A37 Inter-Observer Reliability of the Total Body Score System for Quantifying Human Decomposition

Gretchen R. Dabbs, PhD, Southern Illinois University, Dept of Anthropology, 1000 Faner Drive, MC 4502, Carbondale, IL 62901; Melissa A. Connor, PhD, Colorado Mesa University, 406 Lowell Heiny Hall, 1100 N Avenue, Grand Junction, CO 81501-3122; and Joan A. Bytheway, PhD, Sam Houston State University, College of Criminal Justice, Box 2296, Huntsville, TX 77341-2296*

The goals of this presentation are to: (1) demonstrate inter-observer reliability of the Total Body Score (TBS) system for quantifying decomposition; (2) illustrate where discordance between observers occurs; and, (3) make suggestions to improve the system.[1]

This presentation will impact the forensic science community by demonstrating that the TBS system has low levels of inter-observer error and is a reliable method of quantifying decomposition.

Megyesi and colleagues' TBS system for quantifying decomposition is often cited and is currently used, along with other methods, to record decomposition at several human decomposition research facilities. However, the consistency of observations between two or more observers has never been tested. The use of multiple observers throughout the decomposition process is common at decomposition research facilities. Testing methods to determine rates and potential sources of error are now required in forensic sciences in this post-*Daubert* and National Research Council (NRC) report climate.[2,3] This presentation addresses the observed deficiency.

Sixteen participants scored 59 observation packets using the Megyesi et al. system. The participants included both sexes ranging in education (undergraduate to PhD) and experience (0-6 months to 10+ years). All had some experience working at a human decomposition research facility and using the TBS system. The packets used 13 human cadavers in different stages of decomposition (Postmortem Interval (PMI) 2-186 days) from three human decomposition research facilities. The distribution of the PMI matched the outdoor sample of the original study as closely as possible. Each packet included photographs (averaging nine per packet) that minimally showed the overall body and close-up shots of the head/neck, trunk, and limbs. Where available, notes by the original on-site observer were supplied. Observers were provided the scoring tables from Megyesi et al.'s publication and instructed to follow only those descriptions, disregarding any modifications in use by individual facilities, and to return categorical scores for each bodily area (head/neck, trunk, and limbs), as well as overall TBS scores. When decomposition fit into more than one category, observers recorded both categories and averaged the contribution to TBS, as instructed by the original publication. Data analysis used a two-way random model Intraclass Correlation Coefficient (ICC) in Statistical Package for the Social Sciences (SPSS) (v. 17.0). The ICC is similar but deemed superior to a weighted Cohen's kappa.[4,5]

The TBS method shows "almost perfect" agreement between observers.[6] The overall single measure for Absolute Correlation Coefficient (ACC) is 0.859 for TBS and the Consistency Correlation Coefficient (CCC) is 0.878 for TBS. Assessment of the individual component categories shows variation in the correlation coefficients for each category, with head/neck being the highest (ACC=0.857; CCC=0.875), followed by the limbs (ACC=0.803; CCC=0.817), then trunk (ACC=0.690; CCC=0.720 — only "substantial agreement"). Education impacted the reliability, with individuals holding Master of Arts/Master of Science (MA/MS) degrees (ACC=0.989; CCC=0.923; n=8) or Doctor of Philosophy (PhD) degrees (ACC=0.896; CCC=0.940; n=2) having higher correlation coefficients than those with a Bachelor of Arts (BA) degree or less (ACC=0.801; CCC=0.841; n=6), although all education levels still fall in nearly perfect agreement. No difference in absolute or consistency correlations were observed when the participant sample was divided by experience, with the separation point being more or less than two years of experience (n=8, 8). The individual component scores followed similar patterns as the TBS score when examined under the filter of both education and experience.

Overall, trunk scores were the least concordant. Comments made by observers suggest this may result from difficulty identifying post-bloat release, greater variety of observed color change than described in the TBS system, and differences in decomposition of the upper (ribs) versus lower (abdomen) trunk. Common errors observed during data collation included simple issues, including use of non-existent categories for a particular body portion, probably the result of using the wrong table to make observations; recording point values for the category value (for example, C7 instead of C2, where 7 is the number of TBS points for category C2); and arithmetic errors in calculating TBS. All are easy to remedy (although the second problem could be difficult to parse for earlier stages of decomposition, where the categorical score and TBS point value are similar).

Thus, the TBS system is reliable, with near-perfect agreement between observers. Minor variation exists between observers based on education levels. The trunk category has the lowest level of agreement between observers and may provide an opportunity for improvement.

This study was conducted with the approval of the Southern Illinois University Human Subjects Review Committee, approval 14151.

References:

1. Megyesi MS, Nawrocki SP, Haskell NH. Using accumulated degree-days to estimate the postmortem interval from decomposed human remains. J Forensic Sci 2005;50(3):1-9.
2. *Daubert v. Merrell Dow Pharmaceuticals, Inc.*, 509.US.579,113S.Ct.2786, 125L, Ed.2d 469, 1993.
3. Committee on Identifying the Needs of the Forensic Sciences Community, National Research Council. 2009. *Strengthening Forensic Science in the United States: A Path Forward*. Washington, DC: National Academies Press, 2009.
4. Klales AR, Ousley SD. The Utility of Cohen's Kappa for Testing Observer Error in Discrete Data and Alternatives. 2010. *Proceedings of the American Academy of Forensic Sciences*, *62nd Annual Scientific Meeting*. Seattle, WA. 2010.
5. Haley SM, Osberg JS. Kappa Coefficient Calculation Using Multiple Ratings Per Subject: A Special Communication. Physical Therapy 1989;69(11):970-974.
6. Landis JR, Koch GC. The measurement of observer agreement for categorical data. Biometrics 1977;33:159-174.

Inter-Observer Error, TBS, Forensic Anthropology