either a completely atypical expression as compared to the Hefner standards, or an intermediate expression between two scores. When this was the case, the observers forced the expression into one of the predefined categories, as recommended by J.T. Hefner (pers. comm.), obviously not consistently with one another and thus leading to the lower agreement in this trait. Due to the previously mentioned observations, it is recommended that, if an observed trait is significantly different from the available ordinal scores, it should remain unrecorded instead of being forced into a "closest" available option.

Conclusions

The strength of the Hefner collection procedure is the standardized avenue in which the data are recorded, thus allowing compliance to the rigorous specifications of both *Daubert* and the Scientific Working Group for Forensic Anthropology's best practices (25). The illustrations and descriptions provided vastly improves upon the old typological methods, such as Rhine (26),

which do not express ranges of variation, but rather discrete typologies that generate a suite of traits supposedly characteristic of certain population groups (6). The Hefner data collection procedures do not assume that discrete traits belong only to specific population groups, but rather demonstrates a way to collect data for the analyst who can then use certain statistical methods to classify an individual's ancestry. It is then up to the analyst to apply the appropriate statistical procedures and interpret the data accordingly for classification of ancestry. Based on the present research, ordinal logistic regression is the best method for accurately classifying ancestral groups (86.6% total correct) using the Hefner data collection procedures when males and females can be pooled. Naïve Bayesian performed best in four-way analysis (64.3% total correct) with each ancestral group divided by sex. Results from this study suggest that observers should become familiar with the traits and range of variation present prior to scoring them, as experience impacted classification accuracy and observer agreement in trait scoring. Extremely low observer agreement for some traits suggest that perhaps these should be eliminated during analyses to avoid problems of replicability.

Acknowledgments

The authors would like to thank Lyman Jellema, of the Cleveland Museum of Natural History, for access to the Hamann–Todd Collection and to William Kenyhercz for assisting with data collection and participating in the observer error portion of the research. Thanks also go to Dr. Joseph Hefner for providing the Macromorphoscopic software program and for commentary on the manuscript. Lastly, thank you to the anonymous reviewers for their positive critiques and comments.

References

1. Smay D, Armelagos G. Galileo wept: a critical assessment of the use of race in forensic anthropology. Trans Anthropol 2000;9:19–29.
2. Sauer NJ. Forensic anthropology and the concept of race: If races don't exist, why are forensic anthropologists so good at identifying them? Soc Sci Med 1992;34:107–11.
3. Ousley SD, Jantz R, Freid F. Understanding race and human variation: why forensic anthropologists are good at identifying race. Am J Phys Anthropol 2009;139:68–76.
4. *Daubert v. Merrell Dow Pharmaceuticals*, US Supreme Court 509. U.S.579,113S.Ct.2786, 125L, Ed.2d 469, 1993.
5. Dirkmaat DC, Cabo LC, Ousley SD, Symes SA. New perspectives in forensic anthropology. Yearb Phys Anthropol 2008;51:33–52.
6. Ousley SD, Hefner JT. The statistical determination of ancestry. Proceedings of the 57th Annual Meeting of the American Academy of Forensic Sciences; 2005 Feb 21–26; New Orleans, LA. Colorado Springs, CO: American Academy of Forensic Sciences, 2005.
7. Hefner JT, Ousley SD. Morphoscopic traits and the statistical determination of ancestry II. Proceedings of the 58th Annual Meeting of the American Academy of Forensic Sciences; 2006 Feb 20–25; Seattle, WAColorado Springs, CO: American Academy of Forensic Sciences, 2006.
8. Hefner JT. Cranial nonmetric variation and estimating ancestry. J Forensic Sci 2009;54:985–95.
9. Nafte M. Flesh and bone, 2nd rev. edn. Durham, NC: Carolina Academic Press, 2009.
10. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977;33:159–74.
11. L'Abbé EN, Rooyen CV, Nawrocki SP, Becker PJ. An evaluation of non-metric cranial traits used to estimate ancestry in a South African sample. Forensic Sci Int 2011;209:195.e1–7.
12. Vitek CL. A critical analysis of the use of non-metric traits for ancestry estimation among two North American population samples [thesis]. Knoxville, TN: University of Tennessee, 2012.
13. Osteoware [computer program]. Standardized skeletal documentation software. Washinton, DC: Smithsonian Institution National Museum of Natural History, 2011; http://osteoware.si.edu/ (accessed Septemer 1, 2013).
14. Hefner JT, Ousley SD. Statistical classification methods for estimating ancestry using morphoscopic traits. J Forensic Sci 2014;59:883–90.
15. Walker PL. Sexing skulls using discriminant function analysis of visually assessed traits. Am J Phys Anthropol 2008;136:39–50.
16. Velleman PF, Wilkinson L. Nominal, ordinal, interval, and ratio typologies are misleading. Am Stat 1993;47:65–72.
17. Tabachnick BG, Fidell LS. Using multivariate statistics. Boston, MA: Allyn & Bacon, 2001.
18. SPSS [computer program]. SPSS Statistics for Windows, Version 17.0. Chicago, IL: SPSS Inc., 2008.
19. McCullagh P. Regression models for ordinal data. J R Stat Soc Series B 1980;42:109–42.
20. Breiman L. Random forests. Mach Learn 2001;45:5–32.
21. R [computer program]. R Development Core Team: a language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing, 2008.
22. Hefner JT, Spradley MK, Anderson BE. Ancestry estimation using random forest modeling. Proceedings of the 63rd Annual Meeting of the American Academy of Forensic Sciences; 2011 Feb 21–26; Chicago, IL. Colorado Springs, CO: American Academy of Forensic Sciences, 2011.
23. Hand DJ. Classification and discrimination. New York, NY: Wiley, 1981.
24. Klales AR, Kenyhercz MW. Non-metric assessment of ancestry through cranial macromorphoscopics: a validation of the Hefner (2009) method. Proceedings of the 64th Annual Meeting of the American Academy of Forensic Sciences; 2012 Feb 20–25; Atlanta GA. Colorado Springs, CO: American Academy of Forensic Sciences, 2012.
25. SWGANTH, Scientific Working Group for Forensic Anthropology Ancestry Assessment Draft, 2012; http://swganth.startlogic.com/Ancestry%20Assessment.pdf (accessed September 1, 2013).
26. Rhine S. Non-metric skull racing. In: Gill GW, Rhine S, editors. Skeletal attribution of race: methods for forensic anthropology. Albuquerque, NM: Maxwell Museum of Anthropological Papers, 1990;4:9–20.

Additional information and reprint requests:
Alexandra R. Klales, M.S.
1515 E. Woodbank Way
West Chester, PA 19380
E-mail: alexandra.klales@gmail.com