JOURNAL OF FORENSIC SCIENCES

J Forensic Sci, November 2011, Vol. 56, No. 6
doi: 10.1111/j.1556-4029.2011.01864.x
Available online at: onlinelibrary.wiley.com

PAPER

ANTHROPOLOGY

Polly R. Husmann,[1,†] *Ph.D. and David R. Samson,*[1] *M.A.*

In the Eye of the Beholder: Sex and Race Estimation using the Human Orbital Aperture*

ABSTRACT: From the works of Broca and Krogman to modern-day Jantz and Buikstra, the orbit has been used for both quantitative and qualitative sex and race estimation. This study evaluates the practical value of these estimations. Orbital height and breadth were measured to determine the orbital index and assess differences between men and women or black people and white people in the Hamann–Todd Collection. Replicability of these measures was also examined. Finally, a geometric morphometric study was performed to assess shape differences using the entire margin. Significant differences were found in both the index and the geomorphometric study; however, further investigation revealed that the differences were of little practical use. The measurement differences were found to be smaller than intra-observer error, while the geometric morphometric analysis demonstrated that minimal percentage of variation in shape was attributable to group differences. Thus, these techniques should not be used to estimate sex or race.

KEYWORDS: forensic science, forensic anthropology, sex estimation, race estimation, orbit, replicability

Orbits have been used to estimate the race and sex of individuals for over a hundred years. Scientists have employed numerous methodologies, many of which have since been tested and invalidated, yet some remain in textbooks and manuals today (1,2). Because the *Daubert* decision in 1993 requires scientific testing of all forensic techniques used in court, including peer review, error measurements, and general community acceptance, researchers are actively investigating the validity of these techniques (3). This paper assesses the accuracy and applicability of two orbital metrics: the "orbital index" (the ratio of height to breadth of the orbit) and a new geometric morphometric characterization of the shape of the orbital margin. We then examine how well these two metrics fare at distinguishing between sex and race groups and thus how much the orbit should be utilized in these distinctions.

Background Literature on Human Orbits

In 1875, Paul Broca devised an index to assess orbital size and shape quantitatively using height and breadth measurements (4). It is calculated as the height of the orbit divided by the orbital breadth and then multiplied by 100. Piquet categorized shapes based on this index as high, medium, and low or Hypsiconch, Mesoconch, and Chamaeconch, respectively (5–7). These categories have since been associated with race and/or temporal affiliations. Euro-Asians, for example, supposedly fall into the high category, Europeans and Africans into the medium category, and Aboriginal Australians, Melanesians, and prehistoric populations into the low category (5). Different versions of this system have been elaborated since that time (2,7), including orbital breadth and height use in the latest version of FORDISC (8).

Wilton Krogman developed a different way of assessing orbits. He took a qualitative approach and concluded that northern and southern Europeans had angular orbits, while central Europeans' and Asians' orbits were more rounded. African orbits were deemed more rectangular. Krogman also assessed the differences between the sexes and stated that female orbits were "rounded, higher, relatively larger, with sharp margins," while male orbits were "squared, lower, relatively smaller, with rounded margins" (9–11). These methods, or close variations of them, also continue to be seen in textbooks and research manuals (1,12,13). Most recently, Komar and Buikstra (3) report that male orbits are "squared, low" and with "rounded margins," while female orbits are "rounded, high" and with "sharp margins."

In this study, we assess the index created by Broca to determine its practical value. In addition, we made a geometric morphometric study to characterize the shape of the orbits, including their roundedness, as an objective method to assess Krogman's scheme. Based on the work mentioned above, we will test two main hypotheses with each of these metrics: (i) there will be a significant difference in the shape of orbits between races and (ii) there will be a significant difference in the shape of orbits between the sexes. These hypotheses stand for both the orbital index analysis and the geomorphometric examination.

Methods

For this study, the Hamann–Todd Collection was used to assess usefulness of orbits for both sex and race. The Hamann–Todd Collection at the Cleveland Museum of Natural History has been historically used for studies of race and sex owing to the existence of records indicating sex, race, and age at death. These records were used in this study to determine age over 25, race, and sex. The Hamann–Todd Osteological Collection is composed of more

[1]Indiana University, Student Building 130, Bloomington, IN 47405.

*Presented at the 78th Annual Meeting of the American Association of Physical Anthropologists, March 31–April 4, 2009, in Chicago, IL.

[†]Current address: West Virginia School of Osteopathic Medicine, 400 N. Lee St., Lewisburg, WV 24901.

Received 22 July 2010; and in revised form 28 Sept. 2010; accepted 8 Oct. 2010.

than 3000 individuals that were born between the years 1825 and 1910. They are mostly the remains of unclaimed bodies (most likely from lower socioeconomic classes) that were collected by the anatomists C. A. Hamann and T. W. Todd from Case Western Reserve University in Cleveland, OH (14). We evaluated 762 adults, including 184 black women, 236 black men, 110 white women, and 232 white men.

On these specimens, measurements were taken to a hundredth of a millimeter in order to assess height and breadth of the left orbit whenever possible. In the case of an incomplete left orbit, the right orbit was then used. Digital calipers were utilized to minimize errors. Measurements were taken using the FORDISC skeletal methodology as this was deemed the most commonly utilized source for assessing race in skeletal populations. This methodology stated that orbital breadth is taken from dacryon to ectoconchion. Orbital height was then taken perpendicular to breadth from superior to inferior margin, thus bisecting the orbit. Also, any notches and/or depressions were to be avoided (2). These measurements were then placed into the equation devised by Broca to achieve the orbital index.

For the geomorphometric analysis, the skull was first set up in a Frankfurt horizontal plane with respect to the camera (The superior aspect of the auditory meatus was put directly in a horizontal line with the inferior aspect of the orbit.) (15). It was placed on foam rings to achieve this end, as was the camera that was located 35 cm away. Black velvet was used to decrease the amount of glare on the photographs, and the flash was avoided for the same reason. All of these steps were undertaken in order to standardize the photographs so that as little artifactual variance would be produced in the outlines as possible.

An Olympus Stylus 600 with 6.0 megapixels (Center Valley, PA) was used to capture the image of the orbits. The image was then taken into the Gimp 2.0 program where an outline was manually placed along the orbital margin (16). The outlined images were then compiled into a single file using the tpsUtil program (17). This file was imported into tpsDig2 (18) where geomorphometric outlines were placed on the interior of the painted outline with 200 data points, and the data were saved. A common starting point landmark was assessed, in this case dacryon as defined by Buikstra and Ubelaker (15), and used throughout the analysis (19). The data were then saved as 200 Cartesian *X,Y* coordinates. Returning to tpsUtil, the coordinates were converted into landmarks.

Two geometric morphometric statistical programs were used in the analysis, R (20) and Past (21). The files were first taken into the R program for a Procrustes analysis. This brought all of the objects in the analysis to a standard size, orientation, and position, thus leaving only shape to be further analyzed in the study (22). From this point, the mean centered files were transported and formatted for use in Past, where comparative statistics on the samples were performed. Principal component (PC) analysis was utilized on the sample covariance matrix of eigenshape space to compare the modes of variation in morphology (23,24). The first 12 PCs captured >95% of the variation and were used in the multivariate analysis of variance (MANOVA). Finally, percentage variation attributable was calculated by dividing the model sum of squares by the residual sum of squares and then multiplying by 100 (the model sum of squares representing what is explained by difference in sex or race and residual sum of squares representing what is not explained by sex or race). In addition, the data (in the landmark stage) were taken into tpsRelw (25) by group to visualize the consensus (mean shape) and the variation covered by the first four PCs.

Metric Replicability

All replicability analyses were completed prior to the full study so that the methodology could be adjusted as necessary. For metric replicability, three separate collections from Indiana University were evaluated for adult skulls with at least one complete orbit. The following was the resultant sample. The first collection was the Carolina Biological Supply (CBS) collection made up of 12 skulls largely from India (population 1). The second collection was the teaching collection composed of 13 skulls from a variety of locations across the globe including Africa, Europe, South America, etc. (population 2). The final collection was a North American archaeological collection from the Greenshield site in North Dakota (population 3). This is a 40-acre Arikara site from the Roadmaker phase dating from AD 1785 to 1830 (26). From this collection, 13 skulls were chosen based on adult status, complete orbit, and presence of sex estimation.

Measurements were then taken using the methodology detailed above by three different measurers (both authors and a colleague) using the same vernier scale calipers in the same location and in all but one case on the same day in order to minimize as much error as possible. Each measurer produced a complete set of height and breadth measurements for the Arikara collection (population 3) and two sets for each of the CBS (population 1) and teaching collections (population 2).

From these measurements, SPSS 14.0 was then used to produce paired *t*-tests for both intra- and intero-bserver errors. Technical error of measurement (TEM) was calculated by hand for each variable on each level in order to compare with the Utermohle and Zegura (27) study, which had previously set forth standards for the orbital height and breadth measurements. TEM is traditionally used in anthropometry as an accuracy index that examines the SD of repeated measures (28). Finally, the coefficient of relative variability (CRV) was also calculated by hand to give an expression of the error as a percentage of the measurement itself. This is possible because the CRV is a measure of the SD relative to the mean of the measurements (29).

As we can see from Table 1, only one *t*-test was significant, indicating that these measurements are largely consistent within

TABLE 1—*Measurement intra-observer error.*

	Height					Breadth				
	Paired *t*-Test					Paired *t*-Test				
	R	*T*	Sig	TEM	CRV (%)	*R*	*t*	Sig	TEM	CRV (%)
Observer 1	0.669	−0.331	0.744	1.58	8.77	0.379	2.316	0.030*	1.72	5.69
Observer 2	0.978	−0.16	0.874	0.35	7.05	0.917	−0.378	0.709	0.49	4.74
Observer 3	0.626	0.074	0.942	1.54	7.90	0.458	−0.598	0.556	1.45	5.35

*$p < 0.05$.
CRV, coefficient of relative variability; TEM, technical error of measurement.

each observer. The TEM scores, on the other hand, vary between the observers. These were compared with 0.50 for height and 0.55 for breadth in the study by Utermohle and Zegura (27). Obviously, the majority of the scores reported here are noticeably higher than their scores. Finally, the CRV scores are all above 5% with one exception, so one observer (P.R.H.) was able to produce orbital breadth measurements consistently.

Table 2 with the intero-bserver error scores shows us quite a different picture. The majority of the inter-observer *t*-tests are statistically significant, frequently at the 0.01 level, indicating that often the measurements taken by the two observers were inconsistent. In addition, all of the TEM scores continue to be much higher than those reported by Utermohle and Zegura (27). The CRV scores are now all above 5%, although the breadth measurements are still lower than the height measurements.

From these outcomes, we can draw several conclusions. First, among the intra-observer scores, the results leave a bit to be desired with a number of the TEM scores double or even triple those found in the comparison literature, thus showing just how difficult to replicate these measurements may be (Table 1). The intra-observer numbers demonstrate that while it is possible to make measurements that are consistent, it is also not uncommon to have a significant difference within a single observer's measurements, particularly because these measurements are quite small and the bony markings are not incredibly clear.

The intero-bserver results show an even greater lack of replicability than the intra-observer analysis (Table 2). They are observers who are measuring the same skulls, under the same conditions, on the same day, using the exact same tool, and making sure that the same landmarks are being used by each measurer. The fact that nearly all of the TEM scores are at least double and the CRV scores are all above those found in the comparison literature (plus two significant *t*-tests) deems these measurements, or perhaps this method of taking them, to be largely nonreplicable.

In addition, these replicability tests were run on the individual measurements. These measurements are often put into equations or ratios to assess different aspects of the person being studied. In this case, they are put into a ratio in order to evaluate the orbits. To put multiple measurements with this much error into even a ratio (let alone an equation) is to simply risk-multiplying the amount of error involved. Owing to this information, all further analyses were conducted with measurements taken by one researcher (P.R.H.) to circumvent any unnecessary inter-observer error.

Morphological Replicability

To assess the standardization of the geometric morphometric technique, two different replicability tests were performed. First, the ability of two different observers to place consistent outlines with the Gimp 2.0 program was considered. The methodology for this investigation was as follows: First, pictures were taken of three specimens: two men and one woman, two acquired with postcranial material and one without, all from the CBS collection (the most homogenous collection). These specimens were specifically chosen to demonstrate a normal amount of variation within an average human population.

For each of these pictures ($n = 3$), each observer ($n = 2$) drew an outline on the image 10 separate times resulting in 60 outlines. These were then put through the same process detailed above and sorted into six groups based on specimen and observer. ANOVA tests were run for each observer individually (intra-observer: Figs 1 and 2), and then a MANOVA comparison of the observations of both researchers (inter-observer: Fig. 3) was made.

Both intra-observer results produced a significant score (Table 3). This demonstrates a greater variation between the specimens than

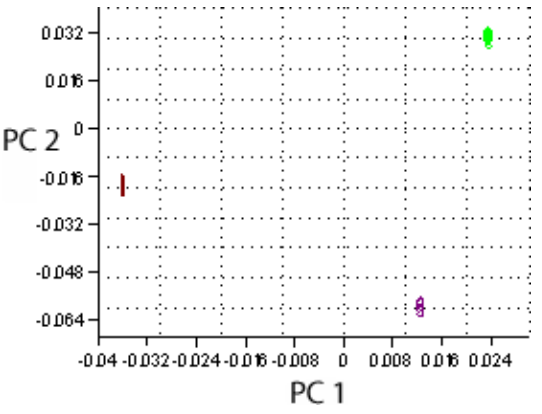

FIG. 1—*Observer 1 morphological intra-observer error.*

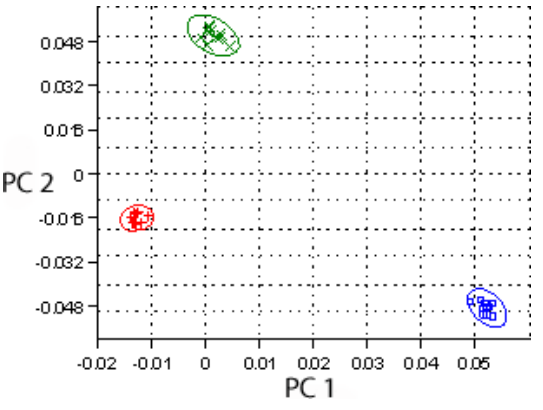

FIG. 2—*Observer 2 morphological intra-observer error.*

TABLE 2—*Measurement intero-bserver error.*

	Height					Breadth				
	Paired *t*-Test					Paired *t*-Test				
	R	*T*	Sig	TEM	CRV (%)	*R*	*t*	Sig	TEM	CRV (%)
Observer 1–Observer 2	0.953	−10.367	0.000*	1.12	8.06	0.892	−1.915	0.063	0.77	6.05
Observer 1–Observer 3	0.636	1.463	0.152	1.61	8.20	0.545	−0.168	0.867	1.57	6.39
Observer 2–Observer 3	0.646	5.476	0.000*	2.02	8.25	0.458	0.683	0.499	1.67	6.18

*$p < 0.01$.
CRV, coefficient of relative variability; TEM, technical error of measurement.

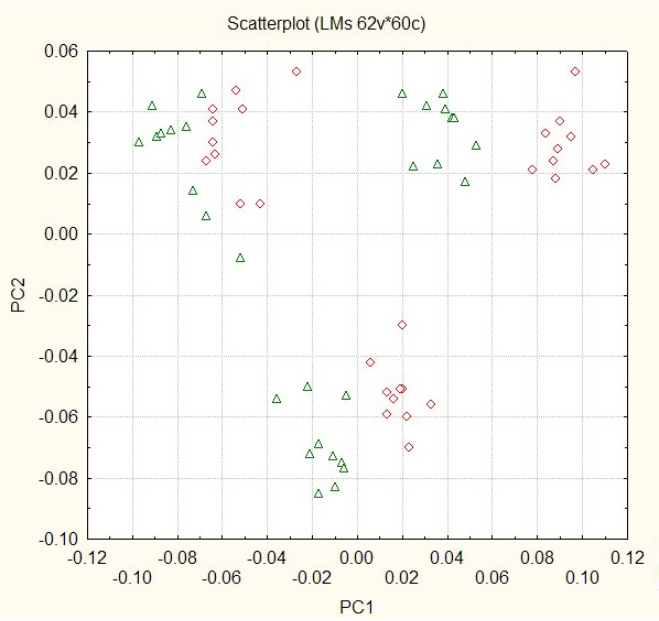

FIG. 3—*Morphological inter-observer error.*

TABLE 3—*MANOVA scores for morphological replicability tests.*

	Principal Components	*F*-value	*p*-Value (Sig.*)
Observer 1	30	11,530	0.00008671*
Observer 2	27	45.34	0.0218*
Inter-observer	35	15.06149	0.000000*

*$p < 0.05$.

there is within the specimens as a result of the outline differences. This shows that each observer is consistent in the manner with which he or she places the outlines. The inter-observer error also returned a significant score on the two-way MANOVA (Table 3). This tells the reader that there is a significant amount of difference between the two observers. For this reason, all outlines for the larger study were placed by one researcher (P.R.H.) so that this interobserver error would no longer be a factor.

Replicability of skull placement was assessed by way of MANOVA to ensure that differential placement of skulls between pictures did not produce significant error (30). Before each picture was taken, the skull was removed and then replaced. This resulted in 10 pictures each of the three different specimens.

Results from this assessment demonstrate that there is a significant difference when each specimen was compared to another. Between-specimen variation is greater than the within-specimen variation without overlap, which indicates that replicability regarding skull placement is not a problem for this methodology.

Results

Returning to the use of these metrics for sex and race distinction, Fig. 4 shows the average orbital index as well as the SD for men, women, black people, and white people. Women and black people tend to have slightly higher orbital indices than men and white people, respectively, although all SD bars overlap. Table 4 demonstrates the results of Student's *t*-tests on these same averages. Both tests indicate the groups were significantly different for orbital index.

Figure 5 shows the female and male consensus shapes. Preliminary visual comparisons suggest that the female shape has more

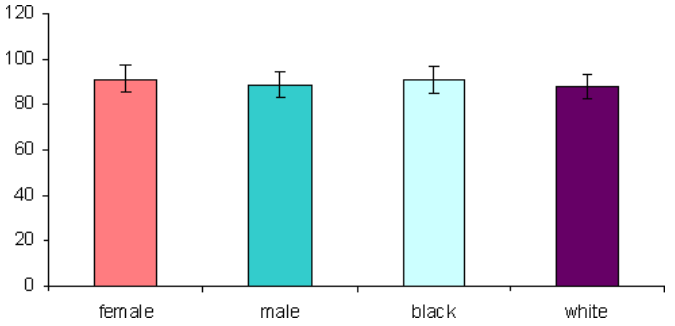

FIG. 4—*Orbital index averages for males, females, blacks, and whites.*

TABLE 4—*t-Test results for the orbital index.*

	t-Test	Effect Size
Female/male	1.6889E−10	0.486
Black/white	8.08835E−15	0.576

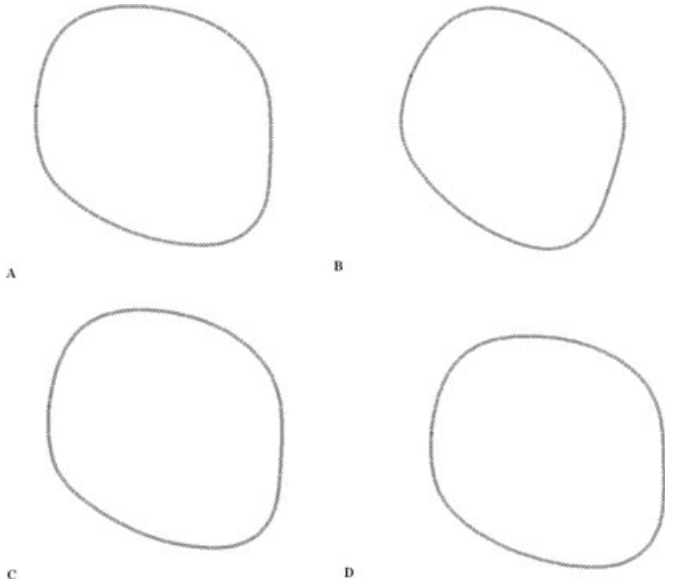

FIG. 5—*Average orbit shape (consensus) for (A) females (both groups), (B) males (both groups), (C) blacks (both sexes), and (D) whites (both sexes).*

TABLE 5—*MANOVA and percentage variation attributable results for principal components 1–12.*

	MANOVA Results	% Variation Attributable
Female/male	1.615E−12	1.07
Black/white	9.624E−13	1.36

inferior droop to the inferior lateral angle. Otherwise, the two shapes remain visually similar. Figure 5 also illustrates the black and white consensus shapes. Visually, these two shapes demonstrate more similarity than the previous comparison. Both demonstrate the inferior deflection at the lateral angle and appear in these images to have similar levels of rounded edges.

MANOVA statistics, using PCs one through 12, and the percentage variation explained are shown in Table 5. Both tests are again

significant. To further visually examine these results, 95% confidence interval graphs were created for PCs 1 and 2 (Figs 6 and 7; PCs 3 and 4 may be seen in Figs S1 and S2).

Discussion

The visual representations and the statistics shown here present two very different notions of these data. The graphs all illustrate that there is nearly complete overlap among these categories, yet the statistics demonstrate significant differences. We believe that this is attributable to the large sample size. Thus, while there is a difference between male and female orbital apertures and between white and black orbital apertures, how useful is this difference for the purposes of estimating sex and race?

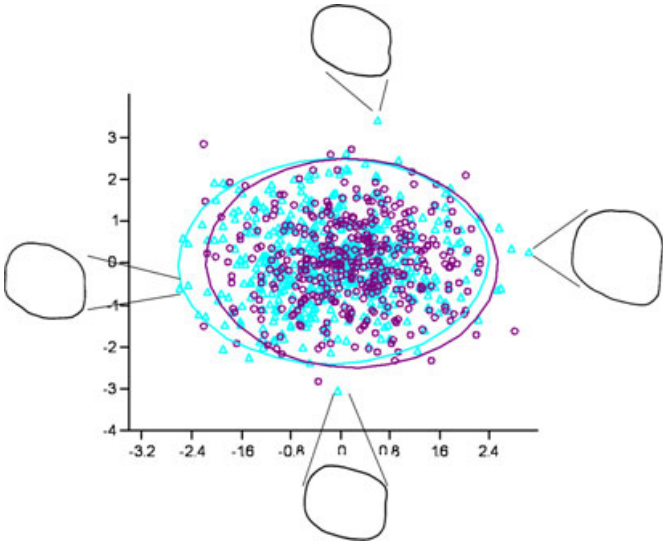

FIG. 6—XY *scatterplot of principal components 1 and 2 for females and males. Circles represent 95% confidence intervals. Ovals (purple)—females, triangles (turquoise)—males.*

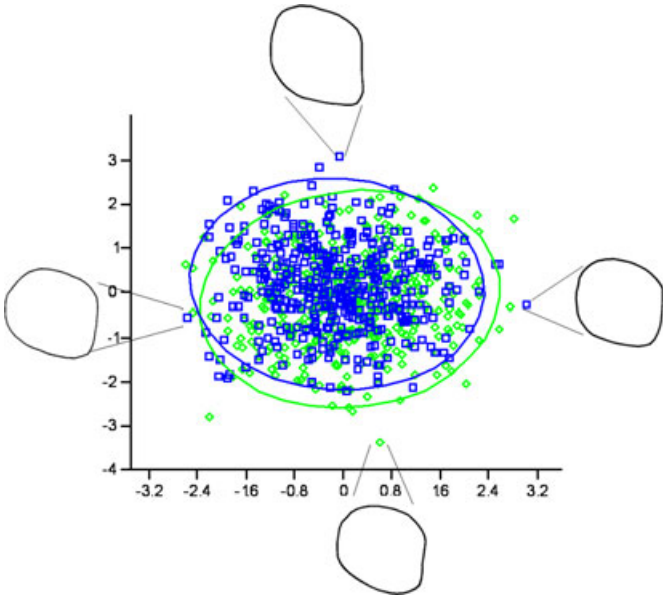

FIG. 7—XY *scatterplot of principal components 1 and 2 for blacks and whites. Circles represent 95% confidence intervals. Diamonds (green)—blacks, squares (blue)—whites.*

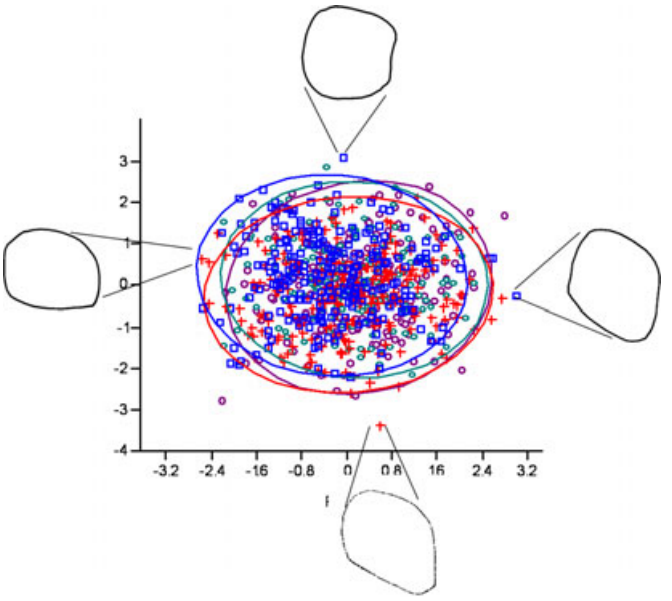

FIG. 8—XY *scatterplot of principal components 1 and 2 for sex and race. Circles represent 95% confidence intervals. Ovals (purple)—black females, Xs (green)—white females, +s (red)—black males, squares (blue)—white males.*

An analysis of the measurements reveals that this difference is not of practical significance. First, if we assume a TEM of *c.* 0.5 as Utermohle and Zegura (27) have suggested for our orbital measurements, which is much less than our analysis here actually found, this means that we allow more than 0.5 mm of error per measurement. This would result in an average orbital index range of more than 0.87–0.92, which would encompass the averages of both sexes and both races. Allowing a CRV of 5%, again less than we found, results in more than a millimeter of error and a range that includes all categories' averages and at least one SD. This suggests that considering the replicability of these measurements makes the statistical differences found here negligible.

The geomorphometric analysis also demonstrates a lack of practical applicability. In addition to the nearly complete overlap of the *XY* scatterplots, the percentage of variation attributable to each category is incredibly small (1.07% for sex, 1.36% for race). Neither of these concerns is assuaged when the variables are combined. Figure 8 illustrates the scatterplots and confidence intervals of PCs 1 and 2 for black women, white women, black men, and white men (PCs 3 and 4 available in Fig S3). Overlap is still virtually absolute. Finally, with both variables combined, the percentage variation attributable to both race and sex is 2.77%, still not nearly enough to maintain practical value in the real world.

Conclusions

The human orbital aperture does not contain practically valid information on race or sex. For orbital height or orbital breadth measurements, even replicability is a serious concern when a reasonable margin of error is more than the differences between the groups in both categories. When considering the overall shape, the orbital aperture demonstrates much more overlap than distinct areas. Thus, in accordance with the standards set forth by the

Daubert decision, the orbit should not be used as a method for assessing either race or sex unless no other options are present.

Acknowledgments

We begin by thanking our anonymous reviewers for their help in the preparation of the manuscript. We also thank the Cleveland Museum of Natural History and Lyman Jellema for access to its collection. Thanks to Della Cook and Kevin Hunt for their guidance. A special thanks to David Polly without whose statistical expertise this study would not have been possible. Finally, thanks to Allison Foley for her input in the origin of the concept, participation in the replicability analysis, and support throughout the process.

References

1. Byers SN. Introduction to forensic anthropology. Boston, MA: Pearson Education, Inc., 2008.
2. Moore-Jansen PM, Ousley SD, Jantz RL. Data collection procedures for forensic skeletal material. Report of investigations. Knoxville, TN: University of Tennessee, 1994.
3. Komar DA, Buikstra JE, editors. Forensic anthropology: contemporary theory and practice. New York, NY: Oxford University Press, 2008.
4. Topinard P, editor. L'anthropologie. Paris, France: C. Reinwald, 1876.
5. Olivier G. Practical anthropology. Springfield, IL: Charles C. Thomas Publisher, Ltd., 1969.
6. Piquet MM. L'indice orbitaire et l'appreciation de la largeur de l'orbite: essai de standardisation. Bull Mem Soc Anthropol Paris 1954;5:100–12.
7. Rogers SL. The human skull: its mechanics, measurements, and variations. Springfield, IL: Charles C. Thomas Publisher, Ltd., 1984.
8. Jantz RL, Ousley SD. FORDISC Ver. 3. Knoxville, TN: The University of Tennessee, 1993, 1996, 2005.
9. Krogman WM. The human skeleton in forensic medicine. I Postgrad Med 1955;17(2):A48; passim.
10. Krogman WM. The human skeleton in forensic medicine. Springfield, IL: Charles C. Thomas Publisher, Ltd., 1962.
11. Krogman WM, Iscan MYS. The human skeleton in forensic medicine, 2nd edn. Springfield, IL: Charles C. Thomas Publisher, Ltd., 1986.
12. Hauser G, DeStefano GF, editors. Epigenetic variants of the human skull. Stuttgart, Germany: E. Schweizerbart Science Publishers, 1989.
13. Szilvassy J. Eine neue methode zur intraserialen analyse von graberfeldern. Mitt Berliner Ges Anthrop Ethnol 1986;7:49–62.
14. Mensforth RP, Latimer BM. Hamann-Todd collection aging studies: osteoporosis fracture syndrome. Am J Phys Anthropol 1989;80:461–79.
15. Buikstra JE, Ubelaker DH, editors. Standards for data collection from human skeletal remains. Fayetteville, AK: Arkansas Archeological Survey, 1994.
16. Kimball S, Mattis P. Gimp 2.0. Ver. 2.2.13, 2004. http://www.gimp.org/ (accessed July 5, 2011).
17. Rohlf FJ. tps utility program. Stony Brook: Suny, 2006. http://life.bio.sunysb.edu/morph/soft-utility.html (accessed July 6, 2011).
18. Rohlf FJ. tpsDig Ver. 2.10. Stony Brook: SUNY, 2006. http://life.bio.sunysb.edu/morph/ (accessed July 6, 2011).
19. Lohmann GP. Eigenshape analysis of micro-fossils—a general morphometric procedure for describing changes in shape. J Int Assoc Math Geol 1983;15:659–72.
20. Gentleman R, Ihaka R. R. Ver. 2.5.1. Vienna, Austria: The R Foundation for Statistical Computing, 2007.
21. Hammer O, Harper D, Ryan PD. PAST: palaeontological statistics software package for education and data analysis. Palaeontologia Electronica 2001;4:9.
22. Hammer O, Harper D. Paleontological data analysis. Malden, MA: Blackwell Publishing, 2006.
23. Cootes TF, Cooper DH, Taylor CJ, Graham J. Trainable method of parametric shape-description. Image Vis Comput 1992;10:289–94.
24. Kent JT. The complex Bingham distribution and shape-analysis. J R Stat Soc Series B Stat Methodol 1994;56:285–99.
25. Rohlf FJ. Relative warps. Ver. 1.44. Stony Brook: SUNY, 2006. http://life.bio.sunysb.edu/morph/soft-tps.html (accessed July 6, 2011).
26. Hildebrant E. Pioneer archaeological excavation: the Greenshield site (32OL17). Mankato, MN: Minnesota Statue University, 2004.
27. Utermohle CJ, Zegura SL. Intra- and interobserver error in craniometry: a cautionary tale. Am J Phys Anthropol 1982;57:303–10.
28. Perini TA, de Oliveira GL, Ornellas JS, de Oliveira FP. Technical error of measurement in anthropometry. Rev Bras Med Esporte 2005;11:86–90.
29. Hendricks WA, Robey KW. The sampling distribution of the coefficient of variation. Ann Math Stat 1936;7:129–32.
30. Bailey RC, Byrnes J. A new, old method for assessing measurement error in both univariate and multivariate morphometric studies. Syst Zool 1990;39:124–30.

Additional information and reprint requests:
Polly Husmann, Ph.D.
West Virginia School of Osteopathic Medicine
400 N. Lee St.
Lewisburg, WV 24901
E-mail: phusmann@osteo.wvsom.edu

Supporting Information

Additional Supporting Information may be found in the online version of this article:

Figure S1. XY scatterplot of PCs 3 and 4 for females and males. Circles represent 95% confidence intervals. Ovals (purple)—females, triangles (turquoise)—males.

Figure S2. XY scatterplot of PCs 3 and 4 for blacks and whites. Circles represent 95% confidence intervals. Diamonds (green)—blacks, squares (blue)—whites.

Figure S3. XY scatterplot of PCs 3 and 4 for sex and race. Circles represent 95% confidence intervals. Ovals (purple)—black females, Xs (green)—white females, plus signs (red)—black males, squares (blue)—white males.

Please note: Wiley-Blackwell is not responsible for the content or functionality of any supporting materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.