2013
NEW FROM THE OFFICE OF RESEARCH • SEPTEMBER

Findings from the 2012 West Virginia Online Writing Scoring Comparability Study

Student responses to the WESTEST 2 Online Writing Assessment are scored by a computer-scoring engine. The scoring method is not widely understood among educators, and there exists a misperception that it is not comparable to hand scoring. To address these issues, the West Virginia Department of Education (WVDE) conducts an annual scoring comparability study that compares scoring by trained human raters to scoring by the computer engine.

Method of study. This year, 45 educators from West Virginia participated in the study. Each scored a set of training essays and operational student essays that also were scored by the scoring engine. Each operational essay was scored independently by two human raters. Human raters' scores were compared to each other and to the engine.

Two research questions were posed: (RQ1) what is the level of calibration to the automated scoring engine that is achieved among human raters as a result of the training provided by the WVDE?; and (RQ2) what is the comparability of scores assigned by human rater pairs as well as between human-to-engine pairs?

Findings. Approximately 58% of human raters met three industry standard calibration criteria for calibration; the remaining 40% did not. Human rater pairs tended to provide the most consistent scores. However, in many cases we found that human raters were more likely to agree with the engine's scores than with each other's. When disagreements did occur though, human raters consistently scored student essays slightly higher than the engine. We believe this outcome should serve to mitigate some concerns that the engine scores student essays wildly differently from regular classroom educators or that the engine scores essays too forgivingly.

Limitations of study. We do not draw definitive conclusions about the consistency of the engine from the results of this study because so few raters met rigorous standards for calibration. However, we note that the test vendor has provided considerable evidence to establish the comparability of the scoring process based upon studies that use only human raters judged to be experts based upon industry standard criteria.

Recommendations. Continue to use the annual comparability study as a professional development experience for educators and additional data collection around educators' perception of the accuracy and fairness of scores assigned by the engine.

For more information, contact coauthor Nate Hixson, Office of Research (nhixson@access.k12.wv.us), or coauthor Vaughn Rhudy, Office of Assessment and Accountability (vrhudy@access.k12.wv.us), or download the full report: Findings from the 2012 West Virginia Online Writing Scoring Comparability Study on Office of Research website (http://wvde.state.wv.us/research/reports2013.html).

This year, 45 educators from West Virginia each scored a set of training essays and operational student essays that also were scored by the scoring engine.

Human rater pairs tended to provide the most consistent scores. However, in many cases we found that human raters were more likely to agree with the engine's scores than with each other's.

When disagreements did occur, human raters consistently scored student essays slightly higher than the engine.

West Virginia Department of EDUCATION
Office of Research