Received: 19 June 2024 | Revised: 8 October 2024 | Accepted: 10 October 2024

DOI: 10.1111/1556-4029.15645

JOURNAL OF FORENSIC SCIENCES AAFS

TECHNICAL NOTE

Anthropology

Sex estimation using metrics of the innominate: A test of the DSP2 method

Kate M. Lesciotto JD PhD[1] | Alexandra R. Klales PhD[2]

[1]University of North Texas Health Science Center, Fort Worth, Texas, USA

[2]Washburn University, Topeka, Kansas, USA

Correspondence
Kate M. Lesciotto, University of North Texas Health Science Center, 3500 Camp Bowie Blvd, Fort Worth 76107, TX, USA.
Email: kate.lesciotto@unthsc.edu

Funding information
Division of Social and Economic Sciences, Grant/Award Number: 2214747; National Institute of Justice, Grant/Award Number: DOJ-NIJ-22-RO-0007

Abstract

Sex estimation is a critical component of the biological profile, and forensic anthropologists may use a variety of sex estimation methods depending upon the degree of completeness and state of preservation of the skeletal remains being analyzed. The innominate is widely accepted to be the most sexually dimorphic skeletal element. The *Diagnose Sexuelle Probabiliste* (DSP) method, which uses 10 measurements of the innominate, was introduced in 2005 and updated as DSP2 in 2017. While DSP2 has been reported to have high classification accuracy rates in studies of South American and European populations, the method has not been widely tested in US samples, and few US practitioners incorporate this method into their casework. The goal of this study was to test the reliability and accuracy of DSP2 using a large, modern sample from the US ($n = 174$). Two observers, blinded from demographic information associated with each specimen, collected the DSP2 metrics. Intra- and interobserver error analyses showed acceptable levels of agreement for all measurements, except for IIMT. Classification accuracies exceeded 95%, with minimal sex bias, for both observers and using various measurement combinations; however, an inclusivity sex bias occurred with more males reaching the 0.95 posterior probability threshold required by DSP2 to provide a sex classification estimate. Based on its high accuracy, forensic anthropologists in the US may consider incorporating DSP2 into their casework, although we recommend excluding IIMT and using SPU with caution. Additional methods will continue to be needed when the posterior probability threshold is not reached.

KEYWORDS
DSP2, forensic anthropology, innominate metrics, interobserver error, intraobserver error, sex estimation

Highlights

- Fewer females reached the posterior probability threshold required by DSP2 for sex classification.

The research presented in this manuscript has been submitted for presentation at the 77th Annual Conference of the American Academy of Forensic Sciences, February 17–22, 2025, in Baltimore, MD.

- Classification accuracy remains high even when using several different measurement combinations.
- DSP2 should be considered when morphological traits are not definitive for sex estimation.

1 | INTRODUCTION

Sex estimation is an integral part of constructing a biological profile in forensic anthropology. After determining whether the individual is an adult (versus subadult), sex estimation is typically the first parameter of the biological profile to be estimated, as many age and stature estimation methods are sex-specific. Although population affinity or ancestry estimation methods are not typically sex-specific, there is some evidence that differences between males and females can affect the traits that are most commonly used for population affinity estimation [1]. As a result, it is imperative that forensic anthropologists have a range of sex estimation methods at their disposal.

Nearly every skeletal element has been evaluated for its utility in sex estimation, although the innominate is largely regarded as the most sexually dimorphic skeletal element and therefore the most useful for sex estimation. Methods for sex estimation using the innominate focus on either metric or morphological traits. The most commonly utilized morphological traits of the innominate were originally described by Phenice [2] and were more recently revised and incorporated into statistical models by Klales et al. [3] and in the MorphoPASSE online program [4, 5]. Several commonly cited references include trait lists for sex estimation with additional morphological traits of the innominate [6–8]. While the program Fordisc 3.1 [9] incorporates two measurements of the innominate in the postcranial module that combines sex and population affinity estimation, there are fewer methods focused solely on measurements of the innominate. Bytheway and Ross [10] and Klales et al. [11] described geometric morphometric approaches to sex estimation using the innominate with high levels of accuracy; however, these approaches have not been translated into formal methods or applications.

In 2005, Murail et al. [12] introduced a new tool named *Diagnose Sexuelle Probabiliste* (DSP), or Probabilistic Sex Diagnosis. DSP was based on Fisher's linear discriminant analysis utilizing 10 linear measurements of the innominate. The freely available online tool was based on an Excel spreadsheet, where a user could enter the data acquired for each specimen, and the tool would provide the posterior probability of each specimen being male or female, with sex being assigned only when the posterior probability exceeded 0.95. DSP could be used with a minimum of four measurements but provided greater accuracy and an increased proportion of individuals meeting the 0.95 posterior probability threshold when using more than the minimum number of variables [12].

The reference samples supporting DSP included data from primarily historic (i.e., early 20th century or prior) skeletal collections from France, England, Portugal, Lithuania, South Africa, and the United States, as well as a late 20th century forensic collection from Thailand, totaling 2040 innominates [12]. Initial investigations determined that the measurements defined for use with DSP represented the full range of sexual dimorphism for all modern human populations, meaning that with DSP population-specific formulae were not required [12]. Although originally envisioned for more archaeological applications, Murail et al. [12] asserted that DSP was applicable to all anatomically modern humans and could be used in forensic contexts. The original DSP tool remains available as a graphical user interface (GUI) online at https://osteomics.com/DSP/.

In 2017, DSP was updated to DSP2, a freely downloadable program (http://projets.pacea.u-bordeaux.fr/logiciel/DSP2/dsp2.html). DSP2 was validated by Brůžek et al. [13] using both the original DSP reference sample and two new target samples from the Maxwell Museum Documented Collection housed at the University of New Mexico and the Simon Collection from the University of Geneva. For all samples tested, DSP2 produced accuracies >95% while remaining inclusive of the total sample, with between 85% and 95% of the tested individuals producing posterior probabilities that exceeded the 0.95 threshold for classification when all 10 measurements were used [13].

Diagnose Sexuelle Probabiliste and DSP2 have been tested in several studies using data from dry bone or virtual CT models from populations not represented in the reference dataset, including modern samples from Brazil, France, Greece, and Romania and archaeological samples from the medieval Eastern Adriatic and pre-Columbian mummies [14–22]. Each of these studies found overall high accuracy rates for DSP or DSP2, relatively low intra- and interobserver error rates, and fairly high levels of inclusivity (i.e., the proportion of individuals with a posterior probability above the 0.95 threshold for whom a sex estimate was provided by the program) when all measurements were available. However, a close examination of several studies purporting to test observer error and accuracy for DSP or DSP2 reveals the use of inappropriate statistics [14, 17], small or unstated observer error sample sizes [16, 21], discrepancies as to whether intra- or interobserver error was being tested [16], or evaluation of only 4 of the defined measurements [14, 15]. Despite these issues, a recent review of sex estimation methods recommended DSP2 as the "method of choice" when the innominate is well preserved in both bioarcheological and forensic contexts [23].

Perhaps as a result of the issues with previous tests of DSP2, this method has not gained widespread acceptance among forensic anthropology practitioners within the United States. A recent survey of forensic anthropology practitioners received responses primarily from individuals performing casework within the United States (95.7% of respondents) [24]. When using morphological traits, the pelvis (innominates and sacrum) was the most preferred skeletal region for sex estimation. However, when using metric methods, the average of respondents' ranking scores placed the pelvis as

the third most preferred, behind the long bones and the skull [24]. Respondents were also asked to rank individual methods in order of preference, and this method ranking reflected the preference for morphological traits of the pelvis for sex estimation [24]. On a Likert scale of 1–5, where 1 indicated that the respondent was "extremely unlikely" to use a method and 5 indicated that the respondent was "extremely likely" to use a method, the average ranking for DSP2 was only 1.9, lower than nearly all morphological methods using the innominate. For comparison, the morphological methods for sex estimation using the innominate described in Klales et al. [3] and the MorphoPASSE program [4] were ranked 4.4 and 4.3, respectively.

Additional insight on US-based practitioner preferences comes from the Forensic Anthropology Database for Assessing Methods Accuracy (FADAMA) [25], an online forensic case database. The FADAMA website allows registered users to submit case information, including which methods were used to estimate parameters of the biological profile. This information is added via drop-down menus, populated with methods that are commonly used by forensic anthropology laboratories with high caseloads and methods identified through a literature review. If a method is not listed, users can request that a new method be added to the drop-down menu. As of May 2024, FADAMA included information from nearly 600 forensic cases from the US in which the submitting user estimated sex; however, neither DSP nor DSP2 was listed as a sex estimation method in any of these cases, and neither method has been included in the drop-down menu of available sex estimation methods [25].

The goals of this project were to evaluate the potential utility of the DSP2 software for forensic anthropological casework in the US by: (1) performing intra- and interobserver error analyses using appropriate statistical analyses on all of the 10 measurements included in DSP2; (2) assessing accuracy and inclusivity of DSP2 on a large sample of known individuals from modern US skeletal collections; and (3) comparing results to previous tests based on the number of measurements available for analysis.

2 | MATERIALS AND METHODS

This study utilized the Southeast Texas Applied Forensic Science Facility Skeletal Collection at Sam Houston State University (STAFS), the Texas State University Donated Skeletal Collection (TXST), and the Documented Skeletal Collection at the Maxwell Museum of Anthropology at the University of New Mexico (Maxwell). While forensic anthropology, as a field, is moving towards the estimation of population affinity (morphological or genetic similarity) in forensic casework, we note that the demographic information maintained by most skeletal collections reflects ancestry (continental origins) or social race identities as reported by the donating individuals or their families [26]. Therefore, the demographics for the samples used in this study reflect the terminology used by each skeletal collection. Two observers collected data at each of these collections. Both observers were blinded to all demographic information associated with each individual during data collection. While 61 individuals from the Maxwell collection were used by Brůžek et al. [13] as a target or test sample, none of those individuals are included in the reference data used by the DSP2 software to estimate sex.

All measurements were taken according to the definitions and images provided in the DSP2 software [13] (Table 1). For this study, the following clarifications were made where the DSP2 measurement definition was silent or when a specimen did not perfectly align with the provided description:

- PUM—measurement taken without requiring the superior pubic ramus to be held in perfect horizontal alignment (as shown in the DSP2 exemplar image).
- SA—if the arcuate line was faint or forked, the auricular point was estimated.
- IIMT—measurement taken using the internal jaws of sliding calipers as suggested by Santos et al. [27] and as shown in the DSP GUI rather than a friction caliper as noted in the DSP2 program [13].
- SCOX—exostoses were excluded along ASIS and PSIS.
- PUM, SPU, ISMM, VEAC, SIS—depending the presence of lipping on the acetabulum, the exact border of the acetabular rim or border used in these measurements was estimated when possible or not scored in cases of severe lipping.

Measurements were taken using the external jaws of digital sliding calipers, except for DCOX (osteometric board), SCOX (spreading calipers), and IIMT (internal jaws of sliding calipers). The left innominate was used except when unavailable or damaged, and then, the right side was substituted. Measurements for each individual were entered into the DSP2 software. The DSP2 software highlights any measurements with out-of-range values compared to the reference sample; even with out-of-range values, DSP2 will still provide a classification and posterior probability. Any out-of-range values were noted. Male and female posterior probability values were recorded, along with whether each individual was classified by DSP2 as Male, Female, or Not Predicted (for individuals who did not exceed the required posterior probability threshold of 0.95).

Intraobserver error was tested on a sample of 35 White individuals (16M, 19F) from the STAFS collection during an initial pilot phase of this research. Each observer collected two trials of data, with approximately 5 days in between trials. During this initial study, IIMT was not measured, as the observers did not have access to divider calipers and adhered to the instructions of Brůžek et al. [13].

Interobserver error and accuracy was tested on a sample of 174 individuals (Table 2), including 80 individuals from the Maxwell collection and 94 individuals from the TXST collection. Demographics of the sample used for the current study are divided by sex and ancestry. The entire set of 10 measurements used by the DSP2 software was collected for the Maxwell and TXST samples, although not every measurement could be taken for every individual. During this phase of the study, following the guidance of Santos et al. [27], IIMT was measured using the internal jaws of sliding calipers. Due to the exclusion of IIMT during the initial pilot portion of the research on

TABLE 1 Abbreviations and definitions for the innominate measurements taken from the DSP2 program [13].

Abbreviation	Measurement	Definition
PUM	Acetabulo-symphyseal pubic length	Minimum distance from the superior and medial point of the pubic symphysis to the nearest point on the acetabular rim at the level of the lunate surface
SPU	Cotylo-pubic width	Pubic breadth between the most lateral acetabular point and the medial aspect of the pubis. Measurement is perpendicular to the major axis of the os pubis. Arms of the sliding caliper are thus parallel to the plan of the obturator foramen
DCOX	Maximum pelvic height	Maximum height of the os coxae measured from the inferior border of the os coxae to the most superior portion of the iliac crest. Can be taken with sliding calipers or osteometric board
IIMT[a]	Depths of the great sciatic notch	Distance from the postero-inferior iliac spine (defined as the point of intersection between the auricular surface and the posterior portion of the sciatic notch) to the anterior border of the great sciatic notch. This dimension must be measured with a divider caliper
ISMM	Post-acetabular ischium length	Distance from the most anterior and inferior point of the ischial tuberosity to the furthest point on the acetabular border
SCOX	Iliac breadth	Distance between the antero-superior iliac spine and the postero-superior iliac spine
SS	Spino-sciatic length	Minimum distance between the antero-inferior iliac spine and the deepest point in the greater sciatic notch
SA	Spino-auricular length	Distance between the antero-inferior iliac spine and the auricular point. Auricular point is defined as the intersection of the arcuate line with the auricular surface
SIS	Cotylo-sciatic breadth	Distance between the lateral border of the acetabulum and the midpoint of the anterior portion of the great sciatic notch. Fixed arm of the sliding caliper is parallel to the acetabular plane
VEAC	Vertical acetabular diameter	Maximum vertical diameter of the acetabulum, measured on the acetabular rim, as a prolongation of the longitudinal axis of the ischium

[a]The definition for IIMT within the DSP2 program differs slightly from the DSP GUI that remains available online, which additionally states: "Axis of the measurement must be perpendicular to the anterior border. Because of the configuration of [the] hip bone, it is easier to use small arms of sliding caliper."

TABLE 2 Sample demographics for interobserver error analysis. Ancestry groups are based on the demographic categories used by the Maxwell and TXST collections.

Ancestry	Female	Male
White	66	61
Black	5	8
Hispanic	10	15
Multiracial[a]	0	5
Native American	1	1
Asian	0	2
Total	82	92

[a]Includes individuals with multiple ancestry groups listed in the collection demographics.

intraobserver error, only the data from the Maxwell and TXST collections were used to evaluate accuracy and inclusivity of the DSP2 software.

Descriptive statistics and boxplots were used to visualize the data and check for any gross measurement errors or errors in data entry. Two-sample *t*-tests were used to determine if there were significant differences between males and females for each measurement. Intra- and interobserver error rates were calculated as the technical error of measurement (TEM), relative technical error of measurement (rTEM), and intraclass correlation coefficient (ICC). TEM is frequently used within anthropology to quantify intra- and interobserver reliability for continuous variables (i.e., measurement data). TEM is calculated by taking the square root of the sum of the squared differences between two measurements divided by the total number of subjects multiplied by two [28]. The resulting TEM value is easily understandable, as it retains the unit of measurement. The rTEM represents the TEM as a percentage relative to the total average. Acceptable limits for rTEM have been cited as <1.5% for intraobserver error and <2.0% for interobserver error [28, 29] or <5.0% for either intra- or interobserver error [30], although others have noted that there is no universal "acceptable" rTEM that can be applied to every study or measurement [31]. The ICC is another commonly used metric to quantify the reliability of measurements by comparing variability in a measurement for the same specimen or subject to the total variation across all measurements and all specimens or subjects. ICC values range from 0 to 1, with the following commonly accepted thresholds: <0.5 indicates poor reliability, 0.5–0.75 indicates moderate reliability, 0.75–0.9 indicates good reliability, and >0.9 indicates excellent reliability [32].

Accuracy rates were calculated based on the subset of individuals for whom sex was correctly predicted by DSP2. The inclusivity rate, or the percent predicted, was calculated based on the number of individuals from the entire sample that had sex predicted by DSP2 (i.e., the proportion of individuals for whom the posterior probability exceeded the 0.95 threshold, regardless of whether the

predicted sex was correct). Since all 10 measurements are not always able to be taken from every individual in forensic casework, several combinations of variables were tested to evaluate the effect of missing data on accuracy and inclusivity, thereby better simulating actual casework. In the development of the DSP2 software, the reference and validation samples were tested with several combinations of variables, including all 10 measurements and eight measurements excluding SIS and VEAC (which were intended to be used only for cases of incomplete preservation) [13]. Additionally, since the method requires a minimum of four measurements, several combinations using only four measurements were tested. The developers of DSP and DSP2 found that the 'best' four measurements were DCOX, PUM, SPU, and IIMT, which still classified 87% of individuals at 99.5% accuracy, and the 'worst' four measurements were SIS, VEAC, SA, and SS, which classified only 42% of individuals at 98.7% accuracy [13]. For this study, each of these analyses was run on a subset of the total sample, using only those individuals who had measurements recorded for each of the required variables.

While efforts were made to include individuals from as many ancestry groups as possible, the sample was predominantly White. Due to the small number of individuals with ancestries other than White, a Fisher's exact test was used to determine whether there was a significant relationship between ancestry and whether an individual's predicted sex was correct, incorrect, or not predicted by the DSP2 software. All statistical analyses were carried out using R and R Studio.

3 | RESULTS

After checking the Maxwell and TXST data for measurement and data entry errors, statistically significant differences between males and females were found for all measurements except SA ($p = 0.1954$) and PUM ($p = 0.05465$) (Table 3). The male mean was larger for all measurements, except for IIMT and PUM. Several measurements were recorded that exceeded the range of the DSP2 reference data: SA (five individuals 94.8–97.3; DSP2 maximum value 94.7), SIS (one individual 52.2; DSP2 maximum value 52.0); SPU (three individuals at 39.0; DSP2 maximum value of 38.5), and SS (two individuals 91.8–93.4; DSP2 maximum value 91.0).

The intraobserver error analyses from the initial pilot study using the STAFS collection ($n = 35$) showed acceptable results for nearly all measurements defined by DSP2 (Tables 4 and 5). As previously mentioned, IIMT was not tested during this initial pilot phase of the study. SPU had the highest rTEM for intraobserver analysis for both observers (Obs 1: 2.23%; Obs 2: 3.58%). The ICC 95% confidence interval (CI) remained within the threshold for excellent reliability for all measurements (>0.9) for both observers, with the exception of PUM, SCOX, and SPU for Observer 2 which encompassed values in the good and excellent reliability categories.

The interobserver error analyses using the Maxwell and TSXT collections ($n = 174$) showed acceptable results for most measurements (Table 6). IIMT had the highest rTEM of 6.56%, which exceeds commonly used thresholds for "acceptable" measurements, and had the lowest ICC of 0.787, with a 95% CI of 0.637–0.867, which spans the moderate and good reliability categories. The rTEM for SPU was also slightly high at 2.83%; however, the ICC was 0.958 with a 95% CI of 0.931–0.973, which remained wholly within the range for excellent reliability.

Fisher's exact tests showed no significant difference between the number of individuals for whom sex was correctly predicted, incorrectly predicted, and not predicted when separated by ancestry group (all $p > 0.05$). Since classification rates were not significantly affected by ancestry, the samples were condensed into female and male groups. Female and male classification rates exceeded 95% accuracy for both observers; however, the number of individuals whose posterior probability exceeded the 0.95 threshold (and were therefore classified by the DSP2 software) was markedly different between males and females. For both observers, females were classified at a lower rate (Obs 1: 75.6%; Obs 2: 82.9%) compared to males (Obs 1: 96.7%; Obs 2: 91.3%) (Table 7).

The data were also examined to test how several combinations of variables impacted classification rates using Observer 1's data (Table 8). Even when using the 'worst' four variables, accuracy rates

TABLE 3 Descriptive statistics for the DSP2 measurements. Range end-points that exceed the DSP2 reference data are highlighted.

	Males			Females			
Measurement	*n*	Range	Mean (SD)	*n*	Range	Mean (SD)	*p*-Value
DCOX	88	191.0–247.0	225.0 (11.6)	81	184.0–222.0	203.0 (8.49)	<2.2e-16
IIMT	88	28.1–58.9	40.8 (5.62)	82	32.3–55.9	45.0 (4.99)	6.697e-07
ISMM	88	98.1–131.0	116.0 (6.01)	76	93.9–112.0	102.0 (4.34)	<2.2e-16
PUM	90	60.7–85.1	72.3 (5.34)	68	62.4–84.9	73.9 (4.76)	0.05465
SA	87	65.7–97.3	81.8 (6.48)	76	63.2–95.7	80.3 (5.76)	0.1232
SCOX	87	137.0–181.0	160.0 (9.36)	70	128.0–173.0	154.0 (8.98)	0.0001672
SIS	76	32.1–52.2	42.3 (3.78)	70	28.3–47.0	37.8 (3.12)	6.004e-13
SPU	91	26.5–39.0	31.9 (3.02)	73	19.3–34.8	26.1 (2.89)	<2.2e-16
SS	91	65.7–93.4	79.7 (5.73)	79	59.3–80.2	71.4 (4.73)	<2.2e-16
VEAC	88	50.9–65.6	57.3 (3.39)	74	44.2–55.8	50.7 (2.58)	<2.2e-16

TABLE 4 Intraobserver error for Observer 1. Higher rTEM values are highlighted.

Measurement	TEM (mm)	rTEM (%)	ICC (95% CI)
DCOX	1.27	0.59	0.993 (0.986–0.996)
ISMM	0.92	0.83	0.99 (0.976–0.995)
PUM	0.86	1.17	0.949 (0.903–0.974)
SA	1.18	1.43	0.964 (0.93–0.981)
SCOX	1.11	0.72	0.982 (0.965–0.991)
SIS	0.37	0.90	0.993 (0.987–0.997)
SPU	0.67	2.23	0.974 (0.94–0.988)
SS	0.49	0.65	0.993 (0.985–0.996)
VEAC	0.64	1.20	0.974 (0.95–0.987)

TABLE 5 Intraobserver error for Observer 2. Higher rTEM values are highlighted.

Measurement	TEM (mm)	rTEM (%)	ICC (95% CI)
DCOX	1.68	0.79	0.987 (0.973–0.994)
ISMM	0.71	0.64	0.994 (0.987–0.997)
PUM	0.88	1.20	0.944 (0.892–0.971)
SA	0.66	0.81	0.988 (0.977–0.994)
SCOX	1.65	1.04	0.962 (0.893–0.983)
SIS	1.41	0.98	0.993 (0.985–0.996)
SPU	1.02	3.58	0.92 (0.85–0.958)
SS	0.68	0.90	0.987 (0.975–0.993)
VEAC	0.76	1.40	0.965 (0.933–0.982)

TABLE 6 Interobserver error analysis. Higher rTEM and lower ICC values are highlighted.

Measurement	TEM (mm)	rTEM (%)	ICC
DCOX	3.442	1.618	0.947 (0.797–0.977)
IIMT	2.860	6.561	0.787 (0.637–0.867)
ISMM	1.213	1.112	0.98 (0.928–0.991)
PUM	1.077	1.470	0.955 (0.932–0.97)
SA	1.555	1.932	0.939 (0.813–0.972)
SCOX	2.122	1.354	0.953 (0.87–0.977)
SIS	0.453	1.131	0.988 (0.983–0.991)
SPU	0.824	2.827	0.958 (0.931–0.972)
SS	0.859	1.131	0.984 (0.978–0.988)
VEAC	0.994	1.835	0.95 (0.932–0.963)

remained above 95% for the total sample, as well as for both females and males. Inclusivity rates dropped only slightly, from DSP2 predicting sex for 92.1% of sample using all 10 measurements down to predicting sex for only 88.1% of the sample using the 'worst' four measurements. Differences in inclusivity between the female and male subsamples remained stable across the tests of different variable combinations, with approximately 15–20% more of the male sample reaching the 0.95 posterior probability threshold and being classified by the DSP2 software. Based on low performance in the interobserver error analysis, an additional test removing IIMT was also performed. The exclusion of IIMT had no effect on either accuracy or inclusivity of the sample compared to using all 10 measurements.

4 | DISCUSSION

The reliability of the DSP/DSP2 measurements have been tested several times in different regional samples, yet differences in statistical methodology and reporting make direct comparisons across studies difficult. Chapman et al. [17] presented an early test of the original DSP spreadsheet and reported "no statistical difference between interobserver measurements" with only a single *p*-value, presumably from the Student's *t*-test described in the methodology section, and "perfect agreement between observers (*Kappa* = 1, $p < 0.000$)." Student's t-tests will test whether a significant difference exists between the means of two groups and is thus not an appropriate statistic for quantifying intra- or interobserver error, and it is unclear how a single *p*-value relates to 10 different measurements. The reported *Kappa* test appears to have tested the agreement between observers on the predicted sex output of DSP, rather than observer error on the actual measurements [17]. These details are significant, as a subsequent test of DSP stated that "excellent interobserver agreement of the measurements has already been demonstrated" [18], while citing Chapman et al. [17]. de Oliveira Lopes et al. [14] examined the performance of the best four of the DSP/DSP2 measurements (DCOX, SPU, PUM, and IIMT) and reported a single "*kappa* concordance measure," but this also seems to be assessing the agreement between observers in the final sex that was predicted by the software rather than a direct test of the reliability of the actual measurements. Kranioti et al. [16] did provide quantifications of what they referred to as "inter-observer" error as TEM, rTEM, and the coefficient of reliability. However, the methods section specified that measurements were taken by a single observer, making it unclear whether the results were truly interobserver error or actually *intra*observer error.

Two studies were identified that individually tested both intra- and interobserver error for each of the 10 DSP2 measurements [19, 20]. In both of these studies, IIMT had the highest levels of intra- and interobserver error. Machado et al. [20] reported an rTEM of 6.26% for intraobserver error and 7.09% for interobserver error, and de Almeida et al. [19] found ICC levels of 0.926 for intraobserver error and 0.837 for interobserver error. While an ICC level of >0.900 is typically considered to be "excellent reliability," IIMT did have the lowest reliability levels out of the 10 measurements [19]. The current study also found IIMT to be an unreliable measure, with an rTEM of 6.56% and ICC of 0.787 for the interobserver error analysis. The lower observer error in this study and previous research is likely due to several factors including the difficulty of properly orienting the calipers, variation in the type of calipers used to collect this measurement, and differences in how DSP/DSP2 defines the posterior

TABLE 7 Classification accuracy and inclusivity rates ($n = 174$). "% Accuracy" was calculated as the proportion of individuals for whom sex was correctly classified by DSP2 with a posterior probability that exceeded 0.95. "% Predicted" was calculated as the proportion of individuals that reached the posterior probability threshold of 0.95 and had sex predicted by DSP2 (either correctly or incorrectly).

	Observer 1		Observer 2	
Sex	% Accuracy	% Predicted	% Accuracy	% Predicted
Females ($n = 82$)	95.2% (59/62)	75.6% (62/82)	98.5% (67/68)	82.9% (68/82)
Males ($n = 92$)	100% (89/89)	96.7% (89/92)	100% (84/84)	91.3% (84/92)
Overall ($n = 174$)	98.0% (148/151)	86.8% (151/174)	99.3% (151/152)	87.4% (152/174)

TABLE 8 Classification accuracy and inclusivity rates based on different combinations of variables using Observer 1's data. "% Accuracy" was calculated as the proportion of individuals for whom sex was correctly classified by DSP2 with a posterior probability that exceeded 0.95. "% Predicted" was calculated as the proportion of individuals that reached the posterior probability threshold of 0.95 and had sex predicted by DSP2 (either correctly or incorrectly).

	Total		Females		Males	
Variables	% Accuracy	% Predicted	% Accuracy	% Predicted	% Accuracy	% Predicted
All 10	100% (105/105)	92.1% (105/114)	100% (41/41)	83.7% (41/49)	100% (64/64)	98.5% (64/65)
9 (w/o IIMT)	100% (105/105)	92.1% (105/114)	100% (41/41)	83.7% (41/49)	100% (64/64)	98.5% (64/65)
8 (w/o SIS and VEAC	100% (123/123)	92.4% (123/133)	100% (45/45)	83.3% (45/54)	100% (78/78)	98.7% (78/79)
Best 4[a]	100% (134/134)	89.9% (134/149)	100% (51/51)	78.5% (51/65)	100% (83/83)	98.8% (83/84)
Worst 4[a]	99.2% (117/118)	88.1% (118/134)	98.0% (50/51)	79.7% (51/64)	100% (67/67)	95.7% (67/70)

[a]According to Brůžek et al. [13].

inferior iliac spine compared to other osteology references. Given the high observer error rates and evidence that the removal of IIMT does not affect either accuracy or inclusivity, we recommend that DSP2 be run without IIMT for forensic casework.

The other DSP2 measurement that warrants further discussion is SPU. Machado et al. [20] reported high rTEM values for SPU, with an intraobserver error of 6.02% and an interobserver error of 4.22%, while de Almeida et al. [19] found more acceptable ICC measures of 0.991 [0.981–0.996] for intraobserver error and 0.937 [0.862–0.972] for interobserver error. In the current study, the rTEM for SPU ranged from 2.23% to 3.58%, which could be interpreted as slightly high or acceptable [28–30]; however, the ICCs for SPU ranged from 0.92 to 0.974, indicating overall excellent reliability. Due to the discrepancies between rTEM and ICC in the current study and previous studies, we recommend that SPU be used with caution, particularly in individuals with lipping or osteophytic growth along the acetabular rim that may affect the ability to locate the most lateral acetabular point. The potential impact of both IIMT and SPU on overall classifications with DSP2 should be investigated further, especially considering that both of these measurements are included in the "best" four measurements by the method developers [12, 13].

The high accuracy of the DSP2 software exceeds the reported accuracy rates of frequently used morphological sex estimation methods using the innominate, including the Klales et al. [3] method (up to 93.5%). However, the higher DSP2 accuracy rate comes at the cost of excluding a large proportion of the original sample due to posterior probabilities that do not exceed the required 0.95 threshold, as sex for these individuals is not predicted by the DSP2 software. There is also a notable sex bias in the individuals who do not reach this 0.95 required threshold, with approximately 10–20% fewer females reaching this threshold and being classified by the DSP2 program compared to males in the current study. This pattern is similar to that reported by Quatrehomme et al. [18]. This bias was not seen among the DSP2 reference samples, where roughly equal proportions of females and males reached the posterior probability threshold (and were therefore classified) regardless of which combination of variables was tested [27]. The opposite pattern, with a higher proportion of males failing to reach the posterior probability threshold, was found in several previous tests of the method [15, 19, 27].

The "Standard for Sex Estimation in Forensic Anthropology" published by the AAFS Standard Board ("ASB") recommends that, when the skeletal elements available for analysis allow for multiple methods of sex estimation, "the method(s) with the greatest accuracy" should be given greater weight in the final conclusion [33]. However, this Standard does not define how "accuracy" should be quantified or the role of posterior probabilities or other measures of confidence in the final classification. The DSP2 software produced classification accuracies exceeding 95%, in agreement with previously published studies. Even using the reportedly worst combination of only four measurements, accuracy remained high, with little sex bias. Removing IIMT, which was the worst-performing variable in the interobserver error study, had no effect on accuracy. Interestingly, DSP/DSP2 is not one of the methods referenced in Standard 090

[33] and is not widely used for US casework [24] despite the high accuracy and reliability.

The recent survey of practicing forensic anthropologists [24] hints at a possible reason why DSP2 is not widely utilized by US practitioners. Respondents reported being more likely to use "user-friendly" methods for sex estimation [24]. Forensic anthropologists may feel that the morphological scoring methods of Klales et al. [3] and the MorphoPASSE program [4] are more "user-friendly," as they do not require any equipment and only assess three traits of the innominate, while DSP2 requires sliding and spreading calipers for taking up to 10 measurements. Alternatively, US practitioners may not be as familiar with DSP2 or its measurements, only two of which are included in the more traditionally taught postcranial measurements used by the program Fordisc [9]. Another possibility may be that US practitioners are aware of DSP2, yet are hesitant to use the method since it will only classify an individual if the 0.95 posterior probability threshold is reached.

The results of this study support the addition of DSP2 to the practicing forensic anthropologist's sex estimation 'toolkit.' When estimating sex in any modern forensic case, the anthropologist must consider each method's known accuracy rate, as well as the confidence in the classification for the specific individual in question. A method may have a high accuracy rate reported in the literature but produce a low posterior probability for the skeletal remains being analyzed. In such cases, the anthropologist should consider using a secondary method of sex estimation. For example, while the Klales et al. [3] and MorphoPASSE [4] methods may be less time-consuming and more user-friendly, the addition of a metric method using the innominate would be beneficial when morphological results produce sex estimates with lower posterior probabilities. Additionally, the observer error results demonstrate that the DSP2 measurements are relatively easy to take, and the required spreading and sliding calipers are standard equipment for a forensic anthropology laboratory. Finally, while the 0.95 posterior probability threshold does reduce the number of individuals for whom sex can be estimated with DSP2, the higher confidence in the final estimation is an acceptable trade-off, particularly when used in conjunction with other sex estimation methods.

5 | CONCLUSION

The goal of creating a biological profile within the field of forensic anthropology is to create the most accurate estimation of an individual's age, sex, ancestry or population affinity, and stature as possible, in order to assist with decedent identification. Towards this goal, Standard 090 "Standard for Sex Estimation in Forensic Anthropology" advises that both morphological and metric methods are acceptable for skeletal sex estimation, so long as the variables are clearly defined [30]. The results from this research suggest that DSP does have clearly defined variables (i.e., innominate measurements) that can be reliably collected, with the exception of IIMT and potentially also SPU. The DSP2 software has also been demonstrated to have high accuracy across multiple global populations, including in the modern US sample from this research. While the posterior probability threshold of 0.95 does significantly limit the number of individuals for whom sex can be predicted, this may not act as a major limitation of this method, as practitioners already tend to use multiple metric and morphological methods for skeletal sex estimation [24]. Therefore, practitioners in the US may include DSP2 within standard practice for forensic anthropological casework when the innominate is sufficiently preserved to collect at least four of the DSP2 measurements, excluding IIMT and SPU, and particularly when morphological methods for sex estimation using the innominate do not produce results at high levels of confidence (e.g., high posterior probabilities).

ACKNOWLEDGMENTS

We want to thank the collection managers at the Southeast Texas Applied Forensic Science Facility at Sam Houston State University, Maxwell Museum at the University of New Mexico, and Forensic Anthropology Center at Texas State University for providing access to the documented skeletal collections. We also want to thank the donors and their families, without whom this research would not have been possible. Finally, we thank the National Science Foundation and the National Institute of Justice for providing financial support.

FUNDING INFORMATION

This research was funded by the National Science Foundation Grant No. 2214747 (co-funded by the National Institute of Justice DOJ-NIJ-22-RO-0007).

CONFLICT OF INTEREST STATEMENT

The authors have no conflicts of interest to declare.

DISCLAIMER

Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation or the National Institute of Justice.

REFERENCES

1. Klales AR, Kenyhercz MW. Morphological assessment of ancestry using cranial macromorphoscopics. J Forensic Sci. 2015;60(1):13–20. https://doi.org/10.1111/1556-4029.12563
2. Phenice TW. A newly developed visual method of sexing the os pubis. Am J Phys Anthropol. 1969;30(2):297–301. https://doi.org/10.1002/ajpa.1330300214
3. Klales AR, Ousley SD, Vollner JM. A revised method of sexing the human innominate using Phenice's nonmetric traits and statistical methods. Am J Phys Anthropol. 2012;149(1):104–14. https://doi.org/10.1002/ajpa.22102
4. Klales AR. MorphoPASSE: the morphological pelvis and skull sex estimation database. Ver 1.0. Topeka, KS: Washburn University; 2018.
5. Klales AR, Cole S. MorphoPASSE: the morphological pelvis and skull sex estimation database manual. Topeka, KS: Washburn University; 2018.

6. Buikstra JE, Ubelaker DH. Standards for data collection from human skeletal remains. Arkansas archaeological survey research series no. 44. Arkansas Archaeological Survey: Fayetteville, AR; 1994.
7. Krogman WM, Isçan MY. The human skeleton in forensic medicine. Springfield, IL: Charles C. Thomas; 1986.
8. Rogers T, Saunders S. Accuracy of sex determination using morphological traits of the human pelvis. J Forensic Sci. 1994;39(4):1047–56. https://doi.org/10.1520/JFS13683J
9. Jantz RL, Ousley SD. FORDISC 3: personal computer forensic discriminate functions. Knoxville, TN: University of Tennessee, Knoxville; 2005.
10. Bytheway JA, Ross AH. A geometric morphometric approach to sex determination of the human adult as coxa. J Forensic Sci. 2010;55(4):859–64. https://doi.org/10.1111/j.1556-4029.2010.01374.x
11. Klales AR, Vollner JM, Ousley SD. A new metric procedure for the estimation of sex and ancestry from the human innominate. Proceedings of the 61st annual scientific meeting of the American Academy of Forensic Sciences. Colorado Springs, CO: American Academy of Forensic Sciences; 2009. p. 311.
12. Murail P, Brůžek J, Houët F, Cunha E. DSP: a tool for probabilistic sex diagnosis using worldwide variability in hip-bone measurements. Bull Mém Soc Anthropol Paris. 2005;17(3–4):167–76. https://doi.org/10.4000/bmsap.1157
13. Brůžek J, Santos F, Dutailly B, Murail P, Cunha E. Validation and reliability of the sex estimation of the human os coxae using freely available DSP2 software for bioarchaeology and forensic anthropology. Am J Phys Anthropol. 2017;164(2):440–9. https://doi.org/10.1002/ajpa.23282
14. Lopes ARDO, Silva EML, Nascimento MMDS, Silva MC, Magalhães CP, Cerqueira GS. DSP2 for sex determination of miscegenated contemporary hip bones. Anat Histol Embryol. 2024;53(1):e12979. https://doi.org/10.1111/ahe.12979
15. Stan E, Muresan C-O, Dumache R, Ciocan V, Ungureanu S, Costachescu D, et al. Sex estimation from computed tomography of os coxae—validation of the Diagnose Sexuelle Probabiliste (DSP) software in the Romanian population. Appl Sci. 2024;14(10):4136. https://doi.org/10.3390/app14104136
16. Kranioti EF, Štovíčková L, Karell MA, Bružek J. Sex estimation of os coxae using DSP2 software: a validation study of a Greek sample. Forensic Sci Int. 2019;297:371. https://doi.org/10.1016/j.forsciint.2019.02.011
17. Chapman T, Lefevre P, Semal P, Moiseev F, Sholukha V, Louryan S, et al. Sex determination using the probabilistic sex diagnosis (DSP: Diagnose Sexuelle Probabiliste) tool in a virtual environment. Forensic Sci Int. 2014;234:189.e1–189.e8. https://doi.org/10.1016/j.forsciint.2013.10.037
18. Quatrehomme G, Radoman I, Nogueira L, Du Jardin P, Alunni V. Sex determination using the DSP (probabilistic sex diagnosis) method on the coxal bone: efficiency of method according to number of available variables. Forensic Sci Int. 2017;272:190–3. https://doi.org/10.1016/j.forsciint.2016.10.020
19. De Almeida SM, De Carvalho MVD, De Lyra Menezes MCT, Petraki GGP, Cunha E, Soriano EP. Validation of the DSP2 tool in a contemporary identified skeletal collection from northeastern Brazil. Adv Anthropol. 2020;10(2):169–80. https://doi.org/10.4236/aa.2020.102010
20. Machado MPS, Costa ST, Freire AR, Navega D, Cunha E, Júnior ED, et al. Application and validation of Diagnose Sexuelle Probabiliste V2 tool in a miscegenated population. Forensic Sci Int. 2018;290:351. https://doi.org/10.1016/j.forsciint.2018.06.043
21. Mestekova S, Brůžek J, Veleminska J, Chaumoitre K. A test of the DSP sexing method on CT images from a modern French sample. J Forensic Sci. 2015;60(5):1295–9. https://doi.org/10.1111/1556-4029.12817
22. Chapman T, Tilleux C, Polet C, Hastir J-P, Coche E, Lemaitre S. Validating the probabilistic sex diagnosis (DSP) method with a special test case on pre-Columbian mummies (including the famous Rascar Capac). J Archaeol Sci Rep. 2020;30:102250. https://doi.org/10.1016/j.jasrep.2020.102250
23. Kotěrová A, Rmoutilová R, Brůžek J. Current trends in methods for estimating age and sex from the adult human skeleton. Anthropologie. 2022;60(2):225–52. https://doi.org/10.26720/anthro.22.10.05.1
24. Klales A, Lesciotto KM. Sex estimation methods in forensic anthropology: current practice and trends. Proceedings of the 76th Annual Scientific Meeting of the American Academy of Forensic Sciences. Colorado Springs, CO: American Academy of Forensic Sciences; 2024. p. 117.
25. Hughes C, Juarez C. Learning from our casework: the forensic anthropology database for assessing methods accuracy (FADAMA). NIJ 2018-DU-BX-0213. 2018. https://www-app.igb.illinois.edu/sofadb. Accessed 10 Oct 2024.
26. Spradley K, Jantz RL. What are we really estimating in forensic anthropological practice, population affinity or ancestry? Forensic Anthropol. 2022;4(4):309–18. https://doi.org/10.5744/fa.2021.0017
27. Santos F, Guyomarc'h P, Cunha E, Brůžek J. DSP: a probabilistic approach to sex estimation free from population specificity using innominate measurements. In: Klales AR, editor. Sex estimation of the human skeleton: history, methods, and emerging techniques. San Diego, CA: Academic Press; 2020. p. 243–69. https://doi.org/10.1016/B978-0-12-815767-1.00015-8
28. Langley NR, Meadows Jantz L, McNulty S, Maijanen H, Ousley SD, Jantz RL. Error quantification of osteometric data in forensic anthropology. Forensic Sci Int. 2018;287:183–9. https://doi.org/10.1016/j.forsciint.2018.04.004
29. Perini TA, de Oliveira GL. Technical error of measurement in anthropometry. Rev Brasil Med Esporte. 2005;11(1):86–90. https://doi.org/10.1590/S1517-86922005000100009
30. Weinberg SM, Scott NM, Neisanger K, Marazita ML. Intraobserver error associated with measurements of the hand. Am J Hum Biol. 2005;14(3):368–71. https://doi.org/10.1002/ajhb.20129
31. Fancourt HSM, Stephan CN. Error measurement in craniometrics: the comparative performance of four popular assessment methods using 2000 simulated cranial length datasets (g-op). Forensic Sci Int. 2018;285:162–71. https://doi.org/10.1016/j.forsciint.2018.02.008
32. Koo TK, Li MY. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. J Chiropr Med. 2016;15(2):155–63. https://doi.org/10.1016/j.jcm.2016.02.012
33. American Academy of Forensic Sciences Standards Board. Standard 090: standard for sex estimation in forensic anthropology. Colorado Springs, CO: AAFS Standards Board; 2019.

How to cite this article: Lesciotto KM, Klales AR. Sex estimation using metrics of the innominate: A test of the DSP2 method. J Forensic Sci. 2025;70:249–57. https://doi.org/10.1111/1556-4029.15645