JEOC REVIEW

March 14, 2018

PART I: Comparative performance of Artificial Intelligence Engines in scoring state assessments

Source: This review is based on two papers by Mark Shermis. Shermis is an academic expert on artificial intelligence scoring of student's constructed responses to items appearing on state tests. Presently Dr. Shermis is Dean of the College of Education at the University of Houston Clear Lake. He also is a member of ODE's Technical Advisory Committee that advises the Department on assessment practices when advice is sought. The focal papers are (paper 1) Shermis, M.D. (2014). State-of-the-art automated essay scoring: Competition, results, and future directions from a United States demonstration. Assessing Writing 20:53-76 and (paper 2) Shermis, M.D. (2015). Contrasting State-of-the-Art in the Machine Scoring of Short-Form Constructed Responses. Educational Assessment 20:46-65.

Findings in paper 1. Nine scoring engines were compared using eight different essays with an average length of response from 94 words to 622 words. Sample sizes ranged from 918 to 1805 student responses for training the scoring engines and from 304 to 601 student responses for evaluating the scoring engines. All but one of the essays were scored using two human raters; human scoring is the baseline for comparison because humans have comprehension that the scoring engine does not have.

In general the average of assigned scores from all eight engines, including the one developed by the American Institutes for Research (AIR), are comparable to both the other engines and to human raters. Sometimes AIR's averages tended to be a small amount higher and sometimes a small amount lower. Exact agreement rates, a traditional measure of hand-scoring quality, were often higher for human to human scoring than for machine to human scoring; and AIR's engine was typical in performance. Comparisons based on Cohen's kappa – a measure of chance adjusted rater agreement – showed all scoring engines to be inferior compared to human raters. While these observations hold for the data, in general, there was some variation that may have been sensitive to rubric design. For example, it is much easier to get agreement on a rubric that scores 0, 1, or 2 than on a rubric that also includes a 3 and a 4. Shermis' conclusion is that scoring engines can be effective at scoring and may be necessary if a state implements multiple essay assessments throughout a school year.

Findings in paper 2. Paper 2 is similar to paper 1 except the student responses are much shorter and in the range of 22 to 58 words. While both papers source data from competitions to produce prize awards of as much as $100,000, paper 2 was a competition among individuals instead of corporate entities. Corporate entities chose to not participate to protect their proprietary code. To quote Shermis, "The winning public competitor ($50,000) was an Ecuadorian student studying data science as an undergraduate at the University of New Orleans. Second place ($25,000) went to a graduate student in Slovenia, and third place ($15,000) was awarded to French actuary in Singapore." Like paper 1, agreement rates and kappa coefficients favored human scoring over machine scoring.

Shermis' conclusions to paper 2 are written as commentary aimed at the PARCC and SBAC consortia:

> *With regard to recommendations for both PARCC and Smarter Balanced, we suggested that machine performance for the top vendors of automated essay scoring was ready to be used, pending additional validity studies, as a second reader for high-stakes assessments and possibly as a first reader for low-stakes assessments (Shermis, 2014). Presently that technology does not yet have the capacity to determine if a good argument (or conclusion) has been made though it can, in a rudimentary way estimate the degree to which one has made a coherent argument (Burstein, Tetreault, Chodorow, Blanchard, & Andreyev,*

2013). Our recommendation to the two major Race-to-the-Top consortia regarding short-answer scoring is that the technology has to undergo some significant additional development before it can be deployed operationally. In its current form, it might be used as a second reader on an experimental basis or phased in slowly after states gain significant experience with the technology.

The challenges for scoring short-answer constructed responses are quite different from that of evaluating essays. Current essay algorithms are focused on grading writing ability and content, whereas the emphasis for scoring short-answer responses is on grading content and response correctness. With regard to content, the essay writer has more writing space in which to establish relevant information, whereas the scoring algorithms for short-answer responses have to make their estimates on less information.

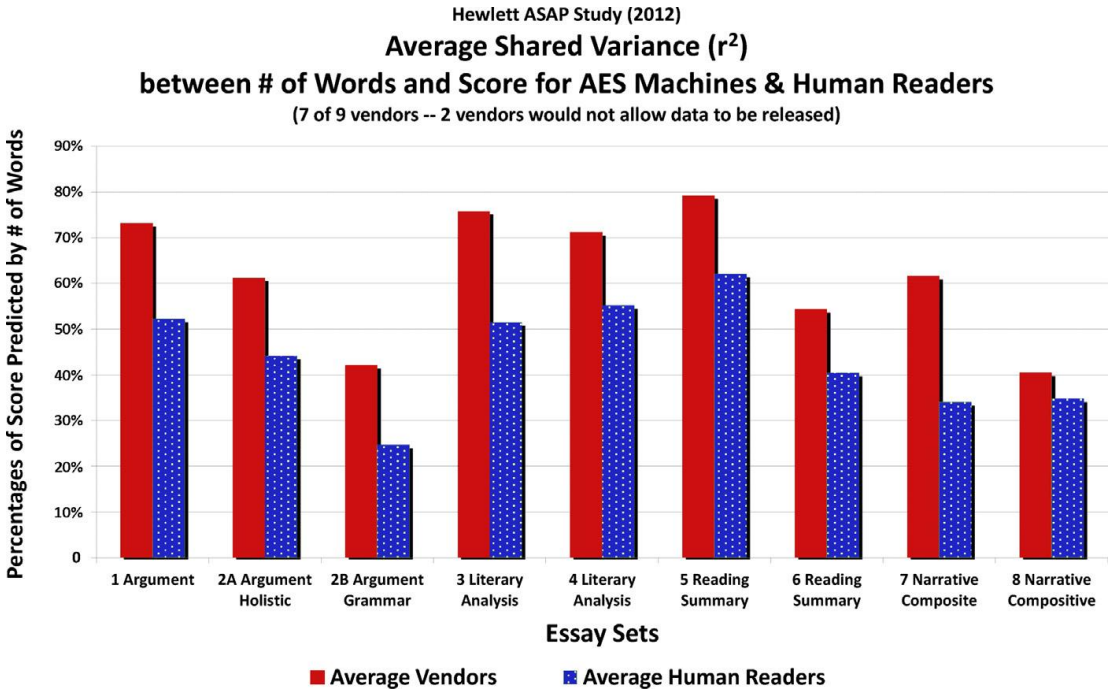
PART II: A rebuttal on the capabilities of Artificial Intelligence Machine scoring of test essay responses

Source: The focal paper, by Les Perelman, "When 'the state of the art' is counting words". (2014, Assessing Writing, Vol. 21) is a response to Shermis' paper, "State-of-the-art automated essay scoring: Competition results and future directions from a United States demonstration." (2014, Assessing Writing, Vol. 20). Dr. Perelman is the now retired director of Writing Across the Curriculum at MIT and a consistent critic of the use of Artificial Intelligence Machines (AIMs) for scoring of essays. He, along with students from MIT and Harvard, developed a computer program called BABEL that can generate gibberish capable of receiving high scores from AIMs.

Perelman's contention. Perelman and Shermis agree that there is considerable economic pressure to replace human rater scoring of test constructed response items with AIM scoring. Some of that pressure stems from the increased use of essay item types that are thought to be responsive to common core standards. As stated in Perelman, "…there were large incentives … to believe Professor Shermis' assertion in his article in this journal that 'Automated essay scoring appears to have developed to the point where it can consistently replicate the resolved scores of human raters in high-stakes assessment' (Shermis, 2014, p. 75). Unfortunately, the data provided in that article and in the link to the raw data provided do not substantiate this claim."

Perelman shows the relationship between length of response and test scores reproduced as Figure 1

Figure 1. From Perelman [Red is AIM scored; blue is for human scored]



As shown in Figure 1, as much as 80 percent of the variance in AIM scores is associated with the number of words in the student response (see essay set 5, Reading Summary). The peak for human raters is just over 60 percent (also set 5). Perelman suggests that the length of response and score interaction occurs "…only for timed-impromptu writing, a genre that does not exist outside of the standardized writing test." He also points out that agreement between human raters remains superior to agreement between human raters and machines – facts shown in Shermis' paper and speculatively attributed, by Perelman, to the capacity of humans to actually understand the response. Understanding, Perelman asserts, is more important when the task is based on a literary passage instead of when based on informational text.

Observations.

1. The scoring engines described do not "read" the responses in the sense-making way that a human might. Instead the scoring engine is trained to identify patterns in responses that are associated with particular human assigned scores found in the training papers and then predict the score that a human might assign to the paper being scored. Because of the way the scoring engines work, as described above, student responses that do not exhibit the same patterns shown in the student responses used for training may be scored differently than if a human (that tries to make sense of the response) had scored the paper. Careful selection of student responses used for training and having a sufficient sample can mitigate this problem to some extent.
2. The assertion that there is an association between length of response and score is true for both human and AIM raters. Does that mean that a higher association for AIM scoring indicates a defect? Isn't it true that more elaboration in a response provides more content as evidence to justify a score? According to the data presented, humans and machines agree on that point.
3. At least two factors are worth some consideration. First, it is clear that Perelman's BABEL experiment producing Jabberwocky-like sentences (some even less meaningful) of varying length increase the opportunity for a response to use words in ways that might fool the AIM. Second, even though the AIM does lack reasoning and common sense it is not necessarily true that students are trying to game the AIM when writing their responses.
4. The source of authority for scoring student-produced constructed responses is human because the AIM "learns" by characterizing student responses and the scores already assigned by humans. [The technical approach is called "latent semantic analysis".] If the use of a test item is predicated on ability for human raters to agree, should the criterion be the same for AIM scoring? Why should agreement between AIM and human raters be higher or lower? Are there practices for writing items where human to AIM score agreement is enhanced?
5. For states that implement assessments that include multiple writing samples, Shermis states the AIM scoring can be effective (Paper 1). Shermis recommended (passage from paper 2) using human raters for high stakes tests with machine scoring as the second or confirmatory rater. In Ohio, the high stakes tests include the high school end-of-course tests and the third grade English language arts test. Ohio used two raters for every Ohio Graduation Test constructed response item scored throughout the life of the test.
6. From personal correspondence with Dr. Shermis: *[Shermis] I just completed a study for NCES on NAEP items and the results from the two vendors (ETS and Pearson) were even better. Most vendors can flag anomalous papers (off-topic, prompt copy, and threats) for human review. Based on the AIR results, it looked like most of the PCM papers are with third-grade writers who may simply be at a scaffolding point where copying the prompt is part of a routine they learned from their teacher. There are vendors that can flag this while the student is writing. So for example the ETS product e-rater can provide qualitative feedback during the writing process that would tell the writer that s/he is copying the prompt and that there is not a unique enough response.*
7. Unstated in both the Perelman paper and in the paper by Shermis is the effect of "condition codes". Condition codes indicate why no score was assigned to a student response. Of particular interest is the condition code for plagiarism. There are at least two cases where the plagiarism code might be assigned. The first is where two or more students provide responses that are so similar that they suggest copying among students. The second is the assignment of a plagiarism code (like Prompt Copy Match or PCM) to a student response because the response contains too much content from the prompt or the passage. Neither paper indicated the extent to which human raters agree with AIMs on condition codes. It seems possible that some condition codes might have the effect of raising human to AIM agreement rates by "coding out" ambiguous responses where humans and machines might disagree on a score.