Educational Assessment, 20:46–65, 2015
Copyright © Taylor & Francis Group, LLC
ISSN: 1062-7197 print/1532-6977 online
DOI: 10.1080/10627197.2015.997617

Routledge
Taylor & Francis Group

Contrasting State-of-the-Art in the Machine Scoring of Short-Form Constructed Responses

Mark D. Shermis
University of Houston—Clear Lake

This study compared short-form constructed responses evaluated by both human raters and machine scoring algorithms. The context was a public competition on which both public competitors and commercial vendors vied to develop machine scoring algorithms that would match or exceed the performance of operational human raters in a summative high-stakes testing environment. Data ($N = 25{,}683$) were drawn from three different states, employed 10 different prompts, and were drawn from two different secondary grade levels. Samples ranging in size from 2,130 to 2,999 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. Machine performance on all of the agreement measures failed to match that of the human raters. The current study concluded with recommendations on steps that might improve machine-scoring algorithms before they can be used in any operational way.

INTRODUCTION

This study is an examination of machine scoring performance on short-form constructed responses sponsored by the Hewlett Foundation. Both commercial and independent competitors from throughout the world created scoring engines to compare summative scoring performance against that of trained human raters in the evaluation of U.S. statewide high-stakes testing prompts across a number of different educational domains. The study was conducted at the behest of the two major Race-to-the-Top consortia, Partnership for Assessment of Readiness for College and Careers (PARCC) and Smarter Balanced that represent efforts of 35 subscribing states to create and score new assessments based on the Common Core State Standards. One concern had to do with the states' capacity to score the millions of written documents that will be generated by the new statewide testing programs once the assessments become operational. For example, if the state of Florida (which has approximately 180,000 students per grade level and tests in Grades 3–10) were to require only five written answers per year for their new high-stakes assessments, the requirement would result in the generation of approximately 7.2 million responses. The ability of testing companies to adequately meet the human scoring demands for such volume is very limited.

Correspondence should be sent to Mark D. Shermis, School of Education, 1236 Bayou, University of Houston—Clear Lake, Houston, TX 77058. E-mail: mshermis@uhcl.edu

The Hewlett Trials Phase 1 (Part 1)

This study is a follow-up to the Hewlett Foundation–sponsored demonstration that contrasted eight commercial vendors and one university laboratory's performance on automated essay scoring with that of human raters (Phase 1; Shermis & Hamner, 2012, 2013). That study employed eight different essay sets drawn from six states representing the PARCC and Smarter Balanced consortia. In that first study, four of the essays were "source based." A student was asked to read an artifact and then respond with an essay. The remaining four essays reflected prompts of a more traditional variety (i.e., narrative, descriptive, persuasive). More than 17,000 essays were randomly divided into two sets, a training set ($n = 13{,}336$) in which the vendors had 1 month to model the data and a test set for which they were required to make score predictions within a 59-hr period. The training set consisted of two human rater scores, a so-called resolved score, and the text of the essay. The resolved score reflected the final score assigned by the state. To resolve differences between the two human raters, states applied a set of adjudication rules (e.g., ask a third rater to evaluate the essay) that may or may not have reflected the judgments of the original human raters. In data sets 2a and 2b of this study, the first human rater determined the essay score assignment and the second score was used only as a check on the reliability of the first human rater. The test set consisted only of the text of the essay. Six of the eight essays were transcribed from handwritten documents using one of two transcription services. Accuracy rates of transcription were estimated to be greater than 98%. The challenge to the nine teams was to predict scores based on essay performance that matched the ratings assigned as the resolved score.

Performance for Phase 1 was evaluated on five different measures of a single evidentiary criterion—agreement with human raters. One set of measures focused on agreement at the distributional level and the other set on agreement at the individual response level. The individual-response-level measures included exact agreement, exact + adjacent agreement, kappa, quadratic weighted kappa, and the Pearson product moment correlation. The automated essay scoring engines performed well on the distributional measures. With a high degree of consistency, all nine demonstrators were able to replicate the means and standard deviations for the scores assigned by the states. With regard to agreement measures, there was some variability, but the automated essay scoring engines performed well on three of the five measures (exact + adjacent agreement, quadratic weighted kappa, correlation). On the two measures where the performance was not as high (exact agreement and kappa), there was also high variability among the performance of operational human raters. In addition, scaling artifacts attributable to the way the state scores were adjudicated may have contributed to the relative lack of precision on predicted scores. In sum, on five of the seven measures, machine scores reasonably approximated those of human raters, and on a few of the data sets even performed better than their human counterparts. The conclusion of that study was that with additional work, automated essay scoring could be a viable solution in some aspects of evaluating high-stakes long-form writing assessments.

The previous study did not evaluate other aspects of machine scoring that are important in establishing the validity of the predicted scorings including evaluating the construct relevance of the process by which the software makes its predictions, the generalizability of the resultant scores, the relationship between the scores obtained by machine scoring and other independent variables, and the impact on decision making the consequences of using automated scoring

(Bennett, 2011; Williamson, Xi, & Breyer, 2012; Yang, Buckendahl, Juszkiewicz, & Bhola, 2002). Nor did the study evaluate the potential for score bias in group differences, and it employed existing essays from states as the source of information rather than obtaining text as a computer-administered assessment. All of these constraints limit the generalizability of the results.

The Hewlett Trials (Phase 1, Part 2)

A second part (Part 2) to Phase 1 was conducted shortly after the conclusion of the vendor demonstration. The second study was similar to the first, with some minor exceptions. It involved recruiting participants from throughout the world using the Kaggle Forum competition platform (www.kaggle.com) to compete for the $100,000 prize sponsored by the Hewlett Foundation. The international competitors had $2\frac{1}{2}$ months to create their scoring engines, the code for which is now available in the public domain. This was a first attempt to include the collective talent to build knowledge around automated evaluation of written responses (Shermis, Burstein, Elliot, Miel, & Foltz, in press). During the course of the $2\frac{1}{2}$ months, an additional set of randomly selected essays was utilized to establish a public leaderboard, which was used to monitor and display competitors' progress, a feature not available for the vendor demonstration. One other change from Part 1 was that personal identifiers were removed from the writing samples used in the public competition using software developed by Kaggle based on the Stanford Named Entity Anonymizer (Finkel, Grenager, & Manning, 2005). An internal study to Part 1 showed that there were no statistically significant differences in machine score predictions based on the original versus anonymized text (Shermis & Hamner, 2012, 2013).

The group of international competitors had 2 weeks to submit their final entries rather than the shorter time limit established for the commercial vendors. Results showed that human raters established a benchmark of reliability, as measured by quadratic weighted kappa of .75. The best of the commercial vendors was able to achieve a kappa of .78. The public competitors, working on a different but equivalent data set, were able to achieve a kappa of .81. Because the average quadratic weighted kappa had met and exceeded the .70 threshold, the study authors concluded that automated essay scoring held the potential to be a viable method to scoring writing samples in situations where scores were to be used to make inferences about the writing of individual students in summative assessment contexts. The conclusion was based on the assumption that additional research on score validity would be conducted.

Short-Form Constructed Responses

In the current study, PARCC and Smarter Balanced were interested in determining whether machine-scoring algorithms might have a role in evaluating short-form constructed responses in a high-stakes setting. Items of this type address a number of educational domains including English language arts, reading, and science. They might also be relevant to other areas that are not currently assessed in high-stakes testing programs. Items of this nature typically come in two forms. In the first form, the student might be given a short reading passage and asked to respond to a question based on the passage. In this study these items are referred to as source-dependent questions. The second question type might simply ask the student to answer a question that they should be familiar with based on the curriculum of class for which they are enrolled. For example, in biology:

The nervous system interacts with other body systems to maintain homeostasis.

a. Describe how the nervous and respiratory systems interact to maintain homeostasis when a person exercises. Explain how this interaction maintains homeostasis.
b. Describe how the nervous and muscular systems interact to maintain homeostasis when a person's body temperature drops. Explain how this interaction maintains body temperature. (Massachusetts Department of Elementary and Secondary Education, 2011. Used by permission.)

Rubrics for such items typically incorporate a scale range from 0 to 2 or 3 and specify the response elements that need to be in place to receive a specific score. In contrast to essay scoring rubrics, there is little or no emphasis on writing ability but rather the degree to which the content is present and correct (or in correct form). Spelling and mechanical errors are generally overlooked to the degree that the rater can make a reasonable inference that the student "knows" the response at a given performance level.

Word-length ranges for these items vary greatly. Some short-answer responses require a word or a phrase or they may be as long as 125 words. In the context of most high-stakes testing programs, these types of responses generally require one or two sentences to answer and generally range between 30 and 50 words per response. The focus of this study is aligned with evaluating the one- to two-sentence responses.

A few investigations have examined the reliability of the machine-scoring for short-answer responses. Attali, Powers, Freedman, Harrison, and Obetz (2008) described a study where the ETS engine *c-rater* was used in an experiment for the GRE Subject tests in biology and psychology. In the experiment participants received immediate feedback on the correctness of their answers (one to three sentences long) and had an opportunity to revise their responses. Eleven questions were given in biology (N=331) and 11 questions were administered in psychology (N = 640). Results showed that the score reliability coefficients were α = .53 for the biology items (human score α = .57) and an α = .59 for the psychology items (human score α = .61). The authors noted that these subject areas were a "good fit" for *c-rater* in that they had a strong content basis, focus on a limited number of concepts, and required a short answer (Attali et al., 2008). The major concern about the study had to do with the limited number of training essays that were available for model building and a subsequent concern about the quality of models.

In an earlier study with NAEP math items, *c-rater* was used to evaluate a small set of fourth- and eighth-grade exams (N = 100 per question; five questions) where math or logical reasoning was being evaluated. The average number of words for each response was 15. *c-rater* obtained average kappas in the range of .75 compared to .82 for human raters (Leacock & Chodorow, 2003b). No comparisons were provided for mean and standard deviation estimates comparing machine and human rater performance.

In a similar study, Sukkarieh and Blackmore (2009) evaluated seven reading comprehension and five math comprehension items from a Maine high-stakes assessments for seventh- and eighth-grade students. *c-rater* was trained on 130 to 150 responses for each item and blind-tested on 61 to 114 new responses. Human-rater unweighted kappas ranged from 0.71 to 1.0 across both sets of items, and *c-rater obtained* unweighted kappas that ranged from 0.55 to 0.94. The authors identified several technical challenges that included indistinct concepts, uncorrected spelling mistakes (or corrections to an unintended word), unexpected phrase variations that a model did not predict, the insufficiency of similar lexicons (e.g., a definition was needed instead of a synonym), linguistic phenomena that were not addressed in the *c-rater* algorithms, the need

for a reasoning/inference model, the general nature of some model sentences that produced false positives, and inconsistencies of concept-based scoring (Sukkarieh & Blackmore, 2009, p. 294). Nevertheless, the authors characterized the results as "promising."

A pilot study in Indiana that evaluated a high-stakes reading comprehension test was conducted on a sample of responses across seven questions. The average length of response was 2.8 sentences or 43 words. One hundred responses for each question were randomly sampled and scored by both *c-rater* and a human judge. A second human rater was utilized to resolve discrepant scores. The average unweighted kappa was .74 for *c-rater*, and when scores did not match, the *c-rater* score assignment was off by 1 point on a 3-point scale (Leacock & Chodorow, 2003b).

The research to date suggests that although pilot studies of machine scoring for short-form constructed responses looks promising, it would appear as if performance is inconsistent, at least in the K-12 environment. Moreover, there have been no large-scale studies that have compared performance across multiple scoring platforms using a variety of prompts.

The purpose of Phase 2 was to evaluate short-form constructed responses using a slight variation of the Phase 1 design and evaluation metrics. The question was, To what degree machine scoring software predictions agree with the score assignments of operational human raters in a high-stakes environment? This study was executed in the form of a public competition sponsored by the Hewlett Foundation and administered on the Kaggle Forum. The competitor that produced the best verifiable scoring algorithm won.

METHOD

Participants

Student short-form constructed responses ($N = 25{,}683$) were collected for 10 different prompts representing one PARCC state and two Smarter-Balanced states. The short-answer responses were obtained from the states' high-stakes assessments administered during the 2010–2011 school year where there was a second rater score. States use a "read behind" process to evaluate the reliability of human rater performance and take a random sample of constructed responses (proportions differ by states) to accomplish this. To the extent possible, an attempt was made to make the identity of the participating states anonymous. One state was located in the northeastern part of the United States, one from the midwest, and one from the West Coast. Because no demographic information was provided by the states, student characteristics were estimated from a number of different sources, as displayed in Table 1. Student writers were drawn from two different grade levels (8, 10), and the grade-level selection was generally a function of the testing policies of the participating states (e.g., short-answer component as part of a 10th-grade exit exam). The estimated population characteristics suggest that the sample was ethnically diverse and evenly distributed between male and female individuals.

Samples ranging in size from 2,130 to 2,999 were randomly selected from the data sets provided by the states and then randomly divided into three sets: a training set, a test set (aka public leaderboard set), and a validation set (aka private leaderboard set). The training set was used by the vendors to create their scoring models and consisted of a score assigned by a human rater and the text of the response. Most vendors process training essay text by parsing it, dividing it into writing components, adding information based on natural language processing, and then creating models (statistical or response templates) upon which they evaluate new candidate

TABLE 1
Sample Characteristics

	Data Set No.									
Variable	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
State	A				B					C
Grade	10				10					8
Grade-level *N*	44,485				81,245					78,902
N	10,966				11,983					2,734
Training *n*	6,579				7,189					1,640
Test *n* (public leaderboard)	2,195				2,395					546
Validation *n* (private leaderboard)	2,192				2,399					548
Estimated gender (%): Male/Female	51 .2/48.8				51.4/48.6					51.5/48.5
Estimated Race (%): White/Non-White	63.8/36.2				77.8/22.2					61.3/38.7
Estimated free/Reduced lunch (%): Not for free lunch	32.9/67.1				40.0/60.0					43.7/56.3

Note. Gender, race, and free-lunch status for the state populations in question were taken primarily from National Center for Education Statistics, Common Core of Data, (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 web site information or annual reports for each participating state.

responses. Indexing the amount and relevance of content in the response is established through a variety of techniques that reference the content in the training sample against new candidate short answers. The most popular approach was that of Latent Semantic Analysis (Landauer, Foltz, & Laham, 1998) or one of its derivatives, a reduced dimensionality variant of the vector space model that preceded it (Salton, Wong, & Yang, 1975; cf. https://www.kaggle.com/c/asap-sas/details/winners). The public leaderboard set consisted of a written response only and was used as part of a blind test for the score model predictions. During the 2-month course of the competition, the public leaderboard provided competitors feedback on how well their models were performing compared to the other teams. The private leaderboard set consisted of written responses only and was employed to verify the scoring code by the competition administrators. The distribution of the samples was split in the following approximate proportions: 60% training sample, 20% public leaderboard, 20% private leaderboard sample. The actual proportions vary slightly due to the elimination of cases containing either data errors or text anomalies. The distribution of the samples is displayed in Table 1.

Instruments

Three of the short-answer prompts were drawn from the general sciences area, two from biology, and five from English/language arts. Eight of the 10 prompts were "source dependent"—that is, the questions asked in the prompt referred to a source document that students read as part of the assessment. In the test set, average word lengths varied from $M = 22.62$ ($SD = 21.83$) to $M = 58.36$ ($SD = 23.71$). The biology short-answer responses (Data Sets 5 & 6; $M = 23.35$, $SD = 19.70$) were significantly shorter than those from the other prompts ($M = 47.70$, $SD = 25.35$), $t(5098) = 30.54$, $p < .0001$. Testing times allotted for the high-stakes assessments varied depending on the context of the administration. Constructed responses that were

handwritten were given during a 1-day period, whereas those that were typed into a computer had a restricted range of dates for completion.

All the prompts employed a holistic scoring rubric. Each rubric specified multiple criteria for the human rater to consider but asked the rater to make one overall score assignment. The ranges for scores in the rubric ran from 0–2 to 0–3. The scale ranges, scale means, and standard deviations are reported in Table 2. Table 2 shows the characteristics of the private leader board or cross-validation set. Human rater agreement information is also reported in Table 2 with associated data for exact agreement, kappa, and quadratic-weighted kappa. Quadratic-weighted kappas ranged from 0.75 to 0.97, a somewhat high range for human rater performance in statewide high-stakes testing programs (Kirst & Mazzeo, 1996; Massachusetts Department of Education, 2005). Because the training, public leaderboard, and private leaderboard sets were randomly selected from the entire data set provided by the state, the characteristics for the training and public leaderboard sets were similar to those of the private leaderboard.

Procedure

Four of the data sets were transcribed from their original paper-form administration in order to prepare them for processing by automated essay scoring engines. At a minimum, the scoring engines require the written responses to be in ASCII format. This process involved retrieving the scanned copies of written responses from the state or a vendor serving the state, randomly selecting a sample of written responses 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 written responses been directly typed into the computer by the student, the normal procedure for automated scoring. Written responses 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 written response could not be scored by human raters (i.e., short answer was off-topic or inappropriate), the written response 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 hand-wrote their responses using a print style in which all letters were capitalized. To address this challenge, we instructed the transcribers to capitalize beginning of sentences, proper names, and so on. This modification may have corrected errors that would have otherwise been made but limited the overidentification of capitalization errors that might have been made otherwise by the automated scoring engines.

Transcribed responses all came from one state, addressed four different prompts, and consisted of 10,966 written responses. To assess the potential impact of transcription errors, a random sample of 220 responses was retranscribed 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 99.38%.[1]

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

TABLE 2
Phase 2 Private Leaderboard Characteristics

Variable	*Data Set No.*									
	1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
N	558	426	318	250	599	599	601	601	600	548
Grade	10	10	10	10	10	10	10	10	10	8
Type of answer	Source dependent	Source dependent	Source dependent	Source dependent	Nonsource dependent	Nonsource dependent	Source dependent	Source dependent	Source dependent	Source dependent
Subject	Science	Science	English -LA	English-LA	Biology	Biology	English-LA	English-LA	English-LA	Science
M no. of words	48.47	58.36	48.51	38.62	22.62	24.08	40.96	54.18	51.33	41.13
SD no. of words	21.92	23.71	15.10	16.26	18.14	21.26	24.82	33.53	40.09	27.33
Type of rubric	Holistic	Holistic	Holistic	Holistic	Holistic	Holistic	Holistic	Holistic	Holistic	Holistic
Range of rubric	0–3	0–3	0–2	0–2	0–3	0–3	0–2	0–2	0–2	0–2
Range of score	0–3	0–3	0–2	0–2	0–3	0–3	0–2	0–2	0–2	0–2
M score	1.53	1.71	0.98	0.68	0.31	0.27	0.76	1.16	1.11	1.22
SD score	1.01	1.03	0.67	0.65	0.65	0.67	0.83	0.85	0.78	0.68
Exact agreement	0.90	0.85	0.80	0.82	0.96	0.95	0.96	0.85	0.81	0.88
kappa	0.86	0.80	0.66	0.69	0.90	0.84	0.93	0.77	0.72	0.80
Quadratic wgt kappa	0.95	0.93	0.77	0.75	0.95	0.93	0.96	0.86	0.84	0.87

Note. Score distribution (means/standard deviation) and agreement statistics (exact agreement, kappa, *r*, and quadratic weighted kappa) are based on human ratings.

The remaining written responses were provided in ASCII format by their respective states. The eighth- and 10th-grade students in those states had typed their responses directly into the computer using web-based software that emulated a basic word processor. Except that the test had been administered by a computer, the conditions for testing were similar to those in states where the written responses had been transcribed.

All responses used in the samples were double-scored by human raters under production conditions. However, only the score assigned by the first human rater was used in determining the score assigned to the essay. The second human rating was employed as a "read behind" on a randomly selected sample of all responses as part of a quality assurance check. The second rater had no influence on modifying the score that was assigned by the state. Agreement rates for the data used in the private leaderboard are given in Table 2.

For this study, three participating commercial vendors demonstrated their system capabilities, among a total pool of 189 other entrants. The commercial vendors were given an advantage, in that they began with models that had already been developed and/or supported by products with significant field applications. Those other 186 entrants were challenged to build new scoring algorithms, and they submitted 1,887 applications over the course of the competition. The challenge was to determine how short-answer machine scores compare to those assigned by human raters in terms of agreement. Competitors were given $2\frac{1}{2}$ months to train their engines on data sets where they were provided two human rater scores and the text of the response. Simultaneously, teams were provided a test set that illustrated how well their engines performed on a public leaderboard with regard to quadratic weighted kappa, the metric chosen to determine the competition winner. The short-answer competition also included an additional validation set (e.g., a private leaderboard) that was used to determine the actual winners on a verified machine scoring algorithm. Verification involved an independent running of the private leaderboard data using the source code provided by the competitor. At the close of the competition, public winners were asked to reveal their source code (under a GSLv3 license) and produce a "technical methods paper" to detail how their particular approach was designed and executed. The commercial vendors who participated in this competition included Educational Testing Service (*c-rater*), Measurement Incorporated, (*Project Essay Grade*), and MetaMetrics (*Lexile® Writing Analyzer).*

Up to 10 weeks were allowed to statistically model the data during the "training" phase of the demonstration. In addition, the competitors were provided with cut-score information (used in determining score assignments) along with any scoring guides that were used in the training of human raters. This supplemental information was employed by most of the competitors to better model score differences for the score points along the state rubric continuum. All prompts were scored using a holistic scoring approach.

During the training period, a website forum was used to address questions or concerns raised by competitors about the nature of the data or to discuss the merits of certain approaches. For example, the forum publicly discussed the relationship between response length and score assignments (e.g., $r = .32$, statistically significant, but not particularly high). Another lengthy discussion centered on the use of the so-called bag of words approach. With this technique, one could generate the vocabulary associated with a correct response but not have to make logical or grammatical sense (e.g., $r = .65$, statistically significant, moderately high).

In the public leaderboard phase of the evaluation, competitors were provided data sets that had only the text of responses associated with them and were asked to make integer score

predictions for each response. They were given the entire 10-week period in which to make their predictions and were required to submit 100% of the essay scores in each data set regardless of whether their scoring engine classified the essay as "unscorable." Even though human raters had successfully rated all the short-answer responses in the test set, there were a variety of reasons that any one response might prove problematic for machine scoring. For example, a response 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 situations, provisions would be made for these to be scored by human raters.

In the private leaderboard phase of the evaluation, competitors were provided data sets that had only the text of responses associated with them and were asked to make integer score predictions for each response. They were given 6 days in which to make their final submissions and were required to submit 100% of the essay scores in each data set. The competition winner was determined on the basis of their average quadratic weighted kappa for their predictions across all 10 prompts.

Scoring and Evaluation

For the purposes of this study, machine scoring performance was evaluated on the basis of a subset of measures that are standard in the field, including two types of agreement statistics:

- Distributional differences—correspondence in mean and standard deviation of the distributions of human scores to that of automated scores.
- Agreement—measured by exact agreement, kappa, quadratic weighted kappa.

Each one of these measures is described in some detail next. Given its limitations, the study could not evaluate all aspects related to the validity of machine scores. Rather a set of relevant metrics that were common across all vendors was selected. These represented a subset of a more comprehensive evaluation scheme proposed by Williamson et al. (2012).

RESULTS

Overall, 256 teams of competitors participated in the competition. Of that group, 51 teams (a total of 189 participants) submitted final predictions for the private leaderboard. The results summarized here include those for the three participating commercial vendors and the five top public competitors. Although not the focus of this study, it is interesting to note the progression of prediction improvements for the public leaderboard over the duration of the competition. Competitors' initial submissions started at $\kappa_w = .63$ and concluded with $\kappa_w = .77$ at the top of the leaderboard. Because competitors had only 6 days in which to submit their private leaderboard results, the quadratic weighted kappas obtained at the end of the competition were slightly lower, as they had little opportunity to overfit their models. Moreover, the code used for the private leaderboard results was verified by Kaggle. The winning public competitor ($50,000) was an Ecuadorian student studying data science as an undergraduate at the University of New Orleans. Second place ($25,000) went to a graduate student in Slovenia, and third place ($15,000) was awarded to French actuary in Singapore. Their technical papers that describe the methodology used, source code, and final and final models can be obtained from https://www.

kaggle.com/c/asap-sas/details/preliminary-winners. Because the commercial vendors elected not to release their source code, they were ineligible to win prize money.

Table 3 shows the means of the distributions for each data set by competitor. The number of the data set corresponds to the number of the prompt used in the study. In Table 3, SS stands for the student score, which, in the case of these data sets, is the score assigned by the first human rater. The remainder of the mean estimates in the table is derived from the machine score predictions. The three letter abbreviations at the top of the table represent the self-assigned name of the team. So, for example, the ETS team is abbreviated with the letters HNY, which stands for their team name of "Henry" (ostensibly after Henry Chancey, the founder of ETS). The final three columns in the table shows the average of the competitors' mean estimates, a *t* test that compares the competitors' average against that of the student scores, and the associated *p* values. In this case, none of the average mean scores was significantly different from that of the student scores.

Table 4 shows the standard deviation distributions across all the competitors for the 10 data sets, the second distributional measure. With the exception of Data Set 5, the standard deviations were all close and functionally replicated the distributions established by human raters. This table lists the average standard deviation for the machine score predictions across the competitors. The *F*-Max test addresses the hypothesis that the standard deviations are the same. There was one significant difference. In Data Set 5 (a biology item) ($SD_{SS} = 0.51$, $SD_{CompAvg} = 0.56$), $F(1, 598) = 1.23$, $p = .01$.

Table 5 begins the sequence of agreement statistics for the data sets. Recall that for all the data sets in this study, the first human rater determined the student score (e.g., the state assigned score) and that the second rater had a nondeterministic "read behind" role. The H1H2 (first reader to second reader) human rater agreement statistics are provided throughout the remainder of the Results section for comparison purposes. Exact agreement is calculated in two ways. For human raters it is computed by taking the proportion of short-answer responses that have the same ratings assigned by the two human raters. Machine scoring exact agreement (i.e.,

TABLE 3
Score Means by Competitor and Data Set

Data Set	N	M *No. of Words*	*SS*	*USA*	*TDA*	*ZBR*	*GXV*	*HNY*	*SHS*	*JJJ*	*MM*	*Comp Avg*	t	p
1	558	48.47	1.55	1.57	1.55	1.63	1.54	1.52	1.56	1.58	1.56	1.56	0.16	0.88
2	426	58.36	1.62	1.58	1.58	1.58	1.53	1.72	1.70	1.61	1.69	1.62	0.00	1.00
3	318	48.51	0.92	0.96	0.94	1.01	0.96	0.96	0.91	0.93	1.00	0.96	0.70	0.49
4	250	38.62	0.61	0.58	0.60	0.65	0.60	0.64	0.61	0.59	0.58	0.61	−0.07	0.94
5	599	22.62	0.22	0.29	0.23	0.23	0.22	0.28	0.27	0.28	0.21	0.25	0.96	0.34
6	599	24.08	0.28	0.35	0.27	0.29	0.29	0.28	0.32	0.34	0.31	0.31	0.63	0.53
7	601	40.96	0.76	0.78	0.73	0.72	0.78	0.72	0.78	0.78	0.73	0.75	0.70	0.87
8	601	54.18	1.11	1.06	1.15	1.15	1.06	1.13	1.16	1.16	1.09	1.12	0.21	0.84
9	600	51.33	1.05	1.11	1.08	1.11	1.03	1.09	1.14	1.08	1.11	1.09	0.90	0.37
10	548	41.13	1.18	1.24	1.20	1.20	1.20	1.23	1.19	1.19	1.22	1.21	0.67	0.52

Note. SS = Student scores; the rest of the means are based on machine score predictions of the top competitors; USA = Measurement, Inc.; TDA = Luis Tandalla; ZBR = Jure Zbontar; GXV = Gxav; HNY = Educational Testing Service; SHS = Stefan Henss & SirGuessalot; JJJ = JJJ; MM = MetaMetrics; Comp Avg = top competitors' average.

TABLE 4
Score Standard Deviations by Competitor and Data Set

Data Set	N	M *No. of Words*	*SS*	*USA*	*TDA*	*ZBR*	*GXV*	*HNY*	*SHS*	*JJJ*	*MM*	*Comp Avg*	F	p
1	558	48.47	1.06	1.07	1.02	1.12	1.02	1.05	1.06	1.13	1.07	1.07	1.07	0.87
2	426	58.36	0.98	1.03	0.92	1.03	1.02	0.95	1.00	1.13	1.04	1.02	1.07	0.47
3	318	48.51	0.69	0.70	0.59	0.72	0.69	0.64	0.70	0.69	0.75	0.69	1.04	0.90
4	250	38.62	0.59	0.62	0.54	0.65	0.65	0.64	0.59	0.65	0.63	0.62	1.09	0.42
5	599	22.62	0.51	0.59	0.51	0.59	0.52	0.60	0.57	0.59	0.56	0.57	1.23	0.01
6	599	24.08	0.71	0.72	0.64	0.72	0.71	0.65	0.74	0.76	0.73	0.71	1.00	0.97
7	601	40.96	0.83	0.81	0.71	0.81	0.77	0.78	0.83	0.84	0.78	0.79	1.10	0.24
8	601	54.18	0.863	0.81	0.73	0.84	0.80	0.80	0.78	0.83	0.83	0.80	1.16	0.08
9	600	51.33	0.77	0.78	0.74	0.79	0.79	0.77	0.75	0.77	0.77	0.77	1.00	1.00
10	548	41.13	0.69	0.72	0.67	0.72	0.68	0.66	0.69	0.71	0.69	0.69	1.01	0.93

Note. SS = Student scores; the rest of the means are based on machine score predictions of the top competitors; USA = Measurement, Inc.; TDA = Luis Tandalla; ZBR = Jure Zbontar; GXV = Gxav; HNY = Educational Testing Service; SHS = Stefan Henss & SirGuessalot; JJJ = JJJ; MM = MetaMetrics; Comp Avg = top competitors' average.

comparisons involving human raters and machine score predictions) refers to the proportion of short-answer responses that have same ratings assigned by the state adjudicated score and predicted by the machine scoring algorithms. Human exact agreements averaged .89 across the 10 data sets and ranged from .80 for Data Set 3 to 0.96 on Data Set 5. Average Machine scoring was .70 across the data sets and exact agreements for the top public competitor ran from 0.60 on Data Set 2 to 0.88 on Data Sets 5 and 6. Without exception, the exact agreement statistics are statistically significantly lower for machine scored responses than for those scored by human raters. The chi-square test for proportions was used to evaluate the differences between H1H2 and the competitors' average agreement for each data set.

Kappa is a measure of agreement that takes into consideration agreement by chance alone. It is calculated in the same fashion as an exact agreement statistic but applies a correction for chance agreement. For example, if a scoring rubric has only two categories, there is a 50% probability that the two raters will agree by chance alone. For the student score, based on the first reader rating, the ranges ran from 0.66 on Data Set 3 to 0.90 on Data Set 5. Average machine performance across the data sets was .59 and ran from 0.41 on Data Set 8 to 0.63 on Data Set 10 for the top competitor. The average discrepancy between human rater performance and machine scoring performance on this measure ran from 0.08 on Data Set 3 to 0.44 on Data Set 7, where human rater agreement was 0.93. These are given in Table 6. Using the same chi-square test for proportions, machine rater agreement was always statistically significantly below, and most often considerably below human agreement.

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. These assumptions are met for the scoring rubrics applied to the data for this demonstration. Quadratic weighted kappa has a number of advantages over simple or linear weighted kappa. First it is a metric that is used by most testing companies for the assessment of both essays and short-form constructed responses, along with the other statistics that were applied in the

TABLE 5
Exact Agreement Coefficients by Competitor and Data Set

Data Set	N	M *No. of Words*	*H1H2*	*USA*	*TDA*	*ZBR*	*GXV*	*HNY*	*SHS*	*JJJ*	*MM*	*Comp Avg*	χ^2	p
1	558	48.47	0.90	0.73	0.70	0.70	0.70	0.72	0.69	0.68	0.65	0.70	68.51	<.0001
2	426	58.36	0.85	0.62	0.65	0.62	0.58	0.60	0.58	0.56	0.56	0.60	65.52	<.0001
3	318	48.51	0.80	0.75	0.73	0.73	0.74	0.73	0.74	0.73	0.72	0.73	3.95	.0468
4	250	38.62	0.82	0.76	0.76	0.72	0.75	0.73	0.74	0.72	0.73	0.74	4.21	.0402
5	599	22.62	0.96	0.87	0.91	0.88	0.88	0.88	0.87	0.87	0.86	0.88	24.97	<.0001
6	599	24.08	0.95	0.87	0.91	0.89	0.87	0.88	0.86	0.87	0.86	0.88	17.98	<.0001
7	601	40.96	0.96	0.74	0.71	0.68	0.65	0.69	0.68	0.69	0.60	0.68	157.72	<.0001
8	601	54.18	0.85	0.64	0.62	0.60	0.62	0.60	0.60	0.63	0.58	0.61	86.60	<.0001
9	600	51.33	0.81	0.75	0.76	0.76	0.75	0.73	0.74	0.75	0.71	0.74	8.03	.0046
10	548	41.13	0.88	0.78	0.78	0.76	0.78	0.77	0.77	0.77	0.75	0.77	22.21	<.0001
DS Avg	—	—	0.89	0.75	0.76	0.74	0.74	0.74	0.73	0.73	0.70	0.70	68.51	<.0001

Note. H1H2 = Rater 1 with Rater 2; the rest of the means are based on machine score predictions of the top competitors; Comp Avg = top competitors' average; DS Avg = competitor average across data sets; USA = Measurement, Inc.; TDA = Luis Tandalla; ZBR = Jure Zbontar; GXV = Gxav; HNY = Educational Testing Service; SHS = Stefan Henss & SirGuessalot; JJJ = JJJ; MM = MetaMetrics.

TABLE 6
Kappa Coefficients by Competitor and Data Set

Data Set	N	M *No. of Words*	*H1H2*	*USA*	*TDA*	*ZBR*	*GXV*	*HNY*	*SHS*	*JJJ*	*MM*	*Comp Avg*	X^2	p
1	558	48.47	0.86	0.64	0.60	0.60	0.59	0.62	0.58	0.57	0.52	0.59	100.67	<.0001
2	426	58.36	0.80	0.48	0.51	0.48	0.43	0.45	0.42	0.41	0.40	0.45	109.84	<.0001
3	318	48.51	0.66	0.59	0.53	0.56	0.58	0.54	0.63	0.62	0.56	0.58	3.99	.0459
4	250	38.62	0.69	0.57	0.56	0.49	0.55	0.52	0.52	0.50	0.51	0.53	12.79	.0003
5	599	22.62	0.90	0.61	0.72	0.61	0.63	0.64	0.61	0.61	0.52	0.62	127.20	<.0001
6	599	24.08	0.84	0.61	0.69	0.63	0.58	0.60	0.57	0.59	0.56	0.60	84.39	<.0001
7	601	40.96	0.93	0.59	0.55	0.48	0.45	0.50	0.49	0.51	0.38	0.49	280.42	<.0001
8	601	54.18	0.77	0.47	0.43	0.38	0.44	0.39	0.39	0.44	0.36	0.41	159.51	<.0001
9	600	51.33	0.72	0.63	0.63	0.63	0.62	0.59	0.59	0.62	0.56	0.61	15.80	<.0001
10	548	41.13	0.80	0.65	0.64	0.61	0.64	0.62	0.63	0.63	0.60	0.63	38.03	<.0001
DS Avg	—	—	0.81	0.59	0.59	0.55	0.55	0.55	0.54	0.55	0.49	0.59	100.67	<.0001

Note. H1H2 = Rater 1 with Rater 2; the rest of the means are based on machine score predictions of the top competitors; Comp Avg = top competitors' average; DS Avg = competitor average across data sets; USA = Measurement, Inc.; TDA = Luis Tandalla; ZBR = Jure Zbontar; GXV = Gxav; HNY = Educational Testing Service; SHS = Stefan Henss & SirGuessalot; JJJ = JJJ; MM = MetaMetrics.

evaluation of scoring performance. Second, it is numerically equal to the Pearson correlation coefficient. This permitted an *r* to *z* conversion for the averaging of performance over data sets. Finally was the metric that was used by Kaggle to determine the competition winners.

Human quadratic-weighted kappas averaged .89 across the data sets and ranged from 0.75 on Data Set 4 to 0.96 on Data Set 7, whereas the average competitor agreements was .72 across the data sets and ranged from .60 on Data Set 8 to .82 on Data Set 1. The average discrepancy between human rater performance and machine scoring performance on this measure ran from 0.06 on Data Sets 3 and 9 to 0.30 on Data Set 7 where human rater agreement was 0.96. The *z* test for correlation differences was used to evaluate the differences between H1H2 and the competitors' average agreement for each data set. Table 7 provides the values for quadratic weighted kappa. There were significant differences on nine of the 10 comparisons.

DISCUSSION

The results for the short-answer competition revealed that the machine scoring algorithms did not achieve the same level of agreement as their human rater counterparts. A number of different measures were utilized to evaluate the performance of the short-answer scoring engines including mean, standard deviation, exact agreement, kappa, and quadratic weighted kappa. The scoring engines closely matched the means of the human rater distributions for each item with neglible differences. Machine performance on replicating the standard deviation was very similar with only minor differences from the human rater dispersion measures.

With regard to estimates for exact agreement and kappa, machine performance was consistently less than that of the human raters. The average human rater performance across the 10 data sets was .90, as measured by quadratic weighted kappa, whereas the top competitor's machine algorithms registered in at .76 across the 10 data sets. There was variability among the data sets for both human and machine ratings. For the machine algorithms, the quadratic weighted kappas ranged in the low 60s for Data Set 8 to the low 80s for Data Set 6. The reliability in performance was in contrast to the Phase 1 competition (essays) where the average quadratic weighted kappa for the human raters was .75 but reached .81 for the top machine competitors. The pattern would appear to be, at least for the data sets used in the two competitions, that human raters do a relatively better job on evaluating short-answer responses and that the machine algorithms have a slight advantage for longer responses. In terms of patterns of agreement coefficients, one noteworthy difference is that the ranges for agreement coefficients were higher in Phase 1, the evaluation of essay scoring. In part this difference in range may be a function of scaling issues with the more extensive rubrics employed in the evaluation of essays. However, the pattern for Phase 2 is that while agreements weren't as high as they were for essays, the variability in agreement rates for short-answer responses (Phase II) was considerably lower. That is, the lows in agreement coefficients across the scoring engines weren't as low and the highs were not as high. Noteworthy also is the discrepancy between the two separate tasks. In terms of quadratic weighted kappa for these studies, human performance was different on the order of .15 (in favor of short-answer responses), whereas the machine algorithms performed better by .05 (in favor of essays). The design in this study does not permit a direct comparison of the two results as the raters for Phase 1 and Phase 2 were different. Moreover, the scales used in the short-form constructed response study were restricted in range

TABLE 7
Quadratic Weighted Kappa Coefficients by Competitor and Data Set

Data Set	N	M *No. of Words*	*H1H2*	*USA*	*TDA*	*ZBR*	*GXV*	*HNY*	*SHS*	*JJJ*	*MM*	*Comp Avg*	z	p
1	558	48.47	0.95	0.84	0.81	0.83	0.81	0.84	0.82	0.83	0.81	0.82	11.24	<.0001
2	426	58.36	0.93	0.77	0.78	0.77	0.76	0.75	0.75	0.75	0.71	0.76	9.63	<.0001
3	318	48.51	0.77	0.73	0.65	0.71	0.71	0.69	0.75	0.73	0.71	0.71	1.67	.095
4	250	38.62	0.75	0.67	0.63	0.62	0.67	0.64	0.62	0.61	0.63	0.64	2.39	.017
5	599	22.62	0.95	0.78	0.82	0.78	0.78	0.79	0.77	0.78	0.73	0.78	13.58	<.0001
6	599	24.08	0.93	0.81	0.84	0.82	0.80	0.80	0.80	0.81	0.78	0.81	9.17	<.0001
7	601	40.96	0.96	0.68	0.67	0.65	0.65	0.67	0.67	0.66	0.58	0.65	20.24	<.0001
8	601	54.18	0.86	0.61	0.62	0.60	0.61	0.61	0.60	0.59	0.53	0.60	10.38	<.0001
9	600	51.33	0.84	0.78	0.79	0.79	0.78	0.76	0.77	0.78	0.75	0.78	3.04	.002
10	548	41.13	0.87	0.73	0.75	0.72	0.74	0.72	0.74	0.72	0.71	0.73	7.36	<.0001
DS Avg	—	—	0.89	0.74	0.75	0.74	0.73	0.73	0.73	0.73	0.70	0.72	—	—

Note. H1H2 = Rater 1 with Rater 2; the rest of the means are based on machine score predictions of the top competitors; Comp Avg = top competitors' average; DS Avg = competitor average across data sets; USA = Measurement, Inc.; TDA = Luis Tandalla; ZBR = Jure Zbontar; GXV = Gxav; HNY = Educational Testing Service; SHS = Stefan Henss & SirGuessalot; JJJ = JJJ; MM = MetaMetrics.

compared those used in the essay competition. The general task of evaluating text was similar in nature, and the raters, although different, had the same type of training. Moreover, both studies included rubrics as guidelines for evaluating the responses.

With regard to recommendations for both PARCC and Smarter Balanced, we suggested that machine performance for the top vendors of automated essay scoring was ready to be used, pending additional validity studies, as a *second reader* for high-stakes assessments and possibly as a first reader for low-stakes assessments (Shermis, 2014). Presently that technology does not yet have the capacity to determine if a good argument (or conclusion) has been made though it can, in a rudimentary way estimate the degree to which one has made a coherent argument (Burstein, Tetreault, Chodorow, Blanchard, & Andreyev, 2013). Our recommendation to the two major Race-to-the-Top consortia regarding short-answer scoring is that the technology has to undergo some significant additional development before it can be deployed operationally. In its current form, it might be used as a second reader on an experimental basis or phased in slowly after states gain significant experience with the technology.

The challenges for scoring short-answer constructed responses are quite different from that of evaluating essays. Current essay algorithms are focused on grading *writing ability* and *content*, whereas the emphasis for scoring short-answer responses is on grading *content* and *response correctness*. With regard to content, the essay writer has more writing space in which to establish relevant information, whereas the scoring algorithms for short-answer responses have to make their estimates on less information. When the domain of content is one or two words (or their variations), the response might be easy to program; however, if there are multiple elements involved in the correct answer, it may be more difficult to capture and characterize the writer's intent in 40 words or less.

The dimension of response correctness is difficult because human writers can produce many variations of words or phrases that can be easily recognized by human raters, but be easily missed by machine scoring algorithms. So, for example, Leacock and Chodorow (2003b) investigated the number of combinations of an answer in which the phrase "President Ronald Reagan" was the correct response. Their investigation suggested that, from 9,000 responses, there were approximately 67 variations of the name "Reagan" that could be correctly deciphered by human raters but would have to be programmed into machine scoring algorithms. Some of the variations can be accounted for by misspellings for which there is a corpus that can easily be tapped, but other variations might not be accounted for by these lists (e.g., proper names).

The machine scoring algorithms also have to take into account other challenges such as the use of passive voice, syntactic variety, pronoun resolution, morphology, the inclusion of negatives (or double-negatives), filling in semantic gaps, and concept mapping (Leacock & Chodorow, 2003b). So the correct answer on a short-answer question might be "Rome destroyed Carthage," but the scoring algorithm would have to take into account "Carthage was destroyed by Rome" and "Rome did not destroy Alexandria, but rather Carthage," and all the variations of these answers (Leacock & Chodorow, 2003b).

These results are a bit less impressive than what has been found in other field trials of short-form automated scoring. For example, Butcher and Jordan (2010) found in a study of that evaluated undergraduate responses in an introductory science module that both human rater and machine software scoring agreements ranged from the mid-70s to the high 90s. Across all items, machine scores were slightly higher. These tasks were similar to the kinds of responses

generated in the current study's sample but used a study sample that incorporated participants that were slightly older.

A number of authors have described best practices for developing short-form constructed responses. These can be distilled in to four general principles. First, outline what constitutes an expected answer, such as the criteria for knowledge and skills (McMillian, 2010). Second, select an appropriate scoring method that is based on the criteria (McMillian, 2010). Developing a rubric that specifies the educational criteria for scoring can be especially helpful (McMillian, 2010). Third, clarify the role of writing mechanics, such as grammar, spelling, punctuation, and so on, and other factors independent from the educational outcomes being measured (Pointek, 2008). Fourth, use a systematic process for scoring each item, and last, creating anonymity for the students' responses during scoring (Pointek, 2008).

Although these guidelines may be helpful for human-graded responses, one additional guideline might be needed for short-answer responses graded by machines: the need to narrow the range of possible responses to a known, finite pool of correct answers. Consider the question from Data Set 6 (modified for this illustration):

> Identify ONE trait that can describe Sara based on her conversations with Gail or Aunt Rosie. Include ONE detail from the story that supports your answer.

One of the possible problems for this question is that the trait that a student might list has a large range of possibilities based on the reader's interpretation AND the identified trait is adequately supported by the one story detail. Not all combinations of traits and supporting information are likely identified in a modestly sized training set, and therefore new unique answers would be difficult to evaluate on operationally scored answers. Consider a different question, however, where the task is to discover, through experimental methodology, what an unknown chemical element is and why the particular element was chosen. As of this writing, there are 118 chemical elements, each of which can be made available to machine scoring algorithms along with multiple potential justifications for choosing it. In this scenario, the correctness of the answer can be more easily determined and the justification becomes the variable aspect of scoring.

Besides an assessment of the "state of the art" in the field of short-form constructed responses, this study has implications for other areas of short-answer engine scoring research that may be worth noting. First, short-answer machine scoring quality is affected by item/prompt type. As seen from Table 7, the engine performance on the 10 prompts can be divided into three groups using quadratic weighted kappa 0.75 and above, 0.70 to 0.74, and below 0.70. If we look at Prompts 4, 7, and 8, these three prompts have engine scoring performance in the low 60s. These all correspond to "source dependent/English-LA" prompt types. Is there something about this type of prompt that makes them harder to score using machine scoring technology? The relationship between item quality and prompt type has been made in Leacock, Messineo, and Zhang (2013).

Second, the human scoring rubric also matters. This article discusses "best practices" and recent research on short-answer engine scoring also provides related findings (Brew & Leacock, 2013). The three low-performing prompts might be improved by revising human rubrics for more effective engine training (Leacock, Gonzalez, & Conarroe, 2013).

Third, at least for automated essay scoring, the technology can quickly identify and highlight "aberrant" responses. Such responses might be flagged "unscorable," "off-topic," or threats to

self/others. Given that engine can be developed to forward responses that require human review (a feature available in most of the commercial scoring engines), this area deserves further research for developing methods of identifying short-answer responses required for human rating. Automated scoring engines will not replace human rating completely but can be implemented with human ratings in a hybrid approach to increase the validity of scores.

Finally, most human rubrics for short-form constructed responses overlook spelling and mechanical errors. Engines can be developed to include spelling error and grammar error correction algorithms to improve their performance (Leacock & Chodorow, 2003a).

Limitations

This study was conducted with a number of limitations identified in the introduction. It was based on existing high-stakes short-answer responses drawn from PARCC and Smarter Balanced states, incorporated a subset of possible distributional and agreement measures, used human ratings as basis for comparisons, and did not address other validity issues such as the potential for differential functioning of automated scoring among vulnerable groups. Nevertheless, it provides some insight as to the state of the art for machine scoring of short-form constructed responses in high-stakes testing in the United States.

REFERENCES

Attali, Y., Powers, D. E., Freedman, M., Harrison, M., & Obetz, S. (2008). *Automated scoring of short-answer open-ended GRE Subject test items*. Princeton, NJ: Educational Testing Service.

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

Brew, C., & Leacock, C. (2013). Automated short answer scoring: Principles and prospects. In M. D. Shermis & J. C. Burstein (Eds.), *Handbook of automated essay evaluation: Current applications and new directions* (pp. 136–152). New York, NY: Routledge.

Burstein, J., Tetreault, J., Chodorow, M., Blanchard, D., & Andreyev, S. (2013). Automated evaluation of discourse coherence quality in essay writing. In M. D. Shermis & J. Burstein (Eds.), *Handbook of automated essay evaluation: Current applications and new directions* (pp. 267–280). New York, NY: Routledge.

Butcher, P. G., & Jordan, S. E. (2010). A comparison of human and computer marking of short free-text student responses. *Computers & Education*, *55*, 489–499.

Finkel, J. R., Grenager, T., & Manning, C. (2005, June). *Incorporating non-local information into information extraction systems by Gibbs Sampling*. Paper presented at the 43nd Annual Meeting of the Association for Computational Linguistics, Ann Arbor, MI. Available from http://nlp.stanford.edu/~manning/papers/gibbscrf3.pdf

Kirst, M. W., & Mazzeo, C. (1996, April). *The rise, fall, and rise of state assessment in California, 1993–1996*. Paper presented at the Annual Meeting of the American Educational Research Association, New York, NY.

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). Introduction to latent semantic analysis. *Discourse Processes*, *25*, 259–284.

Leacock, C., & Chodorow, M. (2003a). Automated grammatical error detection. In M. D. Shermis & J. Burstein (Eds.), *Automated essay scoring: A cross-disciplinary perspective* (pp. 195–208). Mahwah, NJ: Erlbaum.

Leacock, C., & Chodorow, M. (2003b). C-rater: Scoring of short-answer questions. *Computers and the Humanities*, *37*, 389–405.

Leacock, C., Gonzalez, E., & Conarroe, M. (2013, April). *Developing effective scoring rubrics for automated short-response scoring*. Paper presented at the National Council on Measurement in Education, San Francisco, CA.

Leacock, C., Missineo, D., & Zhang, X. (2013, April). *Issues in prompt selection for automated scoring of short-answer questions*. Paper presented at the National Council on Measurment in Education, San Francisco, CA.

Massachusetts Department of Education. (2005). *2005 MCAS Technical Report*. Boston, MA: Author.

Massachusetts Department of Elementary and Secondary Education. (2011). *Release of February 2011 MCAS biology test items*. Malden, MA: Author.

McMillian, J. H. (2010). *Classroom assessment: Principles and practice for effective instruction*. Boston, MA: Allyn and Bacon.

Pointek, M. E. (2008). *Best practices for designing and grading exams* (Vol. 24). Ann Arbor: Center for Research on Learning and Teaching, University of Michigan.

Salton, G., Wong, A., & Yang, C. S. (1975). A vector space model for automatic indexing. *Communications of the ACM*, *18*, 613–620.

Shermis, M. D. (2014). 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., Burstein, J. C., Elliot, N., Miel, S., & Foltz, P. W. (in press). Instructional applications for automated writing evaluation. In C. A. McArthur, S. Graham, & J. Fitzgerald (Eds.), *Handbook of writing research* (2nd ed.). New York, NY: Guilford.

Shermis, M. D., & Hamner, B. (2012, April). *Contrasting state-of-the-art automated scoring of essays: Analysis*. Paper presented at the National Council of Measurement in Education, Vancouver, British Columbia, Canada.

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.

Sukkarieh, J. Z., & Blackmore, J. (2009, May). *c-rater: Automatic content scoring for short constructed responses*. Paper presented at the the 22nd International FLAIRS Conference for the Florida Artificial Intelligence Research Society, Sanibel Island, FL.

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*, 2–13.

Yang, Y., Buckendahl, C. W., Juszkiewicz, P. J., & Bhola, D. S. (2002). A review of strategies for validating computer automated scoring. *Applied Measurement in Education*, *15*, 391–412.

Copyright of Educational Assessment is the property of Taylor & Francis Ltd and its content may not be copied or emailed to multiple sites or posted to a listserv without the copyright holder's express written permission. However, users may print, download, or email articles for individual use.