Testing the Reliability of Morphological Sex Estimation Methods used in Forensic Anthropology

Mackenzie Walls and Alexandra Klales, Ph.D., Forensic Anthropology Program, Washburn University

Introduction

- Establishing a biological profile is an important component of a forensic anthropologist's job → helps narrow list of missing persons and with positive ID
- Sex estimation is especially important because other profile methods are sex-specific (i.e., ancestry, stature, age)
- Walker (2008) and Klales et al. (2012) are popular methods used by practitioners to estimate sex
- Testing reliability of methods is key for acceptance and compliance with the *Daubert* ruling and NAS report
- Present research tests reliability of the methods and examines the role of experience

Materials and Methods

Sample

- Data collected from 222 black and white males and females from the historic Hamann-Todd (HTH) and the modern Bass donated skeletal collections (UTK)(Table 1)

Table 1. Sample size for each collection.

Method	Females	Males	Total
HTH	56	56	112
UTK	54	56	110

Traits

- Phenice (1969) traits as described in Klales et al. (2012) (Figure 1)
  - subpubic concavity/contour (SPC)
  - ventral arc (VA)
  - medial aspect of the ischio-pubic ramus (MA)
- Walker (2008) traits as found in Buikstra & Ubelaker (1994) (Figure 2)
  - nuchal crest (N)
  - mastoid process (M)
  - supra-orbital margin(SO)
  - glabella (G)
  - mental eminence (ME)

Scoring

- Traits were scored on an ordinal scale from one to five by three observers with multiple levels of experience using the descriptions/illustrations from each method

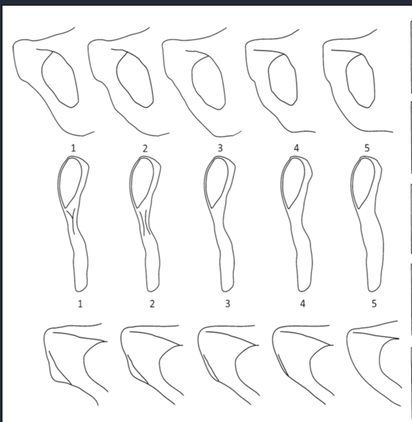

Fig 1. Klales et al. (2012) traits. Top: SPC, Middle: MA, Bottom: VA.

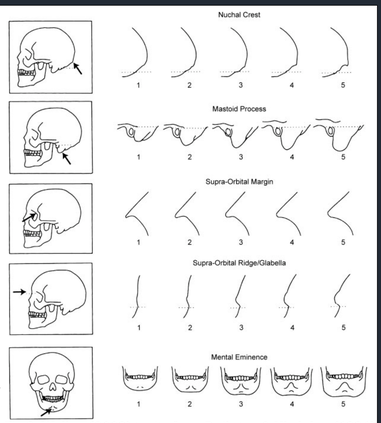

Fig 2. Walker (2008) traits from Buikstra & Ubelaker (1994).

Analyses

- Interobserver (between multiple scorers) was assessed using the intraclass correlation coefficient (ICC) with a two-way random consistency model, agreement 95%
  - Scorer A: Expert (forensic anthropologist)
  - Scorer B: Experienced (second year masters student)
  - Scorer C: Inexperienced (undergraduate)
- Intraobserver error (one person, multiple trials) was calculated using Cohen's weighted Kappa (K)
- Logistic regression analysis (LR) was used to examine method classification accuracy between scorers of multiple experience levels

Results

- ICC was acceptable (i.e., > 0.7) for all traits when scored by all three experience levels (Table 2, yellow), except SO in the UTK collection
  - Experienced observers (A&B) had higher ICC (optimal >0.8 or excellent >0.9) for all traits except G & SO in HTH (Table 2, green)
- K results indicate substantial agreement for all three pelvic traits and fair to moderate agreement for the skull traits based on the Landis & Koch (1977) parameters

Table 2. ICC results by collection.

Trait	Scorer	HTH	UTK
VA	All 3	0.910	0.889
	A/B	0.923	0.942
	A/C	0.828	0.741
	C/B	0.842	0.795
SPC	All 3	0.837	0.886
	A/B	0.893	0.951
	A/C	0.657	0.737
	C/B	0.700	0.776
MA	All 3	0.847	0.870
	A/B	0.852	0.906
	A/C	0.727	0.719
	C/B	0.770	0.796
N	All 3	0.870	0.854
	A/B	0.849	0.838
	A/C	0.820	0.748
	C/B	0.773	0.806
G	All 3	0.875	0.786
	A/B	0.840	0.757
	A/C	0.788	0.649
	C/B	0.851	0.729
SO	All 3	0.779	0.484
	A/B	0.725	0.749
	A/C	0.657	0.281
	C/B	0.731	0.046
M	All 3	0.877	0.822
	A/B	0.885	0.811
	A/C	0.765	0.731
	C/B	0.814	0.709
ME	All	0.749	0.775
	K-C	0.685	0.763
	K-W	0.642	0.635
	W-C	0.669	0.680

Table 3. Classification accuracy (%) for each scorer by method using LR. HTH sample (white), UTK sample (dark gray).

		Klales et al. (2012) Method				Walker (2008) Method - All Traits			
Scorer	**Experience Level**	**Males**	**Females**	**Combined**	**Sex Bias**	**Males**	**Females**	**Combined**	**Sex Bias**
Scorer A	Expert	90.0	93.9	91.9	-3.9	96.4	87.5	92.0	8.9
Scorer B	Experienced	98.0	91.8	94.9	6.2	87.5	89.3	88.4	-1.8
Scorer C	Inexperienced	84.0	87.8	85.9	-3.8	85.7	85.7	85.7	0.0
Scorer A	Expert	94.5	98.0	96.2	-3.5	82.4	86.5	84.5	-4.1
Scorer B	Experienced	90.7	96.0	93.3	-5.3	74.5	80.8	77.7	-6.3
Scorer C	Inexperienced	81.8	78.0	79.2	3.8	78.4	82.7	80.6	-4.3

- Classification accuracy varied by experience level (Table 3)
  - Overall the more experienced observers achieved higher classification accuracy for both methods, with the exception of the Walker method for the UTK collection in which Scorer C achieved better classification than Scorer B

Discussion & Conclusions

- Overall agreement between and within observers was at or above acceptable levels indicating that the Klales et al. (2012) and Walker (2008) methods are reliable
  - However, reliability was higher for the two experienced observers for most traits suggesting that the practitioner should have some familiarity and practical experience with the methods and traits prior to using the method
  - Score differences between observers were random (i.e., no one scorer was consistently scoring higher or lower for traits)
- Experience and greater training increased the validity (classification accuracy) of the methods
  - Classification was higher for the pelvis than the skull
  - Sex bias was low (< 9.0%)
- The ME has been considered difficult to score and reliability in other research has been lower than those found here
  - Reliability was likely high in this study because the majority of individuals were scored between 2-4 (cluster around the median) by all observers
  - The lack of extreme scores (1 or 5) indicate Walker's images may not include the full range of variation found in this trait and/or it is difficult to score

Acknowledgements

This research was funded by National Institute of Justice grant 2015-DN-BX-K014 entitled *An Interactive Morphological Database for Estimating Sex in Modern Adults* awarded to Dr. Klales, September 2015. Thanks go to Lyman Jellema for access to the Hamann-Todd Collection & Dr. Dawnie Steadman for access to the William M. Bass Donated Collection. Thanks also go to Stephanie Cole and Alexis Winter for participating in data collection.

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