Machine Scoring of Student Responses on the CAT

In January 2017, the Center for Assessment and Improvement of Learning (CAIL) began investigating the use of machine learning to score student essay responses to the CAT. A large set of student tests administered at a variety of institutions across the United States whose scoring accuracy was checked by our center experts were transcribed to create a database of responses for each CAT question. Two sets of test responses were randomly selected from the database; a large primary corpus to develop the machine learning models, and a smaller test set of approximately 500 tests to be used test the generalizability of the machine scoring models.

In comparing the machine scoring model to the expert scoring for the total CAT score the percent error = -1.39%. This error is well within the acceptable margin of error (5.0%) set for evaluating the accuracy of institutional scoring sessions. Analysis of the range and variability of the total CAT score assigned to tests by expert scorers and by machine scoring showed little to no difference. The minimum score (0) and the maximum score (33) was the same for both the machine scores and the expert scores. Standard deviations were relatively similar for machine scores (5.86) and expert scores (6.02).