2012

Reliability Study of Methods for Scoring a Non-Metric Human Osteological Trait

Shannon Freire
Ashley Dunford

Follow this and additional works at: https://dc.uwm.edu/fieldnotes

Recommended Citation
Freire, Shannon and Dunford, Ashley (2012) "Reliability Study of Methods for Scoring a Non-Metric Human Osteological Trait," Field Notes: A Journal of Collegiate Anthropology: Vol. 4 : Iss. 1 , Article 8. Available at: https://dc.uwm.edu/fieldnotes/vol4/iss1/8

This Article is brought to you for free and open access by UWM Digital Commons. It has been accepted for inclusion in Field Notes: A Journal of Collegiate Anthropology by an authorized administrator of UWM Digital Commons. For more information, please contact open-access@uwm.edu.
Reliability Study of Methods for Scoring a Non-Metric Human Osteological Trait

Shannon Freire and Ashley Dunford

Abstract: To gain meaningful insights from non-metric trait analysis in the field of human osteological study, issues of reliability and context need to be addressed, especially for the analysis of discrete cranial traits. A preliminary study tested the reliability of different methods of quantifying wormian bones, with the purpose of establishing a consistent method that would enable further applicability for this and other non-metric traits in mortuary analysis. The determination of reliability for both methods was made using Olsson and Janson’s (2001) iota statistic together with Pearson’s product-moment correlation. This study examines the reliability of scoring methods on an interobserver scale, an imperative step for the utility of these techniques for the wider archaeological community, as the majority of current data collection is a collaborative effort. The results of this study support the initial application in a case study involving a Bayesian probability analysis utilizing individuals from the Milwaukee County Institutional Grounds collection.

Key words: wormian bones, reliability, non-metric traits, digitization, interobserver, intraobserver

Introduction

Given the multi-layered nature of this topic, it may be helpful to provide a structured “road-map” for the various topics discussed in this article. Non-metric human osteological traits have archaeological utility if scored using reliable methods. Up to this point, there have not been many methods that have provided the necessary reliability and replicability to support this potentially productive data source. This study was designed to test the reliability of a new method to score non-metric human osteological traits. The non-metric trait used in this study is wormian bones, specifically those extra-sutary ossicles located within the lambdoid suture. This new method, which incorporates digital photography and tracing software, should theoretically provide individual researchers (intraobservers) with a reliable scoring method, as well as allow for greater research possibilities within interobserver research.

This study compares two methods to determine which is the more reliable. The first method involves a traditional scoring method (manual) where wormian bones are quantified individually, and each wormian bone equals one. The second method involves a digital scoring of wormian bones, where the bones and lambdoid suture are traced and quantified based on a percentage: the wormian bones represent a percentage within 0-100% of the total lambdoid suture. Both methods are based on the use of one sample of 20 crania that have been digitally photographed. These photographs are used for the manual and digital scoring.

The authors’ hypothesis is that the digital method will be more reliable on an intraobserver level. However, science and archaeology are not conducted within a vacuum. These techniques need to be able to be replicated by other researchers, especially when non-
metric traits are being scored by multiple individuals within the same project. The second component of the testing within this project involves repeating the entire reliability study using two individuals (judges) instead of one. This is a better representation of the actual potential use of non-metric traits.

We provide two examples of the archaeological applications of non-metric trait research. First, we have provided an archaeological case study using the same population from the Milwaukee County Institutional Grounds collection used as the sample for the reliability study. This case study is important in two ways: first, it shows why utilizing non-metric traits would have potential uses in an archaeological context, and second, why it is vital to have reliable results when dealing with sensitive populations. The second example of archaeological utility incorporates the possibilities of digital scoring within the shifting demands of repatriation. Non-metric human osteological trait research can and should experience a Renaissance if paired with new technologies that can enhance reliability.

Background

Though Berry and Berry (1967) provided a well-timed study pertaining to the analysis of “minor morphological variations in the human skull”, they vastly underrepresented what had been a recent deluge of publications focusing on these “rare”, “occasionally found”, and “neglected anthropological markers” (Caroline-Berry 1967). Berry and Berry 1967 focused their analysis on compiling a definitive list of non-metric traits. Non-metric traits are “morphological variations of anatomy, typically of a feature, or an anatomical landmark” (Saunders 2008:533). Non-metric traits are also referred to as discontinuous, or discrete traits (Corruccini 1974; DeStefano 1984; Prowse 1996). Adding or using the terms discontinuous or discrete to describe a trait signals a more genetically-based perspective (Buikstra 1990; Caroline-Berry 1967; Mays 1998). A discontinuous trait will generally have few possible phenotypic expressions (Ossenberg 1969; Rubini 1997).

The introduction of genetic theory helps to explain the rapid acceptance of analyzing non-metric markers within the human cranium and skeleton. The assumption was made that these traits actually represented genetic markers and hence were direct reflections of the genotype (Buikstra 1990; Dawson 1977). This would enable anthropologists to recreate biological relationships between prehistoric groups, an undeniably exciting proposition. The relatively low cost, ease of scoring, the effective use on fragmentary remains, and the development of a statistically significant measure of population distance added to the appeal of analyzing non-metric traits (Tyrell 2000). Scholarly debate peaked in the 1970s as researchers tested the degrees of heritability in discontinuous traits (Buikstra 1990), and the “logico-positivist New Archaeology” theoretical climate of the time was “highly conducive for the development of esoteric analytical techniques” (Tyrell 2000:292). Eventually, scholars began to debate whether genetic control existed at all (Bennett 1965; Gottlieb 1978; Ossenberg 1970), and methods issues relating to inter- and intra-population variation complicated research. The conclusion that the heritability of discrete traits was either only partial or entirely unknown left many researchers reluctant to pursue the study of discontinuous traits, convinced that there was little or no utility for using these traits to study past human populations (Buikstra 1990).
However, recent studies demonstrating the success of partially heritable non-metric traits in documenting population structure in living communities (Brasili 1999; Sjvøld 1984; Torgersen 1951) has led to a resurgence of biological distance studies using the traits (Buikstra 1990).

Buikstra (1990) states three motivations for examining population variation among prehistoric peoples. First, the results of analysis can be used to answer basic questions of evolutionary history, as shown by the work of scholars including Scuilli (1988) and Harding (1989). Second, biological distance studies can be used to “address basic archaeological questions, such as the nature of residence patterns” (Buikstra 1990:6). As a final point, biological distance analyses can also provide a background against which to examine paleopathology and paleodemography (Buikstra 1990). Given the potentials of this research, it is worthwhile to develop standardizations of method that can provide the most reliable representations of data samples.

This study will first examine reliability of two different methods of quantifying the discontinuous supernumerary ossicles known as wormian bones, with the intention of establishing a consistent method that would enable further applicability for this trait and other potential non-metric human osteological traits. Second, this study will examine the reliability of scoring methods on an interobserver scale, an imperative step for the utility of these techniques for the wider archaeological community, as the majority of current data collection is the result of collaborative effort. If meaningful insight can be applied to mortuary analysis and to biological distance studies via non-metric traits, as Buikstra (1990) and Hanihara (2003) suggest, then the analysis of discrete and quantitative cranial traits deserves a reexamination.

Wormian Bones

In the past, scholarly disagreement about the basic properties and expression of non-metric traits had created a two-part issue. First, researchers attempting tease out the etiology and genetic relationships between individuals, traits, and populations encountered some difficulty; a consensus of results was stymied by differing methods and sample selection. Second, researchers began to doubt the efficacy of using non-metric characteristics to study population relationships. Inter-population and intra-population variation made determining the genetic component of non-metric traits difficult. By the time Hauser et al. 1989 published their extensive guide, including the guidelines for sex, age, symmetry, laterality, and inter-trait associations when analyzing non-metric characters, interest had waned (Buikstra 1990). In an effort to recognize the current revived scholarly interest in analyzing these traits, this study examines in more detail a specific non-metric trait, extra-sutuary ossicles (wormian bones, as seen in the example below in Figure 1), in an attempt to answer method replicability challenges, and create a relationship with larger theoretical concerns of mortuary analysis.

For the purposes of this study, the term wormian bone will be used to indicate extra-sutuary ossicles located at both lambda and within the lambdoid suture between the skull measurement points lambda and asterion, as per Hanihara (2001a). Ossicles located at asterion, the juncture of the lambdoid, mastoid and occipito-mastoid sutures, are not included within the heading of wormian bones (Hanihara 2001a).
The first major scholarly works focusing on wormian bones were published at the turn of the 20th century, by Dorsey 1897 and Parker 1905. Studies following this initial work have since proposed a wide variety of hypotheses regarding the nature of wormian bones, reflecting paradigm shifts within bioarchaeology (Bennett 1965; Pucciarelli 1978; Rubini 1997), increasing knowledge of genetics (Hanihara 2001b; Lorentz 2009; Sjvøld 1984), and contributions from other fields of research, including neuroscience, prenatal care, and plastic surgery (Jeanty 2000; Lekovic 2007; Nitchter 1986). Proposed hypotheses of etiology include a range from solely environmental stressors affecting the formation and incidence of wormian bones to the primary influence of genetic factors with little or no relationship to environmental influences, as with many other non-metric traits (Sanchez-Lara 2007). As Jeanty (2000) indicates, “the prevalence of wormian bones has varied with the bias of the reporting authors” (Jeanty 200:866).

The incomplete understanding of wormian bone morphogenesis continues to fuel debate regarding their causation. The importance of the debate cannot be understated; whether wormian bones are genetically determined or environmentally influenced affects hypothesis of population distance or distribution, which assumes a genetic basis for the trait. Scholars such as Wileczak (2009) see this as an opening to hypothesize about the strains of environmental factors, such as intentional cranial deformation; Hanihara (2001a, 2001b, 2003) believes that genetic studies, such as those by Torgersen 1951 and Sjvøld 1984, which both focus on extra sutuary ossicles within the lambdoid suture that have heritability of at least .50, provide evidence of primary genetic control (O'Loughlin 2004; Sjvøld 1984; Torgersen 1951).

A synthesis of these two ideas represents the essential position taken in much of the work by current scholars, including Das 2005; Del Papa 2007; Hanihara 2003; Jeanty 2000; Lorentz 2009; O’Loughlin 2004; Peregrine 2002; Sanchez-Lara 2007. Hanihara’s (2001a;2001b) work shows the potential uses that result from a more inclusive conception of causation.
While it is encouraging to see wormian-bone related scholarship begin to move beyond the fierce internal debate of genetics vs. environment, the issue of problematic method still remains. The revival and success of using non-metric traits, wormian bones included, is dependent on both the precision used to record data and the reliability of our methods for scoring and quantifying these traits. As the lines of evidence that can be compared to wormian bones become more sophisticated (i.e. ABO, Rh, and Kell systems) (Brasili 1999), so too should our data collection and analysis. This study first examines (on an intraobserver level) the reliability of the traditional hand-count method of quantifying wormian bones, as compared to a new method to quantify the bones through digitization. It next provides an interobserver examination to compare the reliability of both methods between scholars.

Methods

For the purposes of both portions of the study, the precedent set by Antón 1992 for lowest scoring threshold for wormian bones was used. All clearly visible wormian bones completely limited by the lambdoid suture were scored. In an effort to avoid scoring complex sutures as possible extra sutuary ossicles, Wilczak’s 2009 method of following the entire border of the suture to confirm a bony island was followed for both the manual scoring of traits and the digital tracing of the traits. The sample of 20 crania was selected from a larger group of 365 examined crania. Samples were selected primarily for the presence of wormian bones, as the original sample of 32 individuals was reduced to 20 in an effort to select the crania with complete lambdoid sutures on the right and left side that were not entirely illegible. Crania that were in more than two fragments were also excluded. Given the precedent for excluding sex as a variable in the occurrence of wormian bones (Caroline-Berry 1967; Hanihara 2001a; Hanihara 2001b; Hanihara 2003; O'Loughlin 2004; Ossenberg 1970), the crania were not separated by sex. Any bones located at the point of asterion were not included, per Brasili’s (1999) finding that these ossicles were more common in males. Age was considered only in that juveniles were excluded, a practice also supported by the literature (Baker 2005; Caroline-Berry 1967).

After the sample crania were selected, photographs were taken of left and right sides of the lambdoid suture. Of the many photographs taken, those of the best quality were selected to represent each cranium, one image per side. These photographs constituted the sample used for the three manual and Canvas digital trials of the first and second components of the study. Photographs were utilized in an effort to both reduce the amount of stress on fragile remains and provide an equivalent sample for the manual and digital trials.

Three trials occurred with an interval of seventy-two hours between them, as per the intraobserver guidelines established by Gualdi-Russo 1999, Brasili 1999 and Sanchez-Lara 2007. These trials included one session of manual scoring using the digital photographs, followed by digital scoring of the same photographs uploaded to the Canvas program as PDFs. The Canvas portion of the scoring was quite time consuming. In total, scoring both the left and right portions of the lambdoid suture of the 20 sample skulls took approximately 8-10 hours. The second author (Dunford) fulfilled the role of Observer 2 for the interobserver portion of the study; providing additional osteological experience and familiarity with the nature of the research problem. In order to maintain comparability with the results of the intraobserver portion, the same schedule was used for the interobserver portion. The following procedures were followed exactly for both
intra and interobserver trials, with the obvious addition of a second researcher for the interobserver trials. Images for the manual trials were examined and the number of wormian bones on each side was recorded and added to give a total number of wormian bones per cranium.

The digital tracing of the lambdoid suture and enclosed extra sutuary ossicles was accomplished with the use of the ACDSee Systems program, Canvas, Version 9 for Mac OS X. The Canvas program is designed for technical and general drawing applications, as well as imaging, creating, and editing vector and raster graphics (ACDSee International 2011). The photographic images for each crania, left and right side, were individually loaded onto the Canvas pasteboard as PDF documents. The crania had been measured individually to create a scale for each side. Photographs were sized accordingly before tracing began. Using the free-form trace tool, the full sutures involving the perimeters of independent wormian bones were outlined. Next, the remaining segments of the lambdoid suture until the point of asterion were traced. Using the inquiry feature, the full length in centimeters of the both the non-wormian and wormian suture segments were obtained. A percentage of the wormian bones within the lambdoid suture was calculated.

As per Molto’s (1979) findings, each intraobserved trial was treated as a separate event and thus separate judge. The mean of bivariate correlations, iota of quantitative data, and Pearson’s produce moment correlation were computed for the intraobserver manual and Canvas digital trials using the computer program, R. This series of statistical procedures was followed exactly for the second (interobserver) component of the study.

Both components of this study are concerned with the measurement of reliability in a case where two or more observers, or, more specifically, data trials, have evaluated a sample on several dimensions, the manual quantification vs. Canvas digitization. In order to evaluate the extent of agreement within this study, it is important to consider chance-correction in analysis, in that the amount of agreement is greater than that which could be expected from random chance. As Olsson and Janson (2001) indicate, “it is also of advantage for an agreement measure to be directly applicable to reliability assessment” (Olsson 2001:278). Berry and Mielke (1988) developed an extension of kappa measurement represented by the following equation:

\[ R = 1 - \frac{\delta}{\mu} \]

where \( \delta \) represents the average Euclidean distance between the data produced by the two or more evaluators, and \( \mu \) represents the distance between one evaluator’s rating of a specific object and any other evaluator’s rating of any other object (Berry 1988). This functions for this study’s case of several observers and multivariate interval data. Olsson and Janson (2001) are under the opinion that Berry and Mielke’s (1988) use of Euclidean distance as a measure of disagreement results in the “uncertain interpretation of the resulting coefficient, \( R \), in conventional reliability terms” (Olsson 2001:278). Based on the Berry and Mielke (1988) equation and an extension of Cohen’s (1960) version of a kappa equation, the Olsson and Janson (2001) iota equation fulfills the desire for a multivariate agreement measure created in consideration of intraclass correlations to test the reliability of a population and kappa coefficients to quantify agreement in the sample.
in excess of chance (Cohen 1960). Olsson and Janson (2001) refer to observers or raters as ‘judges’ and the objects rated within the sample as ‘targets’ (Olsson 2001). This nomenclature will be used for the remainder of this section. The primary difference between the Olsson and Janson (2001) equation and the Berry and Mielke (1988) equation is that Olsson and Janson (2001) utilize squared Euclidean distance as the distance measure (rather than Euclidean distance) (Olsson 2001). Observed disagreement, as expressed by Olsson and Janson (2001) is the “the average (over targets and pairs of judges) of the squared Euclidean distances between judge’s ratings of the same target” (Olsson 2001:279). This is expressed in the following equation,

\[ d_o = \left( \frac{n}{2} \right)^{-1} \sum_{r<s} \sum_{j=1}^{n} \sum_{i=1}^{c} (x_{rij} - x_{rjs})^2, \]

where \( n \) represents the number of targets, \( b \) represents the number of judges, and \( c \) the number of variables. \( \Sigma \) is presented as the sum over all \( r \) and \( s \), such that \( 1 \) is greater than or equal to \( r \), \( s \) is greater than \( r \), and \( b \) is greater or equal to \( s \), or \( 1 \leq r < s \leq b \).

Following the observed disagreement, the expected disagreement becomes “the average of the squared Euclidean distances between one judge’s rating of a target, and any other judge’s rating of any other target” (Olsson 2001:280). This is expressed in the following equation,

\[ d_e = \left( \frac{n}{2} \right)^{-1} \sum_{r<s} \sum_{j=1}^{n} \sum_{i=1}^{c} (x_{rij} - x_{rjs})^2. \]

Olsson and Janson (2001) refer to expected in the context of kappa literature, which frames expected as “expected from the distribution from the cases at hand” (Olsson 2001:280). Thus, the iota measurement, as the measurement of agreement, is defined as

\[ \iota = 1 - \frac{d_e}{d_o}. \]

Once again, iota describes the magnitude of agreement in a sample over and above that expected by chance. An iota of 1 would indicate perfect agreement, whereas an iota of 0 would essentially indicate an agreement equivalent to chance. Because the variables (the manual count and the Canvas digitization) were measured on two different scales, the variance was equalized by standardization, thus accounting for all judges’ ratings on a variable.

The mean of bivariate correlations, the iota for quantitative data, and the Pearson’s product-moment correlation were computed for the manual and Canvas digital intraobserver trials using the open-source computer language R. R is a flexible program designed for statistical computing (Verzani 2005). This series of statistical procedures was followed exactly for the second (interobserver) component of the study.

Statistical analysis began with bivariate correlation computation to test inter-rater (inter-judge) reliability. The null hypothesis in this case was that agreement would be (0.0). Given a significant p-value, that hypothesis was rejected. While the manual intraobserver trials
established a correlation of (0.79), the next step in analysis compared the digital correlation to the manual, to determine whether the digital trials were more reliable. Both the correlation value (0.962) and the z-value for the Canvas digital intraobserver trial were higher than the manual intraobserver trial, thus indicating that the digital trial was more reliable within a bivariate correlation computation.

The next component of statistical analysis performed the Olsson and Janson (2001) iota test for quantitative data. Recall that the purpose of this test was to evaluate the extent of agreement within the study while considering chance-correction in analysis. In other words, it is important that the amount of agreement is greater than that which could be expected from random chance. When the tests were run, the iota value for the manual intraobserver trial was (0.685) to the digital intraobserver trial’s iota value of (0.955). The digital trial’s extent of agreement beyond random chance was considerably higher than the manual value. According to the theory proposed by Olsson and Janson (2001), this agreement measure, through standardization of variables and elimination of random chance, is a direct indicator of reliability (Olsson 2001).

The last computation performed was a Pearson’s product-moment correlation to test the correlation between the output of the manual test and the output of the digital test. Theoretically, the correlation between the outputs of each test should be very high. Both measures were methods to describe the nature of the wormian bones in the sample crania by quantity in integer form. The ratings for each of the lot numbers (three per lot, reflecting the three different trials, or judges) were collapsed into one single score per lot number. This was performed for the manual and digital data. These averaged readings were then supplied as the data for the Pearson’s correlation test. The correlation found the manual and digital ratings to be significantly correlated, however, the moderate value of the test suggested that they were not the same.

Results

The following lots were shown to be the main issue in the intraobserver results; 2 (MCIG Lot # 1021), 13 (MCIG Lot # 5054), and 20 (MCIG Lot # 5210). The manual test gave these lots low values, while the digital test gave them values that were much higher, as seen in Figure 2. The statistical computations involved in the bivariate correlation test and the iota test found that the agreement values for the digital trials were not only greater than those of the manual trials, but were also greater than what might be expected from random chance. This data suggests that the Canvas digital trials were considerably more reliable than the manual count trials. However, the discrepancy highlighted by the Pearson test in the correlation between the manual and digital trials should be explored.

Manual trials gave these three lot numbers very low values, while Canvas digital trials gave them high values. Both trials reflect an interpretation, a judge’s rating of a quantity of wormian bones. It appears that, in the case of these three lot numbers, the size of the wormian bones may have played a role in the different values between the manual and digital trials. A single wormian bone, regardless of size, will receive a manual score of one. However, that same wormian bone will receive a different digital score dependent on size relative to the lambdoid
suture. As an example, Lot Number 1021 consistently received a score of 3 from the manual quantification of wormian bones. In the digital tests, this cranium received an average score of 82.7895% wormian bone to lambdoid suture. In this scenario, the manual and digital tests do not appear correlated with one another. However, the remaining samples that included wormian bones of a much smaller size showed a significant correlation between manual and digital tests.

Figure 2. Scatterplot of average manual scores and average Canvas digital scores, representing the twenty individual samples. Note the cluster of outliers represented by 2, 13, and 20, or Lots 1021, 5054, and 5210. Manual count is represented by m1; digital scoring is represented by m2.

This finding does not necessarily mean that wormian bones of an exceptionally large size should be excluded from a sample, rather it calls attention to the research question at hand: quantifying wormian bones by the most reliable means. A visual inspection of MCIG Lots # 1021, 5054, and 5210 shows that wormian bones comprise a large part of the lambdoid suture. A possible solution to the manual scoring’s low values could involve creating a weighted scale. A threshold could be established whereby a single wormian bone of a certain size would count as 2 or more bones. This would entail measuring the bone and determining the point at which a single bone would be weighted as multiple bones. However, these additional steps have the potential to introduce problematic sources of error. If the purpose of data collection is to test methods of quantification, it appears that the percentage of wormian bone to suture created by digital measurement is a more reliable representation of quantity. After the intraobserver trials provided a clear example of the more reliable method, the decision was made to replicate the trials using two observers. This would further test the utility and reliability of scoring methods when used by researchers working in concert.
As stated previously, the same series of statistical tests were performed for the interobserver trials. This additional step was taken to corroborate the increased reliability of digital trials found in the intraobserver component of the study. In this second inter-observer study, the agreement values for Rater One were higher for the digital trials (as seen in Figure 3). Rater One is represented by M2 and M4 on the scatterplots. This confirmed the results of the initial study. However, the agreement values for the manual trials (as seen in Figure 4) were higher for Rater Two (represented by M1 and M3 on the scatterplots).

Once a Pearson product-moment test was conducted, the results of the test suggested that although the raters differed on which method was more reliable for the intraobservation, the iota values for the digital trials were significantly more reliable when compared to the iota values for
manual counts within the interobservation study (the results of rater one compared to the results of rater two). Thus, when examining the reliability of trials between scholars, the digital scoring format was the more reliable.

Are manual hand counts for analyzing wormian bones obsolete? Given the high correlation between the results of the manual and digital trials in the intraobserver study, as well as the high iota values for the digital trial in the interobserver study, the assumption would be that if target variables are being quantified in a similar way, and one method was significantly more reliable, the more reliable method would be utilized exclusively. This is further supported by the results of the interobserver trials, where digital trials were shown to be more reliable between observers. In a field (non-metrics) plagued by questions of reliability and replicability, this seems to be a reasonable assumption. Before this question is fully addressed, two examples of applications of non-metric research are provided in an effort to elucidate this question.

**Case Study: The Milwaukee County Asylums**

Non-metric traits have their own role to play in mortuary analysis as archaeologists construct and reconstruct models of the past. I believe that there is an association between wormian bones and specific congenital disorders that bears a closer and careful examination, is particularly relevant in the context of this study’s sample, and will illustrate applications for mortuary analysis.

As early as the 16th century, anatomists Andernah and Vesale associated wormian bones with cerebral disorders (Jeanty 2000). The most commonly described disorders associated with wormian bones involve those of the central nervous system. Jeanty (2000), citing Pryles and Khan (1979), supports their assessment that the prevalence of abnormalities within the central nervous system in a population with wormian bones varies between 93 and 100% for a random sample and reaches nearly 100% in a mentally retarded population (Jeanty 2000; Pryles 1979). The high numbers reported by Pryles and Khan (1979) have not been replicated by other authors, however, as Sanchez-Lara (2007) indicates, in a number of genetic syndromes, wormian bones occur so frequently that they may included in diagnostic tables or may be considered primary phenotypic characteristics of the syndrome (Sanchez-Lara 2007).

Wormian bones are associated with genetic syndromes for a variety of reasons. In the case of hydrocephaly, intracranial forces spread open sutures. This effect is similar to metabolic disorders causing a reduction in skull ossification, resulting in wider sutures seen in brachycephalic disorders (Sanchez-Lara 2007). It is also important to remain cognizant of the issue that the wormian bones themselves are not indicators of a negative prognosis, hence any prognosis would depend on the type and severity of a given congenital disorder. In other words, an individual with a wormian bone does not necessarily have a congenital disorder/genetic syndrome, but an individual with a congenital disorder/genetic syndrome is likely to have a wormian bone.

Given the association between wormian bones and genetic syndromes, there is a possibility that wormian bones may be a post-mortem indicator of a congenital nervous system
A RELIABILITY STUDY OF METHODS FOR STUDYING THE METHODS OF SCORING A NON-METRIC HUMAN OSTEOLOGICAL TRAIT

disorder. Obviously, there would be no way to definitively prove this, and it is important to keep the background of the pertinent population in mind; as Hanihara’s (2001a;2003) work has shown, some global populations display higher frequencies of wormian bones than others. However, in this case of the population interred at the Milwaukee County Poor Farm, burials included individuals who had been institutionalized at the Milwaukee County Hospital for Acute Insane (which later became the Milwaukee County Hospital for the Mentally Diseased) and the Milwaukee County Asylum for the Chronic Insane (which later became the Milwaukee County Asylum for Mental Diseases) (Richards 1997).

In 1880, Milwaukee County authorized the construction of two separate hospitals, one for the care of the sick, and the other for the containment of the insane (Richards 1997). Conditions dramatically improved for the inmates of the asylum, and a report issued in 1881 listed 126 persons present. The reasons for becoming institutionalized, according to the 1881 annual report, included, cerebro spinal sclerosis, epilepsy, intemperance, ill health, ill health from overwork and privation, sunstroke, senility, masturbation, injury to head, idiot, and unknown (Richards 1997). Twenty-five patients died in the year the report was produced from a variety of causes, including cerebral hemorrhage, senile gangrene, epilepsy, senile debility, exhaustion from acute melancholia, and others (Richards 1997). The state asylum in Madison also began sending its chronically insane to the asylum, making conditions once again inadequate only eight years later (Richards 1997). Many of those that died at the asylums were buried within the Milwaukee County Poor Farm. However, not all individuals buried within the Milwaukee County Poor Farm cemetery were asylum inmates.

The grave goods recovered from the Great Lakes Archaeological Research Center excavation were divided into three categories. Of specific interest to this sample is Category I, representing the grave goods associated with the internments of the residents of county institutions. According to Richards (1997), Category I grave goods are “defined by association with only safety pins and/or medical equipment. These grave goods include safety pins, fragments of fabric and rubber, medical instruments, and number tags. If safety pins were found with other types of grave goods, in particular, with buttons or shoes, clothing rather than shroud burial should be assumed and the internment should not be considered a Category I. The presence of Category I grave goods suggest burial in shrouds of rubber or cloth secured with pins” (Richards 1997:209).

In an effort to determine the conditional probability that individuals with wormian bones would also be part of an institutional (asylum) population, several lines of data must be utilized. However, this conditional probability is produced with the caveat that the following data samples of Hanihara 2001b and Pryles and Khan 1979 are limited, and burial log figures of the Poor Farm may be underrepresented. To begin, the general population interred at the Milwaukee County Poor Farm is European in origin, especially German, Irish, and Polish (Richards 1997). Hanihara (2001b), in a study of the frequency of wormian bones in major global populations, found that 1.5 % of European populations expressed this trait. According to the burial log of the Poor Farm, between 1881 and 1925, approximately 5,377 individuals were interred, including 144 individuals from both insane asylums. Using Hanihara’s (2001b) figure, of the 5,233 internments not involving an asylum population, 98.5 % of individuals would likely not have wormian bones.
Pryle and Khan’s (1979) found associations between wormian bones and congenital disorders. Approximately 96% of individuals that expressed the wormian bone trait also had some degree of central nervous system disorder. If the percentage of an internment originating from the asylums is divided by the overall percentage of wormian bone expression (asylum and general population), there is a 66.45 % (see Figure 5) probability that an individual with wormian bones in the MCIG collection was also institutionalized at one of the Milwaukee County asylums.

Again, this figure is likely an underrepresentation. In the sample of 20 individuals used for this study, 6 individuals were recovered with safety pins, which fit within the description for a Category I internment. Of those six, one individual was identified as Neltonia Hennigson, who died of broncho pneumonia at the age of 41 at the County Hospital. The other five individuals, given our conditional probability figure and the poor knowledge of treating many of diseases/disorders at the time of this sample’s internment (1859-1925), may be regarded as likely inmates of one or both of the asylums for mental disease. While these estimates are by no means definitive, using a combination of mortuary analysis and non-metric trait research, these five individuals could be linked a bit closer to their identities.
Method of Choice

One way of answering whether manual scoring is obsolete is to ask another question; is digital scoring worth the extra cost, and will the digital scoring add anything to our knowledge? Gualdi-Russo provides an eloquent answer:

In analyses of non-metric traits in the field of human population biology, the scientific goals are strictly dependent on the precision used in the recording of the data. Therefore, in addition to the use of traits with a firm hereditary component, a common effort is required for the definition of a rigorous methodology of the scoring of epigenetic traits. Only by overcoming these problems, by means of adequate standardization and observer experience, will we be able to speculate on the meaning of these biological variations in the differentiation of human populations. (Gualdi-Russo 1999:549)

If there is going to be a future for the use of non-metric traits in archaeology, the methodological problems plaguing past studies need to be addressed. Gualdi-Russo’s (1999) exhortation to standardize methods and become more precise in recording data suggests that there is very little room for error.

The Canvas digitization does take time, and the program currently retails for $599 on the ACDSee website, making it both more expensive and more time consuming than manual scoring. The following section, Applications for Repatriation, outlines some justifications for this cost.

The Canvas digitization method of scoring is clearly more reliable than manual scoring. There may be certain traits, (e.g., a foramen) where it would be beneficial to score as present only if penetrance was complete; a factor potentially difficult to determine from photographs, and hence not conducive to digitization. In this case, it would be reasonable to maintain manual scoring. However, digitization would be applicable to the far majority of non-metric traits. The initial investment in time and resources seems worthwhile when considering the returns and applications of more reliable data. It would be this author’s recommendation to replace manual scoring with digitization.

Applications for Repatriation

In 1990, after years of legislative effort, the 101st Congress of the United States of America passed the Native American Graves Protection and Repatriation Act (NAGPRA). The law, intended to reflect a “national consensus concerning the dignity and respect due American Indians, their property, and their cultures” (Gunn 2005:505) is administered under the National Park Service, a component of the Department of the Interior.

Of special relevance to archaeologists and osteologists, Section 7 of the law establishes the ‘expeditious return’ of certain materials, including human remains, through a process called repatriation (Gunn 2005). Repatriation cases have opened “institutional skeletal collections up to potential claims by individuals who feel spiritual connections to the remains in question” (White...
Many current archaeological excavations that recover human remains are also subject to NAGPRA regulations (White 2000).

Vast amounts of information have been and continue to be gained by the study of human osteological materials. New analytical techniques and instruments continue to be developed, and the theoretical inquiries that utilize remains as a line of evidence continue to evolve. Perhaps most importantly, as White (2000) points out, the character of self-correction is integral to the scientific endeavor. However, past injustices of scientific study necessitated a law that would allow aggrieved tribal parties obtain information and custody of ancestral skeletal remains (White 2000). Compromise becomes necessary under the law (Renfrew 2004).

While scientists like White (2000) state that “no cast, no image, no measurement, no description can adequately record the information potential held by an original osteological specimen” (White 2000:329), the fact remains that some human skeletal remains will be repatriated, and those recovered in current excavations may be immediately reburied. Digitization has vast applications in this case. A possible justification for the cost of a program such as Canvas would be the potential to continue to learn from remains that may no longer be physically available. Available images of remains, along with the correct scale, may be preserved and restudied digitally. While digitization cannot answer questions pertaining to chemical composition or the internal forms of skeletal material, any possibilities for learning and generating information are better than none at all.

**Conclusion**

Non-metric human osteological trait analysis is a field that has been challenged by miscommunication, errors, assumptions, inconsistent methodologies, and poor sampling. In some respects, the patterns of growth and decline in non-metric research resemble the larger field of anthropology, and the paradigmatic shifts that reorganize research priorities. Now, past moments of renaissance, revival, and rediscovery, the time has come to seriously approach the continued methodological issues that plague non-metric analysis.

Through a detailed examination of a sample non-metric trait, the supernumerary ossicles called wormian bones, this study addressed the challenge of reliable methods. This study has compared the most current methods of manual scoring wormian bones with a new approach to quantification by means of digitization in the ACDSee Canvas program. The two methods were then compared using statistical computations that tested bivariate correlation and iota agreement.

Quantification through digitization emerged as the overwhelmingly more reliable method for intraobserver study, and the more reliable method when considering iota values for the interobserver study. However, what do researchers stand to gain from this increased reliability? There certainly may be infrequent cases when manual scoring would be preferable or necessary, but as the Milwaukee County Asylums case study showed, if non-metric traits are going to be utilized in mortuary analysis, there is a need for increasingly precise data collection and reliable results.
With the possibility of widely reliable results between scholars, as shown by the second component of this study, perhaps the potential of non-metric analysis can be fulfilled, and contributions to larger archaeological questions may be made.

References

ACDSee International, Inc.

Baker, Brenda, Tosha L. Dupras, and Matthew Tocheri

Bennett, Kenneth A.

Berry, K.J., and P.W. Mielke Jr.

Brasili, P., Zaccagni, L. and E. Gualdi-Russo.

Buikstra, Jane E.

Caroline-Berry, A. and R. J. Berry

Cohen, J.

Corruccini, R.S.

Das, Srijit, Rajesh Suri, and Vijay Kapur
2005 Anatomical observations on os inca and associated cranial deformities. Folia Morphologica 64(2):118-121.

Dawson, George L. and El-Najjar, Mahmoud Y.

Del Papa, Mariano C. and S. Ivan Perez

DeStefano, G.F., G. Hauser, A. Guidotti, S. Rossi, E. Gualdi-Russo, P. Brasili Gualandi
Dorsey, G.A.

Gottlieb, Karen

Gualdi-Russo, E., M.A. Tasca, and P. Brasili

Gunn, Steven J.

Hanihara, Tsunehiko and Hajume Ishida

Hanihara, Tsunehiko, Hajume Ishida, and Yuiko Dodo

Harding, R.M., F.W. Rosing, R.R. Sokal

Hauser, G., and G. F. De Stefano

Jeanty, Philippe

Lekovic, Gregory P. M.D., Ph.D., J.D.; Baker, Brenda J. Ph.D.; Lekovic, Jill M. M.D.; Preul, Mark C.

Lorentz, Kirsi O.

Mays, Simon

Molto, J.E.

Nitchter, L.
O'Loughlin, V. D.

Olsson, Ulf and Harald Janson

Ossenberg, N.S.
1969 Discontinuous morphological variation on the human cranium, University of Toronto.

Parker, C.A.
1905 Wormian Bones. Chicago: Roberts Press.

Peregrine, Peter

Prowse, Tracy L. and Nancy C. Lovell

Pryles, C.V., and A.J. Khan

Pucciarelli, Hector A.

Renfrew, Colin

Richards, Patricia B.
1997 Unknown Man No. 198 The Archaeology og the Milwaukee County Poor Farm Cemetery, Anthropology, University of Wisconsin- Milwaukee.

Rubini, Mauro, Silvia Mogliazza, Robert Spencer T. Corruccini

Sanchez-Lara, Pedro A., John M. Graham Jr, Anne V. Hing, John Lee, and Michael Cunningham

Saunders, Shelley R. and Dori L. Rainey

Sciulli, Paul W., Charles Janini, and Myra Gleson

Sjvøld, T.
Torgersen, J.  

Tyrell, Andrew  

Verzani, John  

White, Tim D. and Pieter Arend Folkens  

Wilczak, C.A. and S.D. Ousley  