Cambium Assessment

Communicating to the Public About Machine Scoring: What Works, What Doesn't

Mark D. Shermis
University of Houston—Clear Lake

Susan Lottridge
CAI

Paper presented at the annual meetings of the National Council of Measurement in Education

Toronto, Canada, April 7, 2019

Abstract

This paper documents six case studies about how to, and how not to, communicate a testing entity's transition to machine scoring. Data are drawn from four U.S. state-, one Canadian Province-, and one Country's testing programs. Based on the analysis of the six cases, several tentative recommendations can be made.

Keywords: *machine scoring, automated essay scoring, state testing programs*

Communicating to the Public About Machine Scoring: What Works, What Doesn't

In his last major contribution to automated essay scoring, Page (2003) discussed three objections to machine scoring which he labeled as *humanistic*, *defensive*, and *construct*. The *humanistic* objection stipulates that writing is a unique human skill and cannot be evaluated by machine scoring algorithms. *Defensive* objections deal with concerns about "bad faith" or off-topic essays and scoring algorithm vulnerabilities to them. The *construct* argument suggests that what the human rater is evaluating is substantially different than what machine scoring algorithms used to predict scores for the text. Therefore, it may be possible for humans and machines to come to similar score assignments, but evaluate different things.

All of these objections, and some others, have made testing entities reluctant to engage in a wide-spread implementation of machine scoring for either short-form constructed or essay-length artifacts. This paper uses six case studies, three in which the implementation of machine scoring went well, and three where the implementation was blocked or substantially hindered. The purpose of the paper is to help identify steps that can be taken to address objections to implementing machine scoring on a large-scale basis.

Recent History

Perhaps more than anything else, the so-called Hewlett Trials (Shermis & Hamner, 2013) stimulated interest in the application of machine scoring for high-stakes assessment. In the essay scoring trial, eight commercial vendors and one university laboratory participated in a blind study which evaluated eight different essay prompts across multiple genre. For the first part of the trial, the competitors were given data sets containing essay text and at least two human ratings. In the second part of the trial, the competitors were simply given the text and asked to predict human-rater scores. The results showed that, for the most part, machine score predictions were as reliable, and sometimes even more reliable, than human ratings. On that basis, several large testing entities began to investigate machine scoring for high-stakes essay writing. The notable exception was the state of West Virginia which began a version of machine scoring in 2005.

West Virginia

Rich, Schneider, & D'Brot (2013) documented one of the first positive experiences with machine scoring done on a statewide basis. In 2005 West Virginia was one the first states to implement automated essay scoring using software technology that combined both formative and summative assessment. From 2005-2014, WVDE used automated essay scoring as part of its formative and summative assessment programs. From 2015-2017, the state used hand-scoring only. In 2018 WVDE reinstated the use of automated essay scoring with a new vendor (Cambium Assessment, Inc., or CAI). The state has been known for writing innovations since 1984 when it established a stand-alone assessment of writing.

West Virginia and the vendor at the time, CTB, took a three-pronged approach in implementing the testing program. First, it had a large-scale state summative assessment that was scored using CTB's *Bookette* scoring engine in grades 3 and 11. Second, it connected formative writing assessments to summative writing assessments. Finally, it maintained a continuing professional development program for writing teachers that incorporated elements of machine scoring as part of the curriculum.

CTB bolstered their cause in West Virginia through three empirical studies. Rich et al. (C. S. Rich, Harrington, & Kim, 2008) investigated the relationship between the year-end, stand-alone summative assessment administered in 2007 and the formative assessment use of *Writing Roadmap™ 2.0* using a quasi-experimental design. *Writing Roadmap 2.0* is a product that combined the scoring engine *Bookette* and an electronic portfolio system that administers

prompts and provides an enriched writing environment. *Writing Roadmap* users and non-users were matched based on the performance levels for the WESTEST RLA assessment. The study looked at differential performance based on community type of rural versus urban schools, gender, and ethnicity to provide insight into fairness issues in the use of AEE in West Virginia classrooms. Positive score gains on the state writing test were found for students who used *Writing Roadmap* compared to students who did not. The largest gain was found for the lowest performing group with an effect size of 0.7. The first study concluded that writing technology could have a significant impact on writing outcomes if well-integrated with the writing curriculum.

The second study was conducted on five years of integrating automated essay scoring in classroom assessment and summative assessment, White et al. (2010b) investigated the impact of *Writing Roadmap 2.0* on WESTEST 2 OWA scores. Students who had completed five or more *Writing Roadmap* essay assignments during the 2008–2009 school year were randomly selected for the study. As with the first study, students were matched to *Writing Roadmap* non-users based on grade level, geographic location, and socioeconomic status. The final sample in the study included 8,430 randomly selected students in the treatment group and 8,430 students in the comparison group. The summative online writing test score means from the treatment groups and the comparison groups were compared and showed effect sizes from 0.17 for grade 9 to 0.59 for grade 4. The effect sizes tended to be larger among elementary school students compared to middle school and high school students. Based upon the findings, White et al. (2010a) recommended continued use of Writing Roadmap 2.0 as a formative assessment tool in West Virginia.

The following year, White et al. (White, Hixson, & Whisman, 2011) analyzed the impact of automated essay scoring usage in classrooms. In that study the variance of students' prior year academic performance in RLA. A linear regression model was developed to predict 2010 online summative writing scores using five variables: number of writing assignments in *Writing Roadmap*: male gender, low socioeconomic status, special education eligibility, and 2009 RLA scale score. These variables were chosen due to observed performance gaps in writing assessment among these subgroups of students. Samples of 5% of the population were randomly selected from grades 4 through grade 11 based upon students who had taken the WESTEST 2.

Bivariate correlation and multiple regression analyses were performed on the sample data of 8,577 students from grade 4 through grade 11. All variables were significant predictors in the multiple regression equation and accounted for approximately 37% of variance in students' summative writing scores. White et al. (White et al., 2011) concluded that even after controlling for students' prior academic achievement and gender in a representative sample, a modest but statistically significant positive relationship was found between Writing Roadmap™ 2.0 usage and students' subsequent online summative writing scores for grades 4 through 11, with the exception of grade 10.

Presently, West Virginia administers the West Virginia General Summative Assessment and writing tests are scored using CAI's *Autoscore* engine as a first reader. Hybrid (automated-human) score is used whereby the first 500 responses to a prompt are routed for human verification scoring and responses with low confidence indices (as produced by the engine) are also routed for human scoring. West Virginia also uses *Autoscore* in their interim and benchmark assessments. For these assessments, response are routed to *Autoscore*, which provides a predicted score. Responses with low confidence scores are then flagged for teachers to review and provide a verification score.

With the renewal of its use in 2018, the state undertook an effort to build educator trust in the approach. WVDE created an action plan that was multi-faceted and focused on raising awareness

about the scoring and on obtaining feedback from educators, including creating an advisory panel of WV educators; developing resources about the WVDE writing rubrics with annotated student exemplars to illustrate the application of the rubrics; creating more detailed descriptions of how the scoring engine works, and how the engine is used to score writing in WV; creating a description of the scoring condition codes and their thresholds; and conducting a writing study to recommend thresholds for two condition codes (July 2018), and to score responses and compare their scores to the automated engine scores (November 2018).

The threshold study conducted in July 2018 implemented a 'standard-setting-like' workshop to set thresholds for the proportion of copied text and for the minimum number of words considered eligible for rubric-based scoring. This workshop was developed because of educator concerns, particularly about the threshold around the proportion of copied text. Teachers were divided into grade-level groups (3, 5, 7) and reviewed responses to a single prompt for that grade. In separate iterations, teachers reviewed booklets of student responses and set cuts for the minimum word threshold and for the proportion of copied text threshold in three rounds. In the first round, teachers reviewed response independently and set thresholds. Results from this round were collected and presented to the group. In round 2, teachers then reviewed the results and changed (or retained) their thresholds based upon the discussion. Impact data were presented and discussed using round 2 teacher thresholds, and then in round 3, final threshold recommendations were made. Recommendations were based upon the median threshold value of the teachers in the group. Changes to the thresholds were: 1) a slight increase in the proportion of copied text threshold (Grade 3: 74%; Grade 5: 77%; Grade 7: 72%); 2) a substantial change in the number of words thresholds (was 11 words but was changed to accommodate grade-level thresholds (Grade 3: 14 words, Grade 5: 19 words; Grade 7: 28 words). The state reviewed the panel recommendations alongside impact data and used the average across the grade thresholds for the proportion of copied text (74%), and imputed grade-level thresholds for each grade assessed (Grade 3: 14 words; Grade 4: 16 words; Grade 5: 19 words; Grade 6: 23 words; Grade 7: 28 words; Grade 8: 34 words).

The 1.5 day teacher scoring workshop conducted in November 2018 was modelled after the prior workshops using the *Bookette* engine in 2013 and 2014. The workshop had two main goals: to provide teachers with training on the WV writing rubrics and the scoring processes used in the General Summative Assessment and Interim Assessment programs; and, to examine the comparability of the teacher agreement with CAI's automated scoring engine to the agreement of participating teachers. In this workshop, on day 1 teachers underwent mini-trainings on hand-scoring on a single prompt and each scored about 30 papers. Teachers also were introduced to how automated scoring works and how it is implemented in West Virginia assessments. On the second half-day, the teachers discussed their experiences as scorers and were presented their agreement with one another, and with Autoscore. At the end of the workshop, teachers completed a survey that asked about their understanding and confidence in the rubrics, hand-scoring, and automated scoring. Professional hand-scores from Measurement, Inc. (MI) conducted the training, with training leads supplied at each grade (3 through 8), in the morning and teachers scored the responses using an on-line hand-scoring platform in the afternoon. The survey indicated that teachers agreed or strongly agreed that they understood and had confidence in the rubric and exemplars and how hand-scoring is used in the summative assessment. Teachers were very happy with the workshop overall and felt that it provided them with the opportunity to better understand the rubrics and how to apply them to student writing. Most agreed that they understood and had confidence in the automated scoring engine. That said, most teachers did not qualify (i.e., meet the 70% exact agreement rate in each essay dimension) on the reduced qualification sets (5 essays), suggesting that the training did

not result in industry-standard performance. This result is not surprising given the short duration of training. Teachers were allowed to continue on to scoring regardless of their qualification results. In terms of scoring, teachers were almost always within one score point with one another and the engine. The agreement of the teachers with the engine was slightly lower than that with one another, and QWK values of the engine with teachers were within .1 for 13/18 (72%) of items and traits. A review of the items with out of range quadratic weighted kappa (QWK) differences indicated that teachers were more lenient in two dimensions (purpose and organization, and evidence and elaboration) and were more severe in one dimension (conventions). When mean scores did differ, the engine scores tended to align with those in the sample used to validate the engine originally. These aggregate results suggest that the workshop was an excellent professional development tool for the teachers; however, the design was not sufficient to allow for meaningful interpretations of differences in scoring of the teachers with themselves or with the engine given the inadequate time allocated to training and qualification. As a result, the state may elect to spend more time on training and qualification to help ensure alignment with the intended application of the rubric, which would allow for a more rigorous evaluation of the teacher scoring relative to the engine. West Virginia has studied machine scoring for essays more than any other U.S. state.

Louisiana

Louisiana-LEAP (Louisiana Educational Assessment Program). Louisiana's implementation of machine scoring for their LEAP test came about as a quality assurance step for the scoring of their writing program. The Pacific Metric's (now ACT) *CRASE* scoring engine was used as a second reader on 100% of responses that are typed into the on-line administration of the writing program. This approach has three advantages. First, the machine prediction model establishes a baseline upon which human rating performance can be calibrated. If there is a problem for scoring the prompt with human raters, the proportion of second reads can be increased to better monitor scoring reliability. Second, the approach can detect artifacts like rater drift and rater bias, and can prompt scoring administrators to take appropriate measures to address these artifacts. In particular, the engine scores can be used to flag and recalibrate readers early in the testing window – when training is most critical – and then throughout the testing window and across testing windows. Finally, the process can forecast the likelihood/proportion of aberrant responses so that the scoring rubric may be adjusted if necessary. Moreover, it can identify aberrant responses for human follow-up.

The LEAP assessment program has been supported by the Data Recognition Corporation (DRC) since 2016. DRC uses *Project Essay Grade (PEG)* automated scoring engine and uses the engine as the primary scorer, with human raters providing monitoring reads at 20% rate for the first year.

The *CRASE* engine was also used to score responses to Louisiana's PASS program in 2008-2013, which was on online formative assessment available to any Louisiana student at no cost to the student or school. In this program, tests were available at grades 3-11 in science, math, social studies, and English language arts. Many tests had open-ended items, which were scored by a vendor. Increasing student use of the program translated into significant hand-scoring costs, and *CRASE* was used to score the most frequently-used open-ended items as a cost-saving measure.

Utah

Utah has been using some automated scoring since 2008 in summative and formative assessment (Rathke, Palermo, & Wright, 2018). Automated scoring was introduced for a few reasons: 1) to save money on hand-scoring costs; 2) to provide consistent scoring, particularly in situations where teachers are providing scores; and, 3) to enable faster return of scores to better enable the use of the data in instruction.

In the summative program, automated scoring was introduced Spring 2008 and has

been used consistently, although in different combinations with human scoring, throughout the years. In Spring 2008, automated scoring was implemented in grades 5 and 8. The automated scores were combined with a 100% second human read, and scoring results were provided the following fall, and as processes improved, were reported later in the spring. In Spring, 2014, automated scoring was expanded to other grades and included all grades between three and eleven. At this time, automated scoring was the only score provided. This approach enabled the state to returned scores immediately to teachers, who were pleased to not have to wait for their reports. In spring 2015, a 20% random second read by human scorers was instituted. In spring 2017, the automated scoring engine had functionality added to capture types of gaming or aberrant responses, and functionality added to accompany each score with a confidence index. During this year, human verification reads were conducted on responses that the engine had low confidence in. In spring, 2018, the amount of writing was reduced in the overall assessment program, with writing no longer assessed in grades 9 and 10 and one writing prompt (from a total of two) removed from the other grade-level assessments. The state switched vendors for the 2018-2019 academic year, and is now assessing writing in grade 5, 8 and high school. Automated scoring is implemented only in grades 5 and 8.

Utah uses automated scoring in other assessment programs as well. For formative writing assessment, the state uses Utah Compose, a Measurement, Inc. product. This assessment is available to all teachers but is not mandatory. It was introduced in 2008 in the state and made available to all teachers in 2013. Utah Compose consists of an electronic portfolio system whereby teachers can make writing assignments, students can create their responses, and engage in a write-revise cycle of editing through feedback from *Project Essay Grade*. Utah teachers are encouraged to use this program as a way to incorporate technology with language arts instruction. It is still used today and the state plans to continue using it for the next ten years. Automated scoring was introduced in their benchmark assessment in Fall, 2016 and continues today. And automated scoring was introduced in their interim assessment in Fall, 2015 and continues today. The interim assessment program is, like the formative assessment, optional for teachers and intended for instructional feedback and improvement.

On the whole, automated scoring seen by the state as providing more benefits than drawbacks however the transition wasn't completely smooth. One key issue experienced was in the application of condition codes in the summative assessment to manage gaming and aberrant responses, particularly when automated scoring was used as the sole scoring source and used across grade levels. The main problem with relying on the machine scoring prediction model is that the technology is vulnerable to bad faith essays—ones in which students deliberately try to "game" the technology. For instance, Utah had one instance where a student submitted an entire page consisting of "b"s and was able to obtain a good score. Another strategy used by some students was to write a very good paragraph and then copy it four of five times. As the pattern of artifacts manifests itself, the algorithm is updated to detect them in a perpetual cycle of "cat-and-mouse" (Smith, 2018).

The state worked with the vendor (CAI) at the time to implement filters that capture gaming and aberrant responses. Filters were rolled out in the interim test soon before the spring 2017 assessment, although they were intended to be rolled out in the prior fall (2016). The application of these filters resulted in larger proportions of students receiving 0 scores – sometimes as much as 12-14% in the lower grades. The key concern was a filter that flags responses in which 70% more of the text is a direct copy from the stimulus material. The use of this filter coincided with the release of source-based prompts that required that students read stimulus material and develop an argument with evidence that used the material. The filter that captured the proportion of copied text indicated that many students were writing essays that

consisted primarily of copied text. This was a significant public relations issue in the state, in part because it highlighted a disconnect in how educators were instructing students how to use source materials and how the inclusion of source material in responses are scored. To respond to the issue, the state took two approaches. First, the state chose to remove the filters and re-score the summative responses that year without them. Second, the state developed teacher training materials to illustrate to the teachers how much copied text was appropriate for source-based prompts. This included the release of two prompts as training materials and to provide exemplars of various levels of the proportion of copied text so that teachers could see what various proportions looked like. The following year (2017-2018), when the condition codes were rolled out for the interim and summative tests, there was general widespread acceptance by teachers.

The factors assisting the adoption of automated scoring in the state were multi-threaded. First, the transition of the state testing program from paper to online testing went smoothly. There were no major outages, it was seen as working well, and students liked it. There was initial concern about the keyboarding skills of 3rd and 4th grades students impacting performance but these concerns were allayed once scores were returned for these students. This success helped to build trust in online assessment and scoring of writing. In fact, the state summative assessment program was the last program to go online, so schools, teachers, and students were familiar with the processes of online testing from their experiences in formative assessment. Second, the state and teachers were happy about their ability to faster score reports, as it meant that the assessment results could be used. Teachers saw that the implementation of automated scoring enabled this change, as they used to get scores from the spring assessment in the following fall, and then to get them soon after testing. Third, the formative and benchmark systems enabled teachers to see prompts, rubrics, student responses, and scores all together. Up until this point, the scores associated with responses were not available to teachers. Teachers liked having access to this data, and also generally agreed with the scores produced by the engine. Having this level of transparency in the formative and benchmark assessment helped to build trust in automated scoring.

There are several implementations of machine scoring that either went poorly or encountered setbacks.

Ohio

Ohio is an interesting case. It conducted a modestly successful pilot, but failed to fully brief the State School Board about the nature of the pilot, its purpose, and how machine scoring works. In its first year of operational use, there was a difference in scores between human and machine scored essays for third-graders. The ELA assessment for this grade level in Ohio is of a high-stakes nature since children who fail to meet score thresholds can be held back from grade-level promotion. In Columbus city schools, for instance, 48.04% of third-grade students earned a score of zero on the writing portion of the exam whereas only 36.3% received this score before machine scoring was introduced. It turns out that the artifact that influenced the machine scoring algorithm was the amount of the prompt material that was *copied* in an answer. Some educators argued (rightly or wrongly) that this is how they taught very young students to structure their response—in contrast to using their own words. They argued that doing so was mechanism for increasing the fluency of the response with a population that was unfamiliar with academic protocol. The practice of copying the response was diminished in higher grade levels. Human raters appeared to be more tolerant of copying or less sensitive to it.

Here is a situation where an operational pilot would have discovered the discrepancy and there would have been time to formulate a modification or at least limit the impact of scores that were at variance with human raters. The miscue here appears to be attributable to one of three sources: (1) in the rhetoric of what human

raters say they are doing (i.e., devaluating when copying is present) and how they are actually scoring. Alternatively, (2) the algorithm is overly sensitive to the situation when exact prompt wording is also phraseology that would also be used in common expression or thresholds could be tuned so that it better reflects human judgement around the appropriate percentage of copied text. And, (3) tying in the use of such an approach with teacher professional development. Regardless, it would have been relatively trivial to adjust the scoring algorithms to meet some sort of negotiated stance which is actually how the situation was ultimately resolved. However, educators critiqued the situation as reflecting an inherent problem with the technology while the state department portrayed this as way of getting more accurate compliance with the instructions to devalue copying.

There was some significant resistance in moving forward with machine scoring until these issues were resolved and the state is now re-evaluating its position on machine scoring.

Alberta

In 2014 the Province of Alberta ran a modest pilot study on automated essay scoring with LightSide Labs using *LightSide* for scoring high stakes essays. *LightSide* was the only non-commercial product that participated in the vendor demonstration in the Phase I Hewlett Trials (2014a). It performed well and is the only fully functional automated scoring system in the public domain. *LightSide*'s main drawback is that it employs a variety of empirical predictors that do not necessarily parallel traditional writing traits. This makes it difficult to explain to lay individuals how writing models work and what exactly differentiates one model from another. Most commercial vendors employ NLP routines to tease out characteristics that lay audiences can relate to (e.g., grammar errors), though this information does not necessarily correspond to significantly better prediction models (Shermis, 2018).

The results of the pilot showed that *LightSide* performed at least as well as human raters. However, the Alberta Teacher's Federation asked Dr. Les Perlman to weigh in on the study's results. Dr. Perelman, is a retired professor from the Massachusetts Institute of Technology and a perennial critic of AES technology. Dr. Perelman was a vocal opponent of the Hewlett Trial study results with a number of criticisms that reflected a lack of understanding of how large-scale empirical research is conducted to basic logical flaws. For example, it was his opinion that the average essay length on some of the statewide essays was not long enough to qualify as "essays", even though the participating state departments of education had been running these essay writing programs for years. The Hewlett Trial study was tasked with assessing the feasibility of evaluating essays in whatever form they came, and was not in a position to dictate to states how they created their assessments. He also criticized the results because of their modest correlation with word count ($r = .66$). What he fell prey to was the classic correlation fallacy—that correlation does not mean causation (Shermis, 2014b). The correlation is not with word count as he mistakenly believed, but rather the underlying trait of *fluency*. Peter Elbow, in his classic text *Writing without Teachers*, describes the importance of fluency in the instruction and assessment of writing (Elbow, 1973). His developmental model is geared to encouraging the writer to produce more text and then edit it. Because of the critical role of fluency, length of the document will always be an important correlate to an assigned score. Taking a cognitive perspective on writing, McCutchen, Testke, and Bankston (2008) write: "Fluent text production can influence the writing process both directly and indirectly because inefficient text production can consume [cognitive] resources that might otherwise be devoted to higher level processes such as planning and revising" (p. 457).

One of the main criticisms that Dr. Perelman leveled in the Alberta pilot was the notion that some of the older students could easily "game" the system by writing non-sensical essays that

perhaps had good sentence structure and grammar, met some minimum length threshold, and used sophisticated vocabulary. As was mentioned above, "gaming" the system is a vulnerability cycle that parallels the game of "cat and mouse", but its actual prevalence is low. The only well-documented attempt to game a system was with the GRE writing Program in China where some examines added "shell text" (well-written memorized text that was marginally related to the prompt topic) that apparently impacted *e-rater*® scores. This vulnerability was corrected by a reframing of GRE writing prompts. Otherwise, gaming is something that could theoretically happen in the same way that your car could theoretically blow up, but just does not happen in practice. The conventional wisdom is that it is easier to write a good on-topic essay that it is to invent a discombulated off-topic essay that meets all the criteria for good writing.

Automated essay scoring in Alberta has essentially been put on hold for the foreseeable future.

Australia

Early work on machine scoring was successful in Australia, but the testing agency was outmaneuvered by a Teacher's Federation with an agenda to oppose machine scoring.

It all began well enough. ACARA, the Australian Curriculum Assessment and Reporting Authority, commissioned a study with four testing vendors [Measurement, Incorporated; Pacific Metrics (now part of ACT); Pearson; and MetaMetrics] to create scoring models for four sets of essays (year levels 3, 5, 7 and 9) that was part of a NAPLAN persuasive writing task (ACARA NASOP Research Team, 2015). A convenience sample of N = 1353 essays was collected for each grade level and randomly divided into three sets: a training set (N= 674), a test set (N = 340), and a validation set (N = 339). The training test sets consisted of both essay text and scores from two raters, while the validation set consisted of essay text only. Vendors could model with the training set, and test the fit of their models with the test set. The validation set was used by ACARA to judge model agreements with human raters using a variety of measures, but relying primarily on quadratic weighted kappa. This study was roughly modeled after the Hewlett Trials (Shermis, 2014a) which compared eight commercial vendor software packages across eight state testing program essays at multiple grade levels. All four vendors in the ACARA study also participated in the Hewlett Trials. Those results showed that a number of vendors had scoring engines that performed as well, and sometimes even better, than human raters. Given the appropriate caveats for continuing to research aspects of fairness and validity, the recommendation then was to proceed with caution on machine scoring of essays that were similar to those being contemplated for Common Core State Standard assessments.

Students typed in their essays directly into the computer and resulted in average essay lengths of 118.1, 229.6, 342.2, and 371.1 words for years 3, 5, 7 and 9 respectively. Each essay was marked by two trained human raters on ten different traits. The median raw score awarded to students was 19.2, 26.0, 30.8 and 33.3 for years 3, 5, 7 and 9 respectively. Vendors were then given the two data sets (training and test) with both scores and text. The validation set was provided with text only. A detailed description of how vendors approach the construction of their prediction models is given in Shermis (2014a). It is important to note that three of the four vendors use human-rated essays for model construction while one (MetaMetrics) employs a pre-existing model with strong assumptions grounded on the Lexile Scale for Reading (Burdick et al., 2013).

Based on the Total Score, κ_w between Marker 1 and the four AES engines ranged from .73 to .76 and from .72 to .82 with Marker 2. The relationship between Marker 1 and Marker 2 was κ_w = .79. The relationships between the markers and each of the ten traits had similar ranges, sometimes a bit higher and sometimes a bit lower depending on the trait. While these results were not as impressive as in the Hewlett Trials, there were three important differences: (1)

most of the Hewlett Trial essays were holistically scored; those that were scored based on traits tended to have lower quadratic weighted kappas, primarily because the range of scores was greater; (2) the sample sizes in the Hewlett Trials were larger, allowing for more stable estimates; (3) the Hewlett Trial essays included more content-based scoring which obtained better results than with persuasive essays.

Overall the study results suggested that machine scored essays were feasible given additional research on fairness and validity concerns. The ACARA psychometric oversight committee endorsed the effort along with a number of well-known Australian measurement and writing experts (ACARA NASOP Research Team, 2015).

Enter the New South Wales Teacher's Federation. They commissioned a review of the ACARA study by Dr. Perelman. Perelman's unpublished critique listed a number of potential flaws that had no evidentiary basis (Perelman, 2017). For example, he suggested that because the essays were scored by machine algorithms students would not have a legitimate audience for which to write, as if students couldn't imagine a particular audience in a persuasive essay task. There was no empirical evidence that this was a problem. He rightly suggested that computers could not assess creativity, poetry, or irony, or the artistic use of writing. But again, if he had actually looked at the writing tasks given students on the ACARA prompts (or any standardized writing prompt), they do not ask for these aspects of writing—most are simply communication tasks.

Perelman's third criticism focused on "weaknesses in grammatical analysis". This criticism was based on an analysis of a review of a Noam Chomsky article published in the *New York Review of Books* that evaluated a vendor that was not even part of the ACARA study (Perelman, 2016). There was no attempt to actually use data from the study to identify grammatical misspecifications. His fourth criticism was leveled at potential unfairness in the application of machine scoring technology. He referred to an ETS GRE study in China (Ramineni, Trapani, Williamson, Davey, & Bridgeman, 2012) where some candidates obtained slightly higher scores using ETS's e-rater technology than they got from human raters. However, if he had dug deeper, he would found that this discrepancy arose from candidates' use of "shell text" (i.e., memorized text that was tangentially related to the essay topic). In this case, human raters were instructed to ignore the shell text, but ETS had no effective plagiarism subroutines to screen it for e-rater. This operational problem has now been fixed for the GRE testing program. Again, it involved a vendor that was not part of the ACARA trial. Finally, he suggested that all the vendors were subject to gaming by students, again using evidence from a scoring engine that was not part of the ACARA study. Shermis, Burstein, Elliot, Miel, & Foltz (2015) examined the literature on so-called "bad faith" essays and concluded that it is possible for a good writer to create a bad essay that gets a good score, but a bad writer cannot produce such an artifact. That is, an MIT technical writing professor can write a bad essay that gets a good score, but a typical 9th grader does not. The extensiveness of bad faith essays is like voter fraud—there are some people that are convinced it exists in great numbers, but there is little evidence to show for it.

Dr. Perelman goes on to dispute the results of the Hewlett Trials in a section of the ACARA report that he labels "inaccuracies". That is, he attributes his denigration of the first study to elements of the latter report. So basically, Perelman criticized the ACARA study without pointing to *any evidence* in the actual study to justify his claims. It was a simply a hack job.

Predictably, the NSW Teacher's Federation used the study as a vehicle to sow doubt in the lay press about the veracity of the foundational work in this area. This was unfortunate in that it probably resulted in a two-year delay for the operational use of machine scoring in NAPLAN essays.

There are a couple of entities where implementation of machine scoring either flew or is flying under the radar.

Florida.

During 2014-2015 school year the FCAT Writing Program in Florida transitioned to Florida Standards Assessments. Writing assessments are both scored by a human rater and a machine prediction model. If there is a significant difference between the two scores, a second human rater is brought in to adjudicate the situation. In this scheme, only the human ratings "really" counts, but it also provides a monitoring mechanism for human raters and it will allow the state to compile comprehensive data on machine scoring performance. The hope is to eventually have a greater reliance on machine scored prediction models. Because it is the human score assignment that counts, there has been little negative push-back on this approach.

Wyoming—WY-TOPP

WY uses automated scoring with interim. Teacher reaction to interim automated scoring affects their perception of summative automated scoring. Rolling out AS for interims also involves communications that may have helped in WY such as FAQs, rubric/items/annotated at each score point, road-show presentations.

Recommendations

Given the successes and obstacles encountered in the roll-outs in these states, we suggest following steps be applied for a smooth transition to machine scoring. Given the potential for negative public attention around scoring errors, we recommend a phased process, and that any initial implementation rely primarily on human scores, moving toward greater use of automated scoring as initial successes are achieved. This process may be less true for formative assessment given the lower stakes associated with those scores; however, note that teacher perception of formative engine score quality (which impacts their perception of summative engine score quality) is an important consideration.

Phase 1. Start with a research study endorsed by a technical advisory committee.

At a minimum, this study should examine the performance of the engine relative to hand-scorers using the usual evaluation metrics (Exact Agreement, QWK, SMD, SD Ratio) and the impact at the combined score level (trait correlation matrix, comparison of agreements, score distributions at summed score level) {Williamson:2012vm}. It is recommended that this study include performance of the engine on papers used to train and qualify readers (or validity papers) to demonstrate alignment to the intended application of the rubric. This study should also include a review of condition codes provided by the engine and mapped to the rubric codes and an examination of the engine condition code by various thresholds. It should also include a review of how aberrant responses are flagged and data to support that use. The study should examine the extent to which the engine is robust to common aberrant responses (repeated text, copies from the canon ("Four score and seven years ago…"), gibberish, non-English, copies of the prompt/directions, Babel essays, overly long or short essays, off-topic essays, etc.). The state could involve teachers as appropriate in engine modelling (e.g., getting teacher input on condition code thresholds such a prompt copy match).

Phase 2. Design initial scoring plan.

Use the research study to make decisions around implementation, including how and when human and engine scores are combined. Changes in the source of scoring should examine any impacts at the test scaled score level (item/person parameter estimation, distribution of scaled scores), and performance level (changes in percent in category). In addition, these decisions should include a plan for how to monitor both the engine and human performance and how and when those results are communicated to the state entity.

Phase 3. Design a communication plan.

Design a communication plan for school administrators and teachers that outlines for them the rationale and evidence underlying the

adoption of automated scoring. This plan can include a number of elements: 1. A description of the change and how that might impact them; 2. A description of the how the engine works, what codes are used, what responses are flagged; 3. A description of the model that combines the hand- and engine scoring; 4. An emphasis of rubric-based (versus engine-based) scoring that provides items, rubrics, annotated exemplars and potential training on how to interpret the rubric; 5. A description of how essay scoring maps to achievement; and, 6. An opportunity and method for teachers to ask questions.

Phase 4. Propose a pilot.

This pilot could be at the school or district level and the purpose is to examine the performance in an operational-like setting. Such an approach can test both the scoring quality of the engine and any emergent technical matters (system integration, reporting capability, latency, speed, load, and security). Issues will undoubtedly occur and perform a post-mortem of the pilot to address areas of concern.

Phase 5. Implementation.

If the pilot is successful, then deploy the engine for operational use for the entire state. For successful deployment, the following recommendations to: roll out communication plan, with artifacts and presentations to teachers; ensure engine reproduces results from technical reports; ensure monitoring tools are enabled and tested early in the window; create a mechanism whereby teachers can ask questions, provide feedback, and receive information about scoring questions; have mitigation plan in case of scoring errors; and, document results and any emergent issues for further review.

As part of implementation, the state may want to conduct continued writing training/ professional development for teachers around writing standards, how to use them to score student responses, and how to use that information to improve learning. Such an approach could be part of any larger professional development program.

Phase 6. Review and Revise.

Once a deployment is completed, then a debrief of the technical, psychometric results and teacher feedback should occur. This discussion will help refine the communications document, hybrid scoring plan, and engine modelling.

Bibliography

ACARA NASOP Research Team. (2015). *An evaluation of automated scoring of NAPLAN persuasive writing.* Canberra, Australia: Australian Curriculum, Assessment and Reporting Authority. Retrieved from www.nap.edu.au/online-assessment/research-and-development/automated-essay-scoring

Burdick, H., Swartz, C. W., Stenner, A. J., Fitzgerald, J., Burdick, D., & Hanlon, S. T. (2013). Measuring students' writing ability on a computer-analytic developmental scale: An exploratory validity study. *Literacy Research and Instruction, 52*(4), 255–280. http://doi.org/10.1080/19388071.2013.812165

Page, E. B. (2003). Project Essay Grade: PEG. In *Automated essay scoring: A cross-disciplinary perspective* (pp. 43–54). Mahwah, NJ: Lawrence Erlbaum Associates.

Perelman, L. (2016). Grammar checkers do not work. *WLN a Journal of Writing Center Scholarship, 40*(7-8), 11–20. Retrieved from http://lesperelman.com/wp-content/uploads/2016/05/Perelman-Grammar-Checkers-Do-Not-Work.pdf

Perelman, L. (2017). *Automated essay scoring and NAPLAN: A summary report. Unpublished Report.*

Ramineni, C., Trapani, C. S., Williamson, D. M., Davey, T., & Bridgeman, B. (2012). *Evaluation of the e-rater® scoring engine for the GRE® issue and argument prompts. ETS Research Report Series* (Vol. 2012, pp. i–106). John Wiley & Sons, Ltd.

Rathke, K., Palermo, C., & Wright, J. (2018). Automated writing evaluation: Multiple

perspectives. Presented at the National Conference on Student Assessment, San Diego, CA.

Rich, C. H. (2013). Applications of automated essay evaluation in West Virginia. In *Handbook of Automated Essay Evaluation*. New York, NY: Routledge.

Rich, C. S., Harrington, H., & Kim, J. (2008). Automated essay scoring in state formative and summative writing assessment. Presented at the American Educational Research Association, New York, NY.

Shermis, M. D. (2014a). State-of-the-art automated essay scoring: A United States demonstration and competition, results, and future directions. *Assessing Writing, 20*, 53–76.

Shermis, M. D. (2014b). The challenges of emulating human behavior in writing assessment. *Assessing Writing, 22*, 91–99. http://doi.org/10.1016/j.asw.2014.07.002

Shermis, M. D. (2018). International applicaitons of machine scoring. Presented at the National Council on Measurement in Education, New York, NY.

Shermis, M. D., & Hamner, B. (2013). Contrasting state-of-the-art automated scoring of essays. In M. D. Shermis & J. C. Burstein (Eds.), *Handbook of automated essay evaluation: Current applications and new directions.* (pp. 298–312). New York.

Shermis, M. D., Burstein, J. C., Elliot, N., Miel, S., & Foltz, P. W. (2015). Automated writing evaluation: An Expanding Body of Knowledge. In *Handbook of writing research (2nd Edition)* (pp. 395–409). New York, NY: The Guilford Press.

Smith, T. (2018, June 30). More states opting to "robo-grade" student essays by computer. Washington, DC: National Public Radio.

White, I., Hixson, N., & Rhudy, V. (2010a). *WESTEST 2 online writing scoring comparability study.* Charleston, WV: West Virginia Department of Education.

White, I., Hixson, N., & Whisman, S. A. (2011). *Writing Roadmap usage and additional predictors of WESTEST 2 online writing scores.* Charleston, WV: West Virginia Department of Education, Division of Curriculum and Instructional Services, Office of Research.

White, I., Hixson, N., DBrot, J., Perdue, J., Foster, S., & Rhudy, V. (2010b). *Impact of Writing Roadmap 2.0 on WESTEST 2.0 online writing assessment scores.* Charleston, WV: West Virginia Department of Education.