Findings from the 2012 West Virginia Online Writing Scoring Comparability Study

West Virginia Department of EDUCATION
Office of Research

Findings from the 2012 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, Suite 825 State Capitol Complex
1900 Kanawha Boulevard East
Charleston, WV 25305
http://wvde.state.wv.us/research

September 2013

Keywords
Automated Essay Scoring, Reliability, Consistency, Online Assessment, Writing

Suggested Citation

Hixson, N.H., & Rhudy, V. G. (2013). *Findings from the 2012 West Virginia online writing 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

Abstract

Student responses to the WESTEST 2 Online Writing Assessment are scored by a computer-scoring engine. The scoring method is not widely understood among educators, and there exists a misperception that it is not comparable to hand scoring. To address these issues, the West Virginia Department of Education (WVDE) conducts an annual scoring comparability study that compares scoring by trained human raters to scoring by the computer engine. This year, 45 educators from West Virginia participated in the study. Each scored a set of training essays and operational student essays that also were scored by the scoring engine. Each operational essay was scored independently by two human raters. Human raters' scores were compared to each other and to the engine.

Two research questions were posed: (RQ1) what is the level of calibration to the automated scoring engine that is achieved among human raters as a result of the training provided by the WVDE?, and (RQ2) what is the comparability of scores assigned by human rater pairs as well as between human-to-engine pairs?

Approximately 58% of human raters met three industry standard calibration criteria for calibration; the remaining 40% did not. Human rater pairs tended to provide the most consistent scores. However, in many cases we found that human raters were more likely to agree with the engine's scores than with each other's. When disagreements did occur though, human raters consistently scored student essays slightly higher than the engine. We believe this outcome should serve to mitigate some concerns that the engine scores student essays wildly differently from regular classroom educators or that the engine scores essays too forgivingly.

We do not draw definitive conclusions about the consistency of the engine from the results of this study because so few raters met rigorous standards for calibration. However, we note that the test vendor has provided considerable evidence to establish the comparability of the scoring process based upon studies that use only human raters judged to be experts based upon industry standard criteria.

We recommend continued use of the annual comparability study as a professional development experience for educators and additional data collection around educators' perception of the accuracy and fairness of scores assigned by the engine.

Contents

List of Figures

List of Tables

Introduction

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 from high school. Students must be taught to articulate their thoughts and ideas clearly and effectively. To measure this ability, the West Virginia (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. 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 on to a secure computer-based testing website. Each student receives a randomly assigned passage and prompt in one of the following four writing genres: narrative, descriptive, informative, or persuasive. (A student in Grade 3 receives 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 that has been trained to replicate expert human scoring using hand-scored student essays. 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 raters 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 the 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 essays. The WVDE Office of Research then compares the educators' 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 and comparability of the automated scoring engine. 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 raters are better able to reliably and accurately score student essays. The online writing comparability study seeks to address this issue.

The WVDE conducted its fourth WESTEST 2 Online Writing Comparability Study over a 2-day period in October 2012. Participants included 45 human raters selected to participate in the comparability study as described above. Following an explanation of comparability study and artificial intelligence scoring, table leaders led participants through a training process. Table leaders provided participants with copies of the appropriate grade-level West Virginia Writing Rubrics, copies of the secure 2012 operational passages and prompts, and anchor papers representing different score points for the five analytic writing traits for the various genres. Participants hand scored training sets of 14 randomly selected student responses representing the various genres and various levels of student ability. Scorers completed a worksheet containing four guiding questions while they scored training essays. These questions were designed to prompt reflection among participants and to improve the level of calibration to the engine that was achieved by each scorer. They included:

1. Which papers have you scored the same or very close to the same as the engine? (For example, identify papers where your score matched the engine score exactly for all five traits or matched exactly three or four of the five traits and were only one point off on other traits).
2. Which papers did you score where your scores for any given trait are 2 or more points from the engine score for the same trait?”
3. Are there any papers where you are consistently scoring higher or lower than the engine? If so, go back to the anchor papers and training papers to identify possible reasons why you are scoring lower or higher.
4. For any applicable papers, identify what your revised trait scores would be if you were allowed to score it again, but leave your original scores unaltered.

Scorers took notes for each question as they scored training essays. Additionally, table leaders led discussions of each student response, the human scores, and the computer scores as participants progressed through each of the essays included in the training sets.

Participant training and scoring focused entirely on one genre before moving onto a different genre. For example, a grade-level table might have focused first on narrative; therefore, the training and discussions were focused on the characteristics of narrative writing, and then participants scored student narrative essays. Each group then would move onto another genre. This method allowed participants to focus their attention on one genre at a time.

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

We posed two research questions (RQs) as part of this research study.

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

RQ2. What is the comparability of scores assigned by WV human rater pairs as well as between human-to-engine pairs?

Methodology

Participant Characteristics

The Office of Assessment and Accountability invited 46 educators to participate in the annual study. Five educators were assigned to each grade level, Grades 3 through 11—to participate in the annual study with the exception of Grade 4, where an additional educator was assigned to share table leader duties because one person had to leave early. One educator who indicated she would participate canceled, leaving a total of 45 participants. This left only four scorers at Grade 9.

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). We used the full population of 45 raters to address both RQ1 and RQ2.

Measures

Several measures of agreement were used in this study: (1) exact agreement, (2) exact-and-adjacent agreement, (3) standardized mean differences (SMD), (4) quadratic weighted kappa (QWK), and (5) correlation (Pearson's *r*). Each measure is described below.

Exact agreement

Exact agreement is defined as the circumstance when, examining the same essay, a score assigned by one rater is exactly the same as the corresponding score assigned by another rater. In this study, we calculated exact agreement for all five traits. We calculated agreement for the pair of human raters as well as for each of the two possible human-to-engine pairs.

The rate of exact agreement was defined as the percentage of instances of exact agreement across all essays in a given category. For example, in a sample of 150 Grade 3 essays, if we observed exact agreement between two humans in their mechanics scores for 75 essays, the exact agreement rate for mechanics would be 75/150 or 50%. Similarly, if we examined the same 150 essays but examined exact agreement between one human rater and the automated engine and observed 60 exact matches, our agreement rate would be 60/150 or 40%. In this example, the difference between human-to-human and human-to-engine exact agreement in mechanics for Grade 3 would be approximately 10% in favor of human-to-human agreement.

Exact and adjacent agreement

Exact and adjacent agreement was defined as the circumstance when, examining the same essay, a score assigned by one rater is exactly the same as the corresponding score assigned by another rater, or is equal to that score +/- one point. This is similar to applying a margin of error of 1 point. For example, exact-and-adjacent agreement would be met if rater A scored an essay's mechanics at 4 and rater B scored the same essay's mechanics at either a 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-and-adjacent agreement rates were operationalized as the percentage of instances of exact-and-adjacent agreement observed across all essays in a given category.

Standardized mean differences

In this study, the standardized mean difference (SMD) was defined as the difference between the mean scores assigned by two raters divided by the pooled standard deviation. Based upon the recommendations of Williamson, Xi, and Breyer (2012), we used the industry criteria of flagging any SMD value greater than or equal to .15 to identify meaningful differences among raters.

Quadratic weighted kappa

Quadratic weighted kappa is a measure of interrater agreement. It differs from Cohen's kappa in that it employs a weighting scenario to account for the ordinal (rank-ordered) nature of certain data sets. This measure is very appropriate to use in the case of trait score assignments on the writing rubric, which are ordinal. One notable advantage of this metric above agreement rates is that quadratic weighted kappa takes into account the possibility of two raters agreeing on a given score due to chance. Based upon the recommendations of Williamson, Xi, and Breyer (2012), we used the industry standard threshold of at least .70 for acceptable quadratic weighted kappa values among raters.

Correlation

Pearson's *r* is a measure of the direction and strength of the linear relationship between two variables. Values for *r* range from -1.0 to 1.0 with an absolute value of 1.0 signifying a perfect one-to-one relationship. The sign of the correlation indicates whether the relationship is positive or negative. Positive relationships are those where as one variable increases/decreases in value, the other does so in the same direction. A negative relationship is where as one value increases/decreases, the other changes in the opposite direction. In this study, we correlated trait scores from each pair of possible raters. Once again, we used the recommendations of Williamson, Xi, and Breyer (2012) to identify the industry standard criterion of at least .70 for an acceptable level of association among scores.

Research Design

A set of 14 training essays representing all genres was provided to each human rater. The training sets included three or four essays for each genre representing different score points. Because the scoring process focused on each genre separately, during the calibration process, each grade-level table first focused on the three or four training papers for the specific genre being scored. For example, if the participants at Grade 4 scored narrative essays first, then the calibration process focused only on those training papers from the narrative genre.

Each rater was then assigned a packet containing student essays randomly selected from the 2012 operational WESTEST 2 Online Writing Assessment. Student essays were divided into three packets per genre; two packets contained 32 student essays each, and a third packet contained 11 student essays, for a total of 75 essays per genre per grade level. The packets were distributed in a manner such that each essay would be scored independently by two human raters. Human scores were then compared to each other and to the automated engine scores using the agreement statistics described above.

Results

Research Question 1

RQ1 asked: “What is the level of calibration to the automated scoring engine that is achieved among human raters as a result of the training provided by the WVDE?” To address this question we examined the proportion of raters that met industry standard calibration criteria for three metrics: (a) standardized mean differences (SMD), (b) quadratic weighted kappa (QWK), and (c) Pearson’s *r*.

The data set used included the 45 human raters’ scores for all training essays scored and the accompanying trained automated engine scores for those essays, which were provided by the test vendor. The dataset contained a total of 594 records[1].

Table 1 provides the number and percentage of raters that met acceptable thresholds for each metric. One can conclude from these data that the best calibration was achieved with respect to the mechanics trait, where 62.2% of human raters met all three calibration criteria. Conversely, the least calibration was achieved in word choice/grammar usage, where fewer than half of all raters met all three calibration criteria (42.2%). Across all traits, the median percentage of raters that met all three criteria was approximately 58%. Con-

[1] Most raters completed all 14 of the provided training essays. However, some raters did not, resulting in the data set containing 36 fewer records than expected.

versely, nearly 40% of all human raters did not meet sufficient calibration based upon these metrics.

Table 1. Number and Percentage of Human Raters Meeting Industry Standard Calibration Criteria

	Criterion 1 (SMD <.15)		Criterion 2 (QWK ≥.70)		Criterion 3 (r = .70)		All criteria	
Trait	N	%	N	%	N	%	N	%
Organization	26	57.7	42	93.3	42	93.3	25	55.5
Development	28	62.2	42	93.3	44	97.7	27	60.0
Sentence structure	26	57.7	45	100.0	45	100.0	26	57.7
Word choice/grammar usage	20	44.4	42	93.3	44	97.7	19	42.2
Mechanics	30	66.6	42	93.3	44	97.7	28	62.2

We also provide, in Table 2, the percentage of human raters that met at least 50% exact agreement and 90% exact-and-adjacent agreement with the automated engine by trait. We must acknowledge that these thresholds were chosen somewhat arbitrarily to aid in describing the level of calibration achieved at the conclusion of the training. This is because industry experts have recommended against using simple agreement statistics as definitive measures of calibration because these metrics do not take into account several critical factors.

Table 2. Number and Percentage of Human Raters Meeting Agreement Thresholds

	≥50% Exact agreement		≥90% Exact/adj agreement	
Trait	N	%	N	%
Organization	26	57.8	39	86.7
Development	30	66.7	38	84.4
Sentence structure	26	57.8	39	86.7
Word choice/grammar usage	28	62.2	41	91.1
Mechanics	24	53.3	35	77.8

When examining agreement statistics we found that, depending on the trait under examination, approximately half to almost two thirds of all raters met the criteria of at least 50% exact agreement and approximately three fourths to 91% of all raters met the criteria of at least 90% exact-and-adjacent agreement.

Research Question 2

RQ2 asked: “What is the comparability of scores assigned by WV human rater pairs as well as between human-to-engine pairs?” The data set used included the 45 human raters’ scores for all operational student essays scored and the accompanying engine scores for those essays, which were provided by the test vendor. The dataset contained a total of 2,539 records[2].

We begin this section by first presenting agreement statistics for each grade level[3]. These data are presented first in order to provide a high level description of the level of agreement achieved among human-to-human and human-to-engine pairs during the comparability study. To aid in interpreting differences in agreement rates, we present odds ratios. We calculate these ratios in a manner such that they describe how much more or less likely the pair of human raters were to agree with each other than with the automated scoring engine. For example, an odds ratio of 1.0 indicates that the two human raters were no more likely to agree with each other than with the engine. This would be the ideal scenario because it would indicate no difference among the two methods. Conversely, if the odds ratio is below 1.0, it indicates that the two human raters were less likely to agree with each other than with the engine. This information, though not meeting the scientific rigor necessary to fully address RQ2, provides essential context.

Next we present the standardized mean difference, quadratic weighted kappa, and correlation statistics for each trait and grade level. We also provide a detailed analysis of standardized mean differences (SMDs) to better describe the amount of variation observed in human-to-human scores and human-to-engine scores. These data are used to draw conclusions regarding RQ2.

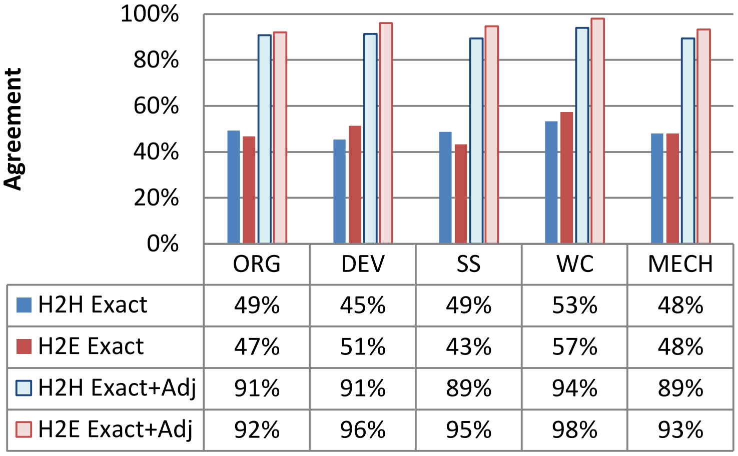
[2] Most raters completed all 14 of the provided training essays. However, some raters did not, resulting in the data set containing 36 fewer records than expected.

[3] For the sake of brevity, we include only the human-to-engine agreement rates for the first sample of human raters. We also compared the scores assigned by the second sample of human raters to the engine scores. The results of these analyses are provided in Appendix A.

Agreement statistics by grade

Grade 3

Figure 1 presents a graphical representation of the agreement rates for Grade 3. The Grade 3 odds ratios for exact and exact-and-adjacent agreement appear in Table 3. When examining exact agreement, the human raters were slightly more likely to agree with each other than with the automated engine for two of five traits (organization and sentence structure). With respect to sentence structure, the human raters were 13% more likely to agree with each other than with the engine. They were no more likely to agree with each other than with the engine for mechanics. For the remaining traits, the raters were less likely to agree with each other than with the automated engine. Of note, they were 12% more likely to agree with the engine than each other on development. With respect to exact-and-adjacent agreement, in all cases the human raters were slightly more likely to agree with the automated engine than with each other.



	ORG	DEV	SS	WC	MECH
🟦 H2H Exact	49%	45%	49%	53%	48%
🟥 H2E Exact	47%	51%	43%	57%	48%
□ H2H Exact+Adj	91%	91%	89%	94%	89%
□ H2E Exact+Adj	92%	96%	95%	98%	93%

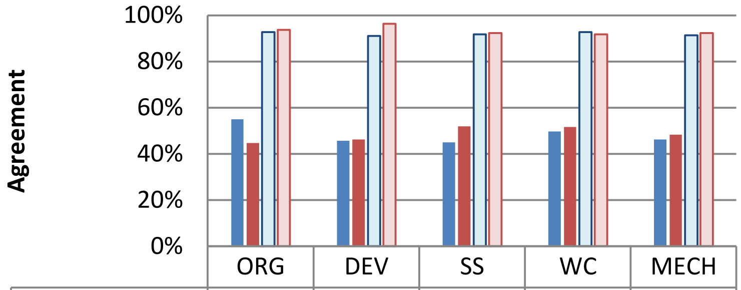
Figure 1. Grade 3 Comparability (H1 Sample)
Agreement rates for Grade 3 organization (ORG), development (DEV), sentence structure (SS), word choice/grammar usage (WC), and mechanics (MECH) scores achieved among human-to-human (H2H) and human-to-engine (H2E) pairs for both exact and exact-and-adjacent agreement.

Table 3. Odds Ratios for Grade 3

	Odds ratio	
Trait	Exact	Exact/adjacent
Organization	1.04	.98
Development	.88	.94
Sentence structure	1.13	.93
Word choice/grammar usage	.92	.95
Mechanics	1.00	.95

Grade 4

Figure 2 presents a graphical representation of the agreement rates for Grade 4. The Grade 4 odds ratios for exact and exact-and-adjacent agreement appear in Table 4. When examining exact agreement, the human raters were slightly less likely to agree with each other than with the automated engine for two of five traits (word choice/grammar usage and mechanics). Specifically, they were 14% less likely to agree with each other than with the engine on sentence structure. However, they were 22% more likely to agree with each other than with the engine with respect to organization and no more likely to agree with each other than with the engine for development. With respect to exact-and-adjacent agreement, in two cases (sentence structure and word choice/grammar usage) the human raters were essentially no more likely to agree with each other than with the automated engine. For the remaining three traits, the human raters were slightly more likely to agree with the engine than with each other.



	ORG	DEV	SS	WC	MECH
■ H2H Exact	55%	46%	45%	50%	46%
■ H2E Exact	45%	46%	52%	52%	48%
□ H2H Exact+Adj	93%	91%	92%	93%	91%
□ H2E Exact+Adj	94%	96%	92%	92%	92%

Figure 2. Grade 4 Comparability (H1 Sample)

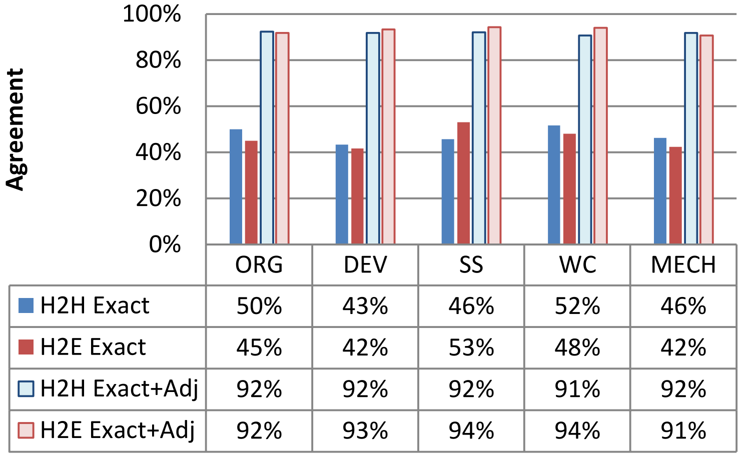
Agreement rates for Grade 4 organization (ORG), development (DEV), sentence structure (SS), word choice/grammar usage (WC), and mechanics (MECH) scores achieved among human-to-human (H2H) and human-to-engine (H2E) pairs for both exact and exact-and-adjacent agreement.

Table 4. Odds Ratios for Grade 4

	Odds ratio	
Trait	Exact	Exact/adjacent
Organization	1.22	.98
Development	1.00	.94
Sentence structure	.86	1.00
Word choice/grammar usage	.96	1.01
Mechanics	.95	.98

Grade 5

Figure 3 presents a graphical representation of the agreement rates for Grade 5. The Grade 5 odds ratios for exact and exact-and-adjacent agreement appear in Table 5. When examining exact agreement, the human raters were slightly less likely to agree with each other than with the automated engine for three of five traits (sentence structure, word choice, and mechanics). Of note, human raters were 11% more likely to agree with each other than with the engine with respect to organization; they were no more likely to agree with each other than with the engine for development; and they were 14% more likely to agree with the engine than with each other with respect to sentence structure. With respect to exact-and-adjacent agreement, there was either no appreciable difference among methods (i.e., sentence structure and word choice/grammar usage) or human raters were slightly more likely to agree with the engine than with each other.



	ORG	DEV	SS	WC	MECH
■ H2H Exact	50%	43%	46%	52%	46%
■ H2E Exact	45%	42%	53%	48%	42%
□ H2H Exact+Adj	92%	92%	92%	91%	92%
□ H2E Exact+Adj	92%	93%	94%	94%	91%

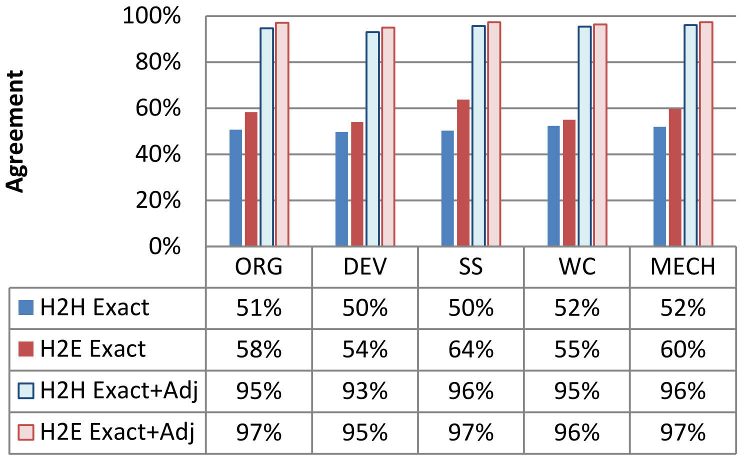
Figure 3. Grade 5 Comparability (H1 Sample)
Agreement rates for Grade 5 organization (ORG), development (DEV), sentence structure (SS), word choice/grammar usage (WC), and mechanics (MECH) scores achieved among human-to-human (H2H) and human-to-engine (H2E) pairs for both exact and exact-and-adjacent agreement.

Table 5. Odds Ratios for Grade 5

	Odds ratio	
Trait	Exact	Exact/adjacent
Organization	1.11	1.00
Development	1.02	.98
Sentence structure	.86	.97
Word choice/grammar usage	1.08	.96
Mechanics	1.09	1.01

Grade 6

Figure 4 presents a graphical representation of the agreement rates for Grade 6. The Grade 6 odds ratios for exact and exact-and-adjacent agreement appear in Table 6. When examining exact agreement, the human raters were slightly less likely to agree with each other than with the automated engine for all five traits. Notably, they were 22%, 13%, and 14% less likely to agree with each other than with the engine with respect to sentence structure, organization, and mechanics, respectively. With respect to exact-and-adjacent agreement, in all cases the human raters were slightly less likely to agree with each other than with the automated engine.



	ORG	DEV	SS	WC	MECH
■ H2H Exact	51%	50%	50%	52%	52%
■ H2E Exact	58%	54%	64%	55%	60%
□ H2H Exact+Adj	95%	93%	96%	95%	96%
□ H2E Exact+Adj	97%	95%	97%	96%	97%

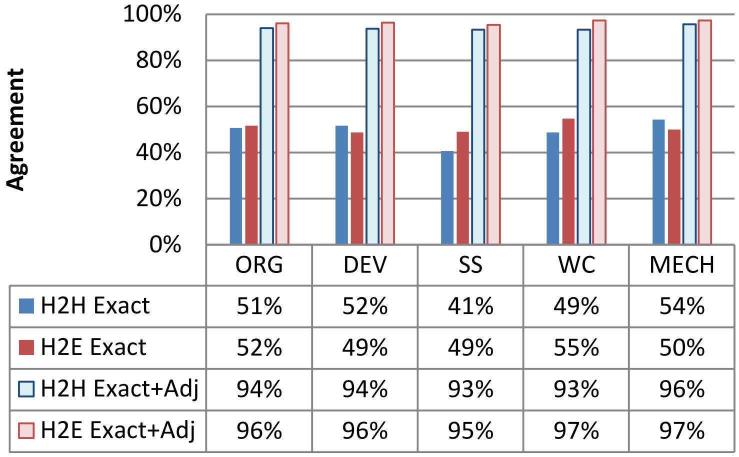
Figure 4. Grade 6 Comparability (H1 Sample)
Agreement rates for Grade 6 organization (ORG), development (DEV), sentence structure (SS), word choice/grammar usage (WC), and mechanics (MECH) scores achieved among human-to-human (H2H) and human-to-engine (H2E) pairs for both exact and exact-and-adjacent agreement.

Table 6. Odds Ratios for Grade 6

	Odds ratio	
Trait	Exact	Exact/adjacent
Organization	.87	.97
Development	.92	.97
Sentence structure	.78	.98
Word choice/grammar usage	.94	.98
Mechanics	.86	.98

Grade 7

Figure 5 presents a graphical representation of the agreement rates for Grade 7. The Grade 7 odds ratios for exact and exact-and-adjacent agreement appear in Table 7. When examining exact agreement, the human raters were slightly less likely to agree with each other than with the automated engine for three of five traits. Notably, they were 17% and 11% less likely to agree with each other than with the engine with respect to sentence structure and word choice/grammar usage, respectively. They were slightly more likely to agree with each other than with the engine with respect to development and mechanics. With respect to exact-and-adjacent agreement, in all cases the human raters were slightly less likely to agree with each other than with the automated engine.



	ORG	DEV	SS	WC	MECH
■ H2H Exact	51%	52%	41%	49%	54%
■ H2E Exact	52%	49%	49%	55%	50%
□ H2H Exact+Adj	94%	94%	93%	93%	96%
□ H2E Exact+Adj	96%	96%	95%	97%	97%

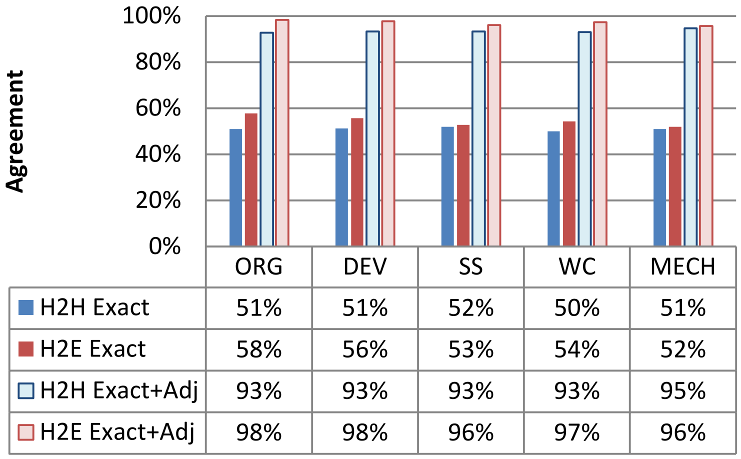
Figure 5. Grade 7 Comparability (H1 Sample)
Agreement rates for Grade 7 organization (ORG), development (DEV), sentence structure (SS), word choice/grammar usage (WC), and mechanics (MECH) scores achieved among human-to-human (H2H) and human-to-engine (H2E) pairs for both exact and exact-and-adjacent agreement.

Table 7. Odds Ratios for Grade 7

	Odds ratio	
Trait	Exact	Exact/adjacent
Organization	.98	.97
Development	1.06	.97
Sentence structure	.83	.97
Word choice/grammar usage	.89	.95
Mechanics	1.08	.98

Grade 8

Figure 6 presents a graphical representation of the agreement rates for Grade 8. The Grade 8 odds ratios for exact and exact-and-adjacent agreement appear in Table 8. When examining exact agreement, the human raters were slightly less likely to agree with each other than with the automated engine for all five traits. Notably, they were 13% less likely to agree with each other than with the engine with respect to organization. With respect to exact-and-adjacent agreement, in all cases the human raters were slightly less likely to agree with each other than with the automated engine.



	ORG	DEV	SS	WC	MECH
■ H2H Exact	51%	51%	52%	50%	51%
■ H2E Exact	58%	56%	53%	54%	52%
□ H2H Exact+Adj	93%	93%	93%	93%	95%
□ H2E Exact+Adj	98%	98%	96%	97%	96%

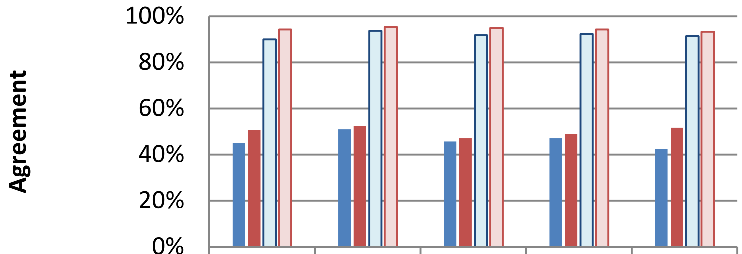
Figure 6. Grade 8 Comparability (H1 Sample)
Agreement rates for Grade 8 organization (ORG), development (DEV), sentence structure (SS), word choice/grammar usage (WC), and mechanics (MECH) scores achieved among human-to-human (H2H) and human-to-engine (H2E) pairs for both exact and exact-and-adjacent agreement.

Table 8. Odds Ratios for Grade 8

	Odds ratio	
Trait	Exact	Exact/adjacent
Organization	.87	.94
Development	.91	.94
Sentence structure	.98	.96
Word choice/grammar usage	.92	.95
Mechanics	.98	.98

Grade 9

Figure 7 presents a graphical representation of the agreement rates for Grade 9. The Grade 9 odds ratios for exact and exact-and-adjacent agreement appear in Table 9. When examining exact agreement, the human raters were slightly less likely to agree with each other than with the automated engine for all five traits. Notably, they were 20% and 12% less likely to agree with each other than with the engine with respect to mechanics and organization, respectively. With respect to exact-and-adjacent agreement, in all cases the human raters were slightly less likely to agree with each other than with the automated engine.



	ORG	DEV	SS	WC	MECH
🟦 H2H Exact	45%	51%	46%	47%	42%
🟥 H2E Exact	51%	52%	47%	49%	52%
□ H2H Exact+Adj	90%	94%	92%	92%	91%
□ H2E Exact+Adj	94%	95%	95%	94%	93%

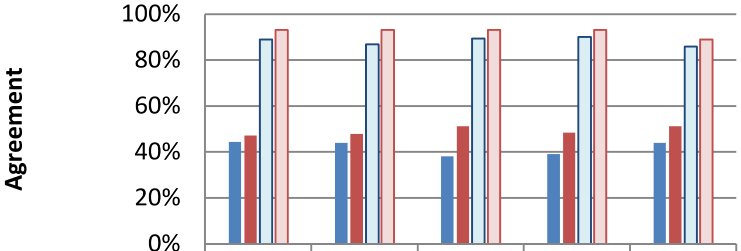
Figure 7. Grade 9 Comparability (H1 Sample)
Agreement rates for Grade 9 organization (ORG), development (DEV), sentence structure (SS), word choice/grammar usage (WC), and mechanics (MECH) scores achieved among human-to-human (H2H) and human-to-engine (H2E) pairs for both exact and exact-and-adjacent agreement.

Table 9. Odds Ratios for Grade 9

	Odds ratio	
Trait	Exact	Exact/adjacent
Organization	.88	.95
Development	.98	.98
Sentence structure	.97	.96
Word choice/grammar usage	.95	.97
Mechanics	.80	.97

Grade 10

Figure 8 presents a graphical representation of the agreement rates for Grade 10. The Grade 10 odds ratios for exact and exact-and-adjacent agreement appear in Table 10. When examining exact agreement, the human raters were slightly less likely to agree with each other than with the automated engine for all five traits. Notably, they were 26%, 19%, and 14% less likely to agree with each other than with the engine with respect to sentence structure, word choice/grammar usage, and mechanics, respectively. With respect to exact-and-adjacent agreement, in all cases the human raters were slightly less likely to agree with each other than with the automated engine.



	ORG	DEV	SS	WC	MECH
■ H2H Exact	44%	44%	38%	39%	44%
■ H2E Exact	47%	48%	51%	48%	51%
□ H2H Exact+Adj	89%	87%	89%	90%	86%
□ H2E Exact+Adj	93%	93%	93%	93%	89%

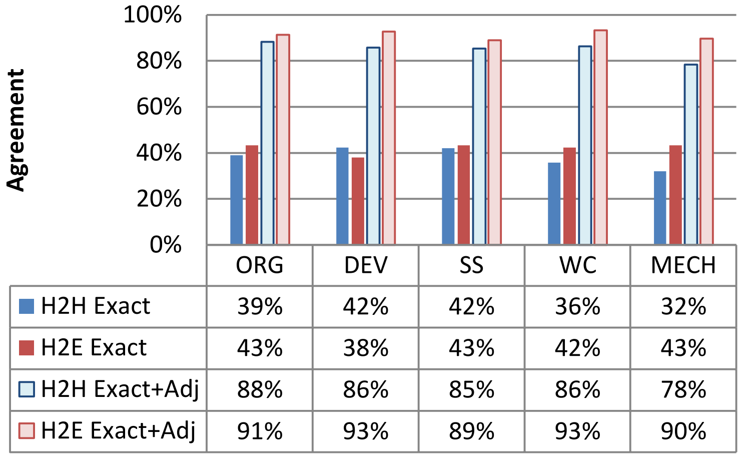
Figure 8. Grade 10 Comparability (H1 Sample)
Agreement rates for Grade 10 organization (ORG), development (DEV), sentence structure (SS), word choice/grammar usage (WC), and mechanics (MECH) scores achieved among human-to-human (H2H) and human-to-engine (H2E) pairs for both exact and exact-and-adjacent agreement.

Table 10. Odds Ratios for Grade 10

	Odds ratio	
Trait	Exact	Exact/adjacent
Organization	.93	.95
Development	.91	.93
Sentence structure	.74	.95
Word choice/grammar usage	.81	.96
Mechanics	.86	.96

Grade 11

Figure 9 presents a graphical representation of the agreement rates for Grade 11. The Grade 11 odds ratios for exact and exact-and-adjacent agreement appear in Table 11. When examining exact agreement, the human raters were slightly less likely to agree with each other than with the automated engine for four of five traits. Notably, they were 26%, 15%, and 10% less likely to agree with each other than with the engine with respect to mechanics, word choice/grammar usage, and organization, respectively. With respect to exact-and-adjacent agreement, in all cases the human raters were slightly less likely to agree with each other than with the automated engine.



	ORG	DEV	SS	WC	MECH
🟦 H2H Exact	39%	42%	42%	36%	32%
🟥 H2E Exact	43%	38%	43%	42%	43%
□ H2H Exact+Adj	88%	86%	85%	86%	78%
□ H2E Exact+Adj	91%	93%	89%	93%	90%

Figure 9. Grade 11 Comparability (H1 Sample)
Agreement rates for Grade 11 organization (ORG), development (DEV), sentence structure (SS), word choice/grammar usage (WC), and mechanics (MECH) scores achieved among human-to-human (H2H) and human-to-engine (H2E) pairs for both exact and exact-and-adjacent agreement.

Table 11. Odds Ratios for Grade 11

	Odds ratio	
Trait	Exact	Exact/adjacent
Organization	.90	.96
Development	1.10	.92
Sentence structure	.97	.95
Word choice/grammar usage	.85	.92
Mechanics	.74	.86

Industry standard agreement criteria by rater pair

Here we present the standardized mean difference, quadratic weighted kappa, and correlation statistics for each trait and grade level. In each section we provide the criteria for each pair of raters. We then provide conclusions about the comparability of scoring methods based upon the extent to which the criteria were met across scoring pairs.

Organization

For no grade level did all scoring pairs meet all three agreement criteria. The most consistent method for scoring organization was human-to-human scoring, which met all three criteria in Grades 3 – 7 and came very close to meeting all criteria in Grades 8 and 9 as well. Notably, human-to-engine scoring met all three criteria in Grade 8 whereas human-to-human scoring did not.

Table 12. Agreement Criteria for Organization Trait by Grade

	SMD			QWK			*r*			All criteria		
Grade	H2H	H12E	H22E	H2H	H12E	H22E	H2H	H12E	H22E	H2H	H12E	H22E
3	0.08	0.13	0.04	0.74	0.64	0.70	0.74	0.65	0.72	YES	NO	YES
4	0.05	0.06	0.01	0.71	0.66	0.64	0.71	0.67	0.65	YES	NO	NO
5	0.01	0.14	0.15	0.74	0.67	0.64	0.74	0.68	0.65	YES	NO	NO
6	0.01	0.19	0.20	0.73	0.76	0.70	0.73	0.79	0.73	YES	NO	NO
7	0.06	0.06	0.13	0.71	0.68	0.68	0.71	0.69	0.70	YES	NO	NO
8	0.01	0.12	0.13	0.69	0.73	0.71	0.69	0.74	0.72	NO	YES	YES
9	0.01	0.19	0.19	0.69	0.72	0.74	0.69	0.74	0.76	NO	NO	NO
10	0.03	0.35	0.37	0.65	0.66	0.60	0.65	0.71	0.65	NO	NO	NO
11	0.08	0.25	0.34	0.67	0.67	0.67	0.68	0.70	0.71	NO	NO	NO

Note: SMD = standardized mean difference; QWK = quadratic weighted kappa; *r* = Pearson product-moment correlation statistic; red = a value that fell below the agreement criteria.

Development

For Grades 7 and 8, all scoring pairs met all three agreement criteria. Across grades, the most consistent method for scoring development was human-to-human scoring, which met all three criteria in Grades 3, 5-9, and 11 and came very close to meeting the criteria in Grades 8 and 9 as well.

Table 13. Agreement Criteria for Development Trait by Grade

	SMD			QWK			*r*			All criteria		
Grade	H2H	H12E	H22E	H2H	H12E	H22E	H2H	H12E	H22E	H2H	H12E	H22E
3	0.05	0.11	0.05	0.72	0.69	0.73	0.72	0.70	0.75	YES	NO	YES
4	0.07	0.04	0.12	0.66	0.67	0.63	0.67	0.69	0.64	NO	NO	NO
5	0.01	0.01	0.00	0.70	0.69	0.64	0.70	0.69	0.64	YES	NO	NO
6	0.03	0.22	0.24	0.73	0.73	0.73	0.73	0.76	0.76	YES	NO	NO
7	0.03	0.02	0.05	0.71	0.70	0.72	0.71	0.71	0.73	YES	YES	YES
8	0.03	0.11	0.14	0.72	0.74	0.72	0.72	0.74	0.73	YES	YES	YES
9	0.02	0.22	0.24	0.79	0.78	0.76	0.79	0.80	0.79	YES	NO	NO
10	0.03	0.24	0.26	0.67	0.71	0.67	0.67	0.74	0.70	NO	NO	NO
11	0.05	0.16	0.23	0.70	0.72	0.71	0.71	0.74	0.73	YES	NO	NO

Note: SMD = standardized mean difference; QWK = quadratic weighted kappa; *r* = Pearson product-moment correlation statistic; red = a value that fell below the agreement criteria.

Sentence Structure

For Grade 9 all scoring pairs met all three agreement criteria. The most consistent method for scoring sentence structure was human-to-human scoring which met all three criteria in Grades 3, 6, and 8-9. Notably, human-to-engine scoring met all three criteria in Grade 8 whereas human-to-human scoring did not. Also notably, human-to-engine scoring met all three criteria in three grades as opposed to four for human-to-human scoring.

Table 14. Agreement Criteria for Sentence Structure Trait by Grade

	SMD			QWK			*r*			All criteria		
Grade	H2H	H12E	H22E	H2H	H12E	H22E	H2H	H12E	H22E	H2H	H12E	H22E
3	0.08	0.16	0.08	0.73	0.68	0.70	0.74	0.70	0.71	YES	NO	YES
4	0.03	0.07	0.04	0.64	0.62	0.62	0.64	0.63	0.62	NO	NO	NO
5	0.05	0.06	0.12	0.69	0.71	0.62	0.69	0.71	0.61	NO	YES	NO
6	0.11	0.04	0.16	0.74	0.79	0.70	0.74	0.80	0.71	YES	YES	NO
7	0.05	0.01	0.04	0.66	0.68	0.68	0.66	0.69	0.69	NO	NO	NO
8	0.02	0.09	0.07	0.71	0.67	0.70	0.72	0.67	0.71	YES	NO	YES
9	0.07	0.00	0.07	0.72	0.71	0.75	0.72	0.71	0.75	YES	YES	YES
10	0.03	0.14	0.18	0.62	0.67	0.67	0.62	0.68	0.68	NO	NO	NO
11	0.04	0.07	0.03	0.65	0.64	0.64	0.65	0.65	0.65	NO	NO	NO

Note: SMD = standardized mean difference; QWK = quadratic weighted kappa; *r* = Pearson product-moment correlation statistic; red = a value that fell below the agreement criteria.

Word Choice/Grammar Usage

For Grade 3 all scoring pairs met all three agreement criteria. The most consistent method for scoring word choice/grammar usage was human-to-human scoring which met all three criteria in Grades 3, 5-6, and 9. Notably, human-to-engine scoring met all three criteria in Grade 8 whereas human-to-human scoring did not.

Table 15. Agreement Criteria for Word Choice/Grammar Usage Trait by Grade

	SMD			QWK			*r*			All criteria		
Grade	H2H	H12E	H22E	H2H	H12E	H22E	H2H	H12E	H22E	H2H	H12E	H22E
3	0.05	0.09	0.04	0.74	0.75	0.78	0.74	0.76	0.78	YES	YES	YES
4	0.07	0.10	0.03	0.66	0.61	0.63	0.66	0.61	0.64	NO	NO	NO
5	0.06	0.17	0.10	0.70	0.70	0.68	0.70	0.71	0.69	YES	NO	NO
6	0.09	0.12	0.22	0.74	0.72	0.69	0.74	0.73	0.71	YES	YES	NO
7	0.07	0.02	0.09	0.63	0.68	0.69	0.64	0.68	0.70	NO	NO	NO
8	0.05	0.04	0.09	0.69	0.69	0.74	0.70	0.69	0.75	NO	NO	YES
9	0.06	0.23	0.28	0.70	0.71	0.72	0.71	0.73	0.75	YES	NO	NO
10	0.02	0.18	0.21	0.63	0.68	0.65	0.64	0.70	0.67	NO	NO	NO
11	0.03	0.13	0.16	0.62	0.68	0.65	0.62	0.70	0.67	NO	NO	NO

Note: SMD = standardized mean difference; QWK = quadratic weighted kappa; *r* = Pearson product-moment correlation statistic; red = a value that fell below the agreement criteria.

Mechanics

For Grade 6 all scoring pairs met all three agreement criteria. The most consistent method for scoring mechanics was human-to-human scoring which met all three criteria in Grades 6-7.

Table 16. Agreement Criteria for Mechanics Trait by Grade

	SMD			QWK			*r*			All criteria		
Grade	H2H	H12E	H22E	H2H	H12E	H22E	H2H	H12E	H22E	H2H	H12E	H22E
3	0.05	0.21	0.15	0.69	0.62	0.72	0.69	0.66	0.75	NO	NO	NO
4	0.03	0.17	0.14	0.66	0.61	0.59	0.66	0.63	0.61	NO	NO	NO
5	0.05	0.13	0.08	0.69	0.62	0.67	0.69	0.63	0.67	NO	NO	NO
6	0.03	0.04	0.07	0.74	0.77	0.72	0.74	0.78	0.73	YES	YES	YES
7	0.02	0.00	0.02	0.71	0.68	0.71	0.71	0.68	0.71	YES	NO	YES
8	0.09	0.02	0.11	0.71	0.67	0.69	0.72	0.67	0.69	YES	NO	NO
9	0.04	0.12	0.15	0.68	0.68	0.76	0.69	0.68	0.77	NO	NO	NO
10	0.02	0.14	0.15	0.65	0.64	0.62	0.65	0.67	0.65	NO	NO	NO
11	0.03	0.08	0.11	0.50	0.63	0.55	0.50	0.64	0.56	NO	NO	NO

Note: SMD = standardized mean difference; QWK = quadratic weighted kappa; *r* = Pearson product-moment correlation statistic; red = a value that fell below the agreement criteria.

Standardized mean differences by rater pair

Below we provide a more nuanced examination of the standardized mean differences (SMD) observed in this study. This examination allows us to quantify the actual difference among scores assigned among human rater pairs and among human and engine pairs.

Organization

Table 17 provides the SMD for each grade for the trait of organization and for all three scoring pairs. Notably, in no case did the human-to-human pairs exceed the threshold of .15 or greater. This indicates that human rater pairs were very consistent in their assignment of organization scores as measured by SMDs.

Table 17. Standardized Mean Difference for Organization Trait by Scoring Pair

	Human to human		Human 1 to engine		Human 2 to engine	
Grade	N	SMD	N	SMD	N	SMD
3	143	.08	149	.13	144	.04
4	297	.05	300	.06	297	.01
5	299	.01	300	.14	299	.15
6	299	.01	300	.19	299	.20
7	299	.06	300	.06	299	.13
8	290	.01	300	.12	290	.13
9	289	.01	297	.19	292	.19
10	284	.03	288	.35	285	.37
11	294	.08	300	.25	294	.34

Note: SMD = standardized mean differences; red = a value that fell outside the range of acceptable tolerance

For the human 1-to-engine and human 2-to-engine pairs, there were several grade levels where the SMD value was outside the range of acceptable tolerance. In all cases, the engine assigned lower scores than the human raters. The average difference for these cases was approximately -.27 points on a 6-point scale. In other words, on average, the engine scored organization slightly more than one quarter point lower than the human raters that participated in this study across the flagged grade levels.

Development

Table 18 provides the SMD for each grade for the trait of development and for all three scoring pairs. Notably, in no case did the human-to-human pairs exceed the threshold of .15 or greater. This indicates that human rater pairs were very consistent in their assignment of development scores as measured by SMDs.

Table 18. Standardized Mean Difference for Development Trait by Scoring Pair

	Human to human		Human 1 to engine		Human 2 to engine	
Grade	N	SMD	N	SMD	N	SMD
3	144	0.05	150	0.11	144	0.05
4	297	0.07	300	0.04	297	0.12
5	299	0.01	300	0.01	299	0.00
6	299	0.03	300	0.22	299	0.24
7	299	0.03	300	0.02	299	0.05
8	291	0.03	300	0.11	291	0.14
9	289	0.02	297	0.22	292	0.24
10	284	0.03	288	0.24	285	0.26
11	294	0.05	300	0.16	294	0.23

Note: SMD = standardized mean differences; red = a value that fell outside the range of acceptable tolerance

For the human 1-to-engine and human 2-to-engine pairs, there were several grade levels where the SMD value was outside the range of acceptable tolerance. In all cases, the engine assigned lower scores than the human raters. The average difference for these cases was approximately -.25 points on a 6-point scale. In other words, on average, the engine scored development one quarter point lower than the human raters that participated in this study across the flagged grade levels.

Sentence structure

Table 19 provides the SMD for each grade for the trait of sentence structure and for all three scoring pairs. Notably, in no case did the human-to-human pairs exceed the threshold of .15 or greater. This indicates that human rater pairs were very consistent in their assignment of sentence structure scores as measured by SMDs.

Table 19. Standardized Mean Difference for Sentence Structure Trait by Scoring Pair

	Human to human		Human 1 to engine		Human 2 to engine	
Grade	N	SMD	N	SMD	N	SMD
3	144	0.08	150	0.16	144	0.08
4	297	0.03	300	0.07	297	0.04
5	299	0.05	300	0.06	299	0.12
6	299	0.11	300	0.04	299	0.16
7	298	0.05	299	0.01	299	0.04
8	291	0.02	300	0.09	291	0.07
9	289	0.07	297	0.00	292	0.07
10	284	0.03	288	0.14	285	0.18
11	294	0.04	300	0.07	294	0.03

Note: SMD = standardized mean differences; red = a value that fell outside the range of acceptable tolerance

For the human 1-to-engine and human 2-to-engine pairs, there were several grade levels where the SMD value was outside the range of acceptable tolerance. In all but one case (i.e., Grade 3 Human 1 to Engine), the engine assigned lower scores than the human raters. The average difference in cases where the engine scored lower than human raters was approximately -.17 points on a

6-point scale. Conversely, for Grade 3 human 1-to-engine agreement, the engine scored sentence structure .17 points higher than human raters.

Word Choice/Grammar Usage

Table 20 provides the SMD for each grade for the trait of word choice/ grammar usage and for all three scoring pairs. Notably, in no case did the human-to-human pairs exceed the threshold of .15 or greater. This indicates that human rater pairs were very consistent in their assignment of word choice/grammar usage scores as measured by SMDs.

Table 20. Standardized Mean Difference for Word Choice/Grammar Usage Trait by Scoring Pair

	Human to human		Human 1 to engine		Human 2 to engine	
Grade	N	SMD	N	SMD	N	SMD
3	144	0.05	150	0.09	144	0.04
4	297	0.07	300	0.10	297	0.03
5	300	0.06	300	0.17	300	0.10
6	299	0.09	300	0.12	299	0.22
7	299	0.07	300	0.02	299	0.09
8	291	0.05	300	0.04	291	0.09
9	289	0.06	297	0.23	292	0.28
10	284	0.02	288	0.18	285	0.21
11	294	0.03	300	0.13	294	0.16

Note: SMD = standardized mean differences; red = a value that fell outside the range of acceptable tolerance

For the human 1-to-engine and human 2-to-engine pairs, there were several grade levels where the SMD value was outside the range of acceptable tolerance. In all cases, the engine assigned lower scores than the human raters. The average difference for these cases was approximately -.22 points on a 6-point scale. In other words, on average, the engine scored word choice/grammar usage slightly less than one quarter point lower than the human raters that participated in this study across the flagged grade levels.

Mechanics

Table 21 provides the SMD for each grade for the trait of mechanics and for all three scoring pairs. Notably, in no case did the human-to-human pairs exceed the threshold of .15 or greater. This indicates that human rater pairs were very consistent in their assignment of development ratings as measured by SMDs.

Table 21. Standardized Mean Difference for Mechanics Trait by Scoring Pair

	Human to human		Human 1 to engine		Human 2 to engine	
Grade	N	SMD	N	SMD	N	SMD
3	144	0.05	150	0.21	144	0.15
4	297	0.03	300	0.17	297	0.14
5	299	0.05	300	0.13	299	0.08
6	299	0.03	300	0.04	299	0.07
7	299	0.02	300	0.00	299	0.02
8	291	0.09	300	0.02	291	0.11
9	289	0.04	297	0.12	292	0.15
10	284	0.02	288	0.14	285	0.15
11	294	0.03	300	0.08	294	0.11

Note: SMD = standardized mean differences; red = a value that fell outside the range of acceptable tolerance

For the human 1-to-engine and human 2-to-engine pairs, there were several grade levels where the SMD value was outside the range of acceptable tolerance. In all cases, the engine assigned lower scores than the human raters. The average difference for these cases was approximately -.17 points on a 6-point scale.

Discussion

The quality of the calibration process has a direct impact upon the validity of the comparability analyses conducted in this study. Only approximately 58% of raters met all three industry standard calibration criteria across traits. The remaining 40% did not. As has been suggested in our previous comparability studies, the inability to meet a highly rigorous standard for calibration is likely due to the limited amount of time and resources available for rater calibration. It would require substantially more time with raters to get them to the level of calibration necessary to make strong claims about comparability. Taking this into account, the fact that we observed generally better consistency among human rater pairs than human-to-engine pairs cannot be assumed to be a totally fair judgment of the engine's comparability to human scoring. Results could have been very different if all scorers met standards for calibration.

To remedy this situation for the annual comparability study would require considerable additional cost. This cost would likely be unjustifiable given that the WESTEST 2 technical report provides objective and rigorous comparability data comparing expert human raters to the automated scoring engine using the 5% read-behind methodology discussed in the introduction of this report. It is worth noting that the annual technical report has repeatedly established the comparability of the engine to expert human raters.

It is our opinion that the purpose of the annual online writing comparability study is most appropriately focused on the professional development experience that it provides for WV educators. This study affords them a unique opportunity to become familiar with the WV writing rubric and to gain a general understanding of how the automated essay scoring engine works. This being said, it is certainly appropriate to continue to provide educators with an overview of the type of agreement achieved via these studies to serve as feedback about the extent to which they were able to reach agreement with the highly trained scoring engine. This year, it is worth noting that in many cases we found human raters were slightly more likely to agree with the automated engine than with each other. When disagreements between human raters and the scoring engine did occur, the differences were often human raters scoring student essays higher than the automated engine by between -.17 and -.27 points on average. This quantity represents approximately 3% to 4.5% of the available 6-point scale. While any difference may appear to be problematic, one must keep in mind the limitations noted above. Further, this finding may actually serve to mitigate some concerns that the automated engine is scoring student essays wildly different from regular classroom educators who have a general understanding of the writing rubric or that the engine scores essays too forgivingly.

One additional limitation that should be noted is that the industry standard criteria used in this study are recommended to be applied to the task level. In the case of West Virginia's online writing assessment, the task level would be considered the prompt type. We did not have sufficient sample sizes to accomplish this given cost limitations. So, we instead aggregated our data to the grade level. Aggregating across genres could have introduced some noise that influenced our ability to accurately gauge consistency.

Recommendations

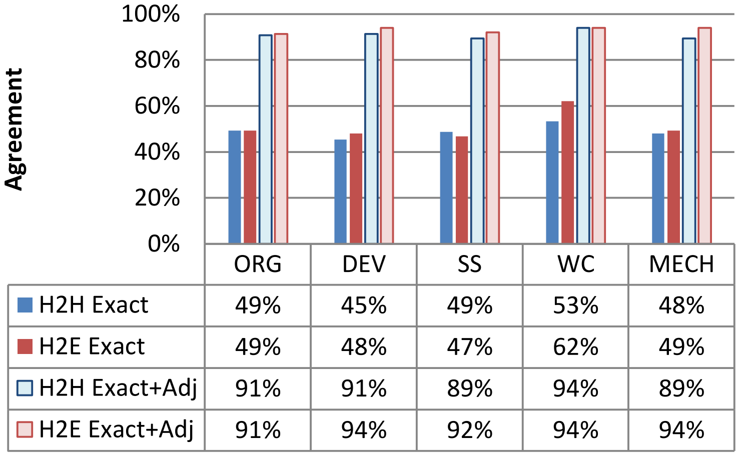
Continue to emphasize the chief purpose of the annual comparability study—to serve as a professional development experience for educators to build their understanding of the automated scoring engine. Build collective understanding in the field that the comparability of the engine to human scoring is well established and that comprehensive evidence is presented each year in the WESTEST 2 technical report. As was recommended last year, consider collecting additional evaluative feedback from study participants about their experiences in participating in the study. Focus this inquiry upon their perceptions about the comparability, rigor, and fairness of the scoring engine and the extent to which they feel the comparability study has improved their ability to recognize and the features of high quality student compositions. Formally collecting such accounts and providing them to the field as part of this annual report could help to dispel unfounded negative sentiments about automated scoring.

References

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

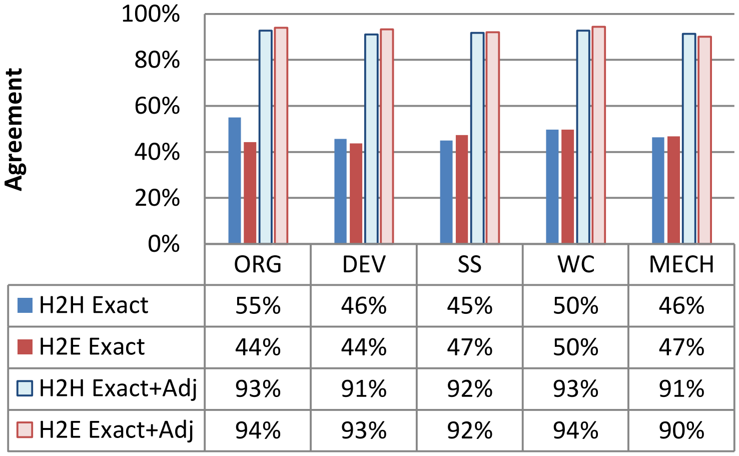
Appendix A. Agreement Statistics for H2 Sample

The following figures provide graphic representations of agreement rates for organization (ORG), development (DEV), sentence structure (SS), word choice/grammar usage (WC), and mechanics (MECH) scores assigned by the second sample of human raters, comparing human-to-human (H2H) and human-to-engine (H2E) pairs for both exact and exact-and-adjacent agreement.



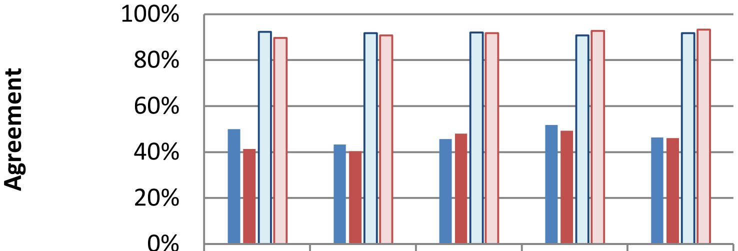
	ORG	DEV	SS	WC	MECH
■ H2H Exact	49%	45%	49%	53%	48%
■ H2E Exact	49%	48%	47%	62%	49%
□ H2H Exact+Adj	91%	91%	89%	94%	89%
□ H2E Exact+Adj	91%	94%	92%	94%	94%

Figure 10. Grade 3 Comparability (H2 Sample)



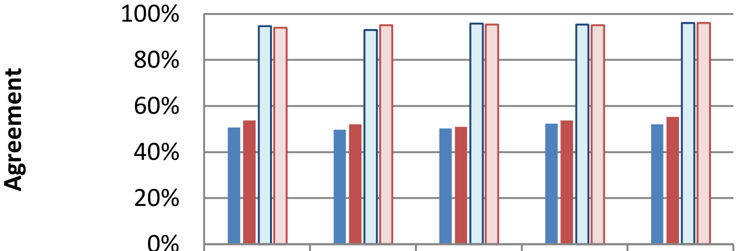
	ORG	DEV	SS	WC	MECH
■ H2H Exact	55%	46%	45%	50%	46%
■ H2E Exact	44%	44%	47%	50%	47%
□ H2H Exact+Adj	93%	91%	92%	93%	91%
□ H2E Exact+Adj	94%	93%	92%	94%	90%

Figure 11. Grade 4 Comparability (H2 Sample)



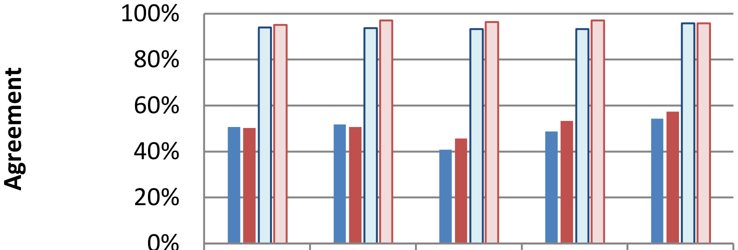
	ORG	DEV	SS	WC	MECH
🟦 H2H Exact	50%	43%	46%	52%	46%
🟥 H2E Exact	41%	40%	48%	49%	46%
□ H2H Exact+Adj	92%	92%	92%	91%	92%
□ H2E Exact+Adj	90%	91%	92%	93%	93%

Figure 12. Grade 5 Comparability (H2 Sample)



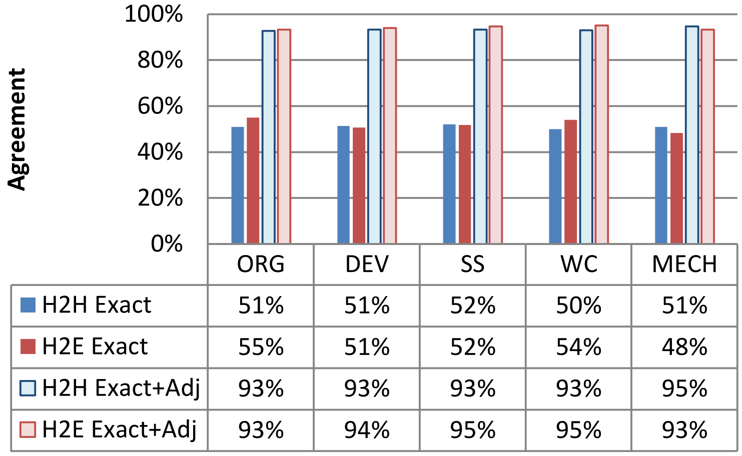
	ORG	DEV	SS	WC	MECH
🟦 H2H Exact	51%	50%	50%	52%	52%
🟥 H2E Exact	54%	52%	51%	54%	55%
□ H2H Exact+Adj	95%	93%	96%	95%	96%
□ H2E Exact+Adj	94%	95%	95%	95%	96%

Figure 13. Grade 6 Comparability (H2 Sample)



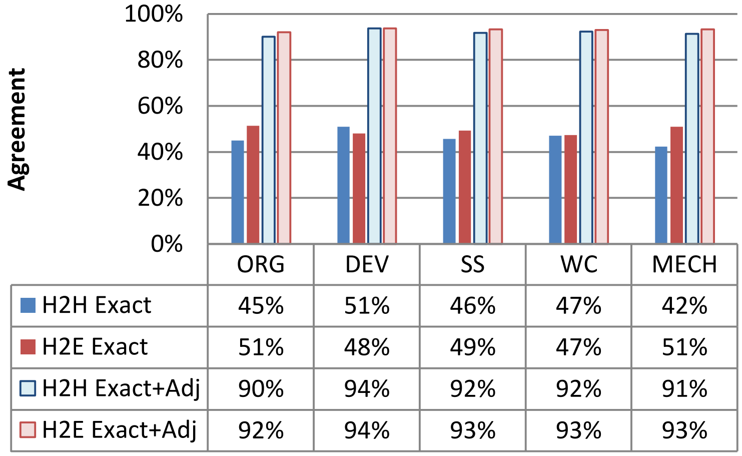
	ORG	DEV	SS	WC	MECH
🟦 H2H Exact	51%	52%	41%	49%	54%
🟥 H2E Exact	50%	51%	46%	53%	57%
□ H2H Exact+Adj	94%	94%	93%	93%	96%
□ H2E Exact+Adj	95%	97%	96%	97%	96%

Figure 14. Grade 7 Comparability (H2 Sample)



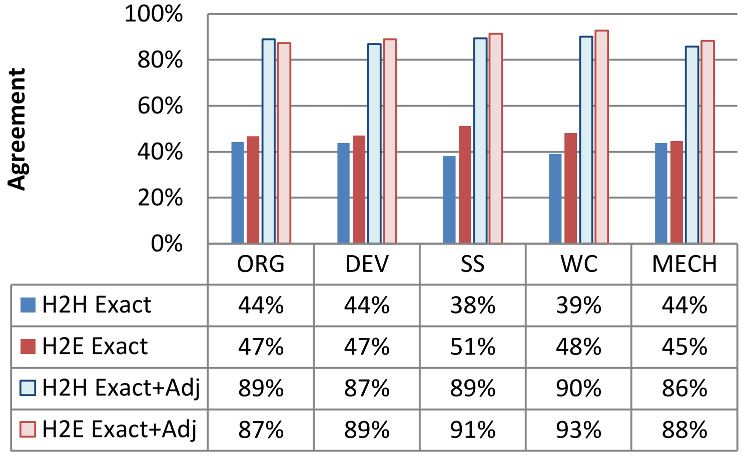
	ORG	DEV	SS	WC	MECH
■ H2H Exact	51%	51%	52%	50%	51%
■ H2E Exact	55%	51%	52%	54%	48%
□ H2H Exact+Adj	93%	93%	93%	93%	95%
□ H2E Exact+Adj	93%	94%	95%	95%	93%

Figure 15. Grade 8 Comparability (H2 Sample)



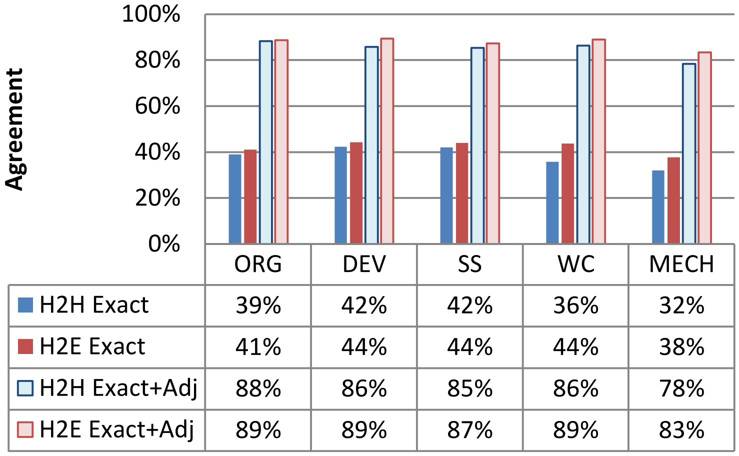
	ORG	DEV	SS	WC	MECH
■ H2H Exact	45%	51%	46%	47%	42%
■ H2E Exact	51%	48%	49%	47%	51%
□ H2H Exact+Adj	90%	94%	92%	92%	91%
□ H2E Exact+Adj	92%	94%	93%	93%	93%

Figure 16. Grade 9 Comparability (H2 Sample)



	ORG	DEV	SS	WC	MECH
■ H2H Exact	44%	44%	38%	39%	44%
■ H2E Exact	47%	47%	51%	48%	45%
□ H2H Exact+Adj	89%	87%	89%	90%	86%
□ H2E Exact+Adj	87%	89%	91%	93%	88%

Figure 17. Grade 10 Comparability (H2 Sample)



	ORG	DEV	SS	WC	MECH
🟦 H2H Exact	39%	42%	42%	36%	32%
🟥 H2E Exact	41%	44%	44%	44%	38%
□ H2H Exact+Adj	88%	86%	85%	86%	78%
□ H2E Exact+Adj	89%	89%	87%	89%	83%

Figure 18. Grade 11 Comparability (H2 Sample)

EducateWV
Enhancing Learning. For Now. For the Future.

West Virginia Department of
EDUCATION
James B. Phares Ed.D.
State Superintendent of Schools