Findings from the 2011 West Virginia Online Writing Scoring Comparability Study

West Virginia Department of EDUCATION
Office of Research

Findings from the 2011 West Virginia Online Writing Scoring Comparability Study

Nate Hixson

Vaughn Rhudy

West Virginia Department of EDUCATION
Office of Research

West Virginia Department of Education
Division of Teaching and Learning
Office of Research
Building 6-Room 722 State Capitol Complex
1900 Kanawha Boulevard East
Charleston, WV 25305
http://wvde.state.wv.us/

October 2012

Jorea M. Marple
State Superintendent of Schools
West Virginia Department of Education

Robert Hull
Associate Superintendent
West Virginia Department of Education

Larry J. White
Executive Director
Office of Research

Keywords
Automated Essay Scoring, Scoring Method Comparability, Writing, Assessment

Suggested Citation

Hixson, N. & Rhudy, V. (2012). *Findings from the 2011 West Virginia Online Writing scoring comparability study*. Charleston, WV: West Virginia Department of Education, Division of Teaching and Learning, Office of Research.

Content Contact
Nate Hixson
Assistant Director
Office of Research
nhixson@access.k12.wv.us

Executive Summary

To provide an opportunity for teachers to better understand the automated scoring process used by the state of West Virginia on our annual WESTEST 2 Online Writing Assessment, the WVDE Office of Assessment and Accountability and the Office of Research conduct an annual comparability study. Each year educators from throughout West Virginia receive training from the Office of Assessment and Accountability and then hand score randomly selected student compositions. The educators' hand scores are then compared to the operational computer (engine) scores, and the comparability of the two scoring methods is examined.

Method of study. A scoring group made up of 43 participants representing all eight regions scored a randomly selected set of student essays using the appropriate grade-level WV Writing Rubrics. A total of 2,550 essays were each scored by two different human scorers to allow for comparison of human-to-human scores as well as human-to-engine scores. Four hypotheses were tested.

Findings. We first sought to determine the extent to which human scorers calibrated their scoring process to align with the automated scoring engine via a series of training papers. We found that calibration was generally quite good in Grades 5-11, but there was room for improvement in Grades 3 and 4. We also found that calibration rates were relatively static across the set of training papers. We next sought to determine the comparability of human-to-human and human-to-engine agreement rates. We examined both exact and exact/adjacent agreement rates (i.e., scores that were exactly matched or within 1 point of each other on a 6-point scale). Looking at well-calibrated human scorers, our analyses showed that, with few exceptions, both exact and exact/adjacent agreement rates were comparable for the human-to-human and human-to-engine pairs. Finally, we examined the average essay scores assigned by the automated scoring engine and those assigned by a sufficiently calibrated human scorer. Our analyses revealed that for four of the available grade levels there were no significant differences. However, for the remaining grades and for all grades in aggregate, differences were statistically significant. In these cases, the difference observed between calibrated human scorers and the automated scoring engine were equivalent to or less than approximately three 10ths of a point (.310) on a 5-point scale or approximately 2% to 5% of the available points, with human scorers typically scoring papers higher. This difference was deemed to be practically insignificant.

Limitations of study. The human essay scores used in similar studies of automated essay scoring are generated by scorers for whom there is strong empirical evidence that indicates they are able to apply a validated scoring rubric in consistent and valid manner. In our case, employing 10 training papers is likely not enough training to ensure our scorers become expert raters. Until the calibration process and measure are improved upon, agreement rates and differences in human and engine scores should be interpreted cautiously.

Recommendations. We recommend improving the calibration process; examining new measures of calibration among scorers to assist in interpreting results; using multiple and different measures to examine agreement between scoring methods; and adding a quali-

tative research component to next year's online writing comparability study to examine teacher outcomes.

Contents

List of Figures

List of Tables

Introduction

The West Virginia Department of Education (WVDE) is committed to providing quality writing instruction in West Virginia schools. Writing is one of the most powerful methods of communication and a vital skill that students must develop throughout their school years to become college and career ready by the time they graduate. Students must be taught to articulate their thoughts and ideas clearly and effectively. To measure this ability, the WVDE began a statewide writing assessment in 1984.

The traditional paper/pencil assessment was administered in Grades 4, 7, and 10 from 1984 through 2004. In 2005, the WVDE led the first administration of a computer-based writing assessment, called the *Online Writing Assessment*. This assessment was expanded to Grades 3 through 11 in 2008 when the department conducted an online writing field test. The Online Writing Assessment then became a session of the West Virginia Educational Standards Test 2 (WESTEST 2) reading/language arts (RLA) assessment in 2009. Student performance on the online writing session is combined with student performance on the multiple choice sessions of the WESTEST 2 RLA assessment to determine students' overall performance levels; therefore, the assessment of student writing ability, in addition to their reading skills, has become an integral part of the state's accountability system.

The WESTEST 2 Online Writing Assessment is administered annually within a 9-week testing window. During the administration of the test, students in Grades 3–11 log onto a secure computer-based testing website. After students confirm their name and grade level, they receive a randomly assigned passage and prompt in one of the following four writing genres: narrative, descriptive, informative, or persuasive. (Students in Grade 3 receive either a narrative or descriptive passage and prompt.) Each student then responds to the prompt by typing his or her composition directly onto the secure website and then submitting that response for scoring.

Student responses are scored by an artificial intelligence computer-scoring engine trained with hand-scored student papers submitted as part of the 2008 field test. Scores are based on grade-level West Virginia Writing Rubrics in the analytic writing traits of organization, development, sentence structure, word choice/grammar usage, and mechanics. Scores range from a low of 1 to a high of 6 in each trait. The average of the five trait scores is then used in the item response theory model by the test vendor to derive students' scale scores for the RLA subtest.

CTB/McGraw-Hill, the state's testing vendor, conducts annual validation studies to confirm and validate the artificial intelligence scoring and to make any necessary adjustments to the scoring engine. Additionally, the vendor conducts a read-behind in which trained human scorers hand score 5% of student submissions each year; the hand scores are compared to the computer scores to ensure accuracy, reliability, and validity.

After the first operational administration of WESTEST 2 Online Writing Assessment in 2009, the WVDE Office of Assessment and Accountability and the WVDE Office of Research began conducting their own annual comparability study, in which selected educators from throughout West Virginia hand score randomly selected student compositions. The WVDE Of-

fice of Research then compares the educators' hand scores to the operational computer scores. The purpose of the comparability study is twofold. First, it serves as a valuable professional development experience for educators in how to appropriately score a student essay based on the grade-level WV Writing Rubrics. Second, it helps to build understanding in the field about the reliability of the automated scoring engine. That is, while automated essay scoring is a very efficient process that allows the test vendor to score several thousand student essays with minimal time requirements, it is sometimes perceived as untrustworthy by educators, some of whom believe human scorers are better able to reliably and accurately score student essays. The online writing comparability study seeks to address this issue.

The WVDE conducted its third WESTEST 2 Online Writing comparability study over a 2-day period in October 2011. Participants included the 43 human scorers selected to participate in the comparability study as described above. Nine educators who had previous scoring experience were invited to serve as table leaders during the 2-day scoring. Following an explanation of comparability study and artificial intelligence scoring, table leaders led participants through a training process. Participants hand scored training sets of 10 randomly selected student responses representing the various genres and various levels of student ability. Table leaders led discussions of each student response, the human scores, and the computer scores as participants progressed through each of the 10 compositions included in the training sets.

After training sets were completed, participants began scoring a randomly selected set of student responses using the appropriate grade-level WV Writing Rubric, recording their scores on a scoring sheet. Table leaders, who also served as scorers, tracked scoring packets to ensure all secure materials were returned as scorers completed their packets. Each essay was scored by two different human scorers to allow for comparison of human-to-human scores as well as human-to-engine scores.

On the second day of scoring, table leaders were provided computer scores for a small sample of student essays for the purpose of recalibrating human scorers. At the completion of the 2-day scoring, all essays and score sheets were collected. The Office of Assessment and Accountability then scanned all score sheets to collect the human scores for all essays.

We posed three research questions (RQs) and four associated hypotheses as part of this research study.

RQ1. What is the level of calibration to the automated scoring engine that is achieved among WV human scorers as a result of the training that is provided by the WVDE?

Hypothesis 1 (*H*1): The median exact agreement rate among human and engine scores will increase as participants score more training papers.

Hypothesis 2 (*H*2)*:* The training will yield an adequate number of scorers in each grade level who are sufficiently calibrated to be compared to the engine.

RQ2. What are the rates of agreement among WV human scorer pairs as well as between a pair consisting of a sufficiently calibrated human scorer and the automated engine?

Hypothesis 3 *(H3):* Human-to-human and human-to-engine exact and exact/adjacent agreement rates will be comparable.

RQ3. What is the level of variability in essay scores assigned by the automated essay scoring engine and sufficiently calibrated human scorers?

Hypothesis 4 (*H*4)*:* The average essay score assigned by the automated scoring engine, defined as the average of the five trait scores, will be comparable to the corresponding score assigned by a sufficiently calibrated human scorer.

Methods

Participant Characteristics

The Office of Assessment and Accountability invited 45 educators—five for each grade level, Grades 3 through 11—to participate in the annual study. Two educators who indicated they would participate canceled, leaving a total of 43 participants. Many of these educators served as members of the WESTEST 2 Online Writing Technical Advisory Committee and had previous scoring experience. The remaining participants were invited from a list of educators recommended by county superintendents and county test coordinators as having expertise in writing instruction and assessment.

Sampling Procedures

The participants were purposely selected to provide representation from all eight of the state's regional education service agencies (RESAs). Table 1 represents the breakdown of participants by RESA.

We utilized the total population of 43 human scorers to address hypotheses related to RQ1. The dataset was constructed with each case representing a single human scorer. Each row contained the five trait scores assigned by the human scorer and the corresponding five trait scores assigned by the automated engine for the same essay. This information was included for each scorer for all ten of the training papers utilized during calibration training.

Table 1. Online Writing Scoring Comparability Participants by RESA.

RESA	No. of participants
Total	43
1	6
2	5
3	5
4	4
5	6
6	5
7	7
8	5

To address RQ2 and RQ3, we constructed another dataset where each case represented a single student essay. Each row contained the five trait scores assigned by the two human scorers assigned to score the essay and the corresponding scores assigned by the automated scoring engine. Addressing *H*3 required examining the agreement rates for the human-to-human pairs as well as a sample of human-to-engine pairs. Because each essay was already assigned to a pair of human scorers at the outset of the comparability study, we were able to simply select all cases for which valid data were obtained and no sampling was necessary to assess human-to-human agreement rates. To assess human-to-engine agreement rates, we had to decide how to select one scorer from among the two available to compare with the automated engine. We did so systematically, by selecting a human-to-engine pair that consisted of the more calibrated of the two human scorers and the engine. In the case of a tie, where both human scorers were calibrated to the same level, we selected the first available human scorer. We also conducted a secondary set of analyses where we filtered the resulting dataset upon the calibration statistic, selecting only those human scorers who met at least 60% median exact agreement during the calibration training. These scorers were then compared with the engine to determine additional human-to-engine agreement rates.

Sample Size, Power, and Precision

Only *H*1 and *H*4 utilized inferential statistics, and therefore, power analyses are only relevant to the tests associated with those hypotheses. For these tests, we used paired-samples *t* tests. We calculated the sample size necessary to achieve 95% confidence in small and moderate effect sizes using this type of test.

A total of 2,550 essays were included in the dataset. An additional 10 papers per grade level (90) were used for calibration training. Table 2 contains the necessary sample size to achieve adequate power for these analyses. For *H*1, we had an effective sample size of 43, enough to detect a medium effect, but not a small effect. For *H*4, our analyses were conducted in aggregate and by grade level. Aggregate analyses well exceeded the sample size necessary to achieve 95% confidence in either a small or medium effect. Most grade level analyses contained approximately 300 cases, and therefore met this condition as well. However, in Grades 3 and 4, we had only 150 essays in our sample. Therefore, we did not have adequate power to detect small effects in these grade levels. However, we did have adequate power to detect medium effects.

Table 2. Sample Size Necessary to Achieve 95% Confidence in Small and Medium Effect Sizes

Hypothesis	Analysis used	Sample size necessary to achieve 95% confidence in a moderate effect (f^2 = .15, d =.5)	Sample size necessary to achieve 95% confidence in a small effect (f^2 = .02, d =.2)
*H*1	Repeated measures ANOVA	15	105
*H*4	Paired-samples *t* test	45	272

Measures and Covariates

Calibration

RQ1 dealt with assessing outcomes of the calibration training provided by the WVDE. To do so, we calculated a calibration statistic for each scorer. *The calibration statistic represents the median proportion of exact agreement with the automated scoring engine across 10 calibration papers scored by each scorer during training*. As noted below, exact agreement occurs when the score assigned by a human scorer is exactly the same as the corresponding engine score (e.g., human rating for mechanics is 3 and engine rating for mechanics is 3).

We used a three-step process to arrive at a measure for each scorer which would represent their overall calibration to the scoring engine. First, we determined if the human score for each trait was an exact match to the corresponding engine score. This was done using a simple logic test (i.e., does X = Y?). Second, we calculated across all five traits, the percentage of traits for which we observed exact agreement among the human scorer and the engine. For example, if the human scorer and engine agreed exactly on the scores for a given paper's mechanics, development, and organization traits, the resulting calibration for that paper would be 3/5 or 60% exact agreement. Third, we operationalized each scorer's overall calibration as the median calibration rate for the 10 papers scored during training.

Table 3 contains an example of the process using simulated data. In this example, Scorer 101 had a range of calibration from 0% on Paper 10 (no exact agreement) to 100% on Papers 1,

4, 5, and 6 (agreement on all 5 traits). To arrive at a single value for this scorer's overall calibration, we calculated the median percent calibration across these 10 training papers. In this case, the median is 60%, meaning that half of the scorer's paper scores were above 60% exact agreement with the engine, and the remaining half of the papers were below this level of agreement.

Table 3. Simulation of Process for Developing a Calibration Statistic for a Human Scorer

Scorer	Human Trait Scores					Engine Trait Scores					Agreement Status					Calibration	
101	O	D	WC	SS	M	O	D	WC	SS	M	O	D	WC	SS	M	CALIB	MEDIAN CALIB
Paper 1	1	2	1	1	2	1	2	1	1	2	Y	Y	Y	Y	Y	100%	60%
Paper 2	2	3	2	2	4	2	1	1	1	1	Y	N	N	N	N	20%	
Paper 3	5	3	5	5	5	5	5	4	4	4	Y	N	N	N	N	20%	
Paper 4	6	5	4	5	5	6	5	4	5	5	Y	Y	Y	Y	Y	100%	
Paper 5	2	2	2	2	2	2	2	2	2	2	Y	N	Y	Y	Y	100%	
Paper 6	5	4	5	6	6	5	4	5	6	6	Y	Y	Y	N	N	100%	
Paper 7	1	2	1	1	1	2	1	1	2	1	N	N	Y	N	Y	40%	
Paper 8	3	4	4	3	3	3	5	4	3	3	Y	N	Y	Y	Y	80%	
Paper 9	6	6	6	6	6	5	6	5	5	5	N	Y	N	N	N	20%	
Paper 10	5	6	5	5	5	3	5	4	3	3	N	N	N	N	N	0%	

Trait abbreviations: O = organization, D = development, WC = word choice/grammar usage, SS = sentence structure, and M = mechanics

Exact agreement

Our strategy to address *H*3 involved examining rates of exact and exact/adjacent agreement for each of the five traits and overall across traits. As noted above, exact agreement was defined as the circumstance when, examining the same essay, a score assigned by one scorer is exactly the same as the corresponding score assigned by another human or engine scorer. In this study, we calculated exact agreement using simple logic tests for each essay in the dataset (i.e., does X = Y?). We did so for all five traits for the pair of human scorers as well as for each of the two possible human-to-engine pairs.

Exact agreement rates were operationalized as the percentage of instances of exact agreement observed across all cases. For example, in a sample of 300 Grade 3 essays, if we observed exact agreement among two humans in their mechanics scores for 150 essays, the exact agreement rate for mechanics would be 150/300 or 50%. Similarly, if we examined the same 300 essays but examined agreement among one human scorer and the automated engine and observed 140 exact matches, our agreement rate would be 140/300 or 47%. In this example, the difference between human-to-human and human-to-engine exact agreement in mechanics for Grade 3 would be approximately 3% in favor of human-to-human agreement.

Exact/adjacent agreement

Exact/adjacent agreement was defined as the circumstance when, examining the same essay, a score assigned by one scorer is exactly the same as the corresponding score assigned by another scorer or is equal to that score +/- one point. We calculated exact/adjacent agreement rates using simple logic tests (e.g., is X = Y or Y+/-1?) This is similar to applying a margin of error of 1 point. For example, exact/adjacent agreement would be met if scorer 101 rated an essay's mechanics at 4 and scorer 102 rated the same essay's mechanics at either 3, 4, or 5. The two scores do not match (exact agreement), but are within one point of each other (adjacent agreement). As with exact agreement rates, exact/adjacent agreement rates were operationalized as the percentage of instances of exact/adjacent agreement observed across all cases.

Table 4 provides an example of how exact and exact/adjacent agreement rates would have been calculated using simulated data. In this example, three essays were scored by a human scorer and the automated engine. The exact and exact/adjacent logic tests indicate, for each trait, if the scores are in agreement. Following each of these tables downward, the percentage of cases where agreement was met is indicated. For example, with respect to the trait of organization (O), the exact agreement rate for these three papers was 100% because for all papers, the human scorer and the engine assigned the same value. The exact agreement rate for the trait of development (D) is 66% because agreement was observed among the scorer and the engine for two of the three essays. However, the exact/adjacent rate for development is 100% because in the one case where the human and the engine diverged on their scores for development (essay 3002) the difference was within 1 point.

In Table 4, below the trait ratings appears an example of the overall exact and exact/adjacent agreement rates for these three essays. These rates are calculated by determining the median percentage of agreement for the traits. In our example, the overall exact agreement rate is 33% because this is the point at which half the agreement statistics for the five trait scores were above and half were below. For exact/adjacent agreement, more than half the scores were at 100%, so the overall exact/adjacent agreement rate for these three papers is 100%.

Table 4. Simulation of Process for Determining Exact and Exact/Adjacent Agreement Rates by Trait and Overall.

	Human Trait Scores					Engine Trait Scores					Exact Agreement					Exact/Adjacent Agreement				
Essay ID	O	D	SS	WC	M	O	D	SS	WC	M	O	D	SS	WC	M	O	D	SS	WC	M
3001	1	2	1	1	4	1	2	1	1	2	Y	Y	Y	Y	N	Y	Y	Y	Y	N
3002	2	3	2	2	4	2	1	1	1	1	Y	N	N	N	N	Y	Y	Y	Y	N
3003	5	3	5	5	5	5	5	4	4	4	Y	N	N	N	N	Y	N	Y	Y	Y
Exact and exact/adjacent rates for three essays by trait (percentage)											100	33	33	33	0	100	66	10	10	33
Overall exact and exact/adjacent rates for three essays (median percentage of trait rates)															33					100

Trait abbreviations: O = organization, D = development, WC = word choice/grammar usage, SS = sentence structure, and M = mechanics

Research Design

We used a mix of descriptive and inferential statistics to test our study hypotheses. Table 5 indicates the research design for each hypothesis.

Table 5. Presentation of Research Methods and Analyses Used for Study Hypotheses

Research question/hypothesis	Method	Analysis
RQ1. What is the level of calibration to the automated scoring engine that is achieved among WV human scorers as a result of the training that is provided by the WVDE?		
*H*1–The median exact agreement rate among human and engine scores will increase as participants score more training papers.	Descriptive statistics Inferential statistics	Graphical representation of calibration statistics over time. General linear model with repeated measures.
*H*2–The training will yield an adequate number of scorers in each grade level who are sufficiently calibrated to be compared to the engine.	Descriptive statistics	Frequency distribution of calibration statistics for scorers by grade level.
RQ2. What are the rates of agreement among WV human scorer pairs as well as between a pair consisting of a sufficiently calibrated human scorer and the automated engine?		
*H*3–Human-to-human and human-to-engine exact and exact/adjacent agreement rates will be comparable.	Descriptive statistics	Graphical representation of agreement rates for scoring pairs.
RQ3.What is the level of variability in essay scores assigned by the automated essay scoring engine and sufficiently calibrated human scorers?		
*H*4–The average essay score assigned by the automated scoring engine, defined as the average of the five trait scores, will be comparable to the corresponding score assigned by a sufficiently calibrated human scorer.	Inferential statistics	Paired samples *t* test to determine presence of statistically significant differences. Tests of effect size to estimate practical significance of differences.

Results

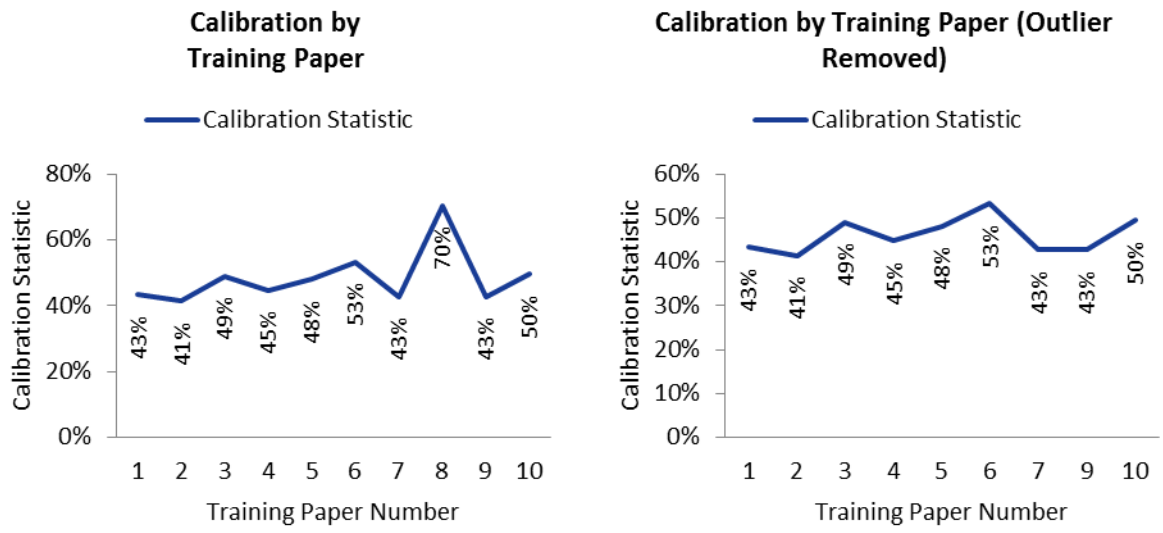
Research Question 1

RQ1 asked, *What is the level of calibration to the automated scoring engine that is achieved among WV human scorers as a result of the training that is provided by the WVDE?* To address this question, we tested two hypotheses.

Hypothesis 1 (*H*1)

*H*1 stated that the median exact agreement rate among human and engine scores will increase as participants score more training papers. Figure 1 represents the trend in median exact average agreement rates (calibration statistics) for the 42 scorers across the 10 training papers. Calibration ranged from 41% to 70%, but the average calibration rate across papers was 43%. Notably, paper 8 appeared to be an anomaly with approximately 70% calibration for the sample. Figure 1 presents the same trend, without paper 8 included. Notably, the trend appears much flatter with this outlier paper removed from the data.

Figure 1. Calibration by Training Paper With and Without Outlier



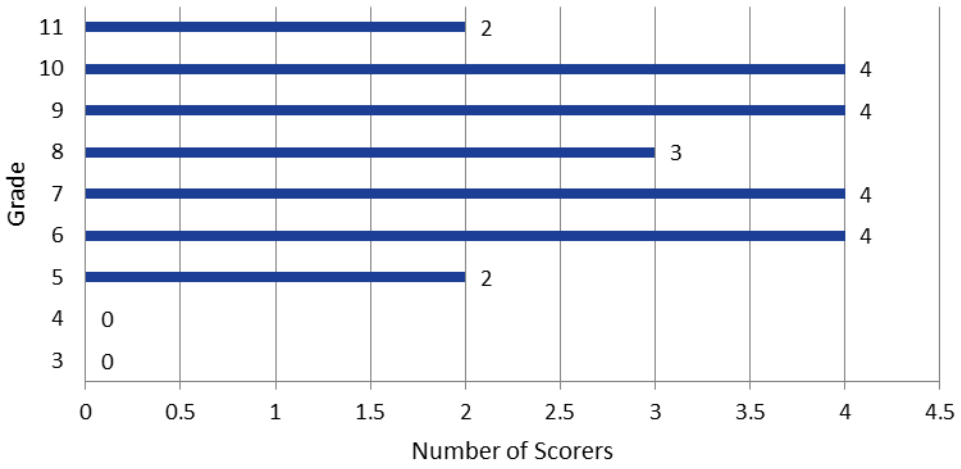
We used repeated measures analysis of variance (ANOVA) to determine if the trend in calibration changed significantly over time. For this analysis, we entered each of the 9 training papers (excluding paper 8) as a 9-level factor labeled *time*, and tested the within-subjects variability on this factor. The test indicated that the average calibration statistic did not improve significantly as human scorers had more opportunities to practice scoring F (8, 328) = .944, p = .480. *Since we were unable to find a significant difference in calibration statistics over time, we rejected* H1.

Hypothesis 2 (*H*2)

*H*2 stated that the training would yield an adequate number of scorers in each grade level who were sufficiently calibrated to be compared with the engine in subsequent analyses. For this analysis, we defined *sufficient calibration* as exceeding at least 60% exact agreement with the engine across the 10 training papers. To satisfy *H*2, we felt we would need to have at least two scorers in each grade level who met this criterion. Ideally, we would have liked to have observed a majority of scorers in each grade level reaching this level of calibration for subsequent comparisons as part of *H*3 and *H*4.

Figure 2 provides the frequency distribution of scorers reaching at least 60% calibration by grade level. Notably, Grades 5–11 had at least one scorer who met this criterion. However, there were no scorers who met the criterion in Grades 3 and 4. *Given these results,* H2 *was only partially supported in this study and we were limited to examining agreement rates for* H3 *using only the more calibrated of the two scorers in Grades 3 and 4.*

Figure 2. Number of Scorers Reaching at Least 60% Calibration



Research Question 2

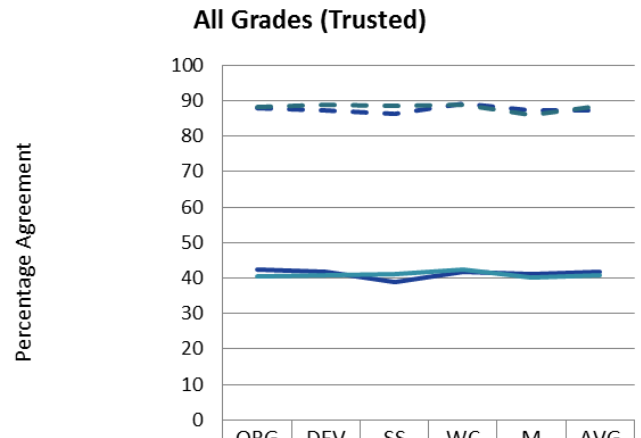
RQ2 asked, *What are the rates of agreement among WV human scorer pairs as well as between a pair consisting of a sufficiently calibrated human scorer and the automated engine?* We tested one hypothesis to answer this question.

Hypothesis 3 (*H*3)

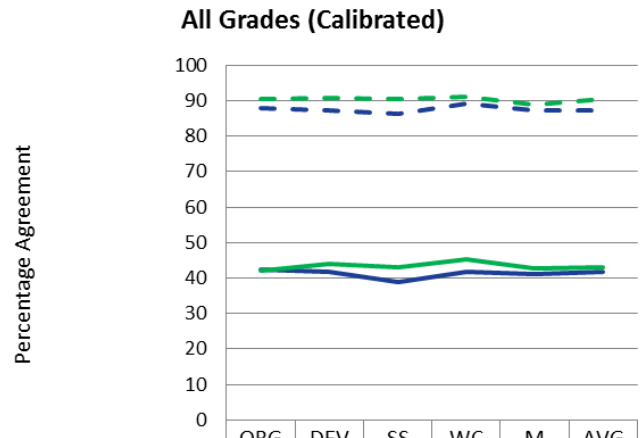
*H*3 stated that human-to-human and human-to-engine exact and exact/adjacent agreement rates would be comparable. Figure 3 represents the agreement rates for all grades aggregated together while Figure 4 through Figure 11 (pp. 15–22) illustrate the agreement rates for each trait by grade level.

For all figures, human-to-human rates of agreement were determined using the pair of human scorers assigned to score a given essay, while human-to-engine agreement rates were determined using either (a) the engine score and the score assigned by the *more calibrated* of the two human scorers assigned to score the essay (hereafter referred to as *trusted* scorers), or, when available, (b) the engine score and the score assigned by raters who reached at least 60% calibration at the conclusion of training (hereafter referred to as *calibrated* scorers).

Figure 3. Comparability Rates for All Grades Combined (Trusted and Calibrated Human-to-Engine)



	ORG	DEV	SS	WC	M	AVG
H2H Exact	42.4	41.9	38.8	41.8	41.1	41.8
H2H Exact+Adj	87.8	87.4	86.2	89.1	87.3	87.4
H2E Exact (Trusted)	40.5	40.7	41	42.3	40.1	40.7
H2E Exact+Adj (Trusted)	88.1	89	88.5	88.7	86	88.5

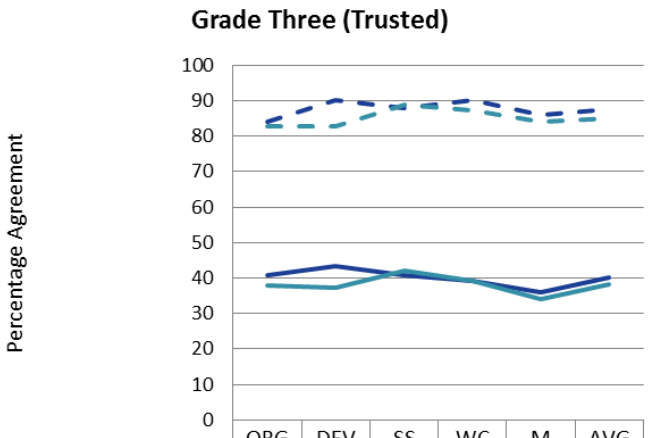


	ORG	DEV	SS	WC	M	AVG
H2H Exact	42.4	41.9	38.8	41.8	41.1	41.8
H2H Exact+Adj	87.8	87.4	86.2	89.1	87.3	87.4
H2E exact (Calibrated)	42.1	44	43.2	45.4	42.6	43.2
H2E Exact+Adj (Calibrated)	90.6	90.8	90.6	91	88.7	90.6

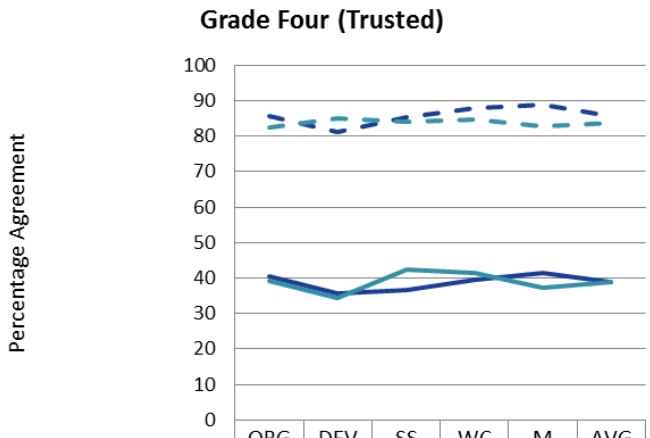
For all grades combined, human-to-engine agreement rates for trusted human scorers were almost identical to human-to-human rates. The average human-to-human exact agreement rate was 42% compared with 41% for trusted human-to-engine agreement. The human-to-human pair average exact/adjacent agreement rate was 87% compared with an average of 88% exact/adjacent agreement for the trusted human-to-engine pair.

When selecting all scorers who met at least 60% calibration, human-to-engine and human-to-human agreement rates were still quite similar. However, calibrated human-to-engine rates were consistently greater than those observed for human-to-human pairs. That is, the average human-to-human exact agreement rate was 42%, compared with 43% for calibrated human-to-engine pairs. The human-to-human pair average exact/adjacent agreement rate was 87% when compared with an average of 91% for the calibrated human-to-engine pair. *Taken together, these results seem to provide initial support for* H3.

Figure 4. Comparability Rates for Grades 3 and 4 (Trusted Human-to-Engine)



	ORG	DEV	SS	WC	M	AVG
H2H Exact	40.7	43.3	40.7	39.3	36	40
H2H Exact+Adj	84	90	88	90	86	87.6
H2E Exact (Trusted)	38	37.3	42	39.3	34	38.12
H2E Exact+Adj (Trusted)	82.7	82.7	88.7	87.3	84	85.08

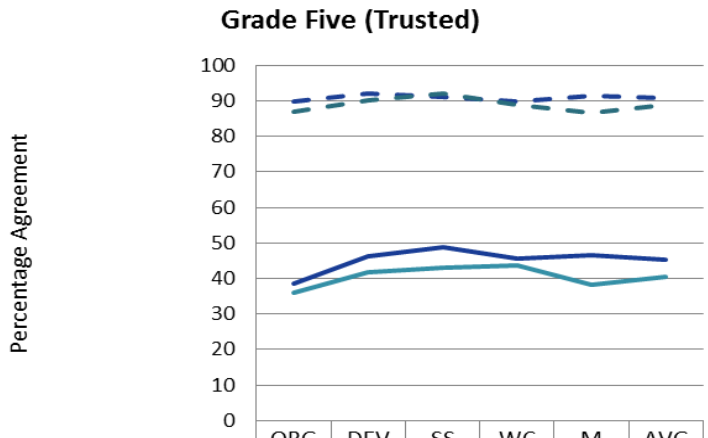


	ORG	DEV	SS	WC	M	AVG
H2H Exact	40.5	35.6	36.6	39.4	41.5	38.72
H2H Exact+Adj	85.6	81	85.2	88	88.7	85.7
H2E Exact (Trusted)	39.3	34.3	42.3	41.3	37.3	38.9
H2E Exact+Adj (Trusted)	82.3	85	84	84.7	82.7	83.74

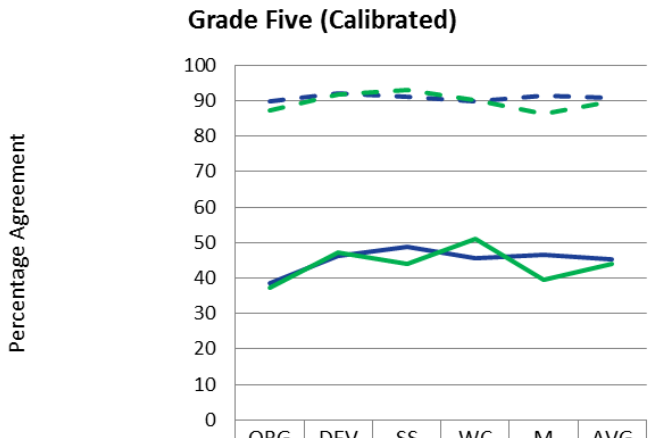
For Grade 3, human-to-engine agreement rates for trusted human scorers were generally quite comparable to human-to-human rates. The average human-to-human exact agreement rate was 40% compared with 38% for trusted human-to-engine agreement. The human-to-human pair average exact/adjacent agreement rate was 88% compared with an average of 85% exact/adjacent agreement for the trusted human-to-engine pair.

For Grade 4, human-to-engine agreement rates for trusted human scorers were also comparable to human-to-human rates. The average human-to-human exact agreement rate was 39% compared with 39% for trusted human-to-engine agreement. The human-to-human pair average exact/adjacent agreement rate was 86% compared with an average of 84% exact/adjacent agreement for the trusted human-to-engine pair. As noted previously, there were no Grade 3 or 4 scorers who met at least 60% calibration at the conclusion of training. Therefore, only trusted human-to-engine agreement rates are presented here.

Figure 5. Comparability Rates for Grade 5 (Trusted and Calibrated Human-to-Engine)



	ORG	DEV	SS	WC	M	AVG
H2H Exact	38.7	46.3	48.7	45.7	46.7	45.22
H2H Exact+Adj	89.7	92	91	89.7	91.3	90.74
H2E Exact (Trusted)	36	41.7	43	43.7	38.3	40.54
H2E Exact+Adj (Trusted)	87	90	92	89	86.7	88.94

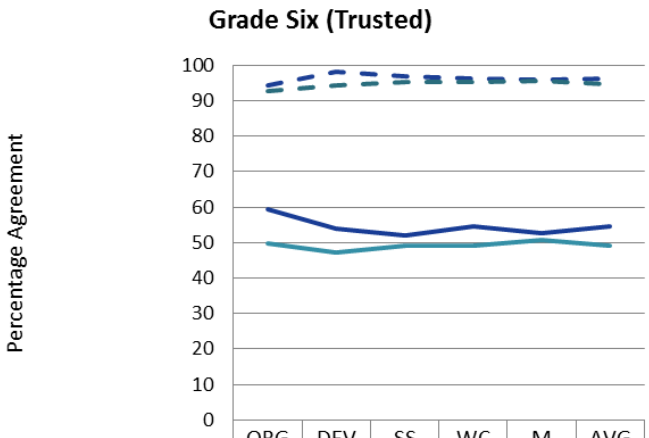


	ORG	DEV	SS	WC	M	AVG
H2H Exact	38.7	46.3	48.7	45.7	46.7	45.22
H2H Exact+Adj	89.7	92	91	89.7	91.3	90.74
H2E Exact (Calibrated)	37.4	47.3	44	51.1	39.6	43.88
H2E Exact+Adj (Calibrated)	87.4	91.8	92.9	90.1	86.3	89.7

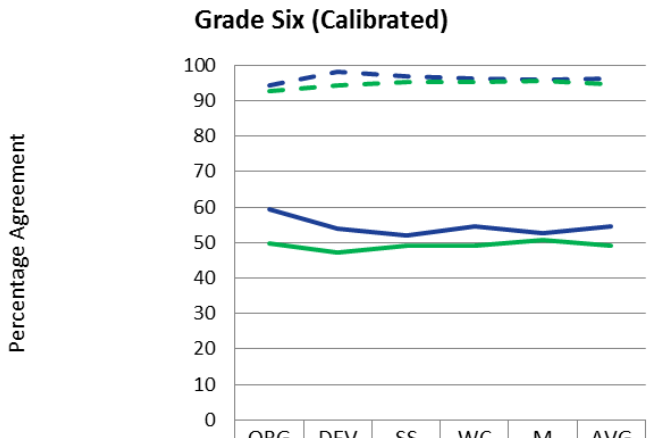
For Grade 5, the human-to-engine exact agreement rates for trusted human scorers were slightly lower when compared with human-to-human exact agreement rates. That is, the average human-to-human exact agreement rate was 45% compared with 40% for trusted human-to-engine agreement, a difference of 5 percentage points. The greatest differences were observed in ratings for development, sentence structure, and mechanics. However, it should be noted that the human-to-human and trusted human-to-engine exact/adjacent agreement rates were generally comparable by trait and overall with averages of 91% and 89%, respectively.

When isolating those Grade 5 human scorers who were calibrated to at least 60% (two of the available five scorers), we observed closer comparability among scorer pairs. That is, the average human-to-human exact agreement rate was 45% compared with a rate of 44% for calibrated human-to-engine pairs. The exact/adjacent rates were also very close at 91% and 90% for human-to-human pairs, and calibrated human-to-engine pairs, respectively.

Figure 6. Comparability Rates for Grade 6 (Trusted and Calibrated Human-to-Engine)



	ORG	DEV	SS	WC	M	AVG
H2H Exact	59.3	54	52	54.7	52.7	54.54
H2H Exact+Adj	94.3	98	97	96.3	96	96.32
H2E Exact (Trusted)	49.7	47.3	49	49	50.7	49.14
H2E Exact+Adj (Trusted)	92.7	94.3	95.3	95.3	95.7	94.66

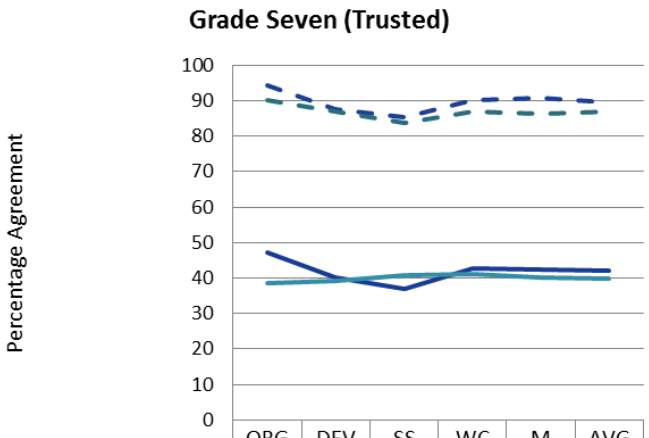


	ORG	DEV	SS	WC	M	AVG
H2H Exact	59.3	54	52	54.7	52.7	54.54
H2H Exact+Adj	94.3	98	97	96.3	96	96.32
H2E Exact (Calibrated)	49.7	47.3	49	49	50.7	49.14
H2E Exact+Adj (Calibrated)	92.7	94.3	95.3	95.3	95.7	94.66

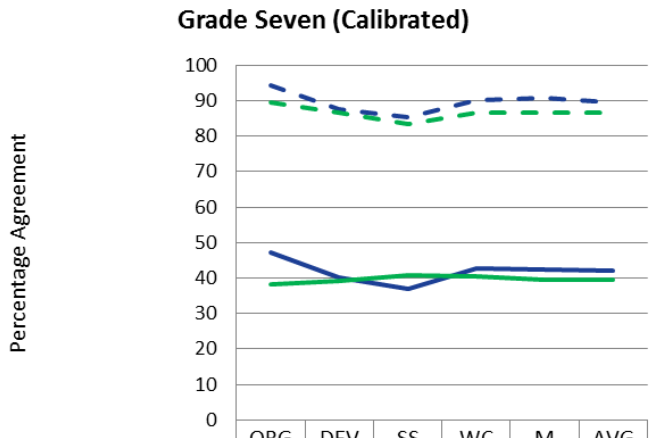
For Grade 6, the human-to-engine exact agreement rates for trusted human scorers were again lower when compared with human-to-human exact agreement rates. That is, the average human-to-human exact agreement rate was 54% compared with 49% for trusted human-to-engine agreement, a difference of 5 percentage points. The greatest differences were observed in ratings for organization, development, and word choice/grammar usage. However, it should again be noted that the human-to-human and trusted human-to-engine exact/adjacent agreement rates were generally comparable by trait and overall with averages of 96% and 95%, respectively.

Notably, the rates for trusted and calibrated scorers in Grade 6 are identical. This is because all four available Grade 6 human scorers met at least 60% calibration at the conclusion of training.

Figure 7. Comparability Rates for Grade 7 (Trusted and Calibrated Human-to-Engine)



	ORG	DEV	SS	WC	M	AVG
H2H Exact	47.3	40.3	37	42.7	42.3	41.92
H2H Exact+Adj	94.3	87.7	85.3	90	90.7	89.6
H2E Exact (Trusted)	38.7	39.3	40.7	41	40	39.94
H2E Exact+Adj (Trusted)	90	87	83.7	87	86.3	86.8

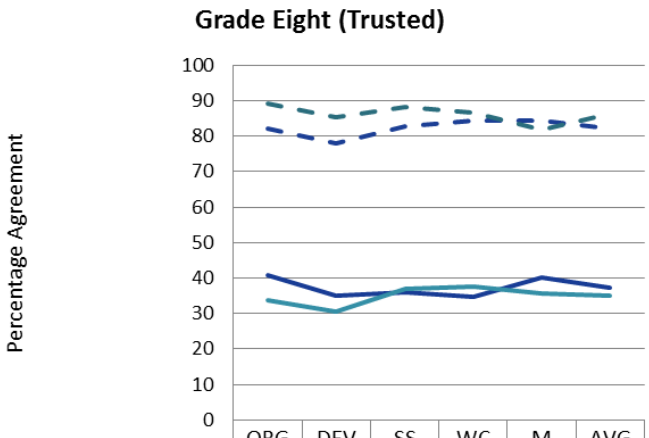


	ORG	DEV	SS	WC	M	AVG
H2H Exact	47.3	40.3	37	42.7	42.3	41.92
H2H Exact+Adj	94.3	87.7	85.3	90	90.7	89.6
H2E Exact (Calibrated)	38.1	39.1	40.8	40.5	39.4	39.58
H2E Exact+Adj (Calibrated)	89.6	86.5	83.4	86.5	86.5	86.5

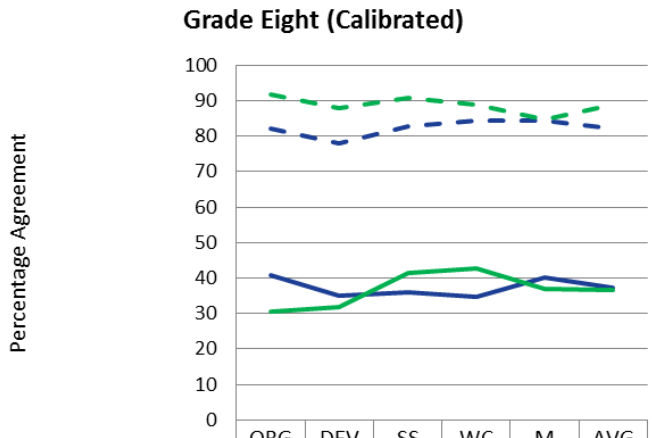
For Grade 7, human-to-engine agreement rates for trusted human scorers were generally quite comparable to human-to-human rates. The average human-to-human exact agreement rate was 42% compared with 40% for trusted human-to-engine agreement. The human-to-human pair average exact/adjacent agreement rate was 90% compared with an average of 87% exact/adjacent agreement for the trusted human-to-engine pair.

When isolating those Grade 7 human scorers who were calibrated to at least 60% (four of the five available scorers), we observed almost identical comparability among scorer pairs. That is, the average human-to-human exact agreement rate was 42% compared with a rate of 40% for calibrated human-to-engine pairs. The exact/adjacent rates were also very close at 90% and 86% for human-to-human pairs, and calibrated human-to-engine pairs, respectively.

Figure 8. Comparability Rates for Grade 8 (Trusted and Calibrated Human-to-Engine)



	ORG	DEV	SS	WC	M	AVG
H2H Exact	40.7	35	36	34.7	40	37.28
H2H Exact+Adj	82	78	82.7	84.3	84.3	82.26
H2E Exact (Trusted)	33.7	30.7	37	37.7	35.7	34.96
H2E Exact+Adj (Trusted)	89.3	85.3	88.3	86.7	81.7	86.26

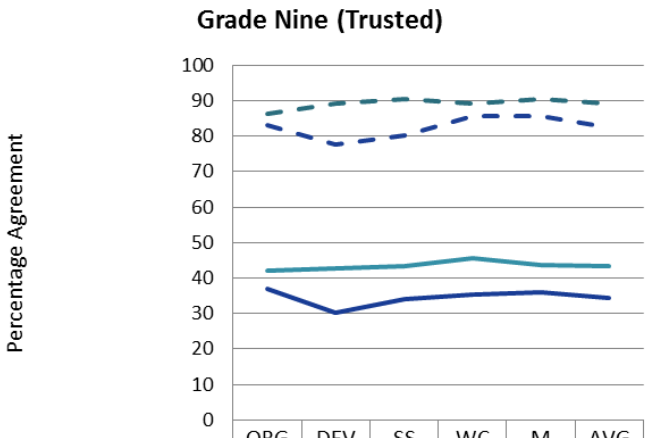


	ORG	DEV	SS	WC	M	AVG
H2H Exact	40.7	35	36	34.7	40	37.28
H2H Exact+Adj	82	78	82.7	84.3	84.3	82.26
H2E Exact (Calibrated)	30.4	31.7	41.3	42.6	37	36.6
H2E Exact+Adj (Calibrated)	91.7	87.8	90.9	88.7	84.8	88.78

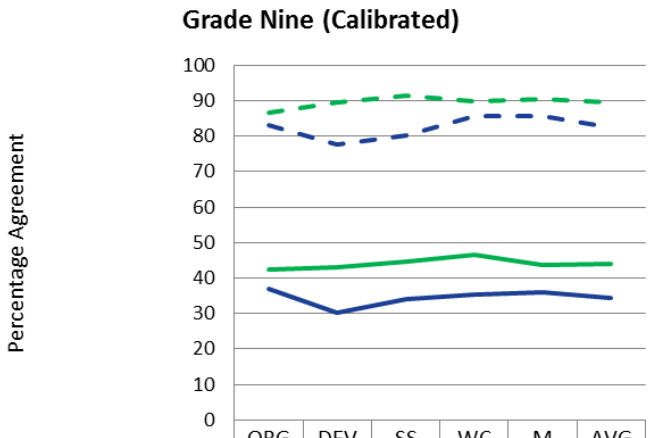
For Grade 8, human-to-engine exact agreement rates for trusted human scorers were generally comparable to human-to-human rates. The average human-to-human exact agreement rate was 37% compared with 35% for trusted human-to-engine agreement. However, the human-to-human pair average exact/adjacent agreement rate was considerably lower (i.e., 82%) when compared with an average of 86% exact/adjacent agreement for the trusted human-to-engine pair.

When isolating those Grade 8 scorers who were calibrated to at least 60% (3 of the 5 available scorers), we observed almost identical exact agreement rates among scorer pairs. That is, the average human-to-human exact agreement rate was 37% compared with a rate of 37% for calibrated human-to-engine pairs. Again, the exact/adjacent rates were considerably different when compared, with an average of 82% and 89% for human-to-human pairs, and calibrated human-to-engine pairs, respectively.

Figure 9. Comparability Rates for Grade 9 (Trusted and Calibrated Human-to-Engine)



	ORG	DEV	SS	WC	M	AVG
H2H Exact	37	30.3	34.2	35.2	35.9	34.52
H2H Exact+Adj	83.1	77.5	80.3	85.6	85.6	82.42
H2E Exact (Trusted)	42	42.7	43.3	45.7	43.7	43.48
H2E Exact+Adj (Trusted)	86.3	89.3	90.3	89.3	90.3	89.1

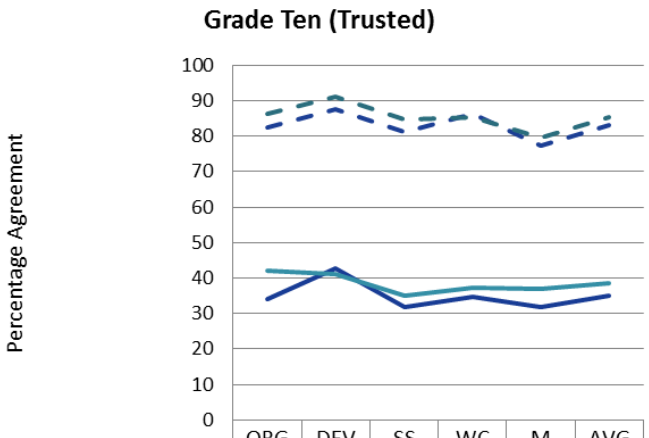


	ORG	DEV	SS	WC	M	AVG
H2H Exact	37	30.3	34.2	35.2	35.9	34.52
H2H Exact+Adj	83.1	77.5	80.3	85.6	85.6	82.42
H2E Exact (Calibrated)	42.3	43	44.7	46.5	43.7	44.04
H2E Exact+Adj (Calibrated)	86.6	89.4	91.5	89.8	90.5	89.56

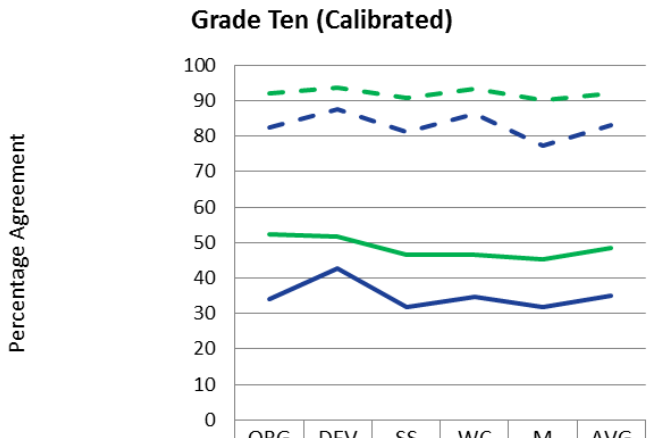
For Grade 9, human-to-engine exact agreement rates for trusted human scorers and human-to-human rates were considerably divergent. The average human-to-human exact agreement rate was 34% compared with 43% for trusted human-to-engine agreement, a difference of 9 percentage points. Likewise, the human-to-human pair average exact/adjacent agreement rate was considerably lower (i.e., 82%) when compared with an average of 89% exact/adjacent agreement for the trusted human-to-engine pair.

When isolating those Grade 9 scorers who were calibrated to at least 60% (four of the six available scorers), we observed almost identical agreement rates. That is, the average human-to-human exact agreement rate was 34% compared with a rate of 44% for calibrated human-to-engine pairs, a difference of 10 percentage points. Again, the exact/adjacent rates were considerably different when compared, with an average of 82% and 89% for human-to-human pairs, and calibrated human-to-engine pairs, respectively.

Figure 10. Comparability Rates for Grade 10 (Trusted and Calibrated Human-to-Engine)



	ORG	DEV	SS	WC	M	AVG
H2H Exact	34.1	42.8	31.8	34.8	31.8	35.06
H2H Exact+Adj	82.3	87.6	81.3	86.3	77.3	82.96
H2E Exact (Trusted)	42	41	35	37.3	37	38.46
H2E Exact+Adj (Trusted)	86.3	91	84.7	85.3	79.7	85.4

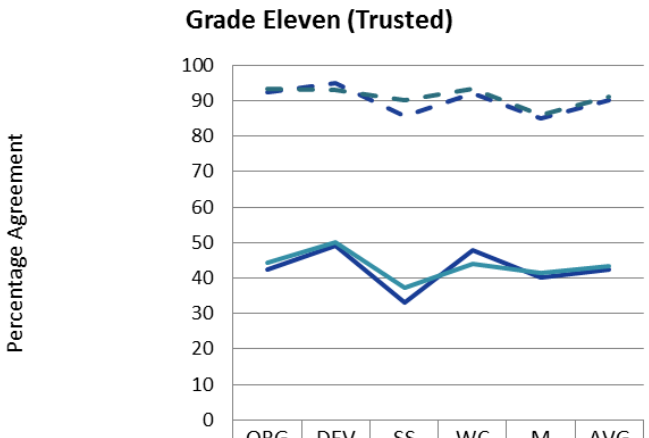


	ORG	DEV	SS	WC	M	AVG
H2H Exact	34.1	42.8	31.8	34.8	31.8	35.06
H2H Exact+Adj	82.3	87.6	81.3	86.3	77.3	82.96
H2E Exact (Calibrated)	52.2	51.6	46.6	46.6	45.3	48.46
H2E Exact+Adj (Calibrated)	91.9	93.8	90.7	93.2	90.1	91.94

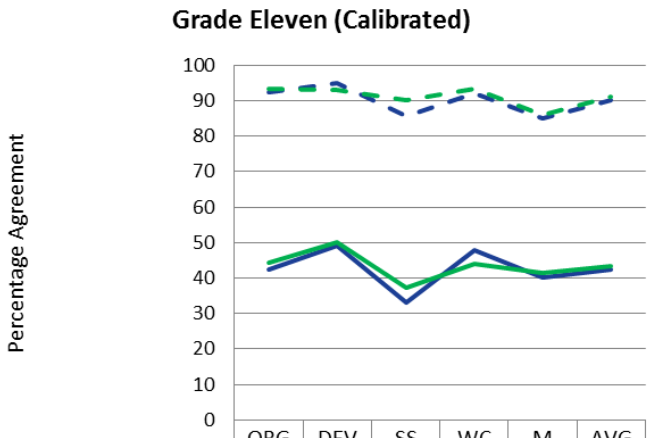
For Grade 10, human-to-engine exact agreement rates for trusted human scorers and human-to-human rates were generally comparable. The average human-to-human exact agreement rate was 35% compared with 38% for trusted human-to-engine agreement, a difference of 3 percentage points. Likewise, the human-to-human pair average exact/adjacent agreement rate was 83% compared with an average of 85% exact/adjacent agreement for the trusted human-to-engine pair.

When isolating those Grade 10 scorers who were calibrated to at least 60% (four of the seven available scorers), we observed quite different results. That is, the average human-to-human exact agreement rate was 35% compared with a rate of 48% for calibrated human-to-engine pairs, a difference of 13 percentage points. The exact/adjacent rates were also considerably different with averages of 83% and 92% for human-to-human pairs, and calibrated human-to-engine pairs, respectively, a difference of 9 percentage points.

Figure 11. Comparability Rates for Grade 11 (Trusted and Calibrated Human-to-Engine)



	ORG	DEV	SS	WC	M	AVG
H2H Exact	42.3	49	33	47.7	40	42.4
H2H Exact+Adj	92.3	95	85.7	92	85	90
H2E Exact (Trusted)	44.3	50	37.3	44	41.3	43.38
H2E Exact+Adj (Trusted)	93.3	93	90	93.3	86	91.12



	ORG	DEV	SS	WC	M	AVG
H2H Exact	42.3	49	33	47.7	40	42.4
H2H Exact+Adj	92.3	95	85.7	92	85	90
H2E Exact (Calibrated)	44.3	50	37.3	44	41.3	43.38
H2E Exact+Adj (Calibrated)	93.3	93	90	93.3	86	91.12

For Grade 11, human-to-engine exact agreement rates for trusted human scorers and human-to-human rates were generally comparable. The average human-to-human exact agreement rate was 42% compared with 43% for trusted human-to-engine agreement, a difference of only 1 percentage point. Likewise, the human-to-human pair average exact/adjacent agreement rate was 90% compared with an average of 91% exact/adjacent agreement for the trusted human-to-engine pair.

Notably, the rates for trusted and calibrated scorers in Grade 11 are identical. This is because both available Grade 11 human scorers met at least 60% calibration at the conclusion of training.

Summary of *H*3 findings

Table 6 presents the difference between average exact and exact/adjacent agreement rates by scoring pair for each grade level and overall. Here, we subtracted the average trusted human-to-engine agreement rate (the average agreement rate across traits) from the corresponding average human-to-human agreement rate. This yielded a measure of the difference between agreement rates that exists among scoring pairs.

Using this metric, a negative value would indicate that the trusted human-to-engine agreement rate was higher than the human-to-human agreement rate. Conversely, a positive value would indicate that the human-to-human agreement rate was higher than the trusted human-to-engine agreement rate.

Examining the table, we observed marginally higher exact and exact/adjacent agreement rates for the human-to-human pair when compared to the trusted human-to-engine pair. This trend was exhibited in five of the nine grade levels (55%). However, conversely we observed higher trusted human-to-engine rates in the remaining four grade levels (45%). *Further, the average difference in exact and exact/adjacent agreement rates across grade levels was less than 0.5% (i.e., 0.30% and -0.39, respectively). The latter particularly provides compelling support for* H*3, which stated the scoring pairs would produce generally comparable rates of agreement.*

It does bear mentioning that, although across grades we observed relatively comparable average agreement rates, we did observe a pattern whereby there were consistently higher agreement rates for the human-to-human pairs in Grades 3 to 8, but for the trusted human-to-engine pairs in Grades 9–11. It is unclear why this trend emerged.

Table 6. Difference in Agreement Rates by Scoring Pair (Human-to-Human–Trusted Human-to-Engine)

Grade	Exact agreement difference (Avg. H2H exact agreement rate–Avg. *trusted* H2E exact agreement rate)	Exact/adjacent agreement difference (Avg. H2H exact/adjacent agreement rate–*trusted* H2E exact/adjacent agreement rate)
3	1.88%	2.52%
4	-0.18%	1.96%
5	4.68%	1.80%
6	5.40%	1.66%
7	1.98%	2.80%
8	2.32%	-4.00%
9	-8.96%	-6.68%
10	-3.40%	-2.44%
11	-0.98%	-1.12%
Average difference observed across grades	0.30%	-0.39%

Table 7 presents the difference between average exact and exact/adjacent agreement rates by scoring pair for each grade level and overall. Here, we subtracted the average calibrated human-to-engine agreement rate (the average agreement rate across traits) from the corresponding average human-to-human agreement rate.

We observed a similar phenomenon as indicated above whereby, for some grades (i.e., 5–9), the average exact human-to-human agreement rate was consistently higher than the average exact calibrated human-to-engine agreement rate. In no case did these differences exceed 5.4%. In contrast, within Grades 10 and 11, the exact agreement rate was higher for the calibrated human-to-engine pair than for the human-to-human pair. This difference was considerable in Grade 10 with the average difference being -13.4% in favor of the calibrated human-to-engine pair.

With respect to exact/adjacent agreement rates, we observed considerably higher rates for the calibrated human-to-engine sample when compared with the human-to-human sample, particularly in Grades 8–11 (57% of the grade levels available for analysis). However, exact/adjacent agreement rates were higher for the human-to-human pair in Grades 5–7 (43% of available grade levels).

Examining the overall average difference among pairs revealed a difference of less than -0.50% among pairs when considering exact agreement and less than -3% with respect to exact/adjacent agreement. As a reminder, negative values indicate a slightly higher agreement rate for the calibrated human-to-engine pair than the human-to-human pair.

Overall, the data indicate that both exact and exact/adjacent agreement rates were quite comparable for these pairs, lending further evidence to support H3. However, it is immediately apparent that Grade 10 was an anomaly whereby agreement rates for the calibrated human-to-engine agreement pair were much higher than for the human-to-human pair. It is unclear why this finding emerged, but it may merit further analysis.

Table 7. Difference in Agreement Rates by Scoring Pair (Human-to-Human–Calibrated Human-to-Engine)

Grade	Exact agreement difference (avg. H2H exact agreement rate–avg. *calibrated* H2E exact agreement rate)	Exact/adjacent agreement difference (avg. H2H exact/adjacent agreement rate–*calibrated* H2E exact/adjacent agreement rate)
3	*	*
4	*	*
5	1.34%	1.04%
6	5.40%	1.66%
7	2.34%	3.10%
8	0.68%	-6.52%
9	2.34%	-7.14%
10	-13.4%	-8.98%
11	-0.98%	-1.12%
Average difference observed across grades	-0.32%	-2.56%

Research Question 3

RQ3 asked, *What is the level of variability in essay scores assigned by the automated essay scoring engine and sufficiently calibrated human scorers?* We tested one hypothesis to answer this question.

Hypothesis 4 (*H*4)

*H*4 stated that the average essay score assigned by the automated scoring engine, defined as the average of the five trait scores, would be comparable to the corresponding score assigned by a sufficiently calibrated human scorer. We conducted a series of 2-tailed paired *t* tests to determine if significant differences existed between scores. When significant differences emerged, we sought to quantify the magnitude of those differences using an estimate of effect size and by estimating the percentage of the available scale represented by the difference between scoring pairs.

Table 8 presents the tests of significance for the mean differences we observed between the scores assigned by pairs of human scorers. The mean differences were observed in both directions (positive and negative) and ranged from -.013 (Grade 4) to .059 (Grade 3). This indicates that the human scorers score essays differently in a manner that is not always predictable (e.g., sometimes lower than another scorer, sometimes higher). However, none of the differences we observed were statistically significant. So, regardless of direction, the differences we observed among human scorers likely do not represent a meaningful proportion of the available points.

Put another way, *for all grade levels, and in aggregate, the average difference observed between human scorers was insignificant, and represented less than approximately one 10th of a point (.10) on a 5-point scale, or less than 1% of the available points*.

Table 8. Tests of Significance for Mean Differences Observed Between Human Scorers

Grade	Mean difference	*SD*	*T*	*df*	Sig (2-tailed)	FLAG	ES	% of Scale
ALL GRADES	.008	.833	.497	2,516	0.79	NO	N/A	N/A
3	.059	.811	.886	149	0.61	NO	N/A	N/A
4	-.013	.846	-.252	283	0.89	NO	N/A	N/A
5	.039	.770	.885	299	0.61	NO	N/A	N/A
6	-.028	.561	-.864	299	0.62	NO	N/A	N/A
7	-.005	.827	-.098	299	0.96	NO	N/A	N/A
8	.019	.914	.354	299	0.85	NO	N/A	N/A
9	.023	1.000	.390	283	0.83	NO	N/A	N/A
10	.032	.945	.588	298	0.75	NO	N/A	N/A
11	-.027	.750	-.631	299	0.73	NO	N/A	N/A

Table 9 presents the tests of significance for the mean differences we observed between the scores assigned by trusted human scorers and the automated scoring engine. Notably, there

were five statistically significant differences (i.e., all grades, Grade 3, Grade 4, Grade 9, and Grade 10). For these grade levels, the mean differences ranged from -.328 to .103. In all cases, the differences were negative, indicating that the trusted human scorers tended to provide slightly higher average scores than the automated scoring engine.

We next conducted a series of analyses to determine the effect size represented by these differences. Effect sizes (Cohen's *d*[1]) ranged from .11 (small) to .30 (small-moderate). To make these measures more interpretable, we also calculated the percentage of available points represented by the average difference in scores. In no case did this percentage exceed 5% of the total available points.

In summary, *for eight of the 10 grade levels in our study (80%), and for all grades in aggregate, there were significant differences observed among the average scores assigned by trusted human scorers and the automated scoring engine. In these cases, the difference observed between trusted human scorers and the automated scoring engine were equivalent to or less than approximately three 10ths of a point (.330) on a 5-point scale, between 2% and 5% of the available points. In all cases, the automated engine assigned slightly lower average scores than trusted human scorers.*

Table 9. Tests of Significance for Mean Differences Observed Between Trusted Human Scorers and the Automated Scoring Engine

Grade	Mean difference	*SD*	*T*	*df*	Sig (2-tailed)	FLAG	ES (Cohen's *d*)	% of scale
ALL GRADES	-.152	.913	-8.387	2549	.000	YES	.13	2
3	-.187	1.04	-2.187	149	.030	YES	.15	4
4	-.233	.960	-4.209	299	.000	YES	.20	3
5	-.157	.906	-2.994	299	.003	YES	.13	2
6	-.033	.757	-.762	299	.447	NO	N/A	N/A
7	-.133	.989	-2.334	299	.020	YES	.11	2
8	.093	.898	1.799	299	.073	NO	N/A	N/A
9	-.330	.850	-6.720	299	.000	YES	.30	5
10	-.277	.953	-5.024	299	.000	YES	.23	5
11	-.127	.840	-2.611	299	.009	YES	.11	2

[1] Guidelines for the interpretation of Cohen's *d* are as follows: .20 = small, .50 = moderate, .80 = large.

Table 10 presents the tests of significance for the mean differences we observed between the scores assigned by calibrated human scores and the automated scoring engine. Notably, there were two significant differences (i.e., all grades and Grade 9). For these grade levels, the mean differences ranged from -.120 to -.311. The significant differences were in opposite directions, indicating that, for all grades, the human scorers tended to provide slightly higher average scores than the automated scoring engine. However, for Grade 9, the automated engine assigned slightly higher average scores than calibrated humans.

We next conducted a series of analyses to determine the effect size represented by these two significant differences. Effect sizes (Cohen's *d*) ranged from .12 (small) to .29 (small-moderate). To make these measures more interpretable, we also calculated the percentage of available points represented by the average difference in scores. As was the case with trusted human scorers, in neither case did this percentage exceed 5% of the total available points.

In summary, *for four of the available grade levels in our analysis (57%), there were no significant differences observed among the average scores assigned by calibrated human scorers and the automated scoring engine. However, for the remaining grades (*Grades *7, 9, and 11) and for all grades in aggregate, the differences were statistically significant. In these cases, the difference observed between calibrated human scorers and the automated scoring engine were equivalent to or less than approximately three 10ths of a point (.310) on a 5-point scale or approximately 2% to 5% of the available points. In all cases, the automated engine assigned slightly lower average scores than calibrated human scorers.*

Table 10. Tests of Significance for Mean Differences Observed Between Calibrated Human Scorers and the Automated Scoring Engine

Grade	Mean difference	*SD*	*T*	*df*	Sig (2-tailed)	FLAG	ES (Cohen's *d*)	% of Scale
ALL GRADES**	-.125	.857	-6.080	1746	.000	YES	.12	2
3	*	*	*	*	*	*	*	N/A
4	*	*	*	*	*	*	*	N/A
5	-.115	.869	-1.792	181	.075	NO	N/A	N/A
6	-.033	.757	-.762	299	.447	NO	N/A	N/A
7	-.152	.991	-2.610	288	.010	YES	.13	2%
8	-.052	.834	-.948	229	.344	NO	N/A	N/A
9	-.310	.850	-6.139	283	0.00	YES	.29	5%
10	-.037	.797	-.593	160	.554	NO	N/A	N/A
11	-.127	.840	-2.611	299	.009	YES	.12	2%

Conclusions

We rejected *H*1 on the basis that the average calibration statistic for our sample did not increase as scorers had more opportunities to evaluate student essays. Contrary to our belief, we found that the average calibration statistic remained relatively static over the course of the training. It is unclear why this is the case. One potential explanation would be variations in the difficulty and/or quality of student papers may have been a confounding factor. Another possible explanation is that the debriefing between each paper was not substantive enough to measurably improve subsequent agreement rates. Still another possibility is that the measure of calibration used in this study was insensitive to improvements in agreement rates. However, because these aspects were either unmeasured or difficult to assess in a post hoc manner, we cannot be sure which, if any, of these explanations are valid.

We found only partial support for *H*2. That is, we did not have scorers in Grades 3 and 4 who met our criterion of at least 60% calibration. However, as mentioned previously, the criterion for acceptable calibration was set in an ad hoc manner, and the measure of calibration may have been insensitive. Revisions to the process for determining calibration are recommended below.

We accepted *H*3 with some cautions. The overall comparability of scoring methodologies appears to be quite good. We found very small differences in overall agreement rates between human-to-human pairs and calibrated human-to-engine pairs (a difference of approximately 1% in average exact agreement rates across traits and 3% in average exact/adjacent agreement rates). However, it bears mentioning here that, when examining aggregated grade level data, the calibrated human-to-engine agreement rates were actually marginally higher than the human-to-human agreement rates. This was also drastically true in some grade levels where the calibrated human-to-engine agreement rates far exceeded human-to-human agreement rates (e.g., Grades 9 and 10). We also observed some differences in the within-trait agreement rates, and agreement rates varied across grades. We may attribute some of these differences to the fact that we observed wide variability in calibration within each grade level. Had we been better able to exercise control over calibration of human scorers, we believe comparability would have been increased.

We were unable to make a definitive conclusion regarding *H*4. We did observe small differences in the average scores assigned by the automated scoring engine and our sample of human scorers. In this examination, the engine tended to evaluate student essays in a more stringent manner, on average scoring those essays between 2% and 5% lower than the corresponding human scorers. While we don't believe these differences to be alarming in terms of their magnitude, we cannot avoid the fact that the differences were statistically significant in some grade levels and in aggregate. Several factors make it difficult for us to make a conclusion regarding the meaningfulness of these differences. First, we observed no statistically significant differences among human rater pairs. This indicates that, without considering validity of score interpretations, human scorers were certainly more consistent in their agreement of the overall quality of student essays. However, we must simultaneously acknowledge that we observed that as scorer calibration increased, the number of statistically significant differences between human and engine scores declined. Based on the latter fact, we hypothesize that if we were able to

exercise better control over scorer calibration among human scorers at the outset of the study, we would have eliminated the remaining significant differences. So, at this time, we are unable to make a sound conclusion about whether or not the overall scores assigned by humans and the engine are comparable. However, we can say that the results observed in this study depict a relatively small practical difference in scores.

Discussion

Interpretation of Findings

Our first research question dealt with assessing the outcomes of the calibration training component of the online writing comparability study. We found that, contrary to our hypothesis, we did not observe progressively higher average rates of agreement as more training papers were scored. Instead, the average exact agreement remained relatively static across all 10 papers. We did not take into account the fact that there would most certainly be variations in essay quality across each of the 10 training papers and that this variation would undoubtedly influence the scorers' ability to accurately score the papers. With this consideration it is not undesirable that the rate of exact agreement appears to be relatively static across papers. In fact, taking essay variability into account, a static agreement rate means that our human scorers were likely quite consistent in their ability to accurately score all papers.

Furthermore, upon examining the distribution of adequately calibrated raters (i.e., those that met or exceeded 60% median exact agreement), we found that sufficiently calibrated scorers were distributed across all grade levels except Grades 3 and 4. This posed some problems in subsequent analyses. However, it should also be noted that we observed some scorers who met even more robust levels of agreement (e.g., 70% to 80%). It would have been ideal if all scorers were trained to calibration before conducting comparisons. However, because this was impractical, and because the comparability study is also intended as a professional development experience, we commenced our research by examining agreement rates for *trusted* and *calibrated* scorers separately.

Doing so, we found what we believe to be relatively comparable rates of agreement across scoring pairs. There was certainly some within-trait variability, but the overall median exact agreement did not vary widely among human-to-human pairs and human-to-engine pairs. This provided some evidence that engine scoring is relatively robust.

Despite comparability of the engine scores to our trained human scores, we did observe statistically significant differences among human and engine scores for some grades. In all cases, it appeared that the automated engine scored student essays marginally lower than a human scorer. However, this difference was often very small, amounting to between 2% and 5% of the available points. We are reluctant to judge this difference as meaningful or alarming given that we observed evidence that suggests that significant differences between human and engine scorers may be increasingly less significant as human scorers are better trained. That is, in our study, the number of significant differences actually decreased quite noticeably when we examined only the most calibrated scorers in our sample. We may anticipate that, had we even more highly calibrated scorers or if we had a stronger measure of scorer calibration, we could have eliminated all significant differences between methods.

Limitations

The automated scoring engines employed in studies of this nature have been rigorously trained to apply a scoring algorithm with the express goal of accurately reproducing those scores which have been assigned by a sample of expertly trained humans. The human essay scores used in such studies are generated by scorers for whom there is strong empirical evidence that indicates they are able to apply a validated scoring rubric in a consistent and valid manner.

If our study were to attempt to replicate this process in a fair manner, we would have to train our human scorers using equally rigorous criteria prior to making any comparisons to the automated engine. In our case, employing 10 training papers is likely not enough training to ensure our scorers become expert raters. This was evidenced in that very few scorers met the criterion of even 60% median exact agreement with the engine. Moreover, even though we observed some degree of evidence of face validity for it, our measure of scorer calibration was potentially too simplistic. Had we a better measure of calibration or a more rigorous training process, we could expect much closer agreement between human scorers and the automated engine. With this in mind, we find our results to be rather promising. That is, we do not expect an engine that is trained to replicate the scoring process of highly trained expert human scorers to agree with absolute consistently with a set of human scorers who have only limited experience in scoring essays according to the rubric. However, as noted previously, when taking into account calibration, we saw generally comparable agreement rates and less significant differences between human and engine scores. This leads us to believe, as mentioned previously, that if we were better able to train our human scorers at the outset of the comparability study, we would have observed even greater agreement.

We should also mention that, because we had a relatively small sample of student responses, we were not able to adequately examine agreement rates or differences in scores within each prompt type. Instead, we examined these characteristics within grade level. Some genres may lend themselves to different scoring conventions. As such, a purer comparison would have examined agreement rates within prompts and grade levels.

Recommendations

Given the results of the current study and the limitations encountered, we recommend first making minor revisions to the training process used for next year's online writing comparability study. Specifically, we suggest that the Office of Assessment and Accountability increases the number of training papers used to at least 15. We hope that doing so will give scorers more opportunities to reach calibration at the outset of the study, allowing us to include more scorers in our subsequent analyses. This is particularly important for Grades 3 and 4, where we had no scorers who met at least 60% calibration in the current examination.

We also suggest examining new measures of calibration in addition to median exact agreement. In the current study, we developed the criterion of 60% in an ad hoc fashion. Using an index of interrater reliability such as Kappa will allow us to use common statistical guidelines that have been historically validated to select our most calibrated scorers (e.g., Kappa >.70).

In light of recent research on automated essay scoring, we also recommend expanding the examination of agreement in our study by using additional statistics such as Kappa and Pearson's correlation coefficient. Using multiple measures to examine agreement may provide a more complete picture of the comparability of scoring methods.

Finally, we recommend adding a qualitative research component to next year's online writing comparability study to examine teacher outcomes. The comparability study presents a unique opportunity to collect feedback from teachers about their perceptions of the fairness and rigor of automated essay scoring. Likewise, the Office of Assessment and Accountability could ask about the extent to which the online writing comparability study helps to build teachers' capacity to use the WV Writing Rubrics and to better understand how to teach and assess writing.

GLOBAL21

Students deserve it • The world demands it

West Virginia Department of EDUCATION

Jorea M. Marple, Ed.D.
State Superintendent of Schools