APPLIED MEASUREMENT IN EDUCATION
2018, VOL. 31, NO. 3, 215–232
https://doi.org/10.1080/08957347.2018.1464452

Routledge
Taylor & Francis Group

The effectiveness of machine score-ability ratings in predicting automated scoring performance

Susan Lottridge[a], Scott Wood[b], and Dan Shaw[c]

[a]Psychometrics, American Institutes for Research, Washington, District of Columbia; [b]ACT, Research Technology, Data Science, & Analytics, Lakewood, Colorado, USA; [c]Writing and Communications, ACT, Iowa City, Iowa

ABSTRACT

This study sought to provide a framework for evaluating machine score-ability of items using a new score-ability rating scale, and to determine the extent to which ratings were predictive of observed automated scoring performance. The study listed and described a set of factors that are thought to influence machine score-ability; these factors informed the score-ability rating applied by expert raters. Five Reading items, six Science items, and 10 Math items were examined. Experts in automated scoring served as reviewers, providing independent ratings of score-ability before engine calibration. Following the rating, engines were calibrated and their performances were evaluated using common industry criteria. Three derived criteria from the engine evaluations were computed: the score-ability value in the rating scale based on the empirical results, the number of industry evaluation criteria met by the engine, the approval status of the engine based on the number of criteria met. The results indicated that the score-ability ratings were moderately correlated with Science score-ability, the ratings were weakly correlated with Math score-ability, and were not correlated with Reading score-ability.

Introduction

A common question asked of vendors who offer automated scoring services is whether an automated scoring engine can score responses to constructed response items in such a way that is comparable to how humans might score them. Answering this question accurately can save on the chain of costs associated with item development, field-testing, hand-scoring, and engine training. This issue is particularly critical for assessment programs seeking to minimize hand-scoring costs and to decrease the time needed to return scores, while also retaining constructed response items.

During item development, an accurate answer to this question stands to expand workflow processes to include automated scoring expert review of items for machine score-ability and potentially result in items that are more likely to be approved for operational use. Here, machine score-ability refers to the ability of an engine to produce scores comparable to humans. Item development is an expensive endeavor, with development costs of upward of $2,500 per item for a large-scale summative test (Rudner, 2010). Thus, efforts to ensure the operational success of items are important.

Additionally, the field-testing and hand-scoring restricted to items identified as machine score-able can save on test administration and scoring costs. Engine training activities can then be restricted to items that are expected to be machine score-able, saving costs for this generally expensive and expert-driven process. Depending on item complexity, engine training can cost between $1,100 and $6,000 per item and per-response human scoring costs can range between

CONTACT Susan Lottridge slottridge@air.org American Institutes for Research, Washington, DC 20007, USA

$0.60 and $2.00 (Topol, Olson, & Roeber, 2014). Item complexity factors include number of response opportunities and scoring dependencies between them, number of traits in the scoring rubric, and the extent to which the item elicits varied and complex responses. In addition, costs are driven by number of human raters, adjudication rules, and read-behind percentages. Often, automated scoring vendors request two independent human ratings with adjudication of discrepant scores to ensure that the best possible score is used to train the engine and evaluate engine performance (Williamson, Xi, & Breyer, 2012).

In practice, automated scoring vendors recommend changes to the item design (e.g., segment the item into text entry boxes with one box per concept), to the rubric (e.g., explicitly define score categories, ensure alignment of rubric and item, and further expand number and range of exemplar papers), or conduct proof-of-concept studies on a set of items to empirically examine score-ability because it can be difficult to predict score-ability of items. The change in item design, while often justifiable when looking through one lens of scoring, can create other issues. Such issues include increased item design complexity and the attendant quality assurance issues, as well as potentially increasing the scoring complexity as the engine or humans may need to explicitly address cross-part overlap and examinee input into the "wrong" text entry box. Such a change also may not necessarily lead to improved machine score-ability because of the potential for added complexity. On the other hand, a discussion about revising the item format to make it more machine score-able may make the item clearer to examinees and/or more defined to those who are scoring the item. Enhancing the rubric to provide more descriptive detail for scorers is another way to improve score-ability for humans, and thus potentially improve score-ability for automated scoring engines (Leacock, Messineo, & Zhang, 2013). One study on human score reliability found that item characteristics such as a smaller number of well-defined concepts can impact human–human quadratic weighed kappa values (Leacock, Messineo, & Zhang, 2013).

The empirical approach of training items of unknown machine score-ability can be expensive and inefficient, as it can require hand-scoring to obtain training and test samples of sufficient size (approximately 750–1,000 total responses scored) and requires expert-trained staff to train the engine and provide reports on engine performance. As mentioned above, this can be a very expensive step to take without some evidence of score-ability.

One study of note has examined the extent to which item feature metadata can predict machine score-ability (McGraw-Hill Education CTB, 2014). This study was part of a larger evaluation study of automated scoring items for the Smarter Balanced Consortia Field test items. The study was conducted on a large item set (657 English Language Arts (ELA)/Literacy and 233 mathematics short-text items) and evaluated the extent to which item grade level, depth of knowledge, predicted difficulty, claim, and rubric score point scale predicted human–human quadratic weighted kappa and engine–human quadratic weighted kappa. Analyses of variance (ANOVAs) were conducted to determine whether the human–human quadratic weighted kappa and engine–human quadratic weighted kappa average results differed for sub-categories of the item features. The focus of the study was not necessary to examine the item-by-item differential between the two quadratic weighted kappas, but rather to examine the extent to which the features are associated with higher quadratic weighted kappas (averaged across items). This study employed multiple engines and used the best-performing engine results in the evaluation. The results indicated that in ELA, grade level, and claim type were predictive of engine–human quadratic weighted kappa. The study also found differences between the human–human quadratic weighted kappa and the engine–human quadratic weighted kappa by content area (with the engine performing better than humans in ELA and worse than humans in math). In ELA, difficulty and depth of knowledge were not found to be predictive of quadratic weighted kappa. In math, mathematical reasoning items were found to be predictive of quadratic weighted kappa. These reasoning items had lower engine–human quadratic weighted kappa than human–human quadratic weighted kappa and the non-math reasoning items were found to have higher quadratic weighted kappas than the reasoning items.

The current state of automated scoring is such that the success of the engines on scoring constructed response items exhibits uneven performance and is empirically driven (McGraw Hill, 2014). Prior research has shown the automated scoring engines perform comparably to humans on essays (Shermis & Hamner, 2013), but generally underperform relative to humans on constructed response items (Shermis, 2015). In addition, recent work on the items from the Smarter Balanced Consortium (McGraw-Hill Education CTB, 2014) found that the best-performing automated scoring engines still only met their performance criteria for 54% of the math reasoning items and 39% of ELA/Literacy items. Thus, the effort expended to identify the machine score-ability of items (or modifying items to be machine score-able) prior to field-testing and training may result in higher percentages.

In this article, we present a framework for evaluating item score-ability using a new, simple score-ability rating scale and then examine the extent to which the ratings are predictive of engine performance. We also list and describe factors that we believe influence machine score-ability. The factors build on those listed in Williamson et al. (2010): number of potential concepts that the item might elicit, the variety of ways concepts can be expressed, and the extent to which the concept expressions are distinct. To our knowledge, no study has yet attempted to empirically verify the use of expert judgment in score prediction. While expert judgment has not demonstrated itself as sufficiently accurate or reliable as statistical prediction when multiple factors need to be weighted in the evaluation (Meehl, 1954), it may provide a lens for identifying features for machine score-ability. The identification of features could then be used to provide an index for machine score-ability ratings during item development. It is our hope that this article provides a first step in the development of such a rating.

The study was conducted on constructed response items—namely, those items that are scored for correctness rather than writing style—because these items have been found to vary in engine performance (McGraw-Hill Education CTB, 2014). The content areas studied were in Math, Reading, and Science. The research questions of the study are presented below and will be answered for each of the content areas of Reading, Science, and Math.

(1) How well do machine score-ability ratings correlate with observed score-ability, as measured by the score-ability index and as measured by common standards (i.e., thresholds set on quadratic weighted kappa, exact agreement, and standardized mean difference) used for evaluating engine performance? Observed score-ability is the extent to which the engine-produced scores actually met the standards.
(2) How well do machine score-ability ratings correlate with quadratic weighted kappa values between the engine and the human scores across the items?
(3) How well do machine score-ability ratings correlate with the relative quadratic weighted kappa and exact agreement values (human minus engine) across the items?
(4) How well do machine score-ability ratings correlate with the standardized mean difference of the engine and human scores across the items?

Method

In this section, we describe the items, examinee sample, hand-scoring methods and data, the machine score-ability index, the panel review process, the engine calibration process, the standards for engine evaluation, and the outcome evaluation measures.

Instruments

Twenty-one items were used in this study: five Reading, six Science, and 10 Math. Items were written at the eighth- and ninth-grade levels and were part of a suite of summative and interim tests for

college and career readiness. The items are confidential, and as a result they cannot be displayed. As an alternative, a description of the expected correct response is presented for each item. Item IDs reflect the subject area (R = Reading, S = Science, M = Math) and the order in which the items were judged to be machine-scorable by the expert raters. A brief description of the item characteristics is presented in Table 1.

The five Reading items were single-part (i.e., presented with a single text entry box) with score points varying from 0-1-2 to 0-1-2-3-4. Items generally had one or two stimuli, which were reading passages. The items prompted examinees to list evidence or reasons, or to provide claims with some justification.

The six Science items were single-part with score points varying from 0-1-2 to 0-1-2-3. Items generally had two or three stimuli, which were tables of data or reading passages. The items prompted examinees to identify a correct answer and provide an explanation for their answer, to describe a relationship between phenomena, or to provide descriptions in response to questions.

The 10 Math items were single-part with score points of 0-1-2-3. Items generally did not have separate stimuli, although sometimes items included number sequences or had small diagrams within the item prompt itself. In these items, examinees were asked to identify a correct answer and provide an explanation, provide a simple proof, or to identify a correct answer and provide a justification for their answer. In two cases, examinees were asked to simply explain a phenomenon. In this sense, the Math items could be described as "Mathematical Reasoning" items.

Examinee sample

Examinees were selected by virtue of their schools' participation in a large-scale pilot of a new summative assessment program. Examinee responses were typed into an online test delivery platform. Each examinee answered between three and four constructed response items for each test (i.e., per subject area). Examinees may have taken tests in more than one subject area. Only a subset of the full set of constructed response items administered in the pilot were selected for automated scoring calibration.

Respondents were students in grades 8 and 9. The total sample sizes across the set of 21 items varied from 1,119 to 7,103. The total examinee sample was divided into training and test samples,

Table 1. Items and descriptions.

Item	Score points	Number of stimuli	Expected elicited response for full credit
R1	0-1-2	One	List two pieces of evidence
R2	0-1-2	One	List two pieces of evidence
R3	0-1-2-3-4	One	List four reasons
R4	0-1-2	One	Provide one claim + one support for claim
R5	0-1-2-3-4	Two	Provide one claim + three supports (at least one from each passage)
S1	0-1-2	Two	Identify and explain
S2	0-1-2	Two	Identify and explain
S3	0-1-2	Two	Describe relationship between phenomena and identify
S4	0-1-2-3	Two	Compare phenomena, explain results, and identify
S5	0-1-2	Three	Identify and explain
S6	0-1-2	Two	Provide two descriptions
M1	0-1-2-3	None	Identify with justification
M2	0-1-2-3	None	Provide proof
M3	0-1-2-3	None	Identify and explain
M4	0-1-2-3	None	Identify and explain
M5	0-1-2-3	None	Identify with justification
M6	0-1-2-3	None	Provide proof
M7	0-1-2-3	None	Identify and explain
M8	0-1-2-3	None	Identify and explain
M9	0-1-2-3	None	Explain
M10	0-1-2-3	None	Explain

Table 2. Examinee total, training, and test sample sizes by item.

Item	Total	Training	Test
R1	2,262	1,585	677
R2	7,103	4,972	2,131
R3	1,210	848	362
R4	5,097	3,568	1,529
R5	2,064	1,446	618
S1	1,657	1,161	496
S2	1,098	770	328
S3	4,950	3,465	1,485
S4	5,056	3,539	1,517
S5	3,973	2,781	1,192
S6	1,048	735	313
M1	2,013	1,430	583
M2	2,650	1,806	844
M3	1,153	810	343
M4	2,189	1,539	650
M5	2,789	1,976	813
M6	1,199	841	358
M7	6,333	4,433	1,900
M8	1,247	874	373
M9	4,507	3,155	1,352
M10	2,168	1,520	648

with 68% to 70% of the sample assigned to the training sample and the remaining 30% to 32% assigned to the test sample. Table 2 lists the total, training, and test sample sizes.

The total sample sizes for the Reading items ranged from 1,210 to 7,103; the training sample sizes ranged from 848 to 4,972 and the test sample sizes ranged from 362 to 2,131. The total sample sizes for the Science items ranged from 1,048 to 5,056; the training sample sizes ranged from 770 to 3,539 and the test sample sizes ranged from 313 to 1,517. The total sample sizes for the Math items ranged from 1,153 to 6,333; the training sample sizes ranged from 810 to 3,155 and the test sample sizes ranged from 343 to 1,900.

Procedure

Human scoring procedures

The responses were scored by at least one and at most four human raters: a first human rater score, a second (independent) human rater score, a read-behind "expert" human rater score, and an "expert" adjudicated score. Each item included an independent second human rater score, although the proportion of scores applied varied across items; some items had a 100% second human rater score while others had a 50% second human rater score. The choice of these percentage allocations was outside of the scope of this study as these data were originally used for the original study.

Read-behind scores were also assigned as part of the scoring process. The assignment of these scores was not random; rather, they were generally applied as a monitoring tool for individual raters and up to the discretion of the table leader. Finally, "expert" adjudication scores were applied when the two independent read scores differed by more than a score point.

The final, resolved scores were assigned using the following rules in sequence: (1) If an adjudicated expert score existed, set the resolved score to the adjudicated score; (2) If a read-behind score existed, set the resolved score to read-behind score; (3) Otherwise, set the resolved score to the first human rater score. The rules for assigning the final, resolved score were intended to ensure that the final score was the best available score for that response.

Readers were trained using multiple practice sets (with annotations), and had to pass two qualification sets to be eligible to score responses. Readers had a four-year college degree, and many had previous experience scoring similar assessments. Practice and qualification sets included

clear examples as well as responses not perfectly aligned to the scoring criteria. In addition, anchor sets were used to monitor scoring during the operational scoring.

Machine score-ability

Machine score-ability was defined as the ability of an automated scoring engine to yield scores that are comparable to human rater scores. Machine score-ability will vary by the methods and criteria used to define "comparable" and in the degree to which the engine meets the criteria. In this section, the development of the score-ability rating scale is described as are the processes used when applying the scale.

Machine score-ability rating scale

The machine score-ability rating scale contains five levels, from very high machine score-ability (5) to not machine-scorable (1). One focus in developing the scale was to allow for reasonable variation in levels while also limiting the complexity of the judgment task. Table 3 presents the rating scale.

As noted in the descriptions, score-ability evaluations were made relative to human scoring, with an emphasis on achieving comparability to human scoring. This approach was chosen because human scoring often serves as the benchmark for automated scoring (Williamson et al., 2012). The comparability measures included exact agreement rates and score point distributions. These measures were chosen because we believe that they are easily conceptualized. Quadratic weighted kappa (QWK) and standardized mean difference (SMD) are very common comparison measures and, in many senses, superior to exact agreement and score point distributions, and thus could be used in the score-ability descriptions. However, while these measures are common in the industry for evaluating engine performance (McGraw-Hill Education CTB, 2014; Williamson et al., 2012), we believe that they are somewhat more difficult for raters to conceptualize and therefore predict. That said, their use in the scale descriptions may provide additional value, particularly if the rating scale is revised to map closely to common industry standards.

Table 3. Machine score-ability rating scale.

Score-ability scale	Score-ability descriptor	Score-ability description
5	Very high	Items for which an automated scoring (AS) engine can yield scores comparable to human scores (HS). Score point distributions from automated scoring should be very similar to distributions obtained by human scorers. Exact agreement rates between an AS engine and a human scorer should be close to, or greater than, the rate between two independent human scorers.
4	High	Items for which an AS engine can yield scores comparable to human scores. Score point distributions from AS should be similar to distributions from human scores, but there are likely some small deviations (e.g., 2 to 4 percentage points). Exact agreement rates for AS will be approximately 2 to 4 percentage points lower than rates from HS.
3	Medium	Items for which an AS engine can yield scores somewhat comparable to human scores. Score point distributions from AS may exhibit 5 to 7 percentage point differences from HS distributions. Exact agreement rates for AS will be approximately 5 to 8 percentage points lower than rates from HS.
2	Low	Items for which an AS engine can yield adequate scores when compared to human scores. Score point distributions from AS may exhibit 8 to 10 percentage point differences from HS distributions. Exact agreement rates for AS will be approximately 8 to 10 percentage points lower than rates from HS.
1	Not score-able	Items are not suitable for automated scoring because AS scores will not be comparable to human scores. AS score point distributions will exhibit 10+ percentage point differences from HS distributions. Exact agreement rates for AS will be approximately 10 or more points lower than rates from HS.

Machine score-ability factors

A set of factors thought to influence machine score-ability were developed and described, and presented to the panel to assist in their evaluations. As mentioned in the literature review, these factors were based on work done by Williamson et al. (2010), MacGraw Hill (2014), and Leacock et al. (2013), as well as the authors' experience with the machine scoring of constructed response items.

These factors included: (1) The number, nature, and relationship of concepts; (2) The alignment of the rubric to the item; (3) The ability of examinees to enter a response; and (4) The alignment of observed scoring behavior to the rubric and the accuracy of rater behavior.

Table 4 presents the list of factors, and their descriptions, considered by panelists in arriving at their machine score-ability ratings.

Review panel composition and methods

Four scoring experts participated in the review. Each panel member had extensive experience in automated scoring of constructed response items. Panel members were asked to provided independent ratings of score-ability using the scale in Table 3 and to provide justification for their ratings. Panelists met in four one-hour sessions to first provide and record their ratings and to discuss their reasoning for the ratings.

Panelists had available to them the following materials for review: item text; item rubric; item stimulus materials, if applicable; and, observed training sample data including score distributions, rater agreements (QWK, exact agreement), and the examinee responses. The reasoning underlying

Table 4. Factors influencing machine score-ability ratings.

Score-ability factor	Factor description
Number of concepts	The fewer the number of concepts elicited in the item, the more likely the item is machine score-able. The greater the number of concepts, the less likely the item is machine score-able.
Concept complexity/ variation	The variation in which concepts can be expressed can impact machine score-ability. Concepts which can be narrowly described using fewer terms are likely to be more accurately scored than those which can be broadly described with a large number of terms. Note that examinee experience with concepts can influence variation; examinees who are not familiar with explaining concepts (e.g., slope) may use a wide variety of terms to describe the concept that are unexpected and atypical.
Concept overlap	An item that elicits distinct concepts is likely to be more machine score-able than one that elicits similar or overlapping concepts.
Concept relationships	An item that asks examinees to compare, contrast, or otherwise describe relationships between concepts is likely to be more difficult to machine score.
Item-Rubric alignment	The extent to which the item is aligned to the rubric impacts machine score-ability. For instance, the item may appear to be written to elicit a certain set of concepts but the set of concepts do not appear aligned to the descriptions in the rubric. Anchor papers at each score point can be helpful in outlining the intended alignment. In addition, the directions in the item may not clearly map to the rubric. For instance, the guide may require one claim and three sources of evidence but the directions may not directly inform students of the requirements to achieve scores. It should be noted that the extent of direction may be part of the construct assessed.
Response entry	Another impact on machine score-ability is the extent to which examinees have the facility to consistently and appropriately enter concepts (e.g., formulas, angles) in plain text using a computer keyboard. This issue tends to apply to certain mathematics items, as well as items that involve quantitative answers (such as the graphs or tables in a science item).
Rater behavior and observed scores	The extent to which raters agree with one another overall and within each rubric score point can impact score-ability. Because automated scoring engines use human scores as the independent variable for training, the quality of human scoring is a critical factor in engine performance. Related to the numeric agreement metrics is also the distribution of scores along the rubric score range; if few responses exist at a score range, the engine will presumably exhibit lower accuracy at these ranges because it does not have sufficient data to "learn" the characteristics of responses at these score points. Last, the observed alignment of rater scoring with the rubric also influences machine score-ability. If raters assign scores inconsistently (or consistently but not aligned with the rubric), then this behavior is likely to impact machine score-ability.

the use of these materials in providing machine score-ability ratings was that we believe machine score-ability is influenced by the item characteristics, rubric characteristics, the human rater scoring process, and by the examinee responses. For instance, the item design, stimulus, and scoring guide are expected to elicit certain characteristics of the responses and certain scoring rules. The responses themselves and the hand-scoring of responses reflect the observed characteristics of the responses and the observed application of the scoring rules. Note that the test sample data were not reviewed during this time to preserve the predictive focus of the study on machine score-ability.

The independently derived panel member judgments were recorded during the review session by the coordinator of the panel sessions (Dr. Wood). The average and standard deviation of the panel ratings were computed for each item. Finally, the average across items within rater were computed also to obtain a sense of the predicted score-ability of the content area.

Automated scoring engines

This work was part of a larger study examining the capabilities of automated scoring engines to score these items. One automated scoring engine (Engine 1) was calibrated on all 21 items, and one engine (Engine 2) was calibrated on a subset of the eight items due to resource constraints for the team supporting that engine. The two engines employed in this study used separate code bases with their own unique processing, feature extraction, and scoring pipelines. Each engine used the same source training data to build the final models, and used the same test data to evaluate the final models (see Table 2 for counts). As noted in the engine descriptions below, each engine used text analytic features—primarily *n*-grams—in the feature extraction method. This pipeline applied to the mathematics items, which did not undergo special treatment for algorithmic evaluation. The engines were trained on the final, resolved score; the purpose of using the final, resolved score was to use to "best" score for any given response. Note that other choices are possible in score prediction (e.g., to use all available scores). Each engine used regression-based approaches to predict the final, resolved score, and then used "cut scores" to convert the predicted continuous scores to the rubric scale. Engine 2 set the cut scores to match the training set score distribution. Engine 1 set the cut scores to match the test set distribution in Science and Reading, but set the cut scores to match the training sample distribution in Math. Typically, the cut scores are set using the training sample for proof-of-concept studies, and the cut scores are set using the test sample (or combined test and training sample) for operational scoring use.

Engine 1 used standard preprocessing steps, including spell correction, key term replacement, punctuation handling, and word tokenization. Following this stage, the engine extracted several features including word counts and complexity, sentence count and complexity, lexical diversity, word frequency, spelling accuracy, proposition density and count, term-document frequency matrices, and word2vec semantic vectors. Gradient boosted machines and convolutional neural nets were used to predict scores.

Engine 2 also used standard preprocessing steps, including spell correction, key term replacement, punctuation handling, and word tokenization. Following this step, term-document matrices were built using either count-based measures or frequency-based measures, with the maximum number of *n*-grams varying by item. Models employed to predict scores were linear regression, random forests, and support vector machines.

Engine evaluation metrics

The analysis methods employed standard agreement and score distribution statistics to measure engine performance on the test sample. Agreement statistics used were exact agreement (EA) and QWK. Score distributional statistics were SMD scores and maximum absolute difference in score point percentages between the human rater scores and the engines (SPD Diff). In addition, the difference between the human–human EA and the engine–human EA (EA Diff) and the difference

between the human–human QWK and the engine–human QWK (QWK Diff) were computed for each engine.

The best engine performance was computed for the measures when two engines were used on an item. The evaluations of engine performance were conducted independently for the different measures. For example, one engine may perform better than another on SMD while another may perform better on the QWK and the different engine results were used for each. The reasoning underlying this approach was to use the best performance in identifying machine score-ability on the given metric, although is not practical in live testing. Human agreement indices were computed using human rater 1 and human rater 2, which served as the basis of comparison. Engine agreement indices were computed using human final, resolved score, and the engine score. Mean scores, standard deviations, score distributions, and standardized mean differences were computed for the human final, resolved score, and the engine score.

Industry standards and thresholds

Commonly used criteria for evaluating scoring automated scoring performance relative to humans employ absolute measures and relative measures. Table 5 presents criteria and thresholds from three recent papers (McGraw-Hill Education CTB, 2014; Pearson and ETS, 2015; Williamson et al., 2012). We used the following criteria to determine whether items could be "approved" for operational use. If the automated scoring engine met all four criteria, then it was considered "approved" for scoring for the purposes of this study: (a) QWK ≥ .70; (b) $QWK_{h1\text{-}h2} - QWK_{engine\text{-}h} < .10$; (c) $EA_{h1\text{-}h2} - EA_{engine\text{-}h} < 5.25\%$; and (d) $|SMD| \leq .15$.

Outcome measures

Several outcome measures were used to evaluate the effectiveness of the machine score-ability ratings in predicting actual engine performance. The outcome measures were based the various statistics computed to arrive at three derived values: (1) The score-ability ratings observed in the data using the machine score-ability rating scale; (2) The number of engine evaluation criteria met; and (3) Whether the engine was approved for use. Correlations of the predicted score-ability ratings with each of these measures were computed on the item scores for the three content areas.

The correlation of the QWK for two human rater scores on the training sample (Train $QWK_{H\text{-}H}$) with the above measures was also computed as a baseline. The reason for using Train $QWK_{H\text{-}H}$ as a baseline is that it is commonly computed and it can be thought to represent level of accuracy in human scoring. In other words, a high Train $QWK_{H\text{-}H}$ suggests that there is strong agreement between raters and thus a better chance for an automated scoring engine to successfully mimic scorers. In this sense, such a measure might function as a reasonable proxy for predicting score-ability. It is important to note that the Train $QWK_{H\text{-}H}$ assumes that rater scores are independent,

Table 5. Criteria and thresholds for flagging automated scoring engines, by source.

	Thresholds		
Criteria	Williamson et al. (2012)	*Smarter balanced* (McGraw-Hill Education CTB, 2014)	PARCC (Pearson and ETS, 2015)
Quadratic weighted kappa for engine score and human scores	QWK ≥ .70	QWK ≥ .70	QWK ≥ .70
Absolute SMD between engine score and human score	$\|SMD\| \leq .15$	$\|SMD\| < .12$	$\|SMD\| \leq .15$
Difference between QWK of two humans and QWK of engine score and human score	$QWK_{h1\text{-}h2} - QWK_{engine\text{-}h} < .10$	$QWK_{h1\text{-}h2} - QWK_{engine\text{-}h} < .10$	$QWK_{h1\text{-}h2} - QWK_{engine\text{-}h} < .10$
Difference between Exact Agreement two humans and Exact Agreement of engine score and human score	n.a.	$EA_{h1\text{-}h2} - EA_{engine\text{-}h} < 5\%$	$EA_{h1\text{-}h2} - EA_{engine\text{-}h} < 5.25\%$

and not influenced by human rater conversations that may artificially inflate the agreement values. Additionally, Train $QWK_{H\text{-}H}$ applies only to human score-ability and, as the above studies have demonstrated, may not translate to machine score-ability. Still, it is commonly used as an initial bar to predict engine performance; that is, if humans can't agree, then we might expect that the engines will struggle to accurately predict scores.

Results

The results present the machine score-ability rating results from the panel, various engine calibration results using the metrics described in the methods section, and finally the relationship of the score-ability ratings with the engine evaluation metrics and the derived criteria.

Machine score-ability ratings

Table 6 presents the individual panel ratings for each item and content area, as well as the average ratings and standard deviations for each item. The tables also include a summary of panelist notes for each item. Finally, the tables also present the average rating across items for each panelist. Recall that the rating scale ranges from 1 to 5, with 5 representing very high machine score-ability and 1 presenting non-machine score-ability.

For the five Reading items, the average panelist ratings ranged from 1.78 to 3.83. The average across all items and panelists was 3.12. The standard deviations ranged from .27 to .70, with average being .46. Panelists generally agreed on items R1-R4 but varied in their judgment on R5 (range of 1 to 2.63). For the six Science items, the average panelist ratings ranged from 2.00 to 4.00. The average across all items and panelists was 3.18. The standard deviations ranged from .00 to .58, with average being .24. Panelists agreed exactly on items S1, S2 and S6, and generally agreed on the other items. For the 10 math items, the average panelist ratings ranged from 1.33 to 3.50. The average across all items and panelists was 2.30. The standard deviations ranged from .00 to .69, with average being .39. Panelists agreed exactly on items M2 and M6, and generally agreed on the other items. The panelists disagreed the most on item M7, with a range of ratings from 1 to 2.67.

Comparing averaged ratings across content areas (Table 7), panelists judged the science items as slightly more score-able (3.18) than the reading items (3.12) and each of reading and science items more score-able than the math items (2.23) and the averaged rater performance (Table 8).

Consistency of rater judgments were computed using the Intra Class Correlation (ICC), with the two-way random model (as each rater examined each item) and the averaged rater performance. The results of these analysis [using the R psych package, with R version 3.4.2 (2017-09-28)] showed excellent agreement in science [ICC = .96, 95% confidence interval (.86, .99)], good to excellent agreement in reading [ICC = .91, 95% confidence interval (.66, .99)], and good agreement in math [ICC = .88, 95% confidence interval (.70, .97)].

Although not presented in the table, single-rater ICCs values, which provide a reliability estimate for any given rater, were .73 in reading, .86 in science, and .66 in math. Thus, reliability estimate in science indicate good reliability, and those in reading and math indicate moderate reliability for a single rater. Math had the lowest single-rater ICC value, suggesting that more than one rater may be warranted for that subject area.

Engine performance

This section presents the performances of the engine on the QWK, EA, Score Means and Standard Deviations, SMD, QWK Diff, and EA Diff Measures.

Table 6. Individual rater and mean machine score-ability ratings with notes.

Item	R1	R2	R3	R4	Mean	*SD*	Notes
R1	4	3.33	4	4	3.83	.34	Small number of statements for evidence. Copies from stimulus accepted as evidence. Inconsistent application of the rubric.
R2	4	4	3	4	3.75	.50	Well-defined concept with small number of answers. Three common answers. Copies from stimulus accepted as evidence.
R3	4	3	3	3	3.25	.50	Very few concepts elicited but many ways to express them. Some concepts overlap.
R4	3	3.33	2.67	3	3.00	.27	Unclear what counts as a "claim." Variation in concept expression. Concept relationships needed. Application of rubric seems inconsistent.
R5	2	1.5	2.63	1	1.78	.70	Many concepts (claims and evidence). Examinee can copy from stimulus for credit.
S1	4	4	4	4	4.00	.00	Negative evidence could be a problem. Concepts are discrete with little variation. Strong rubric alignment.
S2	4	4	4	4	4.00	.00	Few discrete concepts with reasonable variation.
S3	4	4	3.33	4	3.83	.34	Few discrete concepts but examinee must describe relationship between concepts.
S4	3	2	3	3	2.75	.50	Concepts are discrete, but there are many, and the relationships between them may be difficult for the engine.
S5	2	3	3	2	2.50	.58	Concepts are constrained and well-suited for machine-scoring bit response review suggests inconsistent application of the rubric.
S6	2	2	2	2	2.00	.00	Expression of concepts varied considerable, and response review suggests inconsistent application of the rubric.
Item	R1	R2	R3	R4	Mean	*SD*	Notes
M1	4	3	3	4	3.50	.58	Small number of concepts, with some concept variation. Seems more like a 2-point item.
M2	3	3	3	3	3.00	.00	Infinite number of correct answers, but few common ones. Seems more like a 2-point item. Mathematical symbols used inconsistently by examinees.
M3	2	3	3	3	2.75	.50	Few concepts and small concept variation; however, concept overlap may be an issue. Small number of 2 and 3 score points.
M4	2	3	2	2	2.25	.50	Concept variation may be an issue.
M5	2.33	2	2.33	2	2.17	.19	Concepts are well-defined but variation in expression may be an issue.
M6	2	2	2	2	2.00	.00	Mathematical symbols used inconsistently by examinees. Concept variation an issue. Seems like a 2-point item.
M7	1	2	2.67	2	1.92	.69	Small number of top-scoring papers. Concept relationships may be difficult to capture. Range of responses for "1" score point large.
M8	2	2	2	1	1.75	.50	Mathematical symbols used inconsistently by examinees. Small number of concepts but difficult to map to rubric.
M9	1	1.67	2	2	1.67	.47	Mathematical symbols used inconsistently by examinees. Concepts are discrete but variation of expression is high.
M10	1	1	1.33	2	1.33	.47	Large number of concepts with variation in expression. Mathematical symbols used inconsistently by examinees. Small number of 2 and 3 scoring papers.

Table 7. Average machine score-ability ratings across items, by content area.

Content Area	R1	R2	R3	R4	Mean	*SD*
Reading	3.40	3.03	3.06	3.00	3.12	.46
Science	3.17	3.17	3.22	3.17	3.18	.24
Math	2.03	2.27	2.33	2.30	2.23	.39

Table 8. Intra class correlation results by content area.

Content area	ICC	*F*	Df1	Df2	*p*	95% CI Lower	95% CI Upper
Reading	0.91	11	4	12	<.0001	0.66	0.99
Science	0.96	22	5	15	<.0001	0.86	0.99
Math	0.88	8.6	9	27	<.0001	0.70	0.97

Quadratic weighted kappa and exact agreement

Table 9 presents the QWK and EA values for the humans, for each of the two engines, and for the best performing engine on each measure for the three content areas. Engine 1 was primarily used as the "Best Engine" because the Engine 2 was used only for a small subset of items.

Table 9. Human, engine 1, engine 2, and best engine agreements.

	Humans		Engine 1		Engine 2		Best Engine	
Item	QWK	EA	QWK	EA	QWK	EA	QWK	EA
R1	.89	85.4%	.90	85.8%			.90	85.8%
R2	.86	85.2%	.83	81.8%			.83	81.8%
R3	.97	86.5%	.94	81.2%	.94	78.2%	.94	81.2%
R4	.77	76.1%	.80	75.1%			.80	75.1%
R5	.81	70.7%	.83	68.4%	.83	69.4%	.83	68.4%
S1	.96	96.8%	.91	92.3%			.91	92.3%
S2	.96	93.9%	.89	85.4%	.93	91.2%	.93	91.2%
S3	.94	94.2%	.77	74.6%			.77	74.6%
S4	.91	81.6%	.79	62.0%	.85	75.1%	.85	75.1%
S5	.94	97.9%	.72	89.8%			.72	89.8%
S6	.831	81.5%	.740	68.4%			.74	68.4%
M1	.82	80.4%	.81	80.0%	.80	81.3%	.81	81.3%
M2	.89	82.4%	.84	75.9%	.89	81.4%	.89	81.4%
M3	.91	89.2%	.74	76.4%			.74	76.4%
M4	.82	86.2%	.83	85.2%	.84	86.3%	.84	86.3%
M5	.78	83.9%	.74	81.2%	.73	80.2%	.74	81.2%
M6	.76	84.1%	.70	78.8%			.70	78.8%
M7	.79	88.7%	.77	87.1%			.77	87.1%
M8	.91	89.8%	.74	80.4%			.74	80.4%
M9	.73	87.8%	.75	87.9%			.75	87.9%
M10	.72	86.9%	.65	83.2%			.65	83.2%

Table 10. Human, engine 1, engine 2, and best engine score means and standard deviations.

	Human		Engine 1		Engine 2	
Item	Mean	*SD*	Mean	*SD*	Mean	*SD*
R1	.90	.86	.91	.86		
R2	1.02	.77	1.02	.77		
R3	1.31	1.53	1.32	1.53	1.33	1.50
R4	1.02	.86	1.02	.86		
R5	1.08	1.03	1.07	1.02	1.05	.98
S1	.32	.66	.33	.66		
S2	.65	.85	.65	.85	.63	.83
S3	.70	.75	.70	.75		
S4	1.22	1.09	1.22	1.09	1.23	1.09
S5	.19	.54	.19	.54		
S6	.75	.80	.75	.80		
M1	.57	.88	.57	.88	.57	.88
M2	.60	.94	.61	.93	.59	.92
M3	.40	.74	.40	.76		
M4	.38	.71	.38	.71	.36	.68
M5	.45	.61	.47	.62	.47	.61
M6	.52	.59	.48	.60		
M7	.36	.50	.33	.49		
M8	.39	.80	.40	.81		
M9	.27	.51	.26	.51		
M10	.32	.53	.33	.54		

Score means and standard deviations

Table 10 provides score means and standard deviations for the human scores, and each of the engine scores for each item by content area.

Score-ability ratings, performance values, and derived criteria

Table 11–13 present a summary of the results for each item and the three content areas for the predicted score-ability, the engine performance values, and the derived criteria. Each table lists the predicted and observed score-ability ratings, the various metrics used to evaluate the "best" engine

Table 11. Predicted and observed score-ability, performance metrics, and derived criteria for reading.

			"Best" engine performance metrics					Derived criteria		
Item	Train $QWK_{H\text{-}H}$	Score-ability (Pred)	QWK	QWK Diff	EA Diff	|SMD|	SPD Diff	Score-ability	Criteria met	Approved
R1	.88	3.83	.90	−.01	−.4%	.00	0%	5	4	1
R2	.86	3.75	.83	.03	3.4%	.00	0%	4	4	1
R3	.97	3.25	.94	.03	5.3%	.01	3%	3	3	0
R4	.82	3.00	.80	−.03	1.0%	.00	0%	5	4	1
R5	.82	1.78	.83	−.03	1.3%	.03	2%	5	4	1
$r_{Score\text{-}ability}$	*.45*		*.48*	*.55*	*.09*	*−.89*	*−.53*	*−.30*	*−.09*	*−.09*
$r_{TrainQWK}$	*.45*		*.92*	*.72*	*.68*	*−.16*	*.51*	*−.86*	*−.90*	*−.90*

Table 12. Predicted and observed score-ability, performance metrics, and derived criteria for science.

			"Best" engine performance metrics					Derived criteria		
Item	Train $QWK_{H\text{-}H}$	Score-ability (Pred)	QWK	QWK Diff	EA Diff	|SMD|	SPD Diff	Score-ability	Criteria met	Approved
S1	.96	4.00	.91	.06	4.5%	.01	0%	3.5	4	1
S2	.96	4.00	.93	.03	2.7%	.03	2%	4	4	1
S3	.95	3.83	.77	.18	19.6%	.00	0%	1	2	0
S4	.90	2.75	.85	.06	6.5%	.01	1%	3	3	0
S5	.97	2.50	.72	.22	8.1%	.00	0%	2	2	0
S6	.85	2.00	.74	.09	13.1%	.00	0%	1	3	0
$r_{Score\text{-}ability}$	*.70*		*.71*	*−.29*	*−.19*	*.48*	*.39*	*.53*	*.43*	*.73*
$r_{TrainQWK}$	*.70*		*.32*	*.27*	*−.26*	*.20*	*.12*	*.43*	*−.02*	*.44*

Table 13. Predicted and observed score-ability, performance metrics, and derived criteria for math.

			"Best" engine performance metrics					Derived criteria		
Item	Train $QWK_{H\text{-}H}$	Score-ability (Pred)	QWK	QWK Diff	EA Diff	|SMD|	SPD Diff	Score-ability	Criteria met	Approved
M1	0.78	3.50	.81	.01	−.9%	.00	1%	5	4	1
M2	0.92	3.00	.89	.00	1.0%	.00	0%	4	4	1
M3	0.91	2.75	.74	.17	12.8%	.01	1%	2	2	0
M4	0.85	2.25	.84	−.02	−.1%	.01	0%	5	4	1
M5	0.80	2.17	.74	.04	2.7%	.02	1%	4	4	1
M6	0.85	2.00	.70	.06	5.3%	.06	4%	3	3	0
M7	0.81	1.92	.77	.02	1.6%	.06	3%	4	4	1
M8	0.89	1.75	.74	.17	9.4%	.02	1%	2	2	0
M9	0.74	1.67	.75	−.03	−.1%	01	1%	5	4	1
M10	0.62	1.33	.65	.06	3.7%	.01	0%	4	3	0
$r_{Score\text{-}ability}$	*.49*		*.54*	*−.13*	*−.02*	*−.40*	*−.21*	*.13*	*.21*	*.30*
$r_{TrainQWK}$	*.49*		*.55*	*.31*	*.39*	*−.03*	*.12*	*−.48*	*−.25*	*.07*

performance and the "score-ability" ratings, the number of criteria met relative to the thresholds for QWK, QWK Diff, EA Diff, and SMD, and whether the engine would be approved for operational use (1 = yes, 0 = no). The absolute SMD is provided in the table to support the use of the correlation because the magnitude of the difference in standardized means is more important than the direction of the difference. The second-to-bottom row of each table lists the correlation value computed from the predicted score-ability rating with each of the metrics and criteria provided in the table. The tables also present the QWK for humans on the training sample (Train $QWK_{H\text{-}H}$) as a reference point. The bottom row of the table presents correlations of the Train $QWK_{H\text{-}H}$ with the other metrics and derived criteria, again to provide a baseline. If the predicted score-ability ratings are "effective," we would expect that the correlations with the QWK Diff, EA Diff, |SMD|, and SMD Diff would be negative. Namely, we would expect high score-ability ratings be associated with smaller difference scores. In contrast, we would expect the correlations with the derived measures to be

positive. The same reasoning holds true for the Train $QWK_{H\text{-}H}$ correlation values with these measures.

In Reading, the predicted score-ability ratings were moderately correlated with engine–human QWK as well as the QWK difference and were not correlated with EA Diff. The predicted score-ability ratings were strongly negatively correlated with the absolute SMD, and moderately negatively correlated with the SPD Diff. The correlation of the predicted score-ability ratings with the derived score-ability ratings was moderately weak and negative. Looking at individual items, the predicted score-ability ratings diverged from the derived score-ability ratings for items R1 (original: 3.83; derived: 5.00), R4 (original: 3.00; derived: 5.00), and R5 (original: 1.79; derived: 5.00). The correlation predicted score-ability ratings with the number of criteria met and whether an item was approved or rejected for operational use was very low and negative. The average predicted score-ability was 3.12 and the derived score-ability was 4.4, with a difference of −1.28. Thus, the raters tended to under-predict score-ability in Reading. The correlation of the predicted score-ability and the Train $QWK_{H\text{-}H}$ was .45. The correlations of the Train $QWK_{H\text{-}H}$ engine performance metric were strong and positive for QWK, and moderate and positive for QWK Diff and EA Diff. The correlation was weak and negative for |SMD| and moderate and positive for SPD Diff. Finally, the correlations were strong, but negative, for the each of the derived criteria. Thus, interestingly, the Train $QWK_{H\text{-}H}$ were inversely related the overall success in engine scoring for the Reading items.

In Science, the correlation of predicted score-ability ratings was moderately strong and positively correlated with the engine QWK. The predicted score-ability ratings were weakly and negatively correlated with QWK Diff and EA Diff. The predicted score-ability ratings were moderately and positively correlated with the absolute SMD and the SPD Diff values. The correlation of the predicted score-ability ratings with the derived score-ability ratings was moderate and positive. Looking at individual items, the predicted and derived score-ability differed the most for item S3 (predicted: 3.83; derived: 1.00). The correlation of predicted score-ability ratings with the number of criteria met was also moderate and positive, and the correlation with the approval Boolean was moderately strong and positive. The average predicted score-ability was 3.18 and the derived score-ability was 2.42, with a difference of .76. Thus, the raters tended to slightly over-predict score-ability. The correlation of the predicted score-ability and the Train $QWK_{H\text{-}H}$ was .70. The correlations of the Train $QWK_{H\text{-}H}$ with the engine performance metrics were generally weak or moderately weak. The correlation of the Train $QWK_{H\text{-}H}$ with the derived score-ability rating was moderate and positive, but weaker than that of the predicted score-ability ratings. The Train $QWK_{H\text{-}H}$ exhibited no relationship with the number of criteria and exhibited a moderate, positive relationship with the overall item approval.

In Math, the predicted score-ability rating was moderately and positively correlated with QWK. The predicted score-ability ratings were weakly and negatively correlated with QWK Diff and not correlated with EA Diff. The predicted score-ability ratings were moderately and negatively correlated with the absolute SMD and weak and negatively correlated with the SPD Diff values. The criteria derived from the metrics showed weak positive correlations with the predicted score-ability ratings. Looking at individual items, the predicted and derived score-ability differed the most for quite a few math items, including M1, M4, M5, M7, M9, and M10. The average predicted score-ability was 2.23 and the derived score-ability was 3.80, with a difference of −1.57. Thus, the raters tended to under-predict score-ability.

The correlation of the predicted score-ability and the Train $QWK_{H\text{-}H}$ was .49. The correlations of the Train $QWK_{H\text{-}H}$ with the engine performance metrics were almost identical to the score-ability prediction. Other than this, the correlations differed from the predicted score-ability rating correlations. The correlation of the Train $QWK_{H\text{-}H}$ with the derived score-ability rating was moderate and negative, much different than that of the predicted score-ability correlation. The Train QWK exhibited a moderately weak and negative relationship with the number of criteria met in the engine evaluation phase and exhibited no relationship to the overall approval of the item for machine scoring.

Summary by research question

This section presents a brief summary of the results organized by research question.

(1) How well do machine score-ability ratings correlate with observed score-ability, as measured by the score-ability index and as measured by common standards used for evaluating engine performance?

The results suggest that the effectiveness of the machine score-ability ratings in predicting machine score-ability varied by content area. Overall, the ratings were moderately successful in predicting Science score-ability, weakly successful in predicting Math score-ability, and poor in predicting Reading score-ability.

In Reading, the predicted score-ability ratings were negatively correlated with the derived score-ability ratings, the number of criteria met, and the engine approval criteria. The correlation of the predicted score-ability ratings with the derived score-ability ratings was −.30. The correlation with the number of criteria passing was 0.09. And, the correlation with the approval status was −.09. Specifically, items receiving the lowest-scoring ratings (R4 and R5) actually performed well in the engine calibrations. For these items, the raters assigned scores of 3 and 1.79, respectively, and the item engine performance was 5 for each item.

In Science, the predicted score-ability ratings were positively and moderately correlated with the derived criteria. The correlation of the predicted score-ability ratings with the derived score-ability ratings was .53. The correlation with the number of criteria passing was .42. And, the correlation with the approval status was .73. The predictions were well-aligned with derived score-ability values except for item S3 (predicted: 3.83; derived: 1.00). In this case, the raters over-estimated the machine score-ability of the items.

In Math, the predicted score-ability ratings were positively and weakly correlated with the derived criteria. The correlation of the predicted score-ability ratings with the derived score-ability ratings was .13. The correlation with the number of criteria passing was .21. And, the correlation with the approval status was .30. Looking at individual items, the predicted and derived score-ability differed the most of the math items.

(2) How well do machine score-ability ratings correlate with QWK values produced by automated scoring engines?

The correlations of the predicted machine score-ability rating with QWK were moderate across the content areas. In Reading, the predicted score-ability ratings are moderately correlated (.48) with engine–human QWK. In Science, the correlation of predicted score-ability ratings was moderately strong and positively correlated with the engine QWK (.71). In Math, the predicted score-ability rating was moderately and positively correlated with QWK (.54).

(3) How well do machine score-ability ratings correlate with the relative QWK and exact agreement values (human minus engine)?

The correlation of the predicted score-ability ratings with the QWK difference value ranged across the content areas as did the correlation with the EA differences. In Reading, the correlation with the QWK difference was .55. In Science, it was −.29. In Math, it was −.13. We would expect the correlation to be negative and strong given that high predicted score-ability ratings should be associated with small (or negative) QWK differences.

For the exact agreement difference, the correlations were generally weak. In Reading, the predicted score-ability ratings were not correlated with the EA Diff (.09). In Science, the correlation

was −.19. In Math, it was −.03. We would expect the correlation to be negative and strong given that high predicted score-ability ratings should be associated with small (or negative) EA differences.

The general instability of difference values may have influenced these results.

(4) How well do machine score-ability judgments correlate with the standardized mean difference?

The correlation of the predicted score-ability ratings with the absolute SMD ranged across the content areas. In Reading, the correlation with the SMD was −.89. In Science, it was .48. In Math, it was −.40. We would expect the correlation to be positive and strong given that high predicted score-ability ratings should be associated with small SMD. That said, the SMD was not particularly well-estimated because the cut score method was applied to the test sample rather than the train sample.

Discussion

Taken together, the results suggest that the machine score-ability ratings were effective in Science, weak in Math, and not effective in Reading. In Reading and Math, the ratings under-predicted score-ability. In Science, the ratings slightly over-predicted score-ability.

While the study did not incorporate quantitative judgments for the key factors influencing the ratings, a review of the panel discussion notes and our reflection of the process suggests potential reasons for the stronger predictions in science, the weaker predictions in math and the poor predictions in Reading. We believe that rater assumptions around feature extraction, around linguistic variation, and around the commonality of frequent response options were critical. For instance, in Science, we posit that the raters benefited from the more specific vocabulary of Science. In addition, the panel may have benefited from a smaller set of common responses which the engine could capture in the training sample. In Reading, we believe that the raters underestimated the linguistic variation in responses and that this variation would not be captured by the predominantly 'bag of words' methodology. In math, a review of the qualitative discussion indicates that the rater predictions were likely influenced by an expectation that the mathematic logic be captured correctly by the engine, and a reasonable tendency in Math to focus on the range of correct and incorrect responses following from the logic but that may be rare. Again, the use of a 'bag of words' methodology likely influenced this evaluation, as this approach does not specifically capture the logic elicited by the Math items.

Interestingly, the score-ability ratings behaved differently from the human–human QWK based upon the training sample (Train $QWK_{H\text{-}H}$) relative to the key measures. This finding suggests that the ratings provide different information on expected performance relative to the common human rater-based reliability measures. This may suggest that high human rater score agreement is not sufficient for predicting engine performance; namely, there may be more to machine score-ability than human agreement. Put another way, engines may operate differently than humans when predicting scores.

In Reading, the Train $QWK_{H\text{-}H}$ value was highly and negatively correlated with each of the three derived measures and performed quite differently from the predicted score-ability ratings (score-ability: -.86 vs. -.30; criteria met: -.90 vs. -.09; approval: -.90 vs. -.09). This result is somewhat counter-intuitive as one might expect high QWKs to reflect high reliability and thus better capability of the engine to mimic human scoring. In Science, the Train $QWK_{H\text{-}H}$ exhibited a slightly weaker correlation with score-ability (.43 vs. .53), no correlation with the number of criteria met (-.02 vs. .43), and a weaker correlation with approval (.44 vs. .73). In Math, the Train $QWK_{H\text{-}H}$ exhibited a moderately negative correlation with score-ability (-.48 versus .13), a weak negative correlation with the number of criteria met (-.25 vs .21), and no correlation with approval (.30 vs. .07).

While the results were mixed, they do suggest that rater judgments of machine score-ability hold promise in Science, and, to some degree, Mathematics. Including the rater judgments of the factors

associated with the score-ability rating (e.g., number of concepts, concept complexity, concept overlap, concept relationships, item-rubric alignment, response entry, and rater behavior) as Booleans or scaled ratings may help to refine and improve the evaluation process. Including these factors evaluations may provide additional insight into which factors are related to machine score-ability, and thereby improve the item development and/or scoring process. Together with a holistic evaluation, the set of judgments may allow for better predictions of score-ability. In addition, improvements in engine development may also assist, as we believe it is difficult for the human raters to fully appreciate the range of ways examinees can type a response, even with extensive experience with student responses. The newer deep learning methods may hold promise in capturing the variation in phrasing common to student responses (Goldberg, 2017).

Study limitations

This study had several imitations worth mentioning. First, the number of items within each content area was small, and thus the computed statistics likely suffer from small sample variability. The small set of items suited the purpose of developing and studying the score-ability measures but limits the generalizability of the findings.

Second, the application of cut scores to match the test set for the first engine, which served as the 'best' engine for most of the results, meant that the score distribution metrics were artificially inflated for Science and Reading. This is because Engine 1 set the cut scores to match the test set distribution in Science and Reading, but set the cut scores to match the training sample (versus test sample) distribution in Math. The use of the Engine 2 score distributional statistics helped to offset this result, but these comprised only a small proportion of items. In our experience, wide variations in score distributions are not common when cuts are properly applied. That said, keeping the test set independent from any score adjustment would have resulted in better evaluation of those metrics and the derived scores.

Third, the use of qualitative descriptions to explain the ratings was not helpful for diagnosing potential reasons ratings that did not accurately predict score-ability. Having the panel quantitatively identify which factors they felt impacted ratings (positively or negatively) may have provided some additional basis on which to evaluate the quality of the ratings and the extent to which factors may themselves predict score-ability.

Fourth, the study had results from the second engine for only a subset of items. While Engine 1 could have been used by itself, the larger study sought to run competing models through separate pipelines to determine the optimal approach to scoring and the scoring results from the best engine. This study's results may have been different if Engine 2 had been trained on all items, as the engine did provide better performance on some items. While having two engine pipelines may not be typical as scoring vendors use only one engine, it is certainly possible to configure an engine to employ separate preprocessing-feature extraction-modelling pipelines to essentially create separate models. Such an approach can have value when examining the impact of feature groupings on score prediction.

Future work

The study results, while mixed, suggest that the score-ability ratings may provide value in identifying score-able items, particularly in Science, and they provide different information than reliability measures like QWK. Including the score-ability factor booleans in the rating process may help to better understand the factors surrounding score-ability and so will be included into the rating procedure and analysis for future work. In addition, continuing to use and refine the scale and factor evaluations in all training studies will build a larger base of item results and will also help to refine the quality and nature of the judgments. In addition, the scale could be complemented with exemplars and training materials to broaden the rater pool from those who have deep experience in

automated scoring of responses. Ultimately, the hope is that such work will result in a combination of automated and item author-judged metrics that reliably and accurately predict score-ability and items that are developed so that they are optimally score-able.

References

Goldberg, Y. (2017). Neural network methods for natural language processing. *Synthesis Lectures on Human Language Technologies*, *37*, 1–309. doi:10.2200/S00762ED1V01Y201703HLT037

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

Leacock, C., Gonzalez, E., & Conarroe, M. (2013). *Developing Effective Scoring Rubrics for Automated Short-Response Scoring*. Monterey, CA: McGraw-Hill Education CTB Research Report.

McGraw-Hill Education CTB. (2014, December 24). *Smarter balanced assessment consortium field test: Automated scoring research studies (in accordance with smarter balanced RFP 17)*. Retrieved from http://www.smarterapp.org/documents/FieldTest_AutomatedScoringResearchStudies.pdf

Meehl, P. (1954). *Clinical versus statistical prediction: A theoretical analysis and a review of the evidence*. Minneapolis, MN: University of Minnesota.

Pearson and ETS. (2015, March 9). *Research results of PARCC automated scoring proof of concept study*. Retrieved from http://www.parcconline.org/images/Resources/Educator-resources/PARCC_AI_Research_Report.pdf

Rudner, L. (2010). Implementing the graduate management admission computerized adaptive test. In W. Van Der Linden & C. Glas (Eds.), *Elements of adaptive testing* (pp. 151–165). New York, NY: Springer.

Shermis, M., & Hamner, B. (2013). *Contrasting state-of-the-art automated scoring of essay: Analysis. The handbook of automated essay evaluation: Current applications and new directions* (pp. 313–346). New York, NY: Routledge Academic.

Shermis, M. D. (2015). Contrasting state-of-the-art in the machine scoring of short-form constructed responses. *Educational Assessment*, *20*, 46–65. doi:10.1080/10627197.2015.997617

Topol, B., Olson, J., & Roeber, E. (2014, February). *Pricing study: Machine scoring of student essays*. Retrieved from http://www.assessmentgroup.org/uploads/ASAP_Pricing_Study_Final.pdf

Williamson, D., Xi, X., & Breyer, F. J. (2012). A framework for the evaluation and use of automated scoring. *Educational Measurement: Issues and Practice*, *31*(1), 2–13. doi:10.1111/j.1745-3992.2011.00223.x

Williamson, D. M., Bennett, R. E., Lazer, S., Bernstein, J., Foltz, P. W., Landauer, T. K., … Sweeney, K. (2010). *Automated scoring for the assessment of common core standards*. Retrieved December 15, 2010, from http://professionals.collegeboard.com/profdownload/Automated-Scoring-for-theAssessment-of-Common-Core-Standards.pdf