H5 The Utility of Cohen's Kappa for Testing Observer Error in Discrete Data and Alternatives

Alexandra R. Klales, MS, 501 East 38th Street, Erie, PA 16546; and Stephen D. Ousley, PhD, Mercyhurst College, Department of Applied Forensic, Anthropology, 501 East 38th Street, Erie, PA 16546*

After attending this presentation, attendees will learn to use caution in the application of Cohen's Kappa (1960) for assessing interobserver error for ordinal scoring and will learn of the utility in forensic anthropology of alternative methods that can be applied to data for error testing.

This presentation will have an impact on the forensic anthropology community by presenting methods that can be reliably used to test inter- and intra-observer error for discrete data, which is frequently used in sex and age estimation. Additionally, this study will encourage the use of appropriate tests of error for discrete data that can then be applied to make forensic anthropological studies *Daubert* compliant.

Compliance with the *Daubert* criteria requires the validation of all medico-legal methods through error testing. However, Ingvoldstad and Crowder (2009) have revealed a lack of consistency in the field for testing observer error in forensic anthropological research. While there is a clear trend to increase error testing in recent years, many studies fail to test both inter- and intra- observer error, or to use reliable methods for some specific data sets. Among the latter, testing inter- and intra- observer error in discrete data, such as the qualitative or ordinal data often employed to assess sex or age, poses a particular challenge. Cohen's Kappa (1960) is one of the most popular methods to test observer error in discrete data in disciplines such as clinical medicine and psychology, to name a few, yet its application within forensic anthropology has been limited.

The goal of this study is to examine the utility of Cohen's Kappa for assessing observer error and also to compare it with those of other alternative methods. This evaluation is based on the analysis of newly recorded data and the original datasets from Klales et al. (2009) and Vollner et al. (2009) to develop new methods for sex estimation from the human pelvis. A sample of 170 innominates of known, adult individuals from the Hamann-Todd Collection (HTH), housed at the Cleveland Museum of Natural History, was used for this study. Each of the Phenice (1960) traits was scored from the HTH material on separate occasions by two to four different individuals. Scores from one to five were assigned following the scale and corresponding illustrations and written descriptions in the previously mentioned studies. Different estimates of intra- and inter- observer error were obtained, as well as the percentages of correct classification for each specimen observed.

In the prior studies, sex estimation was based on linear discriminant function analysis and provided correct classification rates above 99%. However, Cohen's Kappa in these studies rendered low values: ventral arc 0.53, subpubic concavity 0.40, and medial aspect of the ischio-pubic ramus 0.43 (i.e., high inter-observer errors), therefore questioning the likelihood of replicating this high correct classification rates when the variables are scored by other researchers. Specifically, the low Kappa values did not seem to reflect differences in the end result, a highly accurate method of estimating sex; therefore, using this method of measuring reliability, specifically, inter-observer differences, may itself not be reliable for discrete data, because differences in scoring minimally affected classification accuracy. Unreliable methods cannot be valid and in this case, a low indication of reliability was contradicted by a high indication of validity. When evaluating reliability, one should not be discouraged by a "low" Kappa value, but must also look at other statistics and the practical consequences of inter-observer differences.

Results suggest that in spite of the low Cohen's Kappa figures originally obtained for these data sets, correct classification rates remain high and fairly constant independent of the observer. Cohen's Kappa was clearly outperformed by some of the alternative error estimation methods, which provided results more consistent with the observed correct classification rates. This suggests that Cohen's Kappa should be interpreted with caution, or even abandoned, when analyzing ordinal and other discrete data in forensic contexts.

Observer Error, Cohen's Kappa, Discrete Data

 ** Presenting Author*