RELIABILITY IN EVALUATOR-BASED TESTS: A MODELLING APPROACH FOR INTERPRETING INDICES OF RELIABILITY AND DETERMINING AGREEMENT THRESHOLDS

[1]Dylan Beckler, [1,2]Zachary Thumser and [1,3]Paul Marasco

[1]Laboratory for Bionic Integration, Department of Biomedical Engineering, Lerner Research Institute, Cleveland Clinic
[2]Research Service, Louis Stokes Cleveland Department of Veterans Affairs Medical Center
[3]Advanced Platform Technology Center of Excellence, Louis Stokes Cleveland Department of Veterans Affairs Medical Center

ABSTRACT

Indices of inter-evaluator reliability are used in many fields such as computational linguistics, psychology, and medical science; however, the interpretation of resulting values and determination of appropriate thresholds lack context and are often guided only by arbitrary "rules of thumb" or simply not addressed at all. Our goal for this work was to develop a method for determining meaningful interpretation of values, thresholds, and reliability based on a systematic alteration of the mean and variance within a normally distributed error signal, providing insight into the interplay between bias and error of a hypothetical rater population. As a basic metric for inter-rater reliability we selected Krippendorff's alpha. This is a versatile statistical tool for quantifying the agreement between multiple evaluators on sets of observations or measurements and it is highly flexible in handling multiple raters, missing data, and different scales of measure. We presented a video analysis task to three expert human evaluators and averaged their results together to create an initial dataset of 300 time measurements. We developed a mathematical model that then introduced a unique combination of systematic error and random error onto the original evaluator dataset to generate 4800 new hypothetical raters (each with 300 time measurements).We calculated the percent error and Krippendorff's alpha between the original dataset and each new modified dataset to determine the value envelope of inter-rater agreement. We then used this information to make an informed judgement of an acceptable threshold for Krippendorff's alpha within the context of our specific test. As a marker of utility we calculated the percent error and Krippendorff's alpha between the initial dataset and a new cohort of trained human evaluators, using our contextually derived Krippendorff's alpha threshold as a gauge of evaluator quality. We found that this approach established threshold values of reliability, within the context of our evaluation criteria, that were far less permissive than the typically accepted "rule of thumb" cutoff for Krippendorff's alpha. This procedure provides a less arbitrary method for determining a reliability threshold and can be tailored to work within the context of any reliability index.