West Virginia Department of EDUCATION
Office of Research

2012

NEW FROM THE OFFICE OF RESEARCH • OCTOBER

Findings from the 2011 West Virginia Online Writing Scoring Comparability Study

To provide an opportunity for teachers to better understand the automated scoring process used by the state of West Virginia on our annual WESTEST 2 Online Writing Assessment, the WVDE Office of Assessment and Accountability and the Office of Research conduct an annual comparability study. Each year educators from throughout West Virginia receive training from the Office of Assessment and Accountability and then hand score randomly selected student compositions. The educators' hand scores are then compared to the operational computer (engine) scores, and the comparability of the two scoring methods is examined.

Method of study. A scoring group made up of 43 participants representing all eight regions scored a randomly selected set of student essays using the appropriate grade-level WV Writing Rubrics. A total of 2,550 essays were each scored by two different human scorers to allow for comparison of human-to-human scores as well as human-to-engine scores. Four hypotheses were tested.

Findings. We first sought to determine the extent to which human scorers calibrated their scoring process to align with the automated scoring engine via a series of training papers. We found that calibration was generally quite good in Grades 5-11, but there was room for improvement in Grades 3 and 4. We also found that calibration rates were relatively static across the set of training papers. We next sought to determine the comparability of human-to-human and human-to-engine agreement rates. We examined both exact and exact/adjacent agreement rates (i.e., scores that were exactly matched or within 1 point of each other on a 6-point scale). Looking at well-calibrated human scorers, our analyses showed that, with few exceptions, both exact and exact/adjacent agreement rates were comparable for the human-to-human and human-to-engine pairs. Finally, we examined the average essay scores assigned by the automated scoring engine and those assigned by a sufficiently calibrated human scorer. Our analyses revealed that for four of the available grade levels there were no significant differences. However, for the remaining grades and for all grades in aggregate, differences were statistically significant. In these cases, the differences observed between calibrated human scorers and the automated scoring engine were equivalent to or less than approximately three 10ths of a point (.310) on a 5-point scale or approximately 2% to 5% of the available points, with human scorers typically scoring papers higher. This difference was deemed to be practically insignificant.

Looking at human scorers who had demonstrated their ability to produce well-calibrated scores during the training period, descriptive statistics showed that, with few exceptions, both exact and exact/adjacent agreement rates were comparable for the human-to-human and human-to-engine pairs.

Limitations of study. The human essay scores used in similar studies of automated essay scoring are generated by scorers for whom there is strong empirical evidence that indicates they are able to apply a validated scoring rubric in a consistent and valid manner. In our case, employing 10 training papers is likely not enough training to ensure our scorers become expert raters. Until the calibration process and measure are improved upon, agreement rates and differences in human and engine scores should be interpreted cautiously.

Until the calibration process and measures are improved upon, agreement rates and differences in human and engine scores should be interpreted cautiously.

Recommendations. We recommend improving the calibration process; examining new measures of calibration among scorers to assist in interpreting results; using multiple and different measures to examine agreement between scoring methods; and adding a qualitative research component to next year's online writing comparability study to examine teacher outcomes.

For more information, contact Nate Hixson, Office of Research (nhixson@access.k12.wv.us), or download the full report: *Findings from the 2011 West Virginia Online Writing Scoring Comparability Study* from the Office of Research website.