J Forensic Sci, September 2007, Vol. 52, No. 5
doi: 10.1111/j.1556-4029.2007.00509.x
Available online at: www.blackwell-synergy.com

TECHNICAL NOTE

Gregory E. Berg, M.A.; Sabrina C. Ta'ala, M.A.; Elias J. Kontanis, Ph.D.; and Sardiaa S. Leney, B.S.

Measuring the Intercondylar Shelf Angle Using Radiographs: Intra- and Inter-Observer Error Tests of Reliability

ABSTRACT: This study presents a test of intra- and inter-observer measurement error rates of the intercondylar shelf angle of the distal femur, as measured on lateral radiographs. This measurement is the central element of a method to determine racial affiliation from the distal femur. Four observers independently radiographed a set of 33 complete and partial femora from collections housed at the Joint POW/MIA Accounting Command, Central Identification Laboratory. Each observer then measured the intercondylar shelf angles in each radiograph, following the original method's guidelines. A supplementary inter-observer error test was conducted by four additional observers on one set of radiographs. Statistically significant differences were found for both intra- and inter-observer error based on the results of Student's *t*-tests, paired samples *t*-tests, and ANOVA analyses. The results of this study indicate that further experimentation should be undertaken in order to develop refined measurement techniques that may help improve standardization and reduce the observer error rates.

KEYWORDS: forensic science, forensic anthropology, intercondylar shelf angle, reliability, racial determination, distal femur

With the refinement of legal standards of evidence that has occurred in recent years, particularly since the Daubert decision in 1993 (1), the forensic community has been increasingly concerned with the reliability of methods from which scientific conclusions are drawn. Reliability is how frequently a method yields the same results between successive trials, or its *repeatability*. Without accurate repeatability, a method is considered unreliable. In light of this rising awareness about the importance of method reliability, this study was designed to examine the intra- and inter-observer error associated with a method used for determining race based on radiographs of the distal femur.

Few reliable methods exist for the estimation of race based on postcranial elements. Assessing the racial ancestry of skeletonized remains in the absence of sufficient craniofacial elements is a persistent challenge for forensic anthropologists. Some of the most commonly used methods are based on the observation that femoral morphology tends to vary among racial groups (2–7). One method in particular employs a measurement of the intercondylar shelf angle using lateral radiographs of the distal femur (8). This technique may be particularly valuable in cases where the entire femur is not present, as it can be applied to partial femora.

Craig's (8) method is based on the assertion that the morphology of the intercondylar notch on the distal femur varies between American Blacks and Whites. In true lateral radiographs of the distal femur, the roof of the intercondylar notch is visible as a linear opacity, a feature commonly known as Blumensaat's line. Craig's (8) method comprises a simple system of measurement whereby a straight line is drawn through Blumensaat's line, and another straight line is marked "parallel to the posterior cortex of the bone" along the "distal one third of the femur." The angle between these two lines is then measured using a goniometer or protractor.

Joint POW/MIA Accounting Command—Central Identification Laboratory (JPAC-CIL), Hickam AFB, HI 96853.

Received 1 Sept. 2006; and in revised form 11 Nov. 2006, 8 Mar. 2007; accepted 31 Mar. 2007; published 21 July 2007.

Craig (8) applied this method to a sample of radiographs from 423 White and Black patients from two medical centers in Georgia and Tennessee. The sample included radiographs of either right or left knees from both males and females. *t*-Tests were used to determine the means, standard deviations, and *p*-values for the two racial categories (as well as sex) and a discriminant analysis was used to verify whether the angle measurements could be used to classify race. A test study was then conducted on a skeletal sample ($n = 67$) taken from the William Bass Donated Skeletal collection. Egg crate foam was used to position the skeletonized femora in a true lateral position for radiography. Using an established sectioning point of 141 degrees (with Whites falling above and Blacks falling below), the method was able to correctly classify the race of 57 of the 67 individuals (85%). Craig (8) also conducted an inter-observer error study with three participants, who independently measured 23 radiographs. Reported average variation among the observers was less than 1 degree using the method.

Materials and Methods

This research was designed to test the reliability of the primary measurement employed by the Craig method (8). The intra- and inter-observer error study was conducted by four participants using dry human bone from collections at the Joint POW/MIA Accounting Command, Central Identification Laboratory. Three of the study participants use the method on a regular basis; the fourth had not used the method before. Because the study was intended to test observer error only, and not the actual effectiveness of the method at correctly discerning racial ancestry, neither the race nor the sex of the femora was factored into the analysis. All femora were from adult males. Each observer independently radiographed 33 whole or partial femora using a HOLOGIC RADEX Digital X-ray System (Hologic, Inc., Bedford, MA). The radiographic protocols followed the Craig (8) method (exclusive of radiographic equipment and settings). Egg crate foam was used to support and position each sample. Observers also referred to exemplars of true lateral

positioning from a radiographic textbook (9). Brightness and contrast were adjusted on the digital radiographs as necessary in order to maximize the visibility of Blumensaat's line. In some cases, only one femur was radiographed individually; in other cases two or three femora were radiographed at one time. Radiographs were printed on standard 8 1/2-x-11 inch paper. Each observer printed six copies of the radiographs, giving a copy to each of the other three observers.

Following the method's prescribed instructions, a ruler was used to draw a line through Blumensaat's line, and another line parallel to the posterior cortex of the distal third of the femoral shaft. The resulting angle was then measured using a light table and a protractor. Throughout the radiographic and measuring procedures, consultation between observers was prohibited. Each observer measured the intercondylar shelf angle on their own radiographs three times, with at least 24-h between measuring sessions in order to minimize short-term memory bias. Each observer measured the other observers' set of radiographs once. Thus, nearly 800 observations were conducted on the four independent sets of radiographs. In order to supplement the inter-observer error portion of the study, four additional observers, all board-certified forensic anthropologists, completed measurements on observer #4's radiographic set. Each of these additional observers reviewed the method's instructions prior to analyzing the printed radiographs.

At the completion of data collection, all measurements were collected and entered into a computer database. Intra- and inter-observer error analyses then were conducted through robust statistical analyses including Student's *t*-tests, paired-samples *t*-tests, ANOVA analysis, and summary statistics. All analyses were computed using SPSS® version 10.1 software (SPSS Inc., Chicago, IL). As the reported accuracy using the Craig method (8) had an accuracy rate of 85%, minor data variations from the current study would be magnified if the data were applied in the original study's parameters. This observation indicates that a more conservative approach to the data analysis was warranted. Therefore, the statistical evaluations used a conservative significance level of 0.10, but as will be shown, the results were often highly significant, well below a 0.05 level.

Results

Intra-observer Error

Each observer's recorded measurements were analyzed for intra-observer error through the divergence of the successively recorded measurements. The differences between each trial were calculated and recorded, totaling 99 observations per observer. The differences were calculated in terms of absolute differences between observations rather than by direction (greater than or less than one set of measurements). In this study, directionality is defined as the signed variation between measurements if one specific trial is held constant (e.g., trial 1 for comparisons between trials 1, 2, and 3, and trial 2 for comparisons between trials 2 and 3). The mean differences between each trial, per observer, are listed in Table 1. The absolute variations between measurements ranged from 0 to 15 degrees. If directionality is taken into account, the range of variation was from −15 degrees to +11 degrees. A visual presentation for all participants' intra-observer error, with directionality, is shown in Fig. 1. In either case, one-third of the measurements (132 out of 396, 33%) were greater than 3 degrees different in successive trials.

In order to assess the level of intra-observer variation, Student's *t*-tests were performed on the data. Absolute values were used in all comparisons, rather than values with signed directionality, as the latter would sum toward zero and not adequately reflect the variation present in the data. Several levels of specificity were accepted; as some minor variation is expected in the method (8), we tested for differences greater than 0, 1, and 2 degrees. In all but one instance (Observer 1, 2 degrees) highly significant differences were found between the trials for the study participants (Table 2).

TABLE 1—*Mean differences between successive trials per observer, and the overall mean differences for the complete intra-observer error study.*

Observer	*n*	Trial	Mean difference
1	33	1 and 2	3.9
1	33	1 and 3	3.6
1	33	2 and 3	3.7
1	99	Overall	3.7
2	33	1 and 2	2.6
2	33	1 and 3	2.2
2	33	2 and 3	2.0
2	99	Overall	2.3
3	33	1 and 2	2.5
3	33	1 and 3	3.2
3	33	2 and 3	3.0
3	99	Overall	3.0
4	33	1 and 2	3.0
4	33	1 and 3	3.3
4	33	2 and 3	3.4
4	99	Overall	3.4
Composite	396	All	3.1

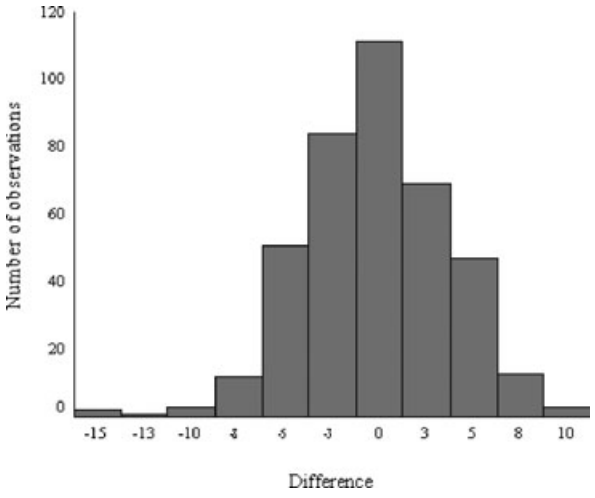

FIG. 1—*Directional differences between trials for all observers.*

TABLE 2—*Student's t-test results for intra-observer error study.*

Observer	Test value (in degrees)	*t*-value	DF	Significance (2-tailed)
1	0	11.44	98	0.000
1	1	6.43	98	0.000
1	2	1.42	98	0.158
2	0	13.60	98	0.000
2	1	10.00	98	0.000
2	2	6.32	98	0.000
3	0	13.38	98	0.000
3	1	8.94	98	0.000
3	2	4.49	98	0.000
4	0	12.95	98	0.000
4	1	8.99	98	0.000
4	2	5.04	98	0.000

Inter-observer Error

All observers measured a radiographic set produced by each study participant for a total of four sets of data (132 observations per set and 528 total observations). Inter-observer error was analyzed in several different ways, all of which used the radiographic sets as a common variable between all observers. Each observer's second trial measurements (from the intra-observer error portion of the study) were used for the observer whose radiographic set was being analyzed. In order to accurately measure the divergence (or lack thereof) between observers, a standardized measurement, or target value, was created. The target value was the mean value of all four observers' measurements over their successive three trials (again, from the intra-observer error portion). The mean value was accepted as an approximation of biological reality as it is not possible to know the actual measurement of the intercondylar shelf angle for any one given femur in nonfleshed remains. All calculated differences between the target measurement and the observers were absolute values rather than signed, directional differences. Using the generated data, the study identified three primary questions regarding inter-observer error:

(1) Is the mean measurement for each observer different from the target mean?
(2) Is the difference between the observer and the target measurement per case different from zero, which would be expected if there was little to no inter-observer error? If there are significant differences from zero, at what point do most differences between the observers and the target value cease?
(3) Are the differences of each observer and the target measurement different between the observers?

Regarding the first question, ANOVA analyses were conducted to determine if there were significant differences in the overall mean between the observers for all femora, as well as between all observers and the mean target measurement. The comparisons between the observers and the target measurement revealed statistically significant differences in three of four cases (Table 3). Only one set of radiographs produced similar results between all observers and the target measurements. Bonferroni *post hoc* tests suggested that the significant differences were not only between the observers and the target response but also between the investigators themselves. When the target response was removed from the analyses, the results were equivocal, with the same three analyses indicating significantly differing means between observers.

If little to no inter-observer error was present within the method, then the differences between each observer and the target value should be close to zero. Therefore, *t*-tests with a test value of zero were conducted for each of the four radiographic sets by observer. The presentation of the observers' distribution on one radiographic set is illustrative of the overall pattern and is presented in Table 4. In each test, highly significant differences were present between the observers and the target value ($p = 0.000$). Significant differences occurred in all radiographic sets until a test value of two was achieved. At this point, 11 of 16 comparisons were still significant; a maximum of two observers per set were not significantly different from the target; one complete radiographic set still retained significant differences for all four observers.

The third question focuses on the differences between the investigators themselves, per case, standardized to the target value. These comparisons were conducted through paired-samples *t*-tests, for a total of six pairs per radiographic set (Table 5). Close examination of the results shows that significant differences between the investigators are present in all but one of the radiographic sets. Observer #2's recorded values were most frequently different from the other observers. Radiographic set #1 fared the best, whereas radiographic set #3 was the worst, in terms of the differences between the observers.

A final series of observations were undertaken by four additional board certified anthropologists. The variation for these observers, per case, ranged from 1 to 20 degrees, with an average range of 6.8 degrees. The original four observers, for the same set of radiographs, ranged from 1 to 17 degrees, with an average of 6.9 degrees. No obvious patterning was apparent in the data; the largest and smallest ranges were not consistent in the cases between the observation groups. Overall, the additional observers'

TABLE 3—*ANOVA results for each observer's radiographic series and the differences between all observers and the target response.*

Observer radiographs	Sum of squares	DF	Mean square	*F*	Significance
Obs1-between groups	76.15	4	19.04	0.857	0.491
Obs1-within groups	3555.03	160	22.22		
Obs1-total	3631.18	164			
Obs2-between groups	299.25	4	74.81	4.83	0.001
Obs2-within groups	2476.72	160	15.48		
Obs2-total	2775.98	164			
Obs3-between groups	208.52	4	52.13	2.81	0.027
Obs3-within groups	2968.55	160	18.55		
Obs3-total	3177.07	164			
Obs4-between groups	180.79	4	45.20	2.28	0.063
Obs4-within groups	3171.94	160	19.83		
Obs4-total	3352.73	164			

TABLE 4—*The absolute differences (in degrees) between the target measurement and each observer for radiographic set #2.*

Case	Obs1	Obs2	Obs3	Obs4
1	2	5	3	3
2	2	2	1	2
3	7	6	11	3
4	1	3	5	5
5	5	4	2	3
6	1	0	3	4
7	11	2	2	0
8	2	1	4	1
9	6	4	5	5
10	6	1	7	3
11	2	4	0	3
12	6	3	6	2
13	2	1	1	3
14	7	2	3	1
15	7	10	3	3
16	9	2	2	2
17	3	4	1	6
18	3	2	3	4
19	4	2	2	3
20	1	1	0	1
21	1	4	1	1
22	1	1	1	5
23	0	4	0	2
24	3	2	1	8
25	13	2	6	3
26	4	2	1	5
27	2	4	5	1
28	2	1	1	2
29	2	4	2	4
30	5	1	7	0
31	0	4	7	1
32	3	5	6	3
33	2	6	4	2
Mean difference	3.8	3.0	3.2	2.9

TABLE 5—*Pair samples t-tests between observers for each radiographic set.*

Radiographic set/pair group	Mean	SD	*t*	DF	Significance (2-tailed)
#1, obs1-obs2	0.94	3.77	1.43	32	0.162
#1, obs1-obs3	0.15	2.64	0.33	32	0.743
#1, obs1-obs4	0.36	3.38	0.62	32	0.541
#1, obs2-obs3	0.79	3.63	1.25	32	0.221
#1, obs2-obs4	0.58	3.33	0.99	32	0.329
#1, obs3-obs4	−0.21	2.93	−0.42	32	0.681
#2, obs1-obs2	−0.09	2.90	−0.18	32	0.859
#2, obs1-obs3	−0.42	3.35	−0.73	32	0.473
#2, obs1-obs4	−0.91	2.59	−2.01	32	0.052
#2, obs2-obs3	0.33	3.17	0.60	32	0.550
#2, obs2-obs4	0.82	2.65	1.77	32	0.086
#2, obs3-obs4	−0.48	3.15	−0.88	32	0.384
#3, obs1-obs2	−0.79	3.29	−1.38	32	0.178
#3, obs1-obs3	1.61	2.98	−1.84	32	0.075
#3, obs1-obs4	−0.06	3.03	2.03	32	0.050
#3, obs2-obs3	0.82	2.31	−0.12	32	0.909
#3, obs2-obs4	−0.85	2.65	3.10	32	0.004
#3, obs3-obs4	1.67	2.45	3.92	32	0.000
#4, obs1-obs2	−0.88	3.33	−1.52	32	0.140
#4, obs1-obs3	0.52	3.17	0.93	32	0.358
#4, obs1-obs4	0.48	2.46	1.13	32	0.267
#4, obs2-obs3	1.39	3.12	2.57	32	0.015
#4, obs2-obs4	1.36	3.10	2.52	32	0.017
#4, obs3-obs4	−0.03	3.09	−0.06	32	0.955

absolute mean measurements were 3.0, 3.4, 3.6, and 4.0 degrees different from the target measurements. The results are in line with the original four observers' results and further confirm this study's findings.

Discussion and Conclusion

Significant amounts of error were encountered when using the measurement guidelines employed by the Craig method (8). Three principal areas of potential methodological error that can be attributed to observer error were identified in the course of this study. First, the femur positioning in the radiographs tended to be slightly different between observers. A "true lateral" position is somewhat difficult to achieve consistently with dry bone specimens. A V-shaped radiopacity, formed by the distal-most portions of the medial and lateral supracondylar ridges, was found to be present in radiographs where the femur was misaligned (Fig. 2). Throughout the study, each participant consciously strove to minimize this error; nevertheless, clear misalignment was apparent in some radiographs.

Two sources of error stem from a certain degree of subjectivity when judging "best-fit" positioning of lines through Blumensaat's line and a line parallel to the posterior cortex of the distal one-third of the femur. The placement of a line through Blumensaat's line tended to be relatively straightforward in most cases, though some variation was present between investigators. The most conspicuous source of intra- and inter-observer error was variation in the placement of the line parallel to the posterior cortex of the distal femur. Determination of a parallel position was extremely difficult, and in nearly every case, varied between trials per observer, and between observers on a single radiograph. This problem was particularly pronounced in instances of highly curvilinear femora. As the reference points are somewhat enigmatic, this line fluctuates considerably between observations.

A final potential source of error was considered in this study—the familiarity of a user with the method. One study participant had never used the method before. That observer had the second least mean amount of intra-observer error recorded, and in general, was usually mid-range in each of the other analyses. Further, two experienced investigators had greater ranges of variation in the intra-observer study. This suggests that the previously identified sources of error had a far greater effect on the measurements than an investigator's experience level.

FIG. 2—*An example of improper positioning of the femur, even though Blumensaat's line is clearly visible. Arrows point to the divergent V-shaped radiopacities formed by the medial and lateral supracondylar ridges.*

Osteological methods to estimate racial ancestry using postcranial skeletal remains are invaluable tools for forensic anthropologists. The Craig (8) method is potentially particularly useful because it utilizes an element that is often more well preserved than others and can frequently be applied even if the femur is partially complete. However, the results of this study indicate that continued reliance on this method requires further experimentation and testing to develop refined measurement techniques that will improve standardization and reduce observer error.

Acknowledgments

The authors thank Drs. Mark Leney and Laura Miller for insightful commentary and critiques. We also extend a grateful thank you to Drs. P. Willey, Ted Rathbun, Michael Finnegan, and Hugh Berryman for participating in the study. Two anonymous reviewers' valuable comments helped us to improve this paper.

References

1. *Daubert V. Merrell Dow Pharmaceuticals, Inc.*, 509 U.S. 579, (1993).
2. Baker SJ, Gill GW, Kieffer DA. Race and sex determination from the intercondylar notch of the distal femur. In: Gill GW, Rhine S, editors. Skeletal attribution of race. Albuquerque, NM: Maxwell Museum of Anthropology, 1990;91–5.
3. Gilbert BM. Anterior femoral curvature: its probable basis and utility as a criterion of racial assessment. Am J Phys Anthropol 1976;45:601–4.
4. Gilbert R, Gill GW. A metric technique for identifying American Indian femora. In: Gill GW, Rhine S, editors. Skeletal attribution of race. Albuquerque, NM: Maxwell Museum of Anthropology, 1990;97–9.
5. Stewart TD. Anterior femoral curvature: its utility for race identification. Hum Biol 1962;34:49–62.
6. Westcott DJ. Population variation in femur subtrochanteric shape. J Forensic Sci 2005;50:286–93.
7. DiBennardo R, Taylor JV. Classification and misclassification in sexing the Black femur by discriminant function analysis. Am J Phys Anthropol 1982;58:145–51.
8. Craig EA. Intercondylar shelf angle: a new method to determine race from the distal femur. J Forensic Sci 1995;40(5):777–82.
9. Yochum TR, Rowe LJ. Essentials of skeletal radiology. 2nd ed. Baltimore, MD: Williams & Wilkins, 1996.

Additional information and reprint requests:
Gregory E. Berg, M.A.
Forensic Anthropologist
Joint POW/MIA Accounting Command—Central Identification Lab
310 Worchester Avenue
Hickam AFB, HI 96853-5530
E-mail: greg.berg@ds.jpac.pacom.mil