Assessing Writing 20 (2014) 53–76

Contents lists available at ScienceDirect

Assessing Writing

State-of-the-art automated essay scoring: Competition, results, and future directions from a United States demonstration

Mark D. Shermis [a,b,*]

[a] Department of Educational Foundations and Leadership, The University of Akron, United States
[b] Department of Psychology, The University of Akron, United States

ARTICLE INFO

Article history:
Received 27 August 2012
Received in revised form 17 April 2013
Accepted 26 April 2013
Available online 30 January 2014

Keywords:
Automated essay scoring
High-stakes assessment
Writing
Race-to-the-Top
Performance assessment
Human raters

ABSTRACT

This article summarizes the highlights of two studies: a national demonstration that contrasted commercial vendors' performance on automated essay scoring (AES) with that of human raters; and an international competition to match or exceed commercial vendor performance benchmarks. In these studies, the automated essay scoring engines performed well on five of seven measures and approximated human rater performance on the other two. With additional validity studies, it appears that automated essay scoring holds the potential to play a viable role in high-stakes writing assessments.

1. Introduction

1.1. Context

A new generation of measurement instruments is being planned for use in the United States as part of the Race-to-the-Top assessments. These instruments will be based on the instructional goals of Common Core State Standards that articulate the required proficiencies for United States students

* Correspondence to: Department of Educational Foundations and Leadership, The University of Akron, 213 Crouse Hall, 44325, United States. Tel.: +1 954 899 8069.
E-mail address: mshermis@uakron.edu

1075-2935/$ – see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.asw.2013.04.001

to be "college-ready" by the time they graduate from high school (Porter, McMaken, Hwang, & Yang, 2011). These new assessments will likely rely less on multiple-choice questions and use performance measures that more closely match the construct under investigation. The move to Common Core State Standards in the U.S. represents a significant departure from a curricular structure that heretofore has been driven by individual states.

Over the past 30 years the high-stakes assessments associated with state objectives have been calibrated to the minimal standards for exiting high school. These standards have not been universal and vary from state to state. In the area of high-stakes state writing assessment, writing objectives can range from the summarization of reading material to the ability to create prose of a particular genre to the mastery of a particular writing form. Writing assessment practices also differ from state to state including the amount and type of writing expected, types of rubrics used, scoring and adjudication protocols, the number and qualifications of raters employed, quality assurance practices, and the reporting of results.

In part because of the emphasis on minimum competency and the varied nature of what a state might emphasize in their high-stakes testing programs there grew a widening pool of college students who had the skill set to graduate from high school yet had to enroll in remedial college classes because that skill set did not include the higher order knowledge or skills required to perform well in entry-level college classes (Attewell, Lavin, Domina, & Levey, 2006) where the curriculum is typically based on the standards of the discipline's national organization. The two major Race-to-the-Top Consortia [Partnership for Assessment of Readiness for College and Careers (PARCC) and SMARTER Balanced Assessment Consortia (SBAC)] and their 46 subscriber states intend to change that pattern by having all students – even those who may wish to pursue a vocational track – work toward college readiness rather than a mastery of basic high school skills (Tucker, 2009).

With regard to assessments in English Language Arts and in many of the science areas, this shift will mean more writing. For instance, students might be given an array of articles in biology to read and then respond to an essay prompt that addresses some conclusion that they might make based on the articles. The essay might ask the student to explain a rationale for a conclusion or to cite evidence in support of an argument. Part of the current debate in planning the new instruments is whether this performance assessment is really a writing task (where the emphasis is on writing ability), reading comprehension (where the emphasis is on understanding the content), or one of critical thinking (where the emphasis is on synthesizing and evaluating information). Two of these options are consistent with Weigle's distinctions regarding the multiple purposes of assessment, assessing writing (AW) and assessing content through writing (ACW) (Weigle, 2013; Weir, 2005). In many cases, the student will be asked to produce a written artifact that must be evaluated – and to do so numerous times throughout the academic year. The sheer number of written responses for high-stakes summative assessments across the grade levels makes it challenging and cost-ineffective to have human raters exclusively score these assessments. For example, the state of Florida has approximately 180,000 students in each grade level. If each student in that one state had five essays graded, the state would be required to evaluate almost 11 million documents per year, raising questions as to the feasibility of recruiting a sufficient number of qualified human graders to provide final scores, read reliably, in a timely manner across the entirety of the United States. The goals of the Consortia have been to strongly encourage the development and use of machine scoring algorithms in order to make it possible to score such volumes in a timely and cost-effective manner.

In order to evaluate the basic feasibility of these goals, the Hewlett Foundation (www.hewlett.org) sponsored a demonstration of existing and emerging automated scoring systems for essays as part of the Automated Student Assessment Prize (ASAP) program (Shermis & Hamner, 2012, 2013), the results of which are reported here. ASAP is an independently funded organization that is exploring the effectiveness of machine scoring in different contexts and sponsors open and public prize competitions to stimulate innovations in machine scoring. Two studies are described below; one is a demonstration of performance for existing essay scoring engines, and another an open competition designed to encourage the creation of new algorithms to score essays. For both studies the goal was to evaluate the extent to which automated scoring systems for essays are capable of producing scores similar to those of trained human graders. The first study focused on pre-existing systems for automated scoring of essays and compared the systems of eight commercial vendors and one university laboratory

to the performance of human raters. The second study encouraged public contribution to the field of automated scoring of essays by providing cash prizes for data scientists who could develop machine scoring approaches that were most similar to the human scores. The intent of this public competition was to drive interest in "pushing the envelope" of machine scoring development based on new perspectives from other fields of study.

These studies target understanding aspects of machine scoring that have not previously been reported in the literature. While there are numerous published studies evaluating the performance of a single machine scoring system, there are few studies that simultaneously evaluate multiple systems. Of those, none have included more than three systems and so the simultaneous demonstration of performance for nine such systems, representing all of the leaders in the field of automated scoring of essays, on a common data set represents a new opportunity for understanding the current state-of-the-art.

One recently published study evaluated the performance of two unnamed automated scoring systems on the writing portion of Australia's AST Scaling Test, an examination used in the selection of tertiary students (McCurry, 2010). The writing portion of the test is conceived of as an assessment of "verbal reasoning and writing ability in which candidates are requested to respond in an argumentative mode of writing to a broad rand of stimulus material on a social and/or political issue" (McCurry, 2010, p. 122). The average human rating performance for two raters was $r = 0.75$ on the ten-point scale used to evaluate the essays. The two automated scoring engines were trained on 187 randomly selected essays that had been evaluated by four raters each. The systems were subsequently provided 63 randomly selected essays on which to make blind predictions. The results of this study showed that inter-human rater performance was significantly better with regard to the correlations used to calculate agreement, and the score distribution resulted in a more discriminating spread. The conclusion was that for closed content driven prompts, the machine scoring of essays fell short of human rater performance. However, this study was limited to one writing prompt and a small training and test sample. Moreover, it incorporated a restricted number of outcome measures, and the approaches taken by the unnamed automated essay scoring engines were unknown.

Finally, the demonstration reported in this paper is moderated by ASAP, which acted as an independent entity with no ties or obligations to any of the developers or purveyors of machine scoring systems for essays. There is little research on how methods from outside the field of educational measurement might be brought to bear on the problem of machine scoring for essays, and this paper reports on an effort to engage a larger community of researchers in this issue. As a result, the outcomes of this work have the potential to inform the public, and the Consortia, regarding the feasibility of their vision of widespread large-scale standardized assessment that includes substantial writing exercises, with fast score reporting, to be ready for operational use by 2014.

While providing new understanding of the current, and near term potential, state-of-the-art of machine scoring for essays to replicate the scores of human raters, the current studies are not comprehensive in their goals and design. In soliciting participation from many different systems, both existing commercial and public systems as well as newly-developed approaches, the basis for the demonstration of these systems was solely established on comparison with human scores on existing state assessments. As such, the study evaluates only one part of the overall set of considerations in validation of machine scoring: Can machine scoring produce scores similar to those of human raters? If machine scoring proves incapable of replicating the scores of human graders then there is little need to pursue deeper questions of the validity of scores as the machines have not yet demonstrated that they are ready for more rigorous evaluation. However, if the machine scoring systems are capable of replicating human scores, then this finding provides (a small) part of the evidence needed for more rigorous validation of machine scoring. Further, the design targets an evaluation of the consistency of this similarity with human scores across multiple types of prompts examined in the study. It is beyond the scope and space limitations of this paper to elaborate on the broader range of evaluations that are needed for a complete validity argument, but the interested reader is referred to Kane (2006) for a fundamental presentation of validity and Williamson, Xi, and Breyer (2012) and Ramineni and Williamson (2013) for perspectives on validation of machine scoring systems in particular. Familiarization with this literature will further sensitize the reader to the fact that this study examines only a portion of what would be required for an overall validity argument for machine scoring. Instead,

the purpose is to demonstrate the extent to which current and near term potential state-of-the-art of machine scoring can satisfy one of the multiple validation criteria for use.

It is also important to represent at the outset that this paper does not address the very important distinction between writing education and summative writing assessment as practiced by states in the United States (Applebee & Langer, 2009). The construct of writing is a rich, nuanced, and sophisticated domain. If actual student writing occurs at all in summative state assessment, the common practice is to administer a single writing prompt and the examinee must write a short (typically with a time limit of around 30 minutes) essay on the given topic, turning in this "first draft" as the final graded submission for the assessment. Naturally, there are quite legitimate concerns that this common practice of summative assessment does not adequately reflect the construct of writing as it is taught in the K-12 American educational system (Condon, 2013). While these concerns are readily acknowledged as important and worthy of debate, the scope of this study is restricted to the extent to which the current state-of-the-art of machine scoring can replicate the scores of human graders on existing writing prompts, regardless of the extent to which these are thought to be optimal measures of the construct of writing.

With these limitations in scope, the following sections more fully describe the design and conduct of the studies.

2. Method

2.1. Study 1 vendor demonstration

2.1.1. Participants

Student essays ($N = 22{,}029$) were collected for eight different prompts representing six (PARCC and SBAC) states (three PARCC states and three SBAC states) that are part of the two Race-to-the-Top assessment consortia. The two consortia made the initial contact with the participating states and ASAP began a series of negotiations that allowed us to screen their availability for participation, the type of essay, grade levels, the presence of multiple ratings, and the likelihood that the essays could be processed and validated in time for the launch of the vendor demonstration. Many states do not employ writing as part of their high-stakes assessment programs, and obviously could not be included in the sample. The states that were chosen and cooperated had the endorsement of the consortia, and simply asked that they not be identified in the report. In formulating the sample, an attempt was made to ensure there were a range of ethnicities and gender representation of students in the sample, though these were not targeted to be representative samples for any particular population of students. Efforts were also undertaken to ensure that the types of prompts represented a range of typical writing tasks through prompts that might be similar to those anticipated in the new assessments. Responses represented a purposely heterogeneous mix in length of response. The sample is composed of essays from volunteer states and therefore cannot be assumed to be a truly representative sample of state practice. Senior representatives from both consortia represented that the participating states were appropriate representatives of the other states in the respective consortia, but due to the sample selection mechanism the results here cannot be considered representative of any particular population since formal sampling designs were not employed for data collection. Instead, these are samples designed to reflect a diversity of scoring policies, demographic makeup, geographic region, prompt type, and other factors of interest for the purposes of a demonstration of a broad range of systems that represent the state-of-the-art for machine scoring of essays.

Three of the states were located in the Northeastern part of the U.S., two from the Mid-west, and one from the West Coast. Because no demographic information was provided by the states, student characteristics had to be estimated based on assumptions related to participating states, as displayed in Table 1. Student writers were drawn from three different grade levels (7, 8, 10) and the grade-level selection was generally a function of the testing policies of the participating states (e.g., a writing component as part of a 10th grade exit exam), were ethnically diverse, and evenly distributed between males and females. No information was obtained regarding the characteristics of students who did not participate in the high-stakes testing programs. U.S. federal guidelines recognize that a very small

proportion of students may be excluded from the regular testing programs based some stringent criteria.

Samples ranging in size from 1527 to 3006 were randomly selected from the data sets provided by the states, and then randomly divided into three sets: a training set, a test set, and a validation set. The training set was used by the participants to create scoring models, and with the exception of one data set, consisted of scores assigned by two human raters, a final or an adjudicated score (referred to as the *resolved score*), and the text of the essay. Most vendors process training essay text by parsing it, dividing it into writing components, adding information based on natural language processing, and then creating statistical models upon which their predictions of score are based. Indexing the amount and relevance of content in the essay is established through a variety of techniques that reference the content in the training sample against new candidate essays. The most popular approach is that of Latent Semantic Analysis (Landauer, Foltz, & Laham, 1998) or one its non-proprietary variants such as Content Vector Analysis (Salton, Wong, & Yang, 1975). The test set consisted of essay text only and was used as part of a blind test to compare different systems on the basis of score predictions. The validation set was used for the second study only and consisted solely of essays, being employed to verify their scoring code by the competition administrators (vendors were not required to show their source code) to prevent the vendors from over-fitting their models. The distribution of the samples was split in the following proportions: 60% training sample, 20% test sample, 20% validation sample. The actual proportions vary slightly due to the elimination of cases containing data errors. The distribution of the samples is displayed in Table 1. The specification of the training set size was predicated on what many states might employ for a pilot sample in pre-testing items for a new version of their high-stakes writing assessment.

2.1.2. Instruments

Four of the essay prompts were drawn from traditional writing discourse modes (persuasive, expository, narrative) and four essay prompts were "source-based"-that is, the questions asked in the prompt referred to a source document that students read as part of the assessment (i.e., sometimes referred to as a reading summary; cf, Weigle, 2013). In the training set, average essay lengths varied from $M = 94.39$ ($SD = 51.68$) to $M = 622.24$ ($SD = 197.08$). Traditional essays were significantly longer ($M = 354.18$, SD 197.63) than source-based essays ($M = 119.97$, SD = 58.88; $t_{(13,334)} = 95.18$, $p < .05$). Testing times allotted for the high-stakes assessments ranged from 45 minutes to one hour. While there may be some

Table 1
Sample characteristics estimated from reported demographics of the state.*

	Data set #															
	1		2		3		4		5		6		7		8	
State	#1		#2		#3		#3		#4		#4		#5		#6	
Grade	8		10		10		10		8		10		7		10	
Grade level *N*	42,992		80,905		68,025		68,025		71,588		73,101		115,626		44,289	
n	2968		3000		2858		2948		3006		3000		2722		1527	
Training *n*	1785		1800		1726		1772		1805		1800		1730		918	
Test *n*	589		600		568		586		601		600		495		304	
Validation *n*[t]	594		600		564		590		600		600		497		305	
Gender M% | F%	51.2	48.8	51.4	48.6	51.0	49.0	51.0	49.0	49.6	50.4	49.2	50.8	51.2	48.8	48.7	51.3
Race % W% | N%	63.8	36.2	77.8	22.2	42.9	57.1	42.9	57.1	70.2	29.9	69.5	30.5	70.2	29.8	66.3	33.7
Free/reduced lunch %	32.9		40.0		32.2		32.2		34.2		34.2		46.6		41.3	

Source: Copyright 2013 from *Handbook of Automated Essay Evaluation: Current Applications and New Directions* by Mark D. Shermis and Jill Burstein. Reproduced by permission of Taylor and Francis Group, LLC, a division of Informa plc.

M, male; F, female; W, white; N, non-white.

* Taken primarily from: National Center for Education Statistics, Common Core of Data (CCD), (2010). State Non-fiscal Survey of Public Elementary/Secondary Education, 2009–10, Version 1a. Washington, DC: U.S. Department of Education. This information was supplemented with state department of education website information or annual reports for each participating state.

[t] The validation set was not used in this study.

debate as to whether writing samples as short as 93 words constitute an "essay," these sample sizes reflect what many states are defining as essays in their high-stakes tests.

Five of the prompts employed a holistic scoring rubric, one prompt was scored with a two-trait rubric, and two prompts were scored with different multi-trait rubrics but reported as a composite score. The holistic scoring rubrics, as labeled by the states, were a holistic-analytic hybrid. Raters were asked to consider different dimensions of writing in making their assessments, but to assign one overall score. The particular analytic dimensions were not scored. As a general rule, the trait rubrics focused on the attributes of performance for a particular audience and writing purpose. The type of rubric, scale ranges, scale means and standard deviations, are reported in Tables 2 and 3. Table 2 shows the characteristics of the training set and Table 3 shows the characteristics of the test set. Human rater agreement information is reported in Tables 2 and 3 with associated data for exact agreement, exact + adjacent agreement, kappa, Pearson *r*, and quadratic-weighted kappa. Quadratic-weighted kappas ranged from 0.62 to 0.85, a typical range for human rater performance in statewide high-stakes testing programs.

In all cases, the classification of prompts as a particular type (e.g. narrative), the particular scoring rubric used by human raters, and the scoring processes followed (including adjudication and resolution procedures) were provided by the states. Since the purpose of this study is to evaluate performance of machine scoring on responses from current practice, no attempt was made to review and/or override decisions made at the state level about prompt classification, rubric design, score resolution procedures or any other aspect of the state testing program. As a result, there is, by design, a diverse set of policies represented in the data sets used for these studies and part of the goal of the research was to see how robust machine scoring is to such variations in policy.

2.1.3. Procedure

Six of the essay sets were transcribed from their original paper-form administration in order to prepare them for processing by automated essay scoring engines, which require the essays to be in ASCII format. This process involved retrieving the scanned copies of essays from the state or a vendor serving the state, randomly selecting a sample of essays for inclusion in the study, and then sending the selected documents out for transcription.

Both the scanning and transcription steps had the potential to introduce errors into the data that would have been minimized had the essays been directly typed into the computer by the student, the normal procedure for automated essay scoring. Essays were scanned on high quality digital scanners, but occasionally student writing was illegible because the original paper document was written with an instrument that was too light to reproduce well, was smudged, or included handwriting that was undecipherable. In such cases, or if the essay could not be scored by human raters (i.e., essay was off-topic or inappropriate as determined by human raters), the essay was eliminated from the analyses. Transcribers were instructed to be as faithful to the written document as possible keeping in mind the extended computer capabilities had they been employed. For example, more than a few students used a print style in which all letters were capitalized. To address this challenge, we instructed the transcribers to capitalize according to conventional practice. This modification may have corrected errors that would have otherwise been made, but limited the over-identification of capitalization errors that might have been made otherwise by the automated essay scoring engines.

The first transcription company serviced four prompts from three states and included 11,496 essays. In order to assess the potential impact of transcription errors, a random sample of 588 essays was re-transcribed and compared on the basis of error rates for punctuation, capitalization, misspellings, and skipped data. Accuracy was calculated on the basis of the number of characters and the number of words with an average rate of 98.12%.[1] The second transcription company was evaluated using similar metrics. From a pool of 6006 essays, a random sample of 300 essays was selected for re-transcription. Accuracy for this set of essays was calculated to be 99.82%.

[1] Error rate was a combination of two components that were equally weighted: (1) the error rate of words over the total number of words and (2) the error rate of characters over the total number of characters. The latter component takes into consideration features such as punctuation.

Table 2
Training set characteristics.

	Data set #							
	1	2		3	4	5	6	7	8
N	1785	1800		1726	1772	1805	1800	1730	918
Grade	8	10		10	10	8	10	7	10
Type of essay	Persuasive	Persuasive		Source-based	Source-based	Source-based	Source-based	Expository	Narrative
M # of words	366.40	381.19		108.69	94.39	122.29	153.64	171.28	622.13
SD # of words	120.40	156.44		53.30	51.68	57.37	55.92	85.20	197.08
Type of rubric	Holistic	Trait (2)		Holistic	Holistic	Holistic	Holistic	Holistic*	Holistic+
Range of rubric	1–6	1–6	1–4	0–3	0–3	0–4	0–4	0–12	0–30
Range of RS	2–12	1–6	1–4	0–3	0–3	0–4	0–4	0–24	0–60
M RS	8.53	3.42	3.33	1.85	1.43	2.41	2.72	19.98	37.23
SD RS	1.54	0.77	0.73	0.82	0.94	0.97	0.97	6.02	5.71
Exact agree	0.65	0.78	0.80	0.75	0.77	0.58	0.62	0.28	0.28
Exact + Adj agree	0.99	0.93	1.00	1.00	1.00	0.98	0.99	0.54	0.49
κ	0.45	0.65	0.66	0.61	0.67	0.42	0.46	0.17	0.15
Pearson *r*	0.72	0.81	0.80	0.77	0.85	0.75	0.78	0.73	0.63
Quadratic weighted κ	0.72	0.81	0.80	0.77	0.85	0.75	0.78	0.73	0.62

Source: Copyright 2013 from *Handbook of Automated Essay Evaluation: Current Applications and New Directions* by Mark D. Shermis and Jill Burstein. Reproduced by permission of Taylor and Francis Group, LLC, a division of Informa plc.

RS, resolved score; Adj, adjacent; Agree, agreement.

* Composite score based on four of six traits.

+ Composite score based on six of six traits.

Table 3
Test set characteristics.

	Data Set #							
	1	2		3	4	5	6	7	8
N	589	600		568	586	601	600	495	304
Grade	8	10		10	10	8	10	7	10
Type of Essay	Persuasive	Persuasive		Source-based	Source-based	Source-based	Source-based	Expository	Narrative
M # of Words	368.96	378.40		113.24	98.70	127.17	152.28	173.48	639.05
SD # of Words	117.99	156.82		56.00	53.84	57.59	52.81	84.52	190.13
Type of Rubric	Holistic	Trait (2)		Holistic	Holistic	Holistic	Holistic	Holistic*	Holistic+
Range of rubric	1–6	1–6	1–4	0–3	0–3	0–4	0–4	0–12	0–30
Range of RS	2–12	1–6	1–4	0–3	0–3	0–4	0–4	0–24	0–60
M RS	8.62	3.41	3.32	1.90	1.51	2.51	2.75	20.13	36.67
SD RS	1.54	0.77	0.75	0.85	0.95	0.95	0.87	5.89	5.19
Exact agree	0.64	0.76	0.73	0.72	0.78	0.59	0.63	0.28	0.29
Exact + adj agree	0.99	1.00	1.00	1.00	1.00	0.98	0.99	0.55	0.49
κ	0.45	0.62	0.56	0.57	0.65	0.44	0.45	0.18	0.16
Pearson *r*	0.73	0.80	0.76	0.77	0.85	0.74	0.74	0.72	0.61
Quadratic weighted κ	0.73	0.80	0.76	0.77	0.85	0.75	0.74	0.72	0.62

Source: Copyright 2013 from *Handbook of Automated Essay Evaluation: Current Applications and New Directions* by Mark D. Shermis and Jill Burstein. Reproduced by permission of Taylor and Francis Group, LLC, a division of Informa plc.

RS, resolved score; Adj, adjacent; Agree, agreement.

* Composite score based on four of six traits.

+ Composite score based on six of six traits.

Two of the essays were provided in ASCII format by their respective states. The 10th grade students in those states had typed their responses directly into the computer using web-based software that emulated a basic word processor. While the test employed digital technology, the conditions for testing were similar to those in states where the essays had been transcribed.

One of the key challenges to both sets of data, those that were transcribed and those that were directly typed, was that carriage returns and paragraph formatting meta-tags were missing from the ASCII text. For some of the scoring engines, this omission could have introduced a significant impediment in the engine's ability to accurately evaluate the underlying structure of the writing, one component in their statistical prediction models.

States subcontracted the scoring of essays to commercial testing vendors who were responsible for recruiting, training, and staffing professional graders on a part- or full-time basis. These individuals typically have college training, are often college graduates, and some are teachers engaging in supplemental work. Each state produces an annual technical report that describes the training of raters and provides descriptions of the quality assurance steps that they take to ensure the efficacy of their work, including reliability and validity checks. States have slight differences in the procedures that they may ask vendors to perform. Much like other characteristics of the data sample the potential variability in recruitment, training and performance characteristics of human graders across the prompts collected is an inherent and unavoidable part of the collection of diverse data from multiple states.

The commercial vendors and the one university laboratory (*LightSIDE*, Carnegie Mellon University, TELEDIA Laboratory) were provided with a training set for each of the eight essay prompts. The commercial vendors were as follows: were:

AutoScore, American Institutes for Research (AIR)
Bookette, CTB McGraw-Hill
CRASE™, Pacific Metrics
e-rater®, Educational Testing Service
Intelligent Essay Assessor (IEA), Pearson Knowledge Technologies
IntelliMetric, Vantage Learning
Lexile® Writing Analyzer, MetaMetrics
Project Essay Grade (PEG), Measurement, Inc.

Up to four weeks were allowed to develop statistical models from the data during the training phase of the demonstration. In addition, the demonstrators were provided with cut-score information along with any scoring guides that were used in the training of human raters. This supplemental information was employed by some of the vendors to better model score differences in the state-provided rubrics. Two of the essay prompts used trait rubrics to formulate a composite score by summing some or all of the trait scores. For these two essays both the holistic and trait scores were provided to the vendors.

During the training period, a series of conference calls with the participants, with detailed questions and answers, were conducted to clarify the nature of the data sets or to address data problems that arose while modeling the data. For example, different states had varying procedures in how to adjudicate discrepant scores. For data sets #1, #7, and #8 the scores from rater #1 and rater #2 were summed to produce the final score. This was problematic because it effectively doubled the range of the original scale. Because data sets #7 and #8 had comparatively wide ranges to begin with, the adjudication procedure would likely impact the calculation of distributional and agreement metrics. However, in order for the outcomes of the study to be reflective of existing state practice we adhered to the adjudication guidelines as set up by each state. For data sets #3 and #4 a third rater was asked to make a final determination of the score, called the *resolved score*. In this case it is possible that the resolved score reflected neither the ratings of rater #1 nor rater #2 if the third rater overrode both previous grades with his or her own judgment. With data sets #5 and #6, the state policy was to take the higher of the two scores as the final score for the essay. This practice poses challenges to the modeling of scores due to the potential to introduce bias into the distribution of the scale relative to the human score distributions for the first two raters. Data set #2 was scored differently altogether. This state used the rating from rater #1 to determine the final score, so no scores in this data set were resolved. The scores from rater #2 were generated as a result of a "read-behind" in which the second

rater independently scores the essay for quality assurance purposes, but the score is never used in the calculation of the final score. Even though the score is never resolved (i.e., because there are no conflicts with the assigned score), it, like all of the final official scores from the states, is still referred to in this manuscript as the resolved score. These adjudication procedures represent a wide range of approaches and substantially complicate the modeling process for the machine scoring systems. While they are accurate representations of current state practice for consequential assessment it is not clear how representative these procedures are for the remaining states, or whether single approach represented could be put forward as a preferred procedure.

In the state that used the adjudication rule to take the higher of the two rater scores, there were a handful of instances where the state did not appear to follow its own rule in resolving the score. Rather than modify the resolved score, the vendors were instructed to use it in their prediction models even though it was apparently inconsistent with the state's guidelines. Naturally, such observations further complicated the modeling task for machine scoring.

The variety of state policies, and deviations from them for particular responses, had the potential to undermine the capacity of machine scoring systems to accurately model the essay scores provided by the states. However, in the context of the current study in which the goal was to replicate state scores these challenges constituted a legitimate test of the flexibility and robustness of machine scoring in meeting state assessment goals. Stated somewhat differently, the consideration of these inconsistencies provided a representation of the typical contextual conditions within which the scoring engines would be deployed if used operationally for state assessments.

In the "test" phase of the evaluation, vendors were provided data sets that had only the text of essays available, and were asked to make integer score predictions for each essay. They were given a 59-hour period in which to make their predictions and were permitted to eliminate up to 2% of the essay score predications in each data set in case their scoring engine classified the essay as "unscorable." Even though human raters had successfully rated all the essays in the test set, there were a variety of reasons that any one essay might prove problematic for machine scoring. For example, an essay might have addressed the prompt in a unique enough way to receive a low human score, but be deemed as "off topic" for machine scoring. In operational deployments of machine scoring provisions would be made for such cases rejected by the machine to be scored by human raters.

2.1.4. Scoring engines

Eight of the nine automated essay scoring engines that were evaluated in the demonstration represented commercial entities which in combination reflect nearly the entire commercial market for automated scoring of essays in the United States. The lone non-commercial scoring engine (*LightSIDE*) from the Teledia Laboratory at Carnegie Mellon University was invited into the demonstration to facilitate a complete survey of the existing field of offerings, although this system is an open-source package that was publicly available on their web site (Mayfield & Rosé, 2010) rather than a commercial system. The scoring engines and the commercial vendor names are shown in Table 4 which also summarizes the key features of the automated scoring engine involved in the study. Due to the number of vendors and space limitations, the operating description of each engine cannot be described here and the reader is referred to Shermis and Burstein (2013) for more information on these capabilities.

2.1.5. Scoring

As indicated above, the sole basis for the comparison of machine scoring methods is in reference to the human scores provided by the states. Rather than try to evaluate the similarity of human and machine scoring on the basis of one metric, this study used a set of measures that are common for such comparisons, including:

- Distributional differences – correspondence in mean and variance of the distributions of resolved scores to that of automated scores. These measures employed the resolved score (i.e., the score assigned by the state) as a benchmark against which to evaluate machine performance,
- Agreement (reliability) – measured by correlation (Stemler, 2004), weighted kappa and percent agreement (exact and exact + adjacent). These measures compared the performance from the two human raters against machine predications to the resolved score.

Table 4
Comparison of AES systems.*

AES System	Developer	Technique	Main focus	Instructional application	Number of essays required for training
AutoScore	American Institutes for Research	NLP	Style and content	N/A	Unavailable
Bookette	CTB McGraw-Hill	NLP	Style and content	Writing Roadmap™	250–500
CRASE™	Pacific Metrics	NLP	Style and content	N/A	100 per score point
e-rater®	Educational Testing Service (Burstein, Kukich, Wolff, Lu, & Chodorow, 1998)	NLP	Style and content	Criterion™	100–1000
Intelligent Essay Assessor (IEA)™	Pearson Knowledge Technologies (Landauer, Laham, Rehder, & Schreiner, 1997)	LSA	Content	Write to Learn	100–300
Intellimetric™	Vantage Learning (Elliot, 2003)	NLP	Style and content	MY Access!e®	300
Lexile® Writing Analyzier	MetaMetrics	NLP	Style and content	N/A	None-uses a fixed model
LightSIDE	Teledia Laboratory (Mayfield & Rosé, 2013)	Statistical	Content	N/A	300
Project Essay Grade (PEG)™	Measurement Incorporated (Ajay et al., 1973)	Statistical	Style	N/A	100–400

* Based on Dikli (2006).

Each of these measures is described below. Given the limitations of this study this evaluation is restricted to the comparison with human scores, a subset of a broader range of evaluations proposed by Williamson et al. (2012).

Supplementary Information

Supplementary information may be found at this link: http://www.scoreright.org/asap.aspx?content=Request_ASAP_Phase_One_Data.

3. Results

Eight of the nine vendors provided scores for all essays in the test set, with the remaining vendor providing scores for all but 10 of the essays. The proportion of non-response was so small (10/4343 = .002) that for comparison purposes, the results are treated as if they had reported at the 100% rate.

The scoring engines were able to replicate the mean scores for all of the data sets. Table 5 illustrates this replication. RS refers to the resolved score (i.e., the official state-adjudicated score) and the abbreviations for the vendors are given above. Note that in three data sets (1, 7, and 8) the resolved score was the equally weighted sum of the two human raters, which were doubled to be on the same scale as the RS.

Prediction accuracy was likely influenced by the range of the scale. For example, all vendor engines generated predicted means within 0.10 of the human mean for Data Set #3, which had a rubric range of 0–3, whereas the difference in mean scores between human and machine scores is larger for prompts with a bigger range of possible scores. However, even when the range was much larger as in Data Set #8 (range of 0–60), mean estimates were generally within 1 point, and usually smaller. As a clarification, Data Sets #7 and #8 were derived from trait scoring rubrics, but were reported as composite (summed) scores. The states that provided these data adjudicated score discrepancies by adding them (in contrast to other possible resolution procedures) thus contributing to the large range of possible scores.

Table 6 shows the results for the standard deviations for each of the data sets. With the exception of data sets 1, 7, and 8, where the two human rater scores were summed to get a resolved score, most of the predicted scores had standard deviations within 0.10 of the resolved scores. The standard deviations were larger for data sets 7 and 8 where the scale ranges were much wider than with some of the earlier data sets.

Table 7 begins the sequence of agreement statistics for the data sets. The human exact agreements ranged from 0.29 on data set #8 to 0.76 for data set #2. As the name implies, exact agreement for the human raters means that the scores assigned by Rater 1 and Rater 2 were exactly the same. Machine scoring exact agreements represented the agreement between the Resolved Score and the predicted machine score and had a range from 0.07 on data set #7 to 0.72 on data sets #3 and #4. A closer inspection shows that machines performed particularly well on data sets #5 and #6, two of the shorter and source-based essays.

In interpreting the results for agreement measures provided in Tables 7–11 the reader is cautioned that some values cannot be directly compared. The values under column headings of H1 and H2 represent the agreement of H1 and H2 (respectively) with the resolved scores. However, since both H1 and H2 directly contribute to the determination of the resolved score these rates of agreement are substantially higher than would be the case between an independent rater (a hypothetical H3) and the resolved score. For correlations (presented later) this phenomenon is referred to as part-whole correlation. Due to the inflation of these measures this paper compares the machine scoring performance to the baseline agreement between the two human scores (H1H2) as the basis for contrasting agreement in Tables 7–11.

Adjacent agreements refer to the combined exact and adjacent score agreements. This calculation is based on the generally accepted testing convention of considering rater score assignments within one score point as being a “match.” If the scores differ by more than one point, many states will ask a third rater to evaluate the essay or they may have some other rule about how to handle the score

Table 5
Test Set Means.

Essay set	*N*	*M* # of words	H1	H2	RS	AIR	CMU	CTB	ETS	MI	MM	PKT	PM	VL	AES mean	AES median
1[*]	589	366.40	8.61	8.62	8.62	8.54	8.51	8.56	8.57	8.53	8.56	8.57	8.49	8.80	8.57	8.56
2a[t]	600	381.19	–	3.39	3.41	3.41	3.36	3.39	3.39	3.37	3.33	3.41	3.36	3.40	3.38	3.39
2b[t]	600	381.19	–	3.34	3.32	3.37	3.18	3.35	3.32	3.21	3.26	3.29	3.32	3.34	3.29	3.32
3	568	108.69	1.79	1.73	1.90	1.90	1.90	1.92	1.88	1.95	1.91	1.84	1.89	1.92	1.90	1.90
4	586	94.39	1.38	1.40	1.51	1.50	1.47	1.50	1.34	1.48	1.46	1.39	1.47	1.57	1.46	1.47
5	601	122.29	2.31	2.35	2.51	2.49	2.51	2.49	2.47	2.51	2.44	2.49	2.50	2.54	2.49	2.49
6	600	153.64	2.57	2.58	2.75	2.79	2.71	2.83	2.54	2.76	2.74	2.76	2.74	2.83	2.74	2.76
7[*]	495	171.28	20.02	20.24	20.13	20.05	19.63	19.46	19.61	19.80	19.63	19.58	19.52	19.91	19.69	19.63
8[*]	304	622.13	36.45	36.70	36.67	37.32	37.43	37.18	37.24	37.23	37.54	37.51	37.04	37.79	37.36	37.32

Source: Copyright 2013 from *Handbook of Automated Essay Evaluation: Current Applications and New Directions* by Mark D. Shermis and Jill Burstein. Reproduced by permission of Taylor and Francis Group, LLC, a division of Informa plc.

H1, human rater 1; H2, human rater 2; RS, resolved score (based on human ratings); AIR, American Institutes for Research; CMU-TELEDIA, Carnegie Mellon University; CTB, CTB McGraw-Hill; ETS, Educational Testing Service; MI, Measurement, Inc.; MM, MetaMetrics; PKT, Pearson Knowledge Technologies; PM, Pacific Metrics; VL, Vantage Learning.

[*] RS score was obtained by summing the equally weighted ratings from H1 and H2. To be on the same scale as the RS, the original ratings for H1 and H2 were doubled.

[t] For data set #2, the first rater determined the score assignment. The second rater was employed as a "read behind", but did not influence the score assignment.

Table 6
Test set standard deviations.

Essay set	*N*	*M* # of words	H1	H2	RS	AIR	CMU	CTB	ETS	MI	MM	PKT	PM	VL	AES mean	AES median
1[*]	589	366.40	1.64	1.68	1.54	1.23	1.45	1.26	1.54	1.51	1.57	1.34	1.44	1.29	1.40	1.44
2a[t]	600	381.19	–	0.78	0.77	0.67	0.84	0.65	0.79	0.69	0.83	0.83	0.78	0.64	0.75	0.78
2b[t]	600	381.19	–	0.73	0.75	0.68	0.83	0.67	0.69	0.84	0.80	0.68	0.72	0.65	0.73	0.69
3	568	108.69	0.79	0.78	0.85	0.76	0.91	0.75	0.79	0.89	0.81	0.72	0.83	0.77	0.80	0.79
4	586	94.39	0.89	0.90	0.95	0.83	0.97	0.88	1.00	0.86	1.12	0.88	0.95	0.82	0.92	0.88
5	601	122.29	0.96	0.97	0.95	0.89	1.00	0.89	1.02	1.08	1.08	0.88	0.94	0.93	0.97	0.94
6	600	153.64	0.90	0.86	0.87	0.82	1.01	0.73	0.95	0.95	1.06	0.85	0.88	0.78	0.89	0.88
7[*]	495	171.28	6.40	6.31	5.89	4.17	6.37	5.30	6.27	6.43	6.51	5.17	5.71	4.99	5.66	5.71
8[*]	304	622.13	5.93	5.68	5.19	4.11	4.44	3.83	4.52	5.38	5.91	4.63	5.16	4.21	4.69	4.52

Source: Copyright 2013 from *Handbook of Automated Essay Evaluation: Current Applications and New Directions* by Mark D. Shermis and Jill Burstein. Reproduced by permission of Taylor and Francis Group, LLC, a division of Informa plc.

H1, human rater 1; H2, human rater 2; RS, resolved score (based on human ratings); AIR, American Institutes for Research; CMU-TELEDIA, Carnegie Mellon University; CTB, CTB McGraw-Hill; ETS, Educational Testing Service; MI, Measurement, Inc.; MM, MetaMetrics; PKT, Pearson Knowledge Technologies; PM, Pacific Metrics; VL, Vantage Learning.

[*] RS score was obtained by summing the equally weighted ratings from H1 and H2. To be on the same scale as the RS, the original ratings for H1 and H2 were doubled.

[t] For data set #2, the first rater determined the score assignment. The second rater was employed as a "read behind", but did not influence the score assignment.

Table 7
Test set exact agreements.

Essay Set	*N*	*M* # of Words	H1	H2	H1H2	AIR	CMU	CTB	ETS	MI	MM	PKT	PM	VL	AES mean	AES median
1[*]	589	366.40	0.64	0.64	0.64	0.44	0.44	0.44	0.42	0.46	0.31	0.43	0.43	0.47	0.43	0.44
2a[t]	600	381.19	–	0.76	0.76	0.68	0.64	0.70	0.69	0.70	0.55	0.64	0.68	0.70	0.66	0.68
2b[t]	600	381.19	–	0.73	0.73	0.68	0.59	0.66	0.69	0.66	0.55	0.66	0.67	0.69	0.65	0.66
3	568	108.69	0.89	0.83	0.72	0.68	0.70	0.66	0.69	0.72	0.63	0.61	0.69	0.69	0.67	0.69
4	586	94.39	0.87	0.89	0.76	0.65	0.68	0.64	0.66	0.72	0.47	0.60	0.64	0.70	0.64	0.65
5	601	122.29	0.77	0.79	0.59	0.71	0.67	0.68	0.65	0.68	0.47	0.68	0.65	0.71	0.66	0.68
6	600	153.64	0.80	0.81	0.63	0.67	0.61	0.63	0.62	0.69	0.51	0.64	0.68	0.69	0.64	0.64
7[*]	495	171.28	0.28	0.28	0.28	0.10	0.15	0.12	0.12	0.17	0.07	0.09	0.12	0.12	0.12	0.12
8[*]	304	622.13	0.35	0.35	0.29	0.12	0.26	0.23	0.17	0.16	0.08	0.14	0.20	0.10	0.16	0.16

Source: Copyright 2013 from *Handbook of Automated Essay Evaluation: Current Applications and New Directions* by Mark D. Shermis and Jill Burstein. Reproduced by permission of Taylor and Francis Group, LLC, a division of Informa plc.

H1, human rater 1; H2, human rater 2; H1H2, human rater1, human rater 2; AIR, American Institutes for Research; CMU-TELEDIA, Carnegie Mellon University; CTB, CTB McGraw-Hill; ETS, Educational Testing Service; MI, Measurement, Inc.; MM, MetaMetrics; PKT, Pearson Knowledge Technologies; PM, Pacific Metrics; VL, Vantage Learning.

[*] RS score was obtained by summing the equally weighted ratings from H1 and H2. To be on the same scale as the RS, the original ratings for H1 and H2 were doubled.

[t] For data set #2, the first rater determined the score assignment. The second rater was employed as a "read behind", but did not influence the score assignment.

Table 8
Test set exact and adjacent agreements.

Essay set	*N*	*M* # of words	H1	H2	H1H2	AIR	CMU	CTB	ETS	MI	MM	PKT	PM	VL	AES mean	AES median
1[*]	589	366.40	0.99	0.99	0.99	0.99	0.99	0.98	0.99	0.99	0.95	0.99	0.99	0.99	0.98	0.99
2a[t]	600	381.19	–	1.00	1.00	1.00	0.99	1.00	1.00	1.00	0.99	0.99	1.00	1.00	1.00	1.00
2b[t]	600	381.19	–	1.00	1.00	0.99	0.98	0.99	1.00	0.99	0.97	0.99	1.00	1.00	0.99	0.99
3	568	108.69	1.00	1.00	1.00	0.98	0.97	0.98	0.98	0.97	0.97	0.98	0.97	0.99	0.98	0.98
4	586	94.39	1.00	1.00	1.00	0.99	0.99	0.98	0.99	0.99	0.96	0.99	0.98	0.99	0.98	0.99
5	601	122.29	1.00	1.00	0.98	0.99	0.99	0.99	0.99	0.99	0.93	0.99	0.99	1.00	0.98	0.99
6	600	153.64	1.00	1.00	0.99	0.99	0.99	0.97	0.99	1.00	0.95	1.00	1.00	1.00	0.99	0.99
7[*]	495	171.28	0.55	0.55	0.55	0.47	0.52	0.50	0.52	0.56	0.38	0.50	0.52	0.56	0.50	0.52
8[*]	304	622.13	0.53	0.52	0.49	0.52	0.51	0.51	0.52	0.52	0.41	0.52	0.48	0.53	0.50	0.52

Source: Copyright 2013 from *Handbook of Automated Essay Evaluation: Current Applications and New Directions* by Mark D. Shermis and Jill Burstein. Reproduced by permission of Taylor and Francis Group, LLC, a division of Informa plc.

H1, human rater 1; H2, human rater 2; H1H2, human rater1, human rater 2; AIR, American Institutes for Research; CMU-TELEDIA, Carnegie Mellon University; CTB, CTB McGraw-Hill; ETS, Educational Testing Service; MI, Measurement, Inc.; MM, MetaMetrics; PKT, Pearson Knowledge Technologies; PM, Pacific Metrics; VL, Vantage Learning.

[*] RS score was obtained by summing the equally weighted ratings from H1 and H2. To be on the same scale as the RS, the original ratings for H1 and H2 were doubled.

[t] For data set #2, the first rater determined the score assignment. The second rater was employed as a "read behind", but did not influence the score assignment.

Table 9
Test set kappas.

Essay Set	*N*	*M* # of words	H1	H2	H1H2	AIR	CMU	CTB	ETS	MI	MM	PKT	PM	VL	AES mean	AES median
1[*]	589	366.40	0.53	0.53	0.45	0.29	0.29	0.25	0.28	0.33	0.16	0.29	0.27	0.32	0.28	0.29
2a[t]	600	381.19	–	0.62	0.62	0.46	0.44	0.49	0.51	0.51	0.30	0.43	0.48	0.50	0.46	0.48
2b[t]	600	381.19	–	0.56	0.56	0.46	0.35	0.42	0.49	0.46	0.27	0.43	0.45	0.48	0.42	0.45
3	568	108.69	0.83	0.77	0.57	0.52	0.56	0.50	0.54	0.59	0.45	0.43	0.55	0.53	0.52	0.53
4	586	94.39	0.82	0.84	0.65	0.49	0.56	0.50	0.53	0.60	0.30	0.44	0.50	0.58	0.50	0.50
5	601	122.29	0.69	0.71	0.44	0.59	0.55	0.55	0.51	0.56	0.28	0.54	0.51	0.59	0.52	0.55
6	600	153.64	0.70	0.71	0.45	0.49	0.44	0.40	0.44	0.55	0.31	0.46	0.51	0.51	0.46	0.46
7[*]	495	171.28	0.23	0.23	0.18	0.05	0.09	0.07	0.08	0.12	0.03	0.05	0.07	0.07	0.07	0.07
8[*]	304	622.13	0.26	0.26	0.16	0.06	0.13	0.11	0.08	0.10	0.04	0.09	0.11	0.04	0.08	0.09

Source: Copyright 2013 from *Handbook of Automated Essay Evaluation: Current Applications and New Directions* by Mark D. Shermis and Jill Burstein. Reproduced by permission of Taylor and Francis Group, LLC, a division of Informa plc.

H1, human rater 1; H2, human rater 2; H1H2, human rater1, human rater 2; AIR, American Institutes for Research; CMU-TELEDIA, Carnegie Mellon University; CTB, CTB McGraw-Hill; ETS, Educational Testing Service; MI, Measurement, Inc.; MM, MetaMetrics; PKT, Pearson Knowledge Technologies; PM, Pacific Metrics; VL, Vantage Learning.

[*] RS score was obtained by summing the equally weighted ratings from H1 and H2. To be on the same scale as the RS, the original ratings for H1 and H2 were doubled.

[t] For data set #2, the first rater determined the score assignment. The second rater was employed as a "read behind", but did not influence the score assignment.

Table 10
Test set quadratic weighted kappas.

Essay set	*N*	*M* # of words	H1	H2	H1H2	AIR	CMU	CTB	ETS	MI	MM	PKT	PM	VL	AES mean	AES median
1[*]	589	366.40	0.77	0.78	0.73	0.78	0.79	0.70	0.82	0.82	0.66	0.79	0.76	0.78	0.77	0.78
2a[t]	600	381.19	–	0.80	0.80	0.68	0.70	0.68	0.74	0.72	0.62	0.70	0.72	0.70	0.70	0.70
2b[t]	600	381.19	–	0.76	0.76	0.66	0.63	0.63	0.69	0.70	0.55	0.65	0.69	0.68	0.66	0.66
3	568	108.69	0.92	0.89	0.77	0.72	0.74	0.69	0.72	0.75	0.65	0.65	0.73	0.73	0.71	0.72
4	586	94.39	0.93	0.94	0.85	0.75	0.81	0.76	0.80	0.82	0.67	0.74	0.76	0.79	0.77	0.76
5	601	122.29	0.89	0.90	0.74	0.82	0.81	0.80	0.81	0.83	0.64	0.80	0.78	0.83	0.80	0.81
6	600	153.64	0.89	0.89	0.74	0.76	0.76	0.64	0.75	0.81	0.65	0.75	0.78	0.76	0.74	0.76
7[*]	495	171.28	0.78	0.77	0.72	0.67	0.77	0.74	0.81	0.84	0.58	0.77	0.80	0.81	0.76	0.77
8[*]	304	622.13	0.75	0.74	0.61	0.69	0.65	0.60	0.70	0.73	0.63	0.69	0.68	0.68	0.67	0.68

Source: Copyright 2013 from *Handbook of Automated Essay Evaluation: Current Applications and New Directions* by Mark D. Shermis and Jill Burstein. Reproduced by permission of Taylor and Francis Group, LLC, a division of Informa plc.

H1, human rater 1; H2, human rater 2; H1H2, human rater1, human rater 2; AIR, American Institutes for Research; CMU-TELEDIA, Carnegie Mellon University; CTB, CTB McGraw-Hill; ETS, Educational Testing Service; MI, Measurement, Inc.; MM, MetaMetrics; PKT, Pearson Knowledge Technologies; PM, Pacific Metrics; VL, Vantage Learning.

[*] RS score was obtained by summing the equally weighted ratings from H1 and H2. To be on the same scale as the RS, the original ratings for H1 and H2 were doubled.

[t] For data set #2, the first rater determined the score assignment. The second rater was employed as a "read behind", but did not influence the score assignment.

Table 11
Test set Pearson product moment correlation, *r*.

Essay set	*N*	*M* # of words	H1	H2	H1H2	AIR	CMU	CTB	ETS	MI	MM	PKT	PM	VL	AES mean	AES median
1[*]	589	366.40	0.93	0.93	0.73	0.80	0.79	0.71	0.82	0.82	0.66	0.80	0.76	0.80	0.78	0.80
2a[t]	600	381.19	–	0.80	0.80	0.68	0.71	0.69	0.74	0.72	0.62	0.70	0.72	0.71	0.70	0.71
2b[t]	600	381.19	–	0.76	0.76	0.67	0.64	0.64	0.70	0.71	0.55	0.65	0.69	0.69	0.66	0.67
3	568	108.69	0.92	0.89	0.77	0.72	0.74	0.69	0.72	0.75	0.65	0.66	0.73	0.73	0.71	0.72
4	586	94.39	0.94	0.94	0.85	0.76	0.81	0.76	0.82	0.82	0.68	0.75	0.76	0.80	0.78	0.76
5	601	122.29	0.89	0.90	0.75	0.82	0.81	0.80	0.81	0.84	0.65	0.80	0.78	0.83	0.80	0.81
6	600	153.64	0.89	0.89	0.74	0.76	0.77	0.65	0.77	0.81	0.66	0.75	0.78	0.77	0.75	0.77
7[*]	495	171.28	0.93	0.93	0.72	0.71	0.78	0.75	0.81	0.84	0.58	0.78	0.80	0.82	0.77	0.78
8[*]	304	622.13	0.87	0.88	0.61	0.71	0.66	0.63	0.71	0.73	0.62	0.70	0.68	0.72	0.69	0.70

Source: Copyright 2013 from *Handbook of Automated Essay Evaluation: Current Applications and New Directions* by Mark D. Shermis and Jill Burstein. Reproduced by permission of Taylor and Francis Group, LLC, a division of Informa plc.

H1, human rater 1; H2, human rater 2; H1H2, human rater1, human rater 2; AIR, American Institutes for Research; CMU-TELEDIA, Carnegie Mellon University; CTB, CTB McGraw-Hill; ETS, Educational Testing Service; MI, Measurement, Inc.; MM, MetaMetrics; PKT, Pearson Knowledge Technologies; PM, Pacific Metrics; VL, Vantage Learning.

[*] RS score was obtained by summing the equally weighted ratings from H1 and H2. To be on the same scale as the RS, the original ratings for H1 and H2 were doubled.

[t] For data set #2, the first rater determined the score assignment. The second rater was employed as a "read behind", but did not influence the score assignment.

difference. As mentioned above, for data sets #1, #7, and #8, the calculation for adjacent agreement was slightly different than for the other data sets. These data sets came from states where the resolved score was the unweighted sum of the two human raters. While the calculation of adjacent agreement could be based on the original scale for the two human raters, the machine score predictions were on the unweighted summed (i.e., doubled) scale. In order to compensate for this scaling difference, the calculation of adjacent agreement for the machine scores were predicated on an adjacent score difference of two points, not one. Adjacent agreements are shown on Table 8, and for data sets #1–#6, these agreements range in the mid-high 90s. Data sets #7 and #8 are lower, but are consistent with human rater performance.

Kappa is a measure of agreement that takes into consideration agreement by chance alone. For the resolved score, based on human ratings, the ranges ran from 0.16 on data set #8 to 0.65 on data set #4. Overall machine performance ran from 0.04 on data set #8 to 0.59 on several data sets. The ranges are shown in Table 9. In general, performance on kappa was lower with the exception of essay prompts #5. On this data set, the AES engines, as a group, matched or exceeded human rater agreement coefficients.

Kappa is typically applied as an agreement measure when the scoring categories have no ordinality associated with them. Quadratic-weighted kappa is appropriate when the categories have some underlying trait that increases as the scale associated with the categories increase. Rather than treating all disagreements the same (as kappa does), quadratic weighted kappa takes into account the weighted distance between the pairs of ratings so that two ratings that are far apart have a more negative impact on the measure than two ratings that are not exact agreements, but are closer. The rubrics used by the states meet the assumptions for applying this metric. Human quadratic-weighted kappas ranged from 0.61 to 0.85 and were closely followed by the machine ranges which went from 0.60 to 0.84. Table 10 shows the values of quadratic weighted kappa across the data sets for each vendor.

The values for the correlation coefficients generally mirror that of quadratic weighted kappa. These values might have been higher except that the vendors were asked to predict integer values only. Had they been given leeway to predict values containing decimal places, the correlation might have been higher. The correlation values are given in Table 11.

3.1. Study 2 public competition

Shortly after the launch of the commercial demonstration, a public competition was initiated using a system provided by Kaggle; a web-based platform for data prediction competitions where organizations can post data for analysis by data scientists throughout the world. This competition used quadratic weighted kappa as the sole evaluation criterion and challenged data scientists to maximize the value of this agreement measure with human scores. The prizes for the top performers on this metric were $60,000 for first place, $30,000 for second place and $10,000 for third place. As such, the presentation of this data analysis challenge was presented as a data modeling competition (onward Kaggle Inc., 2010).

The goal of the public competition was to encourage and make available to the commercial vendors new software technology or programming approaches that would improve the overall agreement of the machine scoring algorithms with human scores. The public competition, which ran in parallel to the commercial demonstration, involved 159 teams of data scientists from around the world.

There were minor, but important differences between the commercial demonstration and the public competition. First, the data scientists had approximately three months to create and train their engines rather than the one month allocated to the commercial vendors (who had existing scoring engines). In this process they used the same training data that the vendors used, but the exception that the data provided to the public competitors had to undergo an anonymization step. This was intended to address concerns that individual students might be identified from details of their essay used in the competition, despite the fact that all of the prompts were designed to elicit either factual or innocuous information. To limit the possibility of a student being identified, the essays were processed by a Kaggle-written anonymizing routine that included the use of the Stanford Named Entity Recognizer (Finkel, Grenager, and Manning, 2005). Some essays were eliminated from the data set due to the

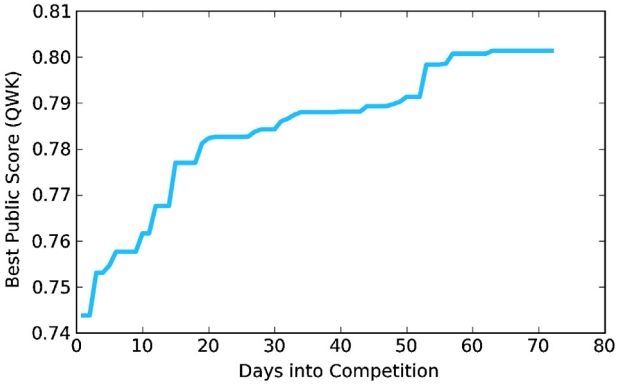

Fig. 1. Public competitor performance over time based on quadratic weighted kappa.

inability of the anonymizing routine to sufficiently mask identifying information provided in those essays.

In order to assess the content equivalency of the original and anonymized data, a small internal study was completed with the *LightSIDE* engine to determine the degree to which there might be differences. The reasoning was that the anonymized data might be harder to model than the original data since it would contain less specific information. However, the *LightSIDE* model showed only a slight drop from .763 to .759 on quadratic weighted kappa across the data sets that was not statistically significant ($p = .15$).

Public competitors had two weeks to process the test data which, like the training data, was the same data used to demonstrate performance from vendors and was likewise anonymized. Unlike the vendor demonstration, this processing was iterative and participants could see daily updates on how their system performed on the test set relative to the other competitors on the Kaggle leaderboard. This leaderboard displayed daily rankings of the competitors based on the agreement (in quadratic weighted kappa) with the human scores, which remained unknown to participants. This allowed participants to continue to refine their procedures and test the performance of their approach against the test data throughout the two week window and to try and outperform their competitors.

The final stage of the competition required that participants process a third data set, the validation set, and submit the essay scores to the Kaggle system. This performance served as the basis for determining the final ranking, and winners, for the competition based on the highest values of quadratic weighted kappa agreement with human scores. Once the winners were determined, a final step included conducting an independent verification of the competitor source code by Kaggle on a set of data randomly selected from the initial pool of 22,029 essays and not used elsewhere in the study. This step was designed to ensure the validity and integrity of the submitted code. Fig. 1 shows the progress of the public competitors over the three months in achieving the final quadratic weighted kappa value.

The outcome of evaluation on the final data set had the top three public competitors achieving an average quadratic weighted kappa of .814, .808, and .806. The first place team consisted of three members: a particle physics engineer from Oxford, a German computer scientist, and a weather analyst working in Washington, DC, at NOAA. The second place team was composed of five team members from Australia, the United States, and Canada. The third place team had two members with business backgrounds.

4. Discussion and study limitations

Several limitations of this study are worth re-emphasizing in the interpretation of results as these may have impacted different aspects of machine scoring performance as well as influencing

interpretation of results. One limitation is that six of the essay sets had to be transcribed from handwriting to ASCII text in order to be processed by automated essay scoring. This step had the possibility of introducing transcription and formatting errors. Further, it is possible that students write differently when writing using pencil and paper than when keyboarding responses. When combined with the elimination of paragraph boundaries there could be non-trivial differences between the data examined in this study and data obtained from direct keyboard entry.

Another limitation stems from the high degree of variability in state practice that is represented in the data sets. The variety of rubrics, human raters, and rules for generating resolved scores applied by different states may have impacted scores and score scales in ways that could have impacted score modeling efforts by machine scoring engines. Together, these limitations suggest that the performance of automated essay scoring engines reported here could be improved upon with data representing more optimal conditions.

Other limitations of this study are inherent in the scope of the study and the resultant constraints on conclusions to be made from these results. These include the following:

- Agreement with human ratings is not necessarily the best or only measure of students' writing proficiency (or the evidence of proficiency in an essay). We can look at other measures as well, including alternate-form reliabilities, and correlations with external measures such as state assessment scores, portfolio assessments, or course grades. The limitation of human scoring as a yardstick for automated scoring is underscored by the human ratings used for some of the tasks in this study, which displayed strange statistical properties and in some cases were in conflict with documented adjudication procedures (Bennett, 2011; Williamson et al., 2012).
- The study did not control for inter-topic reliability, nor was there any attempt to evaluate the impact of discourse mode pairing, prompt type, the prompt topic, circumstances of administration, or other characteristics that could impact how students perform on essays (Ramineni & Williamson, 2013).
- Another issue not addressed by this study is the question of construct validity. A predictive model may do a good job of matching human scoring behavior, but for reasons unrelated (or unsatisfactorily related) to the construct of interest. If accurate predictions of score are achieved by features and methods that do not bear any plausible relationship to the competencies and construct that the item aims to assess, then this prediction, accurate as it may be, is not sufficiently representative of the construct to warrant test use. To the extent that such models are used, such statistical surrogation will limit the validity argument for the assessment (Kane, 2006; Scriven, 1987).
- A related issue is that of potential signaling effects (Bowen, Chingos, & McPherson, 2009, pp. 131–133) on test-taking behavior and instruction. Before using a system operationally, some consideration will need to be given to the question of how the measures used in scoring might be subject to manipulation by test-takers and coaches with an interest in maximizing scores and what message writing to a machine sends. This study does not conduct any evaluations relevant to this question and the data analyzed were produced by students who were writing for a human evaluator. If students knowingly write for an automated scoring system they, and those who coach them, may modify their writing practice to target what they believe the machine scoring values.
- An important aspect of system performance to evaluate before operational use is fairness – whether subgroups of interest are treated differentially by the scoring methodology. This study does not conduct any evaluations relevant to this question though some work has already been performed in this area (Bridgeman, Tripani, & Attali, 2012).
- This scoring study targeted variation among states, prompt types, and scoring procedures that were intended to represent a range of possibilities in essay scoring. However, the data selection procedures were not scientifically sampled and as such cannot be used as the basis for generalizing student writing performance to a larger population as those studies commonly associated with group score assessments (Mazzeo, Lazer, & Ziesky, 2006).
- Finally, the scope of the study was restricted to the question of how well automated scoring of essays can replicate the scores of human raters across different types of prompts and scoring procedures. It does not address the important and very relevant questions of how well the use of essays such as these in a testing program reflects educational practice and aspirations for encouraging writing proficiency (Klobucar et al., 2012).

Despite these limitations some general conclusions are appropriate from these results. 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. While the average performance of vendors does not always match the performance of human raters, the results of the top two to three vendors was consistently good and occasionally exceeded human rating performance. The intent of the two consortia would be to employ the results from the top machine scoring performers, not to incorporate the results of an average performing scoring engine.

5. Summary

Almost 50 years ago, Ellis Page (Page, 1966) prophesied the coming of a "teacher's helper" that would grade essays by computer. Only seven years after that prophecy Page and his colleagues at the University of Connecticut created Project Essay Grade (*PEG*), the first working automated essay scoring engine (Ajay, Tillett, & Page, 1973). The engine worked reasonably well, but the technology for inputting text (e.g., punched cards and tape) had not caught up with Page's ideas (Shermis & Burstein, 2003). *PEG* was shelved for twenty years until the development of microcomputers and ubiquitous access to the Internet. PEG was resurrected in the early 90s and soon thereafter several commercial and not-for-profit vendors began developing an array of machine scoring engines for the English language.

The Hewlett Foundation-sponsored studies described in this article have shown that, with the appropriate qualifications to the validity argument, in a high-stakes testing environment machine-predicted scores came close to matching the distributional and agreement characteristics of scores assigned by human raters (Mislevy, 2007). In order to further establish the validity of machine scores, the kind of experimentation described in this paper must be combined with validation procedures emphasizing explanation (construct, task, and scoring investigation); evaluation (human and automated score agreement); extrapolation (external measure associations); generalization (automated score generalizability across alternate tasks and test forms); and utilization (claims, disclosures, and consequences) (Williamson et al., 2012). The more stakeholders become involved in the validation process, the more likely AES will be understood, refined, and used to target specific traits of the writing construct that are important in instruction and assessment.

Author notes

An earlier version of this work was presented as a paper presented at National Council on Measurement in Education in Vancouver, Canada, 2012. This work was supported through funding from the William and Flora Hewlett Foundation. The opinions expressed in this paper are those of the author and do not necessarily represent the policy or views of the William and Flora Hewlett Foundation or their Board of Directors. The author would like to thank Jaison Morgan (The Common Pool), Tom Vander Ark, Lynn Van Deventer (OpenEd Solutions), Ben Hamner (Kaggle), Tony Albert (SBAC), and Jeff Nellhouse (PARCC) for their tireless efforts in executing this study. The author is also indebted to two anonymous reviewers, and to David Williamson and Norbert Elliot for their encouragement and insightful comments that helped shape the final version of this manuscript.

References

Ajay, H. B., Tillett, P. I., & Page, E. B. (1973). *Analysis of essays by computer (AEC-II)*. Washington, DC: U.S. Department of Health, Education, and Welfare, Office of Education, National Center for Educational Research and Development.231

Applebee, A. N., & Langer, J. (2009). What is happening in the teaching of writing? *English Journal*, *98* (5), 18–22.

Attewell, P., Lavin, D., Domina, T., & Levey, T. (2006). New evidence on college remediation. *Journal of Higher Education*, *55* (5), 886–924.

Bennett, R. E. (2011). *Automated scoring of constructed-response literacy and mathematics items Advancing Consortium Assessment Reform (ACAR)*. Washington, DC: Arabella Philanthropic Advisors.

Bowen, W. G., Chingos, M. M., & McPherson, M. S. (2009). *Crossing the finish line: Completing college at America's public universities*. Princeton, NJ: Princeton University Press.

Bridgeman, B., Tripani, C., & Attali, Y. (2012). Comparison of human and machine scoring of essays: Differences by gender, ethnicity, and country. *Applied Measurement in Education*, *25*, 27–40.

Burstein, J., Kukich, K., Wolff, S., Lu, C., & Chodorow, M. (1998). Computer analysis of essays. In: *Proceedings of Proceedings of the NCME Symposium on Automated Scoring*.

Condon, W. (2013). Large-scale assessment, locally-developed measures, and automated scoring of essays: Fishing for red herrings? *Assessing Writing*, *18*, 100–108.
Elliot, S. (2003). IntelliMetric: From here to validity. In: M. D. Shermis & J. Burstein (Eds.), *Automated essay scoring: A cross-disciplinary perspective* (pp. 71–86). Mahwah, NJ: Erlbaum Associates.
Finkel, J. R., Grenager, T., & Manning, C. (2005). Incorporating non-local information into information extraction systems by Gibbs Sampling. In: *Paper presented at the 43rd annual meeting of the Association for Computational Linguistics* http://nlp.stanford.edu/~manning/papers/gibbscrf3.pdf
Kane, M. T. (2006). Validation. In: R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64). Westport, CT: American Council on Education/Praeger.
Klobucar, A., Deane, P., Elliot, N., Ramineni, C., Deess, P., & Rudniy, A. (2012). Automated essay scoring and the search for valid writing assessment. In: C. Bazerman, C. Dean, J. Early, K. Lunsford, S. Null, P. Rogers, & A. Stansell (Eds.), *International advances in writing research: Cultures, places, measures* (pp. 103–119). Fort Collins, CO: The WAC Clearninghouse.
Landauer, T. K., Foltz, P. W., & Laham, D. (1998). Introduction to latent semantic analysis. *Discourse Processes*, *25* (2–3), 259–284.
Landauer, T. K., Laham, D., Rehder, B., & Schreiner, M. E. (1997). How well can passage meaning be derived without using word order? A comparison of latent semantic analysis and humans. In: M. G. Shafto & P. Langley (Eds.), *Proceedings of the 19th annual meeting of the Cognitive Science Society* (pp. 412–417). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.
Mayfield, E., & Rosé, C. (2010). An interactive tool for supporting error analysis for text mining. In: *Paper presented at the demonstration session at the international conference of the North American Association for Computational Linguistics*.
Mayfield, E., & Rosé, C. (2013). LightSIDE: Open source machine learning for text. In: M. D. Shermis & J. Burstein (Eds.), *Handbook of automated essay assessment: Current applications and new directions* (pp. 124–135). New York, NY: Routledge.
Mazzeo, J., Lazer, S., & Ziesky, M. J. (2006). Monitoring educational progress with group-score assessments. In: R. L. Brennan (Ed.), *Educational measurement* (pp. 469–481). Westport, CT: American Council on Education/Praeger.
McCurry, D. (2010). Can machine scoring deal with broad and open writing tests as well as human raters? *Assessing Writing*, *15*, 118–129.
Mislevy, R. J. (2007). Validity by design. *Educational Researcher*, *36* (8), 463–469.
onward Kaggle Inc. (2010). Home Page. Retrieved from http://www.kaggle.com
Page, E. B. (1966). The imminence of grading essays by computer. *Phi Delta Kappan*, *48*, 238–243.
Porter, A., McMaken, J., Hwang, J., & Yang, R. (2011). Common Core Standards: The new U.S. intended curriculum. *Educational Researcher*, *40* (3), 103–116.
Ramineni, C., & Williamson, D. M. (2013). Automated essay scoring: Psychometric guidelines and practices. *Assessing Writing*, *18*, 25–39.
Salton, G., Wong, A., & Yang, C. S. (1975). A vector space model for automatic indexing. *Communications of the ACM*, *18* (11), 613–620.
Scriven, M. (1987). Validity in personnel evaluation. *Journal of Personnel Evaluation in Education*, *1*, 9–23.
Shermis, M. D., & Burstein, J. (2003). *Automated essay scoring: A cross-disciplinary perspective*. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.
Shermis, M. D., & Burstein, J. (Eds.). (2013). *Handbook of automated essay evaluation: Current applications and new directions*. New York, NY: Routledge.
Shermis, M. D., & Hamner, B. (2012). Contrasting state-of-the-art automated scoring of essays: Analysis. In: *Paper presented at the National Council of Measurement in Education*.
Shermis, M. D., & Hamner, B. (2013). Contrasting state-of-the-art automated scoring of essays. In: M. D. Shermis & J. Burstein (Eds.), *Handbook of automated essay evaluation: Current applications and new directions* (pp. 313–346). New York, NY: Routledge.
Stemler, S. E. (2004). A comparison of consensus, consistency, and measurement approaches to estimating interrater reliability. *Practical Assessment, Research & Evaluation*, *9* (4) http://PAREonline.net/getvn.asp?v=9&n=4
Tucker, B. (2009). The next generation of testing. *Multiple Measures*, *67* (3), 48–53.
Weigle, S. C. (2013). English language learners and automated scoring of essays: Critical considerations. *Assessing Writing*, *18*, 85–99.
Weir, C. J. (2005). Limitations of the Common European Framework for developing comparable examinations and tests. *Language Testing*, *22* (3), 261–300.
Williamson, D. M., Xi, X., & Breyer, F. J. (2012). A framework for the evaluation and use of automated essay scoring. *Educational Measurement: Issues and Practice*, *31* (1), 2–13.

Mark D. Shermis, Ph.D., is a professor at the University of Akron and the principal investigator of the Hewlett Foundation-funded Automated Scoring Assessment Prize (ASAP) program. He has published extensively on machine scoring and recent co-authored the textbook *Classroom Assessment in Action* with Francis DiVesta. Shermis is a fellow of the American Psychological Association (Division 5) and the American Educational Research Association.