CHAPTER 9.

CRITIQUE OF MARK D. SHERMIS AND BEN HAMNER, "CONTRASTING STATE-OF-THE-ART AUTOMATED SCORING OF ESSAYS: ANALYSIS"

Les C. Perelman
MIT

Although the unpublished study by Shermis & Hamner (2012) received substantial publicity about its claim that automated essay scoring (AES) of student essays was as accurate as scoring by human readers, a close examination of the paper's methodology demonstrates that the data and analytic procedures employed in the study do not support such a claim. The most notable shortcoming in the study is the absence of any articulated construct for writing, the variable that is being measured. Indeed, half of the writing samples used were not essays but short one-paragraph responses involving literary analysis or reading comprehension that were not evaluated on any construct involving writing. In addition, the study's methodology employed one method for calculating the reliability of human readers and a different method for calculating the reliability of machines, this difference artificially privileging the machines in half the writing samples. Moreover, many of the study's conclusions were based on impressionistic and sometimes inaccurate comparisons drawn without the performance of any statistical tests. Finally, there was no standard testing of the model as a whole for significance, which, given the large number of comparisons, allowed machine variables to occasionally surpass human readers merely through random chance. These defects in methodology and reporting should prompt the authors to consider formally retracting the study. Furthermore, because of the widespread publicity surrounding this study and because its findings may be used by states and state consortia in implementing the Common Core State Standards, the authors should make the test data publicly available for analysis.

DOI: https://doi.org/10.37514/PER-B.2024.2166.2.09

On April 16, 2012, Mark D. Shermis, Dean of the School of Education at the University of Akron, presented a paper at the annual meeting of the National Council on Measurement in Education on "Contrasting State-of-the-Art in Automated Scoring of Essays: Analysis." Despite its fairly nondescript title, the paper claimed that machines graded essays as well as expert human raters, a claim that was publicized in various press releases and newspaper articles. A press release from the University of Akron, for example, stated, "A direct comparison between human graders and software designed to score student essays achieved virtually identical levels of accuracy, with the software in some cases proving to be more reliable, a groundbreaking study has found" (Man and Machine, 2012). A headline in Inside Higher Ed read, "A Win for the Robo-Readers," and the story included statements such as the following:

> The study, funded by the William and Flora Hewlett Foundation, compared the software-generated ratings given to more than 22,000 short essays, written by students in junior high schools and high school sophomores, to the ratings given to the same essays by trained human readers. The differences, across a number of different brands of automated essay scoring software (AES) and essay types, were minute. (Kolowich, 2012)

Even the venerable British publication, *The New Scientist*, reported "The essay marks handed out by the machines were statistically identical to those from the human graders, says [Jaison] Morgan. 'The result blew away everyone's expectations,' he says." (Giles, 2012) Yet these reports and other statements can best be characterized as unsubstantiated overstatement. The study, however, employs an inconsistent and questionable methodology that favors the machines over the human graders. Even with these biased procedures and results, the data still give some, but lacking the full test essay sets, inconclusive indication that in actual assessments of writing, human scorers were more reliable than machines.

The study derived from the Automated Student Assessment Prize (ASAP), a competition sponsored by the William and Flora Hewlett Foundation, to assess the efficacy of automated scoring engines. The competitions involved evaluating essays from statewide assessments that had already been scored by human readers. Phase One, which dealt with "long-form constructed responses" (although over half of the responses were essentially paragraphs) had two parts. The first involved scoring engines developed by nine testing companies such as the Educational Testing Service, Pearson Knowledge Technologies, and CTB/McGraw-Hill. The second competition was an open contest among software developers. The Shermis and Hamner paper reports on only the first part, the performance of the nine vendors.

A version of the study has subsequently been published as a chapter (Shermis & Hamner, 2013) in the *Handbook of Automated Essay Scoring* co-edited by Mark D. Shermis, the lead author of the original paper, and Jill Burstein (2013). Other contributors in the same volume state explicitly that the study showed Automated Essay Scoring is capable of producing scores similar to those of human readers. All of these studies referenced the original Shermis and Hamner paper. It is the paper's central claim, articulated in the abstract, that has elicited so much publicity: The results demonstrated that overall, automated essay scoring was capable of producing scores similar to human scores for extended-response writing items with equal performance for both source-based and traditional writing genre [sic].

That central claim, however, is clearly not supported by the data. Conversely, the data provide some, although not conclusive, support for the assertion that human scorers performed more reliably than the machines, especially on longer papers that were scored for writing ability rather than solely on content.

NO DEFINITION OF WRITING CONSTRUCT

One major problem with the study is the lack of any explicit construct of writing. Without such a construct, it is, of course, impossible to judge the validity of any measurement. Writing is foremost a rhetorical act, the transfer of information, feelings, and opinions from one mind to another mind. The exact nature of the writing construct is much too complex to outline here; suffice it to say that it differs fundamentally from the Shermis and Hamner study in that the construct of writing cannot be judged like the answer to a math problem or GPS directions. The essence of writing, like all human communication, is not that it is true or false, correct or incorrect, but that it is an action, that it does something in the world. That is what sophists like Protagoras and philosophers, most notably, Aristotle, noted in classical times, and more recently "ordinary language" philosophers like J. L. Austen (1962) and H. Paul Grice (1989), and linguists such as Dan Sperber and Deirdre Wilson (1990), have made apparent in current discussions of language use.

The seriousness of this lack of definition of the writing construct in the study manifests itself in various defects and confusions, beginning with the heterogeneous mix of papers that constitute the study. The study is based on a corpus of eight different essay sets that come from six different states. Each essay set contains a prompt, training information that includes rubrics and in some cases annotated or unannotated training samples along with other materials. Sixty percent of the total papers from each data set were publicly available with scores as training sets. The remaining 40% of papers were divided into two test sets of

20% each. Consequently, of the total sample size of 22,029 papers, only 4,343 papers in eight different essay sets comprised the actual test sets, ranging from 304 to 601 papers each.

However, half the essay sets and over half the aggregate number of papers in the test set were not evaluated on any construct connected to writing. The study defines the four essay sets #3-#6 as source-based writing assignments. Source-based writing assessments measure student writing in response to specific texts or data. A prominent example of this kind of assessment is the Document-Based Question in the Advanced Placement Language and Composition Examination (Perelman, 2008). Students are given a passage or a short essay and then asked to write an argument or analysis about it. Unlike the rubrics that govern the scoring of essay-sets in the Shermis study, the rubrics for these AP Examinations emphasize writing skills such as organization, argument, and expression as well as a student's mastery of content. Essay sets #3-#6, on the other hand, contain prompts and rubrics that are not based on document or source-based writing, however, but are content-based or content dependent exercises that are scored solely on the understanding of content rather than any assessment of writing ability.

Two of these essay sets, #3 and #4, are focused solely on literary analysis. Essay set #3 consists of responses to a prompt based on "Rough Road Ahead" by Joe Kurmaskie: "Write a response that explains how the features of the setting affect the cyclist. In your response, include examples from the essay that support your conclusion." Essay set #4 consists of responses to a prompt based on "Winter Hibiscus" by Minfong Ho. The prompt repeats the last paragraph of the story and then asks students to "Write a response that explains why the author concludes the story with this paragraph. In your response, include details and examples from the story that support your ideas" (see Appendix A).

The rubrics, based on a scale from 0-3, are identical for each of these two essay sets (#3 and #4). The score of 3, the highest score, is defined in each of the rubrics by the following language:

Score 3: The response demonstrates an understanding of the complexities of the text.

- • Addresses the demands of the question
- • Uses expressed and implied information from the text
- • Clarifies and extends understanding beyond the literal

The rubrics and other materials for essay sets #5 and #6 explicitly define them as reading tests, while defining a different scale for writing tests. (See Appendix A; Kaggle-Data 2012.) The prompt for essay set #6 is based on an excerpt discussing the obstacles to putting a mooring mast for dirigibles on top of the Empire State Building:

> Based on the excerpt, describe the obstacles the builders of the Empire State Building faced in attempting to allow dirigibles to dock there. Support your answer with relevant and specific information from the excerpt.

Immediately following the rubric, which focuses on ability to understand the content, not on ability to write, were these scoring notes.

The obstacles to dirigible docking include:

- Building a mast on top of the building
- Meeting with engineers and dirigible engineers
- Transmitting the stress of the dirigible all the way down the building; the frame had to be shored up to the tune of $60,000
- Housing the winches and other docking equipment
- Dealing with flammable gases
- Handling the violent air currents at the top of the building
- Confronting laws banning airships from the area
- Getting close enough to the building without puncturing

I took my first and last programming class in 1967, learning Fortran IV. Even with my rusty recollection of an antique programming language, I am fairly confident that I could construct a program that could also do very well scoring these essays simply by counting strings of key words and phrases, including synonyms. Such a program, however, would not in any way be assessing writing.

LACK OF A CONSTRUCT FOR THE ESSAY

The majority of essay sets in this study are single paragraphs, not essays. Although the study explicitly stated that it was exploring how well machines could grade extended-response writing, (i.e., essays), only three of the eight data sets consisted of what is commonly defined as extended-response writing. The mean number of words for essay sets #3, 4, 5, 6, and 7, ranged from 98.70 to 173.43 words. A typed double-spaced page of prose with 12-point type is approximately 250 words. (Note: the number of words in this paragraph up to and including this note is 96.)

Only three of the eight studies have average word lengths of over 360 words, approximately 1.5 typed double spaced pages. One of the senior managers of the project reported in an email that the team spent three months asking every state for sets of long-form constructed response essays, and even requested papers from international sources (J. Morgan, personal communication, June 19, 2012). He admitted that some of the essay sets defined as “essays” in the study were shorter than the length the team desired, but defended the sample by arguing that even these short pieces of writing are categorized as essays by the states, meaning that at least one of the fifty states defines it as an essay. These essays, however, and the way that they are scored, may be unrepresentative of state writing assessments as a whole, especially since the training materials for essay sets #5 and #6 specifically differentiate these assessments from assessments of writing ability (See Appendix A). To further muddle the study, all of the paragraph markings were removed (Shermis & Hamner, p. 8), converting even the longer, multi-paragraph essays into single paragraphs, making any evaluation of key essay features such as paragraph coherence impossible. In sum, only the scores of three collections of essays out of eight in this study represent actual measures of multi-paragraph writing ability. Moreover, even in those three collections of essays, there were no paragraph markings for computers to use to judge development and coherence - only block text.

FLAWED EXPERIMENTAL DESIGN I: APPLES AND ORANGES

The study employs different methods for humans and machines in computing reliability for all of its measures. This highly unorthodox and statistically indefensible approach substantially and unfairly biases the measures in favor of the machines, artificially increasing their reliability compared to that of the human readers. For all of the measures beginning with those displayed in Table 8 in Appendix B, the study uses the measure H1H2, the comparison between the two human readers, as the measure for reader reliability, while the measure for machine reliability is the comparison between the machines and a construct, the resolved score (RS). (To avoid confusion, tables in this Critique are labeled with letters while tables in Appendix B taken from the Shermis and Hamner Report are labeled with numbers.)

In most essay testing situations, the standard practice is that the resolved score is the sum of the two reader scores if the scores are identical or adjacent; or, if the scores vary by more than one point, the resolved score is established by one or two supervisors rereading the essay. Yet, only one of the essay sets in the study, #1, follows the standard best practice of combing two equal or adjacent scores to compute the RS. Essay set #7 combines composite scores regardless of the size of the difference between them. Essay set #8 also appears to combine composite scores regardless of the size of the difference but has a third reader adjudicating 17.7% of the RS’s randomly. Essay set #2 uses the score of only the first reader as the RS, regardless of the second reader’s score. The remaining four essay sets,

#3, #4, #5, and #6, all compute the RS as the higher of the two scores.2 This procedure, followed by four essay sets - half of the total number - and containing 55% of the essays in the aggregate data sample, skews many of the measures used in favor of the AES scores.

Before going into a more technical explanation, the bias produced by using these two different measures can be best illustrated by a hypothetical example. A company is hiring an additional reader to score essays. It has two applicants for the position and will select the applicant that has the greatest reliability in scoring. A reader who already works for the company has scored all of the essays. The first applicant scores the essays and her reliability is determined by comparing her scores to those of the first reader. The second applicant, however, is told before scoring the essays that his reliability will be determined by comparing his score to the higher of the two previous readers' scores if the scores differ. He realizes that his chances improve dramatically simply by always selecting the higher score in any case in which he is wavering between two scores. He does so, scores more reliably, and gets the job. Clearly, the procedure was biased in favor of the second applicant. This example is completely analogous to the procedure used in the study for essay sets #3, #4, #5, and #6, which are biased towards the machines. If such a procedure as described in the scenario were actually implemented in the real world, it would clearly be an unfair hiring practice. Similarly, this practice used in the Shermis and Hamner study unfairly biases and therefore invalidates half the results.

Essays scores, be they holistic, trait, or analytical, always are continuous variables, not discrete variables (integers), even though graders almost always have to give integer values as scores. The report recognizes this fact in the observation on page 24 that values for the Pearson r "might have been higher except that the vendors were asked to predict integer values only." Each reader has to select a single integer value even though some essays might be on the border between two adjacent integers. Some 3's on a 4-point-scale might be very high 3's bordering on a 4, while other 3's may be very low 3's bordering on a 2. Significantly, some of the training materials for the essay sets included essays scores with plus and minus signs. In the terminology of Classical Test Theory, the True Score might be 3.3 or 2.8. Consequently, adjacent agreement in the correct direction between two readers (e.g. one rater gives an essay a score of 3 and the second rater gives the essay a score of 4) will more closely approximate a True Score of 3.4 than two scores of 3.

Resolving scores merely by selecting the higher one ignores the continuous nature of the scores being measured and penalizes human raters while giving AES algorithms a substantial advantage by allowing them to optimize agreement with the RS by rounding up just like the example of the second job applicant. In the case of the essay that has a True Score of 3.4, for example, there are four likely pairs of scores that would be produced by two human raters: 3-3, 3-4, 4-3, and 4-4. Note that in three of these four cases, selecting the higher score makes 4 the resolved score, and that in two of these three instances one of the two reader scores will be lower than the resolved score. This unjustified score bias can be observed in the data in the tables at the end of the report. Table 9.1 and Table 4 of the study (Appendix B) display the means for the five score sets that ether combine rater scores to compute RS (#1, #7, and #8) or use a single score as the RS (#2A and #2B). In these essay sets, the resolved score does not bias the results against the human readers. Significantly, the means of the human reader scores match the means of the RS much more closely than those of the means of the machine scores. Table 9.2 displays the means for the essay sets that used the higher human rater score as the resolved score. The contrast in the differences between the human reader means and the resolved score means in the two tables is striking and provides a powerful illustration of how computing the RS as the higher scores skews the results against human readers.

Table 9.1. Test Set Mean for Resolved Scores=Sum of Scores or Single Score

	H1	H2	RS	Diff. Avg. Human Rater Means from RS Means	Range of AES Mean Scores	Range of Diff. of AES Mean Scores from RS Means
1	8.61	8.62	8.62	-0.01	8.49-8.80	-0.13-0.18
2A	—	3.39	3.41	-0.02	3.33-3.41	-0.08-0.00
2B	—	3.34	3.32	0.02	3.18-3.37	-0.14-0.05
7	20.02	20.24	20.13	0.00	19.46-20.05	-0.67- -0.08
8	36.45	36.70	36.67	-0.09	37.04-37.79	.037-1.12

Separating the essay sets into two groups, those that use a single human score or a sum of two human scores to compute the resolved score and those that use the higher score as the resolved score, present two very different sets of values of the metrics used in the study. Tables 9.1 & 9.2 also demonstrate the substantial difference for means.

Table 9.2. Test Set Mean for Resolved Scores=Higher Human Rater Scores

	H1	H2	RS	Diff. Avg. Human Rater Means from RS Means	Range of AES Mean Scores	Range of Diff. of AES Mean Scores from RS Means
3	1.79	1.73	1.90	0.14	1.84-1.95	-0.06-0.05
4	1.38	1.40	1.51	0.12	1.34-1.57	-0.17-0.06
5	2.31	2.35	2.51	0.18	2.44-2.54	-0.07-0.03
6	2.57	2.58	2.75	0.18	2.54-2.83	-0.04-0.08

Similar distinctions can be shown in the other tables. Indeed, in the five measures of agreement, exact agreement (Table 8), exact and adjacent agreement (Table 10), Kappas (Table 12), Quadratic Weighted Kappas (Table 14) and the Pearson r (Table 16), the human raters in the group of essay sets clearly outperform the AES engines in the first three and have mixed results for the Quadratic Weighted Kappa and Pearson r. Curiously, for the Quadratic Weighted Kappa (Table 14) the relationship of the two groups is inverted - human raters in two of the four essay sets that use the higher score as the resolved score (#3 and #4) as well as score sets #2A and #2B outperform the AES engines while AES engines outperform human raters in the other essay sets. This anomaly may partially be an artifact of the Quadratic Weighted Kappa measuring correspondence not between two raters, as is its intended use, but between a rater (i.e., the machine score) and the artificial construct of the resolved score as higher of the two scores. Another possible explanation is offered by Brenner & Kliebsch (1996) who noted that that quadratically weighted kappa coefficients tend to increase with larger scales while unweighted kappa coefficients decrease. They noted that "variation of the quadratically weighted kappa coefficient with the number of categories appears to be strongest in the range from two to five categories" (p. 201). As displayed in Table 3 of the report, the scales for essay sets #3 and #4 consisted of a scale of four (0-3), while essay sets #5 and #6 consisted of a scale of five (0-4). With the exception of the four point scale for score set #2B, all the other essay sets had scales greater than five. For essay set #1 the range of the rubric was 1-6 and the range of the resolved score was 2-12. For scoring set #2A, the range was 1-6; for scoring set #7, the range of the rubric was 0-12, and the range of the resolved score was 0-24. For essay set #8, the range of the rubric was 0-30, and the range of the resolved score was 0-30.

The confusion between human scores and resolved score is found throughout the text. The report states, for example, on page 22, "all vendor engines generated predicted means within 0.10 of the human mean for Essay set #3 which had a rubric range of 0-3." The report, however, is referring to the mean of the resolved score not the mean of the human raters, which were, in actuality, lower than the resolved score by 0.11 and 0.17 respectively. (See Table 9.2)

The standard method for comparing the reliability of machine scores to human scores is to compare the reliability of the machine scores to each of the two human scores and then compare those scores to reliability of the human scorers to each other (McCurry, 2010). In McCurry's study, as in many others, humans clearly outperformed machines. Yet the Shermis and Hamner study instead chose to use different variables for humans and machines.

The two readers' individual scores compared to the resolved score (H1 and H2) are consistently higher than those of the machine scores for all of the metrics displayed in all of the tables (Appendix B). This phenomenon could well be an artifact of the individual reader score being a contributing element to the resolved score. However, of the nine score sets, the two scores of H2, the second human reader, for #2A and #2B are completely independent of the resolved score because reader H1 defined the resolved score and H2's scores were used only for computing grading reliability. Consequently, in essay sets #2A and #2B the human reader score and the machine scores are compared to the same measure. That the human rater in essay sets #2A and #2B outperformed all of the machines in every metric except for one machine in Pearson r correlation offers some evidence that the high individual reader scores compared to the resolved score are not solely an artifact of their being a part of the whole. As shown in Table 9.3, #2A, which measured ideas, content, organization, style, and voice, had an exact agreement value of 0.76, compared to the range of machine values of 0.55-0.70. Its Kappa was 0.62, compared to the range of machine values of 0.30-0.51. Its Quadratic Weighted Kappa was 0.80, compared to the range of machine values of 0.62-0.74. And its Pearson r was 0.73, compared to the range of machine values of 0.62-0.74. Similarly, #2B, which measured conventions of grammar, usage, punctuation, and spelling, had an exact agreement value of 0.73 compared to the range of machine values of 0.55-0.69. Its Kappa was 0.56, compared to the range of machine values of 0.27-0.49. Its Quadratic Weighted Kappa was 0.76, compared to the range of machine values of 0.62-0.74. And its Pearson r was 0.76, compared to the range of machine values of 0.55-0.71. Significantly, the prompt in essay set #2 was a traditional argumentative prompt.

Table 9.3. Essay Set #2-H2 Score Compared to Resolved Score vs. Machine Scores

Metric	**2A**		**2B**	
	H2	**2A Range of Machine Scores**	**H2**	**2B Range of Machine Scores**
Exact Agreement	0.76	0.55-0.70	0.73	0.55-0.69
Kappa	0.62	0.30-0.51	0.56	0.27-0.47
Quadratic Weighted Kappa	0.80	0.62-0.74	0.76	0.62-0.74
Pearson *r*	0.73	0.62-0.74	0.76	0.55-0.71

In sum, the use of the artificially inflated resolved scores skews any meaningful analysis. This is particularly serious in the case of the quadratic weighted kappa, which is meant to compare the scores of two autonomous readers, not a reader score and an artificially resolved score. (Sim & Wright, 2005). In a subsequent report (Morgan, Shermis, Van Deventer, & Vander Ark, 2013), this same team uses the quadratic weighted kappa as the single measure of "the concordance between hand scores and machine scores" (p. 11), apparently unaware that they were measuring the concordance between resolved scores and machine scores and that in addition to all the other problems associated with employing resolved scores, the quadratic weighted kappa is an inappropriate measure for such a comparison.

The design of the study allows random chance to produce some seemingly impressive machine scores. Having a pair of readers compete against nine scoring engines is, in essence, like running multiple T-tests or any other kind of multiple comparisons. An occurrence can appear significant but might just be a lucky random occurrence. Any single high machine score among the nine scores by nine different vendors compared by five different metrics - that is 405 individual measures - could possibly be a random anomaly, or to put it in more colloquial terms, a lucky guess, especially since the size of the individual test essay sets were relatively small, ranging from 304 to 601. Statisticians have long known the dangers of producing what is called a Type I Error or False Positive when there are multiple comparisons without any overall testing of the entire model. When deciding if the difference between two variables is significant or possibly due to random chance, the standard statistical practice is to require that the probability of the difference being a product of random chance less than 5% or 1 in 20. But with repeated instances or comparisons, the probability of producing one or more statistically significant events increases. The chance of rolling two dice and getting two sixes is one in thirty six or 2.7%, but if I roll the dice thirty times, there is over a 50% chance I will roll two sixes. Although the case of comparisons in the study is slightly different, the basic analogy holds. The comparison of 405 measures to the resolved scores will produce some high correlations merely by chance.

The standard methodology to prevent these kinds of errors is to perform a test of the model as a whole. Unfortunately, no such tests were performed in the Shermis and Hamner study. Indeed, although various claims were made in the paper, no tests of statistical significance were reported by the authors. Instead, the authors present impressionistic assertions such as "In general, performance on kappa was slightly less with the exception of essay prompts #5 & #6. On these data sets, the AES engines, as a group, matched or exceeded human performance" (p. 23). There are no parameters given on what constituted matching or exceeding human performance.

Eleven months after the paper was presented and widely publicized, the lead author was quoted in the press as stating that he did not perform a regression analysis or any other statistical tests on the data in his study because that was one of conditions imposed upon him by major vendors of essay grading software, including McGraw-Hill and Pearson (Rivard, 2013). Such conditions were not disclosed in the Methods Section of the original paper, even though disclosures of such externally imposed restraints is standard practice in academic publications and especially in empirical studies such as this one.

FAULTY ANALYSIS: SMOKE AND MIRRORS

Overall, the analysis minimizes the accuracy of the human scorers and over-represents the accuracy of machine scoring. The clearest omission is the failure of the authors to report the fairly large percentage of machine values for the Pearson r and the Quadratic Weighted Kappa that fell below the minimum standard of 0.7. This value is used as the absolute minimum because the shared variance or what the machine clearly predicts is the square of that number, or approximately half of the population (Schultz, 2013; Ramineni & Williamson, 2013; Williamson, Xi, & Breyer, 2012). Any value below 0.7 will be predicting significantly less than half the population and, because this is an exponential function, small decreases in value produce large decreases in the percentage accurately predicted. A Pearson r of 0.6, for examples, yields a shared variance or predictive value of 0.36 or slightly more than one-third. A value of 0.5 yields 0.25 or one-quarter of the population. Yet for the Quadratic Weighted Kappa, 28 of the 81 machine scores, 35.6%, are below the minimally acceptable level of 0.7, even though the machines had the advantage in half of the essay sets of matching an inflated Resolved Score. In contrast, the human readers, who had to match each other with no artificial advantage, had only one Quadratic Weighted Kappa below 0.7, for the composite score on essay set # 8 or only 1 out of 9 or 11.1%. Similarly, for the Pearson r, the human readers again had a value below 0.7 for essay set #8 or 11.1% while 23 out of 91 of the machine scores or 28.4% were below the minimum threshold of 0.7.

The claim that the analysis in the paper unfairly underrepresents the performance of the human readers is further substantiated by the comparison between human readers and machine scores condensed in Tables 9.4-9.7. I did not include a table for adjacent and exact agreement because with many of the scales being 1-4, adjacent and exact agreement was often at 0.99 for both humans and machines.

Table 9.4. Exact Agreement Summary

Essay Set	Human Readers			Machines		
	H1	H2	H1H2	Median	Mean	Range
1	0.64	0.64	0.64	0.44	0.43	.31-.47
2a	—-	0.76	0.76	0.68	0.66	.55-.70
2b	—-	0.73	0.73	0.66	0.65	.55-.69
3	0.89	0.83	0.72	0.69	0.67	.61-.72
4	0.87	0.89	0.76	0.65	0.64	.47-.72
5	0.77	0.79	0.59	0.68	0.66	.47-.71
6	0.8	0.81	0.63	0.64	0.64	.51-.69
7	0.28	0.28	0.28	0.12	0.12	.07-.15
8	0.35	0.35	0.29	0.16	0.16	.08-.23

Exact agreement is summarized in Table 9.4. The report aggregates the ranges of agreement for the two human readers H1H2 among all eight essay sets and all nine rows of data, stating on page 22 that "The human exact agreements ranged from 0.28 on essay set #8 to 0.76 for essay set #2." The report then states, "the predicted machine score and had a range from 0.07 on essay set #2 [sic] to 0.72 on essay sets #3 and #4. An inspection of the deltas on Table 9 shows that machines performed particularly well on essay sets #5 and 6, two of the source-based essays."

The report ignores how human scorers performed better than the machines for most of the essay sets. Of the nine scores, the human rater agreement coefficients exceeded the top score of the machines in six of them, tying in a seventh. In essay set # 1 both readers performed .17 better than the best performing machine. In essay set #2A, the single "read-behind" reader performed .06 better than the best performing machine. In essay set #2B, the single "read-behind" reader performed .04 better than the best performing machine. The next four essay sets are content-based reading tests. For essay sets #3 and #4, the agreement of the two readers outperforms all but one of the machines and ties that one. The report also makes the careless error of incorrectly attributing the 0.07 exact agreement to essay set #2 instead of to essay set #7.

Table 9.5. Kappa Summary

Essay Set	Human Readers			Machines		
	H1	H2	H1H2	Median	Mean	Range
1	0.53	0.53	0.45	0.29	0.28	.16-33
2a	—-	0.62	0.62	0.48	0.46	.30-.51

Essay Set	Human Readers			Machines		
	H1	H2	H1H2	Median	Mean	Range
2b	—-	0.56	0.56	0.45	0.42	.27-.49
3	0.83	0.77	0.57	0.53	0.52	.45-.59
4	0.82	0.84	0.65	0.50	0.50	.30-.60
5	0.69	0.71	0.44	0.55	0.52	.28-.59
6	0.70	0.71	0.45	0.46	0.46	.31-.55
7	0.23	0.23	0.18	0.07	0.07	.03-.09
8	0.26	0.26	0.16	0.09	0.08	.04-.13

Table 9.5 summarizes the Kappa scores. On page 23, the report states that "in general, performance on kappa was slightly less with the exception of essay prompts #5 & #6. On these essay sets, the AES engines, as a group, matched or exceeded human performance." While this last claim is true for essay set #5, it was not true for essay set #6, where the value for H1H2 fell right in the middle of the machine scores. Moreover, the machine performance was not "slightly" lower than human performance measured by H1H2, it was substantially lower for all essay sets except 5 & 6 as can be observed simply by comparing H1H2 with the median and range values of the machine scores in Table 9.5.

Table 9.6. Quadratic Weighted Kappa Summary

Essay Set	Human Readers			Machines		
	H1	H2	H1H2	Median	Mean	Range
1	0.77	0.78	0.73	0.78	0.77	.66-.82
2a	—-	0.80	0.80	0.70	0.70	.62-.74
2b	—-	0.76	0.76	0.66	0.65	.55-.69
3	0.92	0.89	0.77	0.72	0.71	.65-.75
4	0.93	0.94	0.85	0.76	0.77	.67-.81
5	0.89	0.90	0.74	0.81	0.79	.64-.82
6	0.89	0.89	0.74	0.76	0.74	.65-.81
7	0.78	0.77	0.72	0.77	0.75	.58-.84
8	0.75	0.74	0.61	0.68	0.67	.60-.73

Tables 9.6 and 9.7 summarize the scores on the quadratic weighted kappa and the Pearson r. As mentioned previously, the machines do better on the quadratic weighted kappa except for score sets #2A and #2B and the literary analysis questions, essay sets #3 and #4. The performance of H1H2, the comparison of the two readers' scores, is mixed against the machine scores.

Table 9.7. Pearson r Summary

Essay Set	Human Readers			Machines		
	H1	H2	H1H2	Median	Mean	Range
1	0.93	0.93	0.73	0.80	0.77	.76-.82
2a	—-	0.80	0.80	0.71	0.70	.62-.74
2b	—-	0.76	0.76	0.67	0.66	.55-.71
3	0.92	0.89	0.77	0.72	0.71	.65-.75
4	0.94	0.94	0.85	0.76	0.77	.68-82
5	0.89	0.90	0.75	0.81	0.79	.65-.84
6	0.89	0.89	0.74	0.77	0.75	.65-.81
7	0.93	0.93	0.72	0.78	0.76	.58-.84
8	0.87	0.88	0.61	0.70	0.68	.62-.73

These results, as stated previously, may simply be the artifact of using different measures for machines and human readers as well as the improper use of the quadratic weighted kappa.

CONCLUSION

The study's numerous and substantial defects clearly undermine its conclusions. Only three of the eight essay sets used in the study contained scores that assessed students' ability to write more than a paragraph, and only one of the five other essay sets contained scores that were concerned with writing ability at all. Even more disturbing was that, with the exception of essay set #2, the study did not measure the correspondence between human readers and machine scores but used different measures for human and machine reliability that artificially inflated machine performance in half the essay sets. For the one essay set, #2, in which the study directly compared human and machine reliability, human readers were clearly more reliable than all of the machines for both of the writing scores contained in this essay set. Moreover, the study failed to follow standard statistical practice to guard against false positives and also made its assertions in the absence of any statistical tests, only based on the impressions of the authors. Consequently, Professor Shermis and Mr. Hamner should consider formally retracting all versions of this study in print or, at a minimum, respond in print to the criticisms enumerated in this article. Even with the flawed overall design of the study, further and rigorous statistical analysis of data may yield some interesting and extremely important information. Moreover, there are pressing policy decisions that argue for further analysis of these data. This paper has been reported to both the Partnership for Assessment of Readiness of College and Careers and the Smarter Balanced Assessment Consortium. The data and conclusions in this report may inform decisions by these two consortia about the use of automated essay scoring in the high stakes testing connected to the Common Core Standards, therefore, it is imperative that the authors publicly post the raw test set data from this study for rigorous statistical analysis.

NOTE

Although the adjudication rules given for the essay set descriptions for Essay Sets #3 and #4 do not mention it, examination of the training set revealed that, like Essay Sets #5 and #6, the resolved score was computed by taking the higher of two adjacent scores. There were no sets of scores in the training sample for Essay Sets #3 and #4 that contained pairs of scores that differed by more than one point, and no third rater scores. Consequently, four of the data sets from, at most, two states computed the resolved score by taking the higher score if the two rater scores were not identical. The authors mention, on page 9, instances in which the higher of the two scores in one essay set (#5) was not the resolved scores. In the two instances I identified, the two readers' scores were not adjacent and the resolved score was probably an adjudicated score.

REFERENCES

Austin, J. L. (1962). *How to do things with words.* Harvard University Press.

Brenner, H., & Kliebsch, U. (1996). Dependence of Weighted Kappa Coefficients on the Number of Categories. *Epidemiology , 7*(2), 199-202.

Giles, J. (2012). AI graders get top marks for scoring essay questions. *The new scientist, 2861.* http://www.newscientist.com/article/mg21428615.000-ai-graders-get-top-marks-for-scoring-essay-questions.html

Grice, H. P. (1989). *Studies in the way of words.* Harvard University Press.

Kaggle. (2012). Data—The Hewlett Foundation Automated Essay Scoring. http://www.kaggle.com/c/asap-aes/data

Kolowich, S. (2012). A Win for the Robo-Readers. I*nside Higher Ed.* http://www.insidehighered.com/news/2012/04/13/large-study-shows-little-difference-between-human-and-robot-essay-graders

Man and machine: Better writers, better grades. (2012). The University of Akron News. http://www.uakron.edu/im/online-newsroom/news_details.dot?newsId=40920394-9e62-415d-b038-15fe2e72a677

McCurry, D. (2010). Can machine scoring deal with broad and open writing. *Assessing Writing, 15(2),* 118-129.

Morgan, J., Shermis, M. D., Van Deventer, L., & Vander Ark, T. (2013). Automated Student Assessment Prize: Phase 1 & Phase 2. http://gettingsmart.com/wp-content/uploads/2013/02/ASAP-Case-Study-FINAL.pdf

Perelman, L. (2008). Information illiteracy and mass market writing assessments. *College Composition and Communication , 60*(1), 128-141.

Ramineni, C. A., & Williamson, D. A. (2013). Automated essay scoring: Psychometric guidelines and practices. *Assessing Writing, 18(1),* 25-39.

Rivard, R. (2013). Professors at odds on machine-graded essays. *Inside Higher Ed.* http://www.insidehighed.com/news/2013/03/15/professors-odds-machine-graded-essays

Schultz, M. T. (2013). The IntelliMetric Automated Essay Scoring Engine—A Review and an Application to Chinese Essay Scoring. In M. D. Shermis & J. Burstein (Eds.), Handbook of automated essay evaluation (pp. 89-98). Routledge.

Shermis, M. D., & Burstein, J. (Eds.). (2013). *Handbook of automated essay evaluation.* Routledge.

Shermis, M. D., & Hamner, B. (2012). Contrasting State-of-the-Art Automated Scoring of Essays: Analysis. http://www.scoreright.org/NCME_2012_Paper3_29_12.pdf

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* (pp. 213-246). Routledge.

Sim, J., & Wright, C. C. (2005). The Kappa Statistic in Reliability Studies: Use, Interpretation, and Sample Size Requirements. *Physical therapy, 85*(3), 257-268.

Sperber, D. (1990). *Relevance: communication and cognition.* Basil Blackwell.

Williamson, D. A., Xi, X., & Breyer, F. J. (2012). A framework for evaluation and use of automated scoring. *Educational Measurement: Issues and Practices , 31*(1), 2-13.

APPENDIX A

Prompts, Rubrics, and Other Materials

From The Hewlett Foundation: Automated Essay Scoring [Data and training material]. http://www.kaggle.com/c/asap-aes/data. Copyright 2012 by Kaggle. Reprinted with permission.

APPENDIX B

Selected Tables

Source: Mark D. Shermis & Ben Hamner, “Contrasting State of the Art Automated Scoring of Essays: Analysis.” https://www.semanticscholar.org/paper/Contrasting-state-of-the-art-automated-scoring-of-Hamner-Shermis/cad818cdb3b8bd2e28374316185782685482O9e1