Reliability and Validity of the Walker (2008) and Klales *et al.* (2012) Methods

Mackenzie Walls & Alexandra R. Klales, PhD, Forensic Anthropology, Washburn University; Kate M. Lesciotto, JD, MS, Anthropology, Pennsylvania State University; Timothy P. Gocha, PhD, Anthropology, University of Nevada, Las Vegas; Heather M. Garvin, PhD, Anatomy, Des Moines University

Introduction

- Many biological profile methods are sex specific (i.e., ancestry, stature, and age) and depend on a correct estimation of sex
- Walker (2008) and Klales et al. (2012) are popular methods for morphological sex estimation

Research Aims: Given that previous studies used varied statistical analyses and report a range of results, this study set out to test the reliability and validity of the Walker (2008) and Klales et al. (2012) sex estimation methods using a large set of data collected from diverse populations and from researchers with varied experience levels and backgrounds. The specific aims of the present paper are to evaluate intra- and interobserver error for the Walker (2008) and Klales et al. (2012) trait scoring methods, assess the role of experience on method validity, and to compare the results of this study to previously published work.

Materials and Methods

- ***Collections:***
  - Hamann-Todd (HT)
  - William M. Bass Donated (UT)
  - Texas State Operation Identification (OI)
  - Texas State Donated Collection (TS)
- ***Traits:***
  - 5 Walker (2008) traits as found in Buikstra & Ubelaker (1994)
  - 3 Phenice (1969) traits as described in Klales et al. (2012)
- ***Total sample size:*** n=346 (skull) and n=442 (pubis) – varied by analysis
- ***Scoring:***
  - 7 observers (Observer A-G) with varying levels of experience (Table 1)
  - Scored traits on an ordinal scale (1-5) using the descriptions/illustrations from each method (Figures 1-2)
- ***Reliability***
  - Intraobserver: linear weighted *Kappa* (w*K*) based on parameters of agreement by Landis and Koch (1977)
  - Interobserver: intra-class correlation coefficient (ICC), a two-way random average model with a 95% absolute tolerance based on parameters of agreement by Cicchetti (1994)

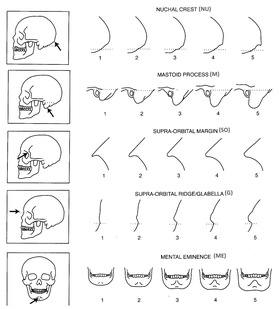

Fig 1. Walker (2008) traits from Buikstra & Ubelaker (1994).

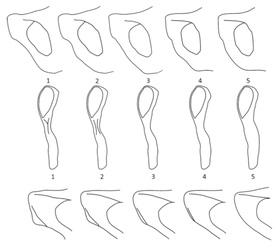

Fig 2. Klales et al. (2012) traits. Top: sub-pubic contour (SPC), Middle: medial aspect of ischio-pubic ramus (MA), Bottom: ventral arc (VA).

- ***Validity:***
  - Classification accuracy calculated using the logistic regression equations provided by the original methods

Table 1. Observer experience level.

	Observer[1]	Experience [2]	Highest Education	Trait/Method Experience[3]	Skeletal Variation Experience[4]
Collected data specifically for NIJ grant (n=222 HTH & Bass Collections)	Observer A (S & P)	Expert	PhD, Practicing Forensic Anthropologist	Scored the traits in >500 individuals and developed the Klales et al. method	Experience working with >500 skeletons from U.S. forensic cases and collections across U.S., Europe, and Egypt.
	Observer B (S & P)	Experienced	MS student, Forensic Anthropology	Scored the traits in approximately 185 individuals.	Experience working with ~200 skeletons from U.S. forensic cases and collections in the U.S.
	Observer C (S & P)	Inexperienced	BS student, Biology	Never scored the traits or used the methods.	N/A
	Observer D (S & P)	Expert	PhD, Practicing Forensic Anthropologist	Scored the traits in approximately 150 individuals	Experience working with >500 skeletons from U.S. forensic cases including >200 presumed Hispanic migrants, skeletal collections in the U.K. and Thailand
Collected skull and pelvic data during previous research/analyses and contributed it to the project	Observer E (S)	Expert	PhD, Practicing Forensic Anthropologist, D-ABFA	Scored cranial traits in approximately 200 individuals.	Experience working with > 500 skeletons from U.S. forensic cases and collections across U.S. and Iberian peninsula.
	Observer F (P)	Experienced	Doctoral Anthropology student	Scored the traits in approximately 50 individuals.	Experience working with approximately 100 skeletons from U.S. forensic cases and skeletal collections.
	Observer G (P)	Inexperienced	Doctoral Anthropology student	Never scored the traits or used the method.	N/A

[1] "S" and "P" refer to whether the Observer contributed skull (S) or pubic (P) trait scores to this study. [2]Experience level was determined based on combined education, trait/method experience, and overall osteological experience and exposure to human variation. [3] Trait/Method Experience refers to the number of individuals formally scored using the traits and/or following the Walker (2008) or Klales et al. (2012) instructions prior to collection of the data included in this study (i.e., it does not include the present study data and does not include any use of a "gestalt" sex method). [4]Variation Experience refers to the number and diversity of skeletons analyzed by the observer prior to collection of data included in this study.

Results

- ***Intraobserver agreement n=222 (Table 2):***
  - Skull: substantial agreement
  - Pelvis: almost perfect agreement
- ***Skull interobserver agreement (Figure 3)***:
  - Ranged from good to excellent for the three experience levels (Obs A-C), except for the SO in pairwise comparisons with the inexperienced observer (red arrows)
  - Pairwise comparisons in Obs A-C were always higher between the expert (A) and experienced (B) observers (blue box) than when either was paired with the inexperienced observer (C)
  - Experts (Obs A,D,E) had excellent agreement for all traits except the ME (green arrows) despite varied educational backgrounds, training, and application of the traits and methods in skeletal collections/casework

Table 2. Intraobserver agreement using linear wK for Observer A.

Trait	w*K*	Agreement Level
VA	0.87	Almost Perfect
SPC	0.89	Almost Perfect
MA	0.84	Almost Perfect
NU	0.72	Substantial
G	0.74	Substantial
SO	0.67	Substantial
M	0.66	Substantial
ME	0.72	Substantial

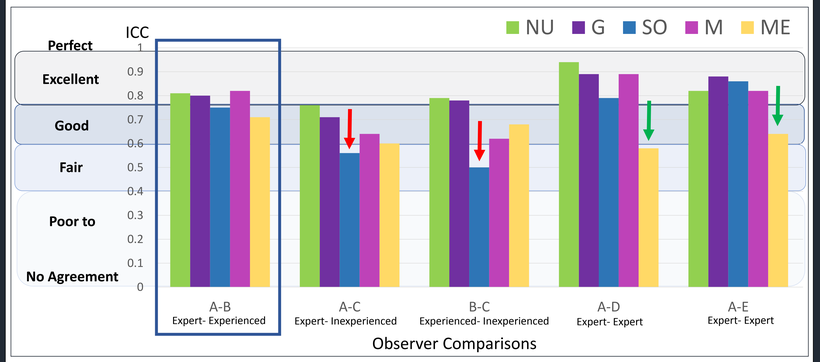

Figure 3. Interobserver agreement between observers of varying experience levels for the Walker (2008) method.

- ***Pelvis interobserver agreement (Figure 4)***:
  - Ranged from good to excellent for three experience levels
  - When expert (A) and experienced observers (B,F) were compared, the observer with experience > n=100 (B) aligned more with the expert (blue box) than experienced observer with < n=100 (F) (green box)
  - Experts (Obs A,D) had excellent agreement for all three traits

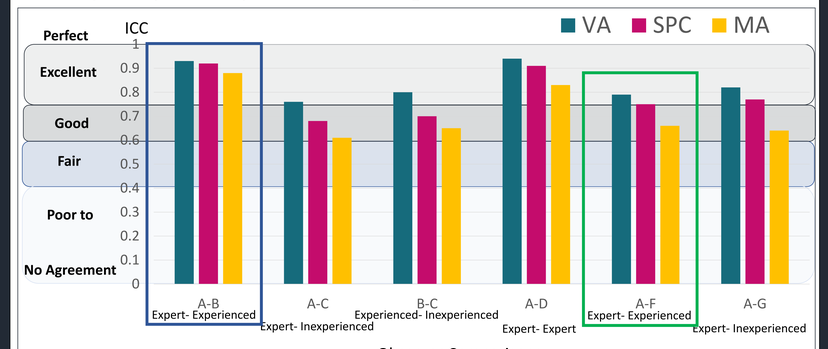

Figure 4. Interobserver agreement between observers of varying experience levels for the Klales et al. (2012) method.

- ***Validity:***
  - Skull: 61.4 to 90.4%
    - Overall lower accuracy rates than reported by Walker (2008)
    - Experts achieved higher accuracy
  - Pelvis: 78.2 to 96.6%
    - Accuracy rates consistent with Klales et al. (2012)- higher in more experienced observers
    - Experts achieved higher accuracy, less experienced had accuracy rates around ~70%
    - Experienced observer with > n=100 also had high accuracy

Collection →	OI & TS		HT & UT		HT
Observer	Skull	Pelvis	Skull	Pelvis	Pelvis
A- expert	84.6	88.5	73.5	96.6	94.5
B- experienced	-	-	61.4	93.6	
C- inexperienced	-	-	70.7	78.2	
D- expert	90.4	86.5	-		-
E- expert	-	-	83.3		-
F- experienced	-	-	-		67.7
G- inexperienced	-	-	-		68.3

Table 3. Classification accuracy (%) by observer using the equations provided by Walker (2008) and Klales et al. (2012).

Discussion & Conclusions

- Traits can be reliably scored by multiple observers with varied backgrounds and experience levels, except for the ME – consistent with previous literature
- Reliability higher for more experienced observers → practitioners should have familiarity and practical experience with methods, traits, and range of skeletal variation in n>100 prior to application
- Experience and greater training increased the validity of the method
- Classification accuracy was higher for the pelvis than the skull
- Exposure to human variation impacted experience level

Acknowledgements

Thanks go to Lyman Jellema for access to the Hamann-Todd Collection, Dr. Dawnie Steadman for access to the William M. Bass Donated Collection, and to Drs. Kate Spradley and Danny Wescott for access to the Operation Identification and Texas State Donated collections. Thanks also go to Stephanie Cole, Alexis Winter, and Lily Doershuk for participating in data collection. This research was partially funded by National Institute of Justice Grant 2015- DN-BX-K014.

Opinions or points of view expressed in this research represent a consensus of the authors and do not necessarily represent the official position or policies of the U.S. Department of Justice or the National Institute of Justice. Any products and manufacturers discussed are presented for informational purposes only and do not constitute product approval or endorsement by the U.S. DOJ or NIJ.

For a full list of references or a copy of the poster, contact: mackenzie.walls@washburn.edu