Toward Theoretically Meaningful Automated Essay Scoring

Randy Elliot Bennett
Educational Testing Service, Princeton, NJ
Anat Ben-Simon
National Institute for Testing and Evaluation, Jerusalem, Israel

Abstract

Automated essay scoring has the potential to reduce processing costs, speed up the reporting of results, and improve the consistency of grading. This study evaluated a "theoretically driven" method for scoring NAEP writing assessments automatically. The method would be usable for any future NAEP writing assessment conducted on computer or conducted with emerging technologies that allow handwriting to be digitized and translated to type. Such a method, if successful, could produce a means for NAEP to score essay responses automatically in a way that can be linked explicitly to the characteristics of good writing.

Existing commercial programs for automated essay scoring have generally used writing features that are empirically weighted to predict the scores of human raters. The selected writing features may or may not have any direct connection to writing theory. This study used variations of an existing commercial program, e-rater®, to compare the performance of three approaches to automated essay scoring: a brute-empirical approach in which variables are selected and weighted solely according to statistical criteria, a hybrid approach in which a fixed set of variables more closely tied to the characteristics of good writing was used but the weights were still statistically determined, and a theoretically driven approach in which a fixed set of variables was weighted according to the judgments of writing experts.

The research questions concerned (1) the reproducibility of weights across writing experts, (2) the comparison of scores generated by the three automated approaches, and (3) the extent to which models developed for scoring one NAEP prompt generalize to other NAEP prompts of the same genre. Data came from the NAEP Writing Online study (Horkay, Bennett, Allen, & Kaplan, 2005), which

included the responses of 1,255 8th grade students to two essays, and from the main NAEP 2002 writing assessment, from which 300 responses to each of four essays were employed. Weights were provided by two committees of writing experts.

Results showed that experts initially assigned weights to writing dimensions that were notably more similar across the two committees than to the empirically derived weights used by the hybrid approach. When one committee was shown the empirical weights and the other committee was not, the differences between the committees increased, with the committee shown the weights moving closer in its judgments to the weights of the hybrid approach. As a consequence, each committee's weights was used separately in the analysis.

The various automated approaches were compared with respect to their relations with human scores, their relations with other indicators, their functioning in NAEP reporting groups, and the resolution of large machine-human score discrepancies. The theoretical approach based on committee judgments informed by the hybrid's empirical weights generally did not operate in a markedly different way from the brute empirical or hybrid approaches. In contrast, many consistent differences with those approaches were observed for the theoretical approach based on the judgments of the committee that was not informed of the empirical weights. For example, this theoretical approach produced mean scores that were significantly lower than human scores; correlated less with human scores than did the hybrid version; had considerably lower exact agreement with humans than did either the brute empirical or hybrid versions; and had a lower between-prompt correlation than observed for human scores.

With respect to generalizability to other prompts, the theoretical approach based on committee judgments informed by empirical weights fared less favorably than the brute empirical and hybrid approaches, but usually by small amounts. In contrast, the theoretical approach based on the judgments of the committee not informed by the empirical weights showed more and larger differences. Should NAEP decide to use automated scoring in future online writing assessments,

empirical weights might provide a useful starting point for expert committees, with the understanding that the weights be moderated only somewhat to bring them more into line with theoretical considerations. Under such circumstances, the results may turn out to be reasonable, though not necessarily as highly related to human ratings as statistically optimal approaches would produce.