Ancestry Estimates: Evaluating the Reliability of Hefner's Cranial Morphoscopic Method

By

Chenée Merchant

A Thesis submitted to the Faculty of Graduate Studies of the University of Manitoba

In partial fulfillment of the requirements of the degree of

MASTER OF ARTS

Department of Anthropology

University of Manitoba

Winnipeg

Abstract

To identify unknown skeletal remains, forensic anthropologists provide police with information of who they might belong to, such as ancestry (someone's familial lineage and geographic origin). The cranium has shape-based traits (morphoscopic traits) that can be scored using visual analysis, and these scores are used to estimate ancestry. The purpose of this study is to evaluate the reliability of Hefner's (2009) morphoscopic trait scoring method, which assesses sixteen traits, as well as the impact of score disagreement among and within observers on the resulting ancestry estimates. Reliability is determined through intra-observer and inter-observer repeatability tests, whereas the impact of score disagreement is observed by comparing ancestry results generated by statistical programs from each observer's scores. In general, most traits have high intra-observer agreement, most trait scores are in agreement, and lower inter-observer agreement. Each trait has their own pattern of disagreement, such as a score of 2 and 3 were always confused with each other for the trait anterior nasal spine, but never a score of 1. Score disagreements caused ancestry estimates to change between observers in most cases. Error causing lower inter-observer agreement included experience, tool use, method iteration, prevalence of traits within the study individuals, vague descriptions, and interpretation differences. This is the first study to thoroughly assess and identify sources of error, as well as provide recommendations for improved descriptions/pictorial representation of all sixteen of Hefner's traits. Overall, Hefner's method requires pictorial and description improvement for the majority of traits before it can be reliably used among practitioners.

Acknowledgements

I extend my gratitude:

To the family who helped me emotionally, financially, or physically when I was struggling with my health or other personal issues: my husband Dana, my parents Shelley and Kelly, and Alex, and Paige.

To the Winnipeg friends who were there for me in study sessions, exercise sessions, tea days, craft nights, and/or game nights to get me through sanely: Nikki, Larissa, Jesse, Felicia, Maiah, Grace, Mahyar, Garth. Each of you has had an impact in my life and I am lucky to have met such wonderful people.

To the family and friends who cheered me on from afar and sent encouraging words: aunts, uncles, grandparents, siblings, parents in-law, Nick, the DnD fam.

To my advisor, Dr. Emily Holland, who had high standards but was completely understanding when I was unable to meet those standards due to personal complications. I thank her for her endless and thorough effort to push me to be my best, as well as the advice that "it is okay to rest when you are feeling good, not just when you feel unwell; you don't always need to use the good days to catch up."

To my committee members who gave me valuable advice and advocated for my employment so I could finish without too much stress: Dr. Rob Hoppa and Dr. Kathleen Buddle

To the department members who gave me valuable experience and encouraged me to carry out my ideas: Dr. Julia Gamble, Dr. Derek Johnson, Dr. Kent Fowler, Dr. Ben Collins, Dr. Rachel ten Bruggencate, Dr. Warren Clarke.

To Dr. Joseph Hefner who gave me his time to talk through statistics and access to the statistical program that was helpful in carrying out part of this research.

To the professors at the University of Alberta who believed I had the potential to do more and encouraged me to pursue this degree: Dr. Lesley Harrington, Pamela Mayne-Correia, Dr. Sandra Garvie-Lok, Dr. Daniel Graf.

Finally, to the funding agencies and opportunities that allowed this research to be completed: SSHRC Canada Graduate Scholarship- Master's (CGS M)- Joseph Armand Bombardier, University of Manitoba Graduate Fellowship (UMGF), and James Gordan Fletcher Graduate Research Award.

This research was a satisfying challenge to take on, and I hope that it provides the forensic anthropology community with valuable information so that families can be reunited in the future.

Table of Contents

List of Figures

List of Tables

Chapter 1: Introduction

The relationship between skeletal traits and genes has been studied by biological anthropologists, bioarchaeologists, and geneticists for decades in order to understand human population relationships and movement (Pilloud & Hefner, 2016). As such, forensic anthropologists use the relationship between skeletal traits and biological origin to estimate relatedness between an unknown individual and population, otherwise known as ancestry (someone's familial lineage or geographic origin) estimation. Translating skeletal traits into ancestry estimation is a challenge that requires multiple perspectives and method testing, such as this research intends to contribute to. Ancestry estimation can be done using both metric (measured) and non-metric (visual) analysis of dental (Edgar, 2013; Kenyhercz, Klales, & Kenyhercz, 2014; Pilloud, Hefner, Hanihara, & Hayashi, 2014; Turner II, Nichol, & Scott, 1991), cranial (Atkinson & Tallman, 2020; Hefner, 2009; Ousley & Jantz, 2012; Rhine, 1990; Stull, Kenyhercz, & L'Abbé, 2014), and post-cranial skeletal material (Christensen, Leslie, & Baim, 2014; Marino, 1997; Spiros & Hefner, 2020). All skeletal data for ancestry methods rely on the same theoretical principles, however, the focus of this research is on cranial non-metric traits used for ancestry estimation.

Non-metric traits are shape-based traits that cannot be measured using units of measurement like metric traits. Instead, they use visual assessment to distinguish between different expressions and are given a numerical value/score. Of particular interest is a sub-category of non-metric traits used by forensic anthropologists called "morphoscopic traits." These are quasi-continuous traits that can be "reflected as soft-tissue differences in the living" (Hefner, Ousley, & Dirkmaat, 2012, p. 295). Quasi-continuous traits have graded expressions (ex. a sliding scale from absent to small to large) that vary in frequency between populations, allowing for the comparison

of an individual's combination of traits to the trait frequencies of reference populations (Go & Hefner, 2020; Hefner, 2009). These comparisons determine which populations have a combination in the highest frequency that matches the unknown individual, resulting in a probability that the individual belonged to a certain population (Pilloud & Hefner, 2016). In this thesis, "non-metric" will be used to describe all non-metric traits, including morphoscopic traits.

In 2009, Hefner (2009) proposed a new ancestry assessment method that works in two parts: trait assessment and ancestry estimation. Trait assessment utilizes a standard data collection system to visually assess and score a suite of cranial morphoscopic traits (referred throughout as the 'scoring method'). There are sixteen traits total, with eleven initially introduced in Hefner (2009) and five more introduced later (Hefner & Linde, 2018; *Osteoware v. 2.4.037*, 2020). Trait scores are then used to estimate ancestry in a statistical framework of comparison. The purpose of this research, in relation to trait assessment, is to test the reliability of the scoring method, including the trait description updates in the most recent publication Hefner and Linde (2018). The reliability of a method depends on its repeatability, which is whether a method can consistently produce the same result. In this case, whether or not the same scores for an individual are generated over multiple assessments.

Testing the scoring method for its ability to provide the same scores across multiple researchers (ie. observers) and scoring periods is calculated as how often observer error, which is when observers have differing opinions on what the correct observation/score is, occurs. Multiple observer error studies can then reveal the error rates associated with the scoring method; error rates are defined as "a continuous, repeatable, consistent action that yields a predictable level of false positive or false negative results" (Budowle et al., 2009, p. 801)." A

qualitative assessment of how the scoring method is interpreted can provide insight to where sources of score disagreement are occurring and how observer error could be reduced.

The repeatability of the scoring method is important because it must produce reliable results in order for ancestry estimations based on trait frequency data to be accurate. Accuracy is defined as whether the results measure what they were supposed to measure (i.e. ancestry), thus, how often ancestry estimates are correct would be a measure of accuracy. However, this research works with skeletal teaching individuals from the University of Manitoba (U of M) and Brandon University (BU) whose ancestral origin is unknown meaning it is difficult to directly test accuracy. Therefore, the second purpose of this research, in relation to ancestry estimation, is to determine the effect differing scores have on ancestry estimations, and whether the results from morphometric ancestry assessment match the results from metric ancestry assessment. Since scores are used in a comparative analysis with reference populations for an ancestry estimate, two statistical programs, *HefneR* and *MaMD Analytical,* are used in this research to estimate ancestry with morphometric data. The reference populations for these programs were documented using the same scoring method for consistent scoring and a reliable comparison (Hefner, 2018). However, the effect of differing scores have on an ancestry estimate is unknown. If the scoring method produces repeatable results, then reference trait frequencies would be reliable and can be used in ancestry assessments. If it does not produce repeatable results, then the differing scores may affect the ancestry estimate. Therefore, comparing ancestry estimates from *MaMD Analytical* and *HefneR* between multiple observers and scoring periods can provide insight to this effect. Furthermore, comparing these results to a commonly used ancestry estimation program that uses metric traits, *Fordisc*, will provide further insight to how well

morphosopic traits estimate ancestry as compared to metric analysis. In the end, this would be an indirect test of accuracy as they should be measuring the same thing (ancestry).

The two components of Hefner's (2009; 2020; Hefner and Linde 2018) method lack extensive testing by different researchers. His scoring method has undergone several intra-observer tests with most indicating that agreement is high (Atkinson & Tallman, 2020; Coelho, Navega, Cunha, Ferreira, & Wasterlain, 2017; Dinkele, 2018; Hefner, 2009; Hurst, 2012; Kamnikar, Plemons, & Hefner, 2018; Kilroy, Tallman, & DiGangi, 2020; L'Abbé, Van Rooyen, Nawrocki, & Becker, 2011; Moffit, 2017; Wang, 2016). However, for a method to be reliable, inter-observer agreement must be high too. Only a handful of researchers studied inter-observer error and the score agreement is much lower than reported intra-observer agreement (Coelho et al., 2017; Dinkele, 2018; Hefner, 2009; Klales & Kenyhercz, 2015; L'Abbé et al., 2011). Moreover, not all traits in the scoring method were tested (Hurst, 2012; Kamnikar et al., 2018; Moffit, 2017; Wang, 2016), and some only tested ancestry estimate accuracy without testing or stating observer agreement (Go & Hefner, 2020; Hefner, Pilloud, Black, & Anderson, 2015; Kenyhercz, Klales, Rainwater, & Fredette, 2017; Monsalve & Hefner, 2016; Redfern et al., 2016). Researchers suggest focusing on sources of error causing disagreement and not just on error rates (Budowle et al., 2009). Only one published study has mentioned sources of error and provided recommendations for the improved description and scoring of two traits (Kamnikar et al., 2018). This research helps fill the gaps of knowledge for how well the method performs among multiple researchers and where sources of error could be occurring for traits not otherwise investigated.

The following are the hypotheses for the quantitative aspects of the research:

1. Previous research shows that intra-observer agreement, one observer scoring the same individuals across multiple scoring periods, is high ($k \geq 0.61$) when using Hefner's scoring

method; in other words, the scores agreed the majority of the time. Agreement values, kappa (k), are calculated by a statistic that determines the level of agreement for trait scores above that of the rate of chance. It is therefore proposed that the results of a repeatability test will show high agreement in scoring. The null hypothesis states that agreement occurs at a rate of chance (k=0), meaning the scoring method is not increasing agreement or causing disagreement within one observer, thus error in the method is random. The null will be rejected if intra-observer agreement is not occurring at a rate of chance.

2. Previous research shows that inter-observer agreement, the same individuals scored by multiple observers, is low (k<0.61) for the majority of traits when using this scoring method, and lower than intra-observer agreement. It is therefore proposed that the results of a repeatability test will show low agreement in scoring as compared to intra-observer agreement. The null hypothesis states that agreement is occurring at a rate of chance (k=0), similar to intra-observer agreement, meaning the scoring method is not increasing agreement or causing disagreement between observers. The null will be rejected if inter-observer agreement is not occurring at a rate of chance, and is lower than that of intra-observer agreement.
3. Since the ancestry estimates are based on the comparison of scores to reference populations, it is proposed that a disagreement of scores between or within observers will affect the ancestry estimates given to study individuals. The null hypothesis is that there will be no change in resulting ancestry estimates when scores are different. The null hypothesis will be rejected if there are changes to resulting ancestry estimates when scores are different.

4. Finally, if morphoscopic traits are accurately translated into a method of ancestry assessment, it can be expected that different methods and software programs used to assess ancestry will return the same result. Therefore, the null hypothesis states that programs using different data to measure the same thing, ancestry, should have no significant difference in ancestry estimate grouping. This is measured through agreement rates, and, therefore, the estimates should agree more than that of a rate of chance. The null hypothesis will be rejected if the rate of ancestry estimate agreement between software is equal to or less than that of chance.

This research is unique as it is the first study to compare the ancestry results between observers and scoring periods to determine if the estimates change upon score disagreement. There is only one study that briefly mentions looking at whether a change in score affects the ancestry estimate, and they concluded that there was "minimal" effect (Kamnikar et al., 2018). Furthermore, it the first study that tests *HefneR* and *MaMD Analytical* in comparison to *Fordisc* to see if their ancestry results agree. If two programs that are supposed to be measuring the same thing are not returning the same answer, this means there is still work to be done on either the scoring method or statistical software. This research is an important step towards the creation of an accurate method by assessing how it performs at several stages and what the impacts of different conditions are. Furthermore, this research contributes to the beta testing of *MaMD Analytical* and will be the first publication to do so in a comprehensive analysis.

It is important to have both a consistent system for trait assessment, as well as reliable reference population data from this system because it could increase the chance of a correct ancestry assessment and ultimately aid in identification. Method testing is not only important for improved accuracy, it is important when justifying the use of that method in court (Christensen,

2004; Christensen & Crowder, 2009; Craig, 2016; Grivas & Komar, 2008; Holobinko, 2012). If a case goes to court, the forensic anthropologist must provide evidence that their methods and the results are reliable and accurate (Craig, 2016). Consequently, Hefner's method (2009; 2020; Hefner and Linde 2018) must be thoroughly tested, have known accuracy and error rates, have established standards, and be widely accepted in the forensic anthropology community (Christensen & Crowder, 2009). Hefner (2009; 2020; Hefner and Linde 2018) has proposed the established standards, but, as of now, the accuracy and error rates are variable and are determined by very few studies. This research acts as an additional validation study of Hefner's (2009; Hefner and Linde, 2018) scoring method and statistical analysis (Hefner, 2020) for assigning ancestry. The results of this research will generate a more comprehensive understanding of his method's error rates. Furthermore, the results of this research will further facilitate the integration of his method as a widely accepted standard in the discipline.

Overall, this research addresses whether Hefner's (2009; Hefner and Linde 2018) scoring method is reliable among observers for all traits, but also if the disagreement in scores affects the resulting ancestry estimate. It reveals sources of error that other studies do not reveal, providing the opportunity for recommendations to reduce this error. Finally, it provides insight to how well different ancestry estimate programs perform in relation to each other, thus providing another discussion point for anthropologists for which traits and programs they should rely on.

Chapter 2: Literature review

2.1 Population Genetics

The theoretical foundation for ancestry estimation originates from the application of population biology and genetics to bioarchaeological research. Population biologists use tools called distance analyses to study dissimilarity (genetic or phenotypic) between multiple pairs of populations, the results of which can provide insight into a population's structure and history (Relethford, 2016). These analyses use gene frequencies to measure the genetic relatedness, or biological affinity, within and among populations (Pilloud & Hefner, 2016; Relethford, 2016), with the expectation that individuals within the same population will be more genetically similar. Since bioarchaeologists often do not work with DNA, they use phenotypic (visual) traits as proxies for genes, and the analyses of these traits are termed biological distances or "biodistances" in bioarchaeology (Relethford, 2016).

2.1.1 Additive gene effect and heritability

Phenotypic traits are used as proxies because genes code for different trait expressions, and these traits are inherited through simple or complex genetic mechanisms (Molnar, 2002). Simple mechanisms include the inheritance of dominant and recessive alleles resulting in one trait that is controlled by one gene. Complex mechanisms, and the most common among humans, result in polygenic traits, which are traits controlled by multiple genes (Molnar, 2002). Metric and non-metric skeletal traits used for ancestry estimations are polygenic traits. The support for using these skeletal traits as genetic proxies comes directly from genetic studies on skeletal trait expressions in mice (Grüneberg, 1952) and anthropologists studying cranial trait heritability in primates (Cheverud, 1982; Cheverud & Buikstra, 1981b, 1981a, 1981c). Support also comes

indirectly through studying frequency distributions of non-metric traits and historical migration data in human populations (Laughlin & Jorgensen, 1956).

Although all skeletal elements can be utilized for ancestry assessments, cranial elements are most often assessed because the craniofacial complex has been shown to be significantly heritable for both metric (Cheverud, 1982; Sherwood et al., 2008) and non-metric traits (Carson, 2006; Cheverud, 1982; Cheverud & Buikstra, 1981a, 1981c; Grüneberg, 1952; Ricaut et al., 2010). Heritability is the contribution of genetics to a trait's expression (Pink, Maier, Pilloud, & Hefner, 2016). Utilizing the theory that visible expression is controlled by multiple genes (Carson, 2006; Grüneberg, 1952), genetic control of phenotypic expression is established through the number and type of genes present along a continuum in a process called the "continuous additive effect" (Carson, 2006; Grüneberg, 1952). For example, if there are more genes present, then the trait expression is bigger (Carson, 2006; Grüneberg, 1952). Additionally, there are thresholds, or checkpoints, along the genetic continuum that indicate a certain number of genes are needed before a trait is expressed (Grüneberg, 1952). A continuum can result in quasi-continuous phenotypes, such as expressed by some morphoscopic traits, which are discrete categories of trait expression with each threshold that is passed (more genes= different expression) (Grüneberg, 1952). For example, rather than a trait increasing in size with each additional gene, the expression differs distinctly from the previous one (ex. absent, square, circle). Furthermore, quasi-continuous traits are inherited in groups due to gene linkage (genes inherited together) or pleiotropy (the same genes controlling multiple traits) (Cheverud & Buikstra, 1981b), which is helpful for determining how alike an individual is to a population. This is because different populations have different genes passed onto their offspring, and, therefore, a different number and combination of genes controlling trait expression. The genes that are inherited are also under the influence of

evolutionary forces such as natural selection, genetic drift, mutation, or gene flow (Molnar, 2002). Over time, the combination of these mechanisms cause specific trait expressions, especially those in the mid-face, to become more frequent in each population (Brettell, 2013; Brues, 1990; Gill, 1998; Klales & Kenyhercz, 2015; Moffit, 2017). This heritability supports the use of morphoscopic traits as visual proxies to genes in order to determine ancestry (Carson, 2006; Laughlin & Jorgensen, 1956),

Studies estimating heritability of non-metric traits tend not to include the traits used in current morphoscopic trait research, therefore, the specific heritability of these traits are unknown (Cheverud & Buikstra, 1981b). However, facial non-metric traits, in general, have shown significant differences in frequency distributions among populations (Cheverud & Buikstra, 1981b). Additionally, correlations between non-metric traits and genetic markers are significant and efficient at detecting larger similar groupings or outlier groups (Ricaut et al., 2010). Although the exact heritability mechanisms that the traits in the current research undergo is not known, the conclusions from these studies can indicate their likely genetic mechanisms.

Since additive genetics result in traits with graded expressions, a numerical scoring system can be used and is useful for tracking trait frequencies. Even though the additive gene mechanism can explain much of the phenotypic variation, there is a lack of understanding of the expression thresholds, the specific genes controlling non-metric traits (Carson, 2006), and the environmental affects on phenotype (Cheverud, 1982; Cheverud & Buikstra, 1981b; Pink et al., 2016; Smith, Hulsey, West, & Cabana, 2016). All these unknown factors may impact expression, in turn affecting the ability to detect and apply grading systems to skeletal traits that correlate with genetic thresholds.

Factors that may affect expression include the functional morphology of a trait. For example, a muscle attachment near or on the trait may be involved in mastication, therefore, the expression may increase in size with an increasing force of the muscle acting on the bone (Hauser & De Stefano, 1989; Hefner & Linde, 2018). Some researchers hypothesize that a trait's etiology, such as genes controlling a trait's developmental process, affects the final expression (Cheverud & Buikstra, 1981a, 1981b). These developmental processes can then be affected by nutrition or intrauterine environment, resulting in individual trait variation. Even after development, traits can be affected by environmental conditions related to disease or hormonal and dietary differences. For example, one study found that changes in morphological trait expression corresponded to advancements in medicine and dietary changes (Kilroy et al., 2020), thus introducing secular change. Secular changes affect trait frequencies which are used for similarity measures, therefore, modern reference populations are recommended as comparison (Spradley, 2014). These changes are seen in craniometric traits in the last 150 years, likely resulting from improved health and nutrition impacting growth and development (Ayers, Jantz, & Moore-Jansen, 1990; Wescott & Jantz, 2005). Similarly, Kilroy et al., (2020) saw secular changes in trait frequencies for seven morphoscopic cranial and mandibular traits in European Americans. Contrary to this research, there has been little to no secular change for one morphoscopic trait, zygomaticomaxillary suture, in some North American Indigenous populations, even post-contact (Maddux, Sporleder, & Burns, 2015).

Ancestry has also been studied in relationship with sex to show that these two aspects interact with contrasting results, much like secular changes. Sexual dimorphism is said to be less pronounced in non-metric traits as compared to metric traits (Corruccini, 1974), with researchers finding no sex differences in the zygomaticomaxillary suture (Maddux et al., 2015), the majority

of cranial morphoscopic traits (Atkinson & Tallman, 2020; Hauser & De Stefano, 1989), and in post cranial non-metric traits (Spiros, 2018). In contrast, some researchers found that ancestry negatively affected the accuracy of sex assessment (Lewis & Garvin, 2016), and that the combination of sex and ancestry for group assignment resulted in reduced accuracy (Murphy & Garvin, 2018). These are just a few examples of how little is known about how and what non-metric trait expressions are affected by internal and external environmental change, if at all.

While it is possible to theorize, it is unknown how and how much these environmental interactions alter trait expressions, which leave anthropologists in charge of establishing visual borders between expressions. These arbitrary borders result in problems with estimating heritability, especially when utilizing 'present' versus 'absent' scoring. Much of the time, bioarchaeologists use this binary system because non-metric traits were initially studied as skeletal anomalies in the form of additional foramina, sutures, ossicles, and Wormian bones (Dunn, Spiros, Kamnikar, Plemons, & Hefner, 2020). However, combining graded expression categories (ex. from four to two) can change the number of individuals considered to have a certain trait. Moreover, if it is unknown when a change in expression is parallel to the change in genetic thresholds, as the scores do not accurately reflect genetic frequencies. Both these instances can skew the resulting heritability score to being more or less heritable than it actually is (Carson, 2006). These borders can also create problems during the scoring session; many expressions show slight differences that may be interpreted differently between observers (observer error). If traits are not scored consistently, there may be problems in the final ancestry assessment. While Hefner (2009; Hefner and Linde, 2018) attempts to define these categories of cranial trait expressions in his scoring method, interpretation can still differ between anthropologists and result in various scores.

The way the trait expressions are determined by genetic mechanisms (additive or not) and the formatting of data by anthropologists (binary or graded) can affect how the data is used to determine biological relationships (Edgar & Ousley, 2016). In general, it seems these compounding effects vary by trait, population, and, possibly, by method of assessment. Regardless of these issues surrounding trait etiology and its impact on visual interpretation, studies using non-metric traits have demonstrated that they are effective for determining population affinity and are used to do so in forensic anthropology (Hefner, 2009; Hurst, 2012; Ricaut et al., 2010). Thus, method testing is important to understand the outcomes with these unknown factors.

2.1.2 Isolation by distance and Biodistance analyses

In order to use trait expressions as genetic proxies to estimate ancestry, distance analyses that have been modified to use phenotypic trait data can be employed (Pilloud & Hefner, 2016). These analyses are based on the genetic theory "isolation by distance," which states that populations that are geographically separated will become less genetically alike over time. Therefore, gene frequencies (or proxies) can be used to study genetic relatedness among populations.

Theoretically, a population's gene frequencies fluctuate around a unique equilibrium due to a net change from immigration, mutation, or selection pressures (Wright, 1943). These equilibriums differ between populations due to restricted gene flow and differing evolutionary pressures that gradually alter gene frequencies (L'Abbé et al., 2011; Wright, 1943). Each population has multiple pressures occurring simultaneously at different rates and intensity (Wright, 1943), which increase or decrease the number of trait-controlling genes present in a population. A gradual change in the number of genes can eventually shift which trait expression is more or less common because different thresholds along the genetic continuum are being reached in each

population (Carson, 2006; Grüneberg, 1952). Therefore, the frequency of each character (trait) expression in relation to each other varies among populations, resulting in unique equilibriums for temporally and geographically different populations (Corruccini, 1974; Edge & Rosenberg, 2015). These equilibriums are the basis for determining similarity of an individual to a population.

Importantly, researchers have provided evidence that neutral evolutionary forces likely shaped much of the human cranial morphology, which means there is no specific selection of traits to increase expression frequency (Smith et al., 2016). Neutral traits are those that give no adaptive benefit or detriment to the individual; therefore, they do not need to be selected for or against in any environment. For example, as some authors put it, there is no functional adaptation to having rounded orbital rims versus angular orbital rims (Brace & Hunt, 1990). This is important because if a certain trait expression was beneficial, then selection would favour that expression and reduce the presence of other expressions. Since directional selection is not driving the changes in gene frequencies for these traits, thus changing equilibriums, this means random gene flow and drift, such as through migration introducing new genes, have resulted in the current trait frequencies (Smith et al., 2016). These neutral traits are shown to have strong geographic patterning that reflect genetics and population histories, allowing for their use in studying genetic relatedness (Ousley, Jantz, & Hefner, 2018).

Population biology studies have demonstrated how limited dispersal results in patterns of isolation by distance (IBD) in just a few generations (Aguillon et al., 2017). The application of this theory suited bioarchaeologists well for understanding human migration because IBD can show local differentiation, thus reveal anthropogenic or geographic barriers to gene flow (Duforet-Frebourg & Blum, 2014). Patterns of IBD between pairs of populations can be detected by using different models (Duforet-Frebourg & Blum, 2014; Meirmans, 2012; Séré, Thévenon, Belem, &

De Meeûs, 2017) that utilize various distance measures, or, in other words, genetic distances. A genetic distance is a translation of the populations genetic "closeness" (Rousset, 1997; Séré et al., 2017), as determined through calculations of allele frequency in each population and the mean frequency over all the populations in the analysis (Relethford, 2016). Simply, a distance measure/genetic distance is the transformation of biological data into some thing tangible that can be compared (Dudzik & Kolatorowicz, 2016).

Each distance measure is calculated differently depending on the information available (genetic or phenotypic) and assumptions about the traits (Relethford, 2016). For example, the mean measure of divergence (MMD) "converts non-metric trait frequencies to a numerical value reflecting similarity" (Pink et al., 2016, p. 99). This measure was, and still is, a popular measure to explore biological relationships (Pink et al., 2016), such as studying whether social barriers caused some historical groups to become more similar or different (Edgar, 2009). It is especially useful for dichotomized non-metric traits in small sample sizes (Velasco, 2018).

Another distance measure that can be calculated is F_{st} (Konigsberg, 2006; Rousset, 1997). F_{st} measures inbreeding and is based on a population genetic model (Relethford, 2016). Inbreeding is used as a measure because it assumes that populations will mate within their own or geographically close populations rather than populations further away. This inbreeding reduces genetic diversity within a population. Conversely, inbreeding increases genetic differentiation with populations that are further away due to lack to gene flow, such as IBD suggests. Therefore, it is not a direct genetic distance, but a measure of genetic differentiation (Séré et al., 2017). If there is small F_{st}, then there is increased gene flow and less differences between populations. Meanwhile, a large F_{st} can indicate that there is increased differentiation (Pink et al., 2016). F_{st} has less power detecting IBD than other distance measures that directly utilize genetic data (Séré et al., 2017), but

it can still be used with indirect genetic data, such as non-metric traits, to give a rough statement about differentiation (Relethford, 2016).

There are also geographic distance measures, such as Mahalanobis distance (D^2). This is the preferred method for estimating biodistance in bioarchaeology because it can be used with non-metric traits to measure divergence between populations (Pink et al., 2016). It is an alteration of another geographic distance, the Euclidean distance (Séré et al., 2017), that takes into account the intercorrelation of traits (Relethford, 2016). This distance measure is an accumulation of similarities and differences that result in a value of how similar an individual is to a group. Mahalanobis distance can be used to find the phenotypic F_{st} (Konigsberg, 2006), and, consequently, find the minimum phenotypic distance between groups (Pink et al., 2016). This distance is often used in ancestry assessments to place an individual in a population whose distance is shortest between them.

Research has shown there is a correlation of genetic data and cranial traits (Smith et al., 2016), between geographic and genetic distances (Séré et al., 2017), and between genetic and phenotypic variance (Pink et al., 2016). If genetic distances are correlated with geographic distances, and the minimum phenotypic F_{st} is assumed to be proportional to the real F_{st} (genetic), then F_{st} can be calculated based on the phenotypic data (Pink et al., 2016). This F_{st} can be calculated from the methods that use phenotypic trait data to estimate distance measures (Pink et al., 2016). In fact, biodistance analyses using craniometric data and coordinate data, to capture phenotypic variation, conform to patterns of IBD (McKeown & Jantz, 2005).

In the end, genetic measures of distance can be applied to metric and non-metric traits (Konigsberg, 2006; Relethford, 2016). However, there is a fundamental disagreement on how human variation across geographic space arose. Some believe isolation by distance while others

say a more complex combination of processes (Edgar & Hunley, 2009). Regardless, IBD has shown to be useful in the study of biological relationships, thus it is useful for comparing similarity of individuals to populations during ancestry estimation.

2.1.3 Application of biodistance analyses

Bioarchaeologists use these models of isolation by distance to determine whether populations had gene flow, understand migration patterns, or understand relatedness within stratified cultures (Dudzik & Kolatorowicz, 2016; Lane & Sublett, 1972). These analyses can also be used to determine when evolutionary forces, such as directional selection or admixture, could have affected frequencies (Konigsberg, 1990). This is particularly important if the traits are not considered neutral because it is unlikely detectable neutral traits would follow the same pattern as selected traits (Ousley et al., 2018). Bioarchaeological research shows patterns of non-metric trait frequencies that vary by region, follow migration routes (Laughlin & Jorgensen, 1956), and are most similar in populations that are located within close temporal and geographic proximity to each other than those further away (Berry & Berry, 1967; Breske, 2018; Hunley, Cabana, & Long, 2016; Kaestle & Horsburgh, 2002; Konigsberg, 1990; Laughlin & Jorgensen, 1956; Relethford, 2004, 2016; Wright et al., 2018; Wright, 1943). These studies agree with results from biological studies on animals showing that limited dispersal contributes to IBD patterns of genetic differentiation (Aguillon et al., 2017). Other researchers, such as Konigsberg (1990) and Macchiarelli, Salvadei, and Bondioli (1995), saw that frequency or distribution changes were not significant across temporal groups in the same space. In contrast, Kilroy et al. (2020) noticed frequency changes in several cranial and mandibular morphometric traits, with most often mandibular traits having significant changes in expression. These studies demonstrate

that some traits may be more useful than others when it comes to determining group affiliation using geographic patterning of frequencies.

Instead of using biodistance analyses to understand relatedness between populations, forensic anthropologists use them to estimate the relatedness of an individual to one or more reference populations, allowing them to place the individual into the population they most likely belong (Pilloud & Hefner, 2016). In other words, to assign them to a population based on their similarity to it (Dudzik & Kolatorowicz, 2016). In fact, forensic anthropologists have tested various statistical methods to determine which one would work best to identify similarities between an individual and populations (Pietrusewsky, 2008; Pink, 2016). The most often used statistic is a linear discriminant function (LDFA), or variants of discriminant function analysis (DFA) (Cunha & Ubelaker, 2020; Geller & Stojanowksi, 2017; Kranioti, García-Donas, Can, & Ekizoglu, 2018; Pilloud et al., 2014), since its first use in forensic anthropology by Giles and Elliott (1962).

Discriminant function analysis is a multivariate statistical method that analyzes reference populations to find the variables that differentiate them the most; variables which are then used to create a discriminant function formula that will maximize differences between groups (Giles & Elliott, 1962). A Mahalanobis distance is usually the measure of biological distance where the closer an individual is to a group, the more similarities they possess to that group (Dudzik & Kolatorowicz, 2016). The formula can then be used to estimate the group affinity of an unknown individual, or place them in a population to which are most similar (Jantz & Ousley, 2012; Pietrusewsky, 2008; Pilloud & Hefner, 2016). The resolution for this similarity measure is dependent on the number of reference populations available for comparison because the fewer the reference groups, the fewer chances of detecting similarities (Hefner, 2018).

A commonly used statistical program that is based on the DFA method for estimating ancestry is *Fordisc 3.1* (Ousley & Jantz, 2012). *Fordisc* is a well-established software that is used to place an individual into a population based on the overall similarity of his/her cranial measurements to reference individuals (Ousley & Jantz, 2012). The classification into a group is partially dependent on having a large sample size. Theoretically, more populations, and individuals within the population, that are available for comparison should increase the accuracy of the estimation. This is because more population differences can be detected. Researchers have shown that the number of subpopulations used in a distance analyses does not affect the performance of detecting IBD, however, the number of individuals in a subpopulation will (Séré et al., 2017). Therefore, when collecting data for trait frequencies, it is important that there are many individuals, otherwise the results from a biodistance analysis may not be accurate.

Furthermore, a method can only identify individuals that belong to a reference population, therefore, many reference populations with different lineages need to be included for comparison (Armelagos & Gerven, 2003; Goodman, 1997). *Fordisc* allows a user to select whether or not an unknown individual is compared to modern or historic reference individuals from two databanks. Historic reference data is largely contributed by Howells who conducted significant studies of craniometric variation from thousands of individuals around the world (Howells, 1973; 1989 as cited in Sauer, Wankmiller, & Hefner, 2016). *Fordisc* also uses metric information from the Forensic Anthropology Data Bank (FDB), which holds morphological data for contemporary Americans (“Forensic Anthropology Databank,” n.d.; Ousley & Jantz, 2012). As a result, *Fordisc* includes a large modern American reference dataset and has become a standard for ancestry assessment. These characteristics make it a good choice for comparison to nonmetric analysis in this research.

2.1.4 Factors affecting biodistance analyses

While a large reference set is useful, a biodistance measure of similarity is also dependent on other factors, such as measurements being affected by cultural practices or environment (Relethford, 2016). A metric or nonmetric trait's vulnerability to damage can reduce the number of traits used for comparison (Relethford, 2016), and missing data can have an effect on the results (depending on the model used to calculate affinity) (Kenyhercz & Passalacqua, 2016). In one morphoscopic trait study, Kenyhercz & Passalacqua (2016) researched whether one could replace missing trait values with a predicted value to improve the analysis. They found that mid-face traits were the most correlated, thus, missing values from this area could be predicted more easily than those outside the mid-face (Kenyhercz & Passalacqua, 2016). Other factors such as unknown mechanisms of population differentiation may contribute to the rough measure of similarity. For example, physical barriers, such as mountains, also reduce gene flow, therefore, populations could be close in geographic distance, but genetically much different. The mechanism of differentiation should not drastically affect the outcome of estimates since population frequency differences are still present. Furthermore, globalization has allowed for the movement of many individuals from various populations over large distances, thereby increasing gene flow among populations of different geographic origins. This becomes an issue because there are individuals with multiple ancestries, making it difficult to place an individual into one population. All these factors in combination cause biodistance analyses to result in only a rough measure of similarity between individuals and populations (Kaestle & Horsburgh, 2002; Pink, 2016).

In response to the global migration issue, researchers have suggested that North America has social barriers that remain in place and occurred during globalization to keep individuals of different geographic origin separate (Edgar, 2009; Ousley, Jantz, & Freid, 2009). Therefore,

forensic anthropologists are still able to place an individual into a population from one of four major geographic regions (Edgar, 2009; Ousley et al., 2009). These barriers are related to the social labelling of "races." Someone's race is currently understood as an outcome of a culturally constructed ordering system (Dasgupta, 2011; Gravlee, 2009) that "utilizes aspects of morphology and culture to define racial groups" (Graves, 2010, p. 51). These barriers include systemic racism separating individuals considered to be of a separate race as well as assortive mating, which is when individuals are more likely to marry someone who is more similar to them culturally or morphologically, both of which can have effects on the genes passed to the next generation (Molnar, 2002). This social labelling system and history of "race" ideas must be discussed in relation to ancestry estimations to understand their effect on trait frequencies. Moreover, traits used in current ancestry estimations have their origin in historical concepts regarding race. These concepts continue to influence research surrounding non-metric traits and the practice of ancestry methods, making it all the more important to discuss the relationship between race and ancestry.

2.2 History of ancestry assessment in forensic anthropology

2.2.1 Historical ideas of group affinity

Scientific theories underlying ancestry assessments, such as IBD, are thoroughly studied and accepted as fact by current forensic anthropologists. However, ancestry assessment was not always understood as population differences through the theory of isolation by distance. Some forensic anthropological methods were not founded on a strong understanding of human variation, therefore, understanding the history of currently used ancestry methods is important. If these methods are not updated to fall in line with the current theoretical foundation underlying

human variation, the methods may result in inaccurate conclusions, thus rendering them invalid and unreliable.

The ancestry methods which forensic anthropologists use are rooted in early biological anthropology studies that were based on theories or ideas that are now considered invalid. One such theory was biological determinism which linked physical and behavioural traits with biological ones because many early practitioners thought that "race" was able to be determined from the skeleton (Sauer, 1992). However, the current definition of race means someone's race can change over time or space depending on the cultural lens applied (Edgar, 2009). In fact, there was debate over how many races existed, ranging from three to up to sixty because the amount of variation found within the human species made it difficult to delineate (Molnar, 2002). However, race is not the result of an ancestry assessment. Instead, forensic anthropologists use skeletal indicators of genetic information to estimate a populational lineage (ex. European). These skeletal indicators do not necessarily reflect the visual appearance or culture of an individual, such as thought by biological determinists. These historical views contribute to an ongoing debate on what race is and its relationship with ancestry methods used in forensic anthropology.

Though ancestry is presently assessed, this belief that race was detectable from the skeleton was a major influence on the studies of human variation and, consequently, the methods that forensic anthropologists currently use (Sauer et al., 2016). Studying the shape and form of the human cranium dates back to the early 18th century when naturalists were interested in systematically ordering the animal kingdom via taxonomy, including humans, to better understand biological variation (Dunn et al., 2020). At this time, race was generally thought to be biological, therefore, some thought this biology determined a number of attributes for each race. For example, Linnaeus described human "subspecies" (later called race) with both physical and

behavioural traits (Armelagos & Gerven, 2003; Caspari, 2009; Molnar, 2002; Ta'ala, 2014). These were considered permanent groups with no overlap in traits (Livingstone & Dobzhansky, 1962; Sauer, 1992; Smay & Armelagos, 2000). Physical traits that were included in each racial category were types of hair, skin colour, nose form, and skull shape (Molnar, 2002). Since the brain gave rise to behavioural traits, the skull shape and size was considered an indicator of brain form, therefore, tied physical traits with behavioural or social traits such as intelligence (Molnar, 2002). Overall, these descriptions of human subspecies contributed to the creation and reinforcement of racial types to try and explain human diversity (Molnar, 2002).

Two schools of thought came from the taxonomy of human groups which influenced the direction of an anthropologist's research in relation to human diversity: monogenists and polygenists. Monogenists thought that all humans, regardless of race, came from one common ancestor. Meanwhile polygenists thought that each race had a different "pure" ancestor or separate evolutionary histories (Ta'ala, 2014). Blumenbach, considered "the father of biological anthropology", subscribed to the idea of polygenism (Ta'ala, 2014). He studied human cranial variation to improve upon Linnaeus' description of human races, and came to the conclusion that there were five human groups that could explain biological variation: black, white, yellow, red, brown (Ta'ala, 2014). His research influenced Morton, the primary founder of biological anthropology in the USA, who collected human skulls for classification and cranial capacity studies (Ta'ala, 2014). He concluded that Blumenbach's five races should be further divided into a number of families (Ta'ala, 2014). Morton also attempted to rank these races from their cranial measurements (Armelagos & Gerven, 2003), linking morphological differences to levels of intelligence and superiority, much like anthropologists before him (Ousley et al., 2018). In fact, Morton had manipulated his results to fit his preconceived notions of race where white men were

superior and more intelligent, and, therefore, had a larger cranial capacity (DiGangi & Hefner, 2013; Dudzik & Kolatorowicz, 2016). Craniometrics were used to reinforce this idea that the larger the skull meant the smarter the race, (Dudzik & Kolatorowicz, 2016) resulting in the view that Africans were impassive and lazy (Caspari, 2009). This binding relationship between behavioural and morphological traits resulted in archaeologists studying cranial forms of ancient civilizations to make inferences about their racial affinity and behaviour (Molnar, 2002). Morton then influenced Broca's work on the brain who developed new techniques for craniometric analysis (Ta'ala, 2014). These analyses resulted in the characterization of a "pure" "negro" race and "American whites" as a distinct subgroup of white people (Ta'ala, 2014).

Since race was often thought of as a subspecies, the idea that there were "entrenched characteristics" of races, and, therefore, "hybridization" resulted in the reduction of "purity," carried on into the early 20th century (Todd, 1929). Hooton was a major contributor to biological anthropology in the USA by training many physical anthropologists as the discipline was becoming established (Caspari, 2009). He also felt that the range of human variation was due to the interbreeding of these "pure" races to create hybrid races (Caspari, 2009; Molnar, 2002). Hooton's work was largely influential to non-metric trait research and methods used in forensic anthropology, even if he did not contribute to forensic anthropology research himself (DiGangi & Hefner, 2013).

While many polygenists subscribed to biological determinism, monogenists also had similar thoughts. Broca's work inspired Hrdlicka, a monogenist who also contributed to American biological anthropology, to describe racial categories and collect craniometric data (Caspari, 2009; Ta'ala, 2014). Some scientists thought that people of colour were "living fossils" and represented earlier evolution stages, or that white people were the original humans and other races were

"degenerated forms" (Armelagos & Gerven, 2003). Therefore, some physical traits, such as broad cheeks, receding foreheads, or left-handedness, were considered remnants of an "ape" past (Molnar, 2002). These traits were then associated with criminals because apes were considered savages in relation to humans (Molnar, 2002). In agreement with the opinions of the time, non-metric traits were also used to describe human "types" (Wood-Jones, 1931c, 1931b, 1933). Overall, many people thought groups of humans (races) could be defined with a set of unchanging biological characteristics (ex. skin colour, intelligence) that made them distinct from another group (Armelagos & Gerven, 2003; DiGangi & Hefner, 2013; Hefner et al., 2012).

Contrary to these widespread beliefs, Boas thought that biological differences between races were small, opposing the use of typologies and challenging the idea of racial determinism (Caspari, 2009; Ta'ala, 2014). He argued that craniometric data was not reliable for determining racial categories (Ta'ala, 2014). Instead, his research was focused on finding the range of human variation and how it arose (Ousley et al., 2018). The results of his research indicated a non-concordance of "racial traits." This research contributed to understanding of geographic variation of traits along with the influence of the environment on this variation (Caspari, 2009).

While race as an indicator of biological human variation was not the only theory in the mid 19th to early 20th century, it was the prominent one that negatively influenced many aspects of society and academia (Caspari, 2009). Biological determinism was used to support social policy asserting that some human groups were inferior (Caspari, 2009). For example, Hrdlicka and Broca used their research to encourage eugenics (Caspari, 2009; Ta'ala, 2014) so undesirable traits from certain groups of individuals would not be inherited (DiGangi & Hefner, 2013). Meanwhile, Morton's work was used as justification for slavery and genocide (DiGangi & Hefner, 2013; Ousley et al., 2018; Ta'ala, 2014). The influence on academia meant that if

human variation was expressed in discrete categories (ex. white, black) (Todd, 1929), and these categories could explain how evolved a race was (Armelagos & Gerven, 2003; DiGangi & Hefner, 2013), then characteristics, such as non-metric traits, had more, or less, evolved expressions that could determine one's race (Armelagos & Gerven, 2003; Ousley et al., 2009; Smay & Armelagos, 2000; Wood-Jones, 1931a). This resulted in articles published in the 1920s and 1930s describing the physical and behavioural characteristics of each "race" (Sauer et al., 2016). For example, Wood-Jones considered individuals to be "pure" or have racial admixture (Wood-Jones, 1931a) and described non-metric traits associated with each race (Wood-Jones, 1931a, 1931b, 1931c, 1933). The idea of being able to determine racial affinity from the skeleton saw its eventual application in the forensic community (Giles & Elliott, 1962; Rhine, 1990).

2.2.2. The debate on race as a component of human variation

As the theoretical shift on how variation arose occurred in the late 20th century, the definition of race and its ability to be determined from the skeleton was and continues to be debated (Armelagos & Gerven, 2003; Gill, 1998; Goodman, 1997; Ousley, Jantz, & Freid, 2009; Sauer, 1992; Smay & Armelagos, 2000). Race definitions have been pointed out to vary in the literature (Newman, 1963, p. 200) because there are disagreements and misunderstandings of what race is and how it relates to ancestry, with some researchers conflating race with ancestry (Maier, Zhang, Manhein, & Li, 2015). The current ideas of how race is an indicator of biological human variation are on a continuum from "races do exist in biology" to "races do not exist in biology." These varying opinions result in inconsistencies for how ancestry is researched or how assessments are conducted.

2.2.2a "Race exists as a measure biological variation"

Race as a measure of biological variation is considered a concept of the Linnean system of classification (Gravlee, 2009; Livingstone & Dobzhansky, 1962; Sauer, 1992; Smay & Armelagos, 2000). This is not a popular view today, but it was heavily argued during the shift in what the theoretical foundation of human variation was thought to be. For example, Newman (1963) thought that biological traits must be re-evaluated in all populations to result in a "proper" understanding and definition of what race is. Race was initially defined as a subspecies (Armelagos & Gerven, 2003; Caspari, 2009; Molnar, 2002; Ta'ala, 2014) but during the theoretical shift, race was argued to be related to, or defined as, populations (Livingstone & Dobzhansky, 1962; Newman, 1963). Definitions ranged from "populations which differ in frequencies of some genes or gene" to " a collection of populations… having features in common, such as a high frequency for Blood group B, and extending over a geographically definable area" (Newman, 1963, p. 200).

Those in favour of this view began to agree that race is not discrete like early researchers thought. However, they adamantly asserted that races still existed, it was just unknown how many races should be labelled based on the extent of variation (Livingstone & Dobzhansky, 1962). The evidence for this view comes from worldwide studies on biological traits that show individuals from the same region reliably cluster together (Gravlee, 2009; Ousley et al., 2018), but it is outweighed by the overwhelming evidence for the view that "races do not exist as a measure of biological variation."

2.2.2b "Races do not exist as a measure of biological variation"

The opposing view that races do not exist in biology takes the stance that it is not possible to divide people into discrete units based on one or a few traits (Livingstone &

Dobzhansky, 1962). Instead, race is viewed as a social construct (Ousley et al., 2018). This view is accepted by the majority of anthropologists (Edgar & Hunley, 2009). In fact, the American Association of Biological Anthropologists released a statement in 1996 stating that genetically "pure" races have never existed, nor do they now. Many morphological differences are influenced by the environment; therefore, biological traits are not necessarily concordant with racial or ethnic groups. Therefore, human species cannot be classified into groups with clear boundaries (AABA, 1996).

The most often cited evidence for this argument comes from a study assessing genetic and blood group (proteins) variation in multiple groups of people; the author specifically added more nationalities within each geographic region in attempts for equality between "races" (Lewontin, 1972). The idea was, if races existed biologically, then the differences among groups of people would account for more of the variation than their similarities. In other words, the genetic diversity would be higher among regions if the total genetic diversity of humans is the sum of variation among geographic regions, among local populations within regions, and within local populations (Relethford, 2002). However, Lewontin found that 85% of the total diversity is found within local populations, approximately 8% among geographic regions, and only 6% among populations (Lewontin, 1972). This means that there is so much overlap between populations that only 6% of variation makes populations different, thus contributing to "racial classification" (Ousley et al., 2018). In other words, humans of different races are more similar than they are different.

Multiple researchers have replicated Lewontin's (1972) results through models used in Zoology to understand subspecies, which can be discretely grouped, as well as with other biological traits (Hunley et al., 2016; Relethford, 1994, 2004; Sauer et al., 2016). Using two

models, Sauer et al. (2016) found that human subspecies or races do not exist. One model stated that if there are subspecies then they should have discrete evolutionary lineages where members interbreed within their lineage more often than between lineages. Instead they found that humans interbred with members of their own species within immediate proximity rather than within their "race" (Sauer et al., 2016). The second model used F_{st} to understand genetic variation by stating that a subspecies F_{st} will exceed a minimum level of differentiation. However, human genetic variation did not exceed this minimum level. In fact, it fell far below it (Sauer et al., 2016). Similarly, F_{st} was used by Relethford (1994) to show that the degree of differentiation (F_{st}) among populations for craniometric data is roughly the same as the genetic differentiation shown by Lewontin (1972). Hunley et al. (2016) also concluded that "human populations are not genetically homogenous within," therefore, race has no genetic or taxonomic significance (Hunley et al., 2016). Using statistics to analyze the structure of genetic variation, the results showed that there was reduced genetic diversity as the distance from populations in Africa increased (Hunley et al., 2016). This supports Lewontin's (1972) results where the majority of the range of human variation is found within populations because there is traceable genetic origin. Importantly, the positive correlation between gene diversity and geographic distance (Hunley et al., 2016) further supports the theory of IBD resulting in the range of biological variation rather than separate "races."

The argument that races do not exist biologically also includes the statement that one cannot use visible traits associated with race to predict other aspects of biology, or vice versa, because traits are not concordant (Goodman, 1997; Gravlee, 2009). Moreover, most variation is clinal with no clear boundaries (Goodman, 1997; Gravlee, 2009). For example, Relethford (2002) studied craniometric data from Howells' databank (1989) with respect to skin colour, the

trait most often associated with race, to understand the pattern of variation. Relethford (2002) found that skin colour as a proxy for other biological traits was weak because it did not reflect genetic or craniometric patterns. In fact, it shows the opposite patterns with 88% of variation among regions (showing high differentiation). In other words, if categorizing race is based on skin colour, other biological data will not have the same pattern of variation (Relethford, 2002). Due to differences among populations only accounting for 5-10% of human variation, race is not considered adequate to describe all of human variation (Gravlee, 2009). Anthropologists conclude there is too much population overlap with one or a few traits to support race as a measure of biology.

Even when finding that there is strong agreement of morphology with social race in Americans, it does not validate biological race; testing European populations against each other will separate populations just as well (Ousley et al., 2009). This is because migration patterns and local gene flow can make distinct populations, but not distinct races (Ousley et al., 2018). Even worldwide variation for craniometric data, apparent from Howells data (1973, 1989 as cited in Sauer et al., 2016), does not prove typological races. Patterns do not mean populations can be naturally divided because not all individuals will fit in the cluster due to too much overlap (Gravlee, 2009; Ousley et al., 2018). Thus, the defence for "race exists because patterns of biological variation, such as clustering (Maier et al., 2015), are present" is not valid because researchers have shown that there is more in common between people of different "race" than there is different.

After all these studies and debates, Ousley et al. (2018) have decided there are four truths to variation: "evolution explains biological variation", "human variation is continuous", "variation involves many traits that usually vary independently", and "genetic variation within

"races" is greater than the variation among them". Therefore, the biological race concept is invalid to explain diversity.

2.2.2c "Race is more complex than "exists" or "does not exist" in biology"

Between the two extremes are various arguments that race is partly biological (Ousley et al., 2018). Some argue that races exist in biology, but do not necessarily include biological determinism as the main component (Ousley et al., 2018). Others argue that race "becomes" biology, where race is a social construct but with biological and social consequences (Gravlee, 2009). Therefore, one must recognize that race is socially important and should be included in research to identify these biological consequences (Gravlee, 2009). As for its use in forensic anthropology, some argue that race is good for small-scale, applied work even if it is known to be more complicated. In other words, it is a "convenient shorthand" to describe variation (Smay & Armelagos, 2000). This shorthand still assumes that individuals of a given race will be similar to others of the same race (Smay & Armelagos, 2000). Another argument is that "race is a necessary evil" where race is a real and social phenomenon, therefore, forensic anthropologists have to use it to replicate social race categories (Smay & Armelagos, 2000). However, questions arise on whether race actually helps or if it hinders the investigation (Smay & Armelagos, 2000). This is because one cannot tell if they are "correct" due to the subjectivity and ambiguity of race (Ousley et al., 2018).

Evidence for race being partly biological rests on the positive correlation of race with some morphological traits. Stull et al. (2014) used craniometrics to study the relationship with self or peer-reported race and demonstrated that there is a correspondence between morphology and peer-reported race. Moreover, Bulbeck (2011) concluded that race and geography have an effect on ancestry classification while using *Fordisc.* This was concluded because they could not

determine if Thai skulls were more similar to "Mongoloid" populations close to Thailand or "Mongoloid" populations in general. Researchers with this perspective assert that categorizing an individual into the correct geographical region with high accuracy using craniometric data does not mean it is a reflection of distinct racial groupings (Ta'ala, 2014) or a validation of biological race (Sauer, 1992). Though there appears to be correlation, it is unclear what the association is (Ousley et al., 2018).

The argument against this view is that people from the same region are more likely to share more genetic loci regardless of socially ascribed "race" (Ousley et al., 2018). Therefore, people within a large region may not look more similar to each other when using a social race category. This is because for any single trait/gene/allele/loci, the total variation within the population is greater rather than among populations (Lewontin, 1972; Relethford, 2002). Pooling multiple loci can provide information about population membership because of the differing evolutionary pressures on each individual locus in each population (Edge & Rosenberg, 2015). In other words, the covariation of traits can be used to group people because there is less overlap between them (Ousley et al., 2009, 2018). These multi-locus methods can detect patterns of variation, with some trait combinations predicting ancestry better than others (Edge & Rosenberg, 2015). Thus, population membership rather than racial affinity is estimated.

2.2.3 How and why forensic anthropologists use race

If there is so much evidence against biological race and agreement that it does not exist, then it would be assumed that race would no longer be a part of current practices (Smay & Armelagos, 2000). This view does not address why people see race and how it is related to ancestry. To this day, forensic anthropologists are commonly asked by law enforcement to determine an individual's "race" in addition to age-at-death, sex, and stature, because it provides

another visual cue for identification. Forensic anthropologists did not use genetics for group affinity research because they were more focused on providing law enforcement with the folk taxonomy labels they wanted (Ousley et al., 2018). While forensic anthropologists reject biological race, social race is real, thus social categories are used by some forensic anthropologists to help identify individuals (Ta'ala, 2014).

Biological traits that are geographically patterned (Irish, 2014; Relethford, 2002) can help forensic anthropologists speculate what an individual's social race label might have been. For example, skin color, usually associated with race, is systematically patterned across geographic space due to strong natural selection at certain latitudes (Relethford, 2002; Sauer, 1992). Though it shows an opposite pattern to other biological traits (Relethford, 2002), the covariation of traits show regional patterns as well that reflect genetic lineages (Relethford, 2002; Sauer, 1992). A likely skin colour can then be inferred based on the regional designation of an individual who was placed using these traits (Relethford, 2002; Sauer, 1992). This inference can facilitate identification, though, it must be used with caution as there is no consistent way to classify humans by race (Ousley et al., 2018). To solidify this point, Goodman (1997) recounted that babies described as Native American on their birth certificate were described as another race on their death certificate (Goodman, 1997).

Other reasons for using social labels include research showing statistically significant differences in craniometric morphology that can be visually appreciated between 'black' and 'white' Americans (Ousley et al., 2009). These differences are attributed to European and African populations being separated for thousands of years before migration to the USA (Ousley et al., 2009). Trait patterns in groups considered separate "races" would have been preserved from this previous geographic separation and through limited gene flow from social factors, such

as systemic racism through eugenics and Jim Crow laws on interracial marriage (Edgar, 2009; Ousley et al., 2009, 2018). Even after the abolishment of these laws, positive assortive mating (Molnar, 2002; Ousley et al., 2009) through religion or ethnicity, as shown by physical indicators (Edgar, 2009), continued to play a part in the separation of these lineages. While classification rates for racial groups is never 100% due to the overlap in variation, the "preserved" geographic pattern can produce high classification accuracy (Ousley et al., 2009)

The current research is based on the consensus that race is a social construct and not rooted in biology. Instead, ancestry is related to the inheritance of gene combinations within a population through complex mechanisms that result in morphological traits, such as those used in this research. The geographic pattern of traits can help infer social race, but this research only accounts for social race as it relates to the statistical programs using social labels.

2.2.4 How past race methods affect current application of ancestry

The history of race research and the debate on race has resulted in forensic anthropology research producing inconsistent and varied methods, and evidence of continued application of invalid methods (Edgar & Hunley, 2009; Gill, 1998; Hughes et al., 2011; Hurst, 2012; Lewontin, 1972; Ousley et al., 2009; Rhine, 1990; Sauer, 1992; Smay & Armelagos, 2000). Typology is a continued influence (Edgar, 2014; Hurst, 2012) with methods for differentiating black and white people popping up due to the apparent link between biological and social race (Ousley et al., 2018). Harris and Foster (2015) created dental metric formulas to discriminate between black and white individuals in the USA. Some researchers broadly apply a racial label to populations with multiple origins and histories instead of studying distinct populations (Edgar, 2013; Spradley, 2014). Edgar (2014) conflates race with ancestry, using these terms interchangeably. These inconsistencies make it difficult to have a standard practice among anthropologists.

At the time when biological determinists were at the forefront of biological anthropology, research on human variation did not occur through theory testing and hypotheses (Ousley et al., 2018). Instead, confirmation bias was often employed meaning that the results of their research are questionable (Ousley et al., 2018). For example, Hooton eyeballed morphological differences and placed people into groups, then used statistics to validate those groups rather than using statistics to find the differences (Ousley et al., 2018). Some anthropologists, such as Kretschmer, studied human form based on what he thought were the "most beautiful specimens" (Molnar, 2002). Goodman (1997) recounts how forensic anthropologists blamed the "melting pot" for any "misclassifications" even though there are no "pure" racial types that could be "blending" the traits. These preconceived ideas of what traits represented each racial group are the same traits that are used in current forensic anthropology research and practice.

These traits were used in typological methods that claimed high accuracies for grouping individuals, but rarely provided explicit instruction for replication (Ousley et al., 2018). One such method is the "trait list;" a list of trait expressions that could identify each race (Brues, 1990; Gill, 1998; Rhine, 1990). These lists devolved from Hooton's "Harvard List," which was an attempt to standardize recording non-metric traits (DiGangi & Hefner, 2013; Dunn et al., 2020). He created a list describing the traits along with recording forms for their scoring, but there were no instructions for how to use it to analyze race (Brues, 1990). Consequently, this list was used for a "trait list" approach to ancestry assessment. This method became prominent, and individual traits became associated with a single race, persisting though there was no consistency in application (Ousley et al., 2018). Moreover, trait lists were often used to support a conclusion *post hoc* (Brues, 1990; Kamnikar et al., 2018). This meant that rather than using trait lists to

assign biological race, the presence of traits were used to support the results of metric race assessment after an individual's race was confirmed (Brues, 1990; Kamnikar et al., 2018).

These trait lists are not considered reliable for several reasons. Some trait descriptions differ between major publications that forensic anthropologists look to for guidance (Bass, 2005; Rhine, 1990). These descriptions were interpreted differently for scoring and had drastically different results in trait distribution (Cataldo-Ramirez, Garvin, & Cabo, 2020). There are also no standard minimum observations needed to make a classification, resulting in inconsistent application (Hughes et al., 2011). The number and types of traits in a list are shown to bias the outcome toward a certain assessment (Hughes et al., 2011). For example, if there is a list that contains sixteen trait expressions and eleven are considered "Black", then it is biasing the estimate towards "Black" even if only five of those traits appear. Moreover, if more weight is given to some traits over others in decision-making, it can also bias the outcome (Hughes et al., 2011). Surprisingly, Hughes et al. (2011) concluded that the trait list method did relatively well regardless of the variations imposed on the method. This gives some validity to the trait's relationship with ancestry even if the method itself does not conform to the current theoretical foundation.

These lists were also not well studied for accuracy with respect to the relationship between an individual's skeletal traits and race (Hughes et al., 2011). When trait lists were used to estimate race, the individual's confirmed race often did not match the assessment (Brues, 1990; Kamnikar et al., 2018). Like most typological approaches, they do not take into account the range of variation in each population (Molnar, 2002). It was initially assumed that one trait would only be found in one race, however, traits have been found in high frequencies within populations that do not match the traditionally associated race (Hurst, 2012; L'Abbé et al., 2011; McDowell, Kenyhercz, & L'Abbé, 2015; Rhine, 1990). For example, a small nasal spine was

found in 59% of a “Caucasoid” sample when a large nasal spine was expected as the defining expression (Rhine, 1990, p. 14,16).

Since human variation is not discretely categorized, the presence of a trait expression cannot definitively determine one’s race as trait lists attest. Instead, human trait variation is continuous and can be studied using frequencies to detect population differences in most common trait combinations. These differences are important because an individual’s ancestry assessment depends on the similarity of their traits to a population. Unfortunately, past methods were often created with few populations and/or grouped geographically different populations together. This underrepresented the range of trait variation in a population or did not capture population differences in trait frequencies (Ayers et al., 1990; Birkby, 1966; Giles & Elliott, 1962; Gill, 1998; Gill & Gilbert, 1990; Rhine, 1990; Spradley, 2014)

While human variation is shown to be continuous, traits from these lists continue to be studied with respect to race (Hurst, 2012; Maier et al., 2015). For example, the palate is described to have three shapes and each shape is associated with a race (Clark, Guatelli-Steinberg, Hubbe, & Stout, 2016; Maier et al., 2015). However, Clark et al. (2016) showed that races did not match the shape, and that this method was inaccurate in both the description of the trait and the ability of an observer to score it. The accuracy of assessments using palate shape was as low as 40% (Maier et al., 2015) and 42% (Clark et al., 2016), which is consistent with research demonstrating that any single characteristic can be found across all populations regardless of the traditionally associated biological race (Hunley et al., 2016; Lewontin, 1972; Relethford, 1994).

The assumption that all individuals of one race shared the same traits meant that methods were also not tested for accuracy on other individuals before application. It has since been shown that a method created on one population cannot be used on another population even if they are

considered the same race because it results in low classification (Birkby, 1966; Fisher & Gill, 1990). This means many individuals in many populations must be studied for a method to be accurately applied across multiple populations. Similarily, analyses created on different temporal populations than the study population cannot be used accurately (Ayers et al., 1990). The need for more, and appropriate, populations within reference databases is recognized by multiple researchers across multiple types of data (Geller & Stojanowksi, 2017; Kindschuh, Dupras, & Cowgill, 2012; Pilloud, Maier, Scott, Hefner, & Scott, 2018; Spiros & Hefner, 2020). This is because the estimation of ancestry depends on these reference populations; results of estimation can change based on the populations to which they are compared to (Geller & Stojanowksi, 2017). The continued study of traits in two racial groups excludes the ability to place an individual into the correct population.

Traits in Hefner's (2009; Hefner and Linde, 2018) scoring method were derived from trait lists, such as Rhine (1990), and race studies (Wood-Jones, 1931a, 1931c, 1931b, 1933). However, Hefner's (2009; 2020; Hefner and Linde, 2018) method, does not associate a single race to each expression due to the current understanding of human variation. Without an associated race, his method avoids confirmation and *post hoc* bias because population-specific information is needed before a conclusion can be made. Furthermore, he includes a clear set of standards for how to assess the traits (Hefner, 2009; Hefner and Linde, 2018) and how to use the trait assessment for ancestry estimation (Hefner, 2018, 2020). To include multiple lineages in his method, numerous researchers have used Hefner's (2009, 2018; Hefner and Linde 2018) scoring method to document archaeological and modern population trait frequencies from 7,397 individuals across forty geographic regions; these data are included in a reference population databank called the Macromorphoscopic Databank (MaMD) that is available upon request

(Coelho et al., 2017; Dinkele, 2018; Go & Hefner, 2020; Hefner et al., 2015; L'Abbé et al., 2011; Moffit, 2017). These changes are attempts to move away from race assessment to ancestry assessment that has scientific foundation.

The movement away from the influence of past methods is an ongoing process. Methods are constantly revised or refined as more research and knowledge accumulate, resulting in a new understanding of an old problem. Ancestry estimation methods are no exception, and have probably undergone the most drastic change in foundational knowledge compared to the other biological profile variables (sex, age). Overall, ancestry methods gradually changed from assigning race (ex. white) to assigning ancestry (ex. European, African), but the component of race still appears. These accuracy issues and inconsistencies in forensic anthropological methods are a problem when the methods need to be justified in court. Consequently, ancestry methods have been called into question by anthropologists as this massive shift in theoretical orientation occurred.

2.3 Overview of ancestry estimation methods

There is a large number of ancestry assessment methods available for forensic anthropologists spanning the entire skeleton. Each method struggles in its own way to overcome the influence of historical processes or thinking. This section will review the variety of ancestry assessment methods and where the current research sits for each of them.

2.3.1 Metric Analysis

2.3.1a Cranial

Cranial metric analysis is most widely used for current ancestry estimations and uses a standard set of cranial measurements (Buikstra & Ubelaker, 1994; Langley, Jantz, Ousley, Jantz, & Milner, 2016a). Traditionally, cranial measurements are taken by placing sliding or spreading

caliper tips at predetermined points (landmarks) on the skull. Alternatively, a recent technology, the *Microscribe 3D-digitizer*, allows researchers to collect x,y,z landmark coordinates from its probe tip and transfer these to a computer (McDowell et al., 2015; Stull et al., 2014). These landmarks are used to take three types of measurements: box, sutural, and extreme curvature (Dudzik & Kolatorowicz, 2016). Box measurements are taken at maximum distances on the skull, such as the maximum distance between two surfaces. Sutural measurements are distances between a suture and another landmark, and, lastly, extreme curvature measurements are taken at the maximum point on a curve in relation to another landmark (Dudzik & Kolatorowicz, 2016).

Like nonmetric traits, an unknown individual's cranial measurements are compared to measurements of other individuals in various reference populations to determine similarity. These analyses result in probabilities that an individual will belong to each group, and can be completed through multiple software programs such as *3D-ID, 3Skull, AncesTrees, CRANID, Fordisc 3.1, (hu)MANid, SkullProfiler, and Locate LAMbDA* (Dunn et al., 2020). While all are intended to estimate ancestry, they vary in the type of statistical analysis, as well as the number and demographics of references groups (Dunn et al., 2020). For example, *CRANID* uses a clustering method called K-nearest neighbour (Dudzik & Kolatorowicz, 2016), *Fordisc* uses discriminant function analysis (Jantz & Ousley, 2012; Ousley & Jantz, 2012), and *AncesTree* uses randomized decision trees (Navega et al., 2015). In relation to reference populations, *AncesTree* has 3,000 individuals (Navega et al., 2015), while *Fordisc* has almost 6,000. *COLIPR* has reference data from modern populations in Prague, Lisbon, and Portugal, therefore, is not likely used in North America (Kranioti et al., 2018). Similarly, *CRANID* is more popular in Europe and Australia, while the program primarily used in North America is *Fordisc;* this is because the reference populations represent the main demographics in each geographic region (Dunn et al., 2020).

Fordisc is the main focus for craniometrics in this research because the results from this method are compared to the results from Hefner's (2009, 2020; Hefner and Linde, 2018) method, assessed through the programs *HefneR* and *MaMD Analytical*. *Fordisc* uses a large number of reference individuals from two databases, resulting in almost 6,000 references from both modern and historical populations (Howell, 1996; Jantz & Ousley, 2012). These reference individuals come from the Forensic Databank (FDB) or Howells Global Craniometric Dataset (Auerbach, n.d.). The FDB holds demographic and skeletal information (both metric and non-metric cranial and post-cranial) from the Terry Collection (born after 1900), and approximately 3,400 modern American individuals from forensic cases with 2,400 having known sex and ancestry ("Forensic Anthropology Databank," n.d.; Ousley & Jantz, 2012). Meanwhile, Howells' data contains 2,504 19th and 20th century individuals from twenty-eight populations worldwide (Auerbach, n.d.; Howell, 1996). These comparative reference populations can be chosen prior to analysis; therefore, anthropologists can use contextual information to make a judgement on which populations to in include or exclude from the analysis. For example, Howells' data is recommended for use if it is suspected the skull originates from an individual with recent populational origins outside of the USA (Ousley & Jantz, 2012). This research uses both databases due to the unknown origin of the sample population.

Fordisc requires the input of values from a list of thirty seven standardized measurements (Jantz & Ousley, 2012; Ousley & Jantz, 2012). The output is "typicality" and "posterior" probabilities for each population that the individual is compared to. A typicality probability (Typ P) shows the percentage of reference individuals within a population that are just as far or further away from the population's average as the unknown individual. This means if there are only a few references individuals who are just as far away from the mean of their own reference population

as the unknown individual, then the probability of belonging to that group is small (Ousley & Jantz, 2012). A posterior probability (PP) is the probability that an individual is member of that population, assuming that the individual came from one reference populations (Ousley & Jantz, 2012; Pietrusewsky, 2008). This probability is based on the average similarities within the population (Ousley & Jantz, 2012; Pietrusewsky, 2008), or, the probability an event will occur based on the number of times it did not occur in the reference populations (Edgar, 2005). Consequently, if the number is high, then the individual has a high probability of belonging to that group because there is a lot of similarity. When interpreting these results an anthropologist will state that the unknown individual is most similar to 'X' reference population (Ousley & Jantz, 2012).

One useful trait of *Fordisc* is that it can still run an analysis despite not having all the measurements (Jantz & Ousley, 2012; Ousley & Jantz, 2012). However, the fewer measurements there are will adversely affect the probabilities associated with each reference population. As the number of measurements decreases, there is a drastic decline of the posterior probabilities and increase of the typicality probabilities, showing that the unknown individual could have come from any population (Bulbeck, 2011). This is because the software cannot detect a pattern within the set of measurements which would have placed the individual closer to one population compared to the others. Therefore, an individual becomes more similar to a greater number of populations as the measurements are fewer. One researcher has determined that sixteen measurements allow for the most reliable result where any more will not provide a better classification, and any less would not provide reliable results (Bulbeck, 2011).

2.3.1b Dental

While teeth have been used to study population differences in bioarchaeology, it has not been adopted by forensic anthropologists as a standard ancestry estimation practice (Pilloud & Kenyhercz, 2016) and is rarely utilized in forensic contexts (Dunn et al., 2020). However, this does not mean dental metric analysis is not able to be used for ancestry estimates. Tooth measurements have been described and standardized since at least the 1950s (Dunn et al., 2020; Moorrees & Reed, 1964). These standard measurements include the maximum crown diameters, in both the mesiodistal and buccolingual planes, and height for each tooth (Moorrees & Reed, 1964). There are also alternative measurements proposed, such as maximums taken on the diagonal (Hillson, FitzGerald, & Flinn, 2005). When teeth are damaged, such as through tooth wear or caries, it is suggested to take measurements at the cervix (Dunn et al., 2020). This is because they can be used as proxies due to their significant correlation with crown size (Hillson et al., 2005).

Using dental metrics, Hanihara and Ishida (2005) have shown that there are global patterns of variation that are useful for assessing relationships among modern human populations. Using these global patterns and DFA, researchers have been able to show that dental metrics could discriminate between populations in three major geographic regions (Asia, Africa, Europe), and classify individuals into these groups with high accuracy (Pilloud et al., 2014). Since these measurements are already standardized and have been well studied between populations, these researchers hope that dental metrics will progress to a similar level as *Fordisc*, with the database already well under way (Pilloud et al., 2014).

Other research has utilized molar size and shape through geometric morphometrics to discriminate between populations because tooth size has been shown to differ across geographically different populations (Kenyhercz et al., 2014). Newer dental metric methods for

ancestry estimation, as discussed by Dunn et al. (2020), include enamel and dentin thickness. These methods still require more research before their utilization.

2.3.1c Post-Cranial

Post-cranial elements are rarely utilized for ancestry estimations (Dunn et al., 2020). Their application to forensic cases arose from documenting post-cranial morphological differences between animal groups, particularly in relation to locomotion, (Dunn et al., 2020). Due to this extensive study on locomotion, there are sets of standardized measurements described for many post-cranial elements, including those used for stature estimation (Buikstra & Ubelaker, 1994; Langley et al., 2016a). These standard measures allow for comparable data sets if post-cranial ancestry estimation becomes more popular. In fact, *Fordisc* allows a user to input measurements for all long bones, scapulae, sacrum, os coxae, and calcanei (Jantz & Ousley, 2012). However, it lacks the reference samples for them to be useful outside of American Black and White groups (Jantz & Ousley, 2012).

Of all post cranial elements, the femur has become the most studied due to its relationship with locomotion (Sauer et al., 2016). Unsurprisingly, researchers have found femora significantly differ in shape between human populations (Shirley, Fatah, & Mahfouz, 2014; Wescott, 2005) and significantly more so than tibiae (Shirley et al., 2014). However, many femoral differences are not entirely genetic and have a lot of known environmental influences, such as biomechanical stress contributing to these differences (Tallman & Winburn, 2015). This has resulted in particular parts of the femur being more indicative of ancestry than others (Tallman & Winburn, 2015). Wescott (2005) validated the use of multiple proximal femur measurements to distinguish between five American groups, whereas individual measurements, such as femoral neck axis length

(Christensen et al., 2014) or sub-trochantric measurements (Tallman & Winburn, 2015), can also discriminate between major geographic populations.

Other elements that have been studied in regard to their ability to discriminate between populations include vertebrae (Marino, 1997; Ünlütürk, 2017) radii, ulnae (Spradley, 2014), the hyoid (Kindschuh et al., 2012), and the hand (Smith, 1996) and foot (Smith, 1997) bones. Some elements have only a few measurements that are significantly different between populations (Kindschuh et al., 2012), and some researchers have tried a combination of lower leg and hip measurements to improve accuracy (Shirley et al., 2014).

Similar to cranial metrics, most post-cranial metric analyses use DFA (Kindschuh et al., 2012; Shirley et al., 2014; Smith, 1996, 1997; Spradley, 2014; Ünlütürk, 2017) with some researchers having success with mid to high accuracy rates (Kindschuh et al., 2012; Ünlütürk, 2017). Elements, measurements, and combinations of measurements have been tested to determine the best measurements to use with multiple elements, such as the hyoid (Kindschuh et al., 2012; Shirley et al., 2014; Ünlütürk, 2017). However, post-cranial research is only just seeing the transition from race estimation to ancestry estimation, unlike cranial metrics and non-metrics that are further along. Studies usually focus on discrimination between White and Black Americans (Spradley, 2014) or South Africans (Ünlütürk, 2017), and some include Hispanics (Spradley, 2014), which is a complicated category due to people with many geographic origins and population histories being grouped together.

2.3.2 Non-metric analysis

While metric analyses have provided high accuracy rates for estimating ancestry, they are often accompanied by non-metric analyses because identification is more effective with multiple lines of evidence (Holobinko, 2012). This is because some research has shown that shape is more

useful than size when it comes to differentiating between individuals of different ancestry (McDowell et al., 2015). Non-metrics analyses encompass the entirety of the surface contour between points of measurement to take the whole shape into account unlike metric analyses which do not take such a fine approach to bone shape. Therefore, using non-metrics in addition to metrics can provide a more complete picture. If multiple methods provide similar conclusions, then ancestry estimation is more likely to be accurate.

2.3.2a Cranial

Non-metric Epigenetic vs. Morphoscopic

The list of cranial non-metric traits is vast (Brues, 1990; Gill, 1998; Gill & Rhine, 1990; Hauser & De Stefano, 1989; Wood-Jones, 1931a, 1931b, 1931c, 1933), and many researchers focus on studying one (Cataldo-Ramirez et al., 2020; Clark et al., 2016; Maddux et al., 2015; Maier et al., 2015; McDowell, L'Abbé, & Kenyhercz, 2012; Sholts & Wärmländer, 2012) or a few (McDowell et al., 2015) traits in their ability to provide ancestry estimates. While useful for understanding whether these traits vary between population, using only one or a few traits for ancestry assessments are not useful because most of the total human variation is accounted for within populations (Lewontin, 1972; Relethford, 2002).

More recently there has been a move to distinguish between non-metric "epigenetic" and "morphoscopic" traits, particularly to differentiate between the traits often used by bioarchaeologists versus forensic anthropologists. Non-metric epigenetic traits are described as discontinuous traits that are able to be scored as present or absent rather than quantified by a measurement, and are often used by bioarchaeologists (Hefner et al., 2012). These traits fall into five categories: Extra-sutural bones, proliferative ossifications, ossification failure, suture variation, and foramina variation (Hefner et al., 2012). These are often used to provide a measure

for relatedness through biodistance analyses, or as genetic proxies to identify familial relationships (Hefner et al., 2012). They have, however, been determined to be unsuitable for ancestry estimation because they fail to reliably place individuals into ancestral groups (Pink, 2016). This is likely due to the collapsed nature of the scoring to present versus absent. It does not take into account that additive genetics can create a range of 'present' expressions occurring in varying frequencies among populations, such as with morphoscopic traits. Morphoscopic traits also fall into five classes: bone shape, bony feature morphology, suture shape, feature prominence/protrusion, and a very few as present/absent data (Hefner, 2009; Hefner & Linde, 2018). These traits are the most often used by forensic anthropologists to estimate ancestry (Hefner et al., 2012).

Morphoscopic traits

Only the traits used in this research (Table 1) will be discussed due to the overwhelming number of non-metric traits. The majority of the traits are found in the mid-face, however, there are a few traits found on the palate and lateral portion of the skull. These traits can be visually assessed with recommendations for tools to assist with some trait's assessment, such as a contour gauge. Some researchers have even tried to quantify non-metric traits

Table 1: Summary of morphoscopic traits, their abbreviations, and the numerical score range which correspond to varying trait expressions. Traits without * are ordinal variables, * are nominal, and ** are binomial. Adapted from (Hefner, 2018).

Trait	Abbreviation	Score Range
Anterior Nasal Spine	**ANS**	1-3
Inferior Nasal Aperture	**INA**	1-5
Interorbital Breadth	**IOB**	1-3
Malar Tubercle	**MT**	0-3
Nasal Aperture Width	**NAW**	1-3
Nasal Bone Contour	**NBC**	0-4
Nasal Overgrowth**	**NO**	0-1
Postbregmatic Depression**	**PBD**	0-1
Supranasal Suture*	**SPS**	0-2
Transverse Palatine Suture*	**TPS**	1-4
Zygomaticomaxillary Suture*	**ZS**	0-2
Nasal Aperture Shape	**NAS**	1-3
Nasal Bone Shape*	**NBS**	1-4
Nasofrontal Suture*	**NFS**	1-4
Orbital Shape*	**OBS**	1-3
Posterior Zygomatic Tubercle	**PZT**	0-3

(Clark et al., 2016; Maddux et al., 2015; McDowell et al., 2012), such as using landmarks to identify the expressions (Maddux et al., 2015), in order to provide more consistent scoring. Similarly, some use geometric morphometrics to try and improve replication for scoring (Clark et al., 2016). Following is an example of how the traits are defined and scored. In this research, the definitions for each trait were taken from Hefner (2009) and/or *Osteoware* (2020), with descriptions for how to score the trait following Hefner (2009) and/or *Osteoware* (2020). Hefner (2009) only includes descriptions for eleven of the sixteen traits, whereas *Osteoware* (2020) hold descriptions of the other five traits between the original publication, Hefner (2009), and the latest publication, Hefner and Linde (2018). Updated descriptions from Hefner and Linde (2018) are also included in this study because observer 1 used these as supplemental training for the method. Any description that is cited with both Hefner (2009) and Hefner and Linde (2018) is because the description did not change between the two publications.

Anterior Nasal Spine

Description

No description is found in Hefner (2009), however, a description is found in Hefner and Linde (2018, p. 13): “Located at the inferior border of the nasal aperture, just anterior to the floor of the nasal cavity.”

How to score

No instructions for how to score found in *Hefner* (2009) or *Osteoware* (2020). However, Hefner and Linde (2018, p. 13) describes: “To view this trait, cranium should be viewed anteriorly and laterally to assess the degree of projection, do not score for edentulous individuals.”

Trait expressions (Figure 1)

The ANS is scored progressively as slight, intermediate, and marked:

1. *Slight: has minimal to no projection of ANS beyond the INA.*
2. *Intermediate: has moderate projection of ANS beyond the INA.*
3. *Marked: has pronounced projection of ANS beyond the INA.* (Hefner, 2009, p. 987; Hefner and Linde, 2018)

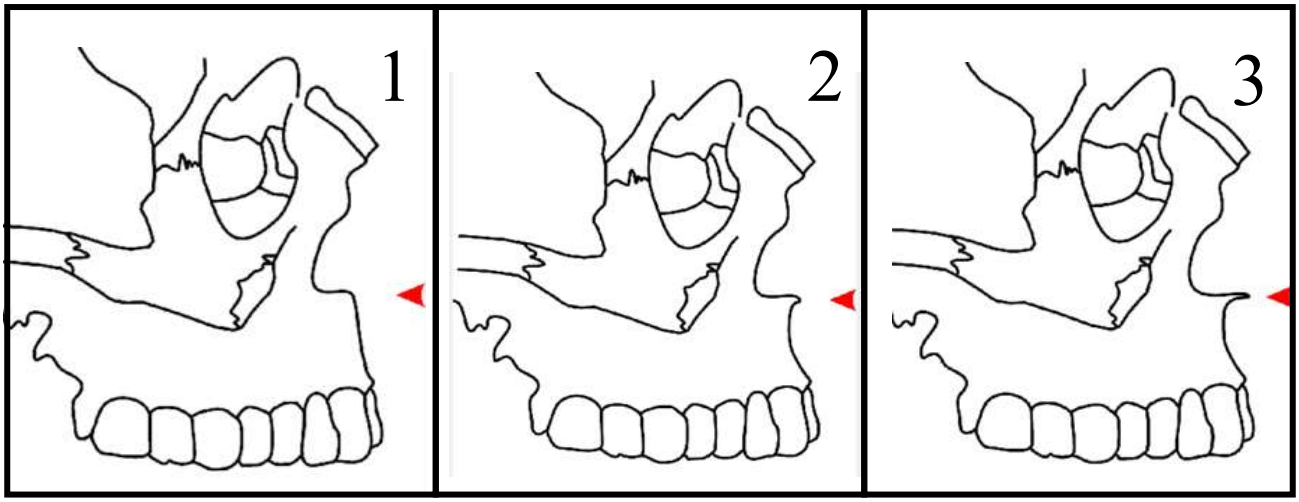

Figure 1: Trait expressions for ANS. Scores given to each expression are indicated as the values 1-3. Red arrows showing where this trait is located. Lateral view of the skull. Images taken from *Osteoware* (2020)

Multiple authors have tested different statistical methods to find the most appropriate one for estimating ancestry from morphoscopic traits (Coelho et al., 2017; Go & Hefner, 2020; Hefner, 2009; Hefner & Ousley, 2014; Hefner et al., 2012, 2015; Hurst, 2012; Kenyhercz et al., 2017; Monsalve & Hefner, 2016; Ousley, 2016; Pink, 2016; Pink et al., 2016). Currently, ancestry assessments using morphoscopic traits can be done using software such as *Combo MaMD Analytical* (Spiros & Hefner, 2019, 2020), *HefneR* (Coelho & Navega, n.d.), *MaMD analytical* (Hefner, 2020; Hefner & Byrnes, 2019), a Naïve Bayesian Classifier (Hermann, n.d.), and OSSA (Hefner, 2019). *Combo MaMD Analytical* combines both cranial and postcranial morphoscopic traits for analysis, and *MaMD Analytical* solely uses the cranial morphoscopic traits. Instead of the

traditionally used discriminant function analysis, an artificial neural network (aNN) is used for both of these programs. An aNN works as a feedback loop that gets better at classification as more reference individuals are included (Spiros & Hefner, 2020). However, it is not yet validated for practitioner use as it is in beta testing. Instead, the easier to use statistical method is the online application *HefneR* that uses the eleven morphoscopic traits from Hefner (2009) to assign an individual to one of four major regional populations. *MaMD Analytical* and *HefneR* both use reference groups from MaM DataBank (Hefner, 2018), which holds 7,397 individuals from archaeological and modern populations across forty geographic regions. In this research these two programs are used for comparison with *Fordisc.*

2.3.2b Dental traits

As with dental metrics, dental non-metrics are not often used and do not have an established method for ancestry estimation. Assessment involves observing structures such as grooves, ridges, and cusps for particular traits in the crown and roots (Edgar, 2013). These traits have been so widely studied that there is an established standard for scoring and recording each trait termed ASUDAS (Edgar, 2013, 2017; Pilloud, Adams, & Hefner, 2019; Turner II et al., 1991). The method involves a series of plaques that visually depict variants of crown and root features accompanied by a designated numerical score for each variant (Edgar, 2013, 2017; Pilloud et al., 2019; Turner II et al., 1991). This system allowed for consistent scoring and recording of traits across multiple studies, resulting in the establishment of global trait frequencies. These frequencies have been used successfully in understanding biological relationships of archaeological populations (Edgar, 2017), thus allowing inferences to be made about the migration of people into America (Turner, 1986). It is these frequencies that would allow their forensic application in the estimation of ancestry (Dunn et al., 2020).

Multiple methods have been proposed for using ASUDAS scores in ancestry estimation (Edgar, 2005, 2013; Irish, 2014), however, some still focus on two racial groups within the USA which are not helpful for other populations (Edgar, 2005, 2013, 2014). Departing from the racial groups, Irish (2014) developed a "summed score" method using ten traits to discriminate between five populations, European, Chinese, Indigenous American, Polynesian, and sub-Saharan African. Irish (2014) used traits that were the least ambiguous to score but varied the most between populations in order to reduce trait combination overlap. The method requires that each trait on an unknown individual be assigned the percentage value corresponding to the reference population with the highest frequency of that trait expression (Irish, 2014). These frequencies are summed to determine the ancestry estimate. For example, if 150% of traits are African and 8% are European, then the ancestry is African.

The most recent proposed method is a software program called *rASUDAS*, but it is still undergoing development (Scott, Pilloud, et al., 2018). It is proposed to complement other ancestry methods as it can accurately classify individuals into groups using 'three regional group' and 'four regional group' models (Scott, Pilloud, et al., 2018). The reference sample is currently based on 30,000 individuals from seven geographic regions and twenty one traits (Scott, Turner, Townsend, & Martinon-Torres, 2018) with future research proposed to collect more information from modern USA individuals (Scott, Pilloud, et al., 2018). These dental methods are promising, but only cranial morphoscopic traits are used in this research.

2.3.2c Post-cranial

Post-cranial elements have not been studied as extensively as cranial elements in relation to ancestry due to the bias of historical research focusing on race in the cranium (Armelagos & Gerven, 2003; Caspari, 2009). Consequently, they were not regularly used for ancestry estimations

as the transition from race to ancestry estimation occurred. However, researchers have shown that some post cranial traits, such as bifidity on cervical vertebrae, also have different frequencies among racial groups in the USA, allowing for high classification accuracy (Duray, Morter, & Smith, 1999). There is currently one set of postcranial traits that have been described specifically for ancestry assessment (Spiros, 2018). Spiros (2018) synthesized the literature on post-cranial research and chose eleven traits to standardize. Four traits were found in the cervical vertebrae, two on the humerus, and one each on the scapula, sternum, femur, patella, and calcaneus. Spiros (2018) defined each trait with a line drawing accompanying the description. Each trait expression was given a numerical score much like Hefner (2009) did with non-metric cranial traits. In a subsequent study, Spiros and Hefner (2020) found that postcranial traits performed very well against cranial traits in Black and White American reference pops, and encourage their use in ancestry estimations.

As expected, there are fewer assessment software programs for post-cranial non-metric traits than cranial traits, and only *Combo MaMD Analytical* is available (Spiros & Hefner, 2020). This application combines both cranial and post cranial traits for assessment, boasting a 15% increased accuracy than just using cranial traits (Spiros & Hefner, 2020). Much like *MaMD Analytical,* it is an aNN and is yet to be used regularly in forensic anthropological casework.

2.3.2d. Recent technology

Recently, new technology has allowed non-metric traits to be analyzed as 3D models to show morphological variation among groups (Carayon et al., 2018; Maddux et al., 2015; Rubin & DeLeon, 2017; Spiros & Hefner, 2020). Geometric morphometrics (GMM) allows for the analysis of both metric and non-metric traits in a new and comprehensive way that traditional techniques could not achieve. It allows someone to find areas of high differentiation in surface features

between groups, or can measure difficult to reach areas (Shirley et al., 2014). GMM has outperformed traditional craniometric analyses (Stull et al., 2014), and has greater scoring consistency than visual analysis, as exemplified in research studying the zygomaticomaxillary suture (Maddux et al., 2015). In addition, it allowed greater success in using the zygomaticomaxillary suture to differentiate European and American Indigenous groups (Sholts & Wärmländer, 2012). This technology been shown to complement as well as point out flaws in the ASUDAS method, such as it does not take into account the overall shape of traits (Carayon et al., 2018). While useful, GMM requires sophisticated software and technology that is not accessible to most practitioners. Until this technology is accessible, it is unlikely to be used practically.

2.3.3 Summary

In general, non-metric traits are well studied in regard to their relationship with ancestry (Armelagos & Gerven, 2003; Caspari, 2009). However, these traits appear to have fewer established standards for assessment and determining likely ancestry as compared to metric methods (Jantz & Ousley, 2005, 2012; Langley et al., 2016a) or compared to age and sex estimation methods that use morphological traits in a similar manner (Bass, 2005; Phenice, 1969). In response to this deficiency of clear standards, Hefner's (2009) method attempts to standardize a suite of cranial traits that are most relevant for ancestry estimations. Due to its relatively recent debut, it needs to be studied further to find and reduce sources of error to ensure it is reliable.

2.4 Assessment standards

Updating ancestry methods based on the change in theoretical foundation, and for improved repeatability is important to improve chances of identifying an unknown individual. It is especially important to update or re-evaluate methods in general because forensic anthropologists may be called to testify to the results of their analysis. These analyses are part of

the evidence to which an anthropologist must testify as an expert witness, and method reliability and accuracy are used by the court to make judgements on their admissibility. An expert witness is an individual who is an expert on a specific aspect of a case that is outside the general knowledge of the public (Auxier, 1989; Auxier & Prichard, 2001; Hunt & Neudorf, 2016). If the methods are invalid or unreliable, this may impact the forensic anthropologist's role as an expert witness or the outcome of the case. Therefore, it is important to understand a forensic anthropologist's legal role and what legal requirements they must meet to ensure they can meet these requirements when updating or using these methods.

2.4.1 Court related standards that forensic anthropological methods must adhere to

A forensic anthropologist's testimony is based on the report that is submitted on the case. They provide their conclusions to the court based on the analysis of the case's physical evidence (ex. skeletal remains), which are followed by an explanation of the foundation for the conclusions, such as the methods and principles applied (Cwik, 1999). Before these conclusions can be testified to in court, they are reviewed in a pretrial admissibility hearing to ensure they meet the legal requirements of relevance, reliability, and validity (Christensen, 2004; Christensen & Crowder, 2009; *Daubert v. Merrell Dow Pharmaceuticals, Inc*, 1993; *General Electric Co. v. Joiner*, 1997; *Kelliher (Village of) v. Smith*, 1931; *Kumho Tire Company, Ltd. v. Carmichael*, 1999; *Regina v. Mohan*, 1994; Grivas & Komar, 2008; Holobinko, 2012). If the judge decides that the testimony does not meet these requirements, the evidence is excluded (Falco, 2016; Hunt & Neudorf, 2016; Page, Taylor, & Blenkin, 2011a, 2011b). Most relevant to this research are the legal requirements of reliability and validity because they require testimony to be based on reliable and valid methods or principles. Examples of reliability and validity include whether ancestry methods are repeatable or are grounded in the theory of isolation by distance,

respectively. Consequently, the methods and principles used to come to the conclusions iterated in the report are important for correctly carrying out a forensic anthropologist's court role.

Expert witnesses are brought in to educate the judge and/or jurors, also known as the trier of fact or fact finders, on complex scientific or technical aspects of a case (Auxier, 1989; Auxier & Prichard, 2001; Beech, 2015; Cwik, 1999). The trier of fact is in charge of coming to a conclusion, or verdict, on what occurred based on what they deem are the facts of the case. Consequently, ensuring they understand all evidence presented in court is paramount. Since the trier of fact does not have the specific knowledge or understanding to offer an opinion on the meaning of scientific or technical evidence, expert witnesses offer their opinion based on their factual observations (Blackwell & Seymour, 2015; Bolhofner & Seidel, 2019; Craig, 2016). In other words, an expert witness helps the trier of fact come to the correct conclusions on a certain aspect of the case so they can make informed judgements. Only expert witnesses can relay both their observations and conclusions or opinions (Eckert & Wright, 1997; Holobinko, 2012), as opposed to material (eye witnesses) or fact witnesses (scene processors) who are limited to relaying personally observed facts without an opinion on the meaning of what they observed (Blackwell & Seymour, 2015; Eckert & Wright, 1997; Holobinko, 2012). This is an important distinction because a forensic anthropologist's opinion is based on forensic anthropological research and methods. Therefore, in order for the opinion/conclusions to be considered reliable and valid, the methods must be reliable and based on valid principles.

It is also important to know how forensic anthropology evidence is used for identification because the expert needs to explain to the jury the limits of the method and how it can and cannot be used. Evidence is categorized as testamentary/material (eye witness or personal observation), physical or demonstrative (Eckert & Wright, 1997) or circumstantial (speculative)

(Holobinko, 2012). More specifically, demonstrative/physical evidence is the evidence that can be seen, felt, touched, smelled, tasted, or heard by the court (Eckert & Wright, 1997). An expert witness often provides testamentary evidence (conclusions) based on examination (methods) of physical evidence (skeletal material) from the case (Cwik, 1999; Eckert & Wright, 1997).

Evidence is further separated into two groups named “class” and “individual” evidence. Class evidence is where an item has similar characteristics to other items (ex. knife marks), while individual evidence has distinct characteristics that differentiate it from other similar items (ex. serrated vs. straight edge knife marks) (Holobinko, 2012). The process of identification using the biological profile involves the inclusion or exclusion of skeletal evidence into these categories to come to conclusions (Holobinko, 2012). Therefore, data from human identification techniques can be used to include or exclude missing individuals, such as including male or excluding female individuals based on sex assessment (class evidence). The evidence in these categories results in a ‘presumptive’ identification as opposed to a ‘tentative’ identification based on circumstantial evidence (ex. drivers license) (Holobinko, 2012). The results of ancestry assessment methods can be used as class evidence to include an individual into a group or population to which they are most similar. Therefore, when searching for matching missing person profiles, identity can be partially based on skeletal evidence indicating an individual of a certain ancestry. If available, individual evidence (ex. dental records) can include or exclude identities from matching the deceased individual because they are based on a set of characteristics considered to be unique or differentiates them from other individuals (Holobinko, 2012). Using ancestry assessments to include an individual in a similar population is dependent on the theory that says forensic anthropologists can do it (ie. IBD) and the method’s reliability and accuracy as stated by the published literature.

To admit the expert and their evidence into court, judges in the USA and Canada use guidelines called the '*Daubert* Criteria' and the '*Mohan* Framework', respectively, to review the scientific evidence and witness for relevance and reliability (Bethard & DiGangi, 2019; Craig, 2016; Cwik, 1999; Eckert & Wright, 1997). In general, the rules state that the expert is reviewed for their qualifications to ensure that they are an expert and that their expertise is relevant to the case (Craig, 2016; Cwik, 1999). Then, the content of the testimony, the evidence, is reviewed for its reliability (Craig, 2016; Cwik, 1999; Lesciotto, 2015). These admissibility rules differ between countries (Craig, 2016), however, since this research is relevant for both the USA and Canada, both countries' admissibility standards are considered.

In the USA, expert witness evidence admissibility is based on the Federal Rules of Evidence, specifically, rules 402, 403, and 702 (*F.R.E*, 2022), as well as the *Daubert* guidelines, which clarify how to determine the reliability of scientific evidence (Auxier & Prichard, 2001; Christensen, 2004; Christensen & Crowder, 2009; Craig, 2016; Cwik, 1999; *Daubert v. Merrell Dow Pharmaceuticals, Inc*, 1993; Eckert & Wright, 1997; *General Electric Co. v. Joiner*, 1997; Grivas & Komar, 2008; Holland & Crowder, 2019; Holobinko, 2012; *Kumho Tire Company, Ltd. v. Carmichael*, 1999; Rogers & Allard, 2004). These rules and guidelines assess expert qualifications, the relevance of testimony, and the reliability and validity of the theories, techniques, and methods used by the expert (Lesciotto, 2015). Rule 402 (*F.R.E*, 2022) states that evidence must be relevant and is admitted unless any other rule excludes it, such as rule 403 (*F.R.E*, 2022) that states evidence will be excluded if it is misleading, prejudicial, or confusing so much so that it "outweighs the probative value" (Holobinko, 2012). Finally, rule 702 (*F.R.E*, 2022) states that an expert's testimony must help the trier of fact (reinforcing relevance) and "is based on

sufficient facts or data, is the product of reliable principles and methods, and the expert has reliably applied these principles and methods to the facts of the case."

In Canada, evidence admissibility is based on a two-stage *Mohan* framework to determine relevance and reliability (Craig, 2016; Falco, 2016; Holobinko, 2012; Hunt & Neudorf, 2016; *Kelliher (Village of) v. Smith*, 1931; *Regina v. Mohan*, 1994; Rogers & Allard, 2004). Similar to rules 402 and 702 (*F.R.E.*, 2022), evidence must satisfy four questions during the first stage: is the expert qualified to give an opinion? Are the opinions relevant to the case? Are the opinions necessary for the trier of fact to come to the correct conclusions of the case? And, does evidence pass other exclusionary rules? The first stage also focuses on the reliability of scientific evidence, where any novel science is subject to "additional scrutiny to determine whether it meets a basic threshold of reliability" (*Regina v. Mohan*, 1994, p. 11). In the *Mohan* ruling, the judge referenced that if there was general acceptance of a method in the relevant discipline, it is important in the decision for determining reliability. However, the framework does not provide specific criteria for determining reliability, thus Canadian courts often look to the *Daubert* standards as a guide (Beech, 2015; Holobinko, 2012).

The *Daubert* guidelines clarify rule 702 (*F.R.E.*, 2022) when determining whether the evidence is scientific and reliable (Christensen, 2004; Holobinko, 2012), and were fine tuned through three cases (*Daubert v. Merrell Dow Pharmaceuticals, Inc*, 1993; *General Electric Co. v. Joiner*, 1997; *Kumho Tire Company, Ltd. v. Carmichael*, 1999) termed the '*Daubert* trilogy' (Grivas & Komar, 2008). First, the *Daubert* guidelines state that in order for the content of the testimony to be reliable, it must be testable or was tested with the scientific method, have been peer reviewed, have established standards, have known or potential error rates, and be widely accepted by the relevant discipline (*Daubert v. Merrell Dow Pharmaceuticals, Inc*, 1993). These

criteria explain how to determine if evidence is reliable according to how scientists consider something reliable or valid (Budowle et al., 2009; Hunt & Neudorf, 2016). This is important because forensic practitioners use the resulting methods and data from scientific research to analyse and determine what likely occurred to give rise to the evidence. If ancestry methods are tested repeatedly and provide consistent and accurate results within the literature, then it can demonstrate to the court that the expert's procedures were the fruition of the scientific method.

Second, the guidelines also state that there needs to be a valid and meaningful connection between the evidence and conclusion (*General Electric Co. v. Joiner*, 1997; Grivas & Komar, 2008). Therefore, an expert must be able to explain how they came to their conclusions. For example, ancestry estimations are based on morphological differences that arose through the theories of additive genetics and isolation by distance. Using these theories and evidence of trait frequencies in global populations, forensic anthropologists can explain how skeletal analysis is able to result in an ancestry estimate. Updating and re-evaluating methods is needed since trait lists are based on outdated theories of biological determinism.

Finally, in Canada, the second stage of admissibility is similar to rule 403 (*F.R.E.*, 2022), which is a cost-benefit analysis of whether the evidence has more probative value than the costs associated with it, such as it may be misleading or prejudicial (Craig, 2016; Falco, 2016; Holobinko, 2012; Rogers & Allard, 2004). If the cost is too high, then the evidence may be excluded. Moreover, if any evidence does not meet all these legal standards, it is considered inadmissible and, subsequently, excluded (Lesciotto, 2015; Page et al., 2011a). These exclusion criteria differ slightly between Canada and the USA for how evidence is handled. In Canada, evidence is excluded if it does not meet all of the criteria, whereas in the USA, evidence is given less weight in decision making. Most relevant to this research is that cases are often excluded due

to reliability issues (Page et al., 2011a, 2011b). For example, such exclusions were due to unfounded or implausible statistics on the method accuracy, error rates, or certainty; the inability to produce method standard procedures and ensure they were followed; or solely relying on general acceptance of the method in the discipline (Page et al., 2011b). These exclusions highlight the need for rigorously tested methods so they can be deemed reliable by the court as it is not up to the scientific community for whether the methodology meets these criteria (Grivas & Komar, 2008). Therefore, to avoid testimony exclusion, forensic anthropologists must make sure they can demonstrate the reliability of their methods to increase the chances of evidence admission.

2.4.2 Ways anthropological methods can meet court standards

One way to demonstrate to the courts that methods are reliable (as outlined by the *Daubert* guidelines) is to have industry (or discipline) standards, which are the result of the scientific method. Standards boast consistency and rigorous testing that adds credibility to the methods employed. Presently, there are no national Canadian or international professional standards to ensure the consistent application of forensic anthropology methods (Bolhofner & Seidel, 2019; Budowle et al., 2009; Christensen & Crowder, 2009). In the USA, a forensic anthropology subcommittee (FAS) has recently (2014) been established in the Organization of Scientific Area Committees (OSAC) for Forensic Science. This organization guides the development of best practice guidelines and standards by informing relevant forensic communities of research needs for standards development. Currently, the FAS is developing method standards to remedy the lack of national standards in the USA (Holland & Crowder, 2019). Since Canada does not have an equivalent committee established in a standards development organization, the research needs and standards published by OSAC are important for Canadian forensic anthropological research.

Without these standards in place, researchers, like Hefner (2009), take it upon themselves to create a standard method.

While there are no published standards for ancestry estimation, the subcommittee does refer to guidelines made by the Scientific Working Group for Forensic Anthropology (SWGANTH), a group that was originally organized by the FBI and Department of Defense in 2008 and is now held under OSAC. The SWGANTH guidelines highlight key properties that ancestry assessment methods need to have, or be founded on, to be considered valid (SWGANTH, 2013). These guidelines can be used by individuals to determine what a valid method of assessment is until official standards are published. The SWGANTH guidelines state that identification methods must be objective, systematic, replicable, validated, and be accompanied by probabilities to indicate certainty (SWGANTH, 2013). Furthermore, the guidelines state that ancestry assessments must have reference samples that are large enough to represent natural variability, clear trait descriptions, and appropriate statistical methods (SWGANTH, 2013). Similarly, the OSAC FAS recognizes that validation studies and understanding population frequencies are important for standards development in anthropological death investigation (OSAC, 2020). If a method meets or is developed following these guidelines, it has a better chance of passing the *Daubert* standards of reliability. Thus, Hefner's (2009) method and this research work towards several of the goals outlined by SWGANTH.

The first relevant procedure to this research, indicated by the process of standards development, is validation. Validation is a process involving multiple researchers testing a method to ensure its reliability, identifying limitations or errors in the method, providing recommendations for improvement, and defining the conditions that are needed to obtain results

(Budowle et al., 2009). Validation studies can also ensure the method works on different samples, otherwise the results of the initial study cannot be validated (Cunha & Ubelaker, 2020). Validation studies must replicate the method with no alterations in order to determine if it is reliable. Additionally, researchers must clearly state the accuracy and precision in a statistically significant and meaningful way (Christensen & Crowder, 2009). To carry out validation studies, anthropologists suggest that researchers use observer error, appropriate sample sizes, and appropriate statistics (Christensen & Crowder, 2009; Dudzik & Kolatorowicz, 2016).

Moreover, multiple forensic anthropologists and the SWGANTH guidelines recommend quality assurance (QA) which would oversee validation processes, finding error rates, and establishing professional standards (Budowle et al., 2009; Christensen & Crowder, 2009; Fleischman, Pierce, & Crowder, 2019; SWGANTH, 2011). These programs would systematically monitor and evaluate the quality of a laboratory's procedures, equipment, and training to ensure industry or professional standards are being met (Budowle et al., 2009; Fleischman et al., 2019). Both internal and external validation processes are used to identify problems because if only internal review is done, error may be overlooked due to systemic internal biases (Budowle et al., 2009; Fleischman et al., 2019). Therefore, external validation is especially important to identify potential error and provide recommendations for resolving it. In the end, quality assurance will bolster the reliability of the expert's methods and increase the chances of evidence admission (Fleischman et al., 2019). Until QA programs and standards are established, researchers can still pursue method validation as it is shown to be an important part of standards development.

This research provides an external validation using observer error to test for the repeatability Hefner's (2009; Hefner and Linde, 2018) scoring method. Much of the literature

discussing Hefner's scoring method (2009; Hefner and Linde, 2018) include quantification of intra- and inter- observer agreement, but do not discuss where error could be occurring so that descriptions can be improved. It is clear from the iterations of the scoring method that appear online (Hefner, Plemons, Kamnikar, Ousley, & Linde, n.d.; *Osteoware v. 2.4.037*, 2020; Wilczak & Dudar, 2020) and in publications (Hefner, 2009; Hefner & Linde, 2018) that there have been improvements in descriptions and attempts to reduce subjectivity. However, low inter-observer agreement suggests that the scoring method still requires some improvement (Coelho et al., 2017; Dinkele, 2018; Hefner, 2009; Klales & Kenyhercz, 2015; L'Abbé et al., 2011). Testing Hefner's (2009; Hefner and Linde, 2018) scoring method for repeatability can help further understand where error is occurring and result in suggestions to reduce errors, moving it towards a more objective application as suggested by SWGANTH (2013). Multiple validation studies, including this research, can also contribute to understanding error rates which are needed to meet legal requirements (Christensen & Crowder, 2009). Finally, the results of this research may help move this method toward its inclusion in future professional standards.

Recognizing error and how often error occurs is an important part of method development, especially when forensic anthropologists must employ a variety of qualitative and quantitative methods due to the variable nature of casework (Kennedy, 2003). These can be based on rigorous methodologies or less rigorous observational studies (Christensen & Crowder, 2009)**.** No matter how rigorous, all methods can produce error. For example, error occurs from inaccurate measurements produced by humans or instruments, such as incorrect caliper calibration (Budowle et al., 2009; Dudzik & Kolatorowicz, 2016; Fleischman et al., 2019)**.** Other sources of error include the preservation of the skull, recording errors, and inexperience (Dudzik & Kolatorowicz, 2016)**.** A validation study, such as the current research, can catch these sources of error.

Subjectivity, such as introduced through bias or observer interpretation, is a common source of error (Budowle et al., 2009). In the application and development of ancestry methods from race assessment, both confirmation and contextual bias were prevalent (Budowle et al., 2009). In fact, Morton had manipulated craniometric results to fit his preconceived notions of race, making it plausible that other researchers did this with non-metric traits (DiGangi & Hefner, 2013). Positive identification was often used to find traits that were already thought to be associated with the confirmed race, thus "validating" the initial assessment. The Harvard List was founded on Hooton's research which used confirmation bias (Hefner, Pilloud, Buikstra, & Vogelsberg, 2016). Furthermore, the racial labels given to specific traits inherently bias the practitioner's estimate because research has shown that any single trait can be found in multiple populations regardless of race (Lewontin, 1972; Relethford, 2002; Rhine, 1990; Spiros, 2018). It is these biases in previous "race" estimation research that fuels the reassessment of these traits and methods in the context of population frequency and isolation by distance (IBD). A reassessment of traits means their application in ancestry estimation is valid in relation to the theoretical foundation. Hefner (2009, 2018) has already attempted to reduce the biases related to traits by removing racial labels and collecting trait frequencies for various populations. Consequently, a forensic anthropologist's determination of ancestry is based on population frequencies under the theory of IBD rather than confirmation bias, bringing it in line with SWGANTH recommendations (SWGANTH, 2012, 2013).

Trait lists from which Hefner's (2009; Hefner and Linde, 2018) scoring method was adapted have unclear assessment standards (Hughes et al., 2011) causing trait descriptions to be interpreted differently among anthropologists and ancestry to be inconsistently estimated (Goodman, 1997; Maddux et al., 2015; Rhine, 1990). Even early researchers such as Wood-Jones

(1931a) who studied traits for racial assessment recognized there was no consistency between trait descriptions. This need for interpretation results in false positives or false negatives due to observer error. For example, the zygomatic projection had two different methods describing how to assess it (Cataldo-Ramirez, Garvin, & Cabo, 2020). Each method resulted in different frequency patterns among populations (Cataldo-Ramirez, Garvin, & Cabo, 2020), reinforcing the importance of standards to ensure consistent application. These lists were introduced in the early 20th century long before the current legal admissibility standards, and were not the fruition of the scientific method (Holland & Crowder, 2019; Holobinko, 2012). As such, if a method like a trait list fails to fulfil the legal requirements of reliability, evidence could be thrown out of court due to the lack of justification for how ancestry was assigned (Holobinko, 2012; Lesciotto, 2015; Page et al., 2011a, 2011b). To avoid exclusion, the redefining of traits with clearer assessment instructions was one of the goals of Hefner's (2009; Hefner and Linde, 2018) scoring method. However, qualitative methods have a higher level of subjectivity than quantitative methods because they rely on this description interpretation. Therefore, greater description clarity is needed for consistent application before a method can be deemed reliable by *Daubert* or before it is integrated as a discipline standard. As an attempt to reduce observer error, Hefner (2009) clearly describes traits and how to assess them with accompanied line drawings, and, more recently, with photographs (Hefner and Linde, 2018). This is in line with the SWGANTH guidelines (2013) that state in order for morphological trait assessment to be used, there must be clear trait descriptions. It also pushes the method towards being more objective, systematic, and replicable as recommended by SWGANTH (2013). To determine if the descriptions are clear enough to produce consistent results between observers, multiple studies need to test his method, such as this research does. If this

research does not produce consistent results or there are issues with understanding trait descriptions, the author can provide comments on which descriptions need to be clearer.

Since professional standards do not currently exist, practitioners rely on published literature for methods. This lack of professional standards in conjunction with the variety of methods means method specific training is not always possible. Therefore, methods must be easily understood through written descriptions and associated images. As recognized by Edgar (2017), until there are professional standards, methods will be used by individuals who have less experience in that method. Therefore, researchers creating the methods should help those who will use the method to collect data that is useful through providing clear instructions. Testing Hefner's (2009; Hefner and Linde, 2018) scoring method can help improve clarity and create a more reliable method that can be accurately applied regardless of training.

The changes that cranial nonmetric trait analyses have undergone in relation to estimating ancestry are a reflection of how other anthropological methods are carried out, thus could be argued are a discipline standard. For example, the ASUDAS plaques were introduced to reduce visual subjectivity and increase precision for dental nonmetric trait identification (Carayon et al., 2018). This plaque system has high replicability in interobserver tests (Irish, 2014) and has allowed for consistent recording of traits across multiple studies, resulting in the production of global trait frequencies (Turner, 1986). More recently, Edgar included error rates to help meet legal standards (Edgar, 2014). Throughout its history, ASUDAS has undergone several iterations and improvements, including a textbook with pictures to reduce problems with visual identification (Edgar, 2017). This process is similar to how Hefner's (2009) scoring method has undergone changes (Hefner & Linde, 2018; Hefner et al., n.d.; *Osteoware v. 2.4.037*, 2020; Wilczak & Dudar, 2020). Hefner (2003) began by selecting traits from other publications and

improving on or using descriptions created by other authors, such as Hauser and De Stefano (1989) and Rhine (1990). A proposed method was published (Hefner, 2009), it was subsequently tested by Kamnikar et al. (2018), and recommendations for improvement were incorporated into the atlas that holds photographic examples of trait expressions (Hefner & Linde, 2018). Similarly, a postcranial method for non-metric trait identification and their utility in ancestry estimates have followed suit; a recent introduction of a standard set of traits for analysis accompanied by descriptions and line drawings (Spiros, 2018). Spiros (2018) found high inter-observer agreement when testing the postcranial method and significant differences in trait frequencies between American black and white individuals for four traits, thus supporting their utility (Spiros, 2018). The consistency of the methodology across multiple skeletal and dental methods lends credibility to anthropology adopting a set of standards, which can help a method meet legal requirements.

While reducing subjectivity is important, subjectivity does not always equate to unreliability (Budowle et al., 2009; Christensen & Crowder, 2009). Many quantitative methods are often seen as more reliable than qualitative methods because they are more objective. A qualitative assessment is still useful to exclude scenarios or explanations to come to a conclusion (Grivas & Komar, 2008). For example, ancestry estimations resulting from trait assessment could exclude ancestries and, therefore, possible identities. To accompany qualitative descriptions, SWGANTH guidelines state that they should be used in conjunction with statistical analysis (SWGANTH, 2012), such as a quantification of how well a qualitative method performs. Importantly, they should be used appropriately to be useful (Budowle et al., 2009). If statistics are not appropriate, qualitative explanations can be used to explain the reliability of a method. For example, explaining what features are used for interpretation, their relative

rarity/commonality, and/or limitations to the method (Budowle et al., 2009). A method will always have some kind of unknown error (Carayon et al., 2018), therefore, error rates can be calculated to convey the ability or strength of method performance. These error rates are also included in the *Daubert* criteria, thus solidifying their usefulness in determining reliability.

In regard to morphological analysis, error can be documented through repeatability tests (Fleischman et al., 2019), such as inter-observer error tests that are recommended when subjectivity cannot be fully eliminated (Rogers & Allard, 2004). There is a wide range of trait agreement when testing Hefner's (2009) scoring method for both intra-observer (Atkinson & Tallman, 2020; Coelho et al., 2017; Dinkele, 2018; Hefner, 2009; Hurst, 2012; Kamnikar et al., 2018; Kilroy et al., 2020; L'Abbé et al., 2011; Moffit, 2017; Wang, 2016) and inter-observer error (Coelho et al., 2017; Dinkele, 2018; Hefner, 2009; Klales & Kenyhercz, 2015; L'Abbé et al., 2011). On average, intra-observer agreement is high and much greater than inter-observer agreement, while inter-observer agreement is often moderate to poor (Coelho et al., 2017; Hefner, 2009; Hurst, 2012; Kamnikar et al., 2018; Klales & Kenyhercz, 2015; L'Abbé et al., 2011). While these studies help the scoring method meet the legal requirements for known error rates, inter-observer agreement must also be high because if scores are not consistent between researchers, then trait frequencies that are used as references may not be reliable. Drost (2011) notes that reliability scores of 0.7 or higher are sufficient, so inter-rater reliability should have an agreement value of 0.7 or more for each trait. Testing Hefner's scoring method can shed light as to why inter-observer agreement rates are low and how to improve them.

Rather than focusing solely on understanding error rates, anthropologists suggest that focus should be on defining what the error is, what caused it, what the consequences are, and what is being done to correct errors because this is considered a part of the validation process

(Budowle et al., 2009). As such, this study identifies sources of error and provides recommendations to correct them. Subsequent inter-observer studies can also help determine if there is a pattern for trait disagreement revealing where error is occurring (Banerjee, Capozzoli, McSweeney, & Sinha, 1999).

While error rates can relay certainty for the scoring method's repeatability, it is not appropriate for statistical ancestry estimation. In fact, SWGANTH guidelines state that when assessing ancestry, probabilities of certainty must be reported (SWGANTH, 2013), which are how likely one outcome is relative to other outcomes. In regard to ancestry, it is how likely the individual belongs to each reference population (Rogers & Allard, 2004). This is because trait expressions and combinations can be found in multiple populations, thus the individual could fit into multiple categories. As a result, probabilities are more appropriate than error rates to relay certainty when multiple answers can be correct (Budowle et al., 2009). This approach is more objective than an anthropologist providing a single ancestry estimate based on subjective interpretation without a level of certainty. It is also more objective than trait lists that assess the racial identity of the individual since these were founded on confirmation bias. Probabilities remove individual interpretation because, as some authors suggest, mathematical means of arriving at a conclusion ensure replicability, make criteria explicit, and provide a method that can be debated and discussed (Hefner, 2009; Rogers & Allard, 2004). To fulfill SWGANTH's recommendations for probabilities of certainty, Hefner's method (Hefner, 2009, 2020; Hefner & Linde, 2018) employs probabilities resulting from the comparison of an individual's trait combination to patterns of trait expressions in each reference population. Using the highest probability of membership for the ancestry estimate fulfills SWGANTH guidelines of a method being objective (SWGANTH, 2012, 2013).

Finally, these statistical methods can be accompanied with accuracy rates that fulfill the *Daubert* standards. The *Daubert* criteria specifically state that methods must also be valid; validity that is based on accuracy (Craig, 2016) is called construct validity (Drost, 2011). This means if ancestry and morphoscopic traits are related, where expression of morphoscopic traits reflect ancestry, inputting trait data into *MaMD Analytical* to estimate ancestry, and measuring how often the program is correct is an indication of how well the concept is translated into a functioning reality. This validity is not the focus of the study, but it is important to mention for the sake of discussing the impact of score disagreement.

Accuracy of the estimates are dependent on comparative data. Forensic anthropologists, along with OSAC, suggest that population frequencies must be used as comparative data for unique feature identifiers (Christensen & Crowder, 2009). Forensic anthropology subcommittees, as well as SWGANTH guidelines on ancestry assessment also recognize that determination of population frequencies are important for standards development in anthropological death investigation methods (OSAC, 2020). This is because positive identification is based on knowing how rare characteristics are in relation to the source (Rogers & Allard, 2004). If everyone has their own experiences with different populations, they have different ideas of what traits are rare or not (Rogers & Allard, 2004). Thus, SWGANTH guidelines state that reference populations should adequately represent natural human variation and are critically evaluated for sample size (OSAC, 2020; SWGANTH, 2013). Since a method can only identify individuals that belong to a reference population, many reference populations need to be included for comparison (Armelagos & Gerven, 2003; Birkby, 1966; Elliott & Collard, 2009; Fisher & Gill, 1990; Goodman, 1997). Theoretically, more populations, and individuals within the population, that are available for comparison should increase estimation accuracy because more population

differences can be detected. However, biological profile categories often lack comparative data and reference samples that can be used to show the range of human variation, resulting in ambiguous estimations (Holobinko, 2012).

Hefner's (2009, 2020; Hefner and Linde, 2018) ancestry estimation method relies on a database holding trait frequencies of multiple reference populations with differing lineages for comparison. Having multiple reference populations fits the SWGANTH guidelines that state reference populations must adequately represent human variation (SWGANTH, 2013). At this time, *MaMD Analytical* only estimates ancestry from a sample of individuals in the MaM Databank, and only results in six ancestry estimates with racial labels, indicating the need for further study before global application. Hefner (2009) has reported high assessment accuracies when using multiple statistical methods to assign ancestry from morphoscopic traits, and, when used in case studies, the first most probable ancestry was correct (Plemons & Hefner, 2016). However, other researchers using Hefner's (2009; Hefner and Linde 2018) method (scoring method plus a statistical determination of ancestry estimation) reported lower ancestry estimation accuracies ranging from percentages in the low 60s to high 80s (Go & Hefner, 2020; Hefner & Ousley, 2014; Hefner et al., 2015; Klales & Kenyhercz, 2015; Monsalve & Hefner, 2016). The range of accuracy from different studies provides known rates that are required to meet legal standards.

For the purposes of this study, accuracy cannot be tested due to the unknown ancestry of the individuals that make up the sample population. Instead, the results from *HefneR* and *MaMD Analytical* in comparison to each other and to *Fordisc* can provide insight to the accuracy of the methods by measuring how often these programs agree. This is because *Fordisc* and *MaMD Analytical* and *HefneR* use different methods and different reference groups for their analyses,

but are intended to measure the same thing, ancestry. If the reference individuals are adequately representing the human variation needed to estimate ancestry accurately, then all the programs should have the same result. The results from this study can add to the knowledge of whether these two different types of data, metric and morphoscopic, result in the same ancestry estimate. Even though this research is unable to determine which software is more accurate, it will indicate how often these two types of data agree. If the results between software agree more than chance, then it can lend credibility to the validity of the software; that they are actually measuring what they intend to measure (Drost, 2011). Agreement can also lend credibility to the scoring method being reliable. If *MaMD Analytical* and *Fordisc* do not agree, this does not necessarily mean that the morphoscopic programs are not measuring ancestry, but rather that the associated reference populations may be too different to return the same results. Furthermore, it does not mean that the morphoscopic traits are not an indicator of ancestry, but that the scoring method may need to be improved so there is higher observer agreement, thus reliability of the scores and, ultimately, the same ancestry assessment.

In summary, using SWGANTH guidelines and OSAC recommendations to address methodological problems and fulfil the legal rulings, Hefner's (2009, 2020; Hefner and Linde, 2018) method departs from the trait list format towards clearer assessment standards that can provide known error rates. Hefner's (2009; Hefner and Linde, 2018) proposed scoring method includes a clear, systematic assessment guide for more objective trait scoring than trait lists. Trait expressions are depicted as line drawings with corresponding numerical scores and brief descriptions for how to assess them. The line drawings reduce description misinterpretations, and the numerical scores allow human variation to be documented without racial bias and converted to frequencies for detecting population differences (Kamnikar et al., 2018; Klales & Kenyhercz,

2015). Finally, statistical methods to estimate ancestry from trait frequencies provides an objective probability of certainty and can be accompanied by accuracy rates. While these adjustments meet the SWGANTH guidelines, his method is relatively new and needs further study before fulfilling the legal standards. Each part of Hefner's (2009) method must be tested multiple times to fulfil the legal standards that require thoroughly studied death investigation methods, and this additional validation study can further understand error in the method, error rates, and how observer errors impact ancestry results if they are not rectified. Testing his method and providing recommendations can lead to improved reliability of the scores and the accuracy of the method in the future, therefore, improved chances of passing admissibility standards.

2.5 Historical context of the study individuals[1]

Due to the unforeseen events of a global pandemic, the original study populations were unavailable. Instead, this research includes skeletal individuals from anatomical teaching collections held at the University of Manitoba (U of M) and Brandon University (BU), whose origins are uncertain. The majority of individuals in these collections have characteristics typical of processed teaching skeletons, such as cranial sectioning and hardware for anatomical articulation. Additionally, the individuals do not have soil staining or evidence of weathering that would be expected for archaeological individuals. While it was a common practice for archaeologists to dig up Indigenous ancestors in North America for research (Breske, 2018; Donlon, 1994; Koehler, 2007; McNiven & Russell, 2005; Thomas, 2014), it does not appear these individuals are the result of that practice. It is known that one individual in the BU collection was bought from a company selling anatomical teaching material, Kilgore International, but the others

[1] Trigger warning, description of violence against humans, particularly minorities

have no such documentation. The U of M individuals also have no documentation of purchase but are consistent with anatomical teaching individuals. Despite the evidence pointing towards purchase from a company, their ultimate point of origin is unclear. Since there is no documentation to trace their origin, the literature on historical practices for obtaining recently deceased individuals for research and teaching collections can provide clues and important context for these individuals.

An increase in the trafficking of human skulls occurred between 1850 and 1930 for race studies (Roque, 2018). In order to obtain skeletons for studies on race, Indigenous peoples, deemed an inferior race, were not only robbed of their recently deceased ancestors, but were murdered for research and display in museums because research facilities were offering payment for them (Breske, 2018; McNiven & Russell, 2005). Additionally, after Indigenous peoples were killed while defending their territory, their heads were taken for this payment (Thomas, 2014). These recently deceased individuals were taken without consent or regard for common Indigenous beliefs that the deceased must not be disturbed (Densmore, 1979; Ferguson, Anyon, & Ladd, 1996; Johnston, 1987; Kluth & Munnell, 1997). Today, stolen remains, both archaeological and non-archaeological, continue to be held as skeletal collections for research in various educational institutions, such as museums and universities, waiting to be repatriated (Bergmann, 2011). Since the recently deceased were never documented, there is no way to trace their origins and determine who is held in these collections. Thus, if the individuals in the study population are of Indigenous ancestry, these events could account for the non-archaeological condition of the remains.

Modern research collections, such as those established by William Bass, were developed from donated individuals with various backgrounds (Campanacho, Alves Cardoso, & Ubelaker, 2021). These collections have “biohistories” associated with them so researchers know who the individual was and where they came from (Campanacho et al., 2021). If this were the case at the

U of M and BU, documentation would be obvious because of the systematic nature of collecting individual information. Due to the lack of documentation and inconsistent preparation of the skulls, this is unlikely the procurement process.

Along with skulls being collected for researchers, individuals would collect trophy skulls from conflicts. Trophy skulls were taken as proof of kill in the context of colonialism in North America and the violence against Indigenous peoples (Yucha, Pokines, & Bartelink, 2017). Another common source of trophy skulls in the USA came from World War 2 and the Vietnam war (Yucha et al., 2017). These skulls were often used for decoration or "practical" purposes such as ash trays (Yucha et al., 2017). Yucha et al. (2017) noted that, in some cases, trophy skulls were incomplete or damaged due to the conflict that produced them. The selling and trading of human remains moved these skulls around the country and globe, meaning it is difficult to ascertain their ancestral origins (Yucha et al., 2017). The individuals in the BU collection that are incomplete and lack hardware for teaching may be trophy skulls, thus may have Vietnamese or European ancestry.

Medical schools in the USA and UK in the 18th and 19th century used cadavers for dissections and subsequently processed them into teaching skeletons (Campanacho et al., 2021; Fabian, 2010; Halperin, 2007; Walker, 2008). Executed criminals were the main source of cadavers because it was considered further punishment to desecrate the body (Halperin, 2007). However, the supply of executed criminals became hard to come by as more medical schools opened and demand for cadavers for study increased. This led to grave robbing and, in some cases, murder because there was payment for obtaining fresh corpses (Halperin, 2007). Individuals would collect bodies off the streets for payment (Fabian, 2010), or medical schools would advertise that slave owners could "dispose of" sick slaves and get paid for them (Halperin, 2007). Wars allowed for a "good supply of "white" men's skulls" (Fabian, 2010, p. 5) whereas funeral homes, hospitals,

the docks, and cemeteries became new sources of cadavers when supplies ran low (Halperin, 2007).

Grave robbing occurred primarily in cemeteries that held the poor or those of African American descent (Halperin, 2007), resulting in hundreds of buried individuals being removed per year in New York alone (Humphrey, 1973). These populations were targeted because very few people had power over what occurred with their dead (Halperin, 2007). Grave robbing only became illegal when white cemeteries or cemeteries of the rich and powerful began to be raided (Humphrey, 1973). In some areas, the laws changed so only individuals who were poor, imprisoned, or African American would be used for dissection (Halperin, 2007; Humphrey, 1973). Regarding research collections, collectors who were colleagues and friends of Morton were even documented to take pleasure in robbing graves for the purposes of racial research (Fabian, 2010). While some cadavers were integrated into teaching collections (Campanacho et al., 2021), there are documented cases of medical cadavers being buried (Murphy, 2011; Stubblefield, 2011). Those that ended up in teaching collections are likely of African American origin since two-thirds of cadavers were this ancestry (Humphrey, 1973), but European ancestry is also a possibility.

Smaller universities that did not have medical schools associated with them would have to obtain teaching skeletons through other means. There are only a few pieces of information that explicitly discuss the buying and selling of human remains between institutions once initially acquired. It has been noted that 19th century newspapers reported crates of bodies being shipped to medical schools (Humphrey, 1973) or bodies placed in barrels to be shipped, but labelled as paint (Halperin, 2007). Halperin (2007) noted that one newspaper article was able to confirm the shipping of African American bodies in barrels to a university in New England. If individuals in

the BU and U of M collections were purchased from other universities, then European and African ancestry is most likely.

Alternatively, many universities' teaching collections are said to be obtained from companies in India after laws made many of the previous sources illegal. Fabian (2010) explained that one skull collector saw that the Malay population had so many dead that he could make profit off their "waste." This collector would gather deceased from hospital waste, along river banks, and from executioners (Fabian, 2010). Due to this opportunity for profit from human remains, bone factories popped up all over India (Carney, 2011). The bone trade from India started in remote villages, and lasted for 150 years until 1985 when exporting human remains from India was banned (Carney, 2011). These bone factories had means of obtaining skeletons similar to medical schools in the 18th and 19th centuries: grave robbing. Grave robbing was not illegal in India during this time compared to colonized countries, therefore, India made up most of the market for human remains from the 20th century (Carney, 2011). Carney (2011) revealed that the Chicago Tribune reported 60,000 skeletons were exported from India in a single year, enough for every medical student to buy their own skeleton. Companies such as Young Brothers would buy skeletons "wholesale" to wire and paint anatomical diagrams on them for the purpose of selling them to medical schools (Carney, 2011). Kilgore International, where one of the study individuals came from, obtained many of these teaching skeletons from the Reknas Company in India (Carney, 2011). Osta International in Canada also had stocked these skeletons from India and sold them throughout the USA and Europe (Carney, 2011). This history presents another possible origin for the individuals in the study population.

The movement of individuals across the country can impede investigations into origin, especially if they were moved without documentation to avoid legal repercussions or had no

original documentation. It is not uncommon that documentation for skeletons was transcribed incorrectly, never received, or misplaced as they changed hands (Bergmann, 2011; Wittwer-Backofen et al., 2014). Therefore, documentation of many teaching collections is lacking or non-existent, such as for BU and U of M. Furthermore, processes of authentication for a skull in racial research left considerable room for bias and falsification, therefore, even if documentation were provided for any current study individuals that exhibited characteristics of a research skull, it may not be correct. For example, authentication of the skull's origins could include, in part, the credibility of the donor, seller, or collector (Roque, 2018). If the authenticity of the skull was doubted, then it would undergo craniometric analysis to confirm or "correct" the origin (Roque, 2018). In response to the lack of documentation, some researchers have completed osteological analyses of previous medical cadavers to determine what demographics made up cadaver populations (Mincer, 2015; Murphy, 2011; Stubblefield, 2011). Murphy (2011) found twice as many female cadavers than males, while Mincer (2015) was unable to determine demographic information for the two individuals they studied. Stubblefield's (2011) research included one individual suspected to be an anatomical cadaver, which was estimated to be an African American female. However, the estimate contrasts with the demographic information during the time that this individual was likely prepared which showed that African American populations were very low in that area. The results of Stubblefield (2011) research indicate that Black cadavers from southern states were shipped to northern schools for teaching. In addition to osteological analysis to determine demographics, historical, morphological, and biological analyses have been completed on other skeletal collections (Wittwer-Backofen et al., 2014). All these analyses can be a base for further investigation into the origins of teaching skeletons.

While there are analyses that can be done to more clearly ascertain who these individuals were, it is outside the scope of this research. Instead, generalizations can be made from the overall appearance and condition of the skulls. Skulls collected for research or as trophies differ from teaching collections with their purposes and, therefore, their preparation methods. Skulls whose specific purpose is teaching often have hardware alterations that allow one to take apart a skull or articulate it with the jaw. Research skulls collected for racial research likely do not have these alterations as they would get in the way of data collection. For example, cutting the skull cap off would introduce measurement error. Finally, trophy skulls can come in a variety of conditions with damage from conflict or alterations for decoration purposes (Yucha et al., 2017). The majority of study individuals have alterations that relate to teaching individuals, indicating origins from medical schools or companies selling teaching individuals. However, there are a few BU individuals that exhibit extensive damage and/or no specific alteration for teaching. These individuals may have come into the collection as trophy skulls or from racial research, however, it is not clear.

Based on the evidence in the literature, teaching collections in Canadian universities could have been acquired through the same means as in the UK and USA, such as graverobbing of the poor or black. Otherwise, they could have been purchased from institutions that historically practiced grave robbing, meaning an individual's possible ancestry would be from Africa or Europe. If the U of M or BU purchased skeletal remains from companies that imported human remains from India (eg., Kilgore or Osta international), individuals may be of Indian descent. They could also be of Asian or Indigenous descent if the individuals without alterations are trophy skulls or skulls collected originally for racial research. However, it is more likely they are of Indian,

European, or African descent. Overall, these individuals could come from several places and this study may shine light on their ancestry.

2.6 Ethical Concerns

The ethical use of human remains in research is an important component to consider because they were previously living individuals with various beliefs. Some central themes, with a Western perspective, related to the study and handling of human remains include consent, dignity and respect, and awareness of the risks and benefits (Tarlow, 2006). While the Western perspective of ethics was followed in the protocols for this research, not everyone agrees on how the deceased should be handled.

Many teaching collections were acquired during the early 20th century when ethics and consent were not a major component to research as it is now. There have been documented cases of individuals being used for educational purposes after the individual has explicitly requested not to be (Alberti, Bienkowski, Chapman, & Drew, 2009), while extreme means of obtaining individuals included grave robbing (Walker, 2008). The origin of the remains at the U of M and BU is unclear because there is no documentation, and there is a concern by the U of M department that companies selling the remains were not honest about their sourcing. Additionally, there is no evidence that consent was obtained for their use in teaching and research suggesting that they are not modern donated individuals. For the purposes of modern research on human remains, the subject's consent to scientific research after death is mandatory. If consent is not obtained before death, the next of kin is normally asked (Tarlow, 2006). However, without this documentation, we have no idea who the next of kin would be and it is not likely possible to determine if these are individuals from the 18th and 19th centuries. In lieu of this documentation, I obtained consent from the laboratory coordinators at BU and U of M and the head of the department at the U of M to

work with these individuals. Furthermore, before research began, an ethics board reviewed the process of my research for approval, which covered both BU and U of M. This was done for the sake of transparency and to practice current ethical standards placed by the Government of Canada Panel on Research Ethics, even though anonymous human remains do not require ethics approval (TCPS, 2018). While it is unknown whether these study individuals gave consent to be used in research and teaching, it is important to acknowledge the contribution they have made to teaching many generations of anthropologists about the complexity of human skeletal anatomy.

Dignity and respect when handling remains varies between different groups of people because they have their own ideas as to what this means (Tarlow, 2006; Walker, 2008). Following cultural protocols or wishes of the deceased would be the most respectful, such as Indigenous peoples who state that rituals must occur if the dead are disturbed (Kluth & Munnell, 1997). However, again, we do not know their origins and cannot assume what these individuals would have considered dignified and respectful treatment because it is dependent on their consent. As a scientist, respect can include handling remains with care (Tarlow, 2006), therefore, I made sure to handle them as delicately as possible. I followed protocols to mitigate any physical damage, such as scratches from equipment (minor) or dropping the skull (severe). These were prevented by using both hands to carry a skull close to the body to the work station, and gently handling the skull with both hands over a padded surface. When using equipment, such as calipers, I would gently place them on the area of interest.

When speaking about the teaching individuals, I actively practice and correct my language so I humanize skeletal individuals rather than objectify them. I also did not do anything outside the scope of my research, such as sex estimations, which would have revealed more about the individual that I did not need to know for the purpose of this research. This can also be considered

respectful of their lived experiences. Since I did not know the origin or ancestry of the teaching individual collections, I could not know the individual's likely wishes for how they were handled after death. The ancestry estimations might provide insight into whether the individuals need to be repatriated and if respectful and dignified treatment more aligned with their possible cultural affiliation can be practiced.

Finally, contemporary ethics (TCPS, 2018) require that work with human individuals or groups should provide benefits not harms. Since these individuals were once alive, and may have living descendants, these benefits and harms should still be taken into consideration. Harms can be physical, such as the scratches on or breaking of bone; or emotional, such as not following cultural protocols. The risks to the deceased teaching individual were low at the time of research, and were mitigated through respectful handling. These harms may become greater if more information is discovered about them, however, until we know more, we cannot predict what these harms would be. Currently, the benefits of having a likely ancestry estimate can help the departments identify if their acquisition was ethical and outweighs the physical harms that were mitigated in the research. An ancestry estimate may help identify the past harm of not obtaining consent and rectify this by putting the individuals to rest. The harms of the past can not be erased, but we can move forward to make reparations. While repatriation would be an ideal goal, it is unclear who these individuals are due to the lack of context and documentation and an ancestry estimate is only one part of the process that could provide greater clarity.

While the U of M is looking for replacement teaching skeletons with more ethical origins, the current individuals remain in the collection for teaching and research purposes. Even if some individuals did donate their remains for educational purposes, some of the individuals may not have ethical origins. I recognize that the research I carried out should have benefits, such as the

improvement of ancestry estimations and putting the individuals to rest, that outweigh the costs of not obtaining the consent of these individuals at the time they were acquired for the teaching collection. I also recognize that I may have not respected the wishes of the deceased and their families. I treated these teaching individuals with what I consider respect and care, and followed protocols that the laboratory coordinators required.

Chapter 3: Methods and Materials

3.1 Materials

3.1.1 Sample populations

The study population includes twenty-seven (N=27) skeletal teaching individuals held at the University of Manitoba (n=10; UofM) and Brandon University (n=17; BU). Only adult individuals, determined by having all their teeth in occlusion (>21 years), were studied because it is unknown whether cranial traits are visible or fully developed in immature individuals. However, there were two exceptions to fully occluded teeth. One individual had unerupted third molars, but upon further inspection, the clavicles were fully fused, indicating adulthood (Scheuer & Black, 2004). The second individual had ¾ erupted third molars and half of the medial clavicular epiphysis formed and fused, indicating this individual is most likely an adult. The younger individual was included as a study individual because of the already reduced sample size. Any trait that was damaged or exhibited pathological changes were not scored for that individual.

3.1.2 Reference populations

Reference populations that were used are found in the databases attached to the statistical software *Fordisc* and *MaMD Analytical*; *HefneR* does not use a database as the equation used in the program was created from frequencies in the original Hefner (2009) publication. *Fordisc* includes metric information from two data banks. The first is Howell's databank, which contains 2,504 19th and 20th century individuals from twenty eight worldwide populations (Howell, 1996). The second is the Forensic Anthropology Databank (FDB), which holds demographic and skeletal information (both metric and non-metric cranial and post-cranial) from the Terry Collection (born after 1900), and approximately 3,400 modern forensic cases (Ousley & Jantz, 2012). Both the

Terry collection and 2,400 individuals from FDB have known sex and ancestry (Ousley & Jantz, 2012).

The software *MaMD Analytical* contains reference populations from the Macromorphoscopic Data Bank (MaMD), which includes 7,397 individuals, with about half being publicly available for use (Hefner, 2018). Some of these population's frequencies were also included in Hefner (2009). Of those used in this software, up to 2,363 are modern American populations, such as American Black and American White individuals, and up to 1,790 are from countries elsewhere, such those in Europe or East Asia (Hefner, 2018). To have comparable representation of global variation, not all populations and individuals were included into *MaMD Analytical* for comparison (Hefner, personal communication).

The *HefneR* application also uses the population frequencies found in Hefner's 2009 article (Coelho, Curate, & Navega, 2020) which are broadly named African, Asian, European, and American Indigenous. In each application, all reference populations were chosen for comparison because the teaching individuals are of unknown origin, with likely ancestry coming from any population such as those of American Indigenous (McNiven & Russell, 2005), African (Halperin, 2007; Humphrey, 1973), Indian (Carney, 2011), or European descent (Fabian, 2010; Halperin, 2007).

3.1.3 Variables

The sixteen cranial traits identified by Hefner (2009) and outlined within *Osteoware* were assessed for this research (Table 2). Eight traits were scored as ordinal values, where as the numerical value increases, there is a progression of the expression (ex. small to large; Table 2). Two traits were scored with binary values that correspond to trait presence or absence (Table 2). Lastly, six traits were scored as nominal scores, which have no progression of trait expression as

the value increases (Table 2). Instead, nominal scores reflect distinct categories of shape expression (ex. square vs. circle) rather than a progression of one expression. All individuals with any number of traits present were assessed. However, *MaMD Analytical* requires at least three trait scores for an estimation, meaning only individuals with three or more can be assessed. For individuals with fewer traits, ancestry may not be accurate because the fewer the traits, the more likely the combinations are found across multiple populations since the majority of the total extent of human variation is found in each population (Lewontin, 1972; Relethford, 2004).

Thirty-six measurements (continuous variables) were taken; twenty-seven for the cranium and nine for the mandible (Appendix 1; Jantz & Ousley, 2012). Age at death and sex were not available and were not be assessed because they have little, if any, effect on trait expressions (L'Abbé et al., 2011; McDowell et al., 2012) or frequency distributions (Hefner, 2016; Maddux et al., 2015). In fact, pooling sexes has resulted in higher estimation accuracy (Hefner, 2009, 2016; Klales & Kenyhercz, 2015).

Below are the descriptions and line drawings for each trait used in this research. The author used all components of the descriptions available for assessment but deferred to

Table 2: List of morphoscopic traits, their abbreviations, and the numerical score range which correspond to varying trait expressions. Traits without * are ordinal variables, * are nominal, and ** are binomial. Adapted from (Hefner, 2018).

Trait	Abbreviation	Score Range
Anterior Nasal Spine	**ANS**	1-3
Inferior Nasal Aperture	**INA**	1-5
Interorbital Breadth	**IOB**	1-3
Malar Tubercle	**MT**	0-3
Nasal Aperture Width	**NAW**	1-3
Nasal Bone Contour	**NBC**	0-4
Nasal Overgrowth**	**NO**	0-1
Postbregmatic Depression**	**PBD**	0-1
Supranasal Suture*	**SPS**	0-2
Transverse Palatine Suture*	**TPS**	1-4
Zygomaticomaxillary Suture*	**ZS**	0-2
Nasal Aperture Shape	**NAS**	1-3
Nasal Bone Shape*	**NBS**	1-4
Nasofrontal Suture*	**NFS**	1-4
Orbital Shape*	**OBS**	1-3
Posterior Zygomatic Tubercle	**PZT**	0-3

the updated descriptions by Hefner and Linde (2018). She also provided one observer, JM, with the descriptions and updated descriptions for data collection. The author's advisor, EH, was familiar with Hefner (2009) and was not provided with the updated descriptions, thus EH used descriptions and line drawings from Hefner (2009).

3.1.3a Anterior Nasal Spine (ANS)

The trait ANS has been described in the previous chapter; please see pages 59-60 and Figure 1 to review the details.

3.1.3.b Inferior Nasal Aperture (INA)

Description

Hefner (2009, p. 987) defined INA as "the most inferior portion of the nasal aperture, which, when combined with the lateral alae, constitutes the transition from nasal floor to the vertical portion of the maxillae, superior to the anterior dentition." This is unchanged in Hefner and Linde (Hefner & Linde, 2018, p. 25).

How to score

Hefner says that if (2009, p. 987) "bilateral asymmetry occurs, the left side is used**"** while Hefner and Linde (2018, p. 25) say that "the entire region is considered unless there is bilateral asymmetry present. In that case, the side with the highest expression is scored." Hefner and Linde (2018, p. 31) also state that "the lateral grooves along the most inferior aspect of the nasal aperture (subnasal grooves) should not be considered when assessing INA."

Trait expressions (Figure 2)

INA is an assessment of the shape of the inferior border of the nasal aperture:

1. *An inferior sloping of the nasal floor which begins within the nasal cavity and terminates on the vertical surface of the maxilla, producing a smooth transition. The*

morphology is distinct from INA 2 regarding the more posterior origin and the greater slope of INA 1.

2. *A sloping of the nasal aperture beginning more anteriorly than in INA 1, and with more angulation at the exit of the nasal opening*
3. *The transition from nasal floor to the vertical maxilla is not sloping, nor is there an intervening projection, or sill. Generally, this morphology is a right angle, although a more blunted form may be observed.*
4. *Any superior incline of the anterior nasal floor, creating a weak (but present) vertical ridge of bone that traverses the inferior nasal border (partial nasal sill).*
5. *A pronounced ridge (nasal sill) obstructing the nasal floor-to-maxilla transition*. (Hefner, 2009, p. 987)

Hefner and Linde (2018, p. 25) updated the descriptions to include more details:

INA is the assessment of the shape of the inferior border of the nasal aperture just lateral to the anterior nasal spine, which defines the transition from the nasal floor to the vertical portion of the maxillae. The morphology of INA ranges from an inferior slope with no delineation of the inferior border (1) to a sharp, vertical ridge of bone, or nasal sill (5).

1. *Has a marked slope of the nasal floor, which starts inside the nasal cavity and terminates on the vertical surface of the maxilla. This gradual slope is a smooth transition between two anatomical areas.*
2. *Has a moderate slope of the nasal floor to the vertical surface of the maxilla. To differentiate from INA, the slope begins more anteriorly in the nasal aperture (more anterior to the insertion of the vomer) and exhibits more angulation at the opening.*

3. *Has an abrupt transition from the nasal floor to vertical maxilla. There is no slope or intervening projection of bone (sill). This morphology resembles a right angle, although a blunted form may present.*
4. *Has any weak ridge of bone that crosses the anterior nasal floor perpendicularly, resulting in a partial nasal sill.*
5. *Has a pronounced ridge (nasal sill) obstructing the nasal floor-to-maxilla*

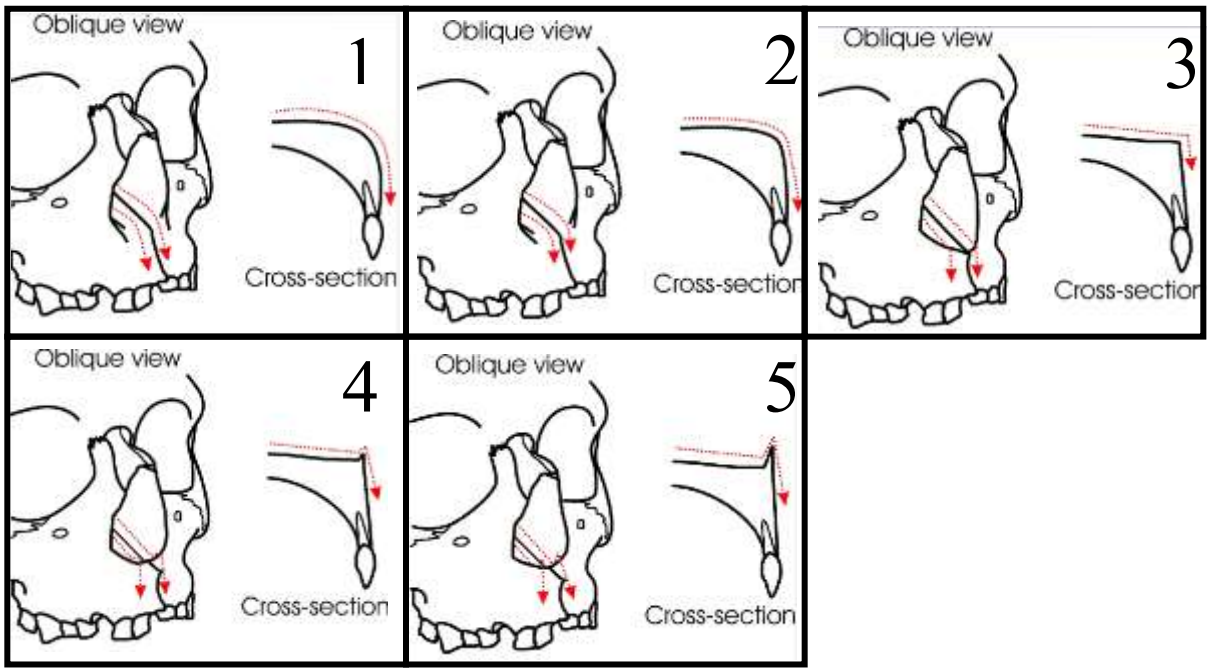

Figure 2: Trait expressions for INA (red dashed arrows). Scores for each expression are indicated by values 1-5. Red arrows are showing the contour of the interior nasal aperture floor. Anterolateral view of the skull. Images taken from *Osteoware* (2020).

transition. (Hefner & Linde, 2018, p. 29)

3.1.3c Interorbital Breadth (IOB)

Description

No description for IOB was found in Hefner (2009). Hefner and Linde (2018, p. 45) say "the relative distance between the orbits."

How to score

Hefner (2009, p. 988) says "this assessment is made relative to the facial skeleton." Hefner and Linde (2018, p. 45) improve the instructions to say that "IOB is the distance between the two orbits in the region of dacryon, relative to the overall breadth of the facial skeleton (from right *frontomalare temporale* to left *frontomalare temporale*)."

Trait expressions (Figure 3)

Hefner (2009, p. 988) originally defined IOB expressions as:

1. *A narrow IOB.*
2. *An intermediate IOB.*
3. *A broad IOB.*

Hefner and Linde (2018) added in ratios to update the description:

1. *Is approximately 1:5 ratio of the interorbital space relative to facial skeleton.*
2. *Is approximately 1:4 ratio of the interorbital space relative to facial skeleton.*
3. *Is approximately 1:3 ratio of the interorbital space relative to facial skeleton*.

(Hefner & Linde, 2018, p. 49)

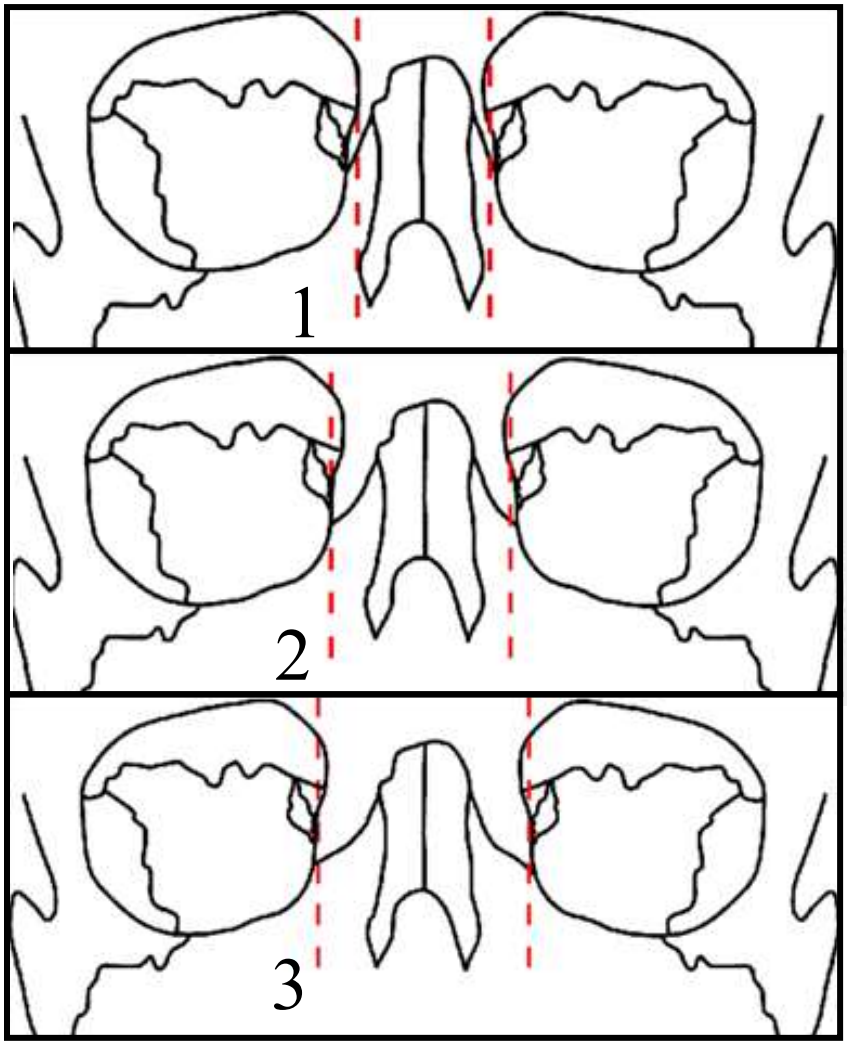

Figure 3: Trait expressions for IOB (red dashed lines). Scores given to each expression are indicated by the values 1-3. Red dashed lines indicate the trait IOB. Anterior view of the skull. Images taken from *Osteoware* (2020).

3.1.3d Malar Tubercle (MT)

Description

"The MT is a caudally protruding tubercle located on the inferior margin of the maxilla and zygomatic bone in the region of the ZS (Hefner, 2009, p. 988)." Hefner and Linde (2018, p. 57) add that it is a "caudally protruding tubercle which may occur on the maxilla, zygomatic, or at their intersection along the inferior aspect of the ZS."

How to score

MT is scored following Hauser and De Stefano (21), who recommend placing a transparent ruler at the intersection of the ZS and the inferior margin of the malar to the deepest incurvature of the maxilla. An assessment is then made on the extent of protrusion beyond the ruler's edge. In instances where the suture is directly on the tubercle, the ruler is placed from the deepest curvature of the maxilla to the deepest anterior curvature on the zygomatic. It should be noted that a MT may be present on the maxilla, the zygomatic, or along the ZS. Observers should not consider the tubercles on the lateral zygomatic arch. A completely absent MT is rare. (Hefner, 2009, p. 988; Hefner & Linde, 2018, p. 57)

Hefner and Linde (2018, p. 63) add that "Slight asymmetry is typical. The side with the greatest expression is recorded." Additionally, "rugosity marking an inferior aspect of the zygomaticomaxillary region is not considered when scoring MT (Hefner & Linde, 2018, p. 70)."

Trait expressions (Figure 4)

0. *No projection of bone.*
1. *A trace tubercle below the ruler's edge (approximately 2 mm or less).*
2. *A medium protrusion below the ruler's edge (approximately 2 to 4 mm).*
3. *A pronounced tubercle below the ruler's edge, more than 4mm below the ruler's edge* (Hefner, 2009, p. 988; Hefner & Linde, 2018, p. 61)

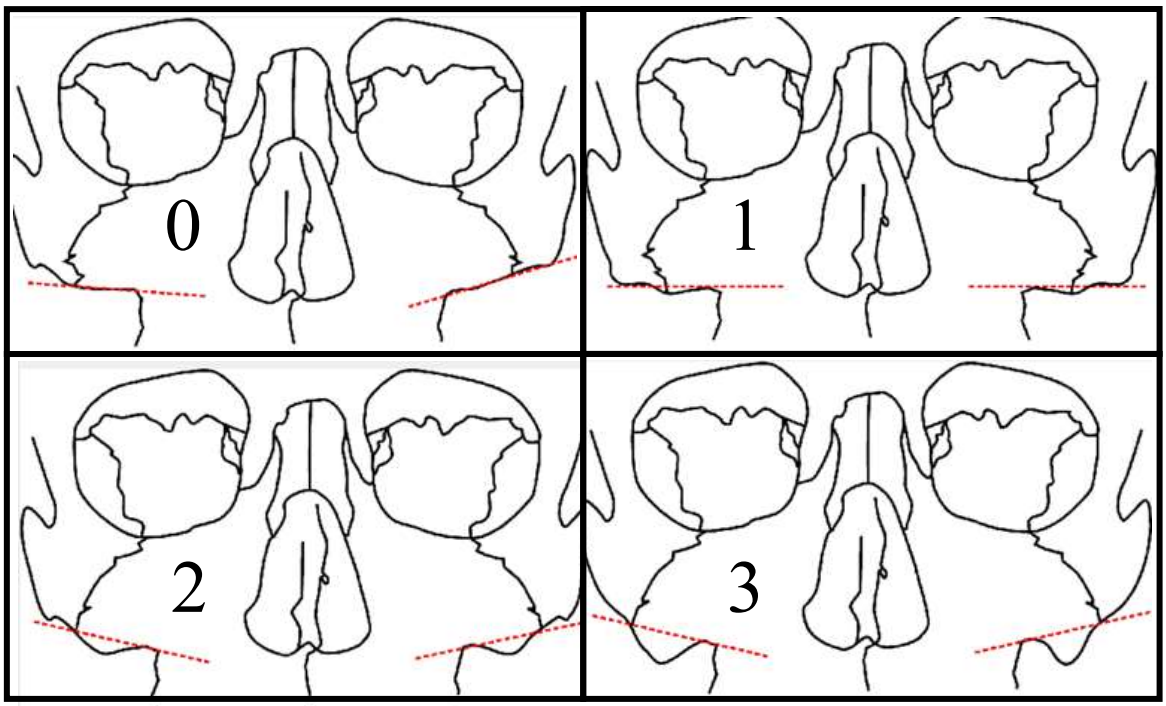

Figure 4: Trait expressions for MT as indicated by the bone extending below the red dotted lines. Scores given to each expression are indicated by the values 0-3. Anterior view of the skull. Images taken from *Osteoware* (2020).

3.1.3e Nasal Aperture Width (NAW)

Description

No description is found in Hefner (2009). Hefner and Linde (2018, p. 89) define NAW as the consideration of "the width of the nasal aperture relative to the entire facial skeleton using a ratio of total facial width to nasal opening."

How to score

Hefner (2009, p. 988) says that "the width of the nasal aperture is assessed relative to the facial skeleton." Hefner and Linde (2018, p. 89) add that it is "scored by viewing the cranium anteriorly and dividing the midfacial region into fractions."

Trait expressions (Figure 5)

Hefner (2009, p. 988) originally defined NAW expressions as:

1. *A narrow NAW.*

2. *An intermediate NAW.*
3. *A broad NAW*.

Hefner and Linde (2018) added in ratios to update the description stating:

To more objectively distinguish character states of NAW, the relative breadth of the nasal aperture should be considered. When assessing each character state, analysis should consider how much of the total facial skeleton the nasal cavity comprises. If, for example, the nasal cavity makes up roughly one-quarter of the upper facial skeleton, the region should be scored medium. (Hefner & Linde, 2018, p. 89)

1. *The nasal aperture takes up 1/5 of the facial skeleton.*
2. *The nasal aperture takes up 1/4 of the facial skeleton*
3. *The nasal aperture takes up 1/3 of the facial skeleton*. (Hefner & Linde, 2018, p. 90)

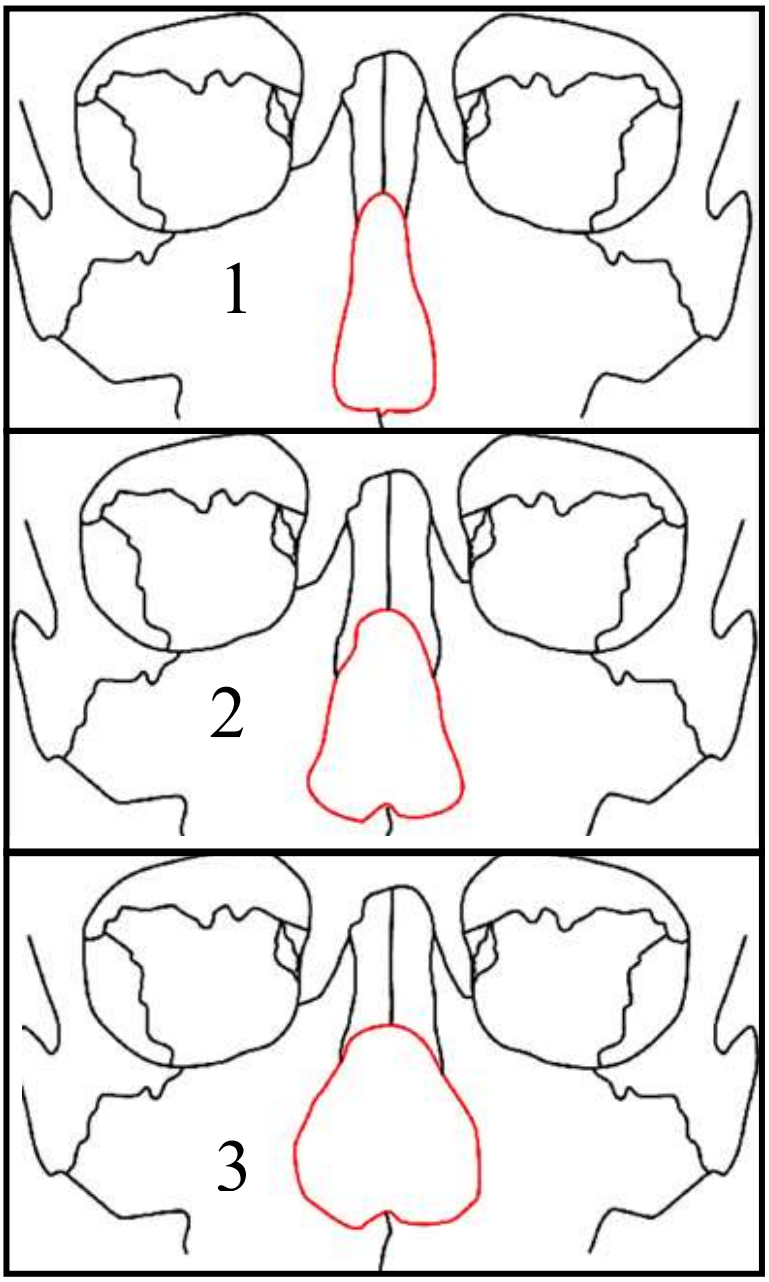

Figure 5: Trait expressions for NAW as indicated by the red lines. Score given to each expression is indicated by the values 1-3. Anterior view of the skull. Images taken from *Osteoware* (2020).

3.1.3f Nasal Bone Contour (NBC)

Description

Hefner (2009, p. 988) describes NBC as "the contour of the midfacial region (particularly the contour of the nasal bones and the frontal process of the maxilla) c. 1 cm below nasion." This is not changed in Hefner and Linde (2018, p. 105).

How to score

Visual interpretation of nasal contour is not the most effective manner of analysis due to high inter- and intra-observer error. The use of a contour gage permits a more reliable and consistent assessment of nasal contour (Fig. 9). To assess NBC, the cranium is placed in a position that allows the observer to gently, but with consistent and balanced pressure, place the contour gage directly on the nasal bones c. 1 cm inferior to nasion, while maintaining the gage roughly perpendicular to the palate and parallel to the orbits. (Hefner, 2009, p. 989).

Hefner and Linde (2018, p. 105,106) add that:

Most observations require approximately 20mm at the deepest point on the contour gauge for proper assessment, although a lower or higher midface may require adjustments. Realign needles after taking each contour, misalignment may result in an incorrect measure.

*Trait expressions (**Figure 6)***

0. *Is low and rounded.*
1. *Is an oval contour, with elongated, high, and rounded lateral walls; NBC 1 presents a circular shape and lacks steep walls.*

2. *Has steep lateral walls and a broad (roughly 7 mm or more), flat superior surface ‘‘plateau,’’ noted on the contour gage as a flat cluster of needles in the midline.*
3. *Has steep-sided lateral walls and a narrow superior surface “plateau.”*
4. *Has a triangular cross section lacking a superior surface “plateau.”* (Hefner, 2009, p. 989)

Hefner and Linde (2018, p. 109) add to expression 2’s description with “a flat cluster of four or more needles.”

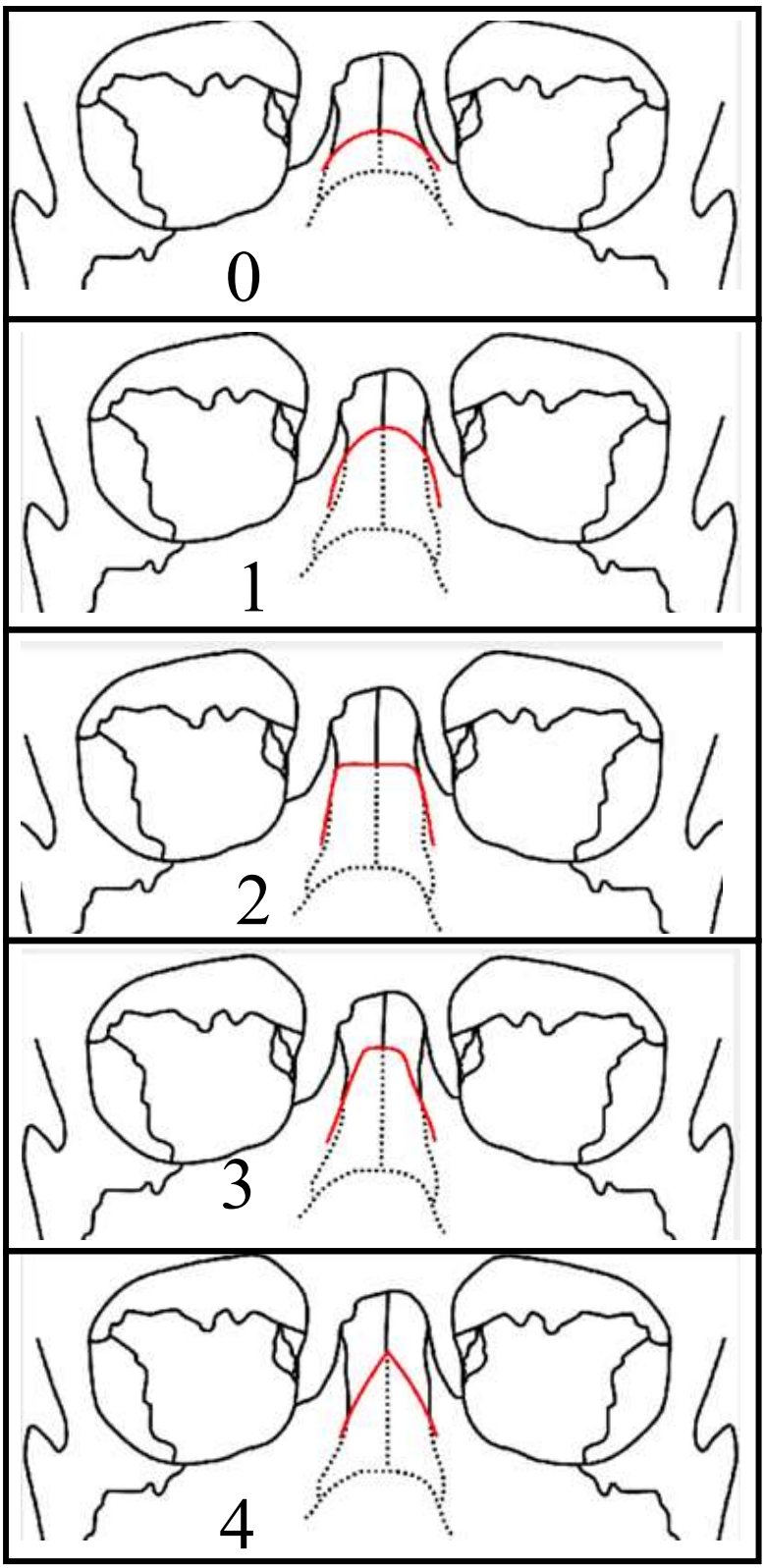

Figure 6: Trait expressions (red lines) for NBC. Score given to each expression is indicated by the values 0-4. Anterior view of the skull. Images taken from *Osteoware* (2020).

3.1.3g Nasal Overgrowth (NO)

Description

Hefner (2009, p. 989) defines NO as "an inferior projection of the lateral border of the nasal bones beyond the maxillae at *nasale inferious*." Hefner and Linde (2018, p. 143) reword it as: "NO is a projection of the lateral border or the nasal bones at their inferior edge, extending beyond the maxilla at the bony landmark '*nasale inferious*.'"

How to score

According to Hefner (2009, p. 989):

assessment of NO does not include anterior bulging of the nasal bones. Observations should be made on the left side. If the left side is damaged, the right side may be substituted. If both nasal bones are missing or fractured (anti-, peri-, or postmortem), the trait is not scored. It is often useful to run a finger along the borders of the maxilla and nasal bones near nasale inferious to determine whether a projection is present.

Hefner and Linde (2018, p. 143) say:

The trait is visualized by close inspection of the inferior lateral border where it meets the maxilla… Gently running your finger along the border of the maxilla and nasal bones at 'nasale inferious' may help to determine whether a projection is present… If asymmetry is present, score the side with the highest expression. If both nasal bones are missing, fractured, or damaged, NO should not be recorded. NO is only an assessment of the nasal bones. Any projection or extension of the maxillae should not be considered as an expression of NO.

Additionally, “even though the inferior aspect of the nasal bone can look separated from the maxilla, there is no extension beyond, (Hefner & Linde, 2018, p. 150)” thus not scored as present.

Trait expressions (Figure 7)

According to Hefner (2009, p. 989):

0. *Is absent*
1. *Is present*

According to Hefner and Linde (2018, p. 147):

0. *Has no bony overgrowth at nasale inferious.*
1. *Has any projection of bone from the lateral border of the nasals beyond the maxillary border at nasale inferious.*

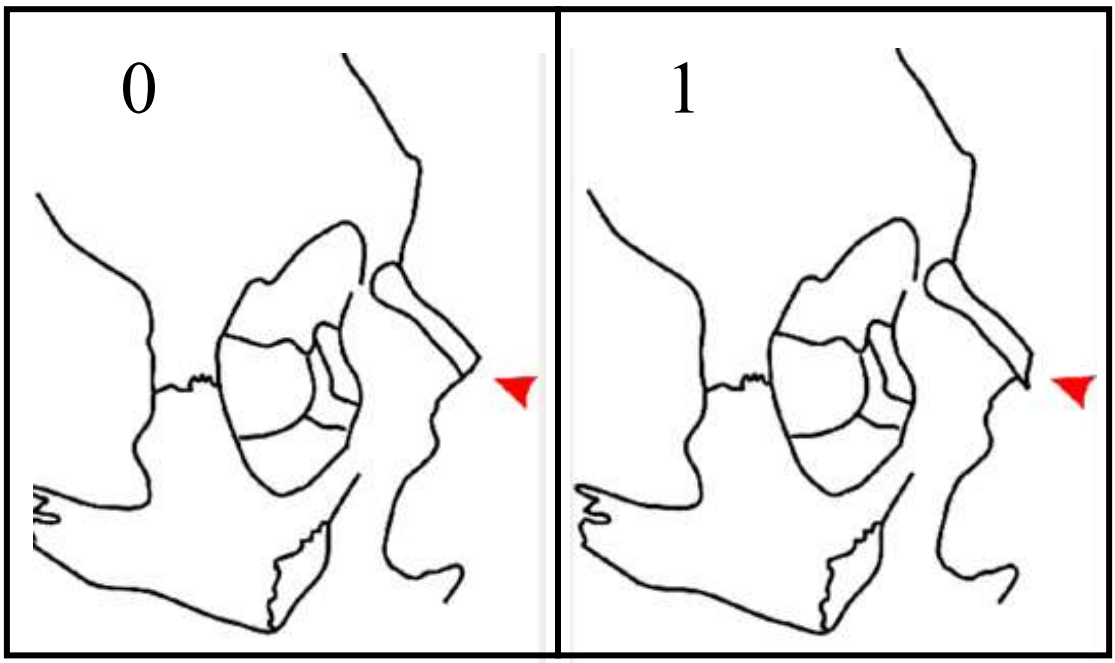

Figure 7: Trait expressions for NO with the red arrow indicating where the bone extends (right) and where it does not (left). Scores given to each expression are indicated by the values 0-1. Lateral view of the skull. Images taken from *Osteoware* (2020).

3.1.3h Post Bregmatic Depression (PBD)

Description

Hefner (2009, p. 989) defines PBD as a "slight to broad depression along the sagittal suture, posterior to *bregma* that is not the result of pathology (e.g., premature synostosis)." This is unchanged in Hefner and Linde (2018, p. 209).

How to score

Hefner (2009, p.989) says to "observe in lateral profile." Hefner and Linde (2018, p. 209) add "to score PBD, hold the cranium in a lateral profile view and look for a depression posterior to *bregma.* It may be helpful to palpitate the area."

***Trait expressions* (*Figure 8*)**

Hefner (2009) originally defines:

0. *Is absent.*
1. *Is present.*

Hefner and Linde (2018, p. 213) add:

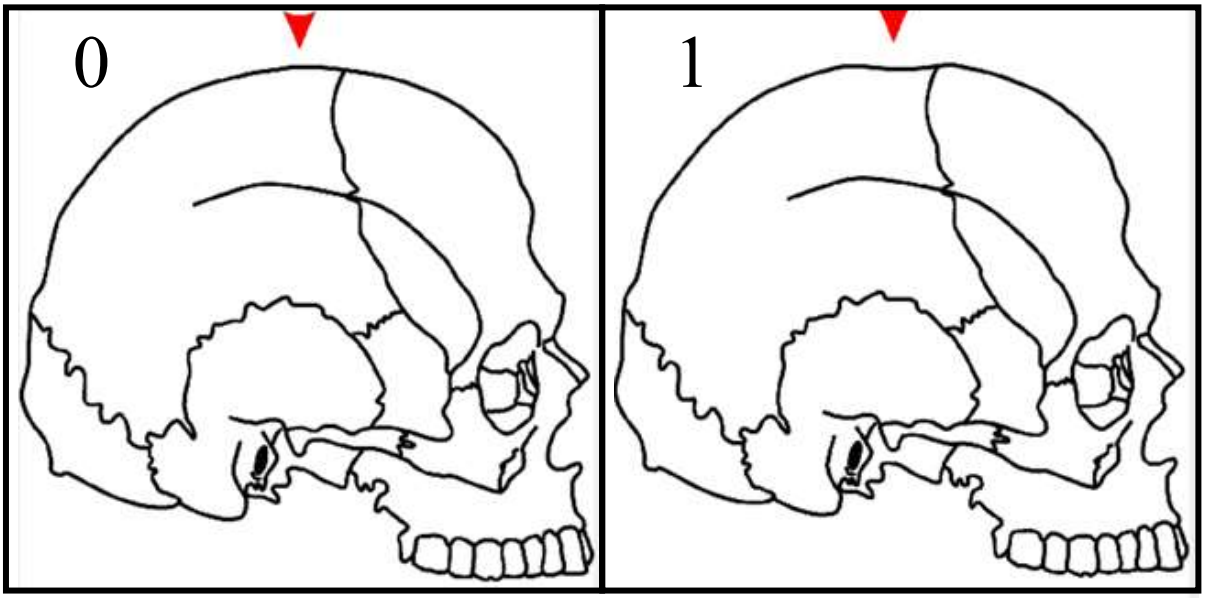

Figure 8: Trait expressions for PBD as indicated by the red arrow where the curvature looks different. Scores given to each expression is indicated by the values 0 and 1 Lateral view of the skull. Images taken from *Osteoware* (2020).

0. *Is absent; no depression posterior to bregma.*
1. *Is present; any depression posterior to bregma along the midsagittal plane.*

3.1.3i Supranasal Suture (SPS)

Description

Hefner (2009, p.989,990) defines the SPS as:

A secondary complex suture may persist, which is generally referred to as the supranasal suture (SPS), or sutura supranasalis. This suture does not represent the nasal portion of a persistent metopic suture, which is usually a single, non-oscillating line. The SPS is the fusion of the nasal portion of a frontal suture that appears as a complex of interlocking bony spicules at 'glabella'.(Hefner, 2009, p. 989,990)

Hefner and Linde (2018, p. 245) add that this "sutural complex is superior to the cranial landmark 'nasion' that may persist into adulthood"

How to score

No description for how to score is found in Hefner (2009), but Hefner and Linde (2018, p. 245) say it is "scored by viewing the cranium anteriorly."

Trait expressions (***Figure 9***)

0. *Is completely obliterated.*
1. *Is open (unfused).*
2. *Is closed, but visible*. (Hefner, 2009, p. 990)

Hefner and Linde (2018, p. 245) add to SPS2 with "although SPS is closed, it remains distinctly visible."

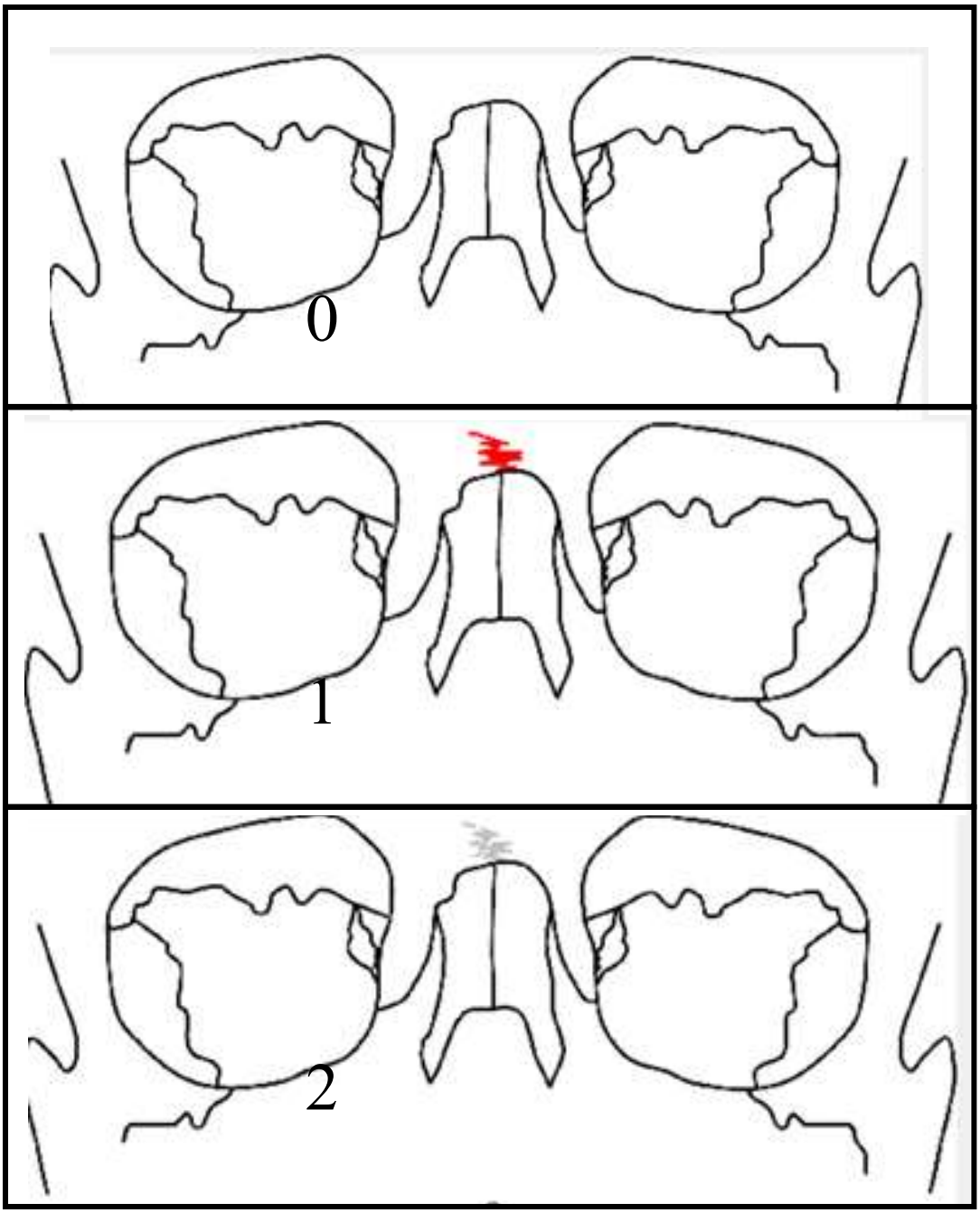

Figure 9: Trait expressions for SPS as indicated by the red and grey lines. Scores given to each expression is indicated by the values 0-2. Anterior view of the skull. Images taken from *Osteoware* (2020).

3.3.1j Transverse Palatine Suture (TPS)

Description

No description for TPS was found in Hefner (2009), however, Hefner and Linde (2018, p. 257) define it as "the course of the TPS on the hard bony palate."

How to score

The course of the TPS (Fig. 13) is highly variable, although certain themes persist. TPS is not scored unilaterally, although asymmetrical sutures are not uncommon. The entire suture is observed, but the medial one-half in the region of the palatine suture is most closely scrutinized. When an asymmetrical suture is present (the two branches of the suture do not come into contact at midline) the general theme is recorded (e.g., straight or jagged). Slight undulations of the suture should not be considered when making a determination. If the suture is obliterated, it is not scored. Hefner (2009, p. 990).

Hefner and Linde (2018, p. 257) add:

To assess this trait, view the cranium inferiorly, at the hard palate. Follow the transverse suture and note if, and how, the deviation occurs near the intersection wit the median palatine suture.

Trait expressions (***Figure 10***)

1. *The suture crosses the palate perpendicular to the median palatine suture with no significant anterior or posterior deviations. If the right and left halves of the suture do not contact each other at midline, but the suture is otherwise straight, score the suture 1.*
2. *The suture crosses the palate perpendicular to the median palatine suture, but near this juncture a significant anterior deviation, or bulging, is present.*

3. *The suture crosses the palate, but deviates anteriorly and posteriorly (e.g., M-shaped) in the region of the median palatine suture. Suture may appear similar to an EKG reading. If the right and left halves of the suture do not contact each other, but the suture is otherwise jagged, a score of 2 is used.*
4. *The suture crosses the palate perpendicular to the median palatine suture, but near this juncture a posterior deviation, or bulging, is present.* (Hefner, 2009, p. 990)

The only difference with Hefner and Linde (2018, p. 260) is that they remove the second sentence of TPS1, and second and third sentence in TPS3.

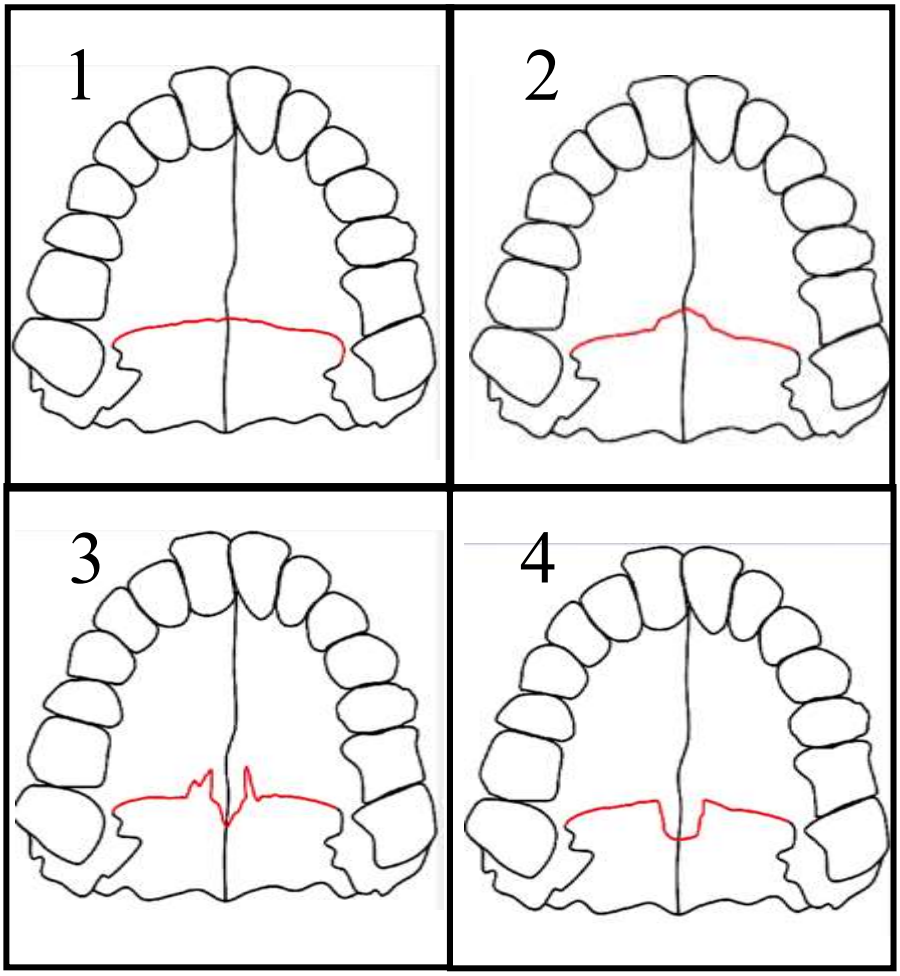

Figure 10: Trait expressions (red lines) for TPS. Scores given to each expression are indicated by the values 1-4. Inferior view of the skull. Images taken from *Osteoware* (2020).

3.1.3k Zygomaticomaxillary Suture (ZS)

Description

Hefner (2009, p. 990) describes ZS as "is the suture between the maxilla and the zygomatic." Hefner and Linde (2018, p. 275) reword it as "the course of the suture between the maxilla and the zygomatic."

How to score

The course of the suture is best observed in the anterior view. In instances of asymmetrical manifestations, the left side is preferred. The infraorbital suture should be ignored when making a determination. Assessment of ZS is based primarily on the approximate location of greatest lateral projection of the suture, and also on the number of major angles present. Hefner (2009, p. 990).

Hefner and Linde (2018, p. 275) add that "when asymmetry occurs, the side presenting the highest expression is recorded. The infraorbital suture should be ignored when making a determination."

*Trait expressions (**Figure 11**)*

0. *Is a suture having the greatest lateral projection at the inferior margin, but a slight angle near the midpoint of the suture.*
1. *Is a suture with one angle and greatest lateral projection near the midline.*
2. *Is a suture with two or more angles (presenting a jagged and/or S-shaped appearance) and a variable position for greatest lateral projection*. (Hefner, 2009, p. 990)
3. Is not defined by Hefner (2009), but an image is shown where the suture is not visible.

Hefner and Linde (2018, p. 279) clarify ZS1 that it is “at the inferior margin of the malar, with no angles,” and do not include expression 3.

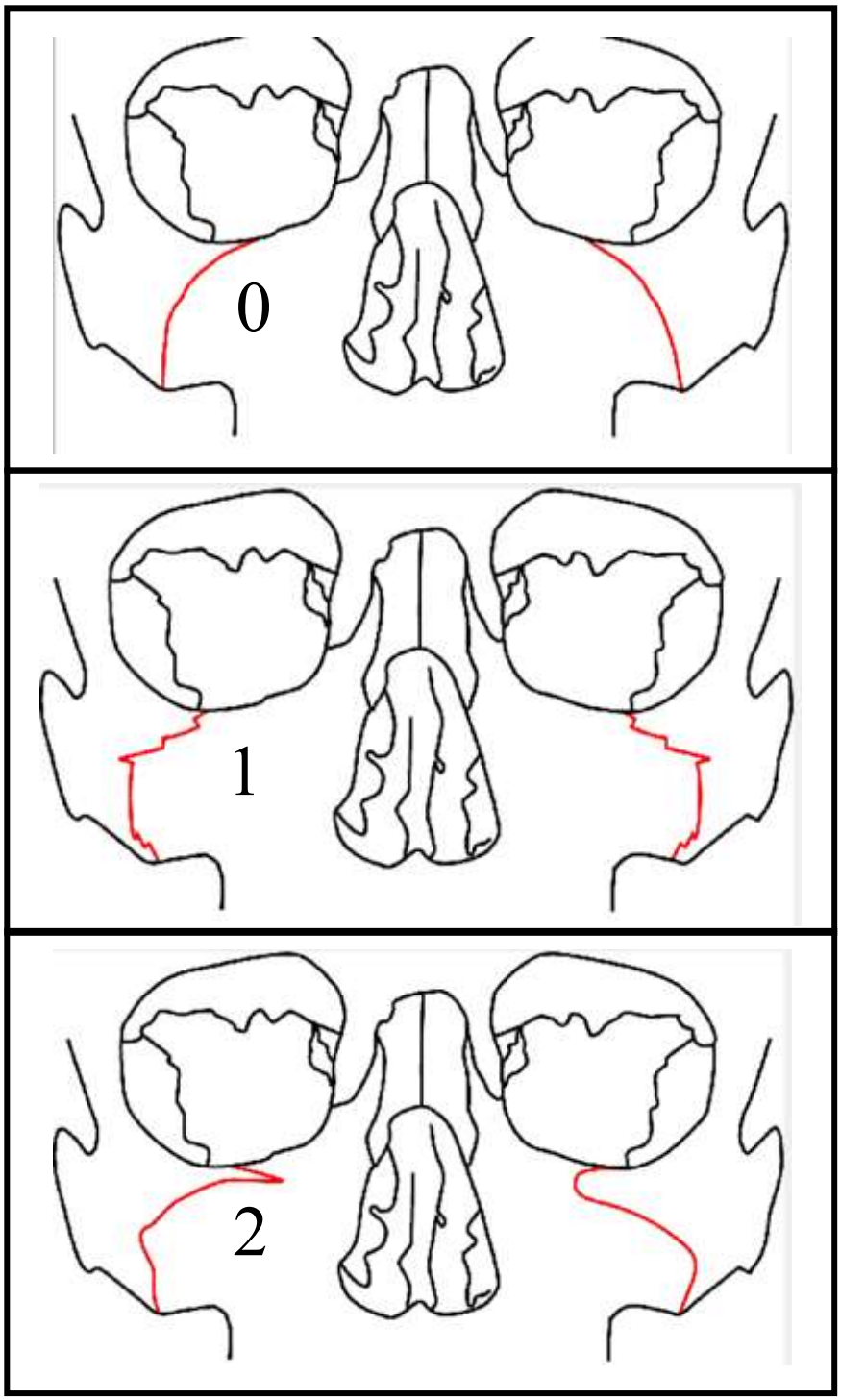

Figure 11: Trait expressions (red lines) for ZS. Scores given to each expression are indicated by the values 0-2. Anterior view of the skull. Images taken from *Osteoware* (2020).

3.1.3l Nasal Aperture Shape (NAS)

Description

No description for NAS was found in Hefner (2009) because it was not included in the article as of yet. *Osteoware* also does not include a description, but does include how to score the trait expressions. Hefner and Linde (2018, p. 73) include a description as such:

> *NAS considers the lateral contours of the nasal opening and the position of the greatest projection along the lateral margins, regardless of overall breadth of the nasal opening. The relative shape of the nasal aperture is defined by the position of greatest lateral projection (The position of the left and right bony landmark 'alare')*

How to score

"The shape of the nasal aperture (NAS) is assessed by observing 1) the lateral contours of the nasal aperture and, directly related, 2) the position of greatest lateral projection of the margin (*Osteoware,* 2020)." Hefner and Linde (2018, p. 73) add to this with the "bilateral asymmetry is not uncommon, but when it does occur, the side presenting the highest expression is recorded."

Trait expressions (***Figure 12)***

Osteoware (2020) originally defined the expressions as:

1. *Teardrop: has a lateral projection intermediate to 2 and 3 (see illustration).*
2. *Bell: has the greatest lateral projection at the inferior margin.*
3. *Bowed: has the greatest lateral projection at midline.*

Hefner and Linde (2018, p. 76) redefined them as:

1. *Teardrop: has the greatest projection of the lateral margin located in the lower 2/3 of the nasal aperture, superior to the location of NAS2 and inferior to the location of NAS3.*

2. *Bell: has the greatest projection of the lateral margin located at the inferior aspect of the nasal aperture.*
3. *Bowed: has the greatest projection of the lateral margin located at the midline of the nasal aperture.*

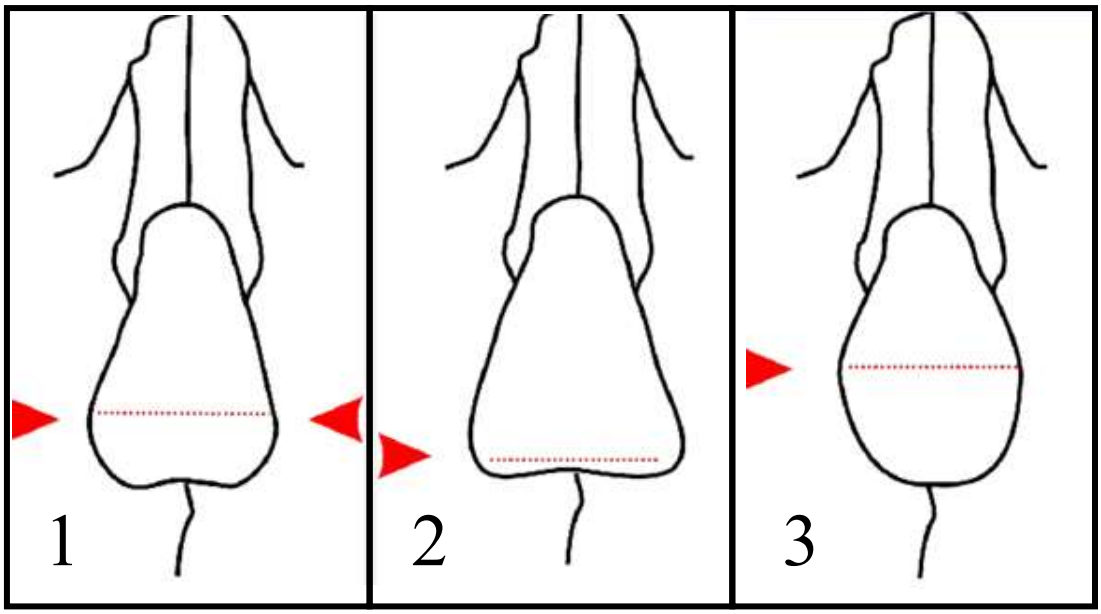

Figure 12: Trait expressions for NAS indicated by the red dotted lines and arrows showing where the greatest lateral margins occur. Scores given to each expression are indicated by the values 1-3. Anterior view of the nasal aperture. Images taken from *Osteoware* (2020).

3.1.3m Nasal Bone Shape (NBS)

Description

No description for NBS was found in Hefner (2009) because it was not included in the scoring method at this time. *Osteoware* also does not include a description. Hefner and Linde (2018, p. 125) say that "it is the contour of the nasal bones along the lateral borders."

How to score

Nasal bone shape (NBS) is assessed from the anterior view with the cranium positioned in approximate anatomical position. A determination is made regarding 1) the position of nasal pinch, if any, and 2) the amount of lateral bulging. While making the assessment, the

observer should not consider the frontonasal suture, the nasal suture, or the symmetry of the nasal bones. Rather, an assessment is made of the lateral contours of the nasal bones. (*Osteoware*, 2020).

Hefner and Linde (2018, p. 125) add to this to also "consider the position of the nasal pinch (if present) and the degree of lateral bulging (if present)."

Trait expressions* (*Figure 13)

1. *Has no nasal pinch. The nasal bones may be wide or narrow.*
2. *Has the nasal bones with a superior pinch and minimal lateral bulging. Note: To differentiate between a score of 2 and 3, the amount of lateral bulging in the inferior region should be assessed.*
3. *Has the nasal bones with a superior pinch and pronounced lateral bulging of the inferior region.*
4. *Has triangular-shaped nasal bones (Osteoware*, 2020).

Hefner and Linde (2018, p. 129) update the descriptions for clarity:

1. *The lateral edges of the nasal bones are not pinched and may be wide or narrow.*
2. *The nasal bones exhibit a superior pinch and only minimum bulging along the lateral edges.*
3. *The nasal bones exhibit a superior pinch and pronounced lateral bulging along the inferior margin.*
4. *The lateral edges of the nasal bones form a triangle*.

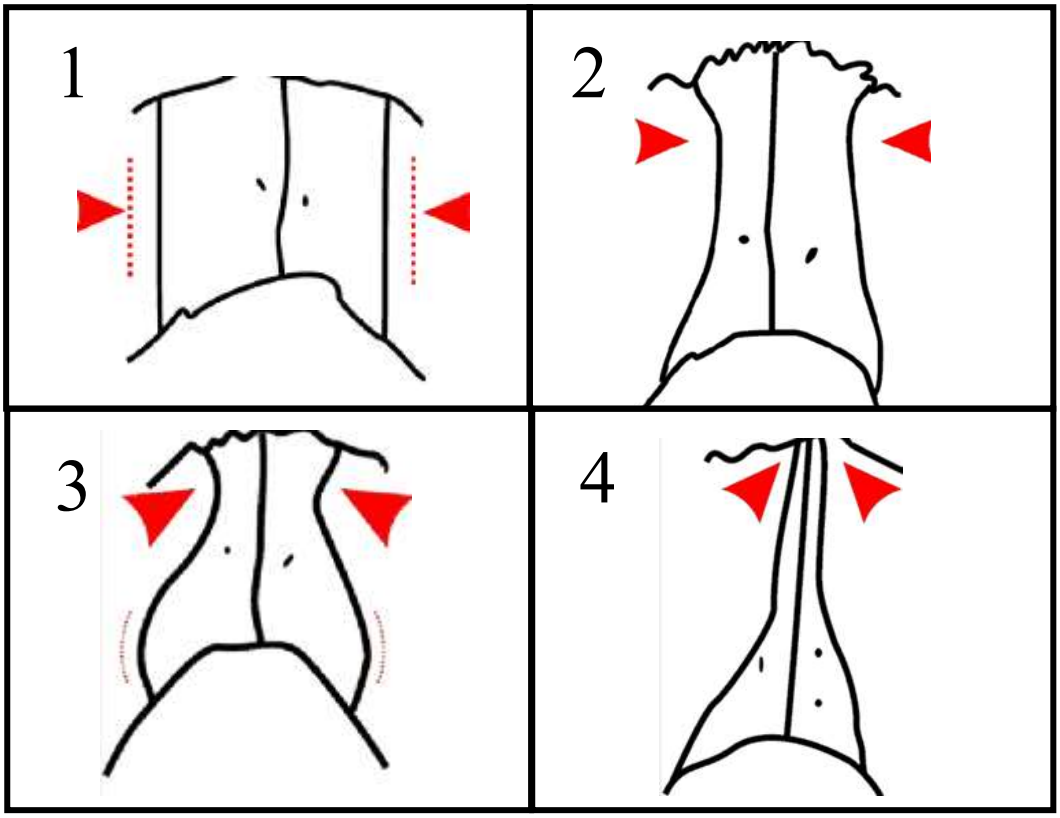

Figure 13: Trait expressions for NBS as indicated by the red dotted lines and arrows showing how the suture line changes shape at different spots. Scores given to each expression is indicated by the values 1-4. Anterior view of the nasal bones. Images taken from *Osteoware* (2020).

3.1.3n Nasofrontal Suture (NFS)

Description

No description for NFS was found in Hefner (2009) because it was not included at this time. *Osteoware* (2020) defines the NFS as "the suture separating the nasal bones from the frontal bone." Hefner and Linde (2018, p. 155) define it as the suture that "connects the nasal bones to the frontal bone at the ethmoidal notch."

How to score

The shape of the suture is assessed. Nonmetric variants include: round, square, triangular, and irregular. Assessment is best made from the anterior view. The symmetry of the nasal bones should be ignored. If nasal bones evince extreme pinching of the superior border (as in NBS 4), observation should be left BLANK – unobservable (*Osteoware*, 2020).

Hefner and Linde (2018, p. 155) say:

To score NFS, hold the cranium in anatomical position, with the anterior portion of the cranium facing the observer. Do not assess NFS from a lateral view. When scoring, disregard the symmetry of the nasal bones. If nasal bones exhibit extreme pinching of the superior border, as it does in NBS4, do not score NFS.

Trait expressions (***Figure 14****)*

1. *Is round and lacks angles.*
2. *Appears square (approximate right angles at nasale superious).*
3. *Appears triangular.*
4. *Is irregular, lacking any definitive shape* (*Osteoware*, 2020; Hefner and Linde, 2018, p. 159).

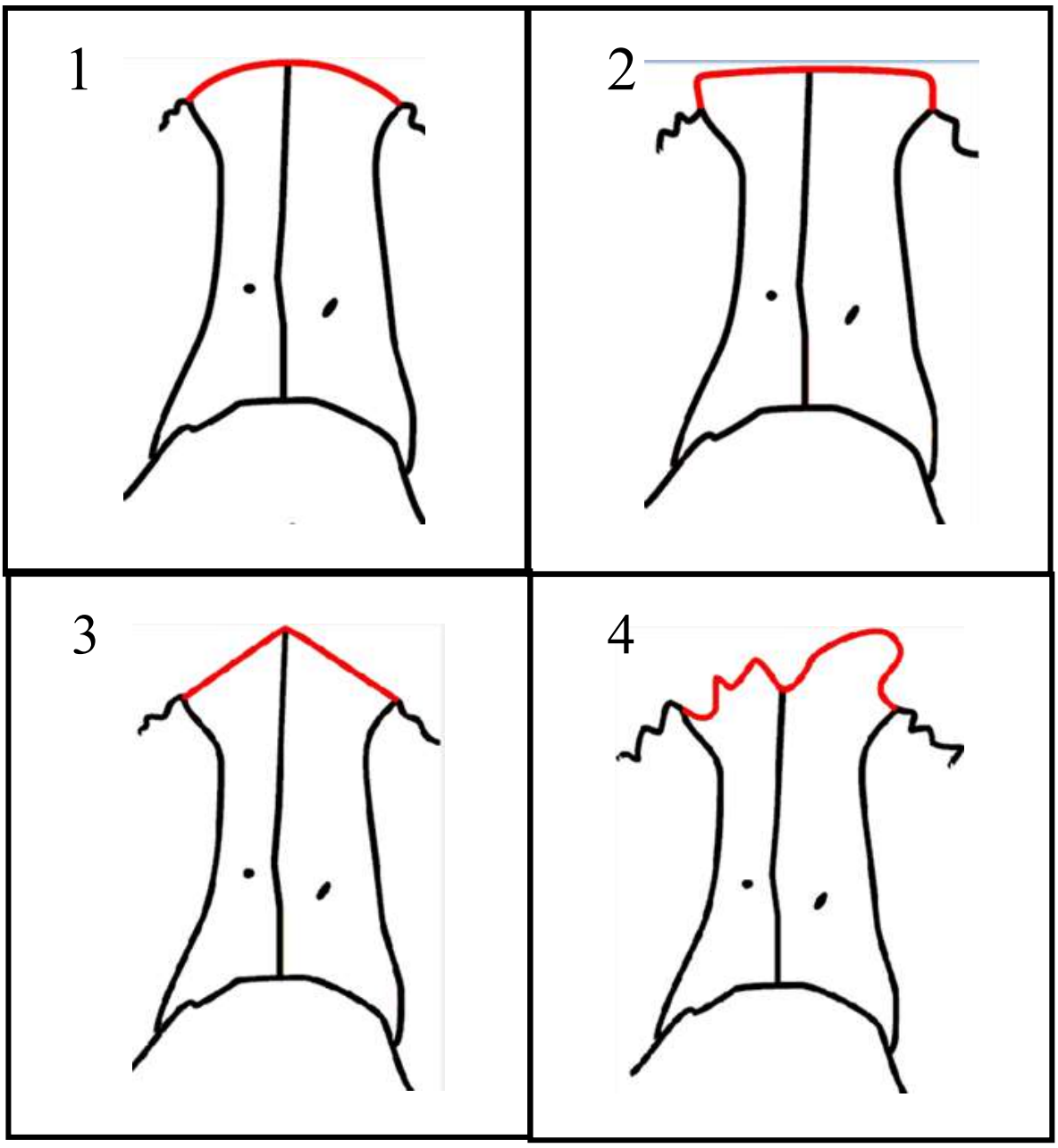

Figure 14: Trait expressions (red lines) for NFS. Scores given to each expression is indicated by the values 1-4. Anterior view of the nasal bones. Images taken from *Osteoware* (2020).

3.1.3o Orbital Shape (OBS)

Description

No description was found for OBS in Hefner (2009) because it was not included at this time. *Osteoware* also does not include a definition. Hefner and Linde (2018, p. 173) define OBS "as the shape of the bony sockets housing the eye."

How to score

Observation is best from the anterior view. The shape of the orbit is defined by the orbital margin of the superior, lateral, and inferior borders. The medial border of the orbit is defined by the anterior lacrimal crest and the maxillary process of the frontal bone. Observers should assess whether the margins are angular (rectangle), curvilinear (circular), or irregular (rhombic). Bilateral asymmetry may occur. All observations should be made from the left orbit (*Osteoware,* 2020).

This is unchanged in Hefner and Linde (2018, p. 173).

Trait expressions (***Figure 15****)*

1. *Has orbits with horizontal margins longer than the vertical margins, but otherwise parallel (i.e., rectangle).*
2. *Has the orbital margin approximately equidistant from center on all sides (i.e., circle).*
3. *Has the medial border height shorter than lateral border height (aviator sunglasses)* (*Osteoware*, 2020; Hefner and Linde, 2018, p. 177).

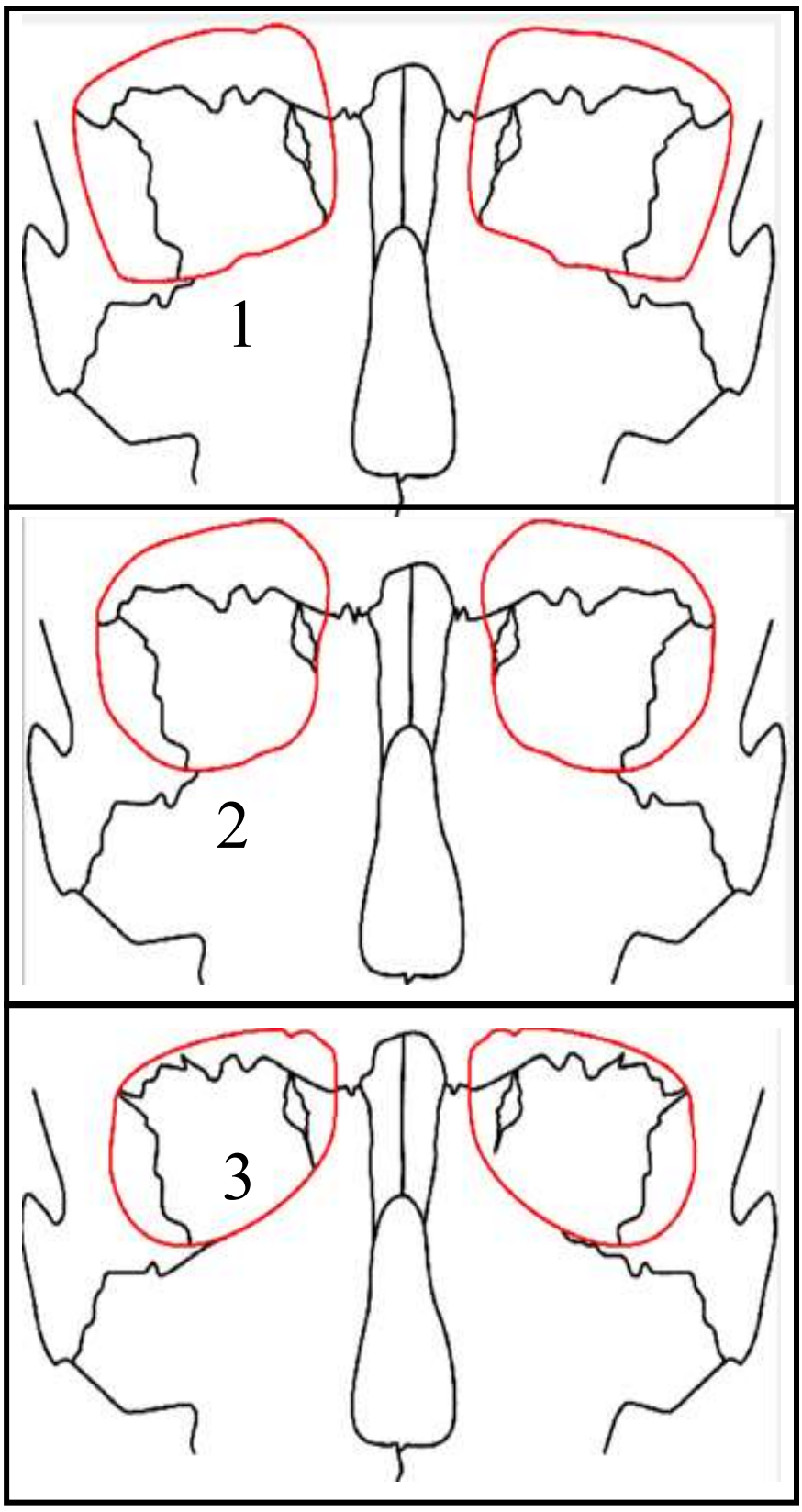

Figure 15: Trait expressions for OBS (red lines). Scores given to each expression is indicated by the values 1-3. Anterior view of the skull. Images taken from *Osteoware* (2020).

3.1.3p Posterior Zygomatic Tubercle (PZT)

Description

No description was found for PZT in Hefner (2009) because it was not included at the time. *Osteoware* (2020) defines PZT as "a posterior projection of the zygomatic bone at approximately midorbit as viewed in lateral plane." This is unchanged in Hefner and Linde (2018, p.223).

How to score

Osteoware (2020) says "To observe the various degrees of expression, a small, transparent ruler is placed on the frontal process of the zygomatic from the landmarks *frontomalare posterale* to *jugale*. The extent of bony protrusion beyond the ruler's edge is then assessed." This is unchanged in Hefner and Linde (2018, p.223).

***Trait expressions* (*Figure 16*)**

0. *Is no projection of bone past the ruler's edge.*
1. *Is a weak projection of bone (less than 4 mm).*
2. *Is a moderate projection of bone (approximately 4 to 6 mm).*
3. *Is a marked projection of bone (generally > 6 mm).* (*Osteoware,* 2020)

Hefner and Linde (2018, p. 227) clarify that the projection of bone is past the ruler's edge.

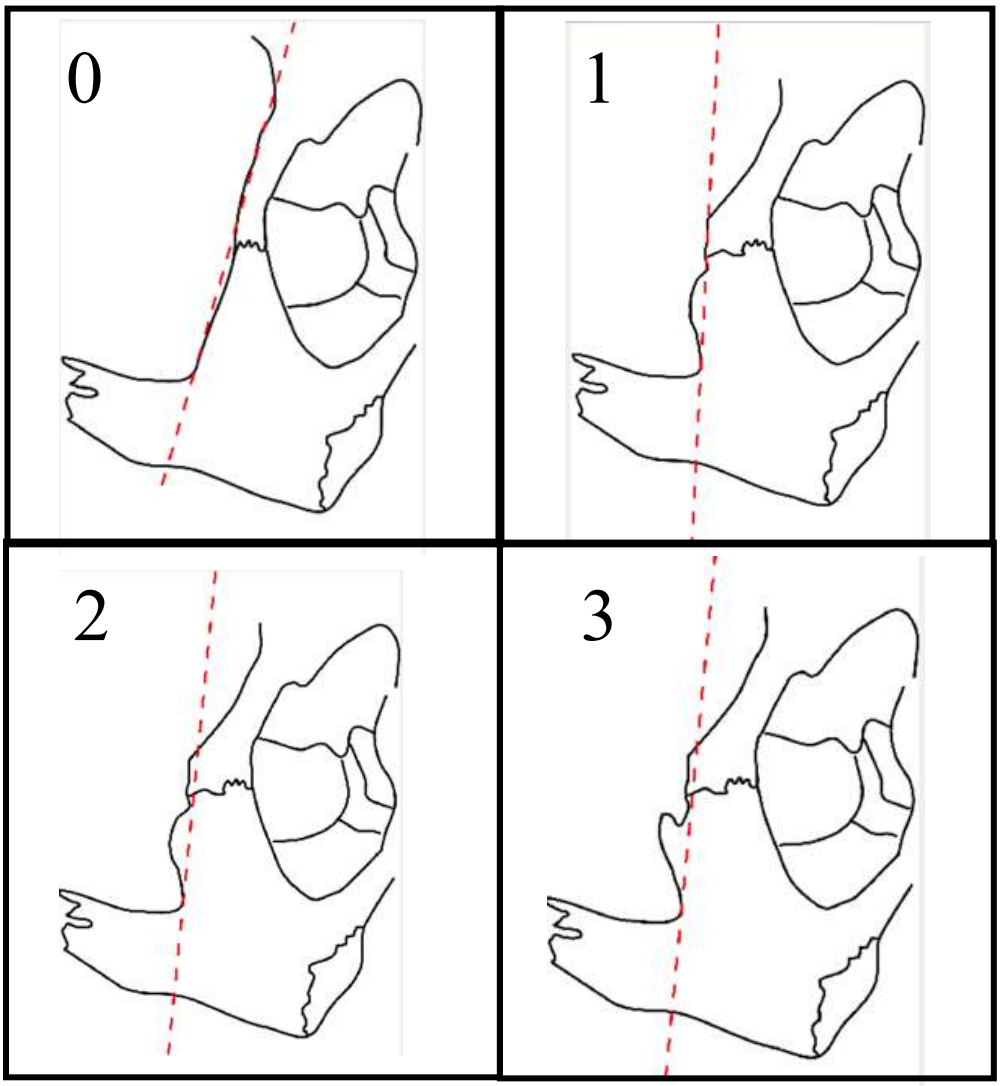

Figure 16: Trait expressions for PZT as indicated by the bone extending left past the red dashed line. Scores given to each expression is indicated by the values 0-3. Lateral view of the skull. Images taken from *Osteoware* (2020).

3.2 Methods

3.2.1 Training

Three observers collected trait expression data: the author and two other observers. The author, observer 1 (CM), is a master's student who has the second most experience with Hefner's scoring method and the variation of these expressions. This experience is an accumulation of studying the traits on casts, real individuals, and pictures from several publications (Hauser & De Stefano, 1989; Hefner, 2009; Hefner & Linde, 2018; Hefner et al., n.d.; *Osteoware v. 2.4.037*, 2020; Wilczak & Dudar, 2020). Other researchers suggest training prior to data collection to become familiar with the trait expressions (Kamnikar et al., 2018; Klales & Kenyhercz, 2015). Therefore, practice using Hefner's (2009; Hefner and Linde, 2018) scoring method and assessment equipment was carried out on the U of M individuals four months prior to the training session. The timing was to ensure there was ample opportunity to ask questions about any parts of a trait description during training. This session was labelled as 'untrained,' or 'scoring period 1' so it could be compared to the datasets after training (scoring periods 2 and 3) which was carried out under the author's advisor, observer 2 (EH).

Observer 2 is a trained forensic anthropologist who uses Hefner's scoring method, as outlined in his 2009 article, during cases; she has the most experience with this method and the variation of these expressions. Finally, observer 3 (JM) is a Master's graduate with the least experience using Hefner's scoring method. Observer 3 was trained in his Master's to assess nonmetric traits using fingers or pencils, such as outlined by Buikstra & Ubelaker (1994), to feel and measure. Observer 1 scored all sixteen traits found within *Osteoware* supplementing the assessment with description updates from the atlas (Hefner and Linde, 2018). These description updates from the atlas as well as figures from *Osteoware* were provided to observer 3, whereas

observer 2 used her own scoring sheets created from Hefner's (2009) article and scored eleven traits.

Prior to data collection, observer 2 trained both observers 1 and 3 on the morphoscopic traits used in Hefner's (2009; *Osteoware,* 2020) method. This training session included the completion of a morphoscopic analysis on one Brandon teaching individual from which observers' scores were compared to each other. Discussions revolved around how observer 2 scored the traits, and what expressions she saw and how these differed from other expressions. This ended up being useful for understanding observer error as well. After training, observer 3 practiced the method on BU individuals before collecting data on the U of M individuals, relying on the line drawings and updated descriptions provided to him.

In order to ensure traits were well understood, publications upon which Hefner's 2009 traits were based on were read (Hauser & De Stefano, 1989; Rhine, 1990). In addition, iterations of Hefner's method that occurred since the first publication were consulted to determine what steps were already taken to reduce subjectivity (Hefner, 2018; Hefner et al., n.d.; *Osteoware v. 2.4.037*, 2020; Wilczak & Dudar, 2020). These iterations included pictures of traits and their expressions on real skulls in each publication. Observing individuals with differing expressions provided the opportunity to see the possible range of expressions.

3.2.2 Data collection

3.2.2a Discrete data

All sixteen traits were visually assessed for each individual following the scoring method as outlined by *Osteoware,* or (Hefner & Linde, 2018). In order to assess traits, a skull was placed in the Frankfurt plane atop a soft cushion and positioned to limit readjustments; only one skull was assessed at a time and the left side was assessed, unless damaged or otherwise indicated in the

instructions. Then, the line drawings depicting each trait were placed next to the skull for visual comparison; the scoring method for visual trait assessment was housed in the program *Osteoware v 2.4.037* (2020) and was used for scoring.

The order of individual assessment and trait assessment for each individual were randomly assigned by rolling a twenty-sided die, with each number between one and sixteen corresponding to a trait or individual. A random order reduced the chance of remembering what each individual was previously scored. The individual's number/name, their trait scores, metric measurements and any other relevant information, such as drawing extra trait expressions, were recorded in a physical notebook. Scores and metric measurements were transferred into *Excel* spreadsheets for subsequent statistical analysis

All selected individuals had traits scored twice on separate occasions at least two months apart by the author in order to test for intra-observer error/agreement between scores. The U of M individuals were assessed three times: one when observer 1 was untrained, and two when observer 1 had training. Three scoring periods occurred because the first one was intended as a practice session to understand areas of confusion based on the published trait descriptions. The two trained scoring periods on the U of M individuals are used as comparable datasets to the two scoring periods with the BU individuals (both named scoring period 2 and 3).

For scoring period 1, only descriptions and line drawings from Hefner's (2009) article, also found in *Osteoware*, were used for scoring because access to atlas (Hefner & Linde, 2018), which contained updated descriptions, was not available. For scoring periods 2 and 3, these line drawings were used along with descriptions from the Hefner (2009) publication, and updated information from the Hefner & Linde (2018) atlas. Scoring period 2 also used the pictures in the atlas along with the line drawings and updated descriptions, which helped to compare the trait

expression observed to comparable expressions in the atlas. Since the atlas descriptions included landmarks for traits related to ratios, math was used to determine these scores even though it was not explicitly mentioned to do the math.

Scoring period 3 used just the updated descriptions and line drawings. This ensured the author was testing the repeatability of the scoring method itself and not relying on photographs. It also mimicked what would normally be available to forensic anthropologists (such as observer 2) because practitioners would rely on their notes and articles after an initial training session if it was available to them. If trait expressions were similar to those in the atlas during scoring period 1, figure numbers were written down for future reference. During scoring period 3, a subset of BU individuals was scored as normal, but then scored again using the atlas pictures to see if the score changed from the initial score. These changed scores were written in a separate column, but were not used for inter-observer comparison or ancestry estimation. This subset was helpful to determine if photographs depicting trait expression improve repeatability.

Overall, there were three data subsets for comparison for intra-observer agreement: one that compares scores taken while untrained to those taken after training; one that compares scores assessed with the help of atlas photos to scores assessed without the atlas photo (after training); and one that compares sets of scores that were both taken with the help of atlas photos (after training).

When a trait's expression was determined, the numerical score was documented in both a physical journal and an *Excel* spreadsheet next to the individual's number or name. Any traits unable to be scored or that had expressions not included in the scoring method were noted and drawn in the physical notebook. When a skull needed readjustment, it was picked up with both hands and carefully placed in the necessary position.

For some traits, a clear ruler or a metal pin contour gauge (*General* no. 837 angle-izer) with thirty-five, 1.4mm pins per inch were use for assessment. Before data collection, the contour gauge was tested with and without a tissue between the pins and the bone. The tissue was to act as a guard against the metal scratching the bone, however, the training session revealed this was a source of error and the gauge was subsequently used without the tissue. An android phone camera was used to take pictures of the contour gauge shape after scoring the traits NBC and PBD. For additional data that might have been useful for understanding error, a protractor was used to measure the angles on the contour gauge for the NBC contours and recorded in the notebook. After one skull was fully assessed, it was carefully placed back in its resting place and another individual was chosen.

Qualitative data was also collected. When observer 3 assessed the U of M individuals, he was provided with the trait descriptions and supplemented notes from the atlas as well as the line drawings from *Osteoware* (2020). On his last individual, he walked the author through how he assessed each trait. His process was written down in the journal, and, when he was finished, both observer's scores were compared to each other and a discussion took place about how the decision was made. This is helpful to determine where error is occurring so it can be reduced, rather than just comparing scores and assigning an inter-observer agreement value to each trait. The BU individuals were assessed by observer 2, but the notes from the training session were used as qualitative data as well.

3.2.2b Continuous data

Measurements for *Fordisc* were taken following the *Fordisc* guide, a guide published by the Forensic Anthropology Centre at the University of Tennessee, Knoxville, specific to the measurements required for the program (Jantz & Ousley, 2012; Langley et al., 2016a). These

measurements were practiced on one BU individual with observer 2 present to ensure observer 1 was taking them correctly. All measurements were collected using a *Mastercraft* digital sliding electronic calipers with a resolution of 0.01mm or spreading calipers. Measurements were only taken once for each individual as the method has been validated and shown to have low intra and inter-observer error rates, in other words, high agreement in values (Langley et al., 2018). Spreading calipers were measured in centimeters, so this was used for the unit when recording all measurements. Measurements were recorded to the first decimal place, including those taken on the sliding caliper that measured in millimeters. Metric data was converted to millimeters when being input into *Fordisc.* Any landmarks that had any damage were not used, and, therefore, any measurements using these landmarks were not taken.

3.2.3 Analysis

Basic statistical analyses, such as the Cohen's Kappa test to determine level of agreement (Cohen, 1960), were completed in *IBM SPSS v.28.0.0.0 (190)* software, however, some trait datasets resulted in unexpected values. These tests were checked by hand calculation to confirm the result. Excel was then used to calculate bias and prevalence indices that were needed to identify bias and prevalence causing the unexpected Kappa values in binomial datasets. A qualitative analysis for whether bias and prevalence were present was completed for datasets with more than two categories.

More advanced statistics for ancestry estimation using metric and nonmetric data were carried out automatically within the appropriate programs *Fordisc 3.1, MaMD Analytical v.0.4.5.*, and the online application *HefneR*. Finally, a descriptive analysis of the agreement, or lack thereof, in ancestry assessment between the programs was conducted.

3.2.3a Observer error for trait scores

To address the objectives of determining the scoring method's repeatability, the datasets from all scoring periods and all observers were analyzed using Cohen's kappa statistics; a statistic to determine the level of agreement for trait scores (Cohen, 1960). This test has been used extensively for intra-observer agreement (Coelho, Navega, Cunha, Ferreira, & Wasterlain, 2017; Dinkele, 2018; Hefner, 2009; Hurst, 2012; Kamnikar, Plemons, & Hefner, 2018; Klales & Kenyhercz, 2015; L'Abbé et al., 2011; Primeau, Arge, Boyer, & Lynnerup, 2015) and inter-observer agreement of morphoscopic traits (Coelho et al., 2017; Dinkele, 2018; Hefner, 2009; Klales & Kenyhercz, 2015; L'Abbé et al., 2011). All assumptions and requirements for Cohen's kappa were met (Banerjee et al., 1999; Cohen, 1960; Lund & Lund, 2020), therefore, kappa tests for each trait and for each study sample were run twice due to the different observers.

Cohen's kappa, unweighted, is determined through the formula:

$$k = \frac{P_o - P_e}{1 - p_e}$$

Where *k* is the kappa value, P_o is the proportion of observed agreement among raters, and P_e is the hypothetical probability of chance agreement. Therefore, the kappa value indicates the agreement that occurs above that of agreement by chance (Cohen, 1960). This was calculated using *SPSS* as well as manually and in excel to confirm the statistics were correct as there were several kappa values that did not go as expected based on the data.

The null hypothesis states that the agreement between observers is no different than the agreement by chance, as expressed by kappa being 0. The alternate hypothesis states that the agreement between observers is different from agreement by chance, as expressed by kappa being a +1 or –1, where +1 is complete agreement and -1 is complete disagreement. The rejection of the null hypothesis means there is a relationship between the trait scores and how

observers scored them. For example, it can indicate if one trait has more disagreement between observers than another trait due to an ambiguous description or different interpretations of a description. Similarly, if there is more agreement than chance, the scoring method is considered to be interpreted the same or traits are scored the same way between observers in order to come to the same result. Qualitative data can complement this analysis to determine why agreement is different than chance, whether higher or lower.

For ordinal data (traits in Table 2 not labelled with asterisks), Cohen's kappa can still be used, but must use a different formula with weights attached to it (Cicchetti & Allison, 1971; Cohen, 1968; Fleiss & Cohen, 1973). Weights are a way to quantify the degree of agreement when multiple categories are close together and have subtle differences among them. This means observers chose different scores based on small changes, so those scores that are closer together have higher agreement than scores that are further apart with more drastic differences.

The assumptions for weighted kappa are the same as the unweighted kappa except that it is on an ordinal scale rather than nominal. Additionally, the null hypothesis and alternative hypotheses are the same as the unweighted kappa.

The formula for determining weighted kappa is:

$$k_w = 1 - \frac{\sum_{ij} w_{ij} p_{ij}}{\sum_{ij} w_{ij} e_{ij}}$$

These weights mean that disagreement for categories that are close together are given less penalty of disagreement than categories that are further apart. The k_w is the kappa value with the weights applied. The w_{ij} indicates the weight applied to the scores as determined by Table 3, the p_{ij} is the proportion of observed scores agreeing, and e_{ij} indicates the proportion of the expected number scores agreeing as calculated by adding the number of scores per each score category for each observer divided by the total number of scores given (Zeman & Benus, 2020).

There are two weights that can be applied: linear and quadratic. A linear weight assumes that for every category of disagreement away from each other, the weight reduces by an equal amount. A quadratic weight applies a more drastic decrease for each category of disagreement because it takes into account that some disagreement may be more drastic between traits that are scored further apart along the scoring range (ex. score of 1 compared to score of 4) than scores closer together (ex. 1 compared to 2) (Watson & Petrie, 2010).

An example for the linear weights between different scores for the trait INA with five categories of scoring is in Table 3. When observers have a perfect agreement, the weight of 1.00 is given, whereas partial agreement is given less weight as the disagreement increases. For example, if observer 1 gives a score of 1 and observer 2 gives a score of 2, then the weight is 0.75 for partial (3/4) agreement. For a quadratically calculated weight, 0.9375 would be the weight applied to one category of disagreement instead of 0.75. Furthermore, instead of 0.25 for three categories of disagreement, it would be 0.4375.

Table 3: Example of how a linear weight would be applied between each category with 1 being no disagreement and 0 being complete disagreement. In general, scores along the diagonal from top left to bottom right are scores that agreed between observers.

		Observer 2				
		Score 1	**Score 2**	**Score 3**	**Score 4**	**Score 5**
Observer 1	**Score 1**	1.00	0.75	0.50	0.25	0.00
	Score 2	0.75	1.00	0.75	0.50	0.25
	Score 3	0.50	0.75	1.00	0.75	0.50
	Score 4	0.25	0.50	0.75	1.00	0.75
	Score 5	0.00	0.25	0.50	0.75	1.00

A linear weight was chosen because a quadratic weight still considered scores that were further apart as more similar than if using a linear weight. This might skew the results to show that there is more agreement between observers than there actually is, especially when the scores that are further apart are for expressions that are drastically different and would likely not be

confused with each other. For example, INA scores of 1-3 do not have a nasal sill and vary with their angle, and scores 4 and 5 have a sill. Scores of 1 and 3 vs. 1 and 4 can mean the difference between low disagreement and high disagreement; a nasal sill is very distinct so the difference between a score of 3 and 4 should not be higher in agreement than between 1 and 3 since the sill is the defining characteristic between the two scores and should indicate higher disagreement. Furthermore, only researchers from two studies on observer error with morphoscopic traits explicitly stated they used a quadratic weight (Kamnikar et al., 2018; Moffit, 2017), all other researchers do not state whether they use a quadratic or linear weight (Atkinson & Tallman, 2020; Coelho et al., 2017; Dinkele, 2018; Hefner, 2009; Hurst, 2012; Kamnikar et al., 2018; Kilroy, Tallman, & DiGangi, 2020; Klales & Kenyhercz, 2015; L'Abbé et al., 2011; Moffit, 2017; Wang, 2016). Therefore, a linear weight is justified based on the argument above.

A linear coefficient adds an equal weighting to each score of disagreement. A trait with a kappa value closer to +1.0 means there is more agreement of scores among observers and the scoring method is able to produce repeatable results (Cohen, 1960). Within SPSS, the weight of 0 is given to scores in complete agreement whereas 1 is the maximum disagreement, thus resulting in a table opposite to Table 3.

To calculate the linear weights, the formula is used below:

$$w_i = 1 - \frac{i}{k-1}$$

Where *i* is the distance between categories and *k* are the number of categories for scoring a particular trait.

Once kappa values were calculated, the degree of agreement was determined following Table 4. In this study, traits with "high" agreement have values that are substantial or precise, in order words, an agreement value greater than 0.61; these are traits where the majority of scores

are agreed upon between observers. Traits with “low” agreement have values that are moderate to poor (k<0.61).

Table 4: : Degree of agreement between observers based on the table by Landis and Koch (1977).

Degree of agreement	K-value
Precise	0.81-1
substantial	0.61-0.8
moderate	0.41-0.60
fair	0.21-0.4
slight	0.00-0.20
poor	<0.00

Kappa values cannot be compared between studies for two reasons: marginal distributions and prevalence (Banerjee et al., 1999). The marginal distribution is the total number of times the raters score each trait in comparison to the other observer (‘total’ columns in Table 5). The marginal distributions differ between studies based on how raters score the traits, which is affected by personal bias (Banerjee et al., 1999). For example, the creator of the scoring method, Hefner, has more experience with the scoring system and has assessed thousands of skulls to understand the full variation in trait expressions; he is more likely to detect subtle differences between trait expressions. When applied to inter-observer kappa values, this means that one observer is going to have more scores in one category than another category as compared to the other observer. This is because they will have a bias towards what they interpret as each score. Instead of having a relatively equal number of scores for a score of 1 (Table 5), one observer will have more than the other. Observer 2 scoring four individuals with a score of 1, whereas observer 1 scored ten individuals with a score of 1. These are the distributions of their scores in the margins of the inter-rater table, also known as the “marginal distributions”.

Throughout the results and discussion, ‘bias’ in relation to scores will pertain to this specific phenomenon unless otherwise noted.

Table 5: Example of how two observers could score trait expressions in fourteen individuals. Diagonal from top left to bottom right are scores that agreed between observers.

Trait A	**Observer 1**			
Observer 2		**Score 1**	**Score 2**	**Total**
	Score 1	4	0	4
	Score 2	6	4	10
	Total	10	4	14

Prevalence also affects the kappa score because two different populations will have differing kappa values based on the trait expressions within that population (Banerjee et al., 1999). In other words, if there is more of one specific trait expression in each population that is constantly disagreed upon, the kappa statistics will differ. For example, if one trait has the expression of 1 that comes up in 90% of the individuals, and all raters agree what ‘1’ looks like, it will have a high kappa value for that trait. However, if ‘1’ is not found in another population, but expressions 2 and 3 are found with a high disagreement about what each of these expressions look like, then the agreement for this trait will be much lower. While kappa values across studies cannot be directly compared, they are still useful to determine general patterns and explain discrepancies in values.

When bias or prevalence is present, Hallgren (2012) suggests to use Byrt, Bishop, and Carlin's (1993) adjustments to the kappa equation. In order to determine if a data set has bias or prevalence, Byrt et al. (1993) have devised formulae for finding the Bias Index (BI) and Prevalence Index (PI) according to the layout of Table 6.

Table 6: Example of how data is organized to find the BI and PI with a-d representing the number of scores given in each category, and N as the total number of individuals scored.

		Observer 1		
Observer 2		**Score 1**	**Score 2**	**Total**
	Score 1	a	b	
	Score 2	c	d	
total				N

$$BI = \frac{a+b}{N} - \frac{a+c}{N} = \frac{b-c}{N} \qquad PI = \frac{\frac{a+b}{N} + \frac{a+c}{N}}{2} - \frac{\frac{c+d}{N} + \frac{b+d}{N}}{2} = \frac{a-d}{N}$$

Once the dataset has been determined if it has bias, prevalence, or both, then the original values are replaced by their average to adjust for these errors, and kappa is calculated with these averages. However, these BI and PI formulae cannot be translated from a 2x2 table to a 5x5 table, therefore, they were only calculated for PBD and NO, traits with two expressions. The other traits were qualitatively assessed. Since the author could only qualitatively assess the marginal totals for bias and prevalence in most traits, the formulae needed to adjust kappa were not used. Qualitatively, the author used two categories of bias. “Great” bias issues are when 50% or more of the scores are in disagreement and/or have scores further apart than one category for ordinal traits. If bias is substantial, Byrt et al. (1993) recommend that an index of agreement may be unnecessary as it should be investigated to determine what this bias is. Prevalence issues were identified if there were expressions completely missing from the data or if an expression appeared less than half the number of times that other expressions did.

Imbalances in these marginal totals can also create paradoxes due to both bias and prevalence, where even if there is a lot of agreement, the analysis will result in a low kappa value (p. 426, Table 10 in Byrt 1993). When prevalence effects are substantial, using Cicchetti and Feinstein’s indices of positive and negative agreement is recommended by Byrt et al. (1993) to

resolve these paradoxes. However, again, the author was unable to translate these to multiple categories and the paradoxes cannot be resolved.

3.2.3b Ancestry estimate

The focus of this study is more on whether ancestry assessment programs agree in their results than on how ancestry assessments are calculated using data. Therefore, there is less focus on explaining the advanced statistics utilized by each program to calculate the probability of group affinity, except in cases when it is needed for the interpretation of the results. Each program utilizes a different statistic to compare the data to reference populations and estimate ancestry. Since each program has a different reference population, there are different populations that the same individual could be grouped into (Table 7) because the programs will still provide an ancestry estimate regardless of whether the individual's ancestry is represented.

To be able to compare the ancestry estimates across programs, large regional designations are given to each population: European, African, Asian, and North American Indigenous. Two populations were difficult to group with one of the four geographic groups due to their different evolutionary histories, these were the Pacific Islander populations and the Guatemalan populations. Since there are four groups, agreement greater than chance would be agreement of more than 25% of the ancestry estimates.

Table 7: Reference groups in each database, and the conversion of those geographic groups into the four major geographic regions so each assessment from each database can be compared.

Database	Ancestry group label in the database	Finer group organization	Ancestry when converted to the 4 major geographic regions
Fordisc- FDB	"White" American		European
	"Black" American		African
	"Hispanic"		Problematic, thus any results that came up with this, used second ancestry estimate and probability
	Guatemalan		American Indigenous
	American Indigenous		American Indigenous
	Japanese		Asian
	Vietnamese		Asian
	Chinese		Asian
Fordisc- Howell's	East Asian	Ainu, Andaman, Anyang, Atayal, Buriat, Hainan, N Japan, S Japan, Philippines	Asian
	Native American	Arikara, Inuit, Peru, Santa Cruz	American Indigenous
	Pacific Islanders	Australia, Easter Island, Guam, Mokapu, Moriori, Tasmania, Tolai	Asian
	African	Bushman, ODgon, Egypt, Teita, Zulu	African
	European	Berg, Norse, Zalavar	European
	20th century black		African
	20th century white		European
HefneR-2009 article	African		African
	American Indian		American Indigenous
	Asian		Asian
	European		European
MaMD analytical- MaMD	American Black		African
	American White		European
	American Indian		American Indigenous
	Asian		Asian

	Guatemalan Southwest Hispanic		American Indigenous Ignored, used second ancestry result in place

Metric Analysis

Fordisc

Fordisc is used as a comparison to the results from *MaMD analytical* and *HefneR*. Therefore, understanding how to read the results are important, which is partially dependent on knowing how the program groups individuals using different probabilities. *Fordisc* uses linear discriminant function analysis and Mahalanobhi's distance when comparing more than two groups. This uses the formula:

$$D_j^2 = d_j(W^{-1})(d_j')$$

Where D_j^2 is the distance of the individual from the jth population, d_j is the difference vector from the mean of the jth population, and W is the within group covariance matric pooled over all groups (Manthey & Jantz, 2020, p. 276)

The smallest Mahalanobis distance (D^2) among reference groups indicates the group that the individual is considered to have the closest affiliation with, and these distance measures are used to calculate posterior and typicality probabilities for each group. In general, the shorter the distance, the higher the posterior probability (PP), which is the probability that the unknown skull belongs to a group. Therefore, the highest PP is the most likely group the unknown individual belongs to, based on the assumption that the unknown individual belongs to one of the reference groups (Jantz & Ousley, 2012). For example, if the posterior probability is 0.65, then the individual has a 65% chance of belonging to that group. This probability is the one that is used to estimate the ancestry of the individual.

Typicality probabilities (Typ P) are helpful to determine whether the individual could belong to one or more groups. There are three of these probabilities: F, Chi, and R (Jantz & Ousley, 2012). Typicality F converts the distance into an F ratio in relation to the size of the reference group and decides whether the individual is typical of each group. When this value is greater than 0.05, then an individual is typical of that group (Jantz & Ousley, 2012), in other words, whether the individual's combination of measurements are typical in that group. Typicality Chi relates to D^2 and decides which group or groups the individual is most typical to and which groups the individual is atypical of (Jantz & Ousley, 2012). This typicality varies with relation to D^2.

The most useful of these typicality probabilities is the Typicality R, which takes the unknown skull, adds it to each group, then calculates the distance of each individual in that reference group, plus the unknown individual, to that group's mean or centroid. This ranks the unknown individual in the group with how far it is from the centroid compared to the individuals in that reference group (Ousley & Jantz, 2012). For example, the unknown skull can rank number 50 out of the 54 individuals, with rank 1 being the closest to the group's centroid. This is interpreted as: the unknown individual falls within the range of that group's measurements but is on the margin of that variation (Jantz & Ousley, 2012). Forty-nine of the reference individuals are closer to the centroid than the unknown skull, and three individuals are just as far or further away from the centroid as the unknown individual. The higher the unknown individual ranks in comparison to the reference individuals, the higher the typical R probability. Therefore, a score above 0.05 means the individual is typical of that group.

The performance of the equation is determined by the number of individuals that are correctly grouped in the reference populations when the unknown individuals is added. In

general, 75% was considered to be a well-functioning equation. This percentage is observed beside the label “Total Correct” and has been cross-validated. The program automatically removes measurements that it deems unusable by the discriminant function, however, the exclusion of measurements by the author was also done if the measurements were three or more standard deviations above or below the mean as recommended when using the program.

If the equation was performing poorly (below 60%), or the typicality probabilities were all below 0.01, then the stepwise function with Forward Wilks L was performed. A minimum and a maximum of six traits were chosen for comparison due to poor computer power to run the stepwise function for multiple measurement combinations above that. This means the software runs combinations of six measurements taken from the total number of measurements through the entire reference database. It runs a different set of six combinations until it is through the total number of trait combinations possible from the total number of measurements.

For the analysis, all measurements were used except those labelled in red or blue in the software program. The coloured labels indicate that if these measurements were used, it would drastically reduce the reference individuals the individual is compared against because not all the measurements are available for all of the reference individuals. Measurements were also excluded if they were three standard deviations away from the mean, or the stepwise function needed to be performed. Measurements that are flagged as problematic may be because some of the individuals were anatomical individuals whose crania were cut and were missing some of the bones, which would affect some measurements. Otherwise, any measurements that are problematic may be due to individual variation that exceeds the variation found within the databases. A total of twenty-six measurements and six angles could be used for comparison;

angles were automatically calculated based on the measurements put into the program, therefore depending on the measurements that were included, they could have zero or six angles.

The individual was compared to both Howell's and FDB reference groups. Outliers were excluded to reveal any group relationships masked by the outliers, then reference populations were excluded to reduce the number of populations in the comparison to ensure higher certainty on the best estimate. Reference populations were excluded based on whether multiple Typ P's were below 0.01. This threshold was chosen rather than 0.05 because many unknown individuals had all of the reference populations with Typ P below 0.05. After reducing the number of comparative populations once, any more outliers and measurements were excluded to reveal any further relationships masked by these outliers.

If the equation performance was not above 75%, then outliers were excluded until there were no outliers left to exclude. If it was not above 75%, and there were still many reference groups (>6), reference groups were then excluded in the order of Typ P's below 0.05, then 0.01, then PP of 0, PP below 0.01, and PP below 0.05. Normally this series of events did not need to run all the way through for the equation to place an individual in one of at least four groups up to ten groups. However, on a few occasions, the equation was still not performing well and other attempts to improve equation performance included running the stepwise function. The stepwise function was run to either choose the top ten groups to compare the individual to if equation performance was especially low (10-40%) or ran to see if it improved equation performance if it was at 70%. When equation performance was especially low, running the stepwise first and using the top ten groups to run the analysis again helped improve the equation performance in most cases.

After each analysis, the best ancestry estimate was written down along with the PP and Typ P's, equation performance, and final number of measurements used for the analysis. When an analysis could not clearly place the individual into one group, the two best ancestry estimates were written down for each reference bank. The ancestry that was chosen for comparison to *MaMD Analytical*'s results from both FDB and Howells was the one with the highest PP with a high Typ F. If the best group had a high PP with a low Typ F, and the second-best group had a lower, but not too much lower PP, with a high Typ F, then that was chosen over the higher PP. This is because Typ F indicates whether the skull is typical of that group, and if it is atypical, then it may not be the best indicator of ancestry as it is not as similar to the group. The highest PP with a Typ F indicating the individual can be a part of that group (>0.05), a Typ Chi with a value showing the group it is most typical of (higher number), and a Typ R that ranks the individual higher than the last individual of that group will be used to determine the ancestry most likely attributed to the unknown individual.

If the two databases had matching ancestry estimates that were in the best two estimates, the final ancestry estimate was chosen based on how well the equation performed with each database. This was determined from the author's notes on whether the individual was placed right away or had results with Typ P below 0.01 and if it needed the stepwise function. When the top results from FDB and Howells did not agree, further analysis was necessary. Similarly, it was based on observer 1's notes of how well each database ran the data, the equation performance from each estimate, and the top three groups that appeared and how close their posterior probabilities were. Justifications for each selection were placed within observer 1's notes.

Since the classifications in *Fordisc* are much finer than the *MaMD Analytical* and *HefneR*, the groups were clumped into the four major geographic regions based on their

geographic locations. For example, Howell's groups "East Asian" were renamed "Asian." When an individual was classified as Hispanic as the best ancestry estimate, the second ancestry estimate was written down. The Hispanic category can create problems within ancestry estimates since it groups many individuals with different geographic origins under one heading that is more cultural in nature (Hefner et al., 2015; Monsalve & Hefner, 2016).

Morphoscopic Analysis

Two programs using morphoscopic traits are useful for comparison with each other, however, there are differences in the traits that can be used for analysis (Table 8).

Table 8: Morphoscopic traits that are used for analysis in each software. The bottom five traits are those introduced after the original 2009 publication.

Morphoscopic traits	MaMD analytical	Osteomics HefneR
Anterior Nasal Spine	Yes	Yes
Inferior Nasal Aperture	Yes	Yes
Interorbital Breadth	Yes	Yes
Malar Tubercle	Yes	Yes
Nasal Aperture Width	Yes	Yes
Nasal Bone Contour	Yes	Yes
Nasal Overgrowth	Yes	Yes
Post-bregmatic Depression	Yes	Yes
Supranasal Suture	No	Yes
Transverse Palatine Suture	No	Yes
Zygomatic Suture	Yes	Yes
Nasal Aperture Shape	No	No
Nasal Bone Shape	No	No
Nasofrontal Suture	No	No
Orbital Shape	No	No
Posterior Zygomatic Tubercle	Yes	No
total	**10**	**11**

HefneR

HefneR is an application hosted through a free, web-based platform called *Osteomics* and is available to practicing forensic anthropologists (Coelho & Navega, n.d.). The application was created by Coelho et al. (2017) using the trait frequencies in Hefner's 2009 article, which means it includes eleven of the described sixteen traits, compared to the ten that *MaMD Analytical* uses. Prediction of ancestry is calculated by a naïve Bayes classifier, which is a probabilistic model based on the Bayesian Theory, that assumes conditional independence (Coelho et al., 2017; Coelho et al., 2020). The Bayesian Theory states it is possible to determine the probability of a specific event occurring using the current data by looking at probabilities of whether it occurred or did not occur in past events (Fielding, 2007a). Therefore, the probability of each ancestry occurring when an individual exhibits certain trait scores is determined based on the likelihood of these trait scores occurring and not occurring in each ancestral reference population. Conditional independence is the assumption that the presence or absence of a trait is not related to the presence or absence of another trait, which is not entirely fulfilled due to the covariation of traits, however, it is argued by Coelho et al. (2017) to work well regardless.

This naïve Bayes classifier (Coelho et al., 2020) is:

$$P(A_K|X_i) = \frac{P(A_K)\prod_{i=1}^{p} P(X_i|A_K)}{P(X_i)}$$

Where *A* is for the ancestral group, *X* for the morphoscopic trait expression, *i* for the number of groups, *p* for the number of traits used in the calculation, and *P* is the probability (Coelho et al., 2017; Coelho et al., 2020). The prior probability, $P(A_K)$, was determined based on the frequencies in Hefner's 2009 paper, and the Bayes' theorem (Fielding, 2007a) was used to invert the probability and create the naïve Bayes classifier (Coelho et al., 2017; Coelho et al., 2020). The classifier allows for the posterior probability, $P(A_K|X_i)$, to be determined, which is

the probability of an individual being part of each ancestral group, *A*, based on the presence of the trait expression, *X*. Much like *Fordisc*, the ancestral group with the highest posterior probability is the most likely ancestry for the unknown individual (Coelho et al., 2020).

MaMD Analytical

MaMD Analytical is a software program that uses an artificial neural network (aNN) using R coding (Hefner, personal communication). An aNN is modelled after a real, biological neural network that has many interconnected and independent neurons (Fielding, 2007b). Essentially it is a network that makes predictions based on the input of known data and compares these predictions to how many times it occurred in a training set. With each added set of traits with a known associated ancestry, it gets better at predicting ancestry of a new set of traits because it adjusts the prediction error for the next calculation (Fielding, 2007b). Not much information is published about this program because it is still in beta testing. It is acceptable to use this program as a comparison to results from *Fordisc* (Hefner, personal communication), and is important for determining whether ancestry estimations using two different data will agree.

Much like *Fordisc*, this program results in posterior probabilities with the highest PP being the most likely ancestry associated with the individual (Hefner, personal communication). The results also indicate the model's accuracy, sensitivity, and specificity. The accuracy is how well the model behaves with the unknown individual's data included, in other words, can the model still correctly identify individuals in the data base with the addition of the unknown individual. The sensitivity is to understand the influence of the variables on the results, or how well the model can identify positive results (few false negatives) (Obertova & Stewarts, 2020). The specificity shows the user how well the classification works, or how often the predicted

membership was correct within the reference group (ability to identify negative results, with few false positives). For example, if the number is 0.92, then the model is 92% correct.

All groups were included in the analysis, but if Southwest Hispanic was the ancestry estimate, it was not taken seriously due to its problematic nature, therefore the second highest PP was used for the estimate. In the case of beta testing, the software version used in this research did not include cross validation, thus ending up with higher-than-normal probabilities (Hefner, personal communication). In the case of this program, only the posterior probability was used to determine ancestry.

Table 9 : Summary of the number of metric and morphoscopic measurements used for ancestry estimation for each skeletal individual at each university.

Collection	Individual	Number of metric measurements (N=26)	Number of morphoscopic traits (N=16)
University of Manitoba	1	20	15
	2	26	16
	3	26	16
	4	26	16
	5	22	16
	6	24	13
	7	26	16
	8	23	15
	9	20	15
	10	18	16
Brandon University	11	26	13
	12	25	16
	13	26	16
	14	26	16
	15	26	14
	16	26	15
	17	26	15
	18	18	10
	19	26	15
	20	17	13
	21	10	3
	22	20	9
	23	22	15
	24	19	11
	25	18	12
	26	17	15
	27	19	13

Chapter 4: Results

4.1 Observer error (Quantitative)

Observer 1 (the author) scored all sixteen traits for twenty-seven individuals from both U of M and BU whereas observer 3 scored all sixteen traits on ten U of M individuals and observer 2 scored eleven traits on seventeen BU individuals. Three observers across these twenty-seven individuals resulted in 308 score pairs for inter-observer comparison of observer 1 to either observer 2 or 3. For intra-observer comparison between scoring periods, the total number of score pairs among twenty-seven individuals are 568. The number of paired data able to be compared was affected by skeletal teaching individuals with missing bone, disagreements on whether the trait could be scored or not, and teaching individuals who were not available for data collection during a collection period. Throughout the results, the capital letter 'X' denotes the total number of score pairs within each trait that can be compared while the number of disagreements/agreements among those pairs is denoted by 'x'.

4.1.1 Intra-observer error- Cohen's Kappa

Intra-observer agreement (Table 10; $k<0.61$) for observer 1 is low when comparing scores taken before (scoring period 1) and after training (scoring period 2). Almost half of the traits (6/16) have fair to poor agreement, two traits have moderate agreement (MT, TPS), three traits have substantial agreement (NBC, NBS, PZT), two traits have precise agreement (NFS, PBD), and two have clear paradox issues that triggered the need for determining the Bias Index (BI) and Prevalence Index (PI).

Many traits had "great" (n=6) or "slight" bias issues (n=3) as well as prevalence issues (n=9) (Table 10). Bias issues are considered "great" when 50% of the scores are in disagreement and/or have more than two scores further apart than one category for ordinal traits. Prevalence

issues were identified if there were expressions completely missing from the score data or if a score appeared less than half the number of times other scores did (Table 11, Table 12). In Table 11 and Table 12, the two 'marginal distributions' of trait scores ('total' columns) indicate whether there are bias or prevalence issues in the data, and these issues are identified based on if the number of individuals assigned a score are similar for each score. Table 11 has bias issues because a total of five individuals were scored as an expression of 1 in scoring period 2, whereas scoring period 3 did not have any individuals scored as 1. Similarly, three individuals in scoring period 1 were scored as an expression of 3 whereas no individuals were scored as an expression of 3 in scoring period 2. Table 12 has prevalence issues because the expression of 3 does not show up as often as a score of 1 or 2. There are no bias issues because almost an equal number of individuals were scored in scoring period 2 and 3 for expressions 1 and 2. See appendix 2 for frequency tables of all other traits that indicate bias or prevalence issues.

In contrast, intra-observer agreement is high after training (scoring period 1 vs. scoring period 3) with the majority of traits (11/16) having substantial agreement (Table 10; $k \geq 61$). Three traits have fair agreement (NO, SPS, PZT), one trait has moderate agreement (MT), and one trait (IOB) has precise agreement. Half of the traits had increased kappa values after training whereas a few remained high (NBS, NBC, NFS), low (MT, NO, SPS), or had reduced kappa values (PBD, NFS, PZT). The higher the kappa values, the higher the agreement between observers, or, in other words, the less observer error that is occurring.

There were far more issues of bias between scoring period 1 and 2 than between scoring period 2 and 3 (Table 10). Instead, scoring period 2 and 3 had more prevalence issues with many expressions appearing in low frequency or not at all (ex. Table 11,Table 12). When comparing scoring period 2 and 3, none of the traits have great bias issues like scoring period 1 and 2 had,

however, six traits have slight bias issues and the majority of traits have prevalence issues (n=12; Table 10).

Table 10: Kappa and p-values for observer 1's (CM) intra-observer tests comparing the untrained (CM1) session and the two trained sessions (CM2 and CM3). PI and BI are abbreviations for Prevalence Index and Bias Index, of which only two traits are calculated whereas the other traits are determined qualitatively. P-values that are labelled 'n/a,' not available, are due to not having the software available at the time of recalculating the kappa value.

	CM1 vs. CM2			CM2 vs CM3		
Traits	**Kappa**	**p-value**	**PI/BI**	**Kappa**	**p-value**	**PI/BI**
ANS	0.255	0.008	Great bias	0.771	<0.001	Slight bias
INA	0.38	0.088	Slight bias	0.694	<0.001	Slight bias and prevalence
IOB	0	n/a	Great bias	0.839	<0.001	Prevalence
MT	0.553	0.016	Great bias and prevalence	0.596	<0.001	Prevalence
NAS	-0.167	0.397	Slight bias and prevalence	0.62	<0.001	Slight prevalence
NAW	0		Prevalence	0.747	<0.001	Prevalence
NBC	0.7	0.005	Prevalence	0.694	<0.001	Prevalence
NBS	0.774	0.008	Prevalence	0.743	<0.001	Prevalence
NO	0.286	0.361	(0.4) Great bias and prevalence (-0.2)	0.317	0.207	Slight bias (0.143) and prevalence (-0.429)
NFS	1	0.001	Prevalence	0.801	<0.001	Slight prevalence
OBS	0.391	0.115	Prevalence	0.758	<0.001	Slight prevalence
PBD	1	0.014	0 (Neither)	0.692	0.002	Slight bias (0.158)
PZT	0.696	n/a	Prevalence	0.326	n/a	Prevalence and slight bias
SPS	0.276	0.322	Great bias	0.275	0.058	Slight bias
TPS	0.447	0.009	Slight bias	0.738	<0.001	Slight bias
ZS	0.067	0.659	Great bias	0.609	<0.001	Prevalence

Table 11: Frequency of trait expressions for IOB and the agreement between scoring periods 1 and 2 to exemplify the bias issues found in the dataset.

	Scoring period 2			total (g)
Scoring period 1	1	2	3	
1	0	0	0	0
2	4	0	0	4
3	1	2	0	3
total (f)	5	2	0	7

Table 12: Frequency of trait expressions for IOB and their agreement between scoring periods 2 and 3 to exemplify the prevalence issues in the dataset.

	Scoring period 3			total (g)
Scoring period 1	1	2	3	
1	12	0	0	12
2	1	9	0	10
3	0	1	0	1
total (f)	13	10	0	23

4.1.1a Patterns spanning all scoring periods and all individuals (BU and U of M)

There were a few major patterns with trait disagreement in ordinal traits across all individuals (N=25) and sessions. For trait ANS, all disagreements were between a score of 2 and 3 (x=7); a score of 1 was completely agreed upon. Furthermore, every score of 2 changed to a score of 3 from scoring period 1 to 2 (x=3) and every score of 3 changed to a score of 2 from scoring period 2 to 3. For both MT and PZT, seven of each of their ten disagreements were between a score of 1 and 2. Finally, IOB only had score disagreements occur between scoring period 1 and 2 while scoring periods 2 and 3 were in full agreement. All scores were disagreed upon between scoring periods 1 and 2, and went down in numerical score from scoring period 1 to 2.

There were also patterns of disagreement in nominal traits. Of the nine disagreements in trait NAS, five were between the scores 1 and 2, and four were between the scores 1 and 3. Additionally, of the seven U of M individuals that were scored three times, three individuals had scores in scoring period 3 that were different from scoring period 2 but matched the score given in scoring period 1, much like ANS. For trait SPS, ten of the twelve disagreements were between scores of 0 and 2. Trait NFS and OBS both had high kappa scores so there were only four disagreements within both traits. Trait NFS had three disagreements between a score of 1 and 4, while OBS had three disagreements between a score of 1 and 2. Scores of 2 and 3 for NFS were all agreed upon (x=8).

For traits INA, NAW, NBC, TPS, and ZS, there were no patterns of disagreement or agreement for particular score pairs across trained and untrained sessions. Both PBD and NO did not have a pattern because they were binomial and scores could only be confused with each other and not any other score. Furthermore, some traits did not have certain scores, for example NAW and IOB did not have a score of 3 in the dataset, and NBS did not have a score of 4. It cannot be determined if there would be confusion when scoring these traits due to the limited variation in the study individuals. The only time a score of 3 showed up for NAW and IOB was in the untrained session; when trained, the score changed and stayed the same score for both trained sessions.

Many disagreements, which were not included in the Kappa analysis, were between whether observer 1 could score the trait or not. There did not appear to be a pattern for this phenomenon except that the majority of disagreements appeared in trait NBS for observer 1, whereas the majority were within trait NO among all observers.

4.1.1b Patterns that appeared between trained and untrained sessions on U of M individuals (Scoring periods 1-3; U of M)

There were also some patterns found (N=7) when comparing all observer 1's scores (X=14) on U of M individuals. For IOB, every score between scoring periods 1 and 2 was disagreed upon, but were all agreed upon between scoring periods 2 and 3, with a score of 3 not appearing in the trained datasets.

A score of 3 for trait NBS was agreed upon across all U of M individuals that had this score (x=6, X=6), and the majority of '2' scores were in agreement (x=6, X=8). The only disagreements for NBS were found on one U of M individual between a score of 1 and 2 (x=2). Three of four NBC disagreements were between a score of 1 and another score (0, 2, and 3). When assessing PZT, three of the four score disagreements were between the two trained sessions, scoring periods 2 and 3. When adding in the additional three U of M individuals that were not scored in scoring period 1, their scores also disagreed between scoring periods 2 and 3.

Nominal traits from U of M individuals also had some patterns. Three of the four NAS disagreements had scores that decreased from scoring period 1 and 2. For NO, two of the three disagreements were a score of 0 in scoring period 1 which changed to a score of 1 in scoring period 2. For TPS, all four disagreements were between a score of 2 and either a score of 1 or 3 (moving 1 score in either direction); three of them were scored 2 in scoring period 1 and changed in scoring period 2.

In general, five ordinal traits (INA, IOB, MT, NAW, NBC) had the majority of score disagreements one score away from each other. In contrast, nominal traits (NAS, OBS, SPS, ZS) had more scores that were further than one score apart. Ordinal traits that had disagreements further apart than one score were very few. For example, out of all fourteen sets of data for U of

M individuals, trained and untrained, two cases were found in INA, one in IOB, one in NBC, and one in PZT; there was with no pattern showing whether it had to do with training (x=3) or not (x=2). Finally, many ordinal traits had the majority of disagreed scores decreasing from scoring period 1 to scoring period 2 (IOB; MT, NAW, NBC, PZT) with only ANS scores increasing.

4.1.1c Patterns that appeared from all trained sessions with BU and U of M individuals

Across all individuals (n=25) in just the trained sessions, scoring periods 2 and 3, a couple patterns arose outside of those outlined in section 4.1.1a Patterns spanning all scoring periods and all individuals (BU and U of M). There was perfect agreement for trait IOB (n= 23 individuals), and a score of 3 for trait OBS was the only score that was always agreed upon (x=3, X=24). Of the four TPS score disagreements, three were between scores of 4 and another score, and three were between a score of 1 and another score. For ZS, four of the five disagreements were between 2 and another score, equally split between 0 and 1.

Similar to the comparison of scoring period 1 to scoring period 2, many ordinal traits had the majority of disagreement in scores that were one score away from each other (ANS, IOB, INA, MT, NAW, NBC, NBS, PZT) while nominal traits had more scores that were further than one score apart (NAS, NFS, SPS, TPS, ZS). Ordinal traits that had disagreements further apart than one score were very few.

4.1.1d Whether scores changed after using the atlas in scoring period 3 with BU individuals

After scoring BU individuals in scoring period 3, the photographic atlas (Hefner & Linde, 2018) was used to see if some scores in a subset of BU individuals would change upon seeing photographs of individuals with the trait expressions; these atlas scores were not used for comparison to the other observers' (observers 2,3) scores. The majority (79.6%) of scores did not

change after viewing the atlas photos (Appendix 3, Table 73). Some pairs of scores for scoring period 2 and 3 did not agree, and the rescoring agreed with scoring period 3. For example, individual 14 had a score of 4 in scoring period 2 and a score of 5 in scoring period 3. Upon viewing the atlas, the score stayed at a 5 and agreed with scoring period 3. Some pairs of scores agreed between the two sessions, and the rescoring using the atlas photographs (Hefner & Linde, 2018) did not agree with them. For example, for INA on individual 11, both sessions were scored a 5, but upon looking at the atlas (Hefner & Linde, 2018) the score changed to a 4. Furthermore, some scores changed even though they were originally scored as "unscoreable." For example, individual 17 had no score for scoring period 2 and a score of 3 in scoring period 3. Upon seeing the atlas pictures, it was scored a 2. This leads to NBS having two scores that changed but have no score to compare it to in scoring period 2.

Of the scores that did change, 40% changed to match scoring period 2 score while 60% changed and did not match scoring period 2. Of particular interest are changes in TPS scores: three of the five scores that changed upon viewing the atlas were scored 4 in scoring period 3, but the changes did not match the scores from scoring period 2.

4.1.2 Inter-observer error

4.1.2a Kappa (Observer 1, scoring period 3 vs observer 2 and 3)

In general, there was more agreement between observers 1 and 2 than there was between observers 1 and 3; eight out of the eleven traits scored by both observers 2 and 3 were higher in agreement between observers 1 and 2. When observer 1's scores were compared to observer 3's, almost half of the traits had slight to poor agreement (n=7), five had moderate agreement, one had fair agreement, and one had precise agreement. Two traits resulted in a kappa value of 0 due to bias issues. Of the 11 traits that observer 2 assessed, four traits had slight to poor agreement

with observer 1, one had fair agreement, four had moderate agreement, and one had substantial agreement. One trait resulted in a kappa value of 0 due to bias issues (NBC). There were also more disagreements that were further apart in score between observers 1 and 3 than between observers 1 and 2.

Furthermore, all Kappa agreement values for both observers 2 and 3 were lower than the intra-observer agreement between scoring periods 2 and 3 for observer 1 except for traits TPS, in the case of observer 3, and SPS in the case of observer 2, both of which had higher agreement between individuals (Table 10,Table 13).

The datasets of observers 2 and 3 compared with observer 1 had bias and prevalence issues, however, observer 3 had greater bias issues than observer 1 had with observer 2. For example, the number of individuals scored by observer 1 as having a different value from that scored by observer 2 or 3 is much greater with observer 3 than with observer 2. The number of times observer 3 had ordinal scores further apart than one score across all the traits was greater than observer 2 whose scores were concentrated to one trait having scores further apart than one score.

Table 13: Kappa and p-values for inter-observer tests between observer 1 (scoring period 3) and observers 2 and 3. PI and BI are abbreviations for Prevalence Index and Bias Index, of which only two traits are calculated and the other traits are determined qualitatively. P-values that are labelled 'n/a,' not available, are due to not having the software available at the time of recalculating the kappa value.

	Observer 1 vs observer 3			**Observer 1 vs observer 2**		
Traits	**Kappa**	**p-value**	**PI/BI**	**Kappa**	**p-value**	**PI/BI**
ANS	0.103	0.389	Great bias	0.65	<0.001	Slight bias
INA	0.355	0.025	Slight bias and prevalence	0.538	<0.001	Slight bias and prevalence
IOB	0	0	Great bias	0.103	0.215	Great bias
MT	0.25	0.114	Great bias and prevalence	0.467	0.003	Prevalence and slight bias
NAS	0.565	0.021	Neither			
NAW	0.091	0.747	Prevalence	0.074	0.634	Great bias
NBC	0.357	0.023	Slight bias	0	1	Great bias
NBS	0.091	0.389	Great bias and prevalence			
NO	0	0	Slight bias (-0.25) Prevalence (-0.75)	-0.077	0.809	Slight bias (-0.286) and slight prevalence (0.143)
NFS	0.345	0.069	Slight bias			
OBS	0.079	0.429	Great bias			
PBD	0.25	0.285	Great bias (-0.375)	0.108	0.428	Great bias (-0.545) and slight prevalence (0.273)
PZT	0.118	0.598	Slight bias and prevalence			
SPS	-0.667	n/a	Great bias, prevalence	0.58	0.001	Prevalence
TPS	0.865	<0.001	Neither	0.561	<0.001	Great bias
ZS	0.194	0.226	Great bias	0.24	0.165	Great bias

Much like the intra-observer results, many ordinal traits had the majority of disagreements that were one score away from each other (ANS, INA, IOB, MT, NAW, PZT, TPS). When looking at all the data, ordinal traits that had disagreements further apart than one

score appeared to be when comparing scores with observer 3 (x=8). Disagreements with observer 2 that were more than one score away were equal in number (x=8) but concentrated to one trait, NBC (x=5, X=11), rather than spread across them all.

Some ordinal traits showed patterns between pairs of observers, where observer 1 and observer 3 score comparisons were from U of M individuals (n=10) and observer 1 and observer 2 score comparisons were from BU individuals (n=17). When scoring ANS (n=10), most disagreements with observer 3 were between a score of 2 and 3 (x=4, X=6), with one disagreement being two scores away. Additionally, observer 1 scored all values lower than observer 3. In contrast, there was much higher agreement with observer 2 for ANS, however, there appeared to be no pattern for score disagreement. The majority of score disagreements with observer 3 for INA were more than one score away from each other (x=3, X=5). Two disagreements were between the scores of 3 and 1 and one between scores of 3 and 5. Comparatively, most disagreements with observer 2 were between a score of 4, with observer 2 scoring 4, and observer 1 scoring another number on either side (3 or 5).

When it came to IOB scores, there was very little agreement between observer 1 and both observer 2 and 3. The majority of disagreements with observer 3 were between a score of 1 and 2 (x=4, X=7), whereas the disagreements with observer 2 were equally split between scores 1 and 2, and scores 2 and 3 (x=10, X=12). Three instances of disagreement between the scores of 1 and 3 occurred between observer 1 and both observer 2 and 3, with both scoring higher than observer 1. Finally, observer 2 had more scores on the higher end of the scoring system. For example, a score of 3 seven times while observer 1 did not have a score of 3.

All disagreements on trait MT between observer 1 and observer 2 or 3 were equally split between scores. However, observer 1 scored all values higher than observer 3 whereas observer

2 scored them all higher than observer 1. For NAW, the majority of disagreements with observer 1 occurred between a score of 1 and 2 for both observer 2 and 3. Generally, observer 2 scored NAW on the extremes of the scale (1 or 3) while observer 1's scores were mainly in the middle (2). Both observer 2 and 3 had scores of 4 for NBC, but observer 1 did not score 4 on several individuals. Of the thirteen NBC disagreements with observer 2, observer 1 had a majority of scores being 1 (n= 11), while observer 2 had a range of scores with no perceived pattern to the disagreement.

For PBD, between observer 1 and both observer 2 and 3, observer 1 thought all disagreed scores were 1 when observer's 2 and 3 scored it as 0. Similarly, with ZS, most disagreements with observers 2 and 3 were between a score of 2 (scored by observer 1) and a score of 0 (scored by observers 2 and 3). Most SPS score disagreements with observer 3 were between 0 and 2, but were equally split for other scores with observer 2, with observer 2 tending to score higher. Instead of scores disagreeing for NO, most disagreements were between whether it could be scored or not, with observer 1 thinking it could not be scored more often.

Only observer 3 had comparative data for traits NAS, NFS, and OBS since observer 2 did not score these. Both disagreements with observer 3 for NAS were between a score of 1 and 2, two of the four disagreements for NFS were more than one score apart, and the majority of disagreements on trait OBS were between a score of 1 and 2. Finally, there did not appear to be a pattern of disagreement for traits TPS with both observer 2 and 3, and NBS and PZT for observer 3; observer 2 did not score these two.

4.2 Observer error (Qualitative)

How often there is disagreement is not the most important part to method testing. It is where individuals are disagreeing that need to be teased apart so potential errors in a method can be identified and rectified.

4.2.1 Observations of trait interpretation before and during training

Prior to training, there were immediately some areas of confusion when trying to apply Hefner's (2009) scoring method. Some confusion was cleared up during the training session, which was a discussion on how observer 2 interpreted and scored the traits compared to how observer 1 and 3 scored by using individual #13 as an example.

Many of the trait expressions that the individual expressed were agreed upon and scored how observer 2 had explained, however, there were differences in interpretation when determining the score. Due to time constraints and the fact some traits were all scored the same, some traits were not discussed or discussed in detail because it was assumed that the expression was understood the same way. It was only in instances of disagreement that discussion took place.

While carrying out the practice round for INA, there were individuals who looked like they had a ridge, but it was horizontal to the INA floor (Figure 17a) rather than vertical (Figure 17b); there were no instructions on how to score this expression. There were also difficulties determining where the start of the floor was to observe a slope. In the line drawings it was unclear what Hefner (2009) meant for the slope of each expression and where they were in relation to other features.

Some of the individuals looked like they had a double sill and it made observer 1 uncertain which ridge to focus on. During training with observer 2, the question of whether the superior incline before the ridge is between these sills or if the incline started further back. Observer 2 said she scored a ridge as **any** dip right in front of the facial surface no matter how small, such as a millimeter or two cavity which decreased rapidly and then increased (Figure 17c). Observer 1 had been looking for a longer decline or incline in the floor of the nasal cavity before the facial surface (Figure 17b). Observer 1 was ignoring these small dips because the ridge of the dip felt like it was in line with the INA floor, and there was no gradual incline, thus scoring it as a 3 (Figure 17c). The individuals who looked like they had a double sill or this small dip could have been the subnasal grooves mentioned by Hefner and Linde, (2018, p. 26,31;

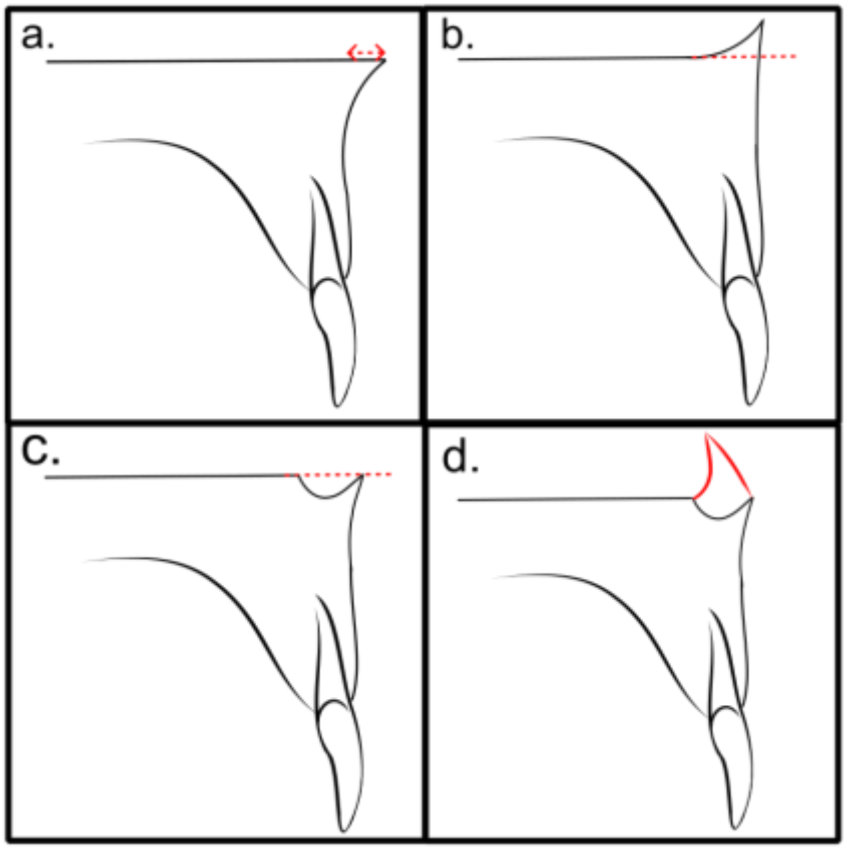

Figure 17: Cross sections of the maxilla comparing expressions of INA and features found around INA in study individuals. Red arrows (a) indicate a horizontal bone ledge jutting from INA. Red dashed lines (b,c) indicate the floor of INA in relation to the expression. Image 'b' is expression 5 as discussed in Hefner and Linde (2009) and how it relates to the INA floor. Image 'c' is a depiction of a dip below the INA floor plane. Solid red lines in 'd' are the assumed lateral subnasal grooves that Hefner and Linde (2018) discuss, but in relation to the dips that the study individuals expressed.

Figure 17d; Figure 18). Even after this discussion, observer 1 had difficulties determining what was an incline before the edge and what were subnasal grooves, thus conforming to observer 2's view of any dip would indicate a ridge.

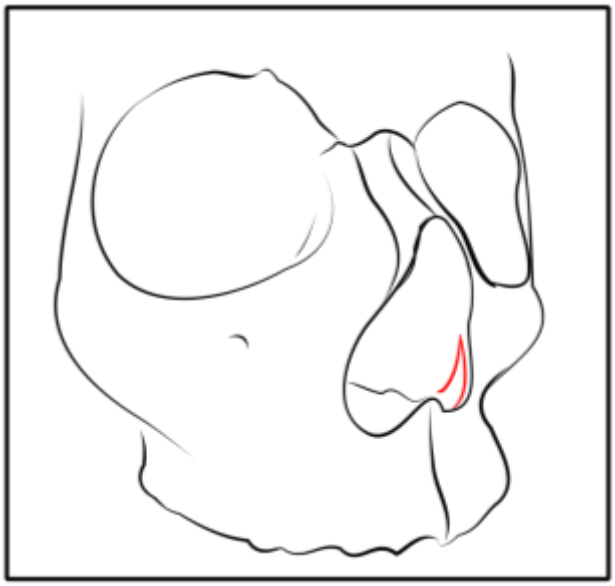

Figure 18: Assumed visual of what subnasal grooves should look like (red lines) when scoring the trait INA. Anterolateral view of the skull.

In the case of NBC, observer 1 and 3 used the contour gauge for scoring whereas observer 2 used her fingers. Prior to training, observer 1 had also used paper towel under the contour gauge to ensure the bone was not scratched. Using these two different techniques, observer 2 had scored a 4 and observer 1 scored a 3 on individual #13. When this disagreement occurred, the contour gauge was used by observer 1 without the paper towel and it changed the score from a score of 3 to a score of 4. However, after data collection this individual ended up not having a score of 4 for either observer 1 or 2.

Prior to the training discussion, IOB and NAW were two traits that gave observer 1 many issues because the original scoring method did not include the ratios and landmarks that were in Hefner and Linde (2018). One difference that arose for scoring NAW is that observer 3 used the outermost edges of the zygomatic arches for comparison to NAW, whereas observers 1 and 2 measured from the point where the zygomatic surfaces change from facial to lateral surfaces. Later, during data collection, observer 3 used *frontomalare posterale* as a landmark for the change from facial to lateral surface.

For IOB, observer 2 looked at the landmarks, but also the angle of the frontal process of the maxilla to help determine if the space between the *dacryon* landmarks wider or narrower, stating that the sharper angled frontal process (Figure 19b**)** meant the *dacryon* landmarks were closer together, and the shallower the curve meant the further apart the dacryon landmarks were (Figure 19a).

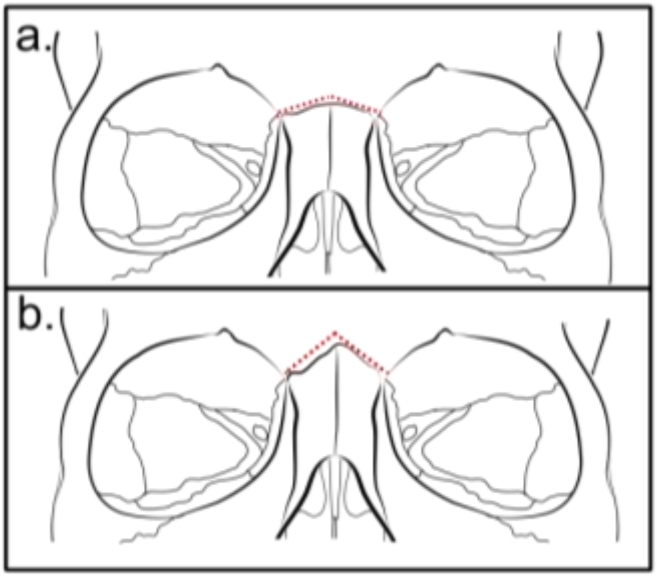

Figure 19: Image showing how the landmarks for scoring IOB (ends of red dashed line) relates to the angle of the frontal process of maxilla (red dashed line) according to observer 2. Anterior view of the skull.

Before training, observer 1 found scoring NBS difficult because the words “pinching” and “bulging” were used to describe traits, but it was not explained what these were in relation to the whole bone. Initially, observer 1 was looking at the suture rather than the 3-D contour of the bone between the sutures. Observer 1 was looking at where the nasolacrimal suture met the inferior edge to see if it flared out in comparison to the superior edge where it meets the frontal suture. Observer 2 described bulging as if someone took their finger and held it under the inferior edge of the bone and lifted; if it was clay, it would follow the contour of your finger. Similarly, pinching was if you held the superior surface of the two nasals between your fingers and pinched; if it was clay, it would follow the contour of your fingers. The line drawings only show an outline and do not encompass the 3-D shape of the bone, which includes the anterior surface contour; therefore, the lines of the suture are the only indication of what pinching and bulging should be by someone who is reading it. If

bulging is encompassing more than the suture pathway, such as the anterior surface, then it is not indicated by the drawing.

During the training session, all observers agreed on the score for MT and did not discuss how it was scored. However, during observer 1's data collection, she ended up having another discussion with observer 2. It was determined that both observers did not understand the directions of the written description for how to place the clear ruler on the "deepest incurvature."

Prior to and after the training discussion, observer 1 was confused on how to score the TPS trait "generally" if there was asymmetry. In many cases, there were individuals where the left and right side were two different expressions. In such cases, it was not scored confidently because of the vague description of "generally." For example, Individual #12 had a score of 3 on the left side and a score of 4 on the right, but was scored 3 for "generally" going anterior and posterior. This particular individual had score disagreements between observers 1 and 2 as well between scoring sessions 2 and 3 for observer 1. Another individual, #25, had a score of 4 for the left side and a score of 2 for the right side resulting in disagreement again between observers 1 and 2.

For ZS, observer 2 explicitly stated she did not understand what the description meant by "angles" because the "no angles" example looked like it had one angle. All observers were confused on how to determine what a new angle was, and this confusion continued for observers 1 and 3 during data collection.

Finally, the projection past the maxilla for NO was unclear and observer 1 could not understand what that meant even after observer 2 explained that she could feel it. The confusion stems from the phrase "extends beyond the border of the maxilla at *nasale inferious*." Observer 1 thinks it can be interpreted two ways. One is based on the line drawing given in the scoring

method: the inferior edge of the nasal bones extends past the inferior edge of the maxilla, if a line was running parallel to the inferior edge at *nasale inferious* (Figure 20a, red dotted line is the angle of the inferior edge of the maxilla), therefore, "extends beyond the border" because the nasals would no longer be connected to the maxilla at the border between them.

The second interpretation is that the inferior edge of the nasal bones, along the border with the maxilla, curves laterally over the maxilla's inferior edge so the border is no longer straight and, therefore, beyond the medial border of the maxilla (Figure 20b). This continued to be an issue for observer 1 throughout collection, especially after reading the atlas where it was mentioned "separation" is not considered overgrowth, without explanation of what this means **(**Figure 20cd).

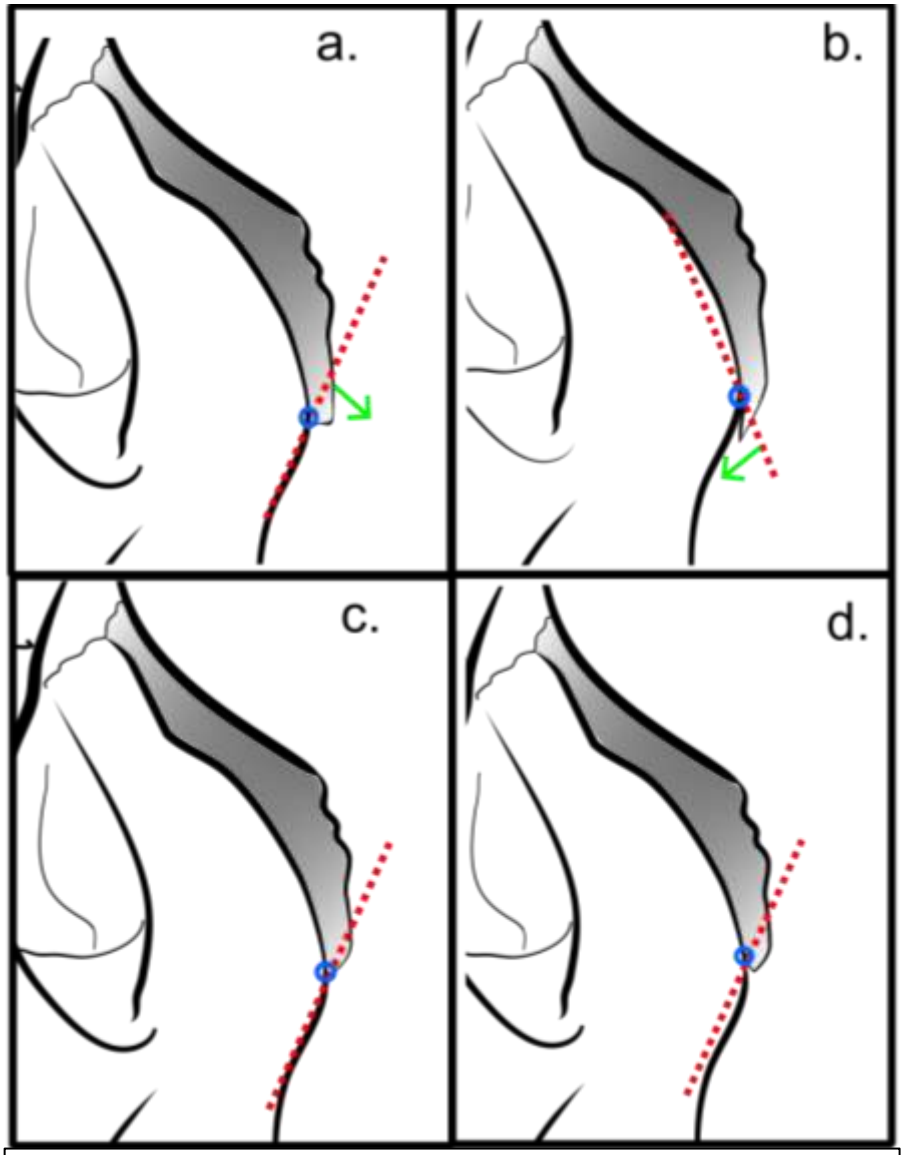

Figure 20: Dashed red lines showing the border from which NO is interpreted to be assessed from. *Nasale inferious* is defined as the 'most inferior point on left nasomaxillary suture', and represented as the centre of the blue dot, this is where the plane (red-dotted line) of the maxilla border is assumed to follow. The green arrows indicate the direction the "overgrowth" will be if the line is following the border. The plane of reference for overgrowth past the maxilla is either past the inferior margin of the maxilla (a) or past the medial border of the maxilla (b). Figures 'c' and 'd' are examples of how the author interprets various expressions of overgrowth, if the inferior border plane (red dotted line) is the correct plane of reference. Lateral view of the skull.

4.2.2 Trait interpretation after reading the atlas for additional training and using it for data collection

Even after reading several iterations of the scoring method, looking at numerous photographed expressions from publications and resources (Hauser & De Stefano, 1989; Hefner, 2009; Hefner & Linde, 2018; Hefner et al., n.d.; *Osteoware v. 2.4.037*, 2020; Wilczak & Dudar, 2020) and discussing with observer 2 how she scored each trait, there were several other areas of confusion that arose during and after data collection.

The photographs in the atlas (Hefner & Linde, 2018) were helpful when scoring ANS because observer 1 felt confident in her scores after using the pictures to match to the expression she was seeing. However, the description and line drawings on their own were not clear because there were no descriptions to differentiate a score of 2, “intermediate,” or 3, “pronounced,” while scoring. After data collection, observer 1 studied the photos from the atlas to attempt to describe observed differences of 2 and 3. An expression of 3 appears to have a longer thinner process than it does base, as well as a more acute angle of the facial surfaces as it turns into the endpoint of ANS (Figure 21c; Figure 22a), whereas a score of 2 has an equal to or longer base than length of spine and a less acute angle (Figure 21b; Figure 22b). Finally, a score of 1 is a right angle or obtuse angle (Figure 21a) with a base longer than length.

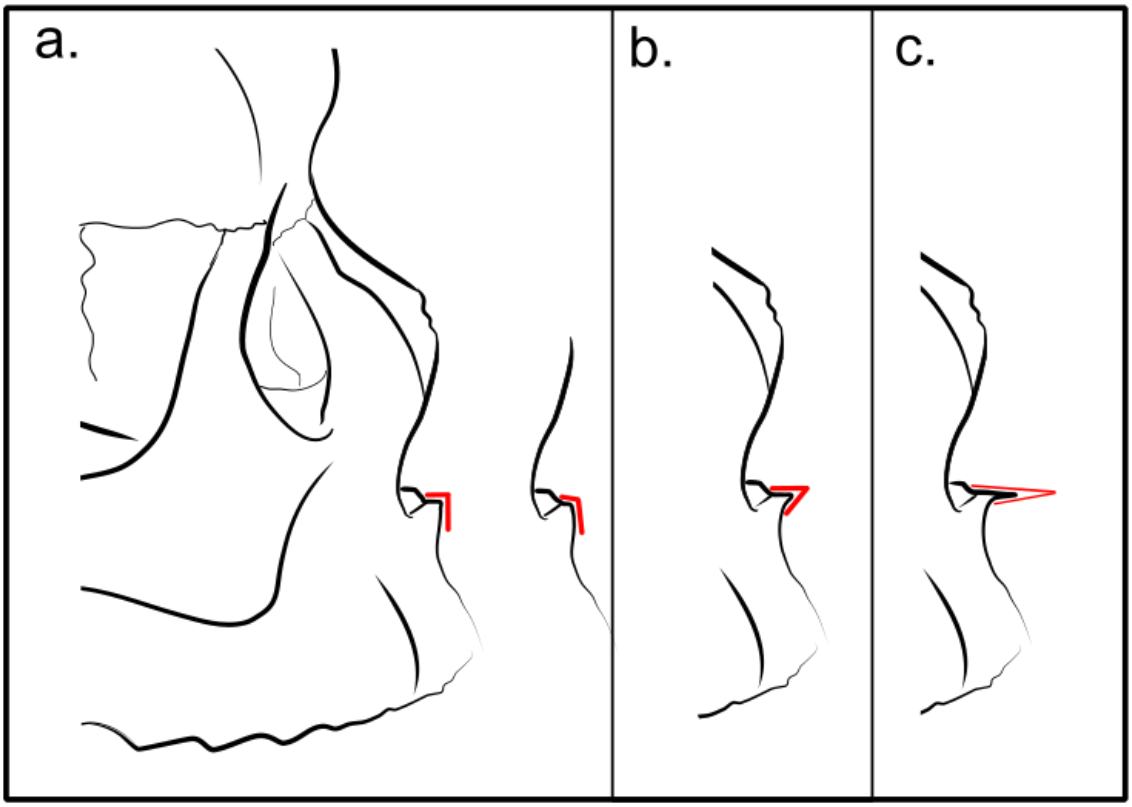

Figure 21: Expressions of ANS and their differences in angle (red line). A score of 1 (a) has an obtuse (right) or right angle (left). A score of 2 (b) has an acute angle (b), and a score of 3 has an even more acute angle where the process is longer than wide (c). Pre-defined angle categories would be determined through study. Lateral view of the skull.

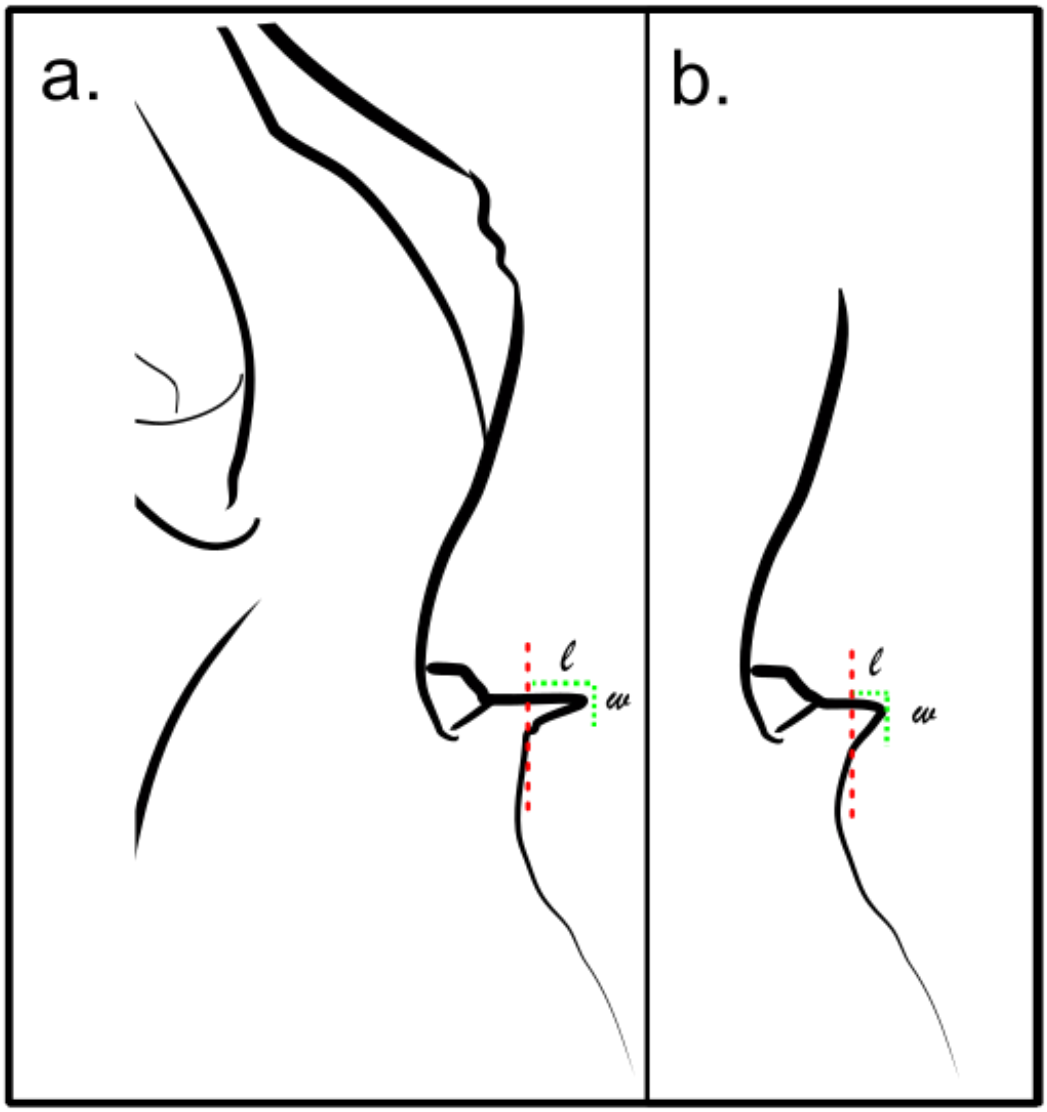

Figure 22: Expressions 3 (a) and 2 (b) for ANS showing the proportions of the length of the process (*l*) to the base (w) of the process (green dotted lines). The red dashed line depicts the point at which the nasomaxillary alveolus surface changes in angle to the ANS, in other words, where the base begins. Lateral view of the skull.

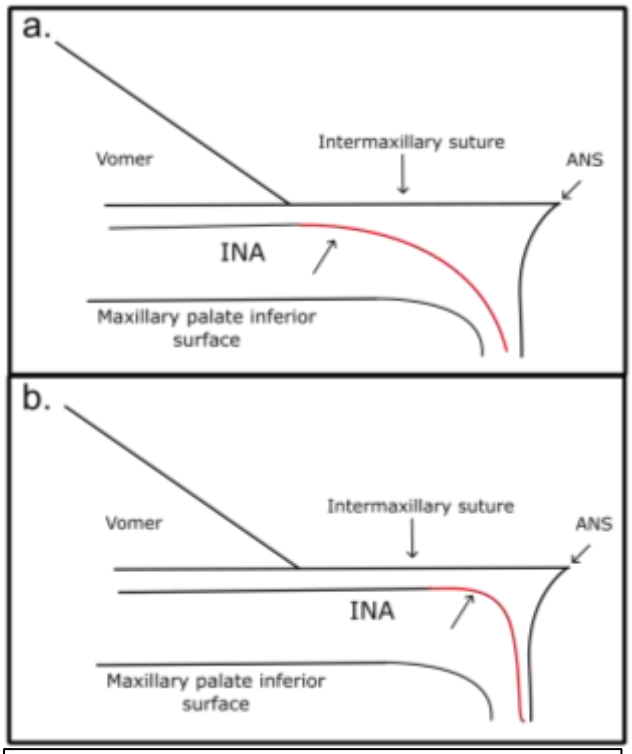

Figure 23: Schematic of what the assumed floor of the INA would look like (red line) in comparison to other features of the nasal cavity (black lines). ANS and Vomer are in midline, whereas INA is on either side of the midline. Red solid lines indicate the expressions 1 (a) and 2 (b) that are described in Hefner and Linde (2018).

The small dips that observer 2 had described for INA during training were still not clear for observer 1 if they were the subnasal grooves or not. It was still unclear where the start of the floor was in relation to other bones to observe where sloping occurred (Figure 23).

To understand where error occurred for IOB, observer 1 visually assessed it without an exact measuring tool to see if visual assessment, even with landmarks in mind, caused a difference subconsciously. The author observed that when visually assessing, it was difficult to focus on where *dacryon* was because the anterior lacrimal crest is the feature closest to this landmark, indicating the margin of the orbit. Any bulging behind this crest that indicated the posterior lacrimal crest extended more laterally than the anterior one was distracting (Figure 24b). This bulge caused observer 1 to think the trait was wider than it actually was when measured with calipers because visually this crest looked like the margin of the orbit (Figure 24). For example, observer 1 noticed that there were a few individuals that, upon initial observation, looked like a clear score of 3, but when using the calipers and the landmarks to measure, it was a score of 2.

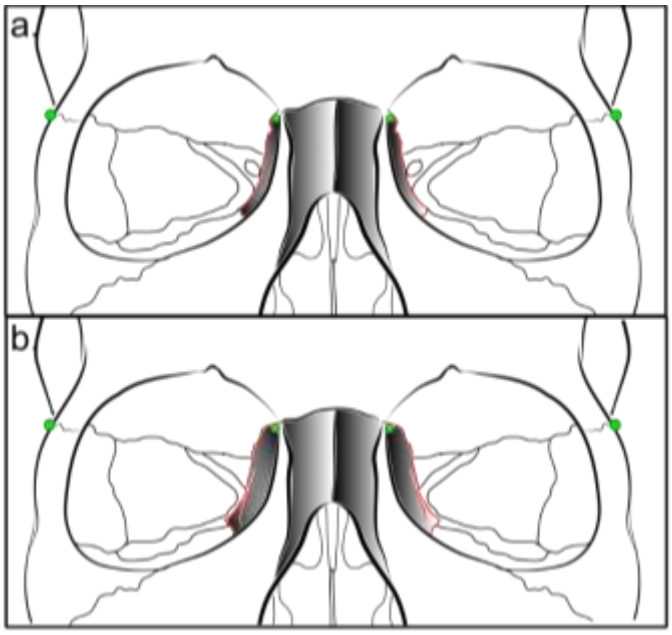

Figure 24: Observed visual difference of bulging at the lacrimals (red outline) that might cause visual bias when scoring. Green circles are the landmarks used when measuring ratios for IOB. Nasals and lacrimals are shaded to show the difference in perceived width of IOB even though landmarks do not change. Anterior view of the skull.

Some individuals' lacrimals were partially absorbed by the maxilla and the suture between these bones occurred more posterior resulting in the landmark being further away and/or wider. Furthermore, individuals with a wider face than at *frontomalare temporale* (*fmt*) would subjectively make the IOB look narrower than it was when measuring *fmt* with calipers, much like NAW (Figure 25). The atlas photographs (Hefner & Linde, 2018) do not mark where *fmt* and *dacryon* are for observers to understand why the individuals were chosen as examples for each trait expression. In the atlas, the line drawings with ratios could also be confusing since the portions are not equal between each section.

Hefner and Linde (2018) had updated the NAW description by adding ratios and landmarks for the width of the nasal aperture as suggested by Kamnikar et al. (2018). However, there is no specification for where to determine the width of the facial region, so the widest part of the face is in different areas for different individuals (Figure 25). Hefner and Linde (2018) did not mark in the atlas where they measured each individual's facial width from so it was unclear why they chose each individual to represent each expression.

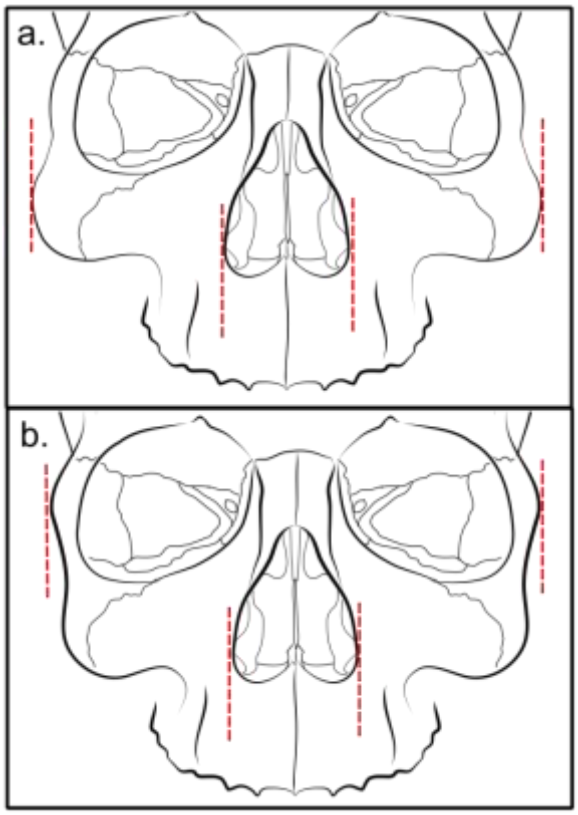

Figure 25: Differences in where the widest part of the facial skeleton is (outer red dashed lines) in comparison to NAW (inner red dashed red lines). Even though the widest part of the face is in two different spots, and at similar width, visually this can affect perceived width. Anterior view of the skull.

The biggest source of confusion in the MT description was what the ‘deepest incurvature’ and ‘deepest anterior curvature’ meant. During data collection, observers 1 and 2 had a lengthy discussion on what it meant and still could not come to a conclusion (Figure 27). One interpretation is placing the ruler edge to the most superior point of the maxilla curvature “incurvature” and then the other edge of the ruler on the most anterior part of the tubercle (Figure 26a). The second way would be placing the edge of the ruler on the lateral surface of the maxilla below the curve, which would be the most medial part of the arch (deepest incurvature), then placing the edge of the ruler on the most anterior part of the tubercle (Figure 26b). Based on

the line drawing provided in *Osteoware* (2020) and the atlas (Hefner & Linde, 2018) it is assumed to be the first interpretation (Figure 26a).

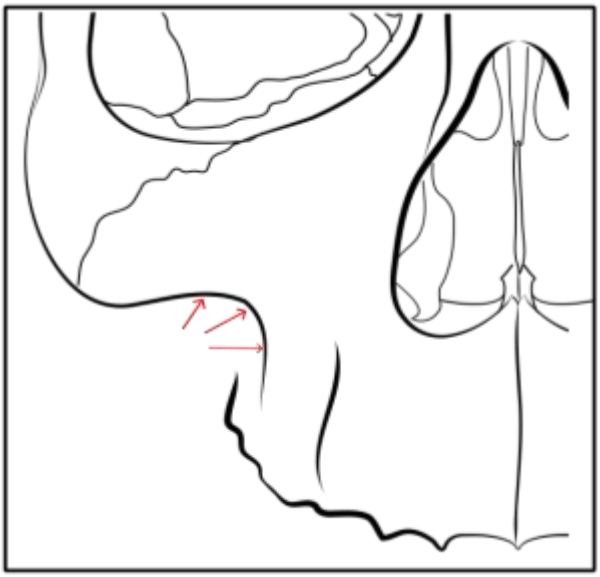

Figure 27: View of the maxillary curve showing where a potential "deepest incurvature" can land (arrows) for MT. Anterior view of the skull.

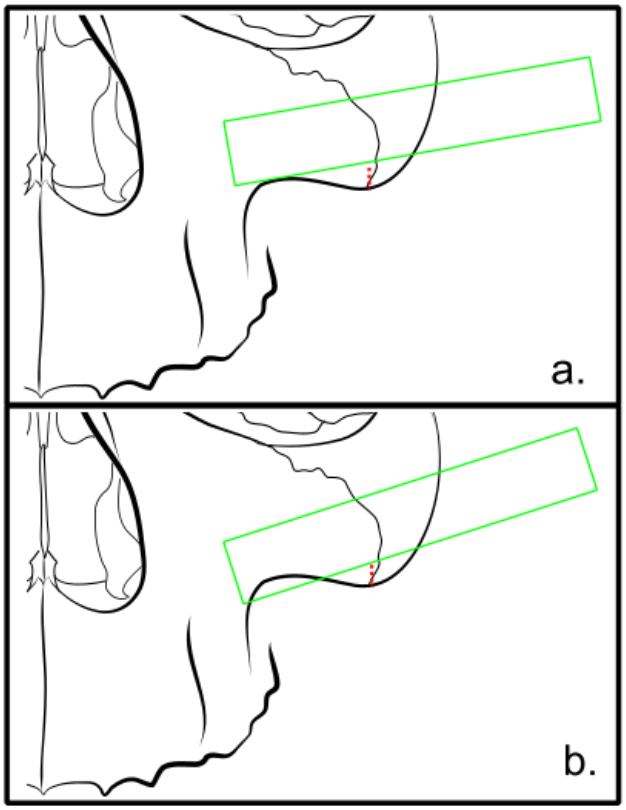

Figure 26: Two interpretations of how a straight edge (green rectangle) can be placed on the "deepest incurvature" for trait MT. One placement of the inferior edge of the ruler can be on the most superior portion of the curve (a) while the other could be placed on the most medial portion of the curve (b). Red dashed line indicates the measurement of the MT based on placement. Anterior view of the skull.

For NBC, the description stated how to use a contour gauge to measure this trait, and provided visuals of expressions through line drawings. However, there was no visual for what the contour gauge should look like for each expression. The line drawings ended the contour at the nasomaxillary sutures, but the description says to measure both the nasals and frontal processes of the maxillae. There is also no indication of what defines a steep wall compared to a

shallow one. This led to another discussion with observer 2 and comparison of several BU individuals. Observers 1 and 2 ended up thinking they all looked the same, even though they were scored differently.

Issues with asymmetry for trait NBS appeared during data collection and there was no indication of how the trait should be scored if one side looked different than the other. Observer 1 continued to have difficulties with what constituted nasal overgrowth (NO) even after looking at photos in the atlas. It was still a question of whether it was growth past the inferior edge of the maxilla, or past the suture dividing the nasal from the maxilla (Figure 20). In one case, Hefner and Linde (2018) states that separation does not constitute overgrowth (p. 149), and in another case looks like separation occurs and is counted as NO being present (Score of 1) (p.153).

For NFS, there was difficulty distinguishing between the triangular (score=3) and round trait expressions (score=1), especially in regard to figures 12.18 and 12.20 in Hefner and Linde (2018). This is because sometimes the superior and inferior edges of the suture would look different (Figure 28, black suture line versus red suture line). Finally, while scoring TPS, there was confusion on what defined slight undulations compared to a significant deviation.

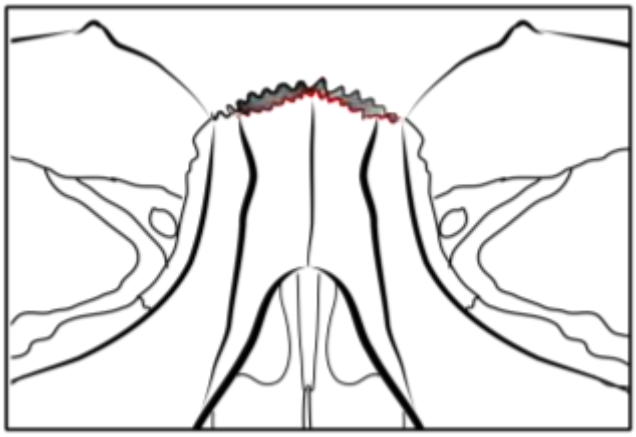

Figure 28: Example of an expression observer 1 scored where the nasofrontal suture had two possible expressions. The red line indicates the inferior edge of the suture at the nasal bones for the NFS, which looks triangular in shape. The black suture line above the red indicated the superior edge of the suture, which was interpreted as a rounded suture shape. The space between the red and black lines was the empty suture space, likely caused by the desiccation of tissue separating the nasal bones from the frontal. Anterior view of the skull.

4.2.3 Observations of trait interpretation with observer 3 after data collection

After observer 3 collected the data on the U of M individuals, another discussion took place with him that revealed there were more differences in technique or interpretation while scoring traits. Observer 3 devised a novel way to score ANS by placing his thumb on the spine and if the pads of the thumb could touch the lip of the INA, then it was scored a 1. If the thumb could be placed under the spine fully, then it was scored as a 3. A score of 2 would be between these two.

In addition to how observer 2 scored INA, observer 3 observed the nasomaxillary alveolus laterally to see if there was a concave curve to the surface. If there was a curve, then the ridge would be visible, and observer 3 would use this characteristic as well as feeling the edge to determine how much of a ridge there was.

During initial discussion with observer 2, the MT score was agreed upon by all observers, therefore, there was no discussion on how to score it. The author discovered that observer 3 placed the top edge of a ruler at the inferior point of the MT and one of the short edges against

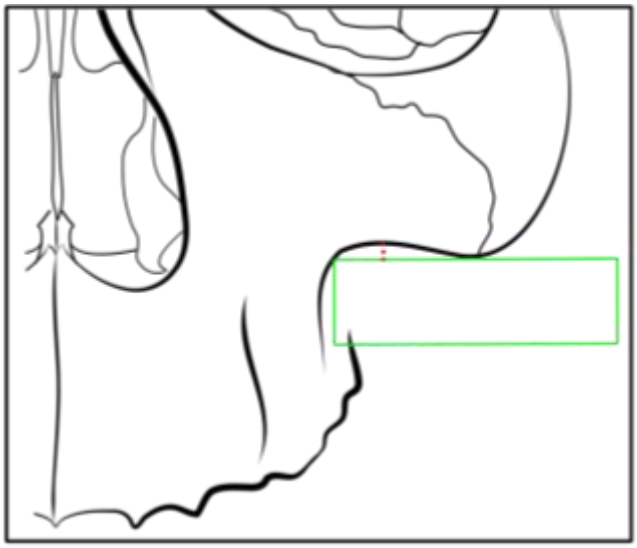

Figure 29: Depiction of how observer 3 placed the straight edge (green rectangle) under zygomatic and measured the height of the MT (red dashed line). Anterior view of the skull.

the anterolateral surface of the maxilla (Figure 29). The ruler highlighted the curve of the inferior edge of the "cheek" maxilla above the ruler and therefore the height of the MT was from the top of the ruler to the most superior part of the curve. Observer 1 did it the opposite way where MT was measured from the bottom of the ruler to the tip of the MT (Figure 26a).

For NBC, observer 3 was confused on where the 1cm was being measured from. It was either 1cm along the surface of the nasal bones (Figure 30, solid red arrow) or 1cm straight down from nasion (Figure 30, dashed blue arrow) then transferred to where it would be on the nasal surface (Figure 30, blue empty circle). The author measured along the surface as it was assumed this was the standard.

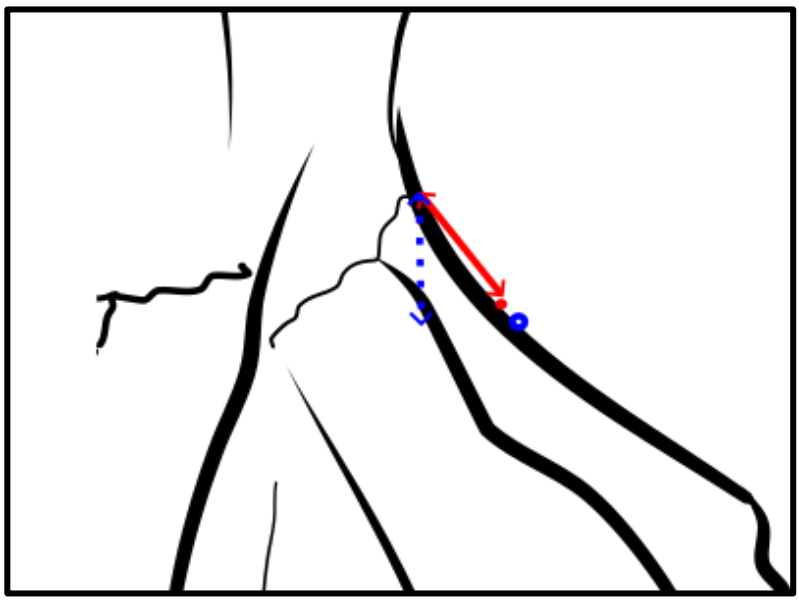

Figure 30: Depiction of how observer 3 was uncertain where to measure 1cm from *nasion* (arrows) and how this uncertainty can affect placement of the contour gauge (circles). Lateral view of the skull, which is facing right, with the orbit on the left side of the drawing. The red solid arrow is measuring 1cm along the nasal bone surface while the blue dashed arrow is measuring 1cm inferior to *nasion* then is translated to the surface of the nasal.

For NBS, observer 3 looked at the suture laterally (Figure 31) rather than anteriorly to see the superior point of the nasomaxillary suture in comparison to the suture pathway. Observer 3 visualized a straight line from the most superior point of the suture where it meets the nasofrontal suture to along the path of the suture. If the suture bent medially (Figure 31, solid arrow) from this imaginary straight line, then it was considered a pinch. If the suture went past the line laterally, then it was a flare (Figure 31, dashed arrow).

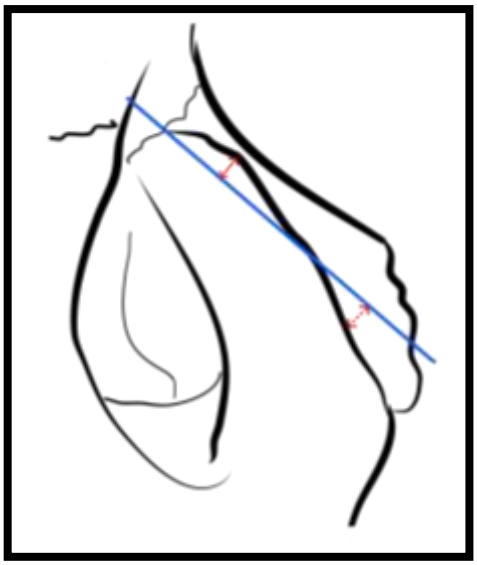

Figure 31: Depiction of how observer 3 placed a straight edge (blue line) to measure whether there was pinching (top solid arrow) and bulging (bottom dashed arrow) for NBS. Lateral view of the skull, with the skull facing right

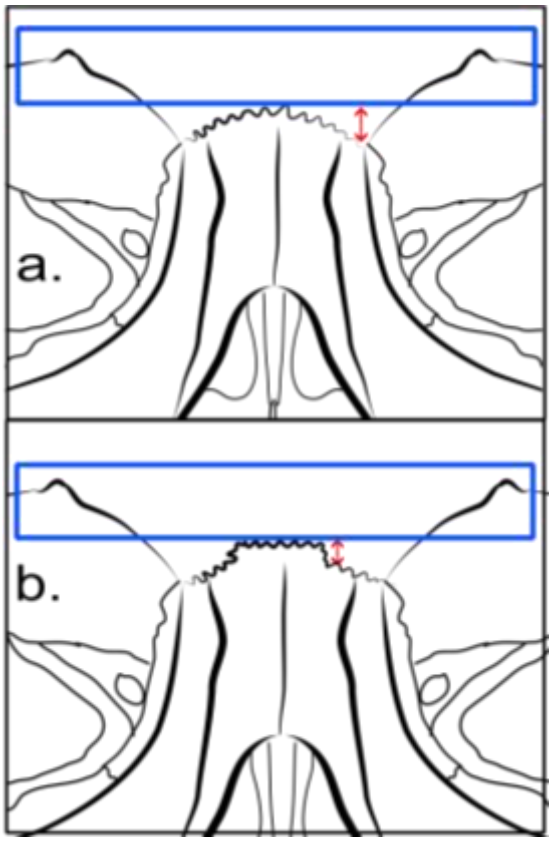

Figure 32: Depiction of how observer 3 placed a straight edge (blue rectangle) to score NFS. Red arrow is indicating how much of the suture is below the line where the suture is at its most superior. The curve in ‘a’ falls away from the ruler quicker than ‘b’ does. Anterior view of skull.

This straight-line technique was also used with trait NFS. A clear ruler was placed lengthwise along the most superior portion of the nasofrontal suture perpendicular to the internasal sutures. The space between the bottom edge of the ruler and *frontomaxillary suture* (*fms*) helped him differentiate expressions based on how much of the *fms* ran with the ruler's edge and what did not (Figure 32). The clear ruler was also used to determine if the superior border of the orbit was angled for OBS. If it was angled, then observer 3 scored it as a 3, but if it was mostly horizontal and the vertical edges were shorter than the horizontal edge, then it was scored a 1.

For PBD, observer 3 described that for a depression to be present, it has to be a concave surface, whereas a depression to observer 1 was a change in curvature from the naturally convex skull; the area could be flattened but interrupts the convexity of the skull looking like a depression (Figure 33).

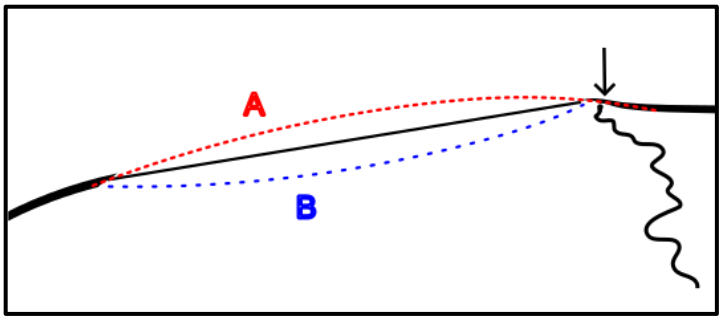

Figure 33: Depiction of what a depression looked like (B) in relation to the natural convex curve of a skull (A). Observer 1 had scored both B and the solid black line as a depression. Black arrow indicates *bregma*. Lateral view of the skull.

There was difference in finding where the landmark *jugale* was for PZT. Observer 3 had *jugale* as where the posterior border of the frontal process **begins** to turn into the superior border of the temporal process (Figure 34, red solid arrow) whereas observer 1 had *jugale* as the **midpoint** between the transition of these edges.

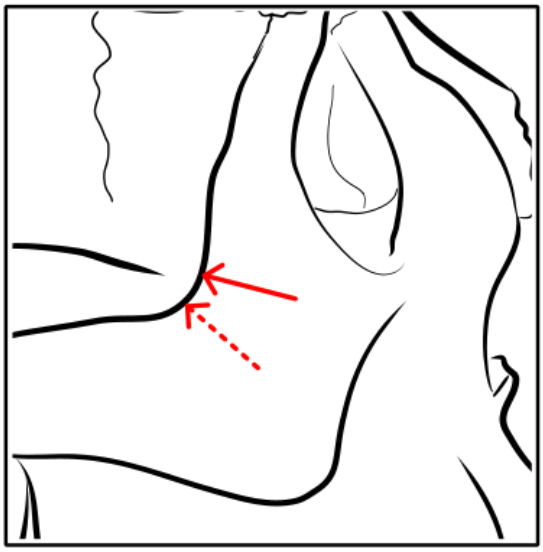

Figure 34: The landmark *jugale* was differentially identified between observer 3 (red solid arrow) and observer 1 (red dashed arrow), thus resulting in different placement of the ruler when scoring PZT. Observer 3 placed the ruler where the curve started and observer 1 placed the ruler at the midpoint of the curve. Lateral view of the skull with the skull facing right.

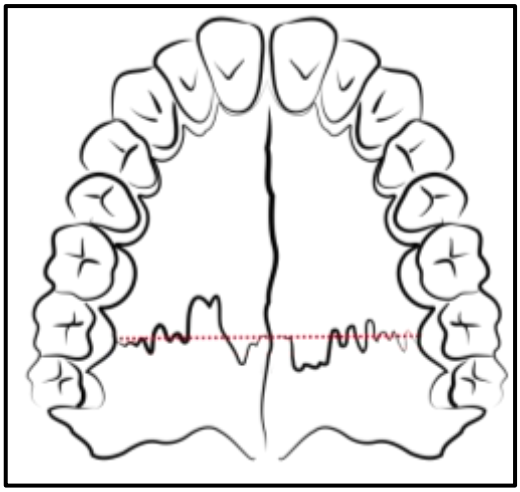

Figure 35: Line of best fit (red dashed line) across the undulations of TPS. This dashed line can be used to determine what a "significant" deviation looks like compared to several smaller undulations.

Finally, another imaginary line was visualized for TPS. Observer 3 visualized a straight line across the suture where most points of the suture cross this line. Then, observer 3 determined that a 'major deviation' away from the line was where the suture did not come back across that imaginary line for longer than the other deviations (Figure 35).

4.2.4 Contour Gauge

A contour gauge helped delineate the walls of NBS, and photos on the contour gauges for NBC and PBD were taken for further comparison. Figure 36 shows that the contour gauge in slightly different places results in variation of contour. A score of 0 was given to individual #1 on the first run, while #3 was given a score of 1 on the first run, but a score of 1 was given to both individuals on the second run.

Figure 37 shows that the contours have slight shape variations but were scored drastically different. Observer 1 scored 1 and observer 3 scored 0 for PBD, and observer 1 scored 1 while observer 3 scored 4 for NBC. Finally, Figure 38 shows the variation in shape for NBC among those that are scored similarly.

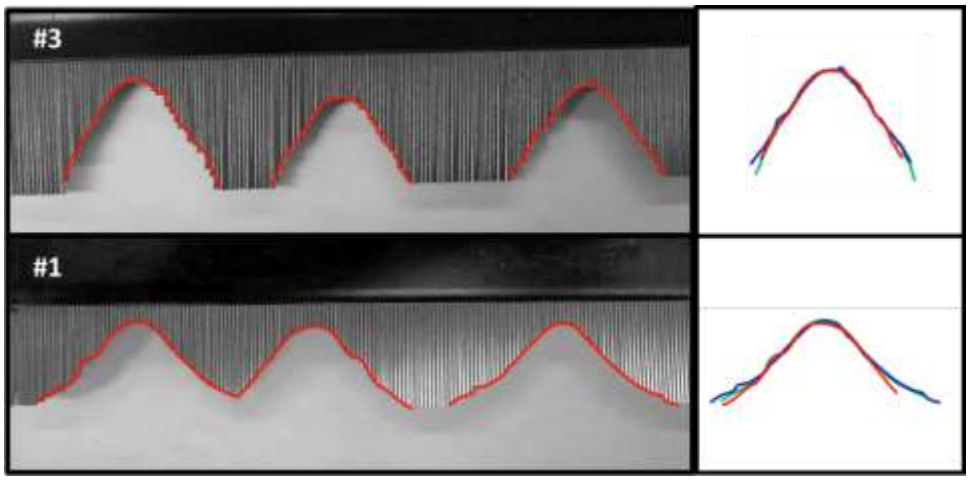

Figure 36: Multiple contour gauge measurements taken at different points of the nasals (red lines) on University of Manitoba individual number 3 (top) and 1 (bottom) by observer 1. Magnification and size are not important in this context, only shape. Red outline is highlighting the contour, and the right most image is superimposing these outlines onto each other, showing there are slight deviations, but generally the same shape. These two individuals were both scored as '1' by observer 1 in scoring period 3.

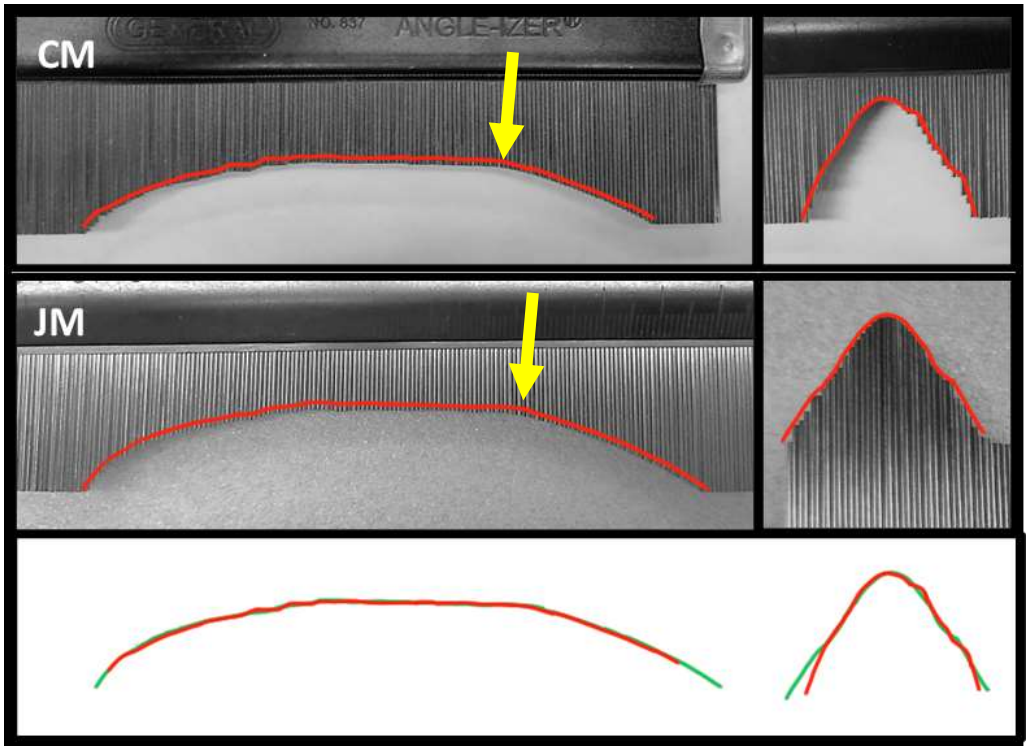

Figure 37: Contour gauge measurements of PBD (left) and NBC (right) from observer 1 (top) and observer 3 (bottom) on individual #3. The contour of the superior surface of the skull (left images) are of the skull facing right, therefore, *bregma*, is located at the arrow. Size and magnification are not important in this context, only shape is. Red line is highlighting the shape of the contour, and the bottom image is superimposing the two contour lines; observer 1 (red) and observer 3 (green). Observer 1 scored the individual as '1' and observer 3 scored the individual as a '0.'

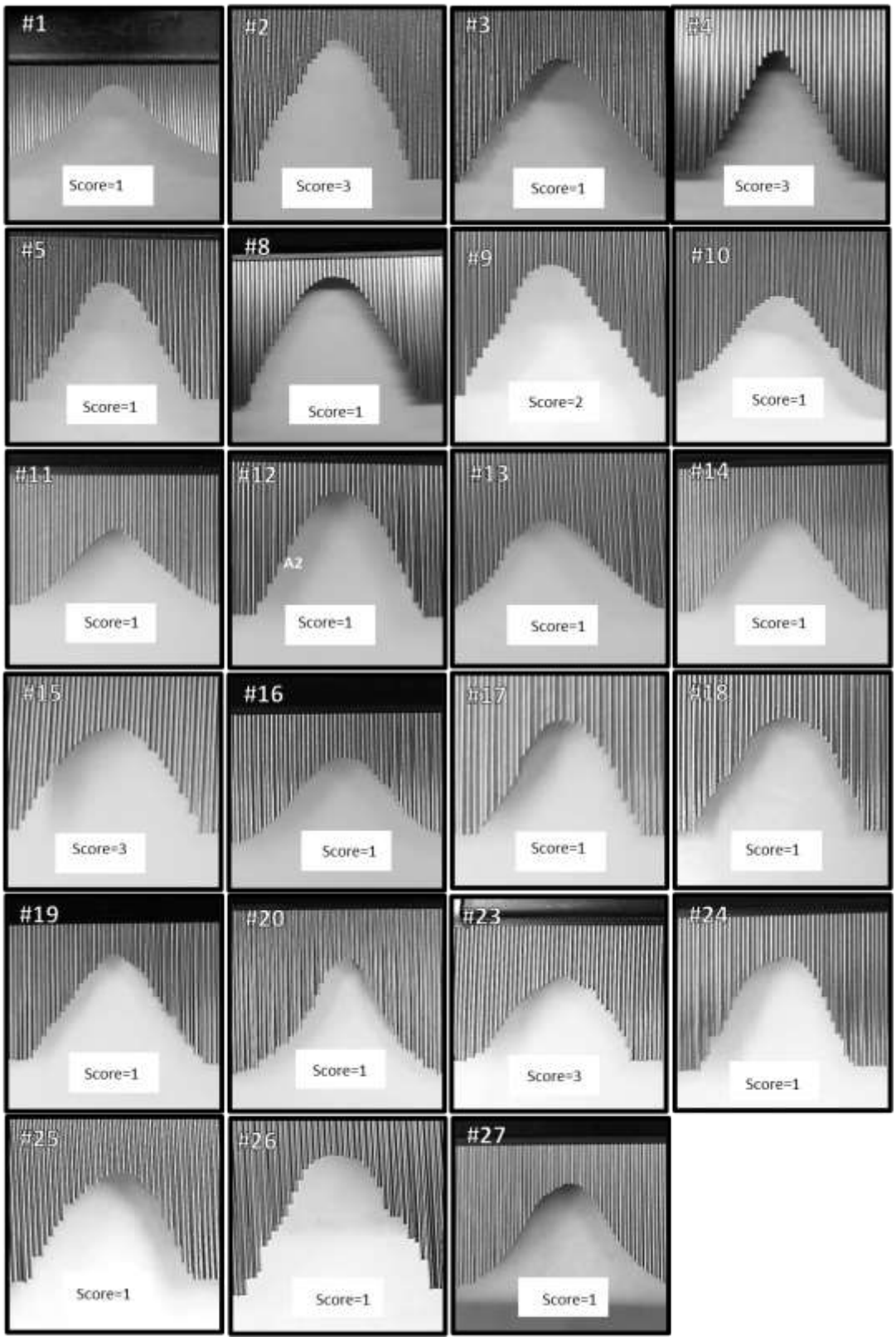

Figure 38: Contour gauge measurements for NBC on each individual by observer 1, scoring period 2; individual numbers are located in top left of each image. Magnification and size is not important in this context, only shape. This image shows the variation in shape among individuals scored the same value.

4.3 Ancestry estimates

The effect that observer error has on the scores is one thing to keep in mind, but the effect that these scores have on the ancestry estimation is another major reason to test this method. If observer error caused score changes so that the ancestry estimates are changed more often than not, this is an issue because disagreements between anthropologists' estimates can change the course of the case since ancestry estimates are used to narrow down missing persons profiles.

4.3.1 Intra-observer data (observer 1's scoring period 2 vs. scoring period 3)

Overall, ancestry estimations were consistent within each program between scoring periods 2 and 3. Within the *HefneR* app, both sets of data resulted in the same ancestry result 60% of the time (n=25), whereas *MaMD Analytical* resulted in the same result 75% of the time (n=24). Only 10% of ancestry assessments agreed between applications (*HefneR* vs *MaMD*) with observer 1's data.

When using both FDB and Howells to run the metric data in *Fordisc*, the ancestry estimate agreed 52% of the time between the two databases. When using scoring period 2's scores, 29% of individuals (n=24) had ancestry estimates that agreed between *Fordisc* and *MaMD*, while 20.8% agreed between *HefneR* and *Fordisc* (n=24). When using scoring period 3's scores, 34.6% of individuals (n=26) had ancestry estimates that agreed between *Fordisc* and *MaMD*, while 19.2% (n=26) agreed between *HefneR* and *Fordisc*. When comparing ancestry estimates from scoring period 3 data between all programs, they only agreed on proposed ancestry 12% (n=25) of the time, which is less than that of chance when comparing four major groups.

Each databank in *Fordisc* had their resulting estimates compared to the other application estimates rather than choosing an ancestry estimate from one of them. Comparing FDB results to

scoring period 3's *HefneR* and *MaMD* results ended up with 19.2% and 23.1% agreement, respectively. Howells results showed slightly higher agreement when compared to scoring period 3's *HefneR* and *MaMD* results, having 26.9% and 34.6% agreement respectively.

4.3.2 Interobserver data

Comparing each observers' results from *HefneR* and *MaMD* had variable agreement. The results of the *HefneR* app for observer 1 and observer 3 agreed 50% of the time (n=10) while the *MaMD* results agreed 30% (n=10) of the time. In fact, there was so much disagreement between observers 1 and 3 that some individuals had all four ancestries estimated across both applications and both observers. When comparing observer 3's results between *HefneR* and *MaMD*, they agreed 20% of the time (n=10). When it came to observer 2's results, they agreed with observer 1's results 43.8% (n=16) of the time in *MaMD*, whereas with *HefneR* they agreed with observer 1's results 40% of the time (n=15). The results between *HefneR* and *MaMD* for observer 2 agreed 18.8% of the time (n=16).

Comparing the morphometric data results to the metric data results showed low agreement. When comparing the ancestry estimations from observer 3, he had 10% of ancestry estimates from *HefneR* agree with *Fordisc* while observer 1 had 0% agree (n=10). Comparatively, observer 3 had 0% of *MaMD* results agree with *Fordisc*, while observer 1 had 30% agree (n=30). In contrast, observer 2 had 43.7% of estimates from *HefneR* agree with *Fordisc* (n=16) while observer 1 had 31.3% of estimates agree (n=16). Observer 2 had 19% of *MaMD* estimates agree with *Fordisc* while observer 1 had 37.5% of estimates agree (n=16)

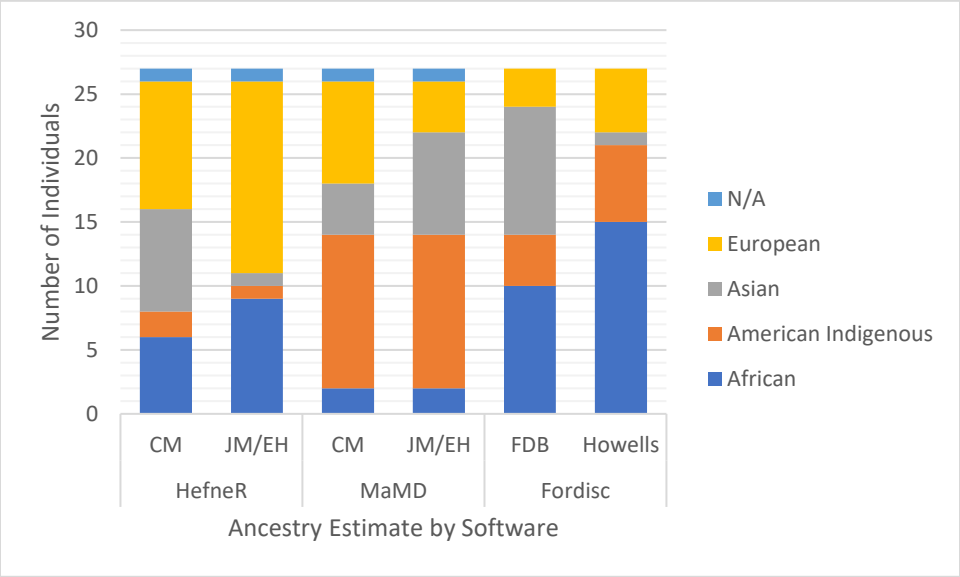

Figure 39: Proportion of ancestry assessments given by software using observer data (Observers 1 (CM), 2(EH), 3 (JM)).

Figure 39 shows that each program has different proportions of ancestry estimates and similar proportions between observers within each program. Discrepancies in the number of individuals used for analysis comes from the programs not being able to place an individual into an ancestry estimate.

4.4. Summary of results

Training increased the score agreement for observer 1 across scoring periods as seen with increased kappa values across the majority of traits (Table 10). However, there were issues of bias and prevalence as seen with the Bias Index (BI) and Prevalence Index (PI) (Table 10). Observers 1 and 2 agreed more often in scores than observers 1 and 3 (Table 13), and disagreements that were more than one score apart occurred more often across all traits between observers 1 and 3 than between observers 1 and 2. Greater issues of bias were present between observers 1 and 3 than between 1 and 2, and prevalence was an issue for all observers (Table 13).

Patterns of agreement were mostly trait specific, but the type of traits, nominal versus ordinal, did have patterns. For example, nominal traits had more scores disagreements that were further apart than one score than the ordinal trait disagreements.

To determine why there were patterns of disagreement, qualitative assessment for each observer's assessment revealed there were differences in interpretation and with technique for assessment. For example, observer 3 placed the ruler's edge under the tubercle and against the maxillary surface for MT (Figure 29) whereas observer 2 placed the ruler's edge above the tubercle and above the maxillary curve (Figure 26). Qualitative assessment also revealed that descriptions used vague words, such as "generally," "facial width" instead of precise landmarks, and "deepest incurvature" instead of precise directional terms. Finally, some atlas photographs were unclear and did not improve the interpretation of the line drawings.

In regard to ancestry estimates that were generated from the scores, disagreement in scores changed some of the ancestry estimates. When comparing ancestry estimates from two programs that generate estimates from morphoscopic trait scores, *HefneR* and *MaMD Analytical*, the results did not agree the majority of the time. When comparing morphoscopic estimates to those generated from metric traits (in *Fordisc*), the results agreed more often with metric estimates than between the morphoscopic estimates, with Howells database (within *Fordisc*) resulting in more estimates agreeing with the morphoscopic estimates than the Forensic Databank (FDB). Finally, observer 1's *MaMD Analytical* ancestry estimates agreed with *Fordisc* estimates more often than observer 2's *MaMD Analytical* ancestry estimates, but observer 2's *HefneR* estimates agreed with *Fordisc* estimates more often than observer 1's *HefneR* estimates.

Chapter 5: Discussion

5.1 Observer error

5.1.1 General trends

This external validation of Hefner's (2009; Hefner and Linde, 2018) scoring method has provided insight regarding where error is occurring in addition to how error can be reduced to improve consistent application among observers. The first hypothesis indicated that the results of a repeatability test would show high score agreement ($k \geq 0.61$). The null hypothesis stated that agreement is occurring at a rate of chance ($k=0$), meaning the scoring method is not increasing agreement or causing disagreement within one observer. It was found that intra-observer agreement was generally high (Table 10, CM2 vs. CM3) as indicated by Table 4, meaning agreement occurred at a rate greater than that of chance (Table 10, $k \neq 0$), therefore, the null hypothesis was rejected. In contrast, the second hypothesis indicated that the results of a repeatability test would show low score agreement between observers as compared to intra-observer agreement. As with the first hypothesis, the null hypothesis stated that agreement is occurring at a rate of chance meaning the scoring method is not increasing agreement or causing disagreement between observers. It was found that there was low score agreement between observers (Table 13; $k<0.61$), and lower agreement than that of intra-observer agreement (Table 10, scoring period 2 vs scoring period 3) meaning error is occurring at a higher rate or greater magnitude between observers. Therefore, the null was rejected.

The reason for low agreement between scoring periods 1 and 2 as compared to scoring periods 2 and 3 was due to the lack of experience of observer 1 during scoring period 1, as well as confusion about the descriptions, resulting in numerous questions on how to score each trait. Intra-observer agreement was improved after training as seen with the combination of higher

kappa values and less bias in the dataset of scoring periods 2 and 3 as compared to scoring periods 1 and 2 (Table 10). This improvement was expected because multiple anthropologists already recommend training prior to use (Kamnikar et al., 2018; Klales & Kenyhercz, 2015) since training is intended to provide application consistency among observers. The improved score agreement in this study can be used as additional evidence that training is needed prior to using Hefner's (2009; Hefner and Linde, 2018) scoring method for case work.

Recommendations for training (Kamnikar et al., 2018; Klales & Kenyhercz, 2015) are also validated through the presence of higher kappa values and less bias among experienced observers (observer's 1 and 2) than between an inexperienced (observer 3) and more experienced observer (observer 1) (Table 13). These results are consistent with Klales & Kenyhercz (2015) who found that observers with experience had a greater number of traits higher in agreement with each other than between an experienced and inexperienced observer. As suggested by Hefner (2014) and Kamnikar et al. (2018), experience with the traits will increase the ability to detect subtle expressions, thus shifting scores from the extremes to intermediate values (Kamnikar et al., 2018). Similarly, observer 3 had more extreme scores compared to observer 1, thus impacting the distance between scores (Appendix 2). Shifting from extremes to intermediate scores also occurred in many ordinal traits in the scoring periods 1 and 2 for observer 1 (IOB; MT, NAW, NBC, PZT). These score shifts from the untrained to trained sessions is a reflection of understanding a greater range of expression after extensively studying images and casts. The exception for inexperience resulting in extreme scores was NBC with observer 2, where five of the eleven disagreements were more than one score away, as opposed to two of the six NBC disagreements with observer 3. This exception is likely due to completely different techniques used for scoring: observer 1 and observer 3 using the contour gage and observer 2 using her

fingers for feeling the trait. There were also a number of disagreements between observer 1 and observer 2 on whether a trait could be scored or not, especially for NO that accounted for four of the twelve disagreements across all sixteen traits. This would indicate some inexperience with how to differentiate broken facial bones from naturally formed edges, and emphasizes the importance of broader osteological training, not just in how to score these traits.

While training can improve observer agreement, method specific training that is recommended under quality assurance programs (Fleischman et al., 2019) may not be available until discipline standards are put into place. A trained osteologist should be able to read the method and apply it with few difficulties since it relies on previously learned skeletal landmarks, features, and tools. However, this study shows that score agreement across most traits is much lower among observers than intra-observer agreement (Table 10;Table 13), which is consistent with previous research intra-observer and (Coelho et al., 2017; Dinkele, 2018; Hefner, 2009; Hurst, 2012; Kamnikar et al., 2018; L'Abbé et al., 2011; Moffit, 2017; Wang, 2016) inter-observer research (Coelho et al., 2017; Dinkele, 2018; Hefner, 2009; Klales & Kenyhercz, 2015; L'Abbé et al., 2011). A higher internal consistency than external consistency seems reasonable because someone who sees a shade of colour as green while another sees it as blue will each continue to score the shade the way they see it. In other words, they will score with high consistency, but have internal bias. It is not until these observers come together and discuss how they determined the shade of colour that will reveal the error or why they see it differently (ex. biological difference in the perception of colour). In this example, correcting the issue can be through metric means, such as using a spectrophotometer, or through a discussion to determine how they will work towards scoring that particular shade as the same score. This low agreement among observers across multiple studies, including this research, shows that Hefner's (2009;

Hefner and Linde, 2018) scoring method needs description and pictorial improvement before it can be used consistently between observers.

In order for an osteologist to apply the method accurately and reliably, the inclusion of specific landmarks in descriptions and drawings that clearly illustrate the location of specific landmarks for visual reference is needed, as are descriptions on how to view the trait, such as the angle at which it is observed, or how to hold an instrument. For example, instructions for how to hold or position the skull to conduct measurements are included in the guide for standard metric measurements (Langley, Jantz, Ousley, Jantz, & Milner, 2016b), which makes data collection consistent. The structure of this guide can be a framework for morphoscopic traits because the collection of metrics boasts low error among observers (Langley et al., 2018; Liebenberg & Krüger, 2020) with only certain types of landmarks producing more error in measurements than others (Ross & Williams, 2008). Without specific instruction on where landmarks are in relation to other features, and how to measure between them, metric analysis would have far greater error. With each revision, it is clear that the descriptions for morphoscopic traits are becoming more precise (Hefner, 2009; Hefner & Linde, 2018; Hefner et al., n.d.; Wilczak & Dudar, 2020). For example, the implementation of ratios in Hefner & Linde (2018), that were proposed for IOB and NAW by Kamnikar et al. (2018), have improved the ability to understand how to score this trait. Further recommendations for how to score the trait, specifically in technique used to visually score it (ex. comparison to a ruler edge), can bring this guide closer to the one used for metric assessment, which is considered a discipline standard.

Kappa values provide an indication of score agreement, but where the disagreement occurs is more crucial to understanding how to reduce or remove the error that alters scores. Bias among observers (outside of inexperience) and prevalence of trait expressions within populations

are issues that fail to be discussed or mentioned in many of the current publications on Hefner's (2009; Hefner and Linde, 2018) scoring method (Andrade et al., 2018; Coelho et al., 2017; Dinkele, 2018; Hefner, 2009; Hurst, 2012; Klales & Kenyhercz, 2015; L'Abbé et al., 2011; Moffit, 2017; Wang, 2016). Kamnikar et al. (2018) was the only study to notice that bias was introduced through various means, such as when photographs were released for use in the *Osteoware* user manual (Hefner et al., n.d.), whether the practitioners used a contour gauge or not, and whether the alteration of scoring scales occurred between iterations of the method. Meanwhile, three studies only briefly mention that their different study individuals resulted in different frequency distributions than found in Hefner (2009), but did not discuss how prevalence might have affected agreement (Hurst, 2012; Klales & Kenyhercz, 2015; L'Abbé et al., 2011). It is suggested by Byrt, Bishop, and Carlin (1993) that these issues must be discussed when testing observer agreement. If possible, BI and PI indices should be used to determine if bias or prevalence are present because these issues can be a source of error contributing to low inter-observer agreement.

In this research, experience bias was heavily apparent between observer 1 and observer 3 even after training on how to score the traits with observer 2 (Table 13). This is seen with both the lower kappa values (Table 13) and the increased frequency of scores that were further apart in disagreement (across all traits) with observer 3 (Appendix 2). These results are consistent with inexperience in the range of expressions, thus not being able to differentiate intermediate scores from each other. The bias was also less extreme between observer 1 and observer 2 because observer 1 had more time to study the traits and their expressions than observer 3 did. These results re-emphasize the importance of method-specific training as well as practice with the method on many individuals to understand expression variation.

The training discussion was useful for identifying error because it provided insight to where observer 1 was misinterpreting or interpreting the trait differently. Observer 2 also clarified some areas of confusion that observer 1 struggled with by explaining what the descriptions were trying to encompass. For example, explaining what the pinching and bulging described for NBS meant helped observer 1 differentiate what these features were whereas before training observer 1 was guessing what it meant. Training was also useful for teasing apart which traits or expressions caused the most difficulties in scoring between observers. For example, the most difficult trait to interpret was INA whose expressions show a gradual slope change from the nasal floor to the vertical maxilla and the introduction of a sill. The discussion with observer 2 showed observer 1, before scoring period 2, that there can be a slight dip right before the increase of the ridge for expressions 4 and 5 (Figure 17c), which observer 1 was ignoring in scoring period 1 since the ridge the dips created felt like it was in line with the rest of the nasal floor (Figure 17d). Knowing this, observer 1 was able to score it how observer 2 scored it throughout scoring period 2 and 3, thus reducing differences in scoring. This discussion was likely the reason for moderate agreement because the photographs in Hefner and Linde (2018) were not particularly helpful for showing each INA expression. It was difficult to gauge depth and where internal nasal cavity features were located to understand why the individuals were chosen as photographic examples. Without knowing why they were chosen as examples, it made it difficult to compare individuals to these photographs. The descriptions also use comparative wording, such as “more angulation,” meaning you would need to see both expressions to determine which one is more angled. Further discussion on how each trait expression was determined could have taken place and likely would have reduced bias between observer 1 and observer 2 even more (Table 13).

Bias can also be in the process of determining the score itself. Kamnikar et al. (2018) saw that new technologies could alter score frequency, such as the use of a contour gauge for NBC. Using fingers to score NBC is a form of technical bias because differences in finger size or sensory differences may affect how a narrow plateau is perceived. This is where the qualitative aspect of this research is useful because observers describing their technique for measuring the trait or how they determine a narrow plateau from a broad one can highlight why differences in scores could occur. Moreover, vague descriptions seemed to facilitate the search for novel techniques that would be helpful in visually scoring the trait. For example, observer 3 used a straight edge to determine NFS shape as opposed to scoring without it (Figure 32). Even if the majority of results between observers are the same, using different tools or techniques to help with visual analysis introduces bias. Since discussion only occurred on traits where there was a disagreement in score, how to score several expressions outside of the agreed scores were missed with observer 2.

Bias can also be in the training individuals had or any previous methods they are familiar with (Rhine, 1990). For example, NAS and NAW used to be one trait, and NBC and NBS used to be one trait (Hefner, 2003). This old description of traits might cause confusion if an observer is familiar with them. For example, observer 3 was trained on identifying a suite of nonmetric traits and seemed to resort to techniques he was trained with, such as using his hands as tools, when the descriptions were too vague (e.g., ANS). More detailed descriptions, or explicit instructions could curb resorting to old habits and therefore limit bias.

Bias between observer 2 and observer 1 likely occurred from the difference in the iteration of the method. Observer 2 used the guide from Hefner's 2009 article that included eleven traits with descriptions that differed slightly from the updated ones that observer 1 used,

which included ratios and landmarks as well as sixteen traits (Hefner & Linde, 2018). Observer 2 used this iteration because these traits correspond to the ones included in the online application *HefneR*; the other five traits do not have practical application so recording them was not necessary or habitual. This is relevant because not all practitioners using Hefner's (2009) scoring method will be updated on how to score the traits, especially when a handful of traits have not been formally defined before the introduction of the MaM Databank (Hefner, 2018). When observer 1 began her research and initial practice without training, she was mainly using Hefner (2009) and *Osteoware* for descriptions. Finding the different iterations (Hefner & Linde, 2018; Hefner et al., n.d.; Wilczak & Dudar, 2020) outside the journal publication took some time and effort, which is likely not the time many practitioners have.

Populations that the observers were trained on may affect how traits are scored on populations outside of the ones with which observers are familiar. In other words, they can become experienced in the range of expression in relation to scoring categories, but not the total variation of, specifically, the score of 3 across populations. The line drawings are an ideal or generalized shape of the score, so observers are placing an individual into a category with which they are most similar. If the trait expressions do not always present themselves exactly the same as the line drawing, but a description can be clear enough to capture everything that characterizes or does not characterize an expression, then how that expression presents itself in different populations should not affect scoring. For example, if measuring the facial width for NAW were to occur at the widest part of the facial skeleton, a demarcation of where the widest points could occur can help reduce guessing. In one population, the widest part of the face could commonly occur at the inferior edge of the zygomatic process of the frontal, whereas another population could have the widest part of the face on the zygomatic body. A description such as "NAW in

relation to the widest part of the facial skeleton, *occurring anywhere between landmarks 'fmt' and the MT*," denotes a particular area that this feature can occur in, and excludes the widest part from occurring above *fmt* where it might be argued is no longer facial skeleton.

Observer error studies for Hefner's scoring method do not talk about how prevalence may have affected agreement. Frequency distributions to determine the prevalence of traits (Hurst, 2012; Klales & Kenyhercz, 2015; L'Abbé et al., 2011) but there is no discussion on how the prevalence might be affecting how the traits are scored. For example, Coelho et al. (2017) have precise agreement for NAW, NO, and PBD, but it is unknown if all the expressions were present. There may have only been one expression on all the individuals and that one expression was the one each observer agreed on. The observers then did not have to decide between the various possible scores if multiple expressions were not present. The same goes for the agreement in the current study. The prevalence of INA expressions 1 and 2 were low compared to 3,4, and 5, therefore, the agreement may have been high with observer 2 because only three of five possible expressions were present to score, thus reducing the chances of disagreement.

The patterns observed in score disagreement might divulge where error is occurring for each trait. In general, it was expected that ordinal traits more commonly had score disagreements only one score away from each other because each new expression is dependent on the previous expression (Carson, 2006; Grüneberg, 1952). It is more difficult to determine the difference between an INA score of 4 or 5 as opposed to a score of 1 or 5 since 1 and 5 are on opposite ends of the phenotypic spectrum. Both 4 and 5 are considered to have a ridge, whereas 1 does not have a ridge. Since these expressions are on a graded scale, confusing whether the ridge is "weak" or "prominent" is more likely to occur than whether there is a ridge or not. Since there is no order to nominal traits, it also makes sense that the disagreed scores are further apart than

ordinal traits because one expression does not require the presence or absence of a characteristic in a previous expression.

The effect these score disagreements have, if any, on ancestry estimates may be different between the type of trait. Small disagreements in ordinal traits may not affect the estimate if they follow a normal distribution, such as occurs in expressions from Hefner (2009). Since there is scale of additive expression (Carson, 2006; Grüneberg, 1952), frequencies will be decreasing as expressions move away from the most frequent expression. Subsequently, a score that is only one category away from another score should not drastically change the probability of an individual belonging to that group; ancestry estimates are probabilities based on expression frequencies. An expression that is close to the most frequent expression can still have a high probability of the individual coming from that population.

Since there is a predicted pattern to expression frequencies in ordinal traits, it is reassuring to know that disagreements further apart than one score were rare. It would make sense that an ancestry estimate would more likely change with greater disagreement because the frequencies decrease in a predictable pattern from the most frequent expression; therefore, a reduced probability of them coming from that population. In contrast, a change of score in nominal traits might drastically change the estimate since it is unknown how these expressions affect probabilities. For example, an individual with ordinal expression 1 can still have high probability of belonging to a group with ordinal expression 2 since expression 2 is closer to 1 than it is to 0. In nominal traits, this is not the case, so it is important to consider since a change of ancestry estimation could alter the outcome of a forensic anthropology case.

5.1.2 Specific Trait Trends

Outside of these general patterns, there are varying patterns of disagreement within each trait that differ in their source of error, such as the technique for scoring or ambiguous descriptions. Therefore, each trait and its expressions are discussed in detail separately. Different patterns in disagreement result in different needs for improvement. The recommendations for each trait description will largely be based on the patterns observed in observer 1 and observer 2's data since there was more experience between these two than with observer 3. However, observer 3's adaptive techniques to measure traits that he was unsure about ended up contributing to the recommendations on how to score some traits. Sometimes, an observer with no familiarity with the method can provide a unique perspective on a problem.

The following descriptions and images are those that are included in Hefner and Linde (2018) because this is the most updated version of the method. Recommendations for improvement will be based on these descriptions. However, the component of the trait that links to error can relate to either the description or images in Hefner (2009) and/or Hefner and Linde (2018) depending on each observer's method.

5.1.2a Trait ANS

Patterns of disagreement

The bias found for the assessment of more scores of 2 and 3 among scoring periods 2 and 1 (Table 14) respectively, and observer 1 and observer 3 (Table 57), respectively, can be attributed to inexperience. The literature says that those with less experience are more likely to score on the extreme ends of the spectrum (ex. score of 3) because they do not have experience with the whole spectrum (Kamnikar et al., 2018; Klales & Kenyhercz, 2015). The inexperience of observer 3 was likely comparable to that of observer 1 in scoring period 1 because he had only

read Hefner's 2009 guide and some research on inter-observer agreement. Even though observer 3 had attended the training session, experience in the range of expression was missing since he did not do as extensive study as observer 1 did up until scoring period 3.

Andrade et al. (2018) also noted in their digital analysis of this trait, which was comparing scores generated via computer analysis to scores generated from an observer, that three disagreements occurred between scores of 2 and 3. They attribute this disagreement to possible fragmentation, which may also be the case for the current study if fragmentation was not clear. The tip of the ANS is very fragile and could easily be broken without detection.

Disagreements could also be due to the vague descriptions of 'moderate' and 'pronounced'; both observer 1 and observer 3 did not use the atlas photos while scoring. Reviewing trait expression photos from several publications was helpful in differentiating the two expressions because every score of 2 changed to a 3 between scoring periods 1 and 2, and remained consistent from scoring periods 2 and 3. Observer 1 felt confident scoring ANS after viewing what expressions of 2 and 3 were on photographs of other individuals. However, the description and line drawings on their own were not clear because there were difficulties determining what constituted a "moderate" or "pronounced" projection. The vague descriptions resulted in observer 3 resorting to the use of his fingers to determine the length. It is not recommended that individuals use their body parts for measuring since every body is different and it cannot remain consistent between observers. This subjective way of scoring likely was the cause of the disagreements further apart in score. No pattern of disagreement present with observer 2 could indicate that disagreement was due to chance rather than a specific reason, such as with observer 3. Otherwise, it may be due to fewer disagreements, thus an inability to detect a pattern.

The description and images have not changed since Hefner (2009), and Hefner (2009) himself saw low intra-observer agreement (k=0.42) in comparison to the other traits (k>0.8), indicating a higher error rate for this trait in comparison to others. This trait has variable agreement (thus underlying error) across multiple studies (k=0.42 to 0.81), likely due to prevalence and bias. This means there cannot be any cross-study comparison because kappa agreement values in one study may be calculated from the presence of one out of three expressions, whereas another study may have the kappa values calculated from the presence all three expressions. However, the low intra-observer agreement in Hefner's (2009) original study, someone who is well versed in the method, can provide insight that this trait needs some major improvement in description or depiction.

Recommendations for description/pictorial improvement

Since the score of 1 was agreed upon among all scoring periods, it does not seem necessary to improve the description or line drawing for the expression of 1 for this trait. However, descriptions for scores 2 and 3 need revision. Upon further study of the atlas photos, it is possible to see some characteristics that Hefner (2009; Hefner and Linde, 2018) might attribute to an 'intermediate' or 'pronounced' ANS even if not described, such as a comparison of the spine length to the spine base as well as the angle of the spine tip. The descriptions can be improved as follows, with consideration that the specific numbers may not be correct, instead, used as a general recommendation of how the description can be more specific.

While a score of 1 appears to be agreed upon, to bring it in line with the following recommended descriptions for scores 2 and 3, it could be described as: *Minimal to no projection of the ANS beyond the INA, with any projection having an obtuse or right angle at the spine tip (Figure 21a) and a base height at least twice the length of the spine (Figure 22 for example on*

how to determine this). The start of the base can be determined by viewing the ANS laterally and identifying where there is a change in angle from the alveolar maxillary surface to the inferior surface of ANS; at this point of angle change, a line parallel to the facial surface can be drawn to denote the beginning of the spine (Figure 22ab, red dotted line).

A score of 2 can be described as: *A moderate projection beyond the INA where the base of the spine is taller or approximately equal in height compared to the length of the spine, but not up to twice the height such as in expression 1. The tip of the spine, the last 1-2mm of the spine, also comes to an angle between 15-89 degrees (Figure 21b).*

A score of 3 can be described as: *A pronounced projection beyond the INA that has a longer spine length than it does base height. The end of the spine also comes to an acute angle approximately 15 degrees or less (Figure 21c). It appears to be more like a needle whereas a score of 2 appears more like a mountain.*

Before accepting this revision, it should be tested to see if these descriptions are adequate for repetition. Furthermore, specific angles that have been determined with additional research should be stated for each expression. Total metric measurement of length may not be helpful because larger individuals will have larger spines, even if they appear a score of 2 in shape. Instead, ratios of base height to spine length should be used so there is a consistent reference for determining what is described as “moderate” or “pronounced” projection past INA.

5.1.2b Trait INA

Patterns of disagreement

Across all of observer 1’s data, there did not appear to be a pattern of disagreement for particular scores. The lack of a pattern could indicate that disagreement is purely because observer 1 could not differentiate between scores closer in value due to their subtle differences.

The low intra-observer agreement between scoring periods 1 and 2 was because observer 1 was confused with where an incline or decline started in relation to other bones and what subnasal grooves were. This confusion made scoring all expressions difficult since each expression relied on the differentiation of these characteristics. However, the observer agreement was substantial for observer 1 in scoring periods 2 and 3 and moderate for observer 1 and observer 2 (Table 10,Table 13) because observer 1 conformed to observer 2's standards of scoring. Even though she conformed to observer 2's standards, the start location of slopes for expressions 1 and 2 were still confusing. This might be why there were cases where disagreement was more than one score apart; observer 1 still thought this trait was difficult to see and feel. These difficulties were not alleviated with the help of the atlas photographs (Hefner & Linde, 2018) because the photographs were not angled the same way as the line drawings. The atlas photographs also did not have arrows pointing to where the defining characteristics of these trait expressions were located, such as where the slope began. Finally, observer 1 disagreeing with observer 2 specifically on what a score of 4 was is likely due to the subjective wording of a "weak" ridge as opposed to a "pronounced" ridge.

Error could also occur from the lack of instructions for what to do with individuals who do not look like the scores depicted, such as those who looked like they had a horizontal ridge (Figure 17a). A horizontal ridge still technically produces an intervening projection of bone between the INA floor and facial surface. This expression excludes it from the score of 3, which states there is no slope or intervening project, and from a score of 4 where a weak ridge is perpendicular to the floor. Without a category to put this expression in, the observer cannot score the trait.

Having disagreements in score that were further apart with observer 3 might be due to observer 3's inexperience with the range of expression, as well as from his idea to take into account a characteristic not mentioned by Hefner (2009; Hefner and Linde, 2018) when looking at this trait from the side. This characteristic was looking at the degree of curvature of the anterior maxillary surface to help observer 3 determine the size of the ridge for expressions 4 and 5 (ex. weak vs. prominent). Additional characteristics such as this one lead back to the proposition of having a reference point for determining how large or tall something is, just like for ANS. It seems to be a natural reaction for an observer to compare a characteristic to another physical feature when the description word is too vague. Unfortunately, comparison of these results to other studies was not possible since they did not talk about INA or which scores were disagreed upon the most.

Recommendations for description/pictorial improvement

There are several expressions that require pictural or description improvement. The following recommendations for expressions 1 and 2 are based on the Hefner and Linde (2018) description and line drawings because there were prevalence issues that hindered the observer from seeing these expressions presented on the study individuals. This means there cannot be further discussion on how the descriptions could be improved for these particular expressions, if it is needed. To improve the clarity of INA line drawings for scores 1 and 2, Figure 23a and 23b are recommended, if it is in fact the correct interpretation. These figures show observers where to look for INA in relation to other features of the nasal cavity, such as the ANS and vomer in the midline where INA is on either side of the midline. The author also suggests the exploration of a smaller contour gage or another piece of equipment/material that can conform to the INA

surface. This would result in an observation outside of the nasal cavity for comparison to the 2-D line drawing.

For expressions 4 and 5, there needs to be clarification of what a weak ridge is compared to a pronounced ridge. It could be defined as follows: *A pronounced ridge is when there is a ridge of bone with two clear anterior and posterior facing surfaces that are vertical and run parallel to each other, and that is taller than it is wide/thick. A weak ridge would have a wall of bone that is wider/thicker than it is tall.*

There is also no clear description of subnasal grooves in any publication relating to Hefner's method, thus a description of these grooves is required if it is noted in the atlas that this could be a point of confusion for identifying a ridge. The subnasal groove description should also be accompanied with a clear depiction in the Hefner and Linde (2018) photographs. For example, superimposing dotted lines over a photograph that outline where the subnasal grooves are located. This might clear up confusion about the small "dips" that were present on the individuals scored in this research.

Finally, the photographs in the atlas (Hefner and Linde, 2018) could be accompanied by line drawings showing cross sections of the individuals' traits, as well as labels on the cross section where an observer can see the features in the photo. Alternatively, a 3-D model or 3-D "line drawing" should be used to guide the observer because it is a 3-D object that is being scored.

5.1.2c Trait IOB

Patterns of disagreement

The change from complete disagreement between scoring period 1 and 2 to precise agreement in scoring period 2 and 3 is likely due to the introduction of ratios in Hefner and

Linde (2018) that were incorporated into observer 1's scoring for scoring periods 2 and 3. The trait expression description was too vague in Hefner's (2009) article and did not define landmarks used for visual reference. Ratios provided a more objective means to scoring the expression, and the use of calipers, though not stated as a requirement, at specific points improved technique consistency between scoring period 2 and scoring period 3, thus reducing error.

The fact that none of the observed expressions in scoring period 2 and 3 were scored a 3 means 'precise agreement' only applies to expressions 1 and 2. It is unknown how the agreement value might change if all three expressions were present. This is another reason why prevalence needs to be addressed in other observer error studies. Not only does it help researchers understand that agreement might not include certain expressions, it might also show that some expressions are more agreed upon than others across studies.

Meanwhile, a score of 3 appeared many times in observer 2's data and multiple times in scoring period 1 by observer 1. Surprisingly, these ended up as three instances of disagreement between the scores 1 and 3 across both scoring period 1 compared to scoring period 2 and observer 1 compared to observer 2, which should be distinctly different expressions. This is likely due to observer 2 and observer 1 during scoring period 1 using Hefner's (2009) original article descriptions without the updated ratios and landmarks. Visually, there are distracting features that may cause an observer to exaggerate what the trait looks like, such as the lacrimal bulging identified in Figure 24. Furthermore, individuals with wider faces would automatically have wider features. If an observer has been scoring individuals who have similar sized faces, they are familiar with what the width of the trait is on those faces. Seeing a wider-faced individual with a wider trait may subconsciously cause the observer to score the trait as wider

than it actually is, especially if the face is wider in a location other than where *fmt* is located (Figure 25). It is known that the brain can perceive things differently based on surrounding characteristics or perceive things that are not there, called "filling-in" (Hopkins & McQueen, 2022). "Filling-in does not require a deficit in visual input, like at the blind spot. Careful arrangements of shapes and colors can induce different kinds perceptual filling-in, such as that of contour and figure, luminance, and color (Hopkins & McQueen, 2022, p. 3)." These visual exaggerations or "filling-in" could also account for observer 2's scores being higher in all cases than observer 1s.

Observer 2 said she looked at the angle of the frontal process of the maxillary bones to determine if the space between *dacryon* and *dacryon* was wider or narrower. If the process is angled to face more anteriorly, the landmark *dacryon* is pushed further apart than if the frontal process was angled more laterally. She then took this observation and compared the width of the IOB to the width of the face for scoring. Observer 1 does not think that the angle of the frontal process ultimately matters in the scoring process, rather it is the exact distance between the landmarks that is important. However, this disagreement in the scoring process is just due to the iteration of the method used because observer 1 was given exact ratios to look for (Figure 19). This difference in observation may or may not affect scoring.

Finally, the atlas photographs (Hefner and Linde, 2018) do not have an indication of where the landmarks are located on each individual; it is unclear looking at the photograph where the *dacryon* landmarks are, and where the widest part of the face is located since you cannot determine where an angle changes from facial to lateral. The photographs of individuals represented in each expression do not have obvious differences which make it clear to an observer why each individual is considered the representation of that particular score.

Recommendations for description/pictorial improvement

This trait description has already been improved in Hefner and Linde (2018) with ratios through the recommendation of Kamnikar et al. (2018), which is what the author would have recommended. It is now recommended by the author that landmarks be clearly defined on the photographs in the atlas (Hefner and Linde, 2018) so the ratios of IOB to facial width are clear. Since the latest publication is relatively recent, the implementation of ratios has not been widely studied and the results from this study cannot be compared to others.

5.1.2d Trait MT

Patterns of disagreement

Among observer 1's data, disagreements occurred most often between the scores of 1 and 2. This was likely because an expression of 3 did not appear for it to be confused with an expression of 2. If a score of 3 had appeared, the same error that resulted in disagreement between 1 and 2, which is the measurement can land directly on the boundary that differentiated two scores, would have occurred between 2 and 3 because they are on the same numerical continuum. In previous iterations of the descriptions, the measurements '2mm' and '4mm' are used as the minimum of one expression and the maximum of the previous expression. A slight deviation in ruler placement would have changed the score, and measurement landing exactly on 2mm or 4mm would have forced the observer to arbitrarily put the individual into a category. The boundary measurement being a characteristic of two expressions may be why the kappa score did not change between untrained and trained data collection sessions. Upon further study, after data collection, a slight rewording in the atlas for expression 3 allows an observer to differentiate a score of 2 and 3, where if it lands exactly on 4mm, it would be scored as 2, whereas "more than 4mm" would place it in a score of 3. However, this wording is not corrected

for the boundary in the descriptions for scores 1 and 2, and 2mm remains a characteristic of both expressions.

The only pattern that was of interest among observers was that observer 2's scores were all higher than observer 1's and observer 1's scores were all higher than observer 3's. This pattern might be due to systematic technical differences, such as the placement of the ruler. Observer 3 scored the trait with the ruler below the curve and tubercle (Figure 29), whereas observer 1 scored with the ruler above the curve and in front of the tubercle (Figure 26a) which is usually higher than where the most superior part of the maxillary curve is. The systematic disagreement with observer 2 might have come down to description interpretation and expression. The biggest source of confusion was what the 'deepest incurvature' and 'deepest anterior curvature' meant. The words "deepest incurvature" resulted in three possible places the ruler could hit (Figure 27), however, observer 2 and observer 1 agreed that it was the superior part of the curve based on the line drawing (Figure 26a). The most likely cause of error is that the specific expression found in the BU individuals resulted in using the modified version of scoring, thus using the wording "deepest anterior curvature". When the tubercle was not directly on the ZS, placing the ruler was straightforward. However, the tubercle more often appeared directly on the ZS, thus resorting to the second scoring technique described. The "deepest anterior curvature of the zygomatic" was confusing because it could be high above the tubercle and above the superior arching curve of the maxilla, thus increasing the measurements of the MT. Likewise, if the most anterior point was on the tubercle itself, but it ended up being below the superior arching curve of the maxilla, measurements would be systematically lower. Finally, the results of this research were not able to be compared to other studies because there did not seem to be any other studies with information about scoring this trait.

Recommendations for description/pictorial improvement

The description for MT has not changed since initial publication (Hefner, 2009) and no one has suggested improvement for this particular trait. This study provides the first recommendations for improvement in trait description to hopefully improve consistency among observers. Since the word "deepest" caused the most discussion on what it meant, it is recommended that it be replaced by directional words: *When the skull is viewed anteriorly, the ruler should extend from, visually, the* ***most superior*** *arching part of the maxillary curve to the* ***most anterior*** *part of the zygomatic, with the most anterior part of the zygomatic being determined from lateral assessment and placement in Frankfurt horizontal plane* (or tubercle, whichever is the correct word to be determined by Hefner).

Since there are two ways to score this trait based on where the tubercle lies in relation to the ZS, it is suggested that this trait should be measured using observer 3's technique (Figure 29). The ruler's short edge would lie against the lateral surface of the maxilla and the long edge would be pointing superior, running perpendicular to the midline, and resting against the most inferior point of the MT. Measurements would be taken from the superior edge of the ruler to the superior most part of the maxillary curve that faces inferiorly (Figure 29, red dotted line). This would result in taking the measurement the same way for both expressions, reducing confusion on where the ruler should be placed. This suggestion should be tested to see if it provides more consistent scoring. If it does prove to be a more reliable technique, then all scores of MT in the MaM Databank should be retaken, if possible.

5.1.2e Trait NAW

Patterns of disagreement

A score of 3 showing up in observer 2's data but not observer 1's is likely due to different scoring techniques. Observer 2 did a visual analysis without using the ratios as described by Hefner and Linde (2018) that observer 1 used. Much like IOB, facial features that are distracting to the eye may be causing over or underestimation. A difference in where the widest part of the face is located could also make the NAW look narrower or wider (Figure 25). Even if observer 1 and observer 2 measured facial width from the same area of the face (ex. zygomatics body vs. frontal process), the point at which the facial surface changes to the lateral surface (ie. the point used for measuring width) can be disagreed upon, thus can account for some error that contributes to score disagreement.

Even after training where observer 2 discussed that she used the change in surface on the zygomatics to measure width, observer 3 decided to use *frontomalare posterale* as his reference for facial width during data collection. This means explicit mention of where the width is measured is necessary to include in the description because notes from training are not necessarily remembered.

The error for this trait was also discussed in Kamnikar et al. (2018) which is where the recommendations for ratios was made. However, there were no other observations discussed, and since the ratios have been implemented, there has not been a publication testing this implementation. Andrade et al (2018) briefly discussed the scoring of this trait using 3-D analysis and comparing it to scores taken by an observer. The 3-D analysis was able to agree with the observer 71% of the time for scores of 1, 100% of the time for scores of 2, and 95% of the time for scores of 3. However, the prevalence of expressions scored of 2 were half that of the

other scores. This pattern is important to note since prevalence can be an issue for agreement, thus similar to the results of this research where agreement can be higher for particular scores only due to fewer expressions being present in the study individuals. This study did not talk about why disagreement may be occurring for the scores of 1 and 3 though, so no comparison with the current study can be made.

Recommendations for description/pictorial improvement

Since observer 1 saw that visual analysis could cause an observer to over or underestimate a ratio, further study should determine if visually assessed trait scores match mathematically derived scores (ex. calipers to measure facial width and NAW for ratios). There also needs to be a description for which landmarks the facial width should be measured from. If it is measuring any part of the face wherever it is widest, then, for example, it should be described as: *The width of NAW is assessed in relation to the widest part of the facial skeleton, which can occur anywhere between the landmarks <u>frontomalare temporale</u> and the inferior tip of the MT.*

This trait requires incorporating ratios and is leaning towards metric analysis since the technique for measuring is metric. Thus, it should be stated that scoring NAW and facial width requires calipers, along with which landmarks are used and an explicit statement that calculations are necessary. If math is used to determine the ratios, then there needs to be a clear cut off of when something is more towards a ratio of 1:4 than a 1:5 or 1:3 ratio. When mathematically deriving the ratios, it is recommended that a value ending with 0.5 and above is rounded up and when it is below 0.5 it is rounded down. Measuring the distances between landmarks will provide a more precise ratio as well as a more objective scoring procedure than visual analysis.

5.1.2f Trait NBC

Patterns of disagreement

NBC is a perfect example of high intra-observer consistency due to internal bias. The intra-observer kappa values remained high during observer 1's untrained and trained sessions (Table 10). It was first thought that there were prevalence issues with NBC in observer 1's data across scoring periods, and bias in observer 2 versus observer 1's data since observer 2 scored differently. However, observer 1 scored the majority of individuals as a score of 1 for scoring periods 2 and 3 whereas observer 2 had a range of scores. Observer 1 in scoring period 3 did have a range of scores in the U of M individuals but did not agree with observer 3 on most of the scores, indicating this might be an issue with observer 1's scoring. Intra-observer agreement was only high because of internal bias. These results are consistent with Hefner (2009) who also found low agreement between observers, likely due to the learning curve associated with the contour gauge. Kamnikar et al. (2018) also recognized that discrepancies likely arose between observation periods due to the implementation of said tool when it was previously not used.

It is clear that the expressions are not distinct enough for observer 1 to detect with a contour gauge, especially an expression of 4 since it appeared for observer 3 and observer 2 but not observer 1. Most disagreements that exist with observer 2's score of 4, that were scored as a 1 by observer 1, are likely due to the fact that the point of the triangle was too subtle for the contour gauge to pick up. Surprisingly, the individual that was scored 4 by all three observers during the training session should have had a score of 4 for the data collection, but it did not occur for either observer 1 or 2. The contour gauge placed at different points by observers can account for these differences, as exemplified by Figure 36 and Figure 37; a millimetre difference

in where the contour gauge or fingers were placed for measurement could mean the difference in shape, thus between scores.

The atlas descriptions and line drawings also contribute to confusion too since there were no visuals of what the contour gage should look like after measurement, or a description of what steep versus shallow meant. Additionally, inconsistent language in the description, such as "steep" and "high," were used. The discussion among observers also revealed that "high and rounded" and "low and rounded" are unclear, and data collection showed a lot of shape variation (Figure 38) for individuals that were scored the same value. The unclear descriptions in conjunction with this variation is likely where error arises. It is confusing to determine what is 'steep' versus 'shallow' on a single individual, but when looking at multiple individuals altogether (Figure 38), differences are more easily seen between the contours.

It is also clear that "approximately 1cm below nasion" can be interpreted differently since observer 3 took it literally, directly below (Figure 30, blue dashed line). In Figure 37, observer 3's contour had a more equally rising slope of the walls, much like that of a triangle, whereas observer 1's looks more like a curved slope. This slight difference in shape may have been due to the difference in where observers are placing the contour gauge. A couple millimetres in either direction along the boney surface may have a different contour (Figure 37) that affects how the expression is perceived by each observer.

Kamnikar et al. (2018) showed that the contour gauge improved the objectivity of the scores, but there is no discussion on where error could be occurring when scoring this trait with the gauge. Therefore, the results of this study could not be compared to the results of other studies since comparing the scores that were derived from use of a contour gauge were not compared to scores derived without a contour gauge within observers.

Recommendations for description/pictorial improvement

The resulting angles on the contour gage might be the best tool for determining whether the walls are steep or shallow. An observer can place the contour gauge over a protractor to find the angle (Figure 40), and 'shallow' or 'steep' can be based on where the angle falls in a predetermined range for each expression. Alternatively, a 'low' contour can be described as having a base width that is greater than or equal to the height of the contour, whereas 'high' can have a height greater than that of the base width. Further research is required to determine what angles and ratios differentiate expressions 1, 2, and 3.

When determining a plateau, the current contour gauge description for a score of 2 is useful since it describes how many needles are needed to make a plateau. However, the resolution of the contour gauge needles should be included (how many needles per centimeter). To bring a score of 3 to a consistent description it should be amended as such: *A narrow surface plateau is described as having a flat cluster of 2-3 contour gauge pins (less than 7mm).*

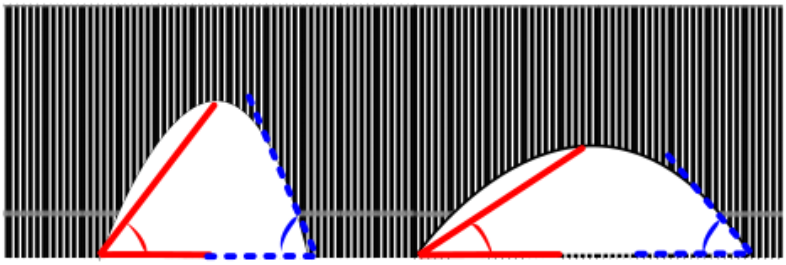

Figure 40: Representation of a contour gauge showing what "steep" (left) and "shallow" (right) walls would look like on a nasal contour. The solid red and dashed blue angles are examples of how the contour gauge shape can be measured to determine whether the walls are steep or shallow, and, therefore, the score given to each shape.

The atlas photos of NBC are not helpful in their current form because an observer cannot see the contour on a flat photograph. It is recommended that the exact location where the contour was assessed on the nasals should be indicated on the photographs. It is also recommended that the photos be accompanied by the associated contour gauge result. This way practitioners can see what each score looks like with and without a contour gauge, particularly since a contour gauge

is required for accurate assessment. The recommendation for using a contour gauge to assess NBC should be changed to ‘mandatory’ to ensure consistent application to decrease error and because it has been found to improve objectivity (Kamnikar et al., 2018).

5.1.2g Trait NO

Patterns of disagreement

A low NO agreement even after training may be due to the description conflicting with photographs in the atlas (Hefner and Linde, 2018), which confused observer 1 as to how to score this trait. Particularly, “*extending beyond the maxilla at the bony landmark ‘nasale inferious’* and images that the author thinks is “extension” but are labelled as “no extension” (Hefner and Linde, 2018, p. 151 fig. 11.8). Hefner and Linde (2018, p. 149, fig. 11.5, 11.8) also mention in a caption figure that “separation is not NO” without explaining what this means, therefore, the figures representing expressions for “absent NO” look similar to figures representing “present NO”.

After much review of the atlas figures, observer 1 thinks the description is trying to explain: *Any projection of the nasal bones past the inferior most part of the maxilla that is connected to the nasals. This means, if a line were drawn following the angle of the inferior border of the maxilla, then extending at this same angle into the nasals (Figure 20a,c,d, dashed red line), the nasals would continue past this border. In other words, it would be as if this extra nasal bone would continue its connection with the maxilla, but it does not.* Therefore, what they mean by “separation does not count as NO” is that there was a connection of the nasals and maxilla, but as the bone dried out, it caused them to separate. Therefore, the remnants of the connection determine that it is no overgrowth. If there is no indication that the maxilla and nasals were connected along this extra nasal bone, then this would be considered overgrowth. However,

due to the fragility and small size of the area, it is difficult to assess this without a magnification lens, at least for the author.

The uncertainty was apparent as observer 1 disagreed more times with herself and other observers on whether it could be scored or not rather than between scores. The thin bone was difficult to determine if it was broken or over growth, thus experience comes into play for a different aspect of scoring error. The dual interpretation that observer 1 described in Figure 20 means there needs to be a clearer description of what 'extending beyond the maxillary border' looks like and what plane of reference should be referred to (Figure 20ab). Since there are no other studies discussing the scoring of this trait, the results of this study could not be compared to others.

Recommendations for description/pictorial improvement

It is recommended that the line drawing or description be altered with a dashed line showing what border is being referred to, such as indicated in Figure 20. If Figure 20a is correct, the description should be explained as above. Furthermore, expansion on what separation is and when NO could look like separation. For example, the inferior nasal edge at the nasomaxillary border curving over the inferior free edge of the maxilla but with a space between them (Figure 20d). This is because if it is recommended to feel the edge, these two can be confused with one another. Perhaps a magnifying glass could be used to assess this trait.

5.1.2h Trait PBD

Patterns of disagreement

The disagreements among observers are fully due to observer 1's misinterpretation of what a depression in the skull was. Observer 1 thought that even though the contour was flat beyond *bregma*, the change in surface of the natural convexity (Figure 33, black line) was

technically depressed from the convex surface. This misinterpretation is clear in Figure 37 because observer 1 contour is the same shape as observer 3's, but the individual was scored differently. Knowing the distinction now, if observer 1 were to measure these individuals again, the agreement may be higher.

This trait was discussed briefly in Andrade et al. (2018) where a digital analysis of the curvature along the sagittal suture was completed. They determined that this trait could be identified by placing a straight line between *bregma* and the most externally projecting part of the convex skull. If the contour landed below this straight line, then it was determined to be present, and the digital analysis agreed with the observer 100% of the time. However, it was only identified three times, which meant that the issue of prevalence appeared. They did not discuss any other issues about where error could be occurring so it could not be compared to this study.

Recommendations for description/pictorial improvement

If observer 1 misunderstood what a depression was, it is likely other people may have this misunderstanding too. It may be helpful to state that a depression would be "a concave contour opposing the convex contour of the skull" and use Figure 33 to exemplify this distinction. An alternative to the contour gauge can be the use of a straight edge placed along the sagittal suture at *bregma* to see if there is a dip below the edge. This tool along with the contour gauge should be noted in the description for how to score the trait.

5.1.2i Trait SPS

Patterns of disagreement

All disagreements between 0 and 2 were likely due to the fact that surface texture of the bone at nasion can be mistaken for suture visibility. Some traits can be completely obliterated except for a difference in bone appearance, such as slightly smoother or shinier. Alternatively, it

can be closed but only visible at a certain angle or in certain lighting, thus being missed by observers if not viewed the same way. The difference, as thought by the author, between "closed but visible" and "obliterated" is actually not that much, therefore, there needs to be further differentiation between these two expressions. Different observers also have varying visual acuity that could contribute to detection or not. This trait was shown to have the second poorest intra-observer agreement in the original publication (Hefner, 2009), meaning there is much to be improved with this trait. The results of this study could not be compared to other studies because there did not appear to be any that tested the scoring of this trait for sources of error.

Upon reading further comments made by observer 2 to the paragraph above, the author noticed there is a difference of opinion on what obliteration of the suture means. The author thinks that an obliterated suture is the filling in of the suture with bone because the suture is defined as a fibrous connection. Therefore, obliteration of the fibrous connection is bone replacing the fibrous tissue and becoming flush with the surface of the original bone (Table 57c). With this in mind, 'closed, but visible' means the fibrous tissue is almost fully replaced by bone except for the surface, which is where the 'dip' occurs in Table 57b. Observer 2 thinks that obliteration means the texture is also altered to match that of the surrounding bone, therefore, even if the new bone is flush with the original bone, a slightly shinier look is not obliterated.

Recommendations for description/pictorial improvement

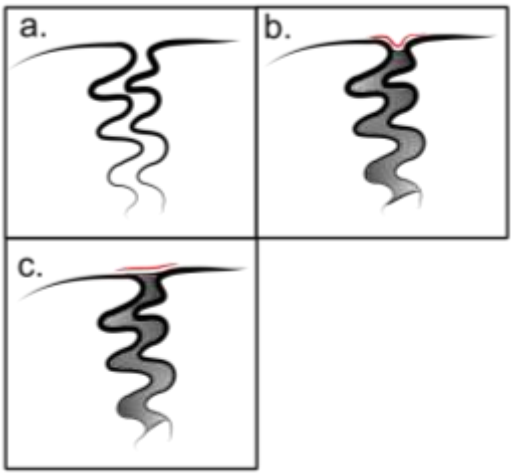

Figure 41: Cross section of the interlocking suture SPS showing an unfused (a) and fused (b,c) appearances, where the dark grey gradient is the bone that fills in or fuses the suture. The red solid line in 'b' indicates that there is space yet to be filled with bone for full fusion. The red solid line in 'c' indicates the bone has filled in the suture all the way for full fusion, but may have a surface appearance difference.

It is clear that observers agreed on what an open suture looks like. However, a clearer description for "closed, but visible" needs to be written, such as: *'Closed' indicates the suture has filled in with bone most of the way, but 'visible' indicates there is a dip in bone where the suture cavity occurs, such that difference in height between suture and surrounding bone casts a shadow (Figure 41b). Completely obliterated should be scored when the bone is completely smoothed over with no dips where the suture was, even if there is a difference in bone appearance such that is slightly shinier or slightly whiter (Figure 41c). It should be viewed anteriorly and no other angle. This way the light does not catch the bone texture in a way that might be perceived as a visible suture.* There also needs to be a discussion on why this trait is not considered ordinal if there is a progression from open, to closed but visible, to closed.

5.1.2j Trait ZS

Patterns of disagreement

It is interesting that most disagreements in score across all observers and scoring periods occurred between a score of 2 or 0 because they should be opposite ends of the spectrum for phenotype. If one trait has no angles, and the other has two angles, how are these getting

confused between observers? It lies in the description where it does not define what an angle is along a curved and undulating continuous line, as well as when undulations are not considered a change in suture direction/angle. First, in Hefner (2009), it was mentioned there was a slight angle in a score of 0, but the description changed to no angles for a score of 0 in Hefner and Linde (2018). This is one cause for error. Second, an angle can be obtuse, such as the one seen in the expression '0' line drawing by Hefner (Figure 11; Figure 42a shows what the angle is for Hefner's line drawing in blue lines), measuring approximately 140 degrees. This obtuse angle likely caused confusion for disagreements on expressions 0 and 1. Furthermore, in many of the study individuals, it looked like they had two angles, but the course did not match the line drawing for expression 2. Instead, these expressions had the suture course starting from the orbit looking like expression 1 (Figure 11) but the angles of expression 0 (Figure 42b shows the combination of expression 0 and 1 from Figure 11). If expression 1 was only supposed to have one angle (Figure 11; Figure 42c shows what the angle is for Hefner's line drawing in blue lines), it was assumed that it was the angle in the middle of the suture course and did not include the part of the suture closest to the orbit or inferior edge. Therefore, these "combination" expressions (Figure 42b) were scored as 0 because it was unclear if an obtuse "angle" was included in the angle Hefner and Linde (2018) were describing (Figure 42a) or scored as 2 because it looked like there were too many angles (Figure 42c). The equal split of disagreed scores between 2 or 0 and 2 or 1 in scoring period 2 and 3 also acts as evidence for the confusion on how a score of 2 is differentiated from the other scores.

Similarly, Hefner's line drawing for expression 1 also looks like it has two or more angles (Figure 11,1; Figure 42c shows what the angles are for Hefner's line drawing in blue lines) and a few expressions on the study individuals looked like they could be scored as 1, but

had too many angles to meet the description requirements (Figure 42b). Which means the same confusion as with expression 0 occurred. Therefore, scores were pushed towards a 0 or 2 because expression 1 was not clearly defined.

Finally, error between observer 1 and 2 also occurs from two different pieces of advice for scoring asymmetry in different publications; Hefner (2009) said to score the left side, and Hefner and Linde (2018) said to score the highest expression. All these reasons could be why there was such low agreement within and among observers across the majority of studies testing the agreement of this trait (Coelho et al., 2017; Dinkele, 2018; Kamnikar et al., 2018; Klales & Kenyhercz, 2015; L'Abbé et al., 2011; Wang, 2016). Other studies only tested the kappa value of agreement and did not discuss sources of error or prevalence, so the results of this study could not be compared to other studies.

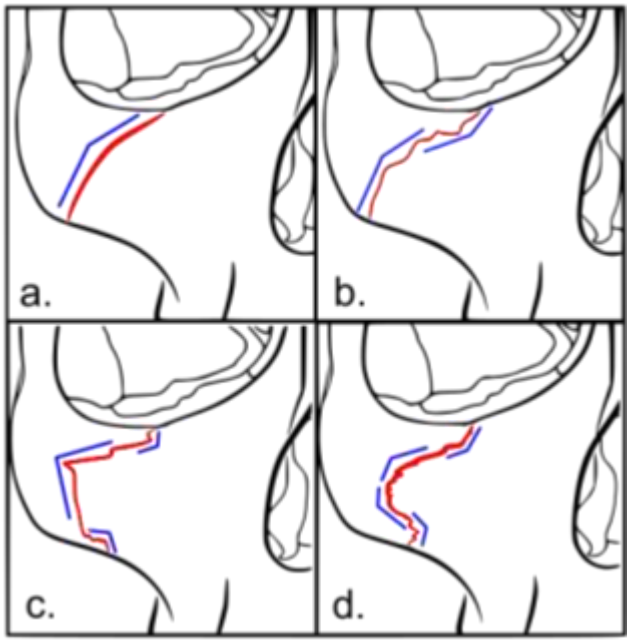

Figure 42: ZS suture course (red line) and the visually identified angles (blue solid lines) that would be used for scoring the expressions. 'a' corresponds to Hefner's line drawing for expression 0 in Figure 11, and 'c' corresponds to expression 1. Anterior view of the skull.

Recommendations for description/pictorial improvement

It is recommended that the atlas (Hefner and Linde, 2018) should include white dashed lines across the photographs to show the course of the suture, as well as arrows indicating where an angle is determined, or the blue lines as indicated by Figure 43. It is also recommended that Hefner clarify what constitutes an angle. For example: *An angle is a change in suture course*

where the suture deviates at least 10 degrees from the previous part of the suture path (Figure 43c, green dashed line showing angle change from the solid blue line). Since the suture course cannot have no angles, expression 0 should include an obtuse or "slight" angle as stated by the original description (Hefner, 2009).

In Hefner's line drawings (Figure 11), it is not clear that expression 1 has one angle (Figure 11; Figure 42c shows what the angle is for Hefner's line drawing in blue lines), therefore, expression 1 should look the same as expression 0 except the inferior edge is more medial rather than lateral, as shown by Figure 43ab. Therefore, there is clear depiction of what one angle is, and that the difference between a score of 0 and 1 is where the greatest lateral projection occurs. Altering expression 1's line drawing brings it in line with the description, and may reduce misunderstanding due to conflicting information.

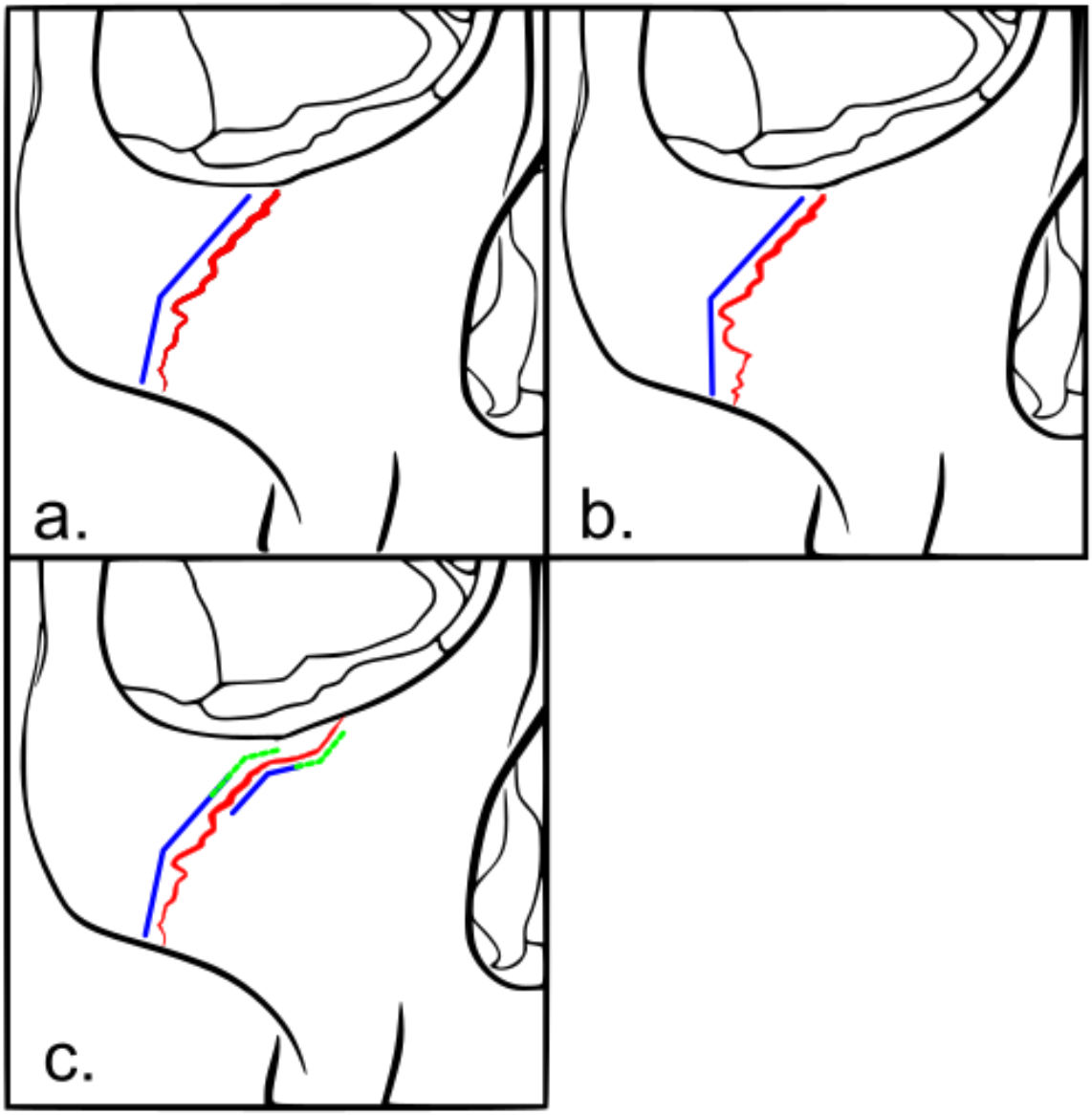

Figure 43: Recommended drawings for ZS that match the expression descriptions (red lines). Figure ‘a’ is expression 0, with the blue line showing one angle but the greatest lateral projection is at the inferior edge. Figure ‘b’ is expression 1, showing one angle with the blue line but the greatest lateral projection is midline. Figure ‘c’ is expression 2 showing two or more angles with the blue solid and green dashed lines. Anterior view of the skull.

5.1.2k Trait TPS

Patterns of disagreement

No distinct patterns of score disagreement among and within observers could mean that the error is due to chance. However, upon further inspection of the individual's expressions, it was found that many individuals had asymmetry in expression on either side of the palatine suture. In the description for how to score this trait, it was noted that in cases of asymmetry an observer should score the trait "generally." The word "generally" left observer 1 feeling conflicted on how to score the trait when both sides had clear expressions that would be scored differently. The majority of disagreements between observer 1 and 2 (x=3, X=5) had individuals with these cases of asymmetry, with the scores that could be attributed to either side appearing for both observers. For example, individual #2 had the left side with a score of "4" and the right side a score of "3", and observer 1 scored the expression as 4 and observer 2 scored it as 3. These cases lend evidence to asymmetry being likely cause of disagreement.

Error can also stem from what constitutes a "slight" or "significant" undulation. A score of 3 was confusing without the atlas photographs because some undulations did not look like they were far enough anterior (ie. "significant" enough undulation) to be considered an anterior deviation, thus only the posterior deviation was identified without the photograph. Moreover, using the atlas changed the majority of the scores of 4 in observer 1's scoring period 3 to a score of 3, however, some of these scores were 1 in scoring period 2 meaning the atlas did not help with differentiation. Both of these results mean there is confusion on what a significant deviation is so that it is detectable apart from slight deviations. In one case of disagreement, individual #17, the expression was atypical with both observers noting the different expression.

Klales and Kenyhercz (2015) mentioned that scoring this trait was difficult only due to the suture being obliterated. In cases of suture obliteration, the suture was also not scored in the current study. However, discussion on where other sources of error for scoring might be occurring did not take place in Klales and Kenyhercz (2015). The results of this study could not be compared to other studies since there did not appear to be studies testing the scoring of this trait.

Recommendations for description/pictorial improvement

It is recommended that the scoring method involves using the left side rather than scoring "generally." If "generally" must be scored, Hefner must provide a definition of what it means with examples of how to score "generally" in the case of asymmetry with two clear scores. In addition, the words "significant" and "slight" can be improved with definitions, or a technique to distinguish significant undulations from slight undulations, such as observer 3 using a line of best fit (Figure 35). A physical line of best fit (LoBF) might be the best course technique for assessment because this provides a consistent plane of reference across observers. If undulations can be considered a series of imperfect parabolas that alternate with bases opening anteriorly and posteriorly, then these parabolas can be used to define undulations in relation to the LoBF. A physical LoBF will help discriminate between undulations that extend further away from the LoBF than other undulations (ex. parabola that is twice the height) and stay away from the LoBF longer than other undulations (ex. parabola that is twice the width), thus improving identification of a "significant deviation." For example: *Using a clear ruler, one edge of the ruler should be placed at the lateral edges of the TPS and perpendicular to the interpalatine suture. Once in place, the ruler is moved anterior and posterior, maintaining perpendicular position, until it*

looks like most undulations in the lateral halves of each bone hit this edge (Figure 35). This is the baseline for measuring deviations that distinguish each expression.

A definition for a significant deviation can be something such as: *A significant deviation occurs when the suture line extends away from the line of best fit at least twice the height and width of the majority of other undulations.* These alterations need to be tested before implementation in the method.

5.1.2l Trait NAS

Patterns of disagreement

It is not surprising that a score of 1 was confused with scores of 2 or 3 across all datasets since it looks like the intermediate between them. It is unknown why the scale is labelled this way because it looks like it could be an ordinal trait with the intermediate expression. Having scores of 2 and 3 never confused with each other, within and among observers (Appendix 2) is a reflection that they are on opposite ends of the visual scale. Therefore, there is less difficulty in differentiating them. These two scores being confused with a score of 1 might also be due to the description of expression 1 including expression 3. It states that the greatest lateral margin is in "lower 2/3" which includes the bottom half, or midline, which is the defining characteristic for expression 3. Even though the description of 1 says "below where a score of 3 would hit", it is confusing why the description is "lower 2/3" instead of "between midline and inferior." Only observer 3 scored this trait for comparison with observer 1, and the confusion between a score of 1 and 2 might further indicate that a score of 1 is harder to differentiate. On the other hand, a score of 3 did not show up as often as a score of 1, indicating prevalence is affecting further analysis of whether an expression of 3 is easily identifiable or not.

This trait was not included in Hefner's (2009) original publication and has only been tested for intra-observer agreement a handful of times (Atkinson & Tallman, 2020; Kilroy et al., 2020; Moffit, 2017; Wang, 2016) and inter-observer agreement once (Klales & Kenyhercz, 2015). While the kappa values cannot be compared across studies, in the case of this trait, the kappa values have similar values across studies, meaning this trait's expressions may be relatively easier to distinguish than others. Until prevalence is discussed in these studies, it cannot be confirmed.

Recommendations for description/pictorial improvement

The first recommendation is to remove "lower 2/3" from score 1's description and replace it with "below midline and above the inferior aspect." The description should also include how to measure the widest part if there is asymmetry; it would be assumed to measure the left side, but repeating this for each trait is necessary so there is no confusion.

The second recommendation is to include how to measure the widest part. For example, calipers to find the widest part of the aperture by taking multiple measurements. Alternatively, the author found it useful to use a toothpick and hold it perpendicular to the midline, moving it up and down the nasal aperture to find the widest point. This allowed the author to visually section the nasal aperture into superior and inferior portions, making it easier to visualize if the aperture was split in half (score of 3), less than half (score of 1), or was at the inferior aspect (score of 2). Alternatively, ratios can be used, such as measuring the total height compared to the height at which the widest portion hits- with the stationary edge of the calipers remaining at the INA. This should not be measured if the nasals are broken and total height is unknown.

Finally, a discussion around why this is not considered an ordinal trait needs to take place. This could be considered ordinal because the widest part goes from the bottom (score=2),

to bottom half (score=1), and to middle (score=3). It is assumed this is considered a nominal trait based on population frequencies, such that trait expressions do not show up in a progression of 2,1,3 or 3,1,2. If there is no reason why these are considered nominal, it is suggested that this trait be considered ordinal and scored in a progression, therefore, when tested for inter-observer agreement, it should be treated with a weighted kappa.

5.1.2m Trait NBS

Patterns of disagreement

This trait was not scored by observer 2, and since observer 3 looked at this score laterally rather than anteriorly this would account for almost all scores disagreeing, and no pattern of disagreement can be detected. Furthermore, a score of 4 was not observed which impacts the ability to determine how an expression of 4 may result in agreement or disagreement among observers.

While the scores themselves are not helpful in teasing apart where error could be occurring, the discussions among observers were helpful. The words “pinching” and “bulging” are likely resulting in error because descriptions do not go into detail to define what pinching and bulging encompasses. The line drawings indicate it may be related to the suture outlining the nasal bones rather than the entire surface plus the suture of the nasals. Observer 1 thought that observer 2 was also describing the entire surface of the bones so she ended up scoring bulging as if the inferior edge was bulging out at a greater degree than the contour of the rest of the bone, including the suture pathway.

Finally, the lack of explicit mention for how to score the trait in the case of asymmetry can introduce difference in technique. For example, an anthropologist can decide to score the left side, the highest expressions, such as done for ZS, or “generally,” such as done for TPS. Since

this trait was not included in the original publication, it has only been tested for inter-observer agreement once (Klales & Kenyhercz, 2015). However, there is no other discussion on where error is occurring when scoring this trait so the results of this study cannot be compared to other studies.

Recommendations for description/pictorial improvement

If pinching and bulging encompass the whole surface and not just the sutures, it is recommended that a 3D line drawing and description are drawn for how the bone surface should look for each feature that defines the expressions. If the surface of the bone is included in the bulging, it can be described as: *Bulging is a rounded convex contour of the bone that is at a convexity of greater degree than that of the surrounding bone. This bulging will occur in the inferior half of the nasals between and including the nasomaxillary sutures and internasal suture.*

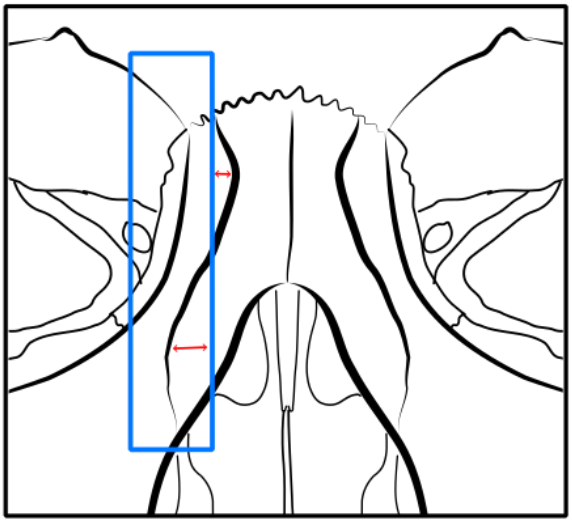

Figure 44: Depiction of how a straight edge can be placed (blue rectangle) to measure whether there is pinching (top arrow) and bulging (bottom arrow) for NBS. Anterior view.

If the nasomaxillary sutures are where observers are supposed to look for pinching and bulging, observer 3 provided a helpful tool for determining this: a straight edge. This will provide a more objective analysis of the curve than the human eye alone. A transparent straight edge can be held parallel to the midline at where the nasomaxillary suture hits the nasofrontal suture down to the inferior edge of the nasal (Figure 44). This will help delineate if there is a medial or lateral curve past the straight edge. The description accompanying the figure can be:

The pinch indicates the shortest distance between the two nasomaxillary sutures and occurs somewhere between the most superior and most inferior points of the nasomaxillary suture, but not at these points. If there is a pinch, there will be a concave curve on the medial side of the ruler, when the ruler is placed at the point where the nasomaxillary suture meets the nasofrontal suture and is held parallel to midline. If there is a bulge, then the suture will flare laterally past the ruler edge. Together these create an hourglass shape to the nasal bones.

Furthermore, the "pronounced lateral bulging" should be defined. It is recommended that research is needed to find the ratios of the lateral margin bulging for expression 2 and 3. For example, in Figure 45a, the bulging (Y) extends laterally past the ruler less than the amount of

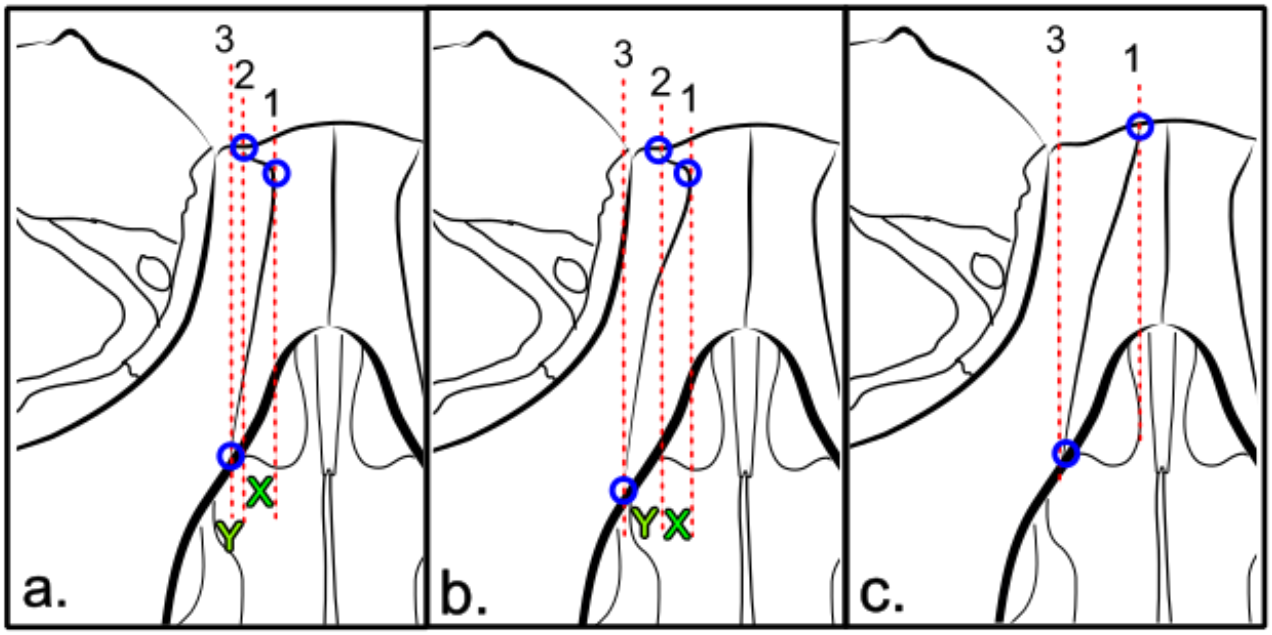

Figure 45: Depiction of some NBS trait expressions and how parallel lines measuring points along the nasomaxillary suture can help determine a pinch and degree of bulging. Line 1 is an extension of the point at which the most medial part of the suture hits (centre of blue circle) parallel to midline, in other words, the shortest distance between the left and right nasomaxillary sutures. Line 2 is an extension of the point where the nasomaxillary suture meets the nasofrontal suture (centre of blue circle) parallel to midline. Line 3 is an extension of the point where the nasomaxillary suture meets the inferior edge of the nasal and maxilla (centre of blue circle) parallel to midline. X is the distance between line 1 and 2, and Y is the distance between line 2 and 3. X can be compared to Y to determine if it is wider (a) or narrower (b), thus indicating whether there is "slight" bulging, as indicated by X>Y (a), or "pronounced" bulging, as indicated by Y>X (b). Image 'c' depicts a trait expression where 1=2, therefore, there is no pinch.

bone that is pinched (X), thus y<x, and has minimal bulging. In Figure 45b, the bulge extends laterally past the ruler greater than the pinch, thus y≥x, and pronounced bulging.

Similarly, the descriptor that there is no "hour glass" flaring can be added to expression 4. Instead: *The shortest distance between the nasomaxillary sutures is at the most superior point of the suture where it meets with the nasofrontal suture. When placing the straight edge at the most superior point, the suture flares laterally past the ruler's edge at this most superior point with no medial flare (Figure 45c).* Finally, there needs to be a description of how to score the trait when asymmetry is present, or if one nasal is missing/broken.

5.1.2n Trait NFS

Patterns of disagreement

Since observer 2 did not score this trait, observer 1 only has observer 3 to which to compare scores. Two of the four disagreements between observer 1 and 3 were more than one score apart and two of the disagreements were between a score of 4 (i.e., irregular) and another score. The uncertainty of when the undulations should be considered irregular is likely why there were disagreements in score. This is also likely the reason why there was no pattern to disagreement with observer 3. Sutures are naturally wavy and do not look straight or smooth like what the line drawings look like; the drawings are an approximation of shape, much like a line of best fit. Some study individuals had many "irregular" looking undulations in the suture, but using a "line of best fit" approach could make it look round. If the line drawings showed an example of a natural suture and how the "round" vs "triangular" trait comes out of these natural undulations, it might reduce interpretation error. It may be the case that the majority of study individuals were characteristically irregular (i.e., a score of 4) making it difficult to place an individual into a category. In fact, Klales and Kenyhercz (2015) also found that there were many "atypical"

expressions and they had to force these expressions into a category, resulting in high disagreement among observers. The current study results having similarities with the results from Klales and Kenyhercz (2015) indicates that there is a need to refine the definition for a score of 4, or that the NFS needs more categories and/or clearer instructions for how to determine general shape from the natural undulations.

When using the atlas photographs, it was easier to place an expression into a category because study individuals could be compared to reference individuals to determine who they were most alike. Therefore, line drawings should reflect reality rather than generalized lines. Of course, technique differences, such as observer 3 using a straight edge to score NFS could also be where error occurs. Since this trait was not included in the original publication (Hefner 2009), it has only been tested once for interobserver error resulting in the lower end of 'fair' agreement (Klales & Kenyhercz, 2015).

Recommendations for description/pictorial improvement

Since there has been no recommendations for improvement, and it has not been tested by many researchers, it is hoped the recommendation here can be implemented immediately. The use of a clear ruler by observer 3 is a useful way to score this trait because it allows comparison of a suture in relation to a tool. Therefore, it is recommended that a clear ruler is used for scoring to introduce consistency between practitioners. Placing this straight edge along the superior most portion of the suture will allow the observer to observe how the suture undulations work towards or away from the ruler edge, and if there is a general curve as opposed to a straight line. Sometimes the human eye cannot distinguish a straight line from a slight curve, so introducing a straight line for comparison can help distinguish these shapes (Figure 32). This straight edge can also determine how much of the suture touches the edge before moving away.

Some of the photographs in the atlas (Hefner and Linde, 2018) were too dark to see suture detail, or it was unclear how the atlas authors saw the exemplified expression on the referenced individual. The atlas then should include dashed lines over top of the photographs, slightly above the suture, to depict the trait clearly. This may make it clearer to an observer why this individual was chosen as the reference for the specific expression. As a final recommendation, there should be instructions on what to do with individuals who have the frontal side of the suture looking different than the nasal side of the suture, such as was found on one individual in this study (Figure 28).

5.1.2o Trait OBS

Patterns of disagreement

The high score agreement (Table 10; k=0.84) is likely due to the fact that observer 1 used a tool to assess this trait because she felt it was too difficult to eyeball the horizontal and vertical border lengths. Using a skewer to roughly measure the horizontal and vertical planes provided a consistent, less subjective technique to score the trait. The guide should have been followed exactly as it was written as the introduction of the tool is not testing the guide's original method. However, because observer 1 resorted to tool use, it allowed for the recognition that the descriptions were not precise enough to assess this trait as they were.

Having only disagreements between 1 and 2 but complete agreement for a score of 3 among observer 1's observations could be due to prevalence issues because a score of 3 only showed up three times. However, it would make sense that 1 and 2 are confused since eyeballing horizontal and vertical measurements is difficult without start and end points to the borders. If there are no points at which to measure this, then it could mean one observer measuring a border further inward or outward than another.

It is likely the disagreements that observer 1 had with observer 3 were caused by observer 3 using the ruler to determine if the superior border was angled for a score of 3 rather than the medial border being shorter than the lateral border. The line drawing also shows that the inferior border is more angled than the superior border in an expression of 3. However, this angle is not discussed in the description, only the distances for the lateral and medial borders, thus this should not have been used as a characteristic for assessment.

Other than observer error studies, whose kappa values cannot be compared, there are no other studies testing the scoring of this trait for where error could be occurring. Therefore, the results of this study are unable to be compared to other studies.

Recommendations for description/pictorial improvement

It is recommended that there is a description of where and how to measure the horizontal and vertical orbits so that there is consistency and reduced subjectivity in determining if they are equal to, greater than, or less than one other. For the horizontal width of the orbit, *ectoconchion* can be used as the landmark. To measure vertical measurements, for example, straight edges can be placed along the orbital borders so that they have the longest straight part of the border following the

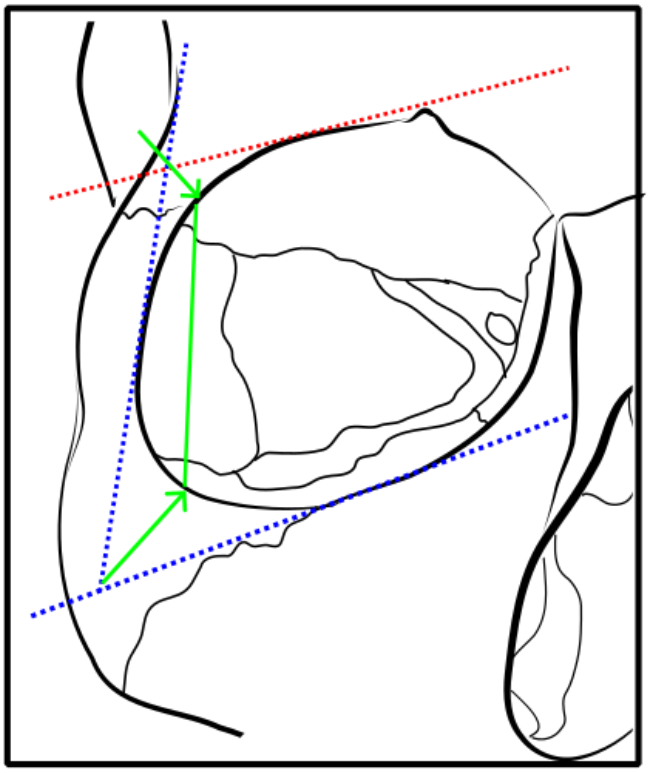

Figure 46: Example of how straight edges can be placed along the orbit margin (dotted lines) to find points (green arrows) where the vertical orbit heights can be measured (green line).

straight edge (Figure 46, dashed lines). Where the straight edges intersect, a line is drawn in the middle of the two to where it would land on the orbit rim (Figure 46, green arrows). Where it lands on the rim is where the measurements would be taken with calipers (Figure 46, solid green line). While there are issues with the proposed change, it may reduce measurement error than if there were no method to measure. This can be the starting point of conversation on how best to measure the vertical borders.

5.1.2p Trait PZT

Patterns of disagreement

Disagreements between a score of 1 and 2 likely occurred from the measurements themselves since four millimetres marks their boundary. Much like MT, a slight placement of the ruler in a different spot can change the measurement, such as one observer measuring 1.9mm and the other 2mm, thus changing the score. Placement of the ruler along the contour of the zygomatic to find *jugale* may be a source of this error because observer 3 placed it on the border where the curve begins to change from the temporal process to the zygomatic process as opposed to observer 1 placing it midway between the changes in contour on the temporal process and zygomatic process (Figure 34). The reason for one individual disagreeing by two scores between observer 1's scoring periods 2 and 3 is odd and the reason is unclear.

There are no other studies that appear to test the scoring of this trait for where error could be occurring, therefore, the results of this study cannot be compared to other studies. Since this trait was only included after the original publication, it has not had time to be tested yet, and it is hoped the results of this study can be compared to the results of any future studies.

Recommendations for description/pictorial improvement

It is recommended that observers review the landmarks before data collection begins to ensure that the landmarks are identified properly. As of now, it is understood by observer 1 and 2 that *jugale* is found in the middle of the arch between the frontal and temporal processes.

5.1.3 Summary of trait scoring

Each of these traits and the techniques for scoring them could use improvement, such as clearer definitions and/or distinct landmarks for measuring. Including reference planes, features, or points for visual analysis will be helpful because they will provide consistency between observers where there was otherwise none. Adding a tool to improve visual analysis, like a clear ruler, is also useful because it requires all observers to use the same technique rather than one observer using their eyes and another observer using their eyes plus a tool. Since it is known the human brain will see things differently based on the surrounding characteristics (Hopkins & McQueen, 2022), it is imperative that there is a constant characteristic across observations that comparisons can be made against.

It is interesting to note that quite a few traits can or do use metric measurements to help categorize a trait, since metric assessment boast greater precision, therefore, objectivity. At this point, some traits could no longer be considered a morphometric trait and, instead, a metric trait, such as PZT and MT. This is because there is no clear justification to why metric analysis should not be employed with these particular traits, or why these categories are more useful in combination with other morphometric traits rather than the measurements in combination with metric traits. That is to say there is nothing wrong with the categorization of expressions, just that if it can be measured on a continuum, then why constrain it to a few categories? Ultimately, the question is, do the traits that require metric assessment to help guide scoring absolutely have

to be used in morphometric analysis, or can they be more useful in metric analysis. There has been no publication, that this author identifies, that describes why these categories have to be used. Rather, it seems that the traditionally studied traits are being molded to fit into morphometric analysis instead of being completely reassessed for their applicability in ancestry assessments, particularly, with how they are used. In other words, are these traits continuing to be treated as morphological traits because there was no other research using them for metric analysis? These questions are brought on by the definition of morphoscopic traits, which are defined as quasi-continuous, not continuous. Therefore, the scales of "less than 2mm, 2-4mm, and greater than 4mm" is in fact an underlying continuous scale. A quasi-continuous categorization would exclude some measurements between the categories. However, it also brings the conversation back to the arbitrary categories that anthropologists are forced to make when it is unclear where the phenotypic shift occurs, which is likely the case. Since categorization of PZT and MT are dependent on a continuous scale, they are a great example for why they could be measured as and analyzed with metric traits. Furthermore, only size is taken into account in these two traits rather than shape, such as in NAW or IOB that use proportions for comparison. It would be preferable that shape is included because shape has been more useful than size to estimate ancestry using traits, such as NAS and NBS (McDowell et al., 2015). Unless there is a shape description, PZT and MT should be metrically analyzed.

The improvement for scoring NAW and IOB with the use of ratios is exciting because research has shown that quantifying morphoscopic traits with metric analysis has higher accuracy for ancestry estimation than simple direct visual analysis (McDowell et al., 2012). McDowell et al. (2012), however, only used one trait for ancestry estimation rather than a suite of traits, which is known to be more valuable for information on population variation (Edge &

Rosenberg, 2015; Ousley et al., 2009). If many of the morphoscopic traits can be quantified, such as McDowell, L'Abbé, and Kenyhercz (2012) are suggesting, perhaps ancestry assessment accuracy can be improved and not just the scoring consistency between observers. All this criticism is just to say that if traits are looking more like they can be measured metrically, it is acceptable to reassess their usefulness in different contexts. In fact, metric analysis should be researched in combination with metric analysis to determine if certain combinations of metric or morphoscopic characteristics are better indicators of ancestry than morphoscopic and metric analysis separately.

5.2 Ancestry estimation

In regard to ancestry estimates, the current research is only focused on assessing whether method reliability can affect ancestry estimate outcome. It was expected in the third hypothesis that a change in scores would affect the resulting ancestry estimate, therefore, the null hypothesis that there would be no change in resulting ancestry estimates when scores are different would likely be rejected. This was found to be true, and the null was rejected, because scores that differed between observers and observation periods resulted in different ancestry estimates. It was observed that the majority of ancestry estimates generated from observer 1's scoring sessions agreed within each program. This is consistent with a high kappa value, or intra-observer agreement for scoring periods 2 and 3 (Table 10). If the disagreements in scores for one individual are few, then it is less likely the ancestry estimate is going to change because combinations of scores, rather than individual scores, are compared to reference populations. Kamnikar et al. (2018) also saw that fluctuation in intra-observer scoring minimally impacted ancestry assessments with only a 1-5% drop in classification accuracy and had similar

distributions. These results might have been due to high internal consistency, such as what the author experienced.

More disagreement between ancestry results generated from observer 1's data and observer 3's is also consistent with lower kappa values (Table 13), especially when comparing the estimates to *Fordisc's* results. More traits with disagreed scores that are further apart also contribute to these ancestry estimate disagreements. Surprisingly, the ancestry estimates for the *HefneR* program generated based on observer 3's data agreed with those generated from observer 1's data more often than those generated from observer 2's data, even though there was low agreement across traits with observer 3. This may be due to sample size because observer 2 scored 1.5x more individuals than observer 3. Alternatively, it may be due to different traits that were disagreed upon, where one trait expression might have a stronger relationship with one population and a change in that expression can drastically change the result. In addition, ordinal trait score disagreements may not change the probability as drastically as a nominal trait score changing. However, there does not seem to be any literature discussing whether the type of trait and their scoring system would affect ancestry estimates differently. Overall, since agreement with observer 2 was higher across traits as well as with ancestry results between metric and morphoscopic traits, this means experience with trait expression plays a large part in the application of the method and its results.

What was surprising was that a majority of the *MaMD Analytical* and *HefneR* estimates did not agree. It was expected that the null hypothesis was correct and programs would return the same result since they used the same data, but this was untrue, and the null was rejected. The estimates came from the same scores from the same observer and, assuming, the programs used a pool of reference data from the same databank. This means these programs should have higher

agreement with each other than with the metric application, *Fordisc*. Similarly, observer 3's and observer 2's results agreed with observer 1's within each software more often than they agreed with their own results between *MaMD Analytical* and *HefneR*, resulting in similar proportions of ancestry assessments within each software (Figure 39). This pattern may lie in the way the different statistics used the data to estimate ancestry. For example, the developers of *HefneR* used frequencies from the 2009 publication as opposed to raw data. They also differ in the type of statistics, such as *MaMD Analytical* that uses a neural network whereas *HefneR* used a simple Bayesian classifier. Finally, *MaMD Analytical* was built on a pool of reference individuals from the MaM Databank that may have differed from the pool used in the 2009 article. Thus, it would be interesting to study the proportions of ancestry estimates within a sample population as compared to the proportion of ancestry of reference individuals. These proportions could be reflecting the variation found within the reference populations (Usher, 2002).

Different iterations of the scoring method also affected ancestry results depending on the program used. Observer 2 used Hefner's 2009 scoring method whereas observer 1 used the updated version in the atlas (Hefner & Linde, 2018). Furthermore, *HefneR* was based on Hefner's 2009 article whereas *MaMD Analytical* is more recent and its database is likely composed of more data from the atlas' scoring method (Hefner & Linde, 2018). These two pieces of information in conjunction with observer 2's *HefneR* results agreeing with *Fordisc* more often than observer 1's, and observer 1's *MaMD Analytical* results agreeing with *Fordisc* more often than observer 2's are evidence for this enhanced ancestry agreement between two programs. Interestingly, a handful of individuals had their *Fordisc* results agree with results from both observer 1's *MaMD Analytical* results and observer 2's *HefneR* results. This result can also be evidence that the iteration of the method works beneficially with the program based on that

same method. If observer 1 has more experience with the updated version, and the updated method corresponds to the data within *MaMD Analytical*, then it makes sense observer 1's *MaMD Analytical* results are in agreement with *Fordisc* more often. Likewise, if observer 2 is more familiar with the 2009 version and *HefneR* is based on this scoring method, then it is expected that her scores are in agreement more often in this application than *MaMD*. The same individuals that have *Fordisc* results agreeing with results from two different programs used by two different observers means that each observer's familiarity with the method increases the accuracy of the statistics that rely on that method.

Finally, the author noticed that a change in one trait expression score could drastically change the ancestry estimate in *MaMD Analytical*. Upon accident, the author input a ZS score of 0 for individual #8 and recorded the result as 'Guatemalan'. When changing the scores for the next individual to be analyzed, the author noticed that ZS was incorrect. The score was corrected to a score of 2 and the data was rerun, resulting in the ancestry of 'American White.' These two expressions were most commonly disagreed upon for ZS and provides an interesting perspective to how disagreement in scores can affect ancestry estimates. Furthermore, individual #7 had one trait with a mistaken score, PBD, that changed the estimated from 100% 'Black' to 73% 'White'. While these may have been simple recording errors, the fact that one of the situations involved scores that were often confused with one another is a concern for inter-observer agreement and its impact on these assessments. This is directly in contrast to Kamnikar et al. (2018)'s intra-observer study that found score differences did not drastically change the resulting ancestry. Thus, another study needs to work out whether inter-observer agreement among experienced observers will impact ancestry estimates the same way that intra-observer agreement does in the current study.

Some other explanations for ancestry estimations not agreeing between software programs can be based on temporal differences in unknown individuals and reference individuals. Ancestry estimates from both *MaMD Analytical* and *HefneR* have a greater number of agreements with estimates generated from Howells database in *Fordisc*, which includes 19th and 20th century individuals, than with the Forensic Databank Bank in *Fordisc*, which include contemporary reference individuals. This agreement between Howells and the morphoscopic programs could be because the teaching collection individuals are historic and the MaM Databank, partially used for *MaMD Analytical,* includes both modern and historic reference individuals (Hefner, 2018). Any secular changes in score frequencies are going to affect estimation, and there is known secular change for some nonmetric traits (Kilroy et al., 2020). Likewise, differences between FDB and Howells results were expected because of the known secular changes in metric measurements over the last couple hundred years (Ayers et al., 1990; Wescott & Jantz, 2005). Furthermore, individuals likely have multiple ancestries and cannot be slotted into one category. The combination of these data should be explored to see if ancestry estimate agreement between different software programs that are using different types of data (morphological vs metric) can be improved.

The author also had difficulties running some data in *Fordisc.* The stepwise function was used in some cases because without it, the program could not place the individual into a population. Unfortunately, the equipment used could only run a combination of six traits for the entire reference database. A greater number of traits would result in too many iterations and would run out of memory before completion. Bulbeck (2011) recommended at least sixteen measurements for the most reliable results, thus some of the individuals likely did not have reliable ancestry results that could be compared to the morphoscopic results.

The most likely explanation for ancestry estimates not agreeing between programs, and within *Fordisc* between Howells and FDB is that the individual's ancestry falls outside of those in the reference populations. India was the largest exporter of teaching individuals at one point (Carney, 2011; Fabian, 2010), and their ancestry is excluded from both morphoscopic and metric reference populations. One BU individual has a company associated with them, Kilgore International, that is known to have bought skeletons from India (Carney, 2011). It is likely many individuals in both teaching collections came from the same company or place of origin. If this is the case, then reference individuals from India need to be included in these databases. They represent variation between East Asia and Africa that likely have a vastly different phenotype as compared to the rest of Asia. As it is known, population variation is expressed on a continuum (Berry & Berry, 1967; Breske, 2018; Hunley, Cabana, & Long, 2016; Kaestle & Horsburgh, 2002; Konigsberg, 1990; Laughlin & Jorgensen, 1956; Relethford, 2004, 2016; Wright et al., 2018; S. Wright, 1943) where there is no distinct line between populations. Therefore, much of the middle East and West Asia being excluded is a large chunk of variation missing from the analysis.

5.2.1 Importance of ancestry estimation programs not matching

This research is not focused on addressing the methodological errors for ancestry estimates. However, the programs disagreeing much of the time is still a concern. Right now, practitioners are likely using *HefneR* because it is freely available and easy to use. When *MaMD Analytical* is published, there needs to be a shift to this software because it will have far more statistical validation due to testing than *HefneR.* This switch will ensure that practitioners are using a consistent statistical method; in other words, the data is being used the same way and compared to the same reference individuals. If there are practitioners that do not make the

switch, the lack of consistency means unreliable results since this research has shown that disagreement occurs between the two programs. Disagreements between ancestry results can set this type of ancestry method back from meeting legal standards of reliability and validity (*Daubert v. Merrell Dow Pharmaceuticals, Inc*, 1993; *General Electric Co. v. Joiner*, 1997; *Kelliher (Village of) v. Smith*, 1931; *Kumho Tire Company, Ltd. v. Carmichael*, 1999; *Regina v. Mohan*, 1994). At this point, issues with disagreement in ancestry estimates between programs that should be measuring the same thing should be fully discussed in the literature. For example, there needs to be a discussion on why they may not agree, such as the author has discussed where it may be due to method iteration.

The implications of *HefneR* not agreeing with *MaMD Analytical* when using the same data is that the ancestry estimates from either software may not be accurately reflecting the individual's ancestry. These inaccurate estimations may impact case progression and reflect poorly on the methods used in forensic anthropology. Ideally, practitioners should wait until both statistical programs are tested and validated on other skeletal populations before use.

Likewise, having very few results agreeing across both metric and morphometric statistical analysis from the same observer is concerning. This is likely due to the reference populations used, but having less than chance agreement means there are some methodological errors that need to be addressed. If two programs are intended to measure the same thing, then they should match in ancestry more often than not. The low matching of ancestry estimates can be problematic in the discipline if one practitioner is using metric measurements while another uses morphometric traits. There needs to be further testing on program agreement and a discussion on why *Fordisc* and *MaMD Analytical* do not agree or why metric and morphometric data do not result in the same estimate. Then, researchers need to discuss how to reconcile these

differences, and what to do in cases where ancestry estimation does not agree. If metrics and morphometrics are used together, having one individual whose ancestry estimate is different between programs can cause confusion on how to record that ancestry estimate and uncertainty on which result to trust.

This is not to say that one of the ancestry estimates are incorrect because populations and groups of people are not mutually exclusive from one another, as learned from Lewontin's (1972) genetic study. Different traits are going to vary independently of other traits in each population due to differing forces acting on them (Molnar, 2002). Instead, the positive impact of having two methods of ancestry estimation is that there is more variation that can be analyzed for estimating ancestry. Having two sets of data that could be used together to understand how metric and morphometric data relate and differ should only improve the accuracy of estimates in the future. Until it is known if and how these traits co-vary, and if one program has a higher accuracy, then estimates should still be produced via metric analysis as this will provide consistency. Morphoscopic traits used for ancestry estimation may be risking evidence being discarded because the scoring method has low inter-observer agreement or because there is disagreement occurring between two programs that should be measuring the same thing.

Only one study has compared the results of metric and morphometric data on the same known individuals, but it was only for one trait rather than a suite of traits (McDowell et al., 2015). This is counter-intuitive because it is known that a trait in isolation is not useful for estimating ancestry because each population holds the majority of the full range of trait expressions (Lewontin, 1972). It is the varying frequencies of traits in combination that help differentiate populations (Brettell, 2013; Brues, 1990; Cheverud & Buikstra, 1981; Edge & Rosenberg, 2015; Gill, 1998; Klales & Kenyhercz, 2015; Moffit, 2017; Ousley et al., 2009;

Ousley, Jantz, & Hefner, 2018). Both types of data and their results should be thoroughly tested and discussed for reliable use.

Some researchers state that some traits are less susceptible to secular changes (Maddux et al., 2015), however, this is for only one trait and two Indigenous populations rather than multiple populations. In contrast, Konigsberg (1990) and Macchiarelli et al. (1995) did not see significant secular changes. This statement needs to be validated or justified before a practitioner can put their faith in morphoscopic traits over metric traits.

Finally, Liebenberg and Krüger (2020) recommend that when methods are altered and improved on, the reference populations should be reassessed with these method improvements so information remains relevant and accurate. Since trait data are used for statistical analysis, these data must be relevant to the current scoring method used. Improvements mean there are changes that might affect how a score is obtained. After Hefner's (2009; Hefner and Linde, 2018) scoring method has been improved and validated, the MaMD Databank reference individuals should be reassessed, or new reference individuals included since *MaMD Analytical* will be using that data to analyze data collected using the updated method.

On one last unrelated note, the introduction of *MaMD Analytical* with results only falling into a few categories with mainly racial labels (American Black, American White, American Indian, Guatemalan, Southwest Hispanic) means the debate of race versus biology rears its ugly head again. If there is access to worldwide data (Hefner, 2018), then why are racial labels grouping these individuals? It would make more sense that there are labels for "European American" and "European" to differentiate the two populations if this was a program geared towards American practitioners but applicability to worldwide populations. Social labels are confusing for worldwide applicability, and are misleading. Someone who is categorized as

'American White' might not actually have "white" skin because it has been shown that people from the same region are more likely to share more genetic loci regardless of socially ascribed "race" (Ousley et al., 2018). The use of these labels should be justified, and the reference populations should be explicitly stated for clarity.

5.3 General recommendations

1. More inter-observer studies on this method need to be conducted among observers that are considered experts in skeletal anatomy outside Hefner's direct training. In these studies, each trait should be discussed regardless of whether the observer's scores agree or not. These discussions should break down how exactly each observer scored the trait, if they used tools or not, and how they differentiate each trait expression from another. It was through these discussions that differences in technique and interpretation revealed itself even if each observers' scores agreed, and even though the same traits were scored and similar guides were used.
2. Every time observer error is researched, prevalence of each score for each trait should be discussed. Simply having high kappa values does not mean that the trait is described and depicted clearly for each expression; a score of 3 may never appear in the sample population, and this does not necessarily mean all observers agree on what a score of 3 looks like.
3. The description of each trait should include an explicit explanation as to how to score the expression when it does not match the expressions listed, and how to score in situations of asymmetry. For example, trait NO is described in both the *Osteoware* manual (Wilczak & Dudar, 2020) and MMS User Manual (Hefner et al., n.d.) that the left side should be scored if undamaged, but it is left out in the atlas (Hefner & Linde, 2018).

4. If the atlas (Hefner & Linde, 2018) is to be used for scoring, it should be updated to include line drawings over top of or beside the photographs to clarify exactly where each landmark is that is used to measure the trait, or outlining the shape on the individual. This makes it explicit where identifying features are located.
5. Before using *MaMD Analytical* for forensic anthropological casework, it should be studied in comparison to *Fordisc*, a program that is supposed to be measuring the same thing, albeit using different data. There needs to be a discussion on the differences between the two programs and what to do when their results do not agree.
6. Ancestry estimates should focus on geographic region rather than socially ascribed race, to ensure consistency between methods and accuracy in different countries. Then, if the practitioner prefers, interpretation can be made for racial labels based on the population an individual is grouped with. While this is not the recommendation of the author, if race is going to be the ascribed label, it should be justified.
7. Finally, since training is an important component to using this method, it is suggested that Hefner offer workshops or provide a training video that can be shared with practitioners who request it. Since the method is visual, he could explain step by step how he scores each trait, showing the audience while he does it. He can also show how he distinguishes different expressions while showing examples of these expressions. This could compliment the line drawings and descriptions since the author found that the atlas answered many questions about how to score certain expressions. In today's world, virtual training is an amazing new option that should be utilized to ensure consistency between practitioners around the world.

5.3.1 Importance of trait descriptions being improved by these recommendations

If both the general and trait specific recommendations are implemented, it will begin to bring the method in line with the reliability and validity standards needed for forensic disciplines (*Daubert v. Merrell Dow Pharmaceuticals, Inc*, 1993; *General Electric Co. v. Joiner*, 1997; *Kelliher (Village of) v. Smith*, 1931; *Kumho Tire Company, Ltd. v. Carmichael*, 1999; *Regina v. Mohan*, 1994). These recommendations, and implementation of previous recommendations (Kamnikar et al., 2018), can be used as evidence that the method is being tested and sources of error are being eliminated or reduced. Not only does testing and implementing changes improve repeatability and ancestry estimation accuracy, it reduces the chance that the results are excluded from evidence (Page et al., 2011a, 2011b). As Hefner (2014, p. 40) himself said, "Learn and understand the traits first, then apply statistical methods of classification." While Hefner knows the traits and what he meant by the descriptions, not all other practitioners do. It is up to him to provide clear and precise descriptions if he wants others to collect the same data. Much like Edgar's (2017) reason to improve clarity on the ASUDAS method (if inexperienced observers will be collecting data regardless of training, at least try to improve the method so they can collect better data) that is what the author recommends. These changes must be made so there is less chance of error, especially by less experienced observers.

Additionally, discipline standards are important for consistency among practitioners. If Hefner's method and statistical program are going to be a discipline standard for morphometric traits, then it should be thoroughly tested. If the MaM Databank is built using a method that has not been thoroughly tested for error, the database will have inaccurate data that will not reflect actual trait frequencies. Low repeatability in scoring means that scores which are generated for an unknown individual and then are compared to the database to determine ancestry may not

necessarily mean the same thing as the scores for the references individuals in the database. This is because reference individuals are scored by different observers (the individual who collected data on the reference collection versus the practitioner) who may have different interpretations of the same scores. Increasing repeatability ensures reduced error in the scores within the database, resulting in a more comparable database and more accurate ancestry estimations. For example, Liebenberg and Krüger (2020) recently tested and recommended description changes for some osteometric traits. They recommended updating certain reference population measurements to take into account the description changes so that the reference data are up to date and consistent with international data collection. *MaMD Analytical*, while useful, may be ahead of its time in the case of method testing. As it was seen with the outcome of the inter-rater reliability, differences in scores can alter the ancestry estimate given. It is not possible to build an accurate morphoscopic reference data pool before the issues with the scoring method are addressed.

5.4 Future directions for method testing

In general, the recommendations outlined in this research should be evaluated by Hefner to determine if these sources of error and recommended improvements are beneficial for improving the method. If determined to be useful, these recommendations should be implemented and the method re-tested. Upon implementation, research needs to focus on inter-observer study; there are enough intra-observer studies on this method. Determining *where* error is occurring is far more important to the future of morphometric trait use than how often practitioners agree within their own scoring sessions. Thorough study and testing by independent practitioners, outside Hefner's direct training, to understand inter-observer agreement can reduce unintentional bias that is not otherwise caught. Discussion around the method, how each

practitioner scores the trait, and how they differentiate each expression is the future direction to improving the repeatability of this ancestry assessment method.

Importantly, a discussion on prevalence of scores is also required. Current publications on observer error do not discuss this issue (C. Coelho et al., 2017; Dinkele, 2018; Hefner, 2009; Hurst, 2012; Kamnikar et al., 2018; L'Abbé et al., 2011; Moffit, 2017; Wang, 2016) (Andrade et al., 2018; C. Coelho et al., 2017; Dinkele, 2018; Hefner, 2009; Klales & Kenyhercz, 2015; L'Abbé et al., 2011). It is unlikely all the trait expressions appear equally as often. Therefore, kappa values could be reflecting that there is high agreement between only two scores amidst five possible scores. It is assumed that Hefner has the most experience with these traits, being involved with multiple publications on various populations (Go & Hefner, 2020; Hefner et al., 2015; Monsalve & Hefner, 2016), thus, prevalence is likely not an issue during his intra-observer testing (Hefner, 2009) and data collection because he is familiar with all the expressions even if they do not appear. This is acceptable for consistency of reference data collection on his part, however, not all practitioners have this experience. The range of expressions found within each population limit the observer to these populations, necessitating a discussion of prevalence to understand agreement among particular expressions. This issue of prevalence is also important for why the method needs to be very clear and easy to use; if practitioners cannot access a population to understand the full extend of variation, the method must makeup for this deficiency. If this author's recommendations are implemented, it is likely the method will have improved repeatability for the sake of the cases where observers do not have access to individuals with the full range of expression.

Since the guide needs to be clear enough to provide consistency across practitioners, one question that could be explored is whether utilizing the atlas will improve scoring agreement.

Using a small subsample in this study, the majority (79.6%) of scores did not change after viewing the atlas photos (Table 73, Appendix 3). This may be because observer 1 used the atlas as a source of training while observer 3 and observer 2 did not view it. Having examples of the variation in trait expressions was helpful when the line drawings did not encompass what observer 1 was looking at. In some cases, the expression on another individual was close to what observer 1 was looking at and helped determine the score. Unexpectedly, of the scores that did change, the majority did not match the scoring session where the atlas (Hefner & Linde, 2018) was used (scoring period 2), indicating that the photos were not helpful in cases of uncertainty.

In regards to using technology for improved objectivity, the contour gauge also needs further testing. There needs to be comparison of how each trait expression appears on the contour gauge, and further description and depiction of how to determine each NBC expression from the contour. The author suggests using angles to measure whether the walls are steep or shallow, therefore, future study can be taking these measurements between individuals to find a range that defines steep or shallow. Furthermore, individuals can take contour gage measurements and compare the contours between scoring sessions to determine whether the score changed along with a shape. This technique should also test out placing the contour gauge in different spots on the nasals that are millimetres apart to see if the shape changes, thus a change in score.

While 3-D study is a popular choice for technological advances in the discipline, the author does not believe this technology or accompanying analyses will be useful for practical purposes. Forensic anthropologists do not carry around the technology that would be used to replicate studies that use 3D analysis (Stull et al., 2014). Instead, this area of research can prove useful for finding further differences in trait expressions that may improve the descriptions for visual analysis. It can also be useful to compare the shape expressions identified visually to

shape expressions identified through geometric morphometrics. New expressions not originally identified through visual analysis might appear, shifting the scale from just a few categories to more.

Finally, this study did not focus on identifying errors in ancestry estimations, only whether estimation disagreement occurred. Therefore, ancestry estimation accuracy using this method should only be tested after the method has thoroughly been tested for inter-observer agreement, and error reduced as much as possible. Subsequently, score differences among observers should be studied further to see what kinds of disagreements will cause more severe ancestry estimation differences.

5.5 Issues in the methods of the current research

As with all research, this research was not carried out perfectly due to unforeseen circumstances (i.e. a pandemic) causing the author to re-evaluate how to continue with testing Hefner's (2009) method. The sample population had several issues. The original sample size would have been four times the size that it ended up being (n=27), likely expressing far more variation in traits than the teaching collection expressed. This variation would have been useful for teasing apart patterns of disagreement and may have had a more accurate reflection of inter-observer agreement. Additionally, some individuals were not available during each data collection session, therefore, there were less chances for the author to agree on what a trait expression looked like. The missing individuals were only able to be assessed once, and occurred at the same time as the second scoring session for the other individuals. In addition, the timing between scoring sessions was not exactly two months in some individuals because of the availability of the laboratory supervisors.

Technical errors include that the teaching individuals were structurally altered for the classroom. These alterations impacted some metric measurements needed for comparison and likely affected the estimates. Unfortunately, the author's laptop could not handle the stepwise function for *Fordisc* and could not do a thorough analysis on all individuals.

In terms of demography, one of the individuals was slightly younger than the rest, and may have affected trait expression. The teaching individuals may also not have been the best option to compare results of *MaMD Analytical* because they are likely historic and any secular changes (Ayers et al., 1990; Spradley, 2014; Wescott & Jantz, 2005) could be affecting the ability for the program to assign ancestry. While *MaM Databank* does include some historic reference populations, and protohistoric Indigenous populations when being used for repatriation, historic individuals are not the intended recipient of the program. If these individuals are early anatomic specimens, then Howells in *Fordisc* is likely the program to use for ancestry estimates and are outside the temporal period intended by *MaMD Analytical*. Finally, the biggest issue with using the teaching individuals is that they did not have known ancestry. The author could not confirm the results of each program. This would have been useful for determining which program was more accurate, and acted as a validation study for *MaMD Analytical.*

The author also noted issues with carrying out Hefner's (2009; Hefner and Linde, 2018) scoring method. The updated method was not used by observer 2 and would have provided data on five other traits. It was not explicitly stated by the author for observer 2 to use the updated method, and was assumed she was going to use the updated method. However, upon reflection, she had just acquired the atlas (Hefner & Linde, 2018) and likely did not have time to look at it, and she regularly uses *HefneR* in casework, which does not require the other five traits. It would

be interesting to know if agreement between observer 1 and observer 2 would be improved if observer 2 reviewed the atlas and rescored the individuals.

The author noted a subconscious decision by herself to alter the method rather than following it exactly, such as the case of using a skewer to score OBS where the instructions did not state to use it. Furthermore, the assumption that the ratios for IOB and NAW were meant to be calculated rather than eyeballed means the technique might have been carried out differently than intended. While these decisions were helpful in determining that the techniques for scoring each trait needed improvement, it should have been carried out exactly so the kappa scores were accurate. The author also noticed that she needed more experience with broken facial bones because it was difficult to tell if the edge was actually broken or just naturally wavy.

Finally, not all traits and trait expressions were discussed with observer 2 during training due to time constraints, and lack of forethought by the author. The discussions that took place were dependent on whether the observers' scores disagreed. It was assumed if all observers had the same score then it was scored the same way, but that was revealed to be incorrect later on. While there were many areas of improvement for this research, the results still allowed for a comprehensive review and study of error that no other study has done up to this point.

5.6 Summary and Conclusion

This research is the first to comprehensively assess Hefner's (2009; Hefner and Linde, 2018) scoring method for error inherent and where it arises between observers. As expected, intra-observer agreement was high ($k \geq 0.61$), and inter-observer agreement was lower than intra-observer agreement; the results reject the null hypothesis that agreement is occurring at a rate of chance. Errors outside of inexperience came from the scoring technique used when descriptions did not outline a specific technique, such as observer 1 using a toothpick for determining ratios

in OBS and NAS, or observer 3 using a ruler for NFS. Error also came from the iteration of the method used, with observer 2 using descriptions from Hefner (2009) whereas observer 1 used Hefner (2009) with updates from Hefner and Linde (2018). It was also revealed that certain traits and trait expressions were more difficult to distinguish, either resulting in greater error or inconsistent application. For example, INA expressions 1 and 2 were difficult to distinguish because it was hard to determine where a slope started within the nasal cavity. Overall, each trait has its own set of inherent errors that need to be rectified and this study attempted to improve the descriptions by studying where error was occurring.

Most of the studies testing Hefner's scoring method (Hefner, 2009; Hefner and Linde, 2018) were intra-observer studies that focused on the percentage of agreement (Atkinson & Tallman, 2020; Coelho et al., 2017; Dinkele, 2018; Hefner, 2009; Hurst, 2012; Kamnikar et al., 2018; Kilroy et al., 2020; L'Abbé et al., 2011; Moffit, 2017; Wang, 2016). These studies are not helpful for determining where error is occurring because it was shown through this research that there is internal bias for description interpretation, as well as high internal consistency for applying the interpretation. Furthermore, Kappa statistics alone, across multiple studies, are not helpful in determining which traits are more or less error prone, even in inter-observer agreement studies (Coelho et al., 2017; Dinkele, 2018; Hefner, 2009; Klales & Kenyhercz, 2015; L'Abbé et al., 2011). This is because there is no discussion on bias or prevalence which might help interpret the Kappa statistics (Banerjee et al., 1999; Byrt et al., 1993; Hallgren, 2012). A trait may have high agreement because there were only two of the four expressions present in the population. Therefore, it is required that there are more inter-observer studies that discuss bias as brought on through interpretation and prevalence of trait expressions in each population. This is the first

study to actively discuss these issues for each trait expression. These issues were also important for identifying where this study is lacking information on error for some trait expressions.

This is also the first comprehensive study to understand how these errors impact scoring, and, ultimately, an ancestry estimate by comparing the results of multiple statistical programs. As expected, the disagreement in scores between observers and scoring periods affected the ancestry estimates, thus rejecting the null hypothesis that there would be no change in estimates if the score changed. Furthermore, one null hypothesis stating that there should be no difference in ancestry estimate grouping between two programs that use the same data to result in the same thing, ancestry, was rejected because the estimate agreement was equal to or less than that of chance. It is important that the error in the scoring method is eliminated or reduced to prevent these issues from occurring. This study illuminated these issues by carrying out practices, such as external validation, that are used in quality assurance programs because these programs are recommended for implementation in the discipline (Fleischman et al., 2019). Not only do these practices improve methods, they help them meet the legal standards of reliability and validity (*Daubert v. Merrell Dow Pharmaceuticals, Inc*, 1993; *Regina v. Mohan*, 1994). Since they do not exist at this time, it is up to external researchers to test the method and discuss with the original method creator where error could be reduced. This study is the first to comprehensively evaluate the scoring method (i.e., all sixteen traits) for error outside of the direct training of Hefner (2009).

Error rates are a component of the legal requirements to determine method reliability (*Daubert v. Merrell Dow Pharmaceuticals, Inc*, 1993) and a main result of validation studies (Budowle et al., 2009; Christensen & Crowder, 2009; Fleischman, Pierce, & Crowder, 2019; SWGANTH, 2011). However, error rates for the scoring method cannot be established from the

current suite of studies using Kappa statistics of agreement since they do not discuss bias or prevalence (Atkinson & Tallman, 2020; C. Coelho et al., 2017; Dinkele, 2018; Hefner, 2009; Hurst, 2012; Kamnikar et al., 2018; Kilroy et al., 2020; Klales & Kenyhercz, 2015; L'Abbé et al., 2011; Moffit, 2017; Wang, 2016). It is unknown how often certain expressions are causing disagreement since this study showed that bias and prevalence affect agreement of particular scores.

This research has the potential to improve ancestry assessment methods and ancestry estimation accuracy if recommendations are implemented. It is the hope that this research and resulting recommendations will be used to improve the scoring method so that it can, ultimately, increase the accuracy of the statistical analysis. It also impacts the research of non-metric ancestry assessment methods in general because it is hoped that future studies will use the structure of this research to focus on qualitative analysis to find error. In addition, it is hoped that this research will promote the explicit discussion of prevalence and bias which will help determine what traits or expressions are causing the most issues among practitioners. This research also impacts the study of this specific method because future studies can focus on the specific trait expressions that did not appear frequently in the current research. The results of this research also have the potential to increase the chances of reuniting skeletal remains with their family members; this is achieved by improving the repeatability of the scoring method so it can improve the accuracy of the estimates. Concurrently, it can help with repatriation efforts where individuals have an unknown origin with no paper trail.

Finally, this research will impact forensic anthropology by helping morphoscopic ancestry assessment methods move towards a discipline standard. This is achieved through the practices used in validation processes, as these processes are within quality assurance programs,

which are recommended for implementation in forensic anthropology. A discipline standard will help bring forensic anthropology in line with other forensic disciplines, and increase the validity of the discipline in court.

Overall, this study has revealed sources of error otherwise not discussed in the current publications. It also has several benefits to the forensic anthropology community, research and practice, and the wider public. The continuation of this research will only improve the method and the practices used by other anthropologists.

References Cited

AABA. (1996). AAPA statement on biological aspects of race. *American Journal of Physical Anthropology*, *101*(4), 569–570. https://doi.org/10.1002/ajpa.1331010408

Aguillon, S. M., Fitzpatrick, J. W., Bowman, R., Schoech, S. J., Clark, A. G., Coop, G., & Chen, N. (2017). Deconstructing isolation-by-distance: The genomic consequences of limited dispersal. *PLoS Genetics*, *13*(8), 1–27. https://doi.org/10.1371/journal.pgen.1006911

Alberti, S. J., Bienkowski, P., Chapman, M. J., & Drew, R. (2009). Should we display the dead? *Museum and Society*, *7*(3), 133–149.

Andrade, B., Dias, P., Santos, B. S., Coelho, C., Coelh, J. D., Navega, D., … Wasterlain, S. (2018). Morphological analysis of 3D skull models for ancestry estimation. *Information Visualisation - Biomedical Visualization, Visualisation on Built and Rural Environments and Geometric Modelling and Imaging, IV 2018*, 567–573. https://doi.org/10.1109/iV.2018.00104

Armelagos, G. J., & Gerven, D. P. Van. (2003). A century of skeletal biology and paleopathology: Contrasts, ontradictions, and conflicts. *American Anthropologist*, *105*(1), 53–64. https://doi.org/10.1525/aa.2003.105.1.53

Atkinson, M. L., & Tallman, S. D. (2020). Nonmetric cranial trait variation and ancestry estimation in Asian and Asian-derived groups. *Journal of Forensic Sciences*, *65*(3), 692–706. https://doi.org/10.1111/1556-4029.14234

Auerbach, B. (n.d.). The William W. Howells Craniometric Data Set. Retrieved September 27, 2022, from http://web.utk.edu/~auerbach/HOWL.htm

Auxier, J. (1989). The role of the expert witness. *Radiation Research Society*, *117*(2), 178–180. Retrieved from https://www.jstor.org/stable/3577317

Auxier, J., & Prichard, H. M. (2001). The role of the expert witness: An update. *Health Physics*, *81*(3), 269–271. https://doi.org/10.1097/00004032-200109000-00008

Ayers, H., Jantz, L. R., & Moore-Jansen, P. (1990). Giles and Elliot race discriminant functions revisited: A test using recent forensic cases. In G. Gill & S. Rhine (Eds.), *Skeletal Attribution of Race: Methods for Forensic Anthropology.* (pp. 65–71). Albuquerque, NM: Maxwell Museum of Anthropology.

Banerjee, M., Capozzoli, M., McSweeney, L., & Sinha, D. (1999). Beyond kappa: A review of interrater agreement measures. *Canadian Journal of Statistics*, *27*(1), 3–23. https://doi.org/10.2307/3315487

Bass, W. (2005). *Human Osteology: A Laboratory and Field Manual* (5th ed.). Missouri Archaeological Society.

Beech, C. (2015). The admissibility of expert opinions in Canada courts. *Nafta - Law and Business Review of the Americas*, *21*(3), 361–368.

Bergmann, K. (Ed.). (2011). *Native American Graves and Repatriation*. Hauppauge, N.Y: Nova Science Publishers.

Berry, A., & Berry, R. (1967). Epigenetic variation in the human cranium. *Journal of Anatomy*, *101*(Pt 2), 361–379. Retrieved from

http://www.ncbi.nlm.nih.gov/pubmed/4227311%0Ahttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC1270890

Bethard, J. D., & DiGangi, E. A. (2019). From the laboratory to the witness stand: Research trends and method validation in forensic anthropology. In L. Fulginiti, K. Hartnett-McCann, & A. Galloway (Eds.), *Forensic Anthropology and the United States Judicial System* (1st ed., pp. 41–52). Chichester, UK: John Wiley & Sons, Ltd.

Birkby, W. (1966). An estimation of race and sex identification from the cranial measurements. *American Journal of Physical Anthropology*, *42*, 21–27.

Blackwell, S., & Seymour, F. (2015). Expert evidence and jurors' views on expert witnesses. *Psychiatry, Psychology and Law*, *22*(5), 673–681. https://doi.org/10.1080/13218719.2015.1063181

Bolhofner, K., & Seidel, A. (2019). Expertise and the expert witness: Contemporary educational foundations of forensic anthropology. In L. Fulginiti, K. Hartnett-McCann, & A. Galloway (Eds.), *Forensic Anthropology and the United States Judicial System* (1st ed., pp. 53–68). Chichester, UK: John Wiley & Sons, Ltd.

Brace, C., & Hunt, K. (1990). A non-racial craniofacial perspective on human variation: A(ustralia) to Z(uni). *American Journal of Physical Anthropology*, *82*, 341–360.

Breske, A. (2018). Politics of repatriation: Formalizing Indigenous repatriation policy. *International Journal of Cultural Property*, *25*, 347–373. https://doi.org/10.1017/S0940739118000206

Brettell, S. A. (2013). A Validation Study Examining Hefner' s "Cranial Nonmetric Variation and Estimating Ancestry." *University of Tennessee Honors Thesis Projects*. Retrieved from http://trace.tennessee.edu/utk_chanhonoproj%5Cnhttp://trace.tennessee.edu/utk_chanhonoproj/1660

Brues, A. (1990). The once and future diagnosis of race. In G Gill & S. Rhine (Eds.), *Skeletal Attribution of Race: Methods for Forensic Anthropology.* (pp. 1–4). Albuquerque, NM: University of New Mexico: Maxwell Museum of Anthropological Papers.

Budowle, B., Bottrell, M. C., Bunch, S. G., Fram, R., Harrison, D., Meagher, S., … Stacey, R. B. (2009). A perspective on errors, bias, and interpretation in the forensic sciences and direction for continuing advancement. *Journal of Forensic Sciences*, *54*(4), 798–809. https://doi.org/10.1111/j.1556-4029.2009.01081.x

Buikstra, J., & Ubelaker, D. H. (Eds.). (1994). *Standards for Data Collection from Human Skeletal Remains: Proceedings of a Seminar at the Field Museum of Natural History*. Retrieved from http://www.amazon.com/Standards-Collection-Human-Skeletal-Remains/dp/1563490757

Bulbeck, D. (2011). Principles underlying the determination of population affinity with craniometric data. *Mankind Quarterly*, *52*(1), 35–89.

Byrt, T. E. D., Bishop, J., & Carlin, J. B. (1993). Bias, prevalence, and kappa. *J. Clinical Epidemiology*, *46*(5), 423–429.

Campanacho, V., Alves Cardoso, F., & Ubelaker, D. H. (2021). Documented skeletal collections and their importance in forensic anthropology in the United States. *Forensic Sciences*, *1*(3), 228–239. https://doi.org/10.3390/forensicsci1030021

Carayon, D., Adhikari, K., Monsarrat, P., Dumoncel, J., Braga, J., Duployer, B., … Zanolli, C. (2018). A geometric morphometric approach to the study of variation of shovel-shaped incisors. *American Journal of Physical Anthropology*, *168*(1), 229–241. https://doi.org/10.1002/ajpa.23709

Carney, S. (2011). *The Red Market: On the Trail of the World's Organ Brokers, Bone Thieves, Blood Farmers, and Child Traffickers*. Toronto: Harper Collins Canada.

Carson, E. A. (2006). Maximum-likelihood variance components analysis of heritabilities of cranial nonmetric traits. *Human Biology*, *78*(4), 383–402. https://doi.org/10.1353/hub.2006.0054

Caspari, R. (2009). 1918: Three perspectives on race and human variation. *American Journal of Physical Anthropology*, *139*(1), 5–15. https://doi.org/10.1002/ajpa.20975

Cataldo-Ramirez, C., Garvin, H. M., & Cabo, L. (2020). A quantitative assessment of zygomatic projection for ancestry estimation. *Journal of Forensic Sciences*, *65*(2), 580–590. https://doi.org/10.1111/1556-4029.14188

Cheverud, J. (1982). Phenotypic, genetic, and environmental morphological integration in the cranium. *Evolution*, *36*(3), 499–516. https://doi.org/10.2307/2408096

Cheverud, J., & Buikstra, J. (1981a). Quantitative genetics of skeletal non-metric traits in rhesus macaques on Cayo Santiago. I. Single trait heritabilities. *American Journal of Physical Anthropology*, *54*, 43–49.

Cheverud, J., & Buikstra, J. (1981b). Quantitative genetics of skeletal non-metric traits in rhesus macaques on cayo Santiago. II. Phenotypic, genetic, and environmental correlations between traits. *American Journal of Physical Anthropology*, *54*, 51–58.

Cheverud, J., & Buikstra, J. (1981c). Quantitative genetics of skeletal non-metric traits in rhesus macaques on cayo Santiago III. Relative heritability of skeletal nonmetric and metric traits. *American Journal of Physical Anthropology*, *59*, 151–155.

Christensen, A. M. (2004). The impact of Daubert : Implications for testimony and research in forensic anthropology (and the use of frontal sinuses in personal identification). *Journal of Forensic Sciences*, *49*(3), 1–4. https://doi.org/10.1520/jfs2003185

Christensen, A. M., & Crowder, C. M. (2009). Evidentiary standards for forensic anthropology. *Journal of Forensic Sciences*, *54*(6), 1211–1216. https://doi.org/10.1111/j.1556-4029.2009.01176.x

Christensen, A. M., Leslie, W. D., & Baim, S. (2014). Ancestral differences in femoral neck axis length: Possible implications for forensic anthropological analyses. *Forensic Science International*, *236*(2014), 193.e1-193.e4. https://doi.org/10.1016/j.forsciint.2013.12.027

Cicchetti, D. V., & Allison, T. (1971). A new procedure for assessing reliability of scoring EEG sleep recordings. *American Journal of EEG Technology*, *11*(3), 101–110. https://doi.org/10.1080/00029238.1971.11080840

Clark, M. A., Guatelli-Steinberg, D., Hubbe, M., & Stout, S. (2016). Quantification of maxillary dental arcade curvature and the estimation of biological ancestry in forensic anthropology. *Journal of Forensic Sciences*, *61*(1), 141–146. https://doi.org/10.1111/1556-4029.12910

Coelho, C., Navega, D., Cunha, E., Ferreira, M. T., & Wasterlain, S. N. (2017). Ancestry

estimation based on morphoscopic traits in a sample of African slaves from Lagos, Portugal (15th–17th centuries). *International Journal of Osteoarchaeology*, *27*(2), 320–326. https://doi.org/10.1002/oa.2542

Coelho, J. D., Curate, F., & Navega, D. (2020). Osteomics: Decision support for forensic anthropologists. In Z. Obertova, A. Stewart, & C. Cattaneo (Eds.), *Statistics and Probability in Forensic Anthropology*. San Diego: Elsevier Inc.

Coelho, J. D., & Navega, D. (n.d.). *HefneR*. Retrieved from https://osteomics.com/hefneR/

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20*(1), 37–46.

Cohen, J. (1968). Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin*, *70*(4), 213–220.

Corruccini, R. S. (1974). An examination of the meaning of cranial discrete traits for human skeletal biological studies. *American Journal of Physical Anthropology*, *40*(3), 425–446. https://doi.org/10.1002/ajpa.1330400315

Craig, A. (2016). Admissibility from the legal perspective. *Forensic Evidence in Court*, 20–38. https://doi.org/10.1002/9781119054443.ch2

Cunha, E., & Ubelaker, D. H. (2020). Evaluation of ancestry from human skeletal remains: A concise review. *Forensic Sciences Research*, *5*(2), 89–97. https://doi.org/10.1080/20961790.2019.1697060

Cwik, C. H. (1999). Guarding the gate: Expert evidence admissibility. *Litigation*, *25*(4), 6–12, 66–67.

Dasgupta, P. (2011). Some theoretical considerations. In A. Smedley (Ed.), *Race in North America: Origin and Evolution of a Worldview* (4th ed., Vol. 4, pp. 11–34). https://doi.org/10.1016/0014-2921(89)90142-6

Daubert v. Merrell Dow Pharmaceuticals, Inc. 509 U.S. 579, 1993

Densmore, F. (1979). *Chippewa Customs*. Saint Paul, MN: Minnesota Historical Society.

DiGangi, E. A., & Hefner, J. (2013). Ancestry estimation. In *Research Methods in Human Skeletal Biology* (pp. 117–149). https://doi.org/10.1016/B978-0-12-385189-5.00005-4

Dinkele, E. (2018). *Ancestral Variation in Mid-Craniofacial Morphology in a South African Sample*. University of Cape Town.

Donlon, D. (1994). Aboriginal skeletal collections and research in physical anthropology: An historical perspective. *Australian Archaeology*, *39*(Dec), 73–82.

Drost, E. (2011). Validity and reliability in social science research. *Education Research and Perspectives*, *38*(1), 105–123.

Dudzik, B., & Kolatorowicz, A. (2016). Craniometric data analysis and estimation of biodistance. *Biological Distance Analysis: Forensic and Bioarchaeological Perspectives*, 35–60. https://doi.org/10.1016/B978-0-12-801966-5.00003-2

Duforet-Frebourg, N., & Blum, M. G. B. (2014). Nonstationary patterns of isolation-by-distance: Inferring measures of local genetic differentiation with bayesian kriging. *Evolution*, *68*(4), 1110–1123. https://doi.org/10.1111/evo.12342

Dunn, R. R., Spiros, M. C., Kamnikar, K. R., Plemons, A. M., & Hefner, J. (2020). Ancestry estimation in forensic anthropology: A review. *WIREs Forensic Science*, *2*(4), 1–26. https://doi.org/10.1002/wfs2.1369

Duray, S. M., Morter, H. B., & Smith, F. J. (1999). Morphological variation in cervical spinous processes: Potential applications in the forensic identification of race from the skeleton. *Journal of Forensic Sciences*, *44*(5), 12020J. https://doi.org/10.1520/jfs12020j

Eckert, W., & Wright, R. (1997). Scientific evidence in court. In W Eckert (Ed.), *Introduction to forensic sciences* (2nd ed., pp. 69–80). CRC Press.

Edgar, H. (2005). Prediction of race using characteristics of dental morphology. *Journal of Forensic Sciences*, *50*(2), 1–5. https://doi.org/10.1520/jfs2004261

Edgar, H. (2009). Biohistorical approaches to race in the United States: Biological distances among African Americans, European Americans, and their ancestors. *American Journal of Physical Anthropology*, *139*(1), 58–67. https://doi.org/10.1002/ajpa.20961

Edgar, H. (2013). Estimation of ancestry using dental morphological characteristics. *Journal of Forensic Sciences*, *58*(SUPPL. 1), 3–8. https://doi.org/10.1111/j.1556-4029.2012.02295.x

Edgar, H. (2014). Dental morphological estimation of ancestry in forensic contexts. *Biological Affinity in Forensic Identification of Human Skeletal Remains: Beyond Black and White*, 191–207. https://doi.org/10.1201/b17832

Edgar, H. (2017). *Dental Morphology for Anthropology: An Illustrated Manual*. New York, NY: Routledge.

Edgar, H., & Hunley, K. L. (2009). Race reconciled?: How biological anthropologists view human variation. *American Journal of Physical Anthropology*, *139*(1), 1–4. https://doi.org/10.1002/ajpa.20995

Edgar, H., & Ousley, S. (2016). Dominance in dental morphological traits: Implications for biological distance studies. *Biological Distance Analysis: Forensic and Bioarchaeological Perspectives*, 317–332. https://doi.org/10.1016/B978-0-12-801966-5.00017-2

Edge, M. D., & Rosenberg, N. A. (2015). Implications of the apportionment of human genetic diversity for the apportionment of human phenotypic diversity. *Studies in History and Philosophy of Biological and Biomedical Sciences*, *52*, 32–45. https://doi.org/10.1016/j.shpsc.2014.12.005

Elliott, M., & Collard, M. (2009). Fordisc and the determination of ancestry from cranial measurements. *Biology Letters*, *5*(6), 849–852. https://doi.org/10.1098/rsbl.2009.0462

Fabian, A. (2010). The skull collectors: Race, science, and America's unburied dead. In *Nuevos sistemas de comunicación e información*. Chicago: University of Chicago Press.

Falco, M. (2016). Unbiased opinion: The objective expert witness in Canada. *Litigation*, *42*(3), 16–17.

Ferguson, J., Anyon, R., & Ladd, E. (1996). Repatriation at the Pueblo of Zuni: Diverse solutions to complex problems. *American Indian Quarterly*, *20*(2), 251–273.

Fielding, A. (Ed.). (2007a). Classification algorithms 1. In *Cluster and Classification Techniques for the Biosciences* (pp. 97–136). Cambridge University Press.

Fielding, A. (Ed.). (2007b). Other classification methods. In *Cluster and Classification*

Techniques for the Biosciences (pp. 137–178).

Fisher, T., & Gill, G. (1990). Application of the Giles and Elliot discriminant function formulae to a cranial sample of Northwestern Plains Indians. In George Gill & S. Rhine (Eds.), *Skeletal Attribution of Race: Methods for Forensic Anthropology.* (pp. 59–64). Albuquerque, NM: Maxwell Museum of Anthropology.

Fleischman, J., Pierce, M., & Crowder, C. (2019). Transparency in forensic anthropology through the implementation of quality assurance practices. In L. Fulginiti, K. Hartnett-McCann, & A. Galloway (Eds.), *Forensic Anthropology and the United States Judicial System* (1st ed., pp. 71–88). Chichester, UK: John Wiley & Sons, Ltd.

Fleiss, J., & Cohen, J. (1973). The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Education and Psychological Measurement*, *33*, 613–619.

Forensic Anthropology Databank. (n.d.). Retrieved September 27, 2022, from Forensic Anthropology Centre website: https://fac.utk.edu/background/

F.R.E. (2022). *Excluding relevant evidence for prejudice, confusion, waste of time, or other reasons, 403*

F.R.E. (2022). *General admissibility of relevant evidence, 402*

F.R.E. (2022).*Testimony by expert witnesses, 702*

Geller, P. L., & Stojanowksi, C. M. (2017). The vanishing Black Indian: Revisiting craniometry and historic collections. *American Journal of Physical Anthropology*, *162*(2), 267–284. https://doi.org/10.1002/ajpa.23115

General Electric Co. v. Joiner. (1997). 522 U.S. 136.

Giles, E., & Elliott, O. (1962). Race identification from cranial measurements. *Journal of Forensic Sciences*, *7*(2), 147–157.

Gill, G. (1998). Criteria in the skeletal attributes of race. In K. Reichs (Ed.), *Forensic Osteology* (pp. 293–318). Springfield, IL: Thomas Publishing.

Gill, G., & Gilbert, B. (1990). Race identification from the midfacial skeleton: American blacks and whites. In George Gill & S. Rhine (Eds.), *Skeletal Attribution of Race: Methods for Forensic Anthropology.* (pp. 47–54). Albuquerque, NM: Maxwell Museum of Anthropology.

Gill, G., & Rhine, S. (Eds.). (1990). *Skeletal Attribution of Race*. New Mexico: Maxwell Museum of Anthropology.

Go, M. C., & Hefner, J. (2020). Morphoscopic ancestry estimates in Filipino crania using multivariate probit regression models. *American Journal of Physical Anthropology*, (January), 1–16. https://doi.org/10.1002/ajpa.24008

Goodman, A. H. (1997). Bred in the bone? *The Sciences*, *37*(2), 20–25. https://doi.org/10.1002/j.2326-1951.1997.tb03296.x

Graves, J. L. (2010). Biological v. social definitions of race: Implications for modern biomedical research. *Review of Black Political Economy*, *37*(1), 43–60. https://doi.org/10.1007/s12114-009-9053-3

Gravlee, C. C. (2009). How race becomes biology: Embodiment of social inequality. *American Journal of Physical Anthropology*, *139*(1), 47–57. https://doi.org/10.1002/ajpa.20983

Grivas, C. R., & Komar, D. A. (2008). Kumho, Daubert, and the nature of scientific inquiry: Implications for forensic anthropology. *Journal of Forensic Sciences*, *53*(4), 771–776. https://doi.org/10.1111/j.1556-4029.2008.00771.x

Grüneberg, H. (1952). Genetical studies on the skeleton of the mouse. *Journal of Genetics*, *51*(1), 95–114. https://doi.org/10.1007/BF02986708

Hallgren, K. A. (2012). Computing inter-rater reliability for observational data: An overview and tutorial. *Tutorials in Quantitative Methods for Psychology*, *8*(1), 23–34. https://doi.org/10.20982/tqmp.08.1.p023

Halperin, E. C. (2007). The poor, the black, and the marginalized as the source of cadavers in United States anatomical education. *Clinical Anatomy*, *20*(5), 489–495. https://doi.org/10.1002/ca.20445

Hanihara, T., & Ishida, H. (2005). Metric dental variation of major human populations. *American Journal of Physical Anthropology*, *128*(2), 287–298. https://doi.org/10.1002/ajpa.20080

Harris, E., & Foster, C. (2015). Size matters discrimination between American Blacks and Whites, males and females, using tooth crown dimensions. In G. Berg & S. Ta'ala (Eds.), *Biological Affinity in Forensic Identification of Human Skeletal Remains: Beyond Black and White* (pp. 209–238). Boca Raton: CRC Press.

Hauser, G., & De Stefano, G. (1989). *Epigenetic Variants of the Human Skull*. Stuttgart: E. Schweizerbart'sche Verlagsbuchhan.

Hefner, J. (2003). *Assessing Nonmetric Cranial Traits currently used in forensic determination of Ancestry*. University of Florida.

Hefner, J. (2009). Cranial nonmetric variation and estimating ancestry. *Journal of Forensic Sciences*, *54*(5), 985–995. https://doi.org/10.1111/j.1556-4029.2009.01118.x

Hefner, J. (2014). Cranial morphoscopic traits and the assessment of American Black, American White, and Hispanic ancestry. *Biological Affinity in Forensic Identification of Human Skeletal Remains: Beyond Black and White*, 27–41. https://doi.org/10.1201/b17832

Hefner, J. (2016). Biological distance analysis, cranial morphoscopic traits, and ancestry assessment in forensic anthropology. In M. Pilloud & J. Hefner (Eds.), *Biological Distance Analysis: Forensic and Bioarchaeological Perspectives* (pp. 301–315). https://doi.org/10.1016/B978-0-12-801966-5.00016-0

Hefner, J. (2018). The macromorphoscopic databank. *American Journal of Physical Anthropology*, *166*(4), 994–1004. https://doi.org/10.1002/ajpa.23492

Hefner, J. (2019). *OSSA nonmetric ancestry*. Retrieved from http://macromorphoscopic.com/links/

Hefner, J. (2020). *MaMD Analytical*. Retrieved from https://github.com/rer145/mamd-analytical

Hefner, J., & Byrnes, J. F. (2019). Globalization, transnationalism, and the analytical feasibility of ancestry estimation. In *Case Studies in Forensic Anthropology: Bonified Skeletons*. Boca Raton: CRC Press.

Hefner, J., & Linde, K. (2018). *Atlas of Human Cranial Macromorphoscopic Traits*. San Diego:

Academic Press.

Hefner, J., & Ousley, S. D. (2014). Statistical classification methods for estimating ancestry using morphoscopic traits. *Journal of Forensic Sciences*, *59*(4), 883–890. https://doi.org/10.1111/1556-4029.12421

Hefner, J., Ousley, S., & Dirkmaat, D. (2012). Morphoscopic traits and the assessment of ancestry. In D. Dirkmaat (Ed.), *A Companion to Forensic Anthropology*. John Wiley & Sons, Inc.

Hefner, J., Pilloud, M., Black, C. J., & Anderson, B. E. (2015). Morphoscopic trait expression in Hispanic populations. *Journal of Forensic Sciences*, *60*(5), 1135–1139. https://doi.org/10.1111/1556-4029.12826

Hefner, J., Pilloud, M., Buikstra, J., & Vogelsberg, C. C. M. (2016). A brief history of biological distance analysis. In *Biological Distance Analysis: Forensic and Bioarchaeological Perspectives* (pp. 3–22). https://doi.org/10.1016/B978-0-12-801966-5.00001-9

Hefner, J., Plemons, A., Kamnikar, K. R., Ousley, S. D., & Linde, K. (n.d.). MMS v1.61: User Manual v. 1.0. Retrieved from http://macromorphoscopic.com/mms-software/

Hermann, N. P. (n.d.). *Naive Bayesian Classifier*. Retrieved from http://macromorphoscopic.com/links/

Hillson, S., FitzGerald, C., & Flinn, H. (2005). Alternative dental measurements: Proposals and relationships with other measurements. *American Journal of Physical Anthropology*, *126*(4), 413–426. https://doi.org/10.1002/ajpa.10430

Holland, T., & Crowder, C. (2019). "Somewhere in this twilight": The circumstances leading to the National Academy of Sciences' report. In L. Fulginiti, K. Hartnett-McCann, & A. Galloway (Eds.), *Forensic Anthropology and the United States Judicial System* (1st ed., pp. 19–40). Chichester, UK: John Wiley & Sons, Ltd.

Holobinko, A. (2012). Forensic human identification in the United States and Canada: A review of the law, admissible techniques, and the legal implications of their application in forensic cases. *Forensic Science International*, *222*(1–3), 394.e1-394.e13. https://doi.org/10.1016/j.forsciint.2012.06.001

Hopkins, A., & McQueen, K. J. (2022). Filled/non-filled pairs: An empirical challenge to the integrated information theory of consciousness. *Consciousness and Cognition*, *97*, 1–48. https://doi.org/10.1016/j.concog.2021.103245

Howell, W. (1996). Notes and comments. *American Journal of Physical Anthropology*, *10*(1), 441–442. https://doi.org/10.1093/past/22.1.93

Howells, W. (1973). Cranial variation in man. *Papers of the Peabody Museum of Archaeology and Ethnology*, *67*, 1–259.

Howells, W. (1989). Skull shapes and the map: Craniometric analyses in the dispersion of modern Homo. *Papers of the Peabody Museum of Archaeology and Ethnology*, *79*, 1–189.

Hughes, C. E., Juarez, C. A., Hughes, T. L., Galloway, A., Fowler, G., & Chacon, S. (2011). A simulation for exploring the effects of the "Trait List" method's subjectivity on consistency and accuracy of ancestry estimations. *Journal of Forensic Sciences*, *56*(5), 1094–1106. https://doi.org/10.1111/j.1556-4029.2011.01875.x

Humphrey, D. C. (1973). Dissection and discrimination: The social origins of cadavers in America 1760-1915. *Bulletin of the New York Academy of Medicine*, *49*(9), 819.

Hunley, K. L., Cabana, G. S., & Long, J. C. (2016). The apportionment of human diversity revisited. *American Journal of Physical Anthropology*, *160*(4), 561–569. https://doi.org/10.1002/ajpa.22899

Hunt, C. D. L., & Neudorf, L. (2016). The expert witness's duty of impartiality in Canada. *The International Journal of Evidence & Proof*, *20*(1), 72–77. https://doi.org/10.1177/1365712715613479

Hurst, C. V. (2012). Morphoscopic trait expressions used to identify Southwest Hispanics. *Journal of Forensic Sciences*, *57*(4), 859–865. https://doi.org/10.1111/j.1556-4029.2012.02080.x

Irish, J. D. (2014). Dental nonmetric variation around the world: Using key traits in populations to estimate ancestry in individuals. *Biological Affinity in Forensic Identification of Human Skeletal Remains: Beyond Black and White*, 165–190. https://doi.org/10.1201/b17832

Jantz, R. L., & Ousley, S. D. (2005). *FORDISC 3.1 Personal Computer Forensic Discriminant Functions*. Knoxville, TN: Knoxville: University of Tennessee.

Jantz, R. L., & Ousley, S. D. (2012). Introduction to Fordisc 3. In M. Tersigni-Tarrant & N. R. Shirley (Eds.), *Forensic Anthropology: An Introduction* (pp. 253–269). Boca Raton: CRC Press.

Johnston, B. (1987). *Ojibway Ceremonies*. Toronto: McClelland and Stewart.

Kaestle, F. A., & Horsburgh, K. A. (2002). Ancient DNA in anthropology: Methods, applications, and ethics. *Yearbook of Physical Anthropology*, *45*, 92–130. https://doi.org/10.1002/ajpa.10179

Kamnikar, K. R., Plemons, A. M., & Hefner, J. (2018). Intraobserver error in macromorphoscopic trait data. *Journal of Forensic Sciences*, *63*(2), 361–370. https://doi.org/10.1111/1556-4029.13564

Kelliher (Village of) v. Smith, S.C.R. 672, Retrieved from http://csc.lexum.org, 1931.

Kennedy, K. (2003). Trials in court: the forensic anthropologist takes the stand. In D. Steadman (Ed.), *Hard Evidence: Case Studies in Forensic Anthropology* (pp. 77–86). https://doi.org/10.5840/raven20111838

Kenyhercz, M., Klales, A. R., & Kenyhercz, W. E. (2014). Molar size and shape in the estimation of biological ancestry: A comparison of relative cusp location using geometric morphometrics and interlandmark distances. *American Journal of Physical Anthropology*, *153*(2), 269–279. https://doi.org/10.1002/ajpa.22429

Kenyhercz, M., Klales, A. R., Rainwater, C. W., & Fredette, S. M. (2017). The optimized summed scored attributes method for the classification of U.S. Blacks and Whites: A validation study. *Journal of Forensic Sciences*, *62*(1), 174–180. https://doi.org/10.1111/1556-4029.13243

Kenyhercz, M., & Passalacqua, N. V. (2016). Missing data imputation methods and their performance with biodistance analyses. *Biological Distance Analysis: Forensic and Bioarchaeological Perspectives*, 181–194. https://doi.org/10.1016/B978-0-12-801966-

5.00009-3

Kilroy, G. S., Tallman, S. D., & DiGangi, E. A. (2020). Secular change in morphological cranial and mandibular trait frequencies in European Americans born 1824–1987. *American Journal of Physical Anthropology*, *173*(3), 589–605. https://doi.org/10.1002/ajpa.24115

Kindschuh, S. C., Dupras, T. L., & Cowgill, L. W. (2012). Exploring ancestral variation of the hyoid. *Journal of Forensic Sciences*, *57*(1), 41–46. https://doi.org/10.1111/j.1556-4029.2011.01962.x

Klales, A. R., & Kenyhercz, M. (2015). Morphological assessment of ancestry using cranial macromorphoscopics. *Journal of Forensic Sciences*, *60*(1), 13–20. https://doi.org/10.1111/1556-4029.12563

Kluth, R., & Munnell, K. (1997). The integration of tradition and scientific knowledge on the Leech Lake reservation. In N. Swidler, K. Dongoske, R. Anyon, & A. Downer (Eds.), *Native Americans and Archaeologists: Stepping Stones to Common Ground* (pp. 112–119). Walnut Creek: AltaMira Press in Cooperation with the Society for American Archaeology.

Koehler, E. (2007). Repatriation of cultural objects to Indigenous peoples: A comparative analysis of U.S. and Canadian law. *International Lawyer*, *41*(1), 1–16.

Konigsberg, L. (1990). Temporal aspects of biological distance: Serial correlation and trend in a pre-historic skeletal lineage. *American Journal of Physical Anthropology*, *83*, 45–52.

Konigsberg, L. (2006). A post-Neumann history of biological and genetic distance studies in bioarchaeology. In J. Buikstra & L. Beck (Eds.), *Bioarchaeology: The Contextual Analysis of Human Remains* (pp. 263–280). Amsterdam: Academic Press.

Kranioti, E. F., García-Donas, J. G., Can, I. O., & Ekizoglu, O. (2018). Ancestry estimation of three Mediterranean populations based on cranial metrics. *Forensic Science International*, *286*, 265.e1-265.e8. https://doi.org/10.1016/j.forsciint.2018.02.014

Kumho Tire Company, Ltd. v. Carmichael, 526 US 137. (1999)

L'Abbé, E. N., Van Rooyen, C., Nawrocki, S. P., & Becker, P. J. (2011). An evaluation of non-metric cranial traits used to estimate ancestry in a South African sample. *Forensic Science International*, *209*(1–3), 195.e1-195.e7. https://doi.org/10.1016/j.forsciint.2011.04.002

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*(1), 159. https://doi.org/10.2307/2529310

Lane, R., & Sublett, A. (1972). Osteology of social organization: Residence pattern. *American Antiquity*, *37*(2), 186–201.

Langley, N., Jantz, L. M., McNulty, S., Maijanen, H., Ousley, S. D., & Jantz, R. L. (2018). Data for validation of osteometric methods in forensic anthropology. *Data in Brief*, *19*, 21–28. https://doi.org/10.1016/j.dib.2018.04.148

Langley, N., Jantz, L. M., Ousley, S. D., Jantz, R. L., & Milner, G. R. (2016a). *Data Collection Procedures for Forensic Skeletal 2.0* (3rd ed.). Knoxville, TN: The University of Tennessee.

Langley, N., Jantz, L. M., Ousley, S. D., Jantz, R., & Milner, G. (2016b). *Data Collection Procedures for Forensic Skeletal 2.0*. Knoxville, TN: Forensic Anthropology Center.

Laughlin, W., & Jorgensen, J.. (1956). Isolate variation in greenlandic Eskimo crania. *Acta Genetica et Statistica Medica*, *6*(1), 3–12.

Lesciotto, K. M. (2015). The impact of Daubert on the admissibility of forensic anthropology expert testimony. *Journal of Forensic Sciences*, *60*(3), 549–555. https://doi.org/10.1111/1556-4029.12740

Lewis, C. J., & Garvin, H. M. (2016). Reliability of the Walker cranial nonmetric method and implications for sex estimation. *Journal of Forensic Sciences*, *61*(3), 743–751. https://doi.org/10.1111/1556-4029.13013

Lewontin, R. C. (1972). The apportionment of human diversity. In T. Dobzhansky, M. Hecht, & W. Steere (Eds.), *Evolutionary Biology* (pp. 381–398). https://doi.org/doi-org.uml.idm.oclc.org/10.1007/978-1-4684-9063-3_14

Liebenberg, L., & Krüger, G. C. (2020). Standardization and quality assurance in skeletal landmark placement and osteometry. *Forensic Science International*, *308*. https://doi.org/10.1016/j.forsciint.2020.110168

Livingstone, F., & Dobzhansky, T. (1962). On the non-existence of human races. *Current*, *3*(3), 279–281. Retrieved from http://www.jstor.org/stable/2739576

Lund, A., & Lund, M. (2020). *Laerd Statistics*. Retrieved from https://statistics.laerd.com/

Macchiarelli, R., Salvadei, L., & Bondioli, L. (1995). Odontometric variation and biological relationships among Italic (Latins, Samnites, Paeligni, Picenes). *Aspects of Dental Biology: Palaeontology, Anthropology, and Evolution*, 419–436.

Maddux, S., Sporleder, A., & Burns, C. (2015). Geographic variation in zygomaxillary suture morphology and its use in ancestry estimation. *Journal of Forensic Sciences*, *60*(4), 966–973. https://doi.org/https://doi.org/10.1111/1556-4029.12774

Maier, C. A., Zhang, K., Manhein, M. H., & Li, X. (2015). Palate shape and depth: A shape-matching and machine learning method for estimating encestry from human skeletal remains. *Journal of Forensic Sciences*, *60*(5), 1129–1134. https://doi.org/10.1111/1556-4029.12812

Manthey, L., & Jantz, R. (2020). Fordisc: Anthropological software for estimation of ancestry, sex, time period, and stature. In Z. Obertova, A. Stewart, & C. Cattaneo (Eds.), *Statistics and Probability in Forensic Anthropology* (pp. 275–287). San Diego: Elsevier Inc.

Marino, E. A. (1997). A pilot study using the first cervical vertebra as an indicator of race. *Journal of Forensic Sciences*, *42*(6), 14271J. https://doi.org/10.1520/jfs14271j

McDowell, J. L., Kenyhercz, M., & L'Abbé, E. N. (2015). An evaluation of nasal bone and aperture shape among three South African populations. *Forensic Science International*, *252*, 189.e1-189.e7. https://doi.org/10.1016/j.forsciint.2015.04.016

McDowell, J. L., L'Abbé, E. N., & Kenyhercz, M. (2012). Nasal aperture shape evaluation between black and white South Africans. *Forensic Science International*, *222*(1–3), 397.e1-397.e6. https://doi.org/10.1016/j.forsciint.2012.06.007

McKeown, A., & Jantz, R. (2005). Comparison of coordinate and craniometric data for biological distance studies. In D. E. Slice (Ed.), *Modern Morphometrics in Modern Physical Anthropology*. New York: Kluwer Academic/ Plenum Publishers.

McNiven, I., & Russell, L. (2005). *Appropriated Pasts: Indigenous Peoples and the Colonial Culture of Archaeology*. Lanham: AltaMira Press.

Meirmans, P. G. (2012). The trouble with isolation by distance. *Molecular Ecology*, *21*(12), 2839–2846. https://doi.org/10.1111/j.1365-294X.2012.05578.x

Mincer, S. (2015). *Skeletons in the Closet: The History and Ethics of Human Skeletal Collections in the United States with an Example from the Vassar College Collection Skeletons in the Closet*. Vassar College.

Moffit, M. (2017). *The Evaluation of Macromorphoscopic Traits To Determine Secular Change Among a Sample of African and American Blacks*. Michigan State University.

Molnar, S. (2002). *Human Variation: Races, Types, and Ethnic Groups* (5th ed.). New Jersey: Pearson Education.

Monsalve, T., & Hefner, J. (2016). Macromorphoscopic trait expression in a cranial sample from Medellín, Colombia. *Forensic Science International*, *266*, 574.e1-574.e8. https://doi.org/10.1016/j.forsciint.2016.07.014

Moorrees, C. F. A., & Reed, R. B. (1964). Correlations among crown diameters of human teeth. *Archives of Oral Biology*, *9*(6), 685–697. https://doi.org/10.1016/0003-9969(64)90080-9

Murphy, C. (2011). What can an osteological investigation reveal about medical education in eighteenth-century Dublin? *Archaeology Ireland*, *25*(3), 30–34.

Murphy, R. E., & Garvin, H. M. (2018). A morphometric outline analysis of ancestry and sex differences in cranial shape. *Journal of Forensic Sciences*, *63*(4), 1001–1009. https://doi.org/10.1111/1556-4029.13699

Navega, D., Coelho, C., Vicente, R., Ferreira, M. T., Wasterlain, S., & Cunha, E. (2015). AncesTrees: Ancestry estimation with randomized decision trees. *International Journal of Legal Medicine*, *129*(5), 1145–1153. https://doi.org/10.1007/s00414-014-1050-9

Newman, M. T. (1963). Geographic and Microgeographic Races. *Current Anthropology*, *4*(2), 189–207. https://doi.org/10.1086/200360

Obertova, Z., & Stewart, A. (2020). Probability distributions, hypothesis testing. In Z. Obertova, A. Stewart, & C. Cattaneo (Eds.), *Statistics and Probability in Forensic Anthropology* (pp. 73–86). San Diego: Elsevier Inc.

OSAC. (2020, March). *OSAC Research and Development Needs*. National Institute of Standards and Technology. Retrieved from: https://www.nist.gov/topics/organization-scientific-area-committees-forensic-science/research-and-development-needs*Osteoware v. 2.4.037*. (2020). Retrieved from https://naturalhistory.si.edu/research/anthropology/programs/repatriation-office/osteoware

Ousley, S. (2016). Forensic classification and biodistance in the 21st century: The rise of learning machines. In M. Pilloud & J. Hefner (Eds.), *Biological Distance Analysis: Forensic and Bioarchaeological Perspectives* (pp. 197–212). London: Academic Press.

Ousley, S., & Jantz, R. (2012). Fordisc 3 and statistical methods for estimating sex and ancestry. In D. C. Dirkmaat (Ed.), *A Companion to Forensic Anthropology* (pp. 311–329). Malden: Wiley-Blackwell.

Ousley, S., Jantz, R., & Freid, D. (2009). Understanding race and human variation: Why forensic

anthropologists are good at identifying race. *American Journal of Physical Anthropology*, *139*(1), 68–76. https://doi.org/10.1002/ajpa.21006

Ousley, S., Jantz, R. L., & Hefner, J. (2018). From Blumenbach to Howells: The slow, painful emergence of theory through forensic race estimation. In *Forensic Anthropology: Theoretical Framework and Scientific Basis* (pp. 65–97). https://doi.org/10.1002/9781119226529.ch5

Page, M., Taylor, J., & Blenkin, M. (2011a). Forensic identification science evidence since Daubert: Part I-A quantitative analysis of the exclusion of forensic identification science evidence. *Journal of Forensic Sciences*, *56*(5), 1180–1184. https://doi.org/10.1111/j.1556-4029.2011.01777.x

Page, M., Taylor, J., & Blenkin, M. (2011b). Forensic identification science evidence since Daubert: Part II-judicial reasoning in decisions to exclude forensic identification evidence on grounds of reliability. *Journal of Forensic Sciences*, *56*(4), 913–917. https://doi.org/10.1111/j.1556-4029.2011.01776.x

Phenice, T. (1969). A newly developed visual method of sexing the os pubis. *American Journal of Physical Anthropology*, *30*(2), 297–301.

Pietrusewsky, M. (2008). Metric analysis of skeletal remains: Methods and applications. In M. A. Katzenberg & R. S. Saunders (Eds.), *Biological Anthropology of the Human Skeleton* (2nd ed., pp. 487–532). https://doi.org/10.1002/9780470245842.ch16

Pilloud, M., Adams, D. M., & Hefner, J. (2019). Observer error and its impact on ancestry estimation using dental morphology. *International Journal of Legal Medicine*, *133*(3), 949–962. https://doi.org/10.1007/s00414-018-1985-3

Pilloud, M., & Hefner, J. (Eds.). (2016). *Biological Distance Analysis: Forensic and Bioarchaeological Perspectives*. London: Academic Press.

Pilloud, M., Hefner, J., Hanihara, T., & Hayashi, A. (2014). The use of tooth crown measurements in the assessment of ancestry. *Journal of Forensic Sciences*, *59*(6), 1493–1501. https://doi.org/10.1111/1556-4029.12540

Pilloud, M., & Kenyhercz, M. (2016). Dental metrics in biodistance analysis. *Biological Distance Analysis: Forensic and Bioarchaeological Perspectives*, 135–155. https://doi.org/10.1016/B978-0-12-801966-5.00007-X

Pilloud, M., Maier, C., Scott, R., Hefner, J., & Scott, G. R. (2018). Advances in cranial macromorphoscopic trait and dental morphology analysis for ancestry estimation. In K. Latham, E. J. Bartelink, & M. Finnegan (Eds.), *New Perspectives in Forensic Human Skeletal Identification* (pp. 23–34). https://doi.org/10.1016/B978-0-12-805429-1.00004-1

Pink, C. (2016). Forensic ancestry assessment using cranial nonmetric traits traditionally applied to biological distance studies. In M. Pilloud & J. Hefner (Eds.), *Biological Distance Analysis: Forensic and Bioarchaeological Perspectives* (pp. 213–230). London: Academic Press.

Pink, C., Maier, C., Pilloud, M., & Hefner, J. (2016). Cranial nonmetric and morphoscopic data sets. In M. Pilloud & J. Hefner (Eds.), *Biological Distance Analysis: Forensic and Bioarchaeological Perspectives* (pp. 91–102). London: Academic Press.

Plemons, A. M., & Hefner, J. (2016). Ancestry estimation using macromorphoscopic traits.

Academic Forensic Pathology, *6*(3), 400–412. https://doi.org/10.23907/2016.041

Primeau, C., Arge, S. O., Boyer, C., & Lynnerup, N. (2015). A test of inter- and intra-observer error for an atlas method of combined histological data for the evaluation of enamel hypoplasia. *Journal of Archaeological Science: Reports*, *2*, 384–388. https://doi.org/10.1016/j.jasrep.2015.03.007

Redfern, R. C., Gröcke, D. R., Millard, A. R., Ridgeway, V., Johnson, L., & Hefner, J. (2016). Going south of the river: A multidisciplinary analysis of ancestry, mobility and diet in a population from Roman Southwark, London. *Journal of Archaeological Science*, *74*, 11–22. https://doi.org/10.1016/j.jas.2016.07.016

Regina v. Mohan, 2 S.C.R. 9 File No. 23063, 1994

Relethford, J. (1994). Craniometric variation among modern human populations. *American Journal of Physical Anthropology*, *95*, 53–62.

Relethford, J. (2002). Apportionment of global human genetic diversity based on craniometrics and skin color. *American Journal of Physical Anthropology*, *118*(4), 393–398. https://doi.org/10.1002/ajpa.10079

Relethford, J. (2004). Global patterns of isolation by distance based on genetic and morphological data. *Human Biology*, *76*(4), 499–513.

Relethford, J. (2016). Biological distances and population genetics in bioarchaeology. In M. Pilloud & J. Hefner (Eds.), *Biological Distance Analysis: Forensic and Bioarchaeological Perspectives* (pp. 23–33). London: Academic Press.

Rhine, S. (1990). Non-metric skull racing. In G. Gill & S. Rhine (Eds.), *Skeletal Attribution of Race: Methods for Forensic Anthropology.* (pp. 9–20). Albuquerque, NM: Maxwell Museum of Anthropological Papers.

Ricaut, F. X., Auriol, V., Von Cramon-Taubadel, N., Keyser, C., Murail, P., Ludes, B., & Crubézy, E. (2010). Comparison between morphological and genetic data to estimate biological relationship: The case of the Egyin Gol necropolis (Mongolia). *American Journal of Physical Anthropology*, *143*(3), 355–364. https://doi.org/10.1002/ajpa.21322

Rogers, T. L., & Allard, T. T. (2004). Expert testimony and positive identification of human remains through cranial suture patterns. *Journal of Forensic Sciences*, *49*(2), 1–5. https://doi.org/10.1520/jfs2003095

Roque, R. (2018). Authorised histories: Human remains and the economies of credibility in the science of race. *Kronos*, *44*(1), 69–85. https://doi.org/10.17159/2309-9585/2018/v44a5

Ross, A. H., & Williams, S. (2008). Testing repeatability and error of coordinate landmark data acquired from crania. *Journal of Forensic Sciences*, *53*(4), 782–785. https://doi.org/10.1111/j.1556-4029.2008.00751.x

Rousset, F. (1997). Genetic differentiation and estimation of gene flow from F-statistics under isolation by distance. *Genetics*, *145*, 1219–1228. https://doi.org/10.1177/000992287301200933

Rubin, K. M., & DeLeon, V. B. (2017). Ancestral variation in orbital rim shape: A three-dimensional pilot study. *Journal of Forensic Sciences*, *62*(6), 1575–1581. https://doi.org/10.1111/1556-4029.13493

Sauer, N. J. (1992). Forensic anthropology and the concept of race: If races don't exist, why are forensic anthropologists so good at identifying them? *Social Science & Medicine*, *34*(2), 107–111. https://doi.org/10.1016/0277-9536(92)90086-6

Sauer, N. J., Wankmiller, J., & Hefner, J. (2016). Assessment of ancestry and the concept of race. In S. Blau & D. H. Ubelaker (Eds.), *Handbook of Forensic Anthropology and Archaeology* (2nd ed., pp. 243–260). New York: Routledge.

Scheuer, L., & Black, S. (2004). Juvenile skeletal remains: Provenance, identification and interpretation. In *The Juvenile Skeleton*. London: Academic Press.

Scientific Working Group for Forensic Anthropology (SWGANTH). (2011). *Laboratory Management and Quality Assurance*. Retrieved from https://www.nist.gov/organization-scientific-area-committees-forensic-science/forensic-anthropology-subcommittee

Scientific Working Group for Forensic Anthropology (SWGANTH). (2012). *Statistical Methods*. Retrieved from https://www.nist.gov/organization-scientific-area-committees-forensic-science/forensic-anthropology-subcommittee

Scientific Working Group for Forensic Anthropology (SWGANTH). (2013). *Ancestry Assessment*. Retrieved from https://www.nist.gov/organization-scientific-area-committees-forensic-science/forensic-anthropology-subcommittee

Scott, G. R., Pilloud, M., Navega, D., D'Oliveira, J., Cunha, E., & Irish, J. D. (2018). rASUDAS: A new web-based application for estimating ancestry from tooth morphology. *Forensic Anthropology*, *1*(1), 18–31.

Scott, G. R., Turner, C. G., Townsend, G., & Martinon-Torres, M. (2018). The anthropology of modern human teeth: Dental morphology and its variation in recent and fossil Homo sapiens. In *The Anthropology of Modern Human Teeth* (2nd ed.). https://doi.org/10.1017/cbo9781316529843

Séré, M., Thévenon, S., Belem, A. M. G., & De Meeûs, T. (2017). Comparison of different genetic distances to test isolation by distance between populations. *Heredity*, *119*(2), 55–63. https://doi.org/10.1038/hdy.2017.26

Sherwood, R., Duren, D., Demerath, E., Czerwinski, S., Siervogel, R., & Towne, B. (2008). Quantitative genetics of modern human cranial variation. *Journal of Human Evolution*, *54*(6), 909–914. https://doi.org/doi:10.1016/j.jhevol.2008.02.009. Quantitative

Shirley, N. R., Fatah, E. E. A., & Mahfouz, M. (2014). Beyond the cranium: Ancestry estimation from the lower limb. *Biological Affinity in Forensic Identification of Human Skeletal Remains: Beyond Black and White*, 133–153. https://doi.org/10.1201/b17832

Sholts, S. B., & Wärmländer, S. K. T. S. (2012). Zygomaticomaxillary suture shape analyzed with digital morphometrics: Reassessing patterns of variation in American Indian and European populations. *Forensic Science International*, *217*(1–3), 234.e1-234.e6. https://doi.org/10.1016/j.forsciint.2011.11.016

Smay, D., & Armelagos, G. (2000). Galileo wept: A critical assessment of the use of race in forensic anthropology. *Transforming Anthropology*, *9*(2), 19–29. https://doi.org/10.1525/tran.2000.9.2.19

Smith, H., Hulsey, B., West, F., & Cabana, G. (2016). Do biological distances reflect genetic distances? A comparison of craniometric and genetic distances and local and global scales.

In M. Pilloud & J. Hefner (Eds.), *Biological Distance Analysis: Forensic and Bioarchaeological Perspectives* (pp. 157–179). London: Academic Press.

Smith, S. L. (1996). Attribution of hand bones to sex and population groups. *Journal of Forensic Sciences*, *41*(3), 469–477. https://doi.org/10.1520/jfs14097j

Smith, S. L. (1997). Attribution of foot bones to sex and population Groups. *Journal of Forensic Sciences*, *42*(2), 186–195. https://doi.org/10.1520/jfs14097j

Spiros, M. (2018). Standardization of postcranial nonmetric traits and their utility in ancestry analysis. *Forensic Anthropology*, *2*(1), 29–44. https://doi.org/10.5744/fa.2018.1031

Spiros, M., & Hefner, J. (2019). *Combo MaMD Analytical*. Retrieved from https://macromorphoscopictraitanalysis.shinyapps.io/combo_mamd/

Spiros, M., & Hefner, J. (2020). Ancestry estimation using cranial and postcranial macromorphoscopic traits. *Journal of Forensic Sciences*, *65*(3), 921–929. https://doi.org/10.1111/1556-4029.14231

Spradley, K. (2014). Metric ancestry estimation from the postcranial skeleton. *Biological Affinity in Forensic Identification of Human Skeletal Remains: Beyond Black and White*, 83–94. https://doi.org/10.1201/b17832

Stubblefield, P. R. (2011). The anatomical diaspora: Evidence of early American anatomical traditions in North Dakota. *Journal of Forensic Sciences*, *56*(5), 1324–1327. https://doi.org/10.1111/j.1556-4029.2011.01738.x

Stull, K. E., Kenyhercz, M., & L'Abbé, E. N. (2014). Ancestry estimation in South Africa using craniometrics and geometric morphometrics. *Forensic Science International*, *245*, 206.e1-206.e7. https://doi.org/10.1016/j.forsciint.2014.10.021

Ta'ala, S. C. (2014). A brief history of the race concept in physical anthropology. *Biological Affinity in Forensic Identification of Human Skeletal Remains: Beyond Black and White*, 1–15. https://doi.org/10.1201/b17832

Tallman, S. D., & Winburn, A. P. (2015). Forensic applicability of femur subtrochanteric shape to ancestry assessment in Thai and White American males. *Journal of Forensic Sciences*, *60*(5), 1283–1289. https://doi.org/10.1111/1556-4029.12775

Tarlow, S. (2006). Archaeological ethics and the people of the past. In C. Scarre & G. Scarre (Eds.), *The Ethics of Archaeology: Philosophical Perspectives on Archaeological Practice* (pp. 199–216). https://doi.org/10.1017/CBO9780511817656.012

TCPS. (2018). Canadian Institutes of Health Research, Natural Sciences and Engineering Research Council of Canada, and Social Sciences and Humanities Research Council, Tri-Council Policy Statement (TCPS): Ethical Conduct for Research Involving Humans, December 2018.

Thomas, M. (2014). Turning subjects into objects and objects into subjects: Collecting human remains on the 1948 Arnhem land expedition. *Circulating Cultures: Exchanges of Australian Indigenous Music, Dance and Media*. https://doi.org/10.22459/cc.12.2014.06

Todd, T. (1929). Entrenched Negro physical features. *Human Biology*, *1*(1), 57–69.

Turner, C. G. (1986). The first Americans; The dental evidence. *National Geographic Research*,

2(1), 37–46.

Turner II, C., Nichol, C., & Scott, R. (1991). Scoring procedures for key morphological traits of the permanent dentition: The Arizona State University dental anthropology system. *Advances in Dental Anthropology*, (JANUARY 1991), 13–31.

Ünlütürk, Ö. (2017). Metric assessment of ancestry from the vertebrae in South Africans. *International Journal of Legal Medicine*, *131*(4), 1123–1131. https://doi.org/10.1007/s00414-016-1483-4

Usher, B. (2002). Reference samples: The first step in linking biology and age in the human skeleton. In R. Hoppa & J. Vaupel (Eds.), *Paleodemography: Age Distributions From Skeletal Samples*. Cambridge: Cambridge University Press.

Velasco, M. C. (2018). Open sepulchers and closed boundaries? Biodistance analysis of cemetery structure and postmarital residence in the late prehispanic Andes. *American Journal of Physical Anthropology*, *166*(4), 906–920. https://doi.org/10.1002/ajpa.23594

Walker, P. L. (2008). Bioarchaeological ethics: A historical perspective on the value of human remains. In *Biological Anthropology of the Human Skeleton* (pp. 1–40). https://doi.org/10.1002/9780470245842.ch1

Wang, K. M. (2016). *A comparison of nonmetric cranial and morphoscopic trait frequencies in Mexican and various Asian populations*. The University of Arizona.

Watson, P. F., & Petrie, A. (2010). Method agreement analysis: A review of correct methodology. *Theriogenology*, *73*(9), 1167–1179. https://doi.org/10.1016/j.theriogenology.2010.01.003

Wescott, D. J. (2005). Population variation in femur subtrochanteric shape. *Journal of Forensic Sciences*, *50*(2), 1–8. https://doi.org/10.1520/jfs2004281

Wescott, D. J., & Jantz, R. (2005). Assessing craniofacial secular change in American Blacks and Whites using geometric morphometry. In D. E. Slice (Ed.), *Modern Morphometrics in Modern Physical Anthropology* (pp. 231–242). New York: Kluwer Academic/ Plenum Publishers.

Wilczak, C., & Dudar, C. (Eds.). (2020). *Osteoware Software Manual Volume 1*. Washington, D.C: Smithsonian Institution.

Wittwer-Backofen, U., Kästner, M., Möller, D., Vohberger, M., Lutz-Bonengel, S., & Speck, D. (2014). Ambiguous provenance? Experience with provenance analysis of human remains from Namibia in the Alexander Ecker collection. *Anthropologischer Anzeiger*, *71*(1–2), 65–86. https://doi.org/10.1127/0003-5548/2014/0382

Wood-Jones, F. (1931a). The non-metrical morphological characters of the skull as criteria for racial diagnosis: Part I-General discussion of the morphological characters employed in racial diagnosis. *Journal of Anatomy*, *65*(2), 179–195.

Wood-Jones, F. (1931b). The non-metrical morphological characters of the skull as criteria for racial diagnosis: Part II- The non-metrical morphological characters of the Hawaiian skull. *Journal of Anatomy*, *65*(Pt 4), 438–43845.

Wood-Jones, F. (1931c). The non-metrical morphological characters of the skull as criteria for racial diagnosis: Part III- The non-metrical morphological characters of the skulls of

prehistoric inhabitants of Guam. *Journal of Genetics*, *65*, 438–445.

Wood-Jones, F. (1933). The non-metrical morphological characters of the skull as criteria for racial diagnosis: Part IV- The non-metrical morphological characters of the Northern Chinese skull. *Journal of Anatomy*, *68*(Pt 1), 96–108.

Wright, J. L., Wasef, S., Heupink, T. H., Westaway, M. C., Rasmussen, S., Pardoe, C., … Lambert, D. M. (2018). Ancient nuclear genomes enable repatriation of Indigenous human remains. *Science Advances*, *4*(12), 1–13. https://doi.org/10.1126/sciadv.aau5064

Wright, S. (1943). Isolation by distance. *Genetics*, *28*, 114–138.

Yucha, J. M., Pokines, J. T., & Bartelink, E. J. (2017). A comparative taphonomic analysis of 24 trophy skulls from modern forensic cases. *Journal of Forensic Sciences*, *62*(5), 1266–1278. https://doi.org/10.1111/1556-4029.13426

Zeman, T., & Benus, R. (2020). Initial assessment: Measurement errors and interrater reliability. In Z. Obertova, A. Stewart, & C. Cattaneo (Eds.), *Statistics and Probability in Forensic Anthropology*. San Diego: Elsevier Inc.

Appendices

Appendix 1: List of metric measurements taken for *Fordisc* ancestry estimation

1. Maximum Cranial Length (g-op, GOL)
2. Nasio-occipital length (NOL)
3. Maximum Cranial Breadth (eu-eu, XCB)
4. Bizygomatic Breadth (zy-zy, ZYB)
5. Basion-Bregma Height (ba-b, BBH)
6. Cranial Base Length (ba-n, BNL)
7. Basion-Prosthion Length (ba-pr, BPL)
8. Maxillo-Alveolar Breadth (ecm-ecm, MAB)
9. Maxillo-Alveolar Length (pr-alv, MAL)
10. Biauricular Breadth (ra-ra, AUB)
11. Nasion-Prosthion Height (n-pr, NPH)
12. Minimum Frontal Breadth (ft-ft, WFB)
13. Upper Facial Breadth (fmt-fmt)
14. Nasal Height (NLH)
15. Nasal Breadth (NLB)
16. Orbital Breadth (d-ec, 0BB)
17. Orbital Height (OBH)
18. Biorbital Breadth (ec-ec, EKB)
19. lnterorbital Breadth (d-d, DKB)
20. Frontal Chord (n-b, FRC)
21. Parietal Chord (b-l, PAC)
22. Occipital Chord (l-o, OCC)
23. Foramen Magnum Length (FOL)
24. Foramen Magnum Breadth (FOB)
25. Mastoid Height (MDH)
26. Biasterionic Breadth (ast-ast, ASB)

27. Bimaxillary breadth (zma-zma, ZMB)
28. Zygoorbitale breadth (zo-zo, ZOB)
29. Chin Height (id-gn)
30. Height of the Mandibular Body
31. Breadth of Mandibular Body
32. Bigonial Breadth (go-go)
33. Bicondylar Breadth (cdl-cdl)
34. Minimum Ramus Breadth
35. Maximum Ramus Height
36. Mandibular Length
37. Mandibular Angle

Appendix 2: Frequency tables

List of Tables

Table 14: Frequency of trait scores for ANS given to U of M individuals and the disagreement between untrained scoring period 1 and trained scoring period 2 for observer 1. Columns are for scores from scoring period 2, and rows are for scores from scoring period 1. The diagonal from top left to bottom right are the number of times the observers agreed on score.

	Scoring period 2			
Scoring period 1	**1**	**2**	**3**	**Total**
1	1	0	0	1
2	0	1	5	6
3	0	0	0	0
Total	1	1	5	7

Table 15: Frequency of trait scores for INA given to U of M individuals and the disagreement between untrained scoring period 1 and trained scoring period 2 for observer 1. Columns are for scores from scoring period 2, and rows are for scores from scoring period 1. The diagonal from top left to bottom right are the number of times the observers agreed on score.

	Scoring period 2					
Scoring period 1	**1**	**2**	**3**	**4**	**5**	**total**
1	0	2	0	0	0	2
2	0	0	0	0	0	0
3	0	0	0	1	0	1
4	0	1	1	0	0	2
5	0	0	0	1	1	2
total	0	3	1	2	1	7

Table 16: Frequency of trait scores for IOB given to U of M individuals and the disagreement between untrained scoring period 1 and trained scoring period 2 for observer 1. Columns are for scores from scoring period 2, and rows are for scores from scoring period 1. The diagonal from top left to bottom right are the number of times the observers agreed on score.

	Scoring period 2			**total**
Scoring period 1	**1**	**2**	**3**	
1	0	0	0	0
2	4	0	0	4
3	1	2	0	3
total	5	2	0	7

Table 17: Frequency of trait scores for MT given to U of M individuals and the disagreement between untrained scoring period 1 and trained scoring period 2 for observer 1. Columns are for scores from scoring period 2, and rows are for scores from scoring period 1. The diagonal from top left to bottom right are the number of times the observers agreed on score.

	Scoring period 2				
Scoring period 1	**0**	**1**	**2**	**3**	**total**
0	2	0	0	0	2
1	0	1	0	0	1
2	0	3	1	0	4
3	0	0	0	0	0
total	2	4	1	0	7

Table 18: Frequency of trait scores for NAW given to U of M individuals and the disagreement between untrained scoring period 1 and trained scoring period 2 for observer 1. Columns are for scores from scoring period 2, and rows are for scores from scoring period 1. The diagonal from top left to bottom right are the number of times the observers agreed on score.

	Scoring period 2			
Scoring period 1	**1**	**2**	**3**	**total**
1	0	0	0	0
2	1	5	0	6
3	0	1	0	1
total	1	6	0	7

Table 19: Frequency of trait scores for NBC given to U of M individuals and the disagreement between untrained scoring period 1 and trained scoring period 2 for observer 1. Columns are for scores from scoring period 2, and rows are for scores from scoring period 1. The diagonal from top left to bottom right are the number of times the observers agreed on score.

	Scoring period 2					**total**
Scoring period 1	0	1	2	3	4	
0	1	0	0	0	0	1
1	1	3	0	0	0	4
2	0	0	0	0	0	0
3	0	0	0	0	0	0
4	0	0	0	1	0	1
total	2	3	0	1	0	6

Table 20: Frequency of trait scores for NO given to U of M individuals and the disagreement between untrained scoring period 1 and trained scoring period 2 for observer 1. Columns are for scores from scoring period 2, and rows are for scores from scoring period 1. The diagonal from top left to bottom right are the number of times the observers agreed on score.

	Scoring period 2		
Scoring period 1	0	1	**total**
0	1	2	3
1	0	2	2
total	1	4	5

Table 21: Frequency of trait scores for PBD given to U of M individuals and the disagreement between untrained scoring period 1 and trained scoring period 2 for observer 1. Columns are for scores from scoring period 2, and rows are for scores from scoring period 1. The diagonal from top left to bottom right are the number of times the observers agreed on score.

	Scoring period 2		
Scoring period 1	**0**	**1**	**total**
0	3	0	3
1	0	3	3
total	3	3	6

Table 22: Frequency of trait scores for SPS given to U of M individuals and the disagreement between untrained scoring period 1 and trained scoring period 2 for observer 1. Columns are for scores from scoring period 2, and rows are for scores from scoring period 1. The diagonal from top left to bottom right are the number of times the observers agreed on score.

	Scoring period 2			
Scoring period 1	0	1	2	**total**
0	2	0	2	4
1	0	1	2	3
2	0	0	0	0
total	2	1	4	7

N (above)

Table 23: **Frequency of trait scores for TPS given to U of M individuals and the disagreement between untrained scoring period 1 and trained scoring period 2 for observer 1** Columns are for scores from scoring period 2, and rows are for scores from scoring period 1. The diagonal from top left to bottom right are the number of times the observers agreed on score.

	Scoring period 2				
Scoring period 1	**1**	**2**	**3**	**4**	**total**
1	2	0	0	0	2
2	1	1	2	0	4
3	0	0	0	0	0
4	0	0	0	1	1
total	3	1	2	1	7

Table 24: Frequency of trait scores for ZS given to U of M individuals and the disagreement between untrained scoring period 1 and trained scoring period 2 for observer 1. Columns are for scores from scoring period 2, and rows are for scores from scoring period 1. The diagonal from top left to bottom right are the number of times the observers agreed on score.

	Scoring period 2			
Scoring period 1	**0**	**1**	**2**	**total**
0	3	1	2	6
1	0	0	1	1
2	0	0	0	0
total	3	1	3	7

Table 25: Frequency of trait scores for NAS given to U of M individuals and the disagreement between untrained scoring period 1 and trained scoring period 2 for observer 1. Columns are for scores from scoring period 2, and rows are for scores from scoring period 1. The diagonal from top left to bottom right are the number of times the observers agreed on score.

	Scoring period 2			
Scoring period 1	1	2	3	**total**
1	3	1	0	4
2	1	0	0	1
3	2	0	0	2
total	6	1	0	7

Table 26: Frequency of trait scores for NBS given to U of M individuals and the disagreement between untrained scoring period 1 and trained scoring period 2 for observer 1. Columns are for scores from scoring period 2, and rows are for scores from scoring period 1. The diagonal from top left to bottom right are the number of times the observers agreed on score.

	Scoring period 2				
Scoring period 1	**1**	**2**	**3**	**4**	**total**
1	0	1	0	0	1
2	0	3	0	0	3
3	0	0	3	0	3
4	0	0	0	0	0
total	0	4	3	0	7

Table 27: Frequency of trait scores for NFS given to U of M individuals and the disagreement between untrained scoring period 1 and trained scoring period 2 for observer 1. Columns are for scores from scoring period 2, and rows are for scores from scoring period 1. The diagonal from top left to bottom right are the number of times the observers agreed on score.

	Scoring period 2				
Scoring period 1	**1**	**2**	**3**	**4**	**total**
1	3	0	0	0	3
2	0	1	0	0	1
3	0	0	2	0	2
4	0	0	0	0	0
total	3	1	2	0	6

Table 28: Frequency of trait scores for OBS given to U of M individuals and the disagreement between untrained scoring period 1 and trained scoring period 2 for observer 1. Columns are for scores from scoring period 2, and rows are for scores from scoring period 1. The diagonal from top left to bottom right are the number of times the observers agreed on score.

	Scoring period 2			
Scoring period 1	**1**	**2**	**3**	**total**
1	4	1	1	6
2	0	1	0	1
3	0	0	0	0
total	4	2	1	7

Table 29: **Frequency of trait scores for PZT given to U of M individuals and the disagreement between untrained scoring period 1 and trained scoring period 2 for observer 1.** Columns are for scores from scoring period 2, and rows are for scores from scoring period 1. The diagonal from top left to bottom right are the number of times the observers agreed on score.

	Scoring period 2				
Scoring period 1	**0**	**1**	**2**	**3**	**total**
0	0	0	0	0	0
1	0	4	0	0	4
2	0	1	2	0	3
3	0	0	0	0	0
total	0	5	2	0	7

Table 30: Frequency of trait scores for ANS given to U of M and BU individuals and the disagreement between trained scoring periods 2 and 3 for observer 1. Columns are for scoring period 3's scores, and rows are for scoring period 2's scores. The diagonal from top left to bottom right are the number of times the observers agreed on score.

	Scoring period 3			
Scoring period 2	**1**	**2**	**3**	**total**
1	6	0	0	6
2	0	4	0	4
3	0	4	5	9
total	6	8	5	19

Table 31: Frequency of trait scores for INA given to U of M and BU individuals and the disagreement between trained scoring periods 2 and 3 for observer 1. Columns are for scoring period 3's scores, and rows are for scoring period 2's scores. The diagonal from top left to bottom right are the number of times the observers agreed on score.

	Scoring period 3					
Scoring period 2	**1**	**2**	**3**	**4**	**5**	**total**
1	1	0	0	0	0	1
2	1	0	2	1	0	4
3	0	0	5	1	0	6
4	0	0	2	5	1	8
5	0	0	0	0	5	5
total	2	0	9	7	6	24

Table 32: Frequency of trait scores for IOB given to U of M and BU individuals and the disagreement between trained scoring periods 2 and 3 for observer 1. Columns are for scoring period 3's scores, and rows are for scoring period 2's scores. The diagonal from top left to bottom right are the number of times the observers agreed on score.

	Scoring period 3			
Scoring period 2	**1**	**2**	**3**	**total**
1	12	0	0	12
2	1	9	0	10
3	0	1	0	1
total	13	10	0	23

Table 33: Frequency of trait scores for MT given to U of M and BU individuals and the disagreement between trained scoring periods 2 and 3 for observer 1. Columns are for scoring period 3's scores, and rows are for scoring period 2's scores. The diagonal from top left to bottom right are the number of times the observers agreed on score.

	Scoring period 3				
Scoring period 2	**0**	**1**	**2**	**3**	**total**
0	5	2	0	0	7
1	0	8	2	0	10
2	0	3	3	0	6
3	0	0	0	0	0
total	5	13	5	0	23

Table 34: Frequency of trait scores for NAW given to U of M and BU individuals and the disagreement between trained scoring periods 2 and 3 for observer 1. Columns are for scoring period 3's scores, and rows are for scoring period 2's scores. The diagonal from top left to bottom right are the number of times the observers agreed on score.

	Scoring period 3			
Scoring period 2	**1**	**2**	**3**	**total**
1	4	0	0	4
2	2	17	0	19
3	0	0	0	0
total	6	17	0	23

Table 35: **Frequency of trait scores for NBC given to U of M and BU individuals and the disagreement between trained scoring periods 2 and 3 for observer 1.** Columns are for scoring period 3's scores, and rows are for scoring period 2's scores. The diagonal from top left to bottom right are the number of times the observers agreed on score.

	Scoring period 3					
Scoring period 2	**0**	**1**	**2**	**3**	**4**	**total**
0	1	1	0	0	0	2
1	0	15	1	1	0	17
2	0	0	0	0	0	0
3	0	0	0	2	0	2
4	0	0	0	1	0	1
total	1	16	1	4	0	22

Table 36: Frequency of trait scores for NO given to U of M and BU individuals and the disagreement between trained scoring periods 2 and 3 for observer 1. Columns are for scoring period 3's scores, and rows are for scoring period 2's scores. diagonal from top left to bottom right are the number of times the observers agreed on score.

	Scoring period 3		
Scoring period 2	**0**	**1**	**total**
0	2	3	5
1	1	8	9
total	3	11	14

Table 37: Frequency of trait scores for PBD given to U of M and BU individuals and the disagreement between trained scoring periods 2 and 3 for observer 1. Columns are for scoring period 3's scores, and rows are for scoring period 2's scores. The diagonal from top left to bottom right are the number of times the observers agreed on score.

	Scoring period 3		
Scoring period 2	**0**	**1**	**total**
0	8	3	11
1	0	8	8
total	8	11	19

Table 38: Frequency of trait scores for SPS given to U of M and BU individuals and the disagreement between trained scoring periods 2 and 3 for observer 1. Columns are for scoring period 3's scores, and rows are for scoring period 2's scores. The diagonal from top left to bottom right are the number of times the observers agreed on score.

	Scoring period 3			
Scoring period 2	**0**	**1**	**2**	**total**
0	5	0	7	12

1	0	3	1	4
2	3	0	5	8
total	8	3	13	24

Table 39: Frequency of trait scores for TPS given to U of M and BU individuals and the disagreement between trained scoring periods 2 and 3 for observer 1. Columns are for scoring period 3's scores, and rows are for scoring period 2's scores. The diagonal from top left to bottom right are the number of times the observers agreed on score.

	Scoring period 3				
Scoring period 2	**1**	**2**	**3**	**4**	**total**
1	5	1	0	2	8
2	0	7	0	0	7
3	0	0	4	1	5
4	0	0	0	1	1
total	5	8	4	4	21

Table 40: Frequency of trait scores for ZS given to U of M and BU individuals and the disagreement between trained scoring periods 2 and 3 for observer 1. Columns are for scoring period 3's scores, and rows are for scoring period 2's scores. The diagonal from top left to bottom right are the number of times the observers agreed on score.

	Scoring period 3			
Scoring period 2	**0**	**1**	**2**	**total**
0	6	0	1	7
1	1	1	2	4
2	1	0	10	11
total	8	1	13	22
				N

Table 41: Frequency of trait scores for NAS given to U of M and BU individuals and the disagreement between trained scoring periods 2 and 3 for observer 1. Columns are for scoring period 3's scores, and rows are for scoring period 2's scores. The diagonal from top left to bottom right are the number of times the observers agreed on score.

	Scoring period 3			
Scoring period 2	**1**	**2**	**3**	**total**
1	11	1	2	14
2	2	3	0	5
3	0	0	4	4

total	13	4	6	23

Table 42: Frequency of trait scores for NBS given to U of M and BU individuals and the disagreement between trained scoring periods 2 and 3 for observer 1. Columns are for scoring period 3's scores, and rows are for scoring period 2's scores. The diagonal from top left to bottom right are the number of times the observers agreed on score.

	Scoring period 3				
Scoring period 2	**1**	**2**	**3**	**4**	**total**
1	0	0	0	0	0
2	1	12	0	0	13
3	0	1	4	0	5
4	0	0	0	0	0
total	1	13	4	0	18

Table 43: Frequency of trait scores for NFS given to U of M and BU individuals and the disagreement between trained scoring periods 2 and 3 for observer 1. Columns are for scoring period 3's scores, and rows are for scoring period 2's scores. diagonal from top left to bottom right are the number of times the observers agreed on score.

	Scoring period 3				
Scoring period 2	**1**	**2**	**3**	**4**	**total**
1	5	0	0	1	6
2	0	6	0	0	6
3	0	0	2	0	2
4	2	0	0	5	7
total	7	6	2	6	21

Table 44: **Frequency of trait scores for OBS given to U of M and BU individuals and the disagreement between trained scoring periods 2 and 3 for observer 1.** Columns are for scoring period 3's scores, and rows are for scoring period 2's scores. The diagonal from top left to bottom right are the number of times the observers agreed on score.

	Scoring period 3			
Scoring period 2	**1**	**2**	**3**	**total**
1	14	1	0	15
2	2	4	0	6
3	0	0	3	3
total	16	5	3	24

Table 45: Frequency of trait scores for PZT given to U of M and BU individuals and the disagreement between trained scoring periods 2 and 3 for observer 1. Columns are for scoring period 3's scores, and rows are for scoring period 2's scores. The diagonal from top left to bottom right are the number of times the observers agreed on score.

	Scoring period 3				
Scoring period 2	**0**	**1**	**2**	**3**	**total**
0	0	0	0	0	0
1	0	10	4	0	14
2	0	2	5	0	7
3	0	1	2	0	3
total	0	13	11	0	24

Table 46: Frequency of trait scores for ANS given to BU individuals and the disagreement between observers 1 and 2. Columns are for observer 1's scores, and rows are for observer 2's scores. The diagonal from top left to bottom right are the number of times the observers agreed on score.CM

	Observer 1			
Observer 2	**1**	**2**	**3**	**total**
1	4	0	0	4
2	2	4	0	6
3	0	2	2	4
total	6	6	2	14

N

Table 47: Frequency of trait scores for INA given to BU individuals and the disagreement between observers 1 and 2. Columns are for observer 2's scores, and rows are for observer 1's scores. The diagonal from top left to bottom right are the number of times the observers agreed on score.

	Observer 2					
Observer 1	**1**	**2**	**3**	**4**	**5**	**total**
1	0	1	0	1	0	2
2	0	0	0	0	0	0
3	0	0	3	2	0	5
4	0	0	0	2	1	3
5	0	0	0	2	4	6
total	0	1	3	7	5	16

Table 48: Frequency of trait scores for IOB given to BU individuals and the disagreement between observers 1 and 2. Columns are for observer 2's scores, and rows are for observer 1's scores. The diagonal from top left to bottom right are the number of times the observers agreed on score.

	Observer 2			
Observer 1	**1**	**2**	**3**	**total**
1	2	5	2	9
2	0	1	5	6
3	0	0	0	0
total	2	6	7	15

Table 49: Frequency of trait scores for MT given to BU individuals and the disagreement between observers 1 and 2. Columns are for observer 2's scores, and rows are for observer 1's scores. The diagonal from top left to bottom right are the number of times the observers agreed on score.

	Observer 2				
Observer 1	**0**	**1**	**2**	**3**	**total**
0	1	3	0	0	4
1	0	6	3	0	9
2	0	0	3	0	3
3	0	0	0	0	0
total	1	9	6	0	16

Table 50: Frequency of trait scores for NAW given to BU individuals and the disagreement between observers 1 and 2. Columns are for observer 2's scores, and rows are for observer 1's scores. The diagonal from top left to bottom right are the number of times the observers agreed on score.

	Observer 2			
Observer 1	**1**	**2**	**3**	**total**
1	2	1	1	4
2	4	4	3	11
3	0	0	0	0
total	6	5	4	15

Table 51: Frequency of trait scores for NBC given to BU individuals and the disagreement between observers 1 and 2. Columns are for observer 2's scores, and rows are for observer 1's scores. The diagonal from top left to bottom right are the number of times the observers agreed on score.

	Observer 2					
Observer 1	**0**	**1**	**2**	**3**	**4**	**total**
0	0	0	0	0	0	0
1	3	2	2	3	2	12
2	0	0	0	0	0	0
3	1	0	0	0	1	2
4	0	0	0	2	4	6
total	4	2	2	5	7	20

Table 52: Frequency of trait scores for NO given to BU individuals and the disagreement between observers 1 and 2. Columns are for observer 2's scores, and rows are for observer 1's scores. The diagonal from top left to bottom right are the number of times the observers agreed on score.

	Observer 2		
Observer 1	**0**	**1**	**total**
0	2	1	3
1	3	1	4
total	5	2	7

Table 53: Frequency of trait scores for PBD given to BU individuals and the disagreement between observers 1 and 2. Columns are for observer 2's scores, and rows are for observer 1's scores. The diagonal from top left to bottom right are the number of times the observers agreed on score.

	Observer 2		
Observer 1	**0**	**1**	**total**
0	4	0	4
1	6	1	7
total	10	1	11

Table 54: Frequency of trait scores for SPS given to BU individuals and the disagreement between observers 1 and 2. Columns are for observer 2's scores, and rows are for observer 1's scores. The diagonal from top left to bottom right are the number of times the observers agreed on score.

	Observer 2			
Observer 1	**0**	**1**	**2**	**total**
0	4	0	1	5

1	0	1	2	3
2	1	0	8	9
total	5	1	11	17

Table 55: Frequency of trait scores for TPS given to BU individuals and the disagreement between observers 1 and 2. Columns are for observer 2's scores, and rows are for observer 1's scores. The diagonal from top left to bottom right are the number of times the observers agreed on score.

	Observer 2				
Observer 1	**1**	**2**	**3**	**4**	**total**
1	2	0	0	0	2
2	1	4	2	0	7
3	0	0	1	0	1
4	0	0	2	3	5
total	3	4	5	3	15

Table 56: Frequency of trait scores for ZS given to BU individuals and the disagreement between observers 1 and 2. Columns are for observer 2's scores, and rows are for observer 1's scores. The diagonal from top left to bottom right are the number of times the observers agreed on score.

	Observer 2			
Observer 1	**0**	**1**	**2**	**total**
0	4	1	0	5
1	0	1	1	2
2	4	1	2	7
total	8	3	3	14

Table 57: Frequency of trait scores for ANS given to U of M individuals and the disagreement between observers 1 and 3. Columns are for observer 3's scores, and rows are for observer 1's scores. The diagonal from top left to bottom right are the number of times the observers agreed on score.

	Observer 3			
Observer 1	**1**	**2**	**3**	**total**
1	0	1	1	2
2	0	0	4	4
3	0	0	4	4
total	0	1	9	10

Table 58: Frequency of trait scores for INA given to U of M individuals and the disagreement between observers 1 and 3. Columns are for observer 3's scores, and rows are for observer 1's scores. The diagonal from top left to bottom right are the number of times the observers agreed on score.

	Observer 3					
Observer 1	**1**	**2**	**3**	**4**	**5**	**total**
1	0	0	0	0	0	0
2	0	0	0	0	0	0
3	2	0	2	0	1	5
4	0	0	1	2	1	4
5	0	0	0	0	1	1
total	2	0	3	2	3	10

Table 59: Frequency of trait scores for IOB given to U of M individuals and the disagreement between observers 1 and 3. Columns are for observer 3's scores, and rows are for observer 1's scores. The diagonal from top left to bottom right are the number of times the observers agreed on score.

	Observer 3			
Observer 1	**1**	**2**	**3**	**total**
1	0	4	1	5
2	0	3	2	5
3	0	0	0	0
total	0	7	3	10

Table 60: Frequency of trait scores for MT given to U of M individuals and the disagreement between observers 1 and 3. Columns are for observer 3's scores, and rows are for observer 1's scores. The diagonal from top left to bottom right are the number of times the observers agreed on score.

	Observer 3				
Observer 1	**0**	**1**	**2**	**3**	**total**
0	2	0	0	0	2
1	3	2	0	0	5
2	0	3	0	0	3
3	0	0	0	0	0
total	5	5	0	0	10

Table 61: Frequency of trait scores for NAW given to U of M individuals and the disagreement between observers 1 and 3. Columns are for observer 3's scores, and rows are for observer 1's scores. The diagonal from top left to bottom right are the number of times the observers agreed on score.

	Observer 3			
Observer 1	**1**	**2**	**3**	**total**
1	1	1	0	2
2	3	5	0	8
3	0	0	0	0
total	4	6	0	10

Table 62: Frequency of trait scores for NBC given to U of M individuals and the disagreement between observers 1 and 3. Columns are for observer 3's scores, and rows are for observer 1's scores. The diagonal from top left to bottom right are the number of times the observers agreed on score.

	Observer 3					
Observer 1	**0**	**1**	**2**	**3**	**4**	**total**
0	1	0	0	0	0	1
1	0	2	1	0	2	5
2	0	0	0	1	0	1
3	0	0	0	0	2	2
4	0	0	0	1	0	1
total	1	2	1	2	4	10

Table 63: Frequency of trait scores for NO given to U of M individuals and the disagreement between observers 1 and 3. Columns are for observer 3's scores, and rows are for observer 1's scores. The diagonal from top left to bottom right are the number of times the observers agreed on score.

	Observer 3		
Observer 1	**0**	**1**	**total**
0	0	0	0
1	2	6	8
total	2	6	8

Table 64: Frequency of trait scores for PBD given to U of M individuals and the disagreement between observers 1 and 3. Columns are for observer 3's scores, and rows are for observer 1's scores. The diagonal from top left to bottom right are the number of times the observers agreed on score.

	Observer 3		
Observer 1	**0**	**1**	**total**
0	4	0	4
1	3	1	4
total	7	1	8

Table 65: Frequency of trait scores for SPS given to U of M individuals and the disagreement between observers 1 and 3. Columns are for observer 3's scores, and rows are for observer 1's scores. The diagonal from top left to bottom right are the number of times the observers agreed on score.

	Observer 3			
Observer 1	**0**	**1**	**2**	**total**
0	2	0	3	5
1	0	0	1	1
2	4	0	0	4
total	6	0	4	10

Table 66: Frequency of trait scores for TPS given to U of M individuals and the disagreement between observers 1 and 3. Columns are for observer 3's scores, and rows are for observer 1's scores. The diagonal from top left to bottom right are the number of times the observers agreed on score.

	Observer 3				
Observer 1	**1**	**2**	**3**	**4**	**total**
1	3	0	0	0	3
2	0	3	0	0	3
3	0	0	2	1	3
4	0	0	0	1	1
total	3	3	2	2	10

Table 67: Frequency of trait scores for ZS given to U of M individuals and the disagreement between observers 1 and 3. Columns are for observer 3's scores, and rows are for observer 1's scores. The diagonal from top left to bottom right are the number of times the observers agreed on score.

	Observer 3			
Observer 1	**0**	**1**	**2**	**total**
0	4	0	0	4
1	0	0	0	0

2	4	1	1	6
total	8	1	1	10

Table 68: Frequency of trait scores for NAS given to U of M individuals and the disagreement between observers 1 and 3. Columns are for observer 3's scores, and rows are for observer 1's scores. The diagonal from top left to bottom right are the number of times the observers agreed on score.

	Observer 3			
Observer 1	**1**	**2**	**3**	**total**
1	6	1	0	7
2	1	0	0	1
3	0	0	2	2
total	7	1	2	10

Table 69: Frequency of trait scores for NBS given to U of M individuals and the disagreement between observers 1 and 3. Columns are for observer 3's scores, and rows are for observer 1's scores. The diagonal from top left to bottom right are the number of times the observers agreed on score.

	Observer 3				
Observer 1	**1**	**2**	**3**	**4**	**total**
1	1	0	0	0	1
2	4	2	0	0	6
3	1	2	0	0	3
4	0	0	0	0	0
total	6	4	0	0	10

Table 70: **Frequency of trait scores for NFS given to U of M individuals and the disagreement between observers 1 and 3.** Columns are for observer 3's scores, and rows are for observer 1's scores. The diagonal from top left to bottom right are the number of times the observers agreed on score.

	Observer 3				
Observer 1	**1**	**2**	**3**	**4**	**total**
1	4	0	0	0	4
2	0	0	1	0	1
3	1	0	0	0	1
4	0	1	1	1	3
total	5	1	2	1	9

Table 71: Frequency of trait scores for OBS given to U of M individuals and the disagreement between observers 1 and 3. Columns are for observer 3's scores, and rows are for observer 1's scores. The diagonal from top left to bottom right are the number of times the observers agreed on score.

	Observer 3			
Observer 1	**1**	**2**	**3**	**total**
1	1	5	0	6
2	0	2	0	2
3	0	2	0	2
total	1	9	0	10

Table 72: Frequency of trait scores for PZT given to U of M individuals and the disagreement between observers 1 and 3. Columns are for observer 3's scores, and rows are for observer 1's scores. The diagonal from top left to bottom right are the number of times the observers agreed on score.

	Observer 3				
Observer 1	**0**	**1**	**2**	**3**	**total**
0	0	0	0	0	0
1	2	2	2	0	6
2	0	2	2	0	4
3	0	0	0	0	0
total	2	4	4	0	10

Appendix 3: Additional tables

Table 73: Table D: List of traits that had a subsample of BU individuals rescored to see if using Hefner and Linde (2018) photographs would change the score. Number of individuals rescored and number of paired scores do not match since some individuals were not available for all data collection sessions or observer 1 did not score the trait because she thought it could not be scored. 'CM2' designates scoring period 2 for observer 1, and 'CM3' designates scoring period 3.

Trait	*# of individuals rescored with atlas*	*# of paired scores for CM2 and CM3*	*# scores that did not change*	*# of scores that changed for individuals with paired scores*	*# of scores that changed to CM2's score*	*# of scores that changed but did not match CM2's score*	*# of times scores did not match between the two sessions*
ANS	12	8	8	4	2	2	2
INA	12	10	10	2	1	1	3
NAS	12	10	10	1	0	1	2
NBS	11	6	9	1	0	1	0
NFS	12	9	12	0	0	0	2
OBS	12	10	10	2	1	1	2
PBD	8	8	8	0	0	0	1
SPS	12	10	10	2	2	0	4
TPS	11	9	6	4	1	3	3
ZS	11	9	7	4	1	3	3
Total	113	89	90	20	8	12	22