Smarter Balanced Assessment Consortium

Field Test: Automated Scoring Research Studies

in accordance with Smarter Balanced RFP 17

McGraw-Hill Education CTB

December 24, 2014

Developed and published under contract with the State of Washington Office of Superintendent of Public Instruction by CTB/McGraw-Hill LLC, 20 Ryan Ranch Road, Monterey, California 93940-5703.

Table of Contents

Table of Tables

Table of Figures

Chapter 1: Automated Research Overview

Executive Summary

Automated Scoring studies conducted as part of the Smarter Balanced Field Test continued and extended the research carried out during the Pilot Study. The Field Test included 683 English language arts (ELA)/literacy short-text, constructed-response, items, 238 mathematics short-text, constructed-response, items, and 66 ELA/literacy essay items. Included with the mathematics items were 41 items developed to test mathematical reasoning. At least one Automated Scoring system was trained for every item. For each item type, four different systems were trained on a subset of 21 items to facilitate evaluation of Automated Scoring systems.

Besides evaluating the performance of Automated Scoring on the Field Test items, four additional research studies focused on two themes: hybrid scoring approaches and extending what can be scored. These themes were selected based on two criteria: the studies could be completed in time for the implementation of 2015–2016 assessment programs, and the results are likely to be directly relevant to assessment programs considering inclusion of Automated Scoring.

Automated Scoring Evaluation

The performance of all Automated Scoring systems was evaluated relative to the final human scores using six criteria, based quadratic weighted kappa (QWK), correlation, perfect agreement rates, and standardized mean difference (SMD). This framework, similar to the framework used in the Pilot Study, is described in more detail in Chapter 3.

Automated Scoring systems were flagged if their scores did not meet performance criteria. The top-performing Automated Scoring system was selected for each item (for each item/trait combination in the case of essay items) based on the number of flags, with ties decided on QWK. The suitability for Automated Scoring for each item was evaluated based on the performance of the top-performing system. The Automated Scoring results for the Field Test items are summarized below. Because the level of human-human inter-rater agreement may influence the performance of Automated Scoring systems, we also summarize human rater performance.

ELA/literacy items

Out of 665 ELA/literacy short-text, constructed-response, items, human-human inter-rater agreement meets all criteria for 268 items (40%), requires review for absolute performance for 6 items (1%), requires review for subgroup performance for 55 items (8%), and does not meet performance criteria for 336 items (51%).

Based on the top-performing Automated Scoring system, 261 items (39%) meet all criteria for Automated Scoring, 47 items (7%) require review for absolute performance, 4 items (1%) require review for performance relative to inter-rater agreement, 183 items (28%) percent require review for subgroup performance, and 170 items (26%) are not suited for Automated Scoring.

Mathematical Reasoning

Out of 41 mathematics short-text, constructed-response, items developed to test mathematical reasoning, human-human inter-rater agreement meets all criteria for 29 items (71%), requires review for subgroup performance for 5 items (12%), and does not meet performance criteria for 7 items (17%).

Based on the top-performing Automated Scoring system, 22 items (54%) meet all criteria, 2 items (5%) require review for absolute performance, 2 items (5%) require review for relative performance, 4 items (10%) require review for subgroup performance, and 11 items (27%) are not suited for Automated Scoring.

Remaining Mathematics items

Out of the remaining 192 mathematics items, human-human inter-rater agreement meets all criteria for 164 items (85%), requires review for absolute performance for 1 item (1%), requires review for subgroup performance for 10 items (5%), and does not meet performance criteria for 17 items (9%).

Based on the top-performing Automated Scoring system, 139 items (74%) meet all criteria, 3 items (2%) require review for absolute performance, 9 items require review for performance relative to inter-rater agreement, 12 items (6%) require review for subgroup performance, 24 items (25%) are not suited for Automated Scoring due to absolute performance and 4 items (2%) are not suited due to relative performance.

ELA essay items

Essay responses were scored on three traits: Organization/Purpose, Evidence/Elaboration, and Conventions. Out of 66 essay items,

- For the Organization/Purpose trait, human-human inter-rater agreement meets all criteria for 60 items (91%), requires review for subgroup performance for 4 items (6%), and does not meet performance criteria for 2 items (3%).
- For the Evidence/Elaboration trait, human-human inter-rater agreement meets all criteria for 59 items (89%), requires review for subgroup performance for 4 items (6%), and did not meet performance criteria for 3 items (5%).
- For the Conventions trait, human-human inter-rater agreement meets all criteria for 26 items (39%), requires review for absolute performance for 2 items (3%), requires review for subgroup performance for 2 items (3%), and does not meet performance criteria for 36 items (55%).

Based on the top-performing Automated Scoring system,

- For the Organization/Purpose trait, 44 items (67%) meet all criteria for Automated Scoring, 13 items (20%) require review for performance relative to inter-rater agreement, 6 items (9%) require review for subgroup performance, and 3 items (5%) are not suited for Automated Scoring.
- For the Evidence/Elaboration trait, 45 items (68%) meet all criteria for Automated Scoring, 12 items (18%) require review for performance relative to inter-rater agreement, 5 items (8%) require review for subgroup performance, 2 items (3%) are not suited for Automated Scoring due to absolute performance, and 2 items (3%) due to relative performance.
- For the Conventions trait, 41 items (62%) meet all criteria for Automated Scoring, 5 items (8%) require review for performance relative to inter-rater agreement, 8 items (12%) require review for subgroup performance, 8 items (12%) are not suited for Automated Scoring due to absolute performance, and 4 items (6%) due to relative performance.

Hybrid Scoring Approaches

Hybrid scoring approaches combine human handscoring with Automated Scoring. The first study, *Continued Read-Behind Scenarios*, considered combinations of Automated Scoring and handscoring at the score level. This included using Automated Scoring systems as a read-behind for a human rater. The second study, *Targeting Responses for Human Review*, considered combining Automated Scoring and handscoring at the response level. Specifically, it studied statistical methods and techniques that can be used to select responses for human review.

Continued Read-Behind Studies

High-stakes assessment programs typically use more than one human rater to score constructed-response items to reduce rater effects. There has been increasing interest in alternative scoring scenarios which combine human and Automated Scoring. Eight of these hybrid scoring scenarios were investigated during the Pilot Study. One finding was that using the best-performing Automated Scoring system as a second rater ("read-behind") with a single human rater resulted in a high score quality.

A major methodological limitation of the Pilot Study, however, was that scores by an independent human rater were not available because collecting these data was organizationally challenging. Instead, the first human rater was used as a human score in the alternative scoring scenarios. This may have inflated the score quality of some scoring scenarios, in particular, the human-Automated Scoring read-behind.

During the Field Test, independent human ratings were collected to further investigate and extend the findings of the Pilot Study research. Based on these new data, the Field Test study confirmed:

- Scoring scenarios where an Automated Scoring system serves as a second rater ("read-behind") behind a human rater produce high quality scores.

Targeting Responses for Human Review

Automated Scoring systems are designed to predict the score that a human rater would assign to a given response based on the associated rubric. Although rater agreement statistics used to evaluate Automated Scoring systems generally show scoring consistency, it is a well-known fact that Automated Scoring provides at times a score that is different from human scores.

In a high-stakes scoring scenario, very little to no score deviation between Automated Scoring and human scoring is acceptable. Hence, it is very important to identify responses where Automated Scoring tends to deviate from human scores. This may depend on several factors, including the features used by the Automated Scoring system, the number of observations per score point, and prediction and classification models that were trained.

In order to identify responses that may need human review, Chapter 5 investigated cross-validation based methods on principal components and number of clusters to identify the responses that lie in the sparse regions of the multi-dimensional feature space. The study concludes:

- Performance measures based on exact agreement, adjacent agreement, discrepant agreement, and quadratic weighted kappa shows that the developed procedure has major improvements over the random selection of responses that were sent out for human review.

Extending What Can Be Scored

The third study, *Item Characteristics that Correlate with Agreement for Handscoring and Automated Scoring*, extended similar research done in the Pilot Study to the Field Test items. It investigated which characteristics of items (e.g., Rubric Type, Claim, etc.) are related to score accuracy of human raters as well as Automated Scoring systems. The fourth study, *Generic Scoring Models*, investigated the feasibility of training an Automated Scoring system to score one trait (Conventions) across several essay prompts, rather than training a scoring model for individual prompts.

Item Characteristics that Correlate with Agreement for Handscoring and Automated Scoring

From experience, about one-third to one-half of short-text items developed for an assessment program cannot be scored reliably by Automated Scoring systems. This drop-out rate increases the cost of item development for an assessment program using Automated Scoring. Chapter 6 investigated whether any of the item characteristic of the Field Test items could be used to predict whether a short-text, constructed-response, item can be scored by an Automated Scoring system. The item characteristics were based on item metadata, the items themselves, and characteristics of the item scoring rubrics. Findings include:

- There was a statistically significant decrease in performance of both Automated Scoring and handscoring when the rubrics of ELA/literacy items were generic as opposed to item-specific rubric.
- Further analyses are needed to understand the effect of Writing Purpose on Reading Comprehension and Brief Writes. For Reading Comprehension, there was higher agreement when the text was fictional for Automated Scoring—and handscoring showed a similar trend. With Brief Writes, there was significantly higher agreement for narrative stimuli for both Automated Scoring and handscoring.

Generic Scoring Models

Chapter 7 investigated the development of generic scoring model to score the Conventions trait for Smarter Balanced essays. Scores for this trait are based on writing conventions as opposed to the content of the essay. Thus, we hypothesized that we can develop a single generic scoring model for each grade level as opposed to item-specific models.

We developed a spelling error feature and five new features based on a grammar-checker: grammar/usage errors, white space errors, capitalization errors, punctuation errors, and stylistic suggestions. These features were added to Vendor 1's feature set. For grades 6 and 11, we trained a generic scoring model on all combinations of five essays and tested on three held-out essays. Based on average quadratic weighed kappa, we conclude:

- It is very likely that a generic scoring model can outperform prompt-specific scoring models for the Conventions trait, given a specific grade.

A probable explanation for this effect is that when combining the training sets in generic scoring, the training model observes many more possible types of grammatical and convention-related errors per given score point than it can find in a single item. Hence, the parameters of the scoring models are computed based on a better knowledge of the errors it may encounter when being tested.

Chapter 1: Automated Research Overview

The research studies related to Automated Scoring that were conducted as part of the Smarter Balanced Field Test continue and extend the research that was conducted during the Pilot Study. The current chapter describes the various item types as they are defined in the data files delivered from the American Institutes for Research (AIR). The Smarter Balanced Assessment Consortium Field Test administration data originates from two main sources: an item metadata file that describes the Content Domain, Claim, Target, and Standards an item measures together with other item attributes (e.g., Depth of Knowledge [DOK]), and a student response data file that includes student response information as well as a subset of item metadata information. In addition to describing items, this chapter also describes how items were selected for Automated Scoring studies.

Chapter 2 summarizes the results of the research conducted during the Pilot Study for the readers' convenience. A detailed report on the Pilot research studies is available online (see www.smarterbalanced.org).

Chapter 3 describes the criteria used to evaluate the functioning of the Automated Scoring of items. These criteria follow the framework used for evaluation in the Pilot Study, but include some refinements to the reporting. The performance of the various Automated Scoring systems on the Field Test items is described. This chapter also provides brief descriptions of the Automated Scoring systems that were a component of the Field Test. Vendors reflect on the lessons learned from participating in the Pilot Study and Field Test and describe various enhancements made to the Automated Scoring systems as a result of the Smarter Balanced studies.

Chapters 4–7 report on the research studies that were conducted to enhance Smarter Balanced's efficacy in its use of Automated Scoring. Online appendices present additional details related to these studies.

Item Descriptions

The Field Test Automated Scoring research conducted on behalf of the Smarter Balanced Assessment Consortium focused on short-text, constructed-response, items (English language arts [ELA]/literacy and mathematics) and essay items. During the Pilot Study, the AIR equation engine was validated. Based on the reported results, Smarter Balanced chose not to include equation items in the Field Test Research Studies.

Item metadata were available for 19,619 Field Test items (9,366 ELA/literacy items and 10,253 mathematics items) in the test delivery system. Of these items, 1,869 items required handscoring by McGraw-Hill Education CTB (CTB) and our scoring subcontractors, either to produce the score of record or to prepare Automated Scoring system training and validation sets.

- 51 items were marked do not score (DNS) by Smarter Balanced for various reasons.
- 79 ELA/literacy short-text, constructed-response, items were only included to facilitate comparison with NAEP and/or PISA and were not considered part of the Smarter Balanced item pool.
- 302 mathematics items have been identified as being dependent, in the sense that accurate scoring depends on the student's response to another item, usually as part of a performance task. While handscoring can accommodate such dependent items, many Automated Scoring systems currently are not capable of scoring such dependent items.

Therefore, these mathematics dependent items were handscored only and not included in the Automated Scoring studies.

Overall, 1,437 items were potential candidates for Automated Scoring studies. Table 1.1 shows the constructed-response item types as identified in metadata files, a description of the item type, the number of items coded with the item type by content area, and the available Automated Scoring system developers for each item type. As noted in the table, some of these developers participated in the Automated Student Assessment Prize (ASAP) competition. One ASAP scoring system (ASAP 3) declined the invitation to participate in the Field Test studies due to scheduling conflicts. Another ASAP scoring system (ASAP 4) required considerable resources to train and validate scoring models during the Pilot Study and was not included in the Field Test study due to scalability concerns.

Table 1.1. Description of Field Test Constructed-Response Item Types and Scoring Systems

Response Type	Description	Number of ELA/literacy Field Test Items	Number of MA Field Test Items	Available Scoring Systems*
Short-text	Short constructed-responses, text only.	907	362	Measurement Incorporated-Project Essay Grade (MI-PEG) TurnItIn/LightSide Labs Luis Tandalla (ASAP 1) Jure Zbontar (ASAP 2) Pawel Jankiewicz (ASAP 5)
Essay	Extended constructed-responses, text only.	168	0	AIR-Open Source Engine Measurement Incorporated-Project Essay Grade (MI-PEG) CTB TurnItIn/LightSide Labs Luis Tandalla (ASAP 1) Jure Zbontar (ASAP 2) Pawel Jankiewicz (ASAP 5)

Note: ASAP refers to the Automated Student Assessment Prize, a competition hosted by the William and Flora Hewlett Foundation. The number refers to the place awarded in the second phase of the public competition to this developer.

Response Sampling and Routing

Based on the results from a research study conducted as part of the Pilot Study, 1,000 on-grade responses is a reasonable minimum for a training set for short-text, constructed-response, items; 1,500 on-grade responses is a minimum for essay items. Moreover, 500 responses is a reasonable size for a validation set. Therefore, 1,500 and 2,000 responses are required for the Automated Scoring systems for the short-text and essay items, respectively.

During the Field Test, many items were either over-exposed or under-exposed to the student population. As a result, items differed greatly in the number of available on-grade responses. If the responses in the Standard Setting and Census Sample could have been combined, the following number of items would have met the minimum requirements (that is, 1,500 or 2,000 responses):

- 771 out of 907 ELA/literacy short-text, constructed-response, items,
- 238 out of 345 mathematics short-text, constructed-response, items, and
- 81 out of 168 ELA/literacy essay items.

Unfortunately, combined handscoring of Standard Setting and Census Sample responses turned out to be operationally infeasible. Therefore, all items that met the minimum requirements either in the Standard Setting Sample or in the Census Sample (or both) were selected. This resulted in selection for Automated Scoring studies for the following number of items (percentage of maximum possible number in parentheses):

- 665 out of 771 ELA/literacy short-text, constructed-response, items (86%),
- 238 out of 345 mathematics short-text, constructed-response, items (69%), and
- 66 out of 81 ELA/literacy essay items (81%).

Overall, the Field Test items requiring handscoring were divided into three cases, as follows:

1. *Case 1. Handscoring only*. These items were not eligible for Automated Scoring studies and were handscored only. Ten percent of the responses were scored with a second read for inter-rater reliability purposes. Items included were:
    a. Any grade-level short-text, constructed-response, items with fewer than 1,500 on-grade responses (except mathematics reasoning items, see Case 3).
    b. Any grade-level essay items with fewer than 2,000 on-grade responses.
    c. Any mathematics short-text, constructed-response, items identified as dependent on another item.
2. *Case 2. Automated Scoring training and validation*. CTB selected a random sample of 1,500 responses (short-text, constructed-response, items) or 2,000 responses (essay items) of the available on-grade responses per item. Of these, a random sample of 500 responses were designated as validation responses while the remaining responses were designated as training responses. The training and validation responses received two human reads and adjudication of any non-exact scores by a senior human rater. The score of record was the senior rater for adjudicated responses or the matched score when the two human scores agreed. Scores of record were compared against engine scores for validation purposes. Items included were:

a. Short-text, constructed-response, items with at least 1,500 on-grade responses.

b. Essay responses with at least 2,000 on-grade responses.

3. *Case 3. Mathematics reasoning items.* Forty-one mathematics short-text, constructed-response, items were identified as mathematics reasoning items. A random sample of up to 2,000 responses received two human reads and adjudication of any non-exact scores by a senior human rater. Note that some of the items have fewer than 1,500 on-grade responses, but were nonetheless included in Automated Scoring studies as part of the mathematics reasoning items.

Selection of Responses

The number of on-grade responses for each item varied considerably. To select the scoring samples required for Case 2 and Case 3 items, random samples were drawn from the response data from the available on-grade responses in either the Standard Setting Sample or the Census Sample. Note that both the Standard Setting Sample and the Census Sample are representative of the student population as a whole by design.

Selection of Items for Each Automated Scoring System

Finally, the items were distributed to the Automated Scoring systems, as follows.

- *Case 1. Handscoring only*. These items were not routed to any Automated Scoring system.
- *Case 2. Automated Scoring training and validation.*
  - All short-text, constructed-response, items were routed to MI-PEG. In addition:
    - A subset of 21 ELA/literacy short-text, constructed-response, items were also sent to the ASAP 1, ASAP 2, and ASAP 5 systems.
    - A subset of 21 mathematics short-text, constructed-response, items were also sent to the ASAP 1, ASAP 2, and ASAP 5 systems.
  - All essay items were routed to the AIR-OSE, CTB, and MI-PEG engines. In addition, a subset of 21 selected essay items were routed to TurnItIn/LightSide Labs. Seven items were selected from each of the grade bands 3-5, 6-8, and 9-11.
- *Case 3. Mathematics reasoning items.* All mathematics reasoning items were routed to MI-PEG and TurnItIn/LightSide Labs.

These items were selected to be representative as best as possible within Smarter Balanced preferences and the operational constraints. Short-text, constructed-response, items with a sufficient number of responses available (1,500–2,000) were scored by MI-PEG. Similarly, ELA/literacy essay items with a sufficient number of responses available (2,000) were scored by AIR-OSE, CTB, and MI-PEG.

Table 1.2. Automated Scoring Systems and Item Counts by Score Type

Content	Score Type	AIR-OSE	CTB	LIGHT SIDE	MI-PEG	ASAP 1	ASAP 2	ASAP 5
ELA/literacy	ST				683	21	21	21
	Essay	67	67	21	67			
Mathematics	ST				238	21	21	21
	MR			41	41			

Note: ST = Short-text, constructed-response, item. MR = Mathematical reasoning. Due to delays in delivery of results and timeline constraints, scores from the AIR-OSE system could not be included in this report.

Overview of the Field Test Research Studies

During the Smarter Balanced Pilot Study, several research studies were conducted related to Automated Scoring. These studies addressed the following topics:

- Performance of Automated Scoring systems
- Training, validating, and deploying Automated Scoring systems
- Developing items for Automated Scoring
- Operational scoring

The results of these studies are summarized in Chapter 2. This section provides an overview of the research studies conducted during the Field Test. These studies continue and extend the research started in the Pilot Study. Table 1.3 provides an overview with a short description. Chapters 4–7 contain more detailed reports on the research studies.

Two factors were considered important in the decision to conduct these studies: (1) the results of these studies are likely to be directly relevant to assessment programs considering inclusion of Automated Scoring and (2) the studies can be completed in time for the implementation of the 2015–2016 assessment programs.

The Field Test research studies address three topics:

- Performance of Automated Scoring systems
- Hybrid scoring approaches
- Extending what can be scored

The performance of several Automated Scoring systems on the Field Test items were evaluated, similarly to what was done with the Pilot Study items. An overview of the vendors scoring each item type is given in Table 1.2. The Pilot Study evaluation framework was used in the Field Test (see Table 3.1), but with some additional refinements in the reporting, as described in Chapter 3.

Two studies (Chapters 4 and 5) focused on hybrid scoring approaches, which combine Automated Scoring and handscoring and two studies (Chapters 6 and 7) focused on extending what can be scored.

The first study, *Continued Read-Behind Scenarios*, considered combinations of Automated Scoring and handscoring at the score level. This includes using Automated Scoring systems as a read-behind for a human rater. The results of this study are relevant for assessment programs considering implementing Automated Scoring in various stages.

The second study, *Targeting Responses for Human Review*, considered combining Automated Scoring and handscoring at the response level. Specifically, it studied statistical methods and techniques that can be used to select responses for human review. Several techniques to detect potential outlier responses were studied during the Pilot Study. The Field Test study extended this research and also investigated the impact on score quality. The results of this study can inform operational decision making, for example, when the percentage of responses reviewed needs to be considered in light of scoring costs.

The third study, *Item Characteristics that Correlate with Agreement for Handscoring and Automated Scoring*, extended a similar study done in the Pilot Study to the Field Test items. It investigates which characteristics (e.g., Rubric Type, Claim, etc.) of items are related to score accuracy of human raters as well as Automated Scoring systems. The results of this study are relevant for item development in the future.

The fourth study, *Generic Scoring Models*, investigated the feasibility of training an Automated Scoring system to score one trait (Conventions) across several essay prompts, rather than training a scoring model for individual prompts. If generic scoring models can achieve satisfactory score quality, then it may be possible to reduce costs associated with Automated Scoring by using such models.

Table 1.3. Overview of the Field Test Research Studies

Category	Study	Chapter	Short Description
Automated Scoring evaluation	Automated Scoring evaluation	3	Evaluation of performance of Automated Scoring systems for the Field Test items
Hybrid scoring approaches	Continued read-behind scenarios	4	Score quality of scoring scenarios which combine human and Automated Scoring
	Targeting responses for human review	5	Methods to route responses for human review during Automated Scoring
Extending what can be scored	Item characteristics that correlate with agreement for handscoring and Automated Scoring	6	Investigating characteristics of items that can be scored accurately by Automated Scoring systems
	Generic scoring models	7	Development of Automated Scoring system to score Conventions trait across multiple essay prompts

Chapter 2: Summary of the Pilot Study Research Studies

The research conducted during the Pilot Study on behalf of the Smarter Balanced Assessment Consortium focused on the Automated Scoring of equation, short-text (both English language arts (ELA)/literacy and mathematics), and essay items. Out of a total of 5,412 Pilot Study items, 1,494 items were constructed-response items eligible for Automated Scoring. These items were handscored by two human raters; in case of disagreement the score was resolved by a third (senior) rater. Using the resolved human scores, nine vendors trained Automated Scoring systems to score (a subset of) the Pilot Study items. A wide range of Automated Scoring approaches were employed. Six special studies were conducted to investigate various aspects of Automated Scoring.

This chapter summarizes the main findings of the Pilot Study Research Studies to provide context for the studies conducted during the Field Test. For additional details on the methodology and the results of each study, the reader is referred to the full Smarter Balanced Pilot Automated Scoring Research Studies research report.

Performance of Automated Scoring Systems

A question central to the Pilot Study was: How well can Automated Scoring systems score the constructed-response items compared to the gold standard of the resolved human scores? Several criteria (see Table 3.1) were used to evaluate reliability and validity of scores assigned by Automated Scoring systems, including common agreement statistics such as exact agreement rates, quadratic weighted kappa, and standardized mean differences. The results can be summarized as follows:

- Out of a total of 348 equation items, 302 items (87%) met all criteria for Automated Scoring, 30 items (9%) needed additional review, and 16 items (5%) were not suited for Automated Scoring.
- Out of a total of 396 short-text items, 192 items (48%) met all criteria for Automated Scoring, 187 items (47%) needed additional review, and 17 items (4%) were not suited for Automated Scoring.
- Out of a total of 51 essay items/traits (three traits for each of 17 essays), 21 items/traits (41%) met all criteria for Automated Scoring while 30 items/traits (59%) needed additional review.

In previously reported research (e.g., the 2012 ASAP 2 competition; Shermis, 2013) on Automated Scoring of short-text items, the agreement between two human raters tended to be greater than the agreement between Automated Scoring systems and human raters. In the Pilot Study, the performance of Automated Scoring systems exceeded human inter-rater agreement for many short-text items. This may be an indication that Automated Scoring technology is improving.

To further the state-of-the-art in Automated Scoring, in particular in the context of large-scale operational assessments, several special studies were conducted, organized around three themes: (a) training, validating, and deploying Automated Scoring systems, (b) developing items for Automated Scoring, and (c) operational scoring using Automated Scoring systems. All nine vendors were invited to participate and all but two participated in at least one special study.

Train, Validate, and Deploy Automated Scoring Systems

Automated Scoring systems commonly require a set of handscored responses (training set) to develop a scoring model. A second, separate set of handscored responses (validation set) is usually required to evaluate the performance of the system. Given the cost of handscoring, a special study considered the size of training and validation sets in relation to scoring accuracy. Key findings were:

- About 500 responses was a reasonable lower limit required to train Automated Scoring systems for the Pilot Study items; 750 to 850 responses were necessary to achieve performance within one standard error of the performance achieved with 1,500 responses in the training set. Performance for essay items continued to improve from 1,000 to 1,500 responses in a training set, but short-text items did not require training sets with more than 1,000 responses.
- Uncertainty in quadratic weighted kappa estimates reduced by approximately 50% when the validation set size increased from 100 to 300 responses and by approximately 75% with 500 responses; validation sets with more than 500 responses may not have been efficient.

After training and validation, a large-scale assessment program can deploy Automated Scoring systems in several scenarios, for example, fully automated operational scoring or a combination of Automated Scoring and human scoring. A special study investigated various scoring scenarios. The main conclusions were:

- Score quality was highest when one human rater was paired with the best-performing Automated Scoring system with a third (human) rater adjudicating all disagreements. This produced substantially better quadratic weighted kappa than other studied scenarios. Note however that results may have been influenced by the fact that the first human rater contributed to the baseline score of record.
- Using a second Automated Scoring system to read-behind another Automated Scoring system did not produce agreement rates as high as the combination of a human and an Automated Scoring system. This may be due to the fact that the Automated Scoring systems were independently trained and therefore there may be converging. It may be worthwhile to investigate training two Automated Scoring systems specifically for a read-behind scenario.

Developing Items for Automated Scoring

From experience, about one-third to one-half of short-text items developed for an assessment program cannot be scored reliably by Automated Scoring systems. This drop-out rate increases the cost of item development for an assessment program using Automated Scoring. A special study investigated whether item characteristics can predict whether a short-text item can be scored automatically. The item characteristics were based on item metadata, item stimulus material, and scoring rubrics. Two of the results of this study were:

- There was a statistically significant decrease in performance of Automated Scoring systems for ELA/literacy items when there were many possible text-based key elements (four or more) or when Depth of Knowledge was high (Level 3). Performance was lower for difficult items, but this was not statistically significant.

- The statistical significance of the relationship between item characteristics and Automated Scoring performance could not be tested for mathematics items. However, the trends for the mathematics items were similar to the trends the for ELA/literacy items.

Operational Scoring

Various issues may impact the quality of scores produced by an Automated Scoring system during operational scoring. Responses may require human review, for example, when a response is unlike the responses used in training to the extent that this affects scoring accuracy. Chapter 6 investigated whether responses likely to require human scoring can be identified automatically. Three aspects were investigated:

- Outlier detection methods indicated that discrepancies between human and automated scores tended to occur for responses with atypical feature patterns.
- Three different methods (Logistic Regression, Support Vector Machines, and Random Forests) had little success in detecting which responses would receive a discrepant score by an Automated Scoring system.
- There is a weak relationship between discrepant scores and disagreement between different prediction models for essay items, but not for short-text responses.

Gaming can be defined as the deliberate addition, by an examinee, of construct-irrelevant features to a response in an attempt to increase a score. Similarly, examinees may plagiarize source documents included in the stimulus material of essay items. In either case, score validity may be compromised during operational scoring. Chapter 7 studied the susceptibility of Automated Scoring systems to several forms of gaming. Three results are noteworthy:

- Optimal gaming strategies (relative to all strategies considered) increased the score of low-scoring responses by almost 0.50 score points; on average, gaming increased the score by about 0.25 points.
- Gaming lowered the score of high-scoring responses, indicating that gaming strategies may be detrimental.
- Automated Scoring systems are still susceptible to gaming, but one Automated Scoring system was not affected by the gaming strategy that appends extra copies of the response. This demonstrates that Automated Scoring systems can be made resilient to gaming.

A special study investigated whether latent semantic analysis can be used to detect various instances of source-based plagiarism. A corpus of 95 student essays, 60 percent of which were plagiarized from Wikipedia to various degrees was used (Clough & Stevenson, 2011).The performance of the algorithm on a small set of test documents can be summarized as follows:

- Plagiarized documents were marked as most suspicious using document similarity when compared to the sources. Detection rates on the study corpus were high with few false positives.

Latent semantic analysis did not offer an advantage over basic document similarity in this study. This may be due to the fact that the plagiarism in the corpus was relatively easy to detect. Although a

detection system was relatively successful at identifying the instances of plagiarism in the study corpus, plagiarism is a complex issue that requires further research. For additional details, see the Smarter Balanced Pilot Automated Scoring Research Studies research report.

Chapter 3: Automated Scoring Evaluation

Williamson, Xi, and Breyer (2012) describe a framework based upon multiple statistics (referred hereafter as the Educational Testing Services [ETS] framework) that are used in combination to evaluate the quality of the scores assigned by an Automated Scoring system (also referred as automated scores or engine scores) in comparison to the human rater quality for each item. Engine scores are evaluated and compared to human scores on the item level in order to diagnose whether suboptimal results are due to (a) an engine's inability to reliably score student responses for a particular item or (b) attributes of an item's design that impede reliable scoring by humans (Higgins, 2013). Included in the ETS framework is the standardized mean difference (SMD) between the human scores and engine scores at both the population and subpopulation level. See Appendix 3.D for a brief description of these rater agreement statistics.

In 2011 CTB adopted the ETS framework, with two minor adjustments (see Table 3.1):

- Bridgeman (2013) noted that high agreement between two raters can occur when raters are truncating the rubric score range. CTB has found that an engine's quadratic weighted kappa (QWK) may be high even though the engine exact agreement rate in comparison to humans is low. In this situation, engines are usually giving adjacent scores to humans so that both the percent agreement and kappa statistics are not comparable to humans. For this reason, CTB also monitors engine performance for a notable reduction (greater than 0.05 difference) in perfect agreement rates between the human-human and engine-human scores.
- Williamson, Xi, and Breyer (2012) flag the SMD if the difference between automated scores and human scores is greater than 0.15 in absolute value. Similarly, they flag the SMD for a subgroup if the difference between automated scores and human scores for that subgroup is greater than 0.10 in absolute value. Because the larger the population SMD value the more likely the subpopulation SMD value will be flagged, CTB reduced the amount of SMD separation tolerated by flagging the population SMD if it exceeds 0.12 in absolute value.

The framework with the adjustments was used to evaluate the performance of Automated Scoring in the Smarter Balanced Pilot Study. The ETS framework is described in depth by Williamson, Xi, and Breyer (2012), Ramineni and Williamson (2013), and Higgins (2013), as well as in the Pilot Automated Scoring Research Studies report. Interested readers should refer to these publications and manuscripts for background on the framework.

Data Sources

The results of Automated Scoring were evaluated for all items that received full engine training and validation processes as set forth in Table 1.2. For a sample of 21 English language arts (ELA)/literacy short-text, constructed-response items and 21 mathematics short-text, constructed-response items, four Automated Scoring systems were trained. For 41 mathematics short-text, constructed response items related to mathematical reasoning, two Automated Scoring systems were trained. For the remaining items, one Automated Scoring system was trained. For the ELA/literacy essay items, one of the Automated Scoring systems did participate in the training but did not deliver scores for the validation sets by the end of the scoring period. Thus, this Automated Scoring system is not included in the evaluation. For a sample of 21 ELA/literacy essay items, three

Automated Scoring systems were trained. For the remaining items, two Automated Scoring systems were trained.

The engine-human SMD value was calculated on the total validation sample and for each subgroup of 100 students or more. For many subgroups, however, insufficient numbers of students were available in the validation sample to calculate the SMD. Outside the scope of the studies presented in this report, given the data sources available and our proposal response, was the relationship of Automated Scores to external measures and to indices based on students reported test scores.

Table 3.1. Statistical Criteria for the Evaluation of Automated Scoring Systems

Flagging Criterion	Flagging Threshold
Quadratic weighted kappa for engine score and human score	Quadratic weighted kappa less than 0.70
Pearson correlation between engine score and human score	Correlation less than 0.70
Standardized difference between engine score and human score	Standardized difference greater than 0.12 in absolute value
Degradation in quadratic weighted kappa or correlation from human-human to engine-human	Decline in quadratic weighted kappa or correlation equal to or greater than 0.10
Standardized difference between engine score and human score for a subgroup	Standardized difference greater than 0.10 in absolute value
Notable reduction in perfect agreement rates from human-human to engine-human	Decline equal to or greater than 0.05

Reporting Refinements for the Field Test

Chapter 2 of the Pilot Automated Scoring Research Studies report presents a detailed report on the performance of the Automated Scoring systems for the different item types. Items were broadly classified into three categories (Meets all criteria for Automated Scoring; Needs additional review; Not suited for Automated Scoring) based on the number of flags for the best-performing Automated Scoring system (0 flags, 1–7 flags, 8 or more flags, respectively). The results can be summarized as follows:

- Out of a total of 348 equation items, 302 items (87%) met all criteria for Automated Scoring, 30 items (9%) needed additional review, and 16 items (5%) were not suited for Automated Scoring.
- Out of a total of 396 short-text items, 192 items (48%) met all criteria for Automated Scoring, 187 items (47%) needed additional review, and 17 items (4%) were not suited for Automated Scoring.

- Out of a total of 51 essay items/traits (three traits for each of 17 essays), 21 items/traits (41%) met all criteria for Automated Scoring while 30 items/traits (59%) needed additional review.

A relatively large percentage of short-text and essay items (47% and 59%, respectively) were categorized as needing additional review. Although the Smarter Balanced Automated Scoring Research Studies report provides a detailed analysis of these items, additional information on the summary level would be desirable. To this end, we implemented a refinement in the reporting of the results, while keeping the statistical criteria in the framework the same.

The statistical criteria in Table 3.1 can be divided into three broad categories: evaluated against the final human scores of record, evaluated against the inter-rater performance of the two initial human raters, and evaluated for the performance in different subgroups. Table 3.2 categorizes the criteria from Table 3.1 into these three classes:

- Criteria in the first category evaluate the performance of an Automated Scoring system by comparing the scores assigned by the Automated Scoring system against the final human scores of record (referred as engine-human performance)
- Criteria in the second category evaluate the performance of an Automated Scoring system by comparing engine-human performance against the inter-rater performance of two initial human raters.
- Finally, criteria in the third category evaluate performance of an Automated Scoring system using engine-human performance (that is, statistics from the first category) for different subgroups. The number of responses in a validation set may not be sufficient to evaluate subgroup performance.

Note the difference between the evaluation criteria in the first and second category. For the first category, the scores assigned by the Automated Scoring system are compared against the final human scores of record. For the second category, statistics from the first category are compared against performance of the two human raters (inter-rater agreement). In other words, evaluation of the criteria of the second category should be subsequent to evaluation of the criteria in the first category. Hence, one could argue that these three categories constitute a hierarchy. For example, if an Automated Scoring system does not meet the performance criteria for the entire population, then evaluating its performance on subgroups may be less relevant. Therefore, we reported not only the total number of flags, but also the number of flags in each of the three categories. Note that we reported subgroup performance criteria even when an Automated Scoring system does not meet the absolute or relative criteria.

Besides reporting flags for the Automated Scoring systems and the distribution of the statistics underlying the flagging criteria across items, we also reported the number of flags (total and in each category) for the human ratings (that is, based on the inter-rater comparison of the first and second human rater). Information on the number of flags received by human raters can be combined with the information on the Automated Scoring system to potentially diagnose different types of problems.

For example, if the human raters received fewer flags than the Automated Scoring systems, then this is an indication that the scoring models developed for that item need to be improved. On the other hand, if the human raters received many flags as well, then it could be that performance of the Automated Scoring systems was limited by the quality of the human raters. In that case, perhaps the scoring rubric needs to be reevaluated.

Finally, we refined the summary designations based on the flagging criteria as presented in the Table 3.3. This provided some additional information on type of item review that may be needed before an item can be considered for Automated Scoring. The different criteria in Table 3.3 were evaluated sequentially, in line with the hierarchical nature of the criteria. This ensured that each item was counted in only one reporting category. The reader should note, however, that when an item is designated as needing review for performance relative to final human score, for instance, this does not mean that the item should not also be reviewed for other criteria, for example, for performance for subgroups.

Table 3.2. Grouped Evaluation Criteria

Category	Flagging Criterion	Flagging Threshold
Performance evaluated by comparing against final human score of record	Quadratic weighted kappa for engine score and human score	Quadratic weighted kappa less than 0.70
	Pearson correlation between engine score and human score	Correlation less than 0.70
	Standardized difference (SMD) between engine score and human score	Standardized difference greater than 0.12 in absolute value. (Flags will be separately reported for items with SMD > 0.12 and with SMD < -0.12).
Performance evaluated by comparing against inter-rater agreement of the two initial human raters	Degradation in quadratic weighted kappa from human-human to engine-human	Decline in quadratic weighted kappa greater than or equal to 0.10
	Degradation in correlation from human-human to engine-human	Decline in correlation greater than or equal to 0.10
	Notable reduction in perfect agreement rates from human-human to engine-human	Decline greater than or equal to 0.05
Performance evaluated for different subgroups (when sample size is sufficient).	Standardized difference between engine score and human score for a subgroup	Standardized difference greater than 0.10 in absolute value

Table 3.3. Summary Designations

Flagging Criterion	Summary Designation
No flags	Meets all criteria
No flags from the first category, but one or two flags from the second category	Needs additional review for performance relative to inter-rater agreement
No flags from the first category, but three flags from the second category	Not suited for Automated Scoring
No flags from the first or second category, but one or more flags from the third category	Needs additional review for subgroup performance
One flag from the first category	Needs additional review for performance relative to final human score
Two or more flags from the first category	Not suited for Automated Scoring
Eight or more flags (total)	Not suited for Automated Scoring

Automated Scoring of Short-Text Items

The results are presented in the following order:

- 21 ELA/literacy short-text, constructed response, items for which four Automated Scoring systems were trained and the remaining 45 ELA/literacy short-text, constructed response items for which one Automated Scoring system was trained.
- 41 mathematics short-text, constructed response, items developed to test mathematical reasoning, for which two Automated Scoring systems were trained.
- 21 mathematics short-text, constructed response items for which four Automated Scoring systems were trained and the remaining 176 mathematics short-text, constructed response, items for which one Automated Scoring system was trained.

Table 3.4 presents the summary results for the ELA/literacy short-text, constructed-response, items for the human inter-rater agreement, the top performing Automated Scoring system for each item, the four Automated Scoring systems trained for 21 items, and the human inter-rater agreement and one Automated Scoring system trained for the remaining items.

Details for each item are presented in Appendix 3.A: inter-rater agreement between the two human raters for 21 items (Table 3.A.1), the top-performing Automated Scoring system for each of the 21 items (Table 3.A.2), Vendor 2 (Table 3.A.3), Vendor 3 (Table 3.A.4), Vendor 6 (Table 3.A.5), Vendor 9 (Table 3.A.6), inter-rater agreement between the two human raters for the remaining items (Table 3.A.7) and Vendor 3 for the remaining items (Table 3.A.8).

Table 3.4. Average Rater Agreement Statistics, ELA/literacy—Short-text Items

Rater	SMD	Agreement Statistics				Flags			
		QWK	Correlation	Percent Agree	Percent Adjacent & Agree	Total	Absolute	Relative	Subgroup
H1H2 (21 items)	-0.02	0.69	0.70	79	99	1.76	1.19	0.00	0.57
Top	0.01	0.75	0.76	84	100	0.81	0.33	0.05	0.43
Vendor 2	0.04	0.74	0.75	82	100	1.57	0.52	0.05	1.00
Vendor 3 (21 items)	0.06	0.75	0.75	83	100	1.81	0.48	0.05	1.29
Vendor 6	0.02	0.73	0.74	82	100	2.00	0.62	0.00	1.38
Vendor 9	-0.06	0.73	0.74	84	100	2.91	0.91	0.00	2.00
H1H2 (remaining items)	0.00	0.69	0.69	82	99	1.47	1.02	0.00	0.44
Vendor 3 (remaining items)	0.06	0.74	0.75	85	100	2.24	0.65	0.04	1.55

Note: H1H2 = Human-human inter-rater agreement. Top = Top-performing Automated Scoring system.

Out of 665 ELA/literacy short-text, constructed-response, items, human-human inter-rater agreement meets all criteria for 268 items (40%), requires review for absolute performance for 6 items (1%), requires review for subgroup performance for 55 items (8%), and does not meet performance criteria for 336 items (51%).

Based on the top-performing Automated Scoring system, 261 items (39%) meet all criteria for Automated Scoring, 47 items (7%) require review for absolute performance, 4 items (1%) require review for performance relative to inter-rater agreement, 183 items (28%) percent require review for subgroup performance, and 170 items (26%) are not suited for Automated Scoring.

Table 3.5. Average Rater Agreement Statistics, Mathematical Reasoning and Mathematics—Short-text Items

Rater	SMD	Agreement Statistics				Flags			
		QWK	Correlation	Percent Agree	Percent Adjacent & Agree	Total	Absolute	Relative	Subgroup
H1H2 (MR)	0.00	0.81	0.81	93	100	0.61	0.34	0.00	0.27
Top (MR)	0.00	0.74	0.75	93	99	1.44	0.63	0.29	0.51
Vendor 8 (MR)	-0.09	0.64	0.67	92	99	4.24	1.17	0.59	2.49
Vendor 3 (MR)	0.01	0.74	0.75	93	99	1.68	0.68	0.32	0.68
H1H2 (21 items)	0.00	0.86	0.86	92	100	0.24	0.19	0.00	0.05
Top (21 items)	0.00	0.87	0.87	93	100	0.14	0.10	0.00	0.05
Vendor 2 (21 items)	0.01	0.85	0.85	92	100	0.81	0.19	0.14	0.48
Vendor 3 (21 items)	0.02	0.84	0.84	92	100	0.62	0.24	0.05	0.33
Vendor 6 (21 items)	0.00	0.83	0.84	92	100	1.05	0.29	0.14	0.62
Vendor 9 (21 items)	-0.04	0.84	0.84	93	99	1.24	0.19	0.14	0.91
H1H2 (remaining items)	0.00	0.85	0.86	94	100	0.28	0.18	0.00	0.09
Vendor 3 (remaining items)	0.02	0.82	0.82	93	100	0.95	0.31	0.19	0.45

Note: MR = Mathematical reasoning. H1H2 = Human-human inter-rater agreement. Top = Top-performing Automated Scoring system.

Table 3.5 presents the summary results for the mathematical reasoning and other mathematics short-text, constructed-response, items for the human inter-rater agreement, the top performing Automated Scoring system for each item, the four Automated Scoring systems trained for 21 items, and the human inter-rater agreement and one Automated Scoring system trained for the remaining items.

Details for each item are presented in Appendix 3.B: inter-rater agreement between the two human raters for 21 items (Table 3.B.1), the top-performing Automated Scoring system for each of the 21 items (Table 3.B.2), Vendor 2 (Table 3.B.3), Vendor 3 (Table 3.B.4), Vendor 6 (Table 3.B.5), Vendor 9 (Table 3.B.6), inter-rater agreement between the two human raters for the remaining items (Table 3.B.7) and Vendor 3 for the remaining items (Table 3.B.8). For the mathematical reasoning items, Table 3.B.9 presents inter-rater agreement between the two human raters, Table 3.B.10 the results for the top-performing Automated Scoring system for each item, Table 3.B.11 the results for Vendor 3, Table 3.B.12 the results for Vendor 8.

Out of 41 mathematics short-text, constructed-response, items developed to test mathematical reasoning, human-human inter-rater agreement meets all criteria for 29 items (71%), requires review for subgroup performance for 5 items (12%), and does not meet performance criteria for 7 items (17%).

Based on the top-performing Automated Scoring system, 22 items (54%) meet all criteria, 2 items (5%) require review for absolute performance, 2 items (5%) require review for relative performance, 4 items (10%) require review for subgroup performance, and 11 items (27%) are not suited for Automated Scoring.

Out of the remaining 192 mathematics items, human-human inter-rater agreement meets all criteria for 164 items (85%), requires review for absolute performance for 1 item (1%), requires review for subgroup performance for 10 items (5%), and does not meet performance criteria for 17 items (9%).

Based on the top-performing Automated Scoring system, 139 items (74%) meet all criteria, 3 items (2%) require review for absolute performance, 9 items require review for performance relative to inter-rater agreement, 12 items (6%) require review for subgroup performance, 24 items (25%) are not suited for Automated Scoring due to absolute performance and 4 items (2%) are not suited due to relative performance.

Automated Scoring of Essay Items

The functioning of 66 ELA/literacy essay items was evaluated. Each essay response was scored for three trait scores, namely trait A: Organization/Purpose; trait B: Evidence/Elaboration; and trait C: Conventions (referred hereafter as trait A, B, and C). Thus, this study evaluated the functioning of 198 essay item/trait combinations. The essay items consisted of 66 item/traits worth 2 points (scored 0–2) and 132 item/traits worth 4 points (scored 0–4). According to the scoring rubric, some responses were assigned a non-numeric code ("condition code"), for example, when the response was off-topic. Any condition code was recoded as 0 for the purpose of the Automated Scoring evaluation.

The results are presented in the following order:

- 21 essay items for which three Automated Scoring systems were trained (a fourth Automated Scoring system was trained but did not score the validation sets).
- 45 essay items for which two Automated Scoring systems were trained.

A top-performing Automated Scoring system was identified for each item/trait based on the total number of flags and the QWK values from the agreement between the Automated scores and the final human score of record.

Table 3.6. Average Rater Agreement Statistics, Essay—21 Items

Trait	Rater	SMD	Agreement Statistics				Flags			
			QWK	Correlation	Percent Agree	Percent Adjacent & Agree	Total	Absolute	Relative	Subgroup
A	H1H2	0.00	0.83	0.83	77	99	0.19	0.10	N/A	0.10
A	Top	0.02	0.82	0.82	76	98	0.71	0.19	0.24	0.29
A	Vendor 1	0.04	0.83	0.84	76	99	1.57	0.33	0.19	1.05
A	Vendor 3	0.03	0.83	0.83	77	98	0.86	0.19	0.24	0.43
A	Vendor 8	0.00	0.78	0.78	71	97	1.81	0.43	0.67	0.71
B	H1H2	0.00	0.82	0.82	76	98	0.38	0.19	N/A	0.19
B	Top	0.03	0.84	0.84	77	99	0.95	0.14	0.24	0.57
B	Vendor 1	0.04	0.83	0.83	76	99	1.33	0.19	0.29	0.86
B	Vendor 3	0.03	0.83	0.83	76	98	1.14	0.24	0.24	0.67
B	Vendor 8	0.01	0.77	0.77	70	97	2.14	0.52	0.67	0.95
C	H1H2	0.00	0.71	0.71	72	97	1.62	1.24	N/A	0.38
C	Top	0.01	0.74	0.74	74	98	0.76	0.38	0.10	0.29
C	Vendor 1	0.05	0.74	0.74	73	98	2.67	0.62	0.29	1.76
C	Vendor 3	0.00	0.74	0.74	74	98	1.00	0.43	0.14	0.43
C	Vendor 8	0.00	0.61	0.62	64	95	6.33	2.19	1.10	3.05

Note: H1H2 = Human-human inter-rater agreement. Top = Top-performing Automated Scoring system.

Table 3.6 presents the summary results for the human inter-rater agreement, the top performing Automated Scoring system for each item, and the three Automated Scoring systems trained for these 21 items. Details for each item are presented in Appendix 3.C: inter-rater agreement between the two human raters for 21 items (Table 3.C.1), the top-performing Automated Scoring system for each of the 21 items (Table 3.C.2), followed by the performance of Vendor 1 (Table 3.C.3), Vendor 3 (Table 3.C.4), and Vendor 8 (Table 3.C.5) on 21 items.

Chapter 3: Automated Scoring Evaluation

For traits A and B, average performance of the top Automated Scoring system for each item, Vendor 1, and Vendor 3 as measured by the agreement statistics was comparable to human inter-rater agreement, while Vendor 8 had slightly lower average performance. On average, standardized mean difference of the Automated Scoring systems was slightly above 0 but less than 0.05, with Vendor 8 posting the best SMD, followed by Vendor 3, and finally Vendor 1.

For trait C, average performance of the top Automated Scoring system for each item, Vendor 1 and Vendor 3, as measured by the agreement statistics, exceeded human inter-rater agreement, while Vendor 8 had lower average performance. On average, standardized mean difference of Vendor 3 and Vendor 8 was equal to SMD of the human raters, while Vendor 1 posted a SMD exceeding 0.05.

Based on the inter-rater agreement between two human raters, the following number of items were flagged:

- Trait A: 18 items (86%) meet all criteria, 2 items (10%) need review for subgroup performance, and 1 item (5%) did not meet performance criteria.
- Trait B: 17 items (81%) meet all criteria, 2 items (10%) need review for subgroup performance, and 2 items (10%) did not meet performance criteria,
- Trait C: 7 items (33%) meet all criteria, 1 item (5%) needs review for subgroup performance, and 13 items (62%) did not meet performance criteria.

Based on the top-performing Automated Scoring system for each item, the following number of items were flagged:

- Trait A: 14 items (67%) meet all criteria, 3 items (14%) need review for performance relative to inter-rater agreement, 2 items (10%) need review for subgroup performance, and 2 items (10%) are not suited for Automated Scoring.
- Trait B: 14 items (67%) meet all criteria, 3 items (14%) need review for performance relative to inter-rater agreement, 2 items (10%) need review for subgroup performance, and 2 items (10%) are not suited for Automated Scoring.
- Trait C: 13 items (62%) meet all criteria, 3 items (14%) %) need review for performance relative to inter-rater agreement, and 5 items (24%) are not suited for Automated Scoring.

Table 3.7. Average Rater Agreement Statistics, Essay—45 Items

Trait	Rater	SMD	Agreement Statistics				Flags			
			QWK	Correlation	Percent Agree	Percent Adjacent & Agree	Total	Absolute	Relative	Subgroup
A	H1H2	0.00	0.86	0.86	77	99	0.13	0.04	0.00	0.09
A	Top	0.03	0.86	0.86	77	99	0.69	0.07	0.22	0.40
A	Vendor 1	0.04	0.85	0.85	76	99	0.98	0.11	0.33	0.53
A	Vendor 3	0.03	0.86	0.86	77	99	0.78	0.07	0.22	0.49
B	H1H2	0.01	0.85	0.85	77	99	0.13	0.04	0.00	0.09
B	Top	0.02	0.86	0.86	78	99	0.47	0.04	0.24	0.18
B	Vendor 1	0.04	0.84	0.85	76	99	0.98	0.11	0.36	0.51
B	Vendor 3	0.03	0.86	0.86	78	99	0.78	0.07	0.24	0.47
C	H1H2	0.01	0.73	0.73	73	97	1.31	1.07	0.00	0.24
C	Top	0.02	0.77	0.77	76	98	0.71	0.18	0.24	0.29
C	Vendor 1	0.04	0.76	0.76	74	98	1.78	0.27	0.31	1.20
C	Vendor 3	0.00	0.77	0.77	76	98	1.13	0.22	0.27	0.64

Note: H1H2 = Human-human inter-rater agreement. Top = Top-performing Automated Scoring system.

Table 3.7 presents the summary results for the human inter-rater agreement, the top performing Automated Scoring systems, and the two Automated Scoring systems trained for the remaining 45 items. Details for each item are presented in Appendix 3.C: inter-rater agreement between the two human raters (Table 3.C.6), the top-performing Automated Scoring system for each item (Table 3.C.7), Vendor 1 (Table 3.C.8), and Vendor 3 (Table 3.C.9).

For traits A and B, average performance of the top Automated Scoring system for each item, Vendor 1 and Vendor 3, as measured by the agreement statistics, was comparable to human inter-rater agreement. On average, standardized mean difference of the Automated Scoring systems was slightly above 0 but less than 0.05, with Vendor 3 posting better SMD than Vendor 1.

For trait C, average performance of the top Automated Scoring system for each item, Vendor 1 and Vendor 3, as measured by the agreement statistics, exceeded human inter-rater agreement. On

average, standardized mean difference of Vendor 3 was equal to the SMD of the human raters, while Vendor 1 posted a SMD exceeding 0.04.

Based on the inter-rater agreement between two human raters, the following number of items were flagged:

- Trait A: 42 items (93%) meet all criteria, 2 items (4%) need review for subgroup performance, and 1 item (2%) did not meet performance criteria.
- Trait B: 42 items (93%) meet all criteria, 2 items (4%) need review for subgroup performance, and 1 item (2%) did not meet performance criteria.
- Trait C: 19 items (42%) meet all criteria, 2 items (4%) need review for absolute performance, 1 item (2%) needs review for subgroup performance, and 23 items (51%) did not meet performance criteria.

Based on the top-performing Automated Scoring system for each item, the following number of items were flagged:

- Trait A: 30 items (67%) meet all criteria, 10 items (22%) need review for performance relative to inter-rater agreement, 4 items (9%) need review for subgroup performance, and 1 item (2%) is not suited for Automated Scoring.
- Trait B: 31 items (69%) meet all criteria, 9 items (20%) need review for performance relative to inter-rater agreement, 3 items (7%) need review for subgroup performance, and 2 items (4%) are not suited for Automated Scoring.
- Trait C: 28 items (62%) meet all criteria, 5 items (11%) need review for performance relative to inter-rater agreement, 5 items (11%) need review for subgroup performance, and 7 items (16%) are not suited for Automated Scoring.

Automated Scoring Engine Descriptions

Automated Scoring Engine Technical Description: Luis Tandalla (ASAP 1) (Open-Source Engine)

A. Changes or updates from the Pilot Study:

- The current engine uses the same methods and models for response cleaning, dictionary building, feature selection, training, and ensembling that were described in the Pilot Study. No major changes were made to the overall system.
- There were several updates to the source code that improved the organization and readability, but the structure behind the system is the same.
- Different values were tested for the different parameters of the engine using cross-validation, and the combination that produces the highest quadratic weighted kappa was chosen. Fewer values were tested for the current engine than for the Pilot Study, so that the training time is reduced. The overall performance of the model is expected to be approximately the same.

B. Lessons learned and road blocks.

The following table shows the performance of the engine compared to a human rater:

Content	Item Group	Quadratic Weighted Kappa		
		Engine-Human	Human-Human	Absolute Difference
ELA/literacy	46282	0.57	0.74	-0.17
ELA/literacy	46348	0.79	0.82	-0.03
ELA/literacy	46450	0.55	0.57	-0.02
ELA/literacy	46517	0.77	0.77	0.00
ELA/literacy	48247	0.83	0.80	0.03
ELA/literacy	50932	0.68	0.74	-0.06
ELA/literacy	51416	0.78	0.76	0.02
ELA/literacy	46115	0.75	0.63	0.12
ELA/literacy	46121	0.72	0.73	-0.01
ELA/literacy	46203	0.74	0.79	-0.05

Content	Item Group	Quadratic Weighted Kappa		
		Engine-Human	Human-Human	Absolute Difference
ELA/literacy	49151	0.59	0.52	0.07
ELA/literacy	50577	0.88	0.97	-0.09
ELA/literacy	50786	0.53	0.44	0.09
ELA/literacy	50868	0.72	0.59	0.13
ELA/literacy	53040	0.86	0.82	0.04
Mathematics	45535	1.00	0.99	0.01
Mathematics	48558	0.92	0.96	-0.04
Mathematics	48560	0.94	0.89	0.05
Mathematics	49790	0.94	0.90	0.04
Mathematics	52909	0.83	0.78	0.05
Mathematics	46597	0.70	0.73	-0.03
Mathematics	46619	0.88	0.83	0.05
Mathematics	46793	0.93	0.94	-0.01
Mathematics	51802	0.65	0.70	-0.05
Mathematics	53299	0.64	0.59	0.05

- The performance of the engine is similar to a human rater's performance for most of the items. In general, the engine performs better for mathematics items. For ELA/literacy items, the engine performance differs considerably from a human rater's performance, such as with item 46282.
- The engine rates the responses mostly based on the content. The reason for the difference in performance with ELA/literacy and mathematics items could be that ELA/literacy responses are slightly more open ended than mathematics responses. The lexicon used for those responses are more broad, so it is hard for the engine to learn all the possible words for all the possible correct responses.

C. Suggestions for future research.

- A possible method to improve the performance for ELA/literacy responses could be to treat those responses as long answer essays. An engine that grades based on content and also on style may be able to perform better on ELA/literacy responses. In the current engine, adding additional features, such as labeling each word with part of speech tags, may add the capability of rating style in the responses.

Automated Scoring Engine Technical Description: Jure Žbontar (ASAP 2) (Open-Source Engine)

A. Changes or updates from the Pilot Study:

- This scoring engine is based on machine learning. Five models (Ridge Regression, Support Vector Regression, Gradient Boosting Machines, k-Nearest Neighbors, and Random Forest Regression) were trained with different hyperparameter settings and different preprocessing methods to obtain 30 trained models. After training, the predictions from the individual models were combined with Ridge Regression to obtain the final score. For details on the models and dataset construction, see *Short Answer Scoring by Stacking* (Žbontar, 2012).
- An advantage of using machine learning is that the rules used to score the answers are automatically inferred from the data, rather than coded by hand, which means that relatively little had to be changed from the Pilot Study. For the current study, it was necessary to retrain the models on the 21 response items for this study. New models were also added to the ensemble in order to improve the accuracy from the Pilot Study. The models that gave the overall best score on internal 5-fold cross-validation are shown in the table below.

Name	Description
1w linr	Ridge regression on 1w data
4c linr	Ridge regression on 4c data
6c linr	Ridge regression on 6c data
4cc linr	Ridge regression on 4cc data
6cc linr	Ridge regression on 6cc data
4cp200 linr	Ridge regression on 4cp200 data
4cp500 linr	Ridge regression on 4cp500 data
1w gbr 3 1	Gradient boosting (max depth=3, max features=1) on 1w data
1w gbr 4 0.5	Gradient boosting (max depth=4, max features=0.5) on 1w data
4c gbr 3 1	Gradient boosting (max depth=3, max features=1) on 4c data
4c gbr 4 0.5	Gradient boosting (max depth=4, max features=0.5) on 4c data
4c gbr 5 0.5	Gradient boosting (max depth=5, max features=0.5) on 4c data

Name	Description
4cc gbr 3 1	Gradient boosting (max depth=3, max features=1) on 4cc data
4cc gbr 3 1	Gradient boosting (max depth=3, max features=1) on 4cc data
4cc gbr 4 0.5	Gradient boosting (max depth=4, max features=0.5) on 4cc data
4cc gbr 5 0.5	Gradient boosting (max depth=5, max features=0.5) on 4cc data
4cp200 gbr 3 1	Gradient boosting (max depth=3, max features=1) on 4cp200 data
4cp500 gbr 3 1	Gradient boosting (max depth=3, max features=1) on 4cp500 data
4c libsvm	Support vector machine on 4c data
6c libsvm	Support vector machine on 6c data
4cc libsvm	Support vector machine on 4cc data
6cc libsvm	Support vector machine on 6cc data
4cp200 libsvm	Support vector machine on 4cp200 data
4cp500 libsvm	Support vector machine on 4cp500 data
4c rf sparse	Random forest on sparse 4c data
4cc rf sparse	Random forest on sparse 4cc data
4c knn	k-nearest neighbor on 4c data
6c knn	k-nearest neighbor on 6c data
4cc knn	k-nearest neighbor on 4cc data
4cp200 knn	k-nearest neighbor on 4cp200 data
4cp500 knn	k-nearest neighbor on 4cp500 data

Note: For the description of the different preprocessing methods (1w, 4c, 6c, 4cc, 6cc, 4cp200, and 4cp500) see *Short Answer Scoring by Stacking* (Žbontar, September 2012).

B. Lessons learned and road blocks:

- Use machine learning to build automated scoring engines. With the availability of large datasets it is now possible to build systems that automatically infer scoring rules from training data, instead of relying on human hand-engineering. These systems are cheaper to build, can automatically adapt to new domains, and can outperform other approaches,

as was demonstrated on the ASAP competition where all the top performing entries used some form of machine learning. Systems built with this approach will only get better as bigger datasets become available to train them.

- Instead of training a single model, a better approach is to train an ensemble of models and combine their predictions. It is often the case that the accuracy of the combined model is better than the accuracy of the best model in the ensemble. While this method has been around for more than 20 years (Wolpert, 1992), it tends to be underrated. This is how almost all machine learning competitions today are won; the most famous one arguably being the $1 million Netflix prize.

C. Suggestions for future research:

- Use more models in the ensemble. One particular model to include is Deep Convolutional Neural Networks, which have been beating previous state-of-the-art results in many different tasks (e.g. object classification, image segmentation, and stereo vision). It would be interesting to see how these models could be applied to the problem of automated scoring.
- Experiment with different dataset representations. Each response is currently represented as a bag of character 4-grams and 6-grams. Recently, there has been some interesting work on text representation (Mikolov et al., 2013). Since the source code of word2vec is available online, it should be easy to determine whether the new representation would increase accuracy.
- Change the training procedure to improve the accuracy of the ensemble. When the final prediction is an average of the predictions of many models, it helps if the individual models overfit the training data (i.e. have low bias and high variance). A simple example is the Random Forest classifier which works better with unpruned decision trees, even though the accuracy of the individual trees is higher if they are pruned.
- Choose models that overfit by running the original training procedure and then modifying the hyper-parameters, e.g. for Ridge Regression and Support Vector Regression, decrease the value of the regularization parameter, for k-Nearest Neighbors, increase the value of k, etc.
- Learn a single model to output the same predictions as the ensemble in order to decrease the resources necessary to run the model at test time. When the program is asked to score a response it must run all 30 models in the ensemble, which can be time consuming. It has recently been demonstrated by Hinton (unpublished work, https://www.youtube.com/watch?v=EK61htlw8hY) that it is possible to build a single model that has comparable accuracy to the ensemble. The advantage is lower running time and ease of deployment.

Automated Scoring Engine Technical Description: Pawel Jankiewicz (ASAP 5)
(Open-Source Software)

D. Changes or updates from the Pilot Study and Lessons learned and road blocks:

- Structure of the models.

  In the prior version, the scores were modeled as regression task. The scores were predicted using ordinal grading system. The prediction was a real number which was then transformed to a discrete grade using "grade on a curve" transformation—meaning that the share of each grade was calculated in all observations in the training observations and then applied to new data.

  The previous approach, while giving satisfactory results, is very biased. It evaluates the essays using previous distribution of grades. This can lead to untrue evaluation if the new essays come from a different distribution.

  For this reason and also because for this study, the models had to predict the probability of each score, it was decided to change the structure of the models. To create a model for n scores, an n of independent models were used. Whichever grade had the highest probability was chosen as a final score. It also made blending of the models more obvious.

- Choice of the models

  Apart from using *sklearn* GradientBoostingClassifier (http://scikitlearn.org/stable/modules/generated/sklearn.ensemble.GradientBoostingClassifier.html), it was decided to try:

  - a new boosting library xgboost (https://github.com/tqchen/xgboost)
  - a deep learning library nolearn (https://github.com/dnouri/nolearn)

  xgboost (https://github.com/tqchen/xgboost) is a remarkable implementation of a tree boosting algorithm. It is unique because it allows training trees using many cores.

- Data preprocessing

  Mathematics

  For this study, 50% of the essays were mathematics items. Much time was spent trying to clean the mathematics responses to normalize them in every possible way.

  text to math

  text_to_math is a new module to convert any numbers written in English to a digit representation number.

  Some examples of transformations:

  1. replaces one with 1, two with 2, ...
  2. replaces 2 out of 4 with 2/4

3. replaces thirds with 1/3, fourths with 1/4

Formulas simplifications

After mathematics cleaning it is possible to evaluate the mathematics expressions in the response. Counting the number of true expressions is a powerful feature.

2 PLUS 2 EQUALS 4 -> 4=4

2/2 PLUS 1 EQUALS 2 -> 2=2

2 TIMES 2 PLUS 2 EQUALS 7 -> 6=7

NO BECAUSE 6 AN 6 = 12 IN ADDITRON -> NO BECAUSE 12=12 IN ADDITRON

6 + 6 = 12 AND 9 + 9 = 18 -> 12=12 AND 18=18

- **Pipeline**

  The code needed to transform the data was refactored. Pipelines (pipelines.py) are now lists of steps needed to go from the raw essay text to features.

  A simple pipeline created three feature sets:

  - length of original essay text
  - length of stemmed essay text
  - 1-, 2-, and 3-grams for stemmed essay text

  Each pipeline defined this way can be applied to a response collection. Eight different pipelines were created, which enabled testing different features.

- **New features**
  - **wikipedia *n*-gram coverage.** The pre-calculated 1-, 2-, 3-grams from Wikipedia were downloaded. A coverage of the n-grams in the essay was used as a feature. For example if all 2-grams in the essay were seen on Wikipedia, that is a score of 1.0. If all 2-grams are novel, that is a score 0.0. This gives models some valuable information about the style of the essay.
  - **sentiment analysis.** Using *TextBlob* (https://github.com/sloria/TextBlob) library, additional features likesentiment_polarity and sentiment_subjectivity were calculated. Both features turned out to be quite useful.
  - **character based *n*-grams.** Apart from standard word-based *n*-grams. It was decided to add character based *n*-grams (from 1 up to 4-grams in some pipelines). This proved to be very valuable transformation.
  - **language probability.** To recognized texts written in different languages, a Python library *lang.id* (https://github.com/saffsd/langid.py) was used. The probability of the language being English, French and Spanish was calculated.
  - **word2vec clusters.** *word2vec* (https://code.google.com/p/word2vec/) is a powerful library to calculate word embeddings. Each word is represented as a vector of *n* real numbers. Pre-computed vectors from news articles were used. A quote from the https://code.google.com/p/word2vec/:

"We are publishing pre-trained vectors trained on part of Google News dataset (about 100 billion words). The model contains 300-dimensional vectors for 3 million words and phrases."

It was decided to cluster semantically similar words into clusters. The number of clusters was chosen to be **d** / 4. Where **d** is the number of unique words in the essay dictionary (bag of words). The reason for this was to force to group sparse features into more dense representation.

- Model ensemble

  In total, 107 different model types and parameters were generated. For the most difficult essays, additional models were generated. Altogether there were 2,219 models.

  To merge the results of the models, direct optimization of the weighted kappa metric was tried. Unfortunately this resulted in very poor generalization. The coefficients began to over fit to the training data. To limit its learning potential, the regularization procedure shown below was used:

```
repeat 40 times
     sample 50% of the training observations
     optimize weighted kappa on the sample
     save coefficients
calculate average coefficients for each model
```

  This resulted in better generalization.

Automated Scoring Engine Technical Description: TurnItIn/LightSide Labs (Open-Source Software)

A. Changes or updates from the Pilot Study:

- There is not a great deal of difference between the engine that was used on the Pilot Study and the one used for the Field Test. Techniques have been added for optimizing essay and short answer models, as well as for improving the speed with which the optimum model parameters are found. In addition, various changes and de-bugging have been made to the Researcher's Workbench since the Pilot Study.

B. Lessons learned and road blocks:

- Adding word and part-of-speech trigrams improves model performance, at the cost of processing time. The models generated during the Pilot Study used word unigrams and bigrams, part-of-speech bigrams, and character trigrams and 4-grams. Through experimentation, it was discovered that adding word trigrams and part-of-speech trigrams, while reducing character n-grams to only 4-grams, had a net positive effect on the average quadratic weighted kappa of the models. The cost of adding these trigrams is that a) the feature space is now measured in tens to hundreds of thousands of features for data sets of this size, which requires more computer memory to work with and b) searching over the additional feature sets requires more processing time during the training phase. Fortunately, it does not have a large adverse effect on the processing time of generating predictions, which is where efficiency is critical in order to scale properly.
- By using average essay length of the training set, the naïve Bayes or Logistic Regression classifiers can be intelligently chosen. The TurnItIn/LightSide Labs engine can use many different types of classifiers, from Support Vector Machines (SVMs) to Decision Trees to *k*-Nearest Neighbors and so on. It is generally found that for classifying essay-type data, naïve Bayes tends to outperform the other classifiers. However, during the Pilot Study it was noted that Logistic Regression with a strong regularizer tends to outperform naïve Bayes on short answer data. This insight was used to build an average text length cutoff feature into the proprietary optimizer that can automatically choose whether to try naïve Bayes, Logistic Regression, or both. During the Field Test, that cutoff was set such that if the average text length of the training set was less than 350 characters, the optimizer would only try Logistic Regression; if the average length was greater than 450 characters, it would only try naïve Bayes; and if it was in between those two points, it would try both. This resulted in a significant speed up during training, as many possible optimization paths were able to be culled automatically.
- When using Logistic Regression with TurnItIn/LightSide Labs feature selection process on short answer data, L1 regularization tends to outperform L2 regularization. During the Pilot Study, Logistic Regression was only used with L2 regularization. However, L1 regularization, which is a stronger regularizer than L2, tended to outperform L2 on the mathematics short answer data of the Field Test. The optimizer only preferred L2 regularization over L1 on 8 of the 41 short answer mathematics questions that were

modeled. No correlation was found between the chosen regularizer and external item metadata.

- Increasing the number of options for the chi-squared feature selection threshold improves model performance, at the cost of processing time. The proprietary optimizer uses grid search with 10-fold cross-validation to search over different parameter combinations in order to select the optimal set for a given model. One of those parameters is the number of features selected by the chi-squared feature selection process. It was found that expanding the possible values of that parameter led to better models, at the cost of processing time during the search. The values searched over for the Field Test data sets were 500, 1000, 2500, 5000, and 7500.
- When using Logistic Regression with TurnItIn/LightSide Labs feature selection process on short answer data, tuning the misclassification cost can lead to better models. The Logistic Regression models generated during the Pilot Study used a fixed misclassification cost (the C parameter in LIBLINEAR) of 1.0. It was found that by tuning that parameter, we were better able to fit the data. The parameter was tuned, like all of our model parameters, in the optimizer by using grid search with 10-fold cross-validation. The values that we searched over were 0.1, 1.0, and 10.0.

Automated Scoring Engine Technical Description: CTB
(Proprietary Engine)

A. Changes or updates from the Pilot Study:

- We replaced our methods for prediction, replacing neural networks with a wide range of classifiers that include Random Forests, Support Vector Machines (SVMs), and Logistic Regression.
- We developed new content-based features. We revised the method for feature selection. Because features may not necessarily have good discriminative power to classify or predict scores, it is required to select the features that have good discriminative information across scores. For better classifier performance, the available features are ranked and then selected based on how close or how far the two distributions of scores for a certain feature are.
- We perform automatic model selection through cross-validation, based on a customizable objective function. The models predict class probabilities as well as scores. Final scores are computed from the predictions using bias-reducing algorithm.
- We developed new methods to identify condition code papers.

B. Lessons learned and road blocks:

- During the Pilot Study, it became clear that we needed a more nimble system—and developed a totally new architecture that includes seamless use of multiple feature sets, classifiers, and training sets per prompt, as configured by the user. This system enables CTB to train high-quality automated scoring models tailored to specific prompts.

C. Suggestions for future research:

- We are designing a new set of features that can be used to score source-based essays that are being developed to conform to the Common Core State Standards.

Automated Scoring Engine Technical Description: Measurement Incorporated (PEG) (Proprietary Engine)

A. Changes or updates from the Pilot Study:

- Since the Pilot Study, experience in handling the data flow improved across the board. There were still a few unexpected artifacts in the rules and content (e.g. some reader disagreements that did not have a third reader arbitration, some control characters, such as 0x02 in a few files, etc.), but the import and data handling experience was much smoother than during the Pilot. The actual processing was very similar to the work for the Pilot Study, but using pre-established conventions (for instance, treating condition codes as scores-of-zero) also made the training and prediction process more fluid.

B. Lessons learned and road blocks:

- Aside from the relatively minor data issues mentioned above, there were only a few stumbling blocks. One was the presence of large (occasionally > 1,000,000 character-long) essays. These were invariably copy-paste efforts which were given a zero score, but they slowed down the internal file handling somewhat. A larger stumbling block was the presence of training sets with quite skewed scores (an example might be training set of 1,500 responses for an item with a score range of 0-2, then finding about 1,450 items with a score of 0, perhaps about 50 items with a score of 1, and possibly no items with a score of 2). A number of techniques were used to identify and manage this type of training data, but obtaining a high QWK in such situations was often quite difficult, as there were so few non-zero scores that were available for our Automated Scoring system to train upon. There was also a related problem in predicting condition codes, as often the coded responses would be too few, or different codes might be found for very similar responses.

C. Suggestions for future research:

- We are constantly looking for ways to fold new research into our process. For instance, meta-cognitive evaluation of the methods we use to best accommodate skewed training samples.

Chapter 4: Continued Read-Behind Studies

High-stakes assessment programs typically use more than one human rater to score constructed-response items to reduce rater effects. For instance, performance tasks that are part of admissions or licensure tests are often scored by two human raters (Ramineni & Williamson, 2013). In such a scoring scenario, a senior, expert rater usually adjudicates any disagreement in the scores given by the first and second rater. Averaging the two initial ratings is another possibility when the assessment program allows for fractional scores.

Any multiple-rater scoring scenario (with adjudication) needs multiple reads per response, which increases scoring costs. Therefore, testing programs may consider using a single-rater scoring scenario, in which each response is scored by a single human rater or a single Automated Scoring system. An obvious disadvantage of these scoring scenarios is that second reads can no longer be used to ensure score quality. Instead, assessment programs seek to ensure score quality through other methods. Measures typically implemented include check-reads (raters need to qualify for scoring on a selected set of responses) and a limited percentage of second reads (e.g., a second rater for 10% of the responses).

There is increasing interest in alternative scoring scenarios which combine human and Automated Scoring. These scoring scenarios potentially offer both cost savings as well as high score quality. For example, ETS has been using its Automated Scoring system (e-rater®) in place of a second human rater on the Test of English as a Foreign Language (TOEFL) since 2009 (Trapani, Bridgeman, & Breyer, 2011). Another possibility is to use Automated Scoring systems to detect responses likely to require human review. These latter methods are the topic of other studies (see e.g., Chapter 5 of this report and Chapter 4 of the Pilot Study report) and will not be considered here.

Several scoring scenarios ("read and read-behind scenarios") were investigated in the Pilot Study. These scenarios can be categorized based on

1. the number of raters (one or two),
2. the type of the first and second rater (human or Automated Scoring system), and
3. the adjudication rule which determines when scores from the first and second rater need to be adjudicated by a third rater:
    a. adjudicate when the scores of the first and second rater disagree (non-exact)
    b. adjudicate when the scores of the first and second rater differ by more than 1 score point (non-adjacent)

A total of eight different read and read-behind scenarios were investigated as part of the Pilot Study. These scoring scenarios are summarized in Table 4.1. See Chapter 3 of the Pilot Study report for additional details. Note that the term read-behind was chosen over the more familiar term second-read to emphasize the following point: Consider a scoring scenario with two raters, a human reader and an Automated Scoring system. We say that the Automated Scoring system reads behind the human rater, because the final score of record will always be assigned by a human—either the first human reader or a senior, expert human rater in case of adjudication. In other words, the Automated Scoring system is used only to determine whether a response requires an additional human read.

Chapter 4: Continued Read-Behind Studies

Table 4.1. Read and Read-Behind Scenarios Investigated during the Pilot Study

Scenario	Number of Raters	First Rater	Read-Behind Rater	Adjudication Rule
1	Single rater	Human		
2		Top-performing Automated Scoring system		
3	Two raters	Human	Human	Non-exact
4				Non-adjacent
5		Human	Top-performing Automated Scoring system	Non-exact
6				Non-adjacent
7		Top-performing Automated Scoring system	Second-best performing Automated Scoring system	Non-exact
8				Non-adjacent

An important research question is the practical impact these different scoring scenarios have on the score quality, as compared to a baseline scenario. To answer this question, in the Pilot Study scenario scores of record for each of the read and read-behind scenarios were constructed from available data. This allowed a comparison of the impact each scenario would have had on score quality if it had been used. The final human score of record served as a baseline in the comparisons; these scores were produced using Scenario 3. This scoring scenario was chosen as baseline as it represents a fully human scoring scenario.

Scoring scenarios were compared on five English language arts (ELA)/literacy short-text, constructed-response, items; three mathematics short-text, constructed-response, items; and five ELA/literacy essay items. Table 4.2 presents the results for the mathematics items—the pattern was similar for the ELA/literacy short-text, constructed-response, and essay items (see the Pilot Study report for details). Note that only four read-behind scenarios applied to the mathematics items, because these items were scored 0 or 1. Hence, there were no non-adjacent scores and only non-exact adjudication applied.

Table 4.2. Agreement Statistics Summarized by Scenario, Mathematics

Scenario	Average					
	SMD	Kappa	Quadratic Weighted Kappa	Correlation	Percent Agree	Percent Adjacent & Agree
Human-human	0.01	0.79	0.79	0.76	0.88	1.00
1	0.02	0.90	0.90	0.89	0.94	1.00
2	0.06	0.86	0.86	0.83	0.92	1.00
5	0.02	0.98	0.98	0.96	0.98	1.00
7	0.05	0.89	0.89	0.88	0.94	1.00

Note: See Appendix 3.D for a brief description of some of these rater agreement statistics.

Comparing the single-read scenarios (Scenarios 1 and 2) to the human-human inter-rater agreement statistics, the average score quality of the single-read scenarios was (very) high. An even higher score quality was achieved in the scenario where the best-performing Automated Scoring system served as read-behind for a single human rater (Scenario 5). The increase in rater agreement statistics demonstrates the positive effect read behind and adjudication can have on score quality.

Although these results are promising, a major methodological limitation of the Pilot Study is that the first human rater was used to construct scenario scores of record for Scenarios 1 and 3-6. However, this rating was also used to determine the final human score, which was used as a baseline in score quality comparisons. In other words, these ratings were not independent of the baseline score. Therefore, the score quality of several of the read-behind scenarios may have been inflated; in particular, the benefits of using a single human rater (Scenario 1) and an Automated Scoring system read-behind (Scenario 5) may have been overstated.

To mitigate this effect, ratings from another set of human raters are needed, produced independently from the scoring process resulting in the final scores of record. Such ratings, however, were not available during the Pilot Study, because collecting these data was organizationally challenging.

Purpose

The present study extended the research on read and read-behind scenarios conducted as part of the Pilot Study.

First, ratings from an independent human rater were collected as part of the Field Test, to address the methodological limitations of the research during the Pilot Study. With these data, several read-behind scoring scenarios studied in the Pilot Study (as described in Table 4.1) were reexamined and new scenarios were investigated.

Second, the human-Automated Scoring system read-behind scenarios during the Pilot Study considered only the top-performing Automated Scoring systems as a read-behind for human raters. However, that leaves the question of how the quality of the Automated Scoring system impacts the

results of the read-behind scenarios. Thus, the second-best Automated Scoring system was also considered as a read-behind rater. Finally, the scoring scenario where both Automated Scoring systems serve as read-behind (double read-behind) was also considered.

In summary, eleven scoring scenarios were investigated as part of the Field Test studies (see Table 4.3).

Table 4.3. Read and Read-Behind Scenarios Investigated during the Field Test Study

Scenario	Number of Raters	First Rater	Read-Behind Rater(s)	Adjudication Rule
1	Single rater	Independent human reader		
2		Top-performing Automated Scoring system		
3		Second-best performing Automated Scoring system		
4	Two raters	Independent human reader	Top-performing Automated Scoring system	Non-exact
5				Non-adjacent
6			Second-best performing Automated Scoring system	Non-exact
7				Non-adjacent
8		Top performing Automated Scoring system	Second-best performing Automated Scoring system	Non-exact
9				Non-adjacent
10		Second-best performing Automated Scoring system	Top-performing Automated Scoring system	Non-adjacent
11	Three raters	Independent human reader	Top-performing and second-best Automated Scoring systems	Non-exact

Methodology

The methodology of this study generally follows the methodology of the read and read-behind study conducted during the Pilot Study, except that independent human ratings were also collected. Additionally, more items (21 compared to 13) and more scoring scenarios (11 compared to 8) were considered.

Data Source

The primary data sources for this study consisted of seven ELA/literacy short-text, constructed-response, items; seven mathematics short-text, constructed-response, items; and seven ELA/literacy essay items. The items were selected to cover several different types of short-text, constructed-response, items (reading, research, or writing) and essay items corresponding to different writing purposes (informational, explanatory, argumentative, or narrative).

These items were administered as part of the online Field Test. ELA/literacy short-text, constructed-response, items were scored on a scale from 0 to 2 points, mathematics short-text, constructed-response, items were scored 0 or 1 point, and essay items were scored on three traits (trait A [Organization/Purpose], 1–4 points; trait B [Evidence/Elaboration], 1–4 points; and trait C [Conventions], 0–2 points).

As part of the Field Test, each item was handscored by two human raters. When the two (initial) raters assigned the same score, this was the final score for the response. If the scores of the two human raters did not agree exactly, then the response was routed to a third, expert rater who assigned the final score (non-exact adjudication). This final score served as the baseline score of record. Automated Scoring systems were trained using these final scores.

The responses also received one additional handscore by another trained human rater. This rater was a fully qualified human reader and experienced with the scoring rubric for the item. These ratings were not a part of the final score (and hence not adjudicated) but served as an independent human score in the read and read-behind scenarios. This reflected scoring scenarios where a single human rater scored responses (possibly in combination with an Automated Scoring system). The (possibly adjudicated) score from the other human raters served as the baseline score of record against which different scoring scenarios were compared.

Procedures

For each of the 21 items, about 500 responses in the validation set were available along with scores from each of the Automated Scoring systems that scored that particular item.

- Each short-text, constructed-response, item had scores from four independently developed Automated Scoring systems.
- Each essay item had scores from three independently developed Automated Scoring systems.

For the purpose of this study, after evaluating the score quality of each Automated Scoring system using the evaluation criteria, the two Automated Scoring systems evaluated as the two highest performers were selected (by item for short-text, constructed-response, items and by item/trait for essay items). Performance was determined first by the total number of flags (fewer is better) and,

when Automated Scoring systems had the same number of flags, by quadratic weighted kappa (higher is better).

Score quality for each scenario was evaluated using the criteria in the evaluation framework. This included the following agreement statistics: standardized mean difference (SMD), quadratic weighted kappa (QWK; Cohen, 1960, 1968), Pearson correlation, percent agreement, and percent adjacent and (exact) agreement.

Averages for Pearson correlation and Cohen's quadratic weighted kappa were computed using Fisher's *r*-to-*z* transformation. This may produce less biased estimates of population correlations (Silver and Dunlap, 1987). Specifically, each correlation *r* was converted to a *z*-score before averaging, as follows:

$$z = \frac{1}{2}\ln\frac{1+r}{1-r}$$

The average *z*-score, $\bar{z}$, was then converted back to an average correlation $\bar{r}$:

$$\bar{r} = \frac{e^{2\bar{z}} - 1}{e^{2\bar{z}} + 1}$$

A similar procedure was used to average quadratic weighted kappa statistics.

Scenarios with different adjudication rules differed in the number of responses that required a third read. Adjudication rates were computed for each scenario, where the adjudication rate was defined as the percentage of responses that required a third read under the scenario scoring rules.

Results

Tables 4.A.1–4.A.12 present detailed information by item for human-human inter-rater agreement as well as the eleven scoring scenarios investigated. Details presented in the appendix include mean and standard deviation of the focal and reference scores, standardized mean difference (SMD), quadratic weighted kappa (QWK), correlation, percent agreement, percent adjacent and (exact) agreement, and number of flags (total, absolute, relative, subgroup). The presentation in this chapter focuses on summarizing these results.

ELA/literacy short-text, constructed-response items.

Based on quadratic weighted kappa (QWK), a single independent human rater scoring scenario produced slightly higher quality scores than the human-human inter-rater agreement levels, with a QWK around 0.77. The two single-read scenarios using Automated Scoring systems produced higher quality scores with an average QWK of about 0.80, and moreover, resulted in much fewer flags on average.

Using an Automated Scoring system to read-behind a human rater with non-exact adjudication increased score quality substantially resulting in an average QWK of about 0.92. Using the second-best Automated Scoring system instead of the top-performing Automated Scoring system resulted in slightly more subgroup flags. Using both Automated Scoring systems to read behind a single human rater produced the highest quality scores with an average QWK of 0.94.

When only non-adjacent scores were adjudicated, there was no improvement in score quality. An explanation is that very few if any scores assigned by read-behind Automated Scoring systems were

actually non-adjacent.

Using an Automated Scoring system as a read-behind for another Automated Scoring system did result in an improvement in score quality. However, an Automated Scoring system reading behind a human rater produced better scores.

Table 4.4. Average Rater Agreement Statistics, ELA/literacy—Short-text Items

Scenario	SMD	Agreement Statistics				Average Number of Flags			
		QWK	Correlation	Agree	Adjacent & Agree	Total	Absolute	Relative	Subgroup
Human-Human	0.00	0.76	0.76	79	99	0.43	0.29	0.00	0.14
1	0.00	0.78	0.79	80	100	4.86	0.86	0.00	4.00
2	-0.01	0.81	0.81	83	100	0.29	0.00	0.00	0.29
3	0.00	0.80	0.80	82	99	0.29	0.00	0.00	0.29
4	-0.01	0.93	0.93	93	100	0.00	0.00	0.00	0.00
5	0.00	0.78	0.80	81	100	4.86	0.86	0.00	4.00
6	-0.01	0.92	0.92	93	100	0.14	0.00	0.00	0.14
7	0.01	0.78	0.80	81	100	4.71	0.71	0.00	4.00
8	-0.01	0.87	0.87	88	100	0.00	0.00	0.00	0.00
9	-0.01	0.81	0.81	83	100	0.29	0.00	0.00	0.29
10	0.00	0.80	0.80	82	99	0.43	0.00	0.00	0.43
11	-0.01	0.94	0.94	95	100	0.00	0.00	0.00	0.00

Mathematics short-text, constructed-response items.

The pattern for the seven mathematics short-text, constructed-response, items was similar to the pattern for the ELA/literacy short-text, constructed-response, items.

Table 4.5. Average Rater Agreement Statistics, Mathematics—Short-text Items

Scenario	SMD	Agreement Statistics				Average Number of Flags			
		QWK	Correlation	Agree	Adjacent & Agree	Total	Absolute	Relative	Subgroup
Human-Human	0.00	0.89	0.89	91	100	0.00	0.00	0.00	0.00
1	0.01	0.88	0.89	87	100	3.29	0.43	0.00	2.86
2	0.00	0.92	0.92	93	100	0.00	0.00	0.00	0.00
3	-0.01	0.90	0.90	92	100	0.14	0.00	0.00	0.14
4	0.00	0.97	0.97	98	100	0.00	0.00	0.00	0.00
5	0.02	0.88	0.90	88	100	3.29	0.43	0.00	2.86
6	0.00	0.97	0.97	97	100	0.00	0.00	0.00	0.00
7	0.02	0.88	0.90	88	100	3.29	0.43	0.00	2.86
8	-0.01	0.94	0.94	95	100	0.00	0.00	0.00	0.00
9	0.00	0.92	0.92	93	100	0.00	0.00	0.00	0.00
10	-0.01	0.90	0.90	92	100	0.14	0.00	0.00	0.14
11	0.00	0.98	0.98	99	100	0.00	0.00	0.00	0.00

Here, a single independent human rater scenario produced scores with an average QWK of 0.88, slightly lower than but still comparable to human-human inter-rater agreement. The two Automated Scoring systems produced scores with a slightly higher average QWK of about 0.90. Using an Automated Scoring system as a read-behind improves score quality considerably with an average QWK of about 0.97, provided non-exact adjudication is used. When only non-adjacent scores are adjudicated, the improvements in score quality are minimal if any. As for the ELA/literacy items, an Automated Scoring system reading behind an independent human rater produced scores with a higher quality QWK than when an Automated Scoring system reads behind another system.

ELA essay items.

For the ELA/literacy items, the pattern for the read-behind scenarios resembles that for the previous two item types, but the single-read scenarios are different. For all three traits, these scoring scenarios produce scores with average QWKs below human-human inter-rater agreement.

Table 4.6. Average Rater Agreement Statistics, Essay—Trait A

Trait	Scenario	SMD	Agreement Statistics				Average Number of Flags			
			QWK	Correlation	Agree	Adjacent & Agree	Total	Absolute	Relative	Subgroup
A	Human-Human	0.01	0.86	0.86	79	99	0.00	0.00	0.00	0.00
A	1	-0.01	0.81	0.82	73	98	2.86	0.57	0.00	2.29
A	2	0.00	0.83	0.83	77	99	0.43	0.29	0.00	0.14
A	3	0.02	0.83	0.83	77	99	1.00	0.29	0.43	0.29
A	4	-0.01	0.92	0.92	88	100	0.00	0.00	0.00	0.00
A	5	-0.01	0.83	0.84	73	99	2.71	0.57	0.00	2.14
A	6	0.00	0.93	0.93	89	100	0.00	0.00	0.00	0.00
A	7	0.00	0.83	0.84	74	99	2.57	0.57	0.00	2.00
A	8	0.02	0.90	0.90	87	99	0.29	0.00	0.00	0.29
A	9	0.00	0.84	0.84	78	99	0.57	0.29	0.00	0.29
A	10	0.02	0.84	0.84	77	99	0.57	0.29	0.00	0.29
A	11	0.00	0.95	0.95	92	100	0.00	0.00	0.00	0.00

Table 4.7. Average Rater Agreement Statistics, Essay—Trait B

Trait	Scenario	SMD	Agreement Statistics				Average Number of Flags			
			QWK	Correlation	Agree	Adjacent & Agree	Total	Absolute	Relative	Subgroup
B	Human-Human	0.02	0.85	0.85	78	99	0.14	0.00	0.00	0.14
B	1	0.03	0.81	0.81	72	98	2.71	0.57	0.00	2.14
B	2	0.02	0.85	0.85	78	99	0.14	0.00	0.00	0.14
B	3	0.02	0.83	0.83	76	99	0.86	0.29	0.29	0.29
B	4	0.01	0.92	0.93	89	100	0.00	0.00	0.00	0.00
B	5	0.03	0.82	0.83	74	99	2.57	0.57	0.00	2.00
B	6	0.01	0.92	0.92	89	100	0.00	0.00	0.00	0.00
B	7	0.03	0.83	0.84	74	99	2.57	0.57	0.00	2.00
B	8	0.01	0.90	0.90	86	100	0.00	0.00	0.00	0.00
B	9	0.02	0.85	0.85	79	99	0.14	0.00	0.00	0.14
B	10	0.02	0.83	0.83	76	99	0.57	0.29	0.00	0.29
B	11	0.01	0.94	0.94	92	100	0.00	0.00	0.00	0.00

Table 4.8. Average Rater Agreement Statistics, Essay—Trait B

Trait	Scenario	SMD	Agreement Statistics				Average Number of Flags			
			QWK	Correlation	Agree	Adjacent & Agree	Total	Absolute	Relative	Subgroup
C	Human-Human	0.04	0.73	0.73	72	98	1.57	1.14	0.00	0.43
C	1	-0.09	0.65	0.67	65	96	4.57	1.86	0.00	2.71
C	2	0.00	0.73	0.73	74	98	0.43	0.29	0.00	0.14
C	3	0.01	0.68	0.68	70	97	2.57	0.86	0.57	1.14
C	4	-0.02	0.89	0.89	88	99	0.14	0.00	0.00	0.14
C	5	-0.09	0.69	0.71	67	98	3.57	1.00	0.00	2.57
C	6	-0.02	0.88	0.89	88	99	0.14	0.00	0.00	0.14
C	7	-0.09	0.70	0.72	68	98	4.14	1.29	0.00	2.86
C	8	0.00	0.84	0.84	84	99	0.14	0.00	0.00	0.14
C	9	0.00	0.74	0.74	74	98	0.43	0.29	0.00	0.14
C	10	-0.01	0.71	0.71	71	98	2.43	1.00	0.00	1.43
C	11	-0.01	0.93	0.93	92	99	0.00	0.00	0.00	0.00

Adjudication Rates. Table 4.B.1 in Appendix 4.B presents adjudication rates by item for the eight scenarios involving more than a single rater. Table 4.9 presents average adjudication rate for each of these scenarios for ELA short-text, constructed-response, items; mathematics short-text, constructed-response, items; and ELA essay item by trait.

Table 4.9. Average Percentage of Adjudication

Average Percentage of Adjudication									
Content	Trait	S4	S5	S6	S7	S8	S9	S10	S11
ELA		22.2	0.6	23.1	1.0	10.5	0.1	0.1	27.6
MA		15.1	0.3	15.8	0.3	5.6	0.0	0.0	18.2
Essay	A	25.7	1.7	26.6	1.4	18.1	0.8	0.8	34.7
Essay	B	25.6	1.1	27.3	1.7	17.1	0.1	0.1	34.4
Essay	C	34.8	3.3	36.4	4.3	22.4	1.6	1.6	45.2

Similar to the Pilot Study, average adjudication rates in non-adjacent adjudication scenarios are low (maximum 4.3 percent). As expected, scoring scenarios with non-exact adjudication had higher adjudication rates, ranging from 5.6 percent (two Automated Scoring systems for mathematics short-text, constructed-response, items) to 36.4 percent (second-best Automated Scoring system reading behind a human rater).

Adjudication rates are higher when an Automated Scoring system reads behind a human rater than when an Automated Scoring system reads behind another Automated Scoring system. Using the second-best Automated Scoring system increased adjudication rates slightly. The scoring scenario with three raters had the highest adjudication rates, with over 45 percent of responses adjudicated for essay items, trait C.

Adjudication rates were lowest for the mathematics short-text, constructed-response, items, followed by ELA/literacy short-text, constructed-response, items. Essay items, traits A and B, had slightly higher adjudication rates. A possible explanation may be the number of score points. Essay items, trait C, does not follow this pattern, as it had the highest adjudication rates. This may be connected to the generally lower agreement rates for this trait.

Discussion

Bennett (2011) and Zhang (2013) have suggested using Automated Scoring in combination with human scoring in different read and read-behind scenarios. These scenarios offer potentially higher score quality at lower costs for large-scale assessment programs. To help inform assessment programs that are considering using Automated Scoring systems in combination with human raters, many different scoring scenarios were investigated in the Pilot Study and the Field Test.

The finding of the present studies can be summarized as follows:

- In the Pilot Study, a single human rater scenario resulted in scores with average QWK much

higher than human-human inter-rater agreement. However, the ratings from that scenario were not produced independently. In the Field Test, independent ratings were collected, and average QWK for a single human rater scenario was around or below inter-rater agreement levels.

- As was the case in the Pilot Study, an Automated Scoring system reading behind a single human rater resulted in a substantial improvement in score quality (as measured by average QWK). The difference between using the second-best Automated Scoring system, as opposed to the top-performing system, was minimal.
- An Automated Scoring system reading behind another Automated Scoring system did not improve score quality by much. Note, however, that the systems were not explicitly trained for a read-behind scenario. It could be that improvements can be made if a system is designed for read-behind use.
- Only adjudicating non-adjacent scores did not increase score quality over single-read scoring scenarios. Thus, pragmatically the choice may be between either a single read scenario or a second-read scoring scenario with non-exact adjudication.

Chapter 5: Targeting Responses for Human Review

Automated Scoring systems are designed to predict the score that a human rater would assign to a given response based on the associated rubric. One of the challenges associated with the application of these systems to high-stakes testing is that automated scores will deviate from human scores for some responses. Although metrics to evaluate Automated Scoring systems describe overall scoring consistency compared to humans, such indexes do not identify which responses are likely to receive scores that are discrepant with human scores. In a scenario where most responses are scored by a single Automated Scoring system only, it is important to identify which responses are aberrant to the extent that their automated score is likely to be different from the score a human would have given. In such situations, flagged responses should be routed to humans for scoring.

Hastie, Tibshirani, and Friedman (2009, Equation 7.9, p. 223) provide a decomposition of the expected prediction error of a regression model that is insightful in discussing the reasons why Automated Scoring systems may provide a predicted score that is discrepant from a human score. In a regression model, it is assumed that the output *Y* (i.e., the human score) is a function of the input variables *X* (i.e., response features), plus a random error term ε with an expected value of zero, $E(\varepsilon) = 0$ Using squared-error loss, the expected prediction error at input x_0, $Err(x_0)$, decomposes as a sum of three terms,

$$Err(x_0) = \sigma_\varepsilon^2 + \left[E\left(\hat{f}(x_0)\right) - f(x_0)\right]^2 + E\left(\left[\hat{f}(x_0) - E\left(\hat{f}(x_0)\right)\right]^2\right), \quad (1)$$

where $Var(\varepsilon) = \sigma_\varepsilon^2$ and $\hat{f}(x_0)$ is the estimator of $f(x_0)$.

The first term, σ_ε^2, is irreducible error; it cannot be avoided regardless of how well the Automated Scoring system is trained. The second term is the squared bias, and the third term is the variance. In this study, it is assumed that Automated Scoring systems are sufficiently flexible and that they were trained on a sizeable training sample, and therefore the squared bias and variance are low overall. We can then distinguish between the following two reasons for a high discrepancy between human and automated scores:

1. The prediction error is large due to a high amount of irreducible error. A related situation is where the true mean is not a smooth function of the features in certain regions of the feature space, so that there is always a large interpolation error, even when the data are not sparse in that region.
2. There are regions in the feature space with only few training examples (sparse regions). Predicted scores in those regions are based on extrapolation from regions in the feature space that are more densely populated or on interpolation based on a few cases only. An Automated Scoring system is likely to suffer from high bias or variance in these regions, resulting in a large prediction error.

Chapter 5: Targeting Responses for Human Review

Three different studies approaches were investigated in the Pilot Study. The first approach addressed irreducible error by modeling the agreement between human and automated scores as a function of the input variables that constitute the feature space. This is analogous to the analysis of residuals in linear regression. The second approach focused on the identification of (observations in) sparse regions of the feature space using outlier detection methods, and whether the degree to which a response is atypical in terms of features is related to the amount of discrepancy between human and automated scores. The third approach applied to Automated Scoring systems employ several models in a first stage and treat the scores that result from each of these models as input variables in a second stage. We investigated whether large discrepancies between automated scores obtained in the first stage were predictive for discrepancies between automated and human scores. Among the three approaches, the identification of responses in sparse regions of the feature space using outlier detection methods was the most promising. Therefore, this is the approach that is further investigated in the Field Test study.

Purpose

The first goal of this study is to replicate and fine tune the outlier detection method that was established for the Pilot Study. The number of items considered in this study is more than four times the number of items considered in the previous study. In addition, cross-validation techniques will be used to select optimal choices for settings that were selected in an ad hoc manner before.

Second, we will assess the practical significance of the method by comparing two scenarios for routing a subset of automated scores for a second human read. In the first scenario, responses that are flagged by the outlier detection method will be routed for human review (targeted routing), whereas in the second scenario, the same proportion of scores will be routed for human review, but these responses will be randomly selected (random routing). The random routing scenario is a baseline scenario against which the targeted routing will be compared. We will look at both the proportion of routed scores that needed adjudication, and the agreement statistics between the final score for each scenario and the official score of record.

Methodology

Data Source

We used the 63 Case 2 items that were scored by multiple Automated Scoring systems (see Chapter 1). In particular, there were 21 English language arts (ELA)/literacy short-text, constructed-response items, 21 mathematics short-text, constructed-response, items, and 21 essay items. For the essay items, the scores on all three traits (A: Organization/Purpose; B: Evidence/Elaboration; C: Conventions) were considered. The characteristics of the items are discussed in Chapter 1 and in Table 1.1.

Whether or not a response lies in a sparse region of the feature space (an outlier) obviously depends on the feature space that is specific to the Automated Scoring system. For the essays, we used the features that were extracted by the Vendor 1 engine. For the ELA/literacy and mathematics items, we used the features that were extracted by Vendor 6.

For this study, all responses that received a condition code as the score of record were excluded. In addition, the Vendor 1 engine, but not the Vendor 6 engine, was able to predict a condition code. Responses for which the Vendor 1 engine produced a condition code were excluded as well.

Procedures

In the first part of the study, we by-and-large employed the method developed for the Pilot Study and scaled the method to deal with a larger number of items. The method consisted of several stages. First, a principal component analysis was carried out to reduce the dimensionality of the feature space. In the next step, a partitioning around means analysis was carried out on the principal component scores of the training set. Partitioning around medoids is a statistical technique similar to *k*-means clustering , but it is more robust than *k*-means clustering in the presence of outliers (see Husson, Josse, & Pagès, 2010, for a discussion).

For the Vendor 1 Automated Scoring system, all features were numerical variables. The means and standard deviations varied considerably across features but this variation was at least partially due to differences in scales (i.e., features included frequencies as well as ratios). Therefore, all variables were standardized prior to the principal component analysis. In contrast, all features were binary and reflected presence or absence for the Vendor 6 Automated Scoring system. Because the same scale with the same meaning was in place in this case, the feature variables were not standardized prior to the principal component analysis. The principal component solution obtained from the training data was used to generate principal component scores for both the training and the validation sets.

For the Field Test items, the optimal number of principal components and the optimal number of clusters was determined using 5-fold cross-validation. The number of principal components was varied from 15 to 25, and the number of clusters from 2 to 5, resulting in 44 possible combinations. For each combination, the cross-validated polyserial correlation was computed between the Euclidean distance of the responses to the closest cluster medoid and the absolute value of the difference between human and automated scores. The polyserial correlation is a measure of association between an ordered categorical and a continuous variable (i.e., the absolute value of the difference between human and automated scores, and the distance to the closest cluster medoid, respectively). We opted for the polyserial correlation in order to select the number of principal components and clusters that maximized the relation between the degree of atypicality of a response and discrepancies between human raters and automated scoring systems. Note that for the essay items, even though the feature space is the same across the three traits, the number of principal components and clusters that optimizes the cross-validated polyserial correlation can still be different. Therefore, the number of principal components and the number of clusters was determined separately for each of the three traits.

Once the number of principal components and the number of clusters was determined, one more partitioning around means was carried out on the entire training set. This final partitioning of the training data was used to compute the Euclidean distances between the responses of the validation set and each of the cluster medoids

In the second part of the study, two routing scenarios were compared. In the first scenario, the (100-*y*) % responses of the validation set for which the distance to the closest medoid was larger than percentile *y* were routed for human review. The official score of record was used for the human score. In case of non-exact agreement, the score was adjudicated by taking the official score of record as the final score. This scenario was carried out for a range of percentiles, and for each value the quadratic weighted kappa was computed between the human scores and the (possibly adjudicated) automated scores. This scenario was compared against a baseline scenario where (100-*y*) % of the validation responses were selected at random and routed for human review.

Results

Table 5.1 presents the descriptive statistics for the selected number of principal components and number of clusters for each of the three item types. For the ELA/literacy items (Vendor 6), the selected number of principal components ranged from 15 to 25, with a mean of 21.04, a median of 22, and a standard deviation of 3.77. For the mathematics items (Vendor 6), the selected number of principal components ranged from 15 to 25 with a mean of 19.38, a median of 18, and a standard deviation of 3.93. Finally, for the essays (Vendor 1), the selected number of principal components ranged from 15 to 25, with a mean of 21.34, a median of 22, and a standard deviation of 3.34.

The selected number of principal components and clusters for individual items are presented in Appendix 5.A.

Table 5.1. Descriptive Statistics of Principal Components and Clusters

Content	Mean		Median		Standard Deviation		Range	
	Principle Components	Clusters	Principle Components	Clusters	Principle Components	Clusters	Principle Components	Clusters
ELA	21.05	3.76	22	4	3.77	1.14	10	3
MA	19.38	3.00	18	3	3.93	1.18	10	3
Essay	21.35	3.11	22	3	3.34	1.21	10	3

The degree of success of identifying which responses are likely to receive scores that are discrepant with human scores depends on the association between the distance to the closest cluster medoid and the absolute difference between human and automated scores. Figures 5.1 to 5.5 summarize this association computed on the validation set for each of the three item types. Each panel presents the average distance to the closest cluster medoid for each value of the absolute difference between human and automated scores. For the essay items, the maximal possible absolute difference between human and automated score amounted to three for the traits Organization/Purpose (trait A) and Evidence/Elaboration (trait B). Because an absolute differences of three was rarely observed, absolute differences of two and three were collapsed into a single category for these traits. The mathematics items were scored on a scale of 0–2, and therefore the maximum absolute difference amounted to 2. The average distances were computed separately for each item on the responses of the validation set, and then further averaged across the 21 items of the same item type. For the essay items, the averages were computed separately for each of the three traits. Each panel also presents the polyserial correlation(s), averaged over the 21 items of a given type. For each item type, very commonly, the average distance to the closest cluster medoid increased with an increasing discrepancy between human and automated scores. The results for individual items are presented in Appendix 5.B (discrepancy) and 5.C (polyserial correlations).

Figure 5.1. Discrepancy vs Average distance—ELA/literacy

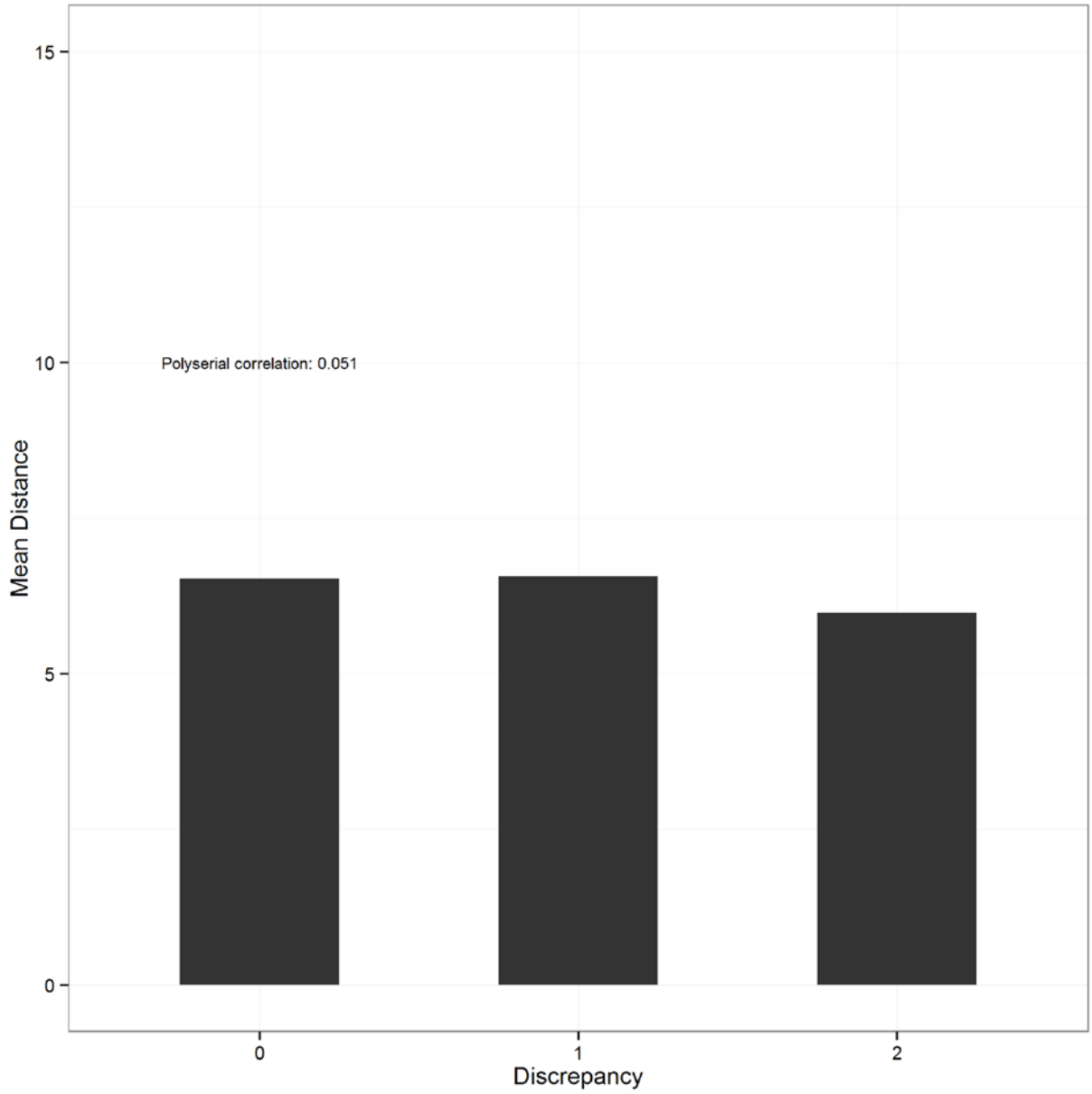


Figure 5.2. Discrepancy vs Average distance—Mathematics

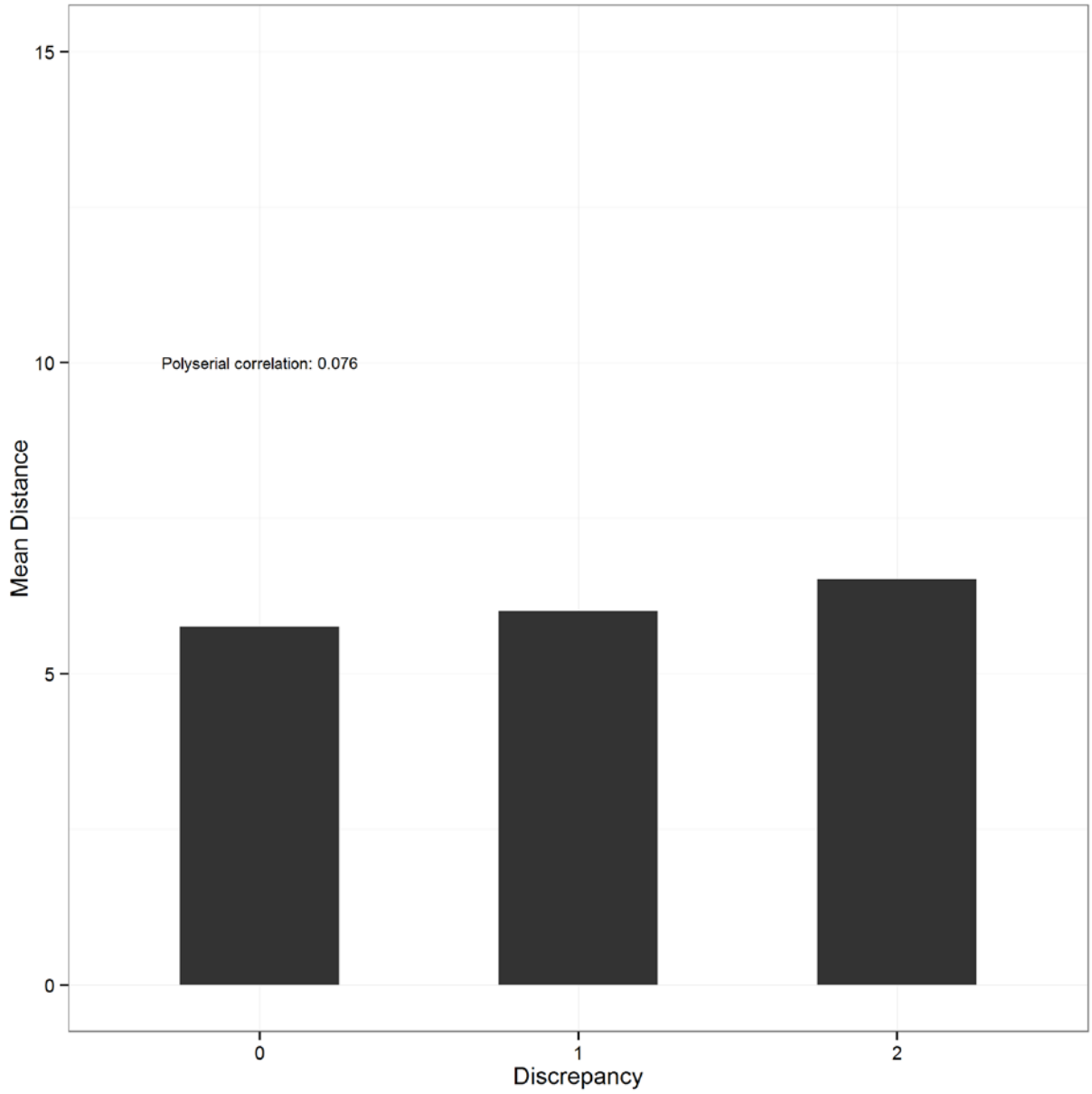


Figure 5.3. Discrepancy vs Average distance—Essay, Trait A

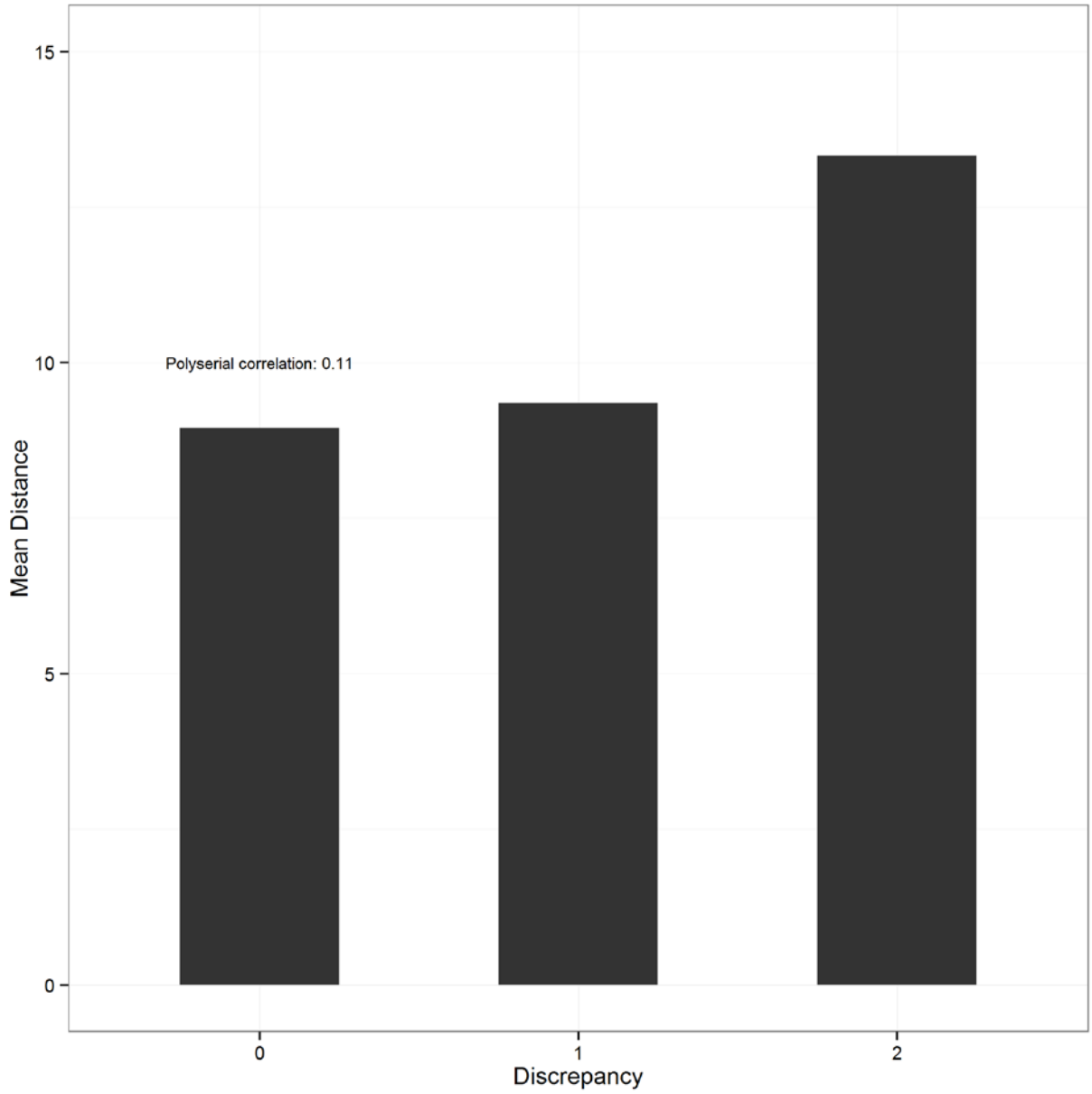


Figure 5.4. Discrepancy vs Average distance—Essay, Trait B

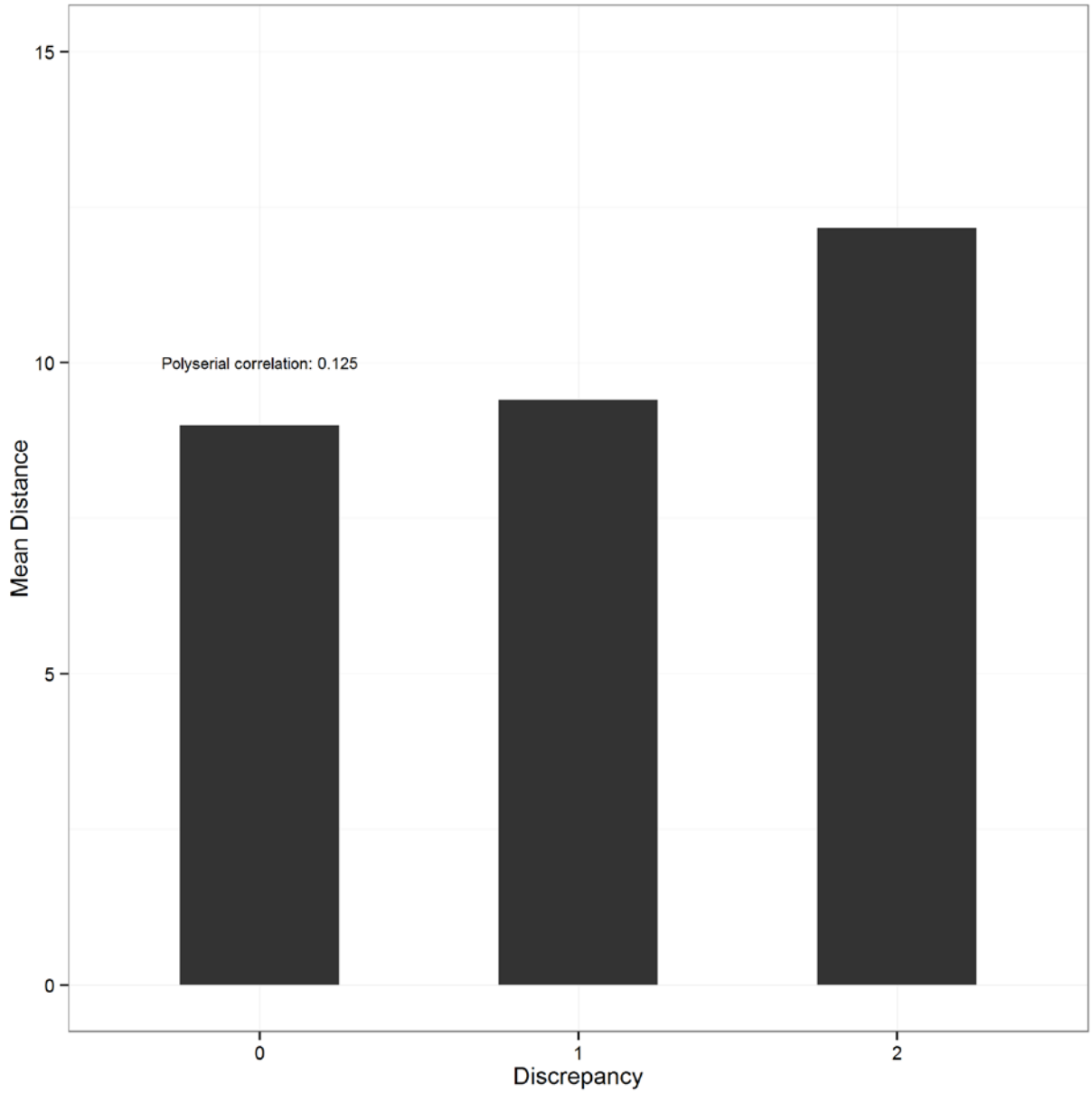
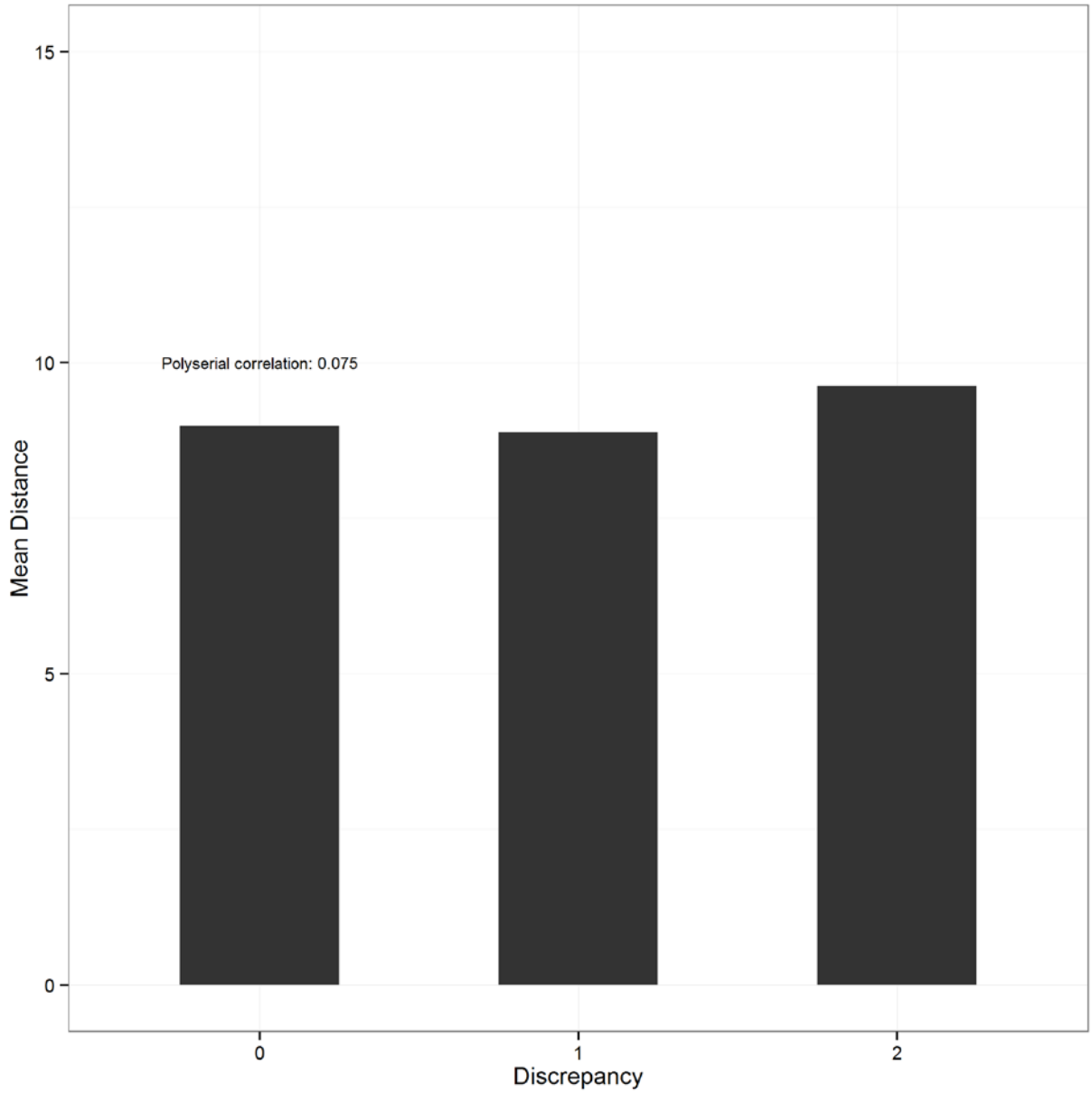


Figure 5.5. Discrepancy vs Average distance—Essay, Trait C



Figures 5.6 through 5.8 summarize the results for the random and targeted routing scenarios. Again, we present the average results for each item type. Each of the figures presents the increase in average quadratic weighted kappa as a function of the proportion of responses that were routed for human review. Because the automated scores were adjudicated whenever there was non-exact agreement between human and automated scores, all quadratic weighted kappas asymptote to the value of 1 as the proportion of routed responses approach 1. The averages are computed over the 21 items within each item type. The results for individual items are presented in Appendix 5.D through Appendix 5.F. For all item types (and traits), the curves for the targeted routing dominate the curves for random routing, indicating that for any given proportion of routing automated scores for a human review, more discrepant scores were resolved under the scenario of targeted routing. For example, at 20% of the data sent for human review, the targeted routing performs 5% better than random routing for ELA/literacy, performs 4.2% better than random routing for mathematics, and 5.1% better than random routing for essay items measured across quadratic weighted kappa.

Figure 5.6. Comparison of Average Performance of Random vs Targeted Routing—ELA/literacy Short-text

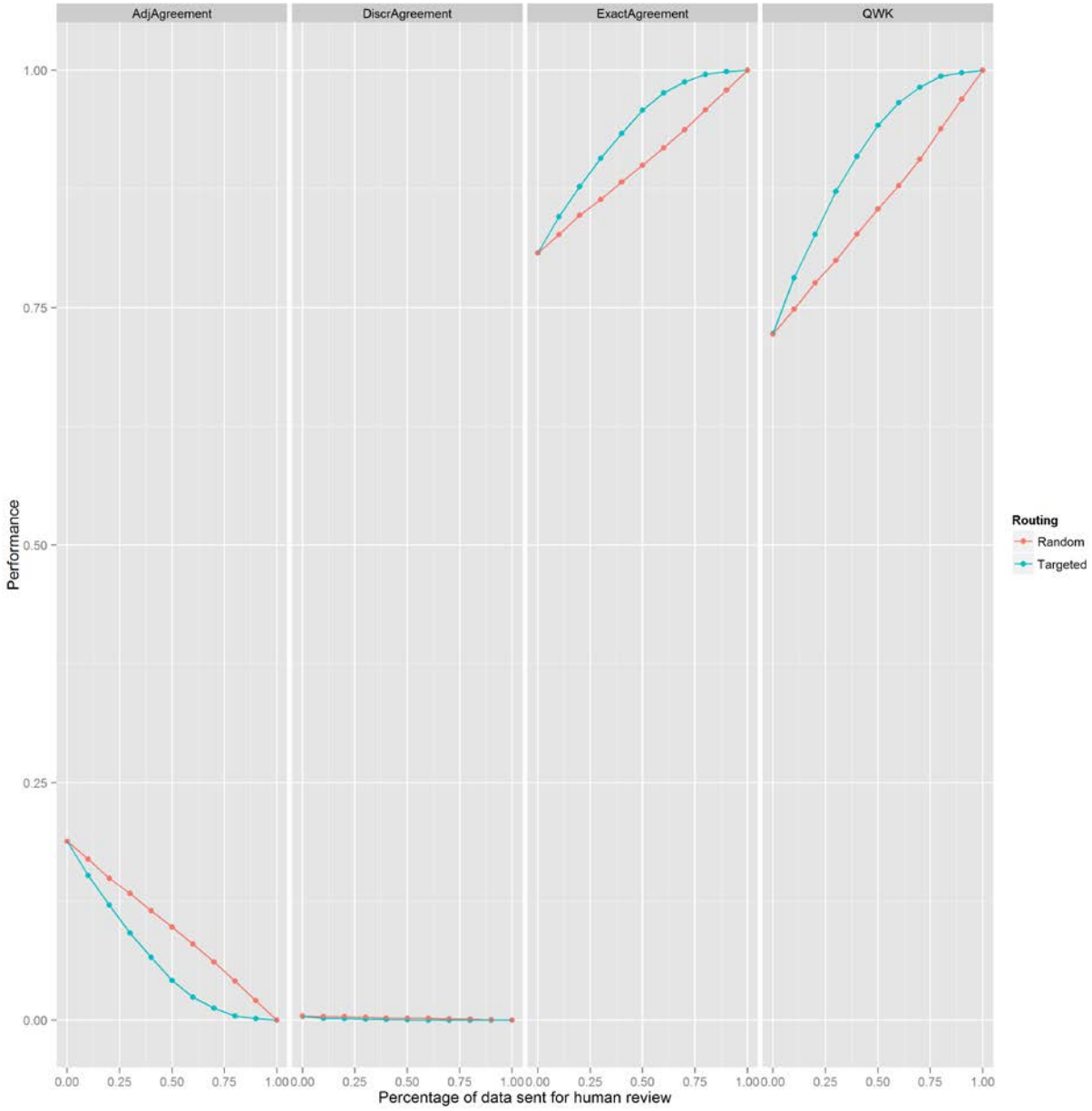


Figure 5.7. Comparison of Average Performance of Random vs Targeted Routing—Mathematics Short-text

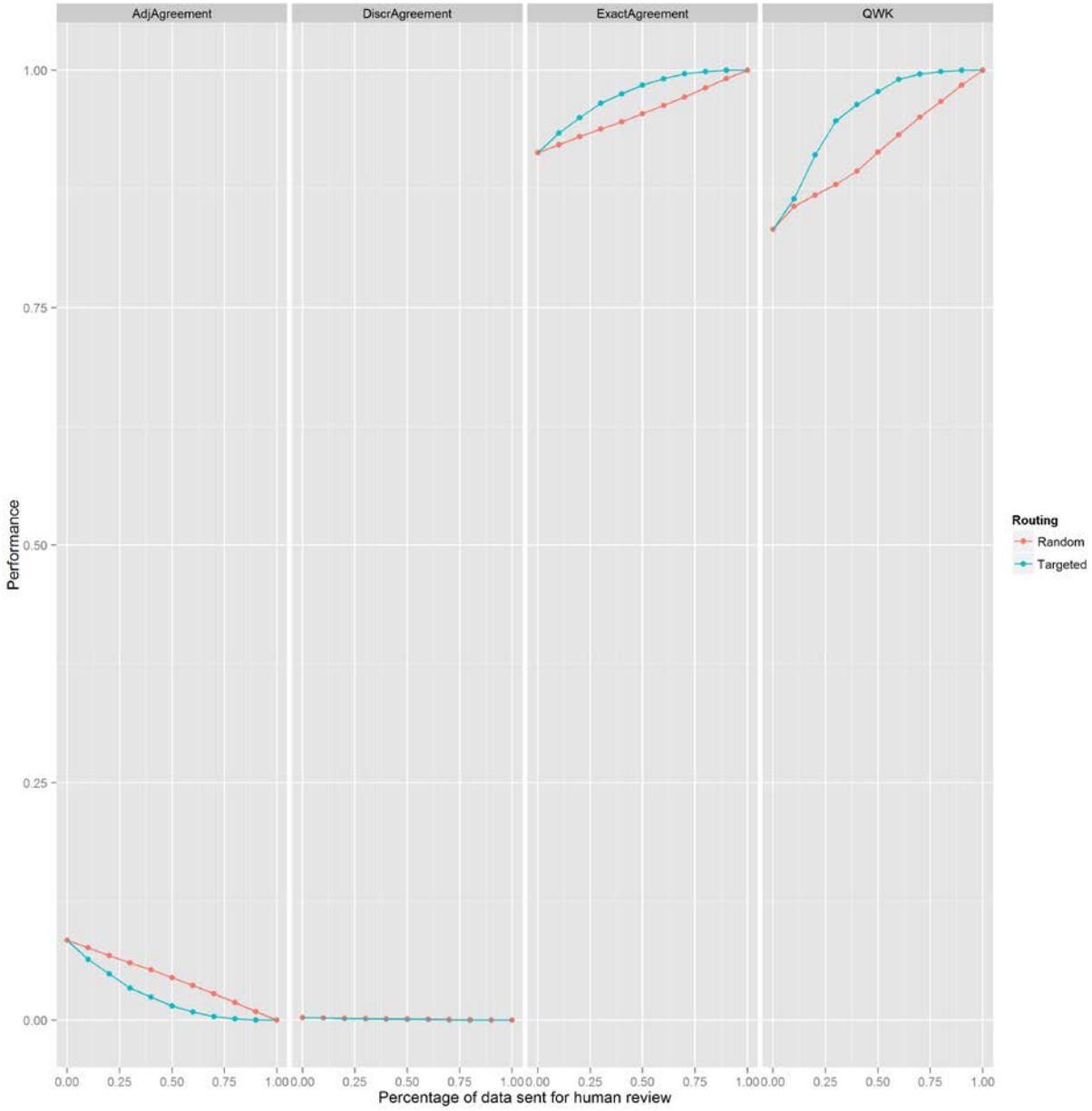
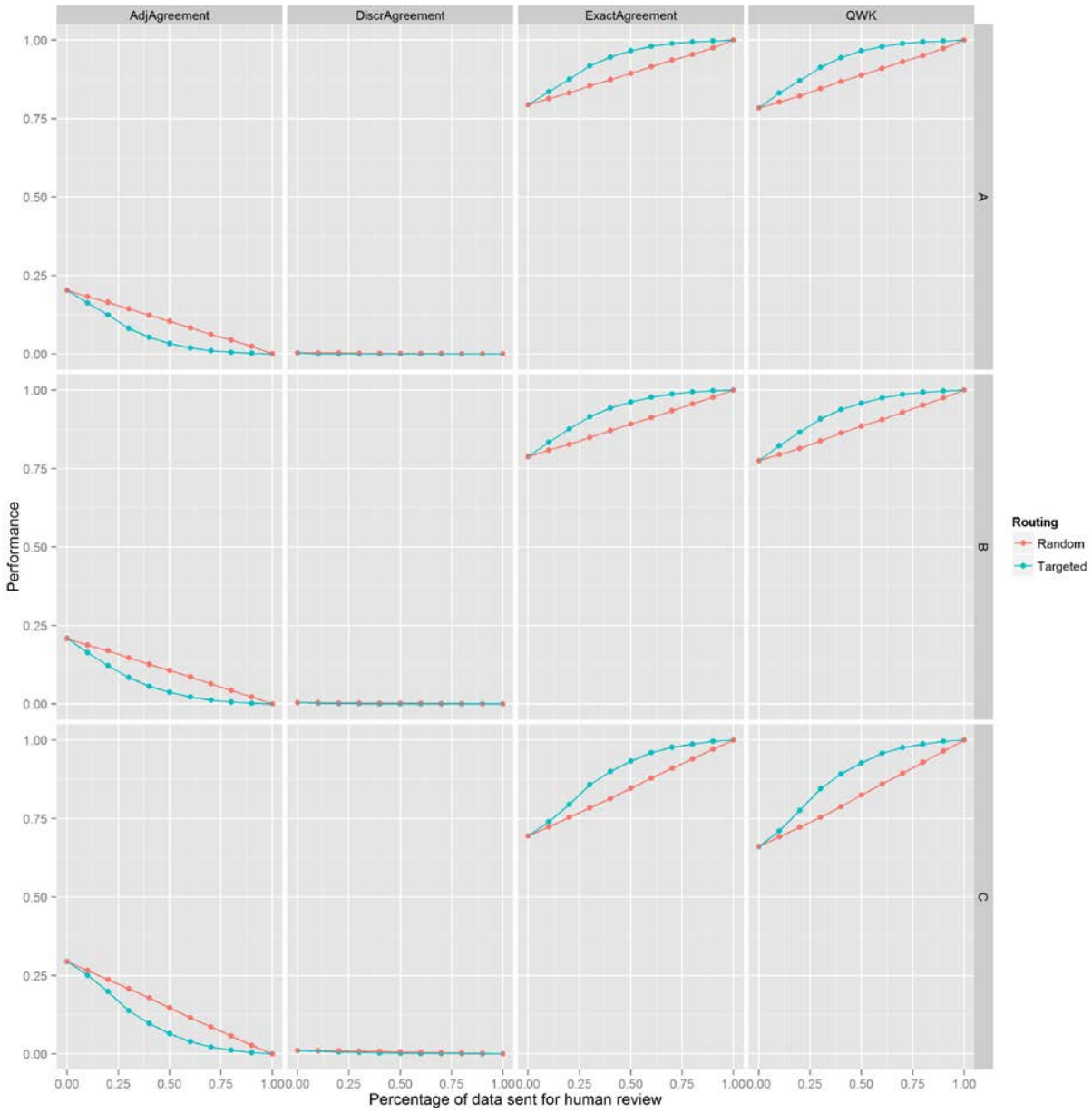


Figure 5.8. Comparison of Average Performance of Random vs Targeted Routing—ELA/literacy Essay



Chapter 5: Targeting Responses for Human Review

Discussion

In this study, we investigated whether it is possible to identify responses that are likely to be scored differently by Automated Scoring systems and humans. In a previous study, three methods were investigated. The most successful method was an outlier detection method to identify responses that were located in sparse regions of the feature space. The method consists of a principal component analysis followed by a partitioning around medoids analysis. The purpose of the principal component analysis is to reduce the dimensionality of the feature space. The partitioning around medoids identifies prototypical responses within the reduced feature space. Responses in sparse regions of the feature space are identified as responses with a relatively large distance to the closest prototypical response. In this study, we further developed this method.

In the previous study, both the number of principal components and the number of clusters were determined ad hoc. In this study, we successfully used cross-validation to determine these parameters. The method was validated on more than 60 items from three different item types and scored by two different vendors of Automated Scoring systems. The method was generally successful across item types and vendors.

Finally, we investigated the practical significance of our proposed method of flagging automatically scored responses for human review. Two scenarios were compared: a scenario where responses were routed based on the atypicality of those responses, and a scenario where responses were routed at random. Across item types and vendors, substantially fewer responses needed to be routed under the targeted routing scenario than under the random routing scenario to obtain a similar gain in agreement between the automated and human scores.

Chapter 6: Item Characteristics that Correlate with Agreement for Handscoring and Automated Scoring

Chapter 6: Item Characteristics that Correlate with Agreement for Handscoring and Automated Scoring

In the Smarter Balanced Pilot Study, *Toward Predicting Whether a Short-Text Constructed-Response Item Can Be Scored Using Automated Scoring Engines*, we found that there are characteristics of items correlated with the quadratic weighted kappas of both human-human agreement and engine-human agreement for English language arts (ELA)/literacy items. We saw similar trends for mathematics items, but the results were not statistically significant, possibly because there were too few mathematics items (see also Leacock, Messineo, & Zhang, 2013).

The item characteristics were derived from (1) the language of the prompt, (2) the item metadata, and (3) the scoring rubrics. That is, the characteristics can be identified before items are administered. In the research done during the Pilot Study, we began with the hypothesis, based on findings in Leacock et al. (2013), that human agreement would be affected by having a large number of relevant text-based examples to draw from in the reading passage (for Reading Comprehension items). As expected, when there are many possible text-based examples in ELA/literacy items, human-human and engine-human agreement both dropped.

Otherwise, the assumption that none of the variables would correlate with low human agreement was not confirmed. For ELA/literacy Depth of Knowledge (DOK), human-human and engine-human agreement was statistically significantly lower for DOK 3. Although statistical significance could not be tested for mathematics, there was a trend of greater disagreement for DOK 3. In the Pilot Study, there were no short-text items with DOK 4.

None of the other variables had statistically significant results for human-human and engine-human agreement but, again, there seem to be some trends. For both ELA/literacy and mathematics, agreement for Predicted Difficulty fell on the items that were predicted to be Hard. Similarly, there was lower agreement when the rubrics contained holistic language and involved clear inferences.

Purpose

The goal of this research study is to determine whether we can learn to predict, in advance of scoring, whether short-text, constructed-response, items can be scored using Automated Scoring systems. More specifically, this study has the following objectives:

1. Study the new ELA/literacy items that were developed for the Field Test: Reading Comprehension (Claim 1), Brief Writes (Claim 2), and Performance Task Research (Claim 4), Within each claim, we also look at the Targets and Writing Purposes (or writing genre) of the stimuli.
2. Include a larger number of mathematics items with the aim of getting statistical significance in the analyses.

There are many differences between the Pilot Study and Field Test items. As such, we have updated and redefined many of the item characteristics and are focusing on those characteristics that can be found in the metadata.

Chapter 6: Item Characteristics that Correlate with Agreement for Handscoring and Automated Scoring

Methodology

Data Source

The data set was comprised of responses to the 657 ELA/literacy and 233 mathematics short-text constructed response Field Test items that were selected for the Automated Scoring Special Studies. The numbers in the parentheses indicate the number of items in the training data.

ELA/literacy:

1. Grade Level:
   - Elementary: grades 3, 4, 5 (231)
   - Middle: grades 6, 7, 8 (217)
   - High: grade 11 (217)
2. There were three Smarter Balanced ELA/literacy *Depth of Knowledge* (DOK) levels. (There were no DOK 1 items among the Field Test items selected for the Automated Scoring studies. Since there were only eight DOK 2 items, they were not included in the study.):
   - DOK 3: Strategic Thinking (502)
   - DOK 4: Extended Thinking (155)
3. Predicted Difficulty:
   - Quintile 1 and 2—very easy and easy (126)
   - Quintile 3—medium (300)
   - Quintile 4 and 5—hard and very hard (239)
4. The Smarter Balanced *Claims* (there were no Claim 3—Speaking and Listening items in the short-response items):
   - Claim 1: Reading Comprehension: Read analytically—students can read closely and analytically to comprehend a range of increasingly complex literary and informational texts (258). The subtasks of summarizing/determining a central idea and inference/explanation will be studied separately.
   - Claim 2: Brief Writes: Write effectively—students can produce effective and well-grounded writing for a range of purposes and audiences. Students are asked to write introductions, conclusions, or are asked to elaborate text (210).
   - Claim 4: Source-based Performance Tasks—Conduct research—students can engage in research/inquiry to investigate topics and to analyze, integrate, and present information. Students are asked to do some research across texts and explain their answers using examples from multiple sources (197).

Claim 1 and Claim 2 used generic rubrics such that a single rubric could be used to score many different items. On the other hand, Claim 4 uses item-specific rubrics. Thus we characterize ELA/literacy Claim 1 and Claim 2 items as ELA-generic-rubric and Claim 4 as ELA-item-specific.

For the ELA/literacy items, each Claim was associated with a set of Targets and a set of Writing Purposes (genres).

The Targets for Claim 1 (Reading Comprehension) are:

- Targets 2&9: Students are asked to either summarize a story, or in most cases asks them to determine a central idea and explain using details from text (114).
- Targets 4&11: Asks for an inference or an explanation of the inference using evidence from the text (144).

The Writing Purposes for Claim 1 are:

- Informational (103)
- Literary (155)

The Targets for Claim 2 (Brief Writes) are:

- Elaboration (96)
- Organization-Conclusion (54)
- Organization-Introduction (60)

The Writing Purposes for Claim 2 are:

- Informational/Explanatory (63)
- Narrative (76)
- Opinion/Argumentative (71)

The Targets for Claim 4 (Performance Tasks) are:

- 2. Interpret and Integrate Information (92)
- 3. Evaluate Information/Sources (43)
- 4. Use Evidence (62)

The Writing Purposes for Claim 4 are:

- Informational/Explanatory (91)
- Narrative (42)
- Opinion/Argumentative (64)

Chapter 6: Item Characteristics that Correlate with Agreement for Handscoring and Automated Scoring

Mathematics: What follows are the item characteristics that were extracted automatically from the metadata for the mathematics short-text items (233).

1. Grade Level:
   - Elementary: grades 3, 4, 5 (111)
   - Middle: grades 6, 7, 8 (49)
   - High: grade 11 (72)
2. Score point range:
   - 1 (50)
   - 2 (155)
   - 3 (27)
3. There are four Common Core State Standards (CCSS) mathematics *Depth of Knowledge* (DOK) levels. DOK 1 (Recall) was not in the items selected for the Special Studies and DOK 4 (Extended Thinking) only had four items, so they were not included in the study.
   - DOK 2: Basic Application (48)
   - DOK 3: Strategic Thinking (181)
4. Predicted Difficulty, as predicted by the item writer:
   - Quintile 1 and 2—very easy and easy (41)
   - Quintile 3 (77) —medium
   - Quintile 4 and 5 (115) —hard and very hard
5. The Smarter Balanced *Claim* has four values. There were no Claim 1 items that were selected for the Special Studies and only eight Claim 2 (Problem Solving), so they were not included in the study.
   - Claim 3: Communicating reasoning—students can clearly and precisely construct viable arguments to support their own reasoning and to critique the reasoning of others (172).
   - Claim 4: Modeling and data analysis—students can analyze complex, real-world scenarios and can construct and use mathematical models to interpret and solve problems (53).

   All mathematics items have item-specific (as opposed to generic) rubrics with example answers and scoring rules for that item. Thus they were characterized as Math-item-specific-rubric.
6. Mathematical Reasoning (MR) item. Smarter Balanced contracted with Illustrative Mathematics to develop mathematics items to "push the field forward" in terms of finding ways to automatically score constructed responses. Thus they were, in part, designed to test the limits of automated scoring.

- Yes—a MR item (41)
- No—not a MR item (192)

To determine whether any of the characteristics can be used to predict accurate Automated Scoring of items, we used the general linear model to conduct an analysis of variance (ANOVA) analysis. For each characteristic, we performed the ANOVA test to compare the engine-human quadratic weighted kappa (QWK; Cohen, 1960, 1968) mean differences. A significant *p*-value would lead to the rejection of the null hypothesis and would indicate whether engines had more difficulty assigning scores when the feature characteristic had a certain classification. This process was repeated for human-human agreement.

Automated Scoring Engines

Five automated scoring vendors were used for this study. All used the same basic approach to Automated Scoring which is called a "bag of words" approach. First, they transformed the words and textually-based statistics into features and used them to generate prediction models using many machine learning methods, such as random forests, support vector machines, etc. The predicted scores of each method were subsequently treated as higher-order features that were combined in a second prediction model.

All of the Automated Scoring engines did some kind of preprocessing of the student responses including, but not limited to:

1. Spelling error correction. Unlike essays, where spelling and grammar are an integral part of the score, with short-text items, raters are typically instructed to look exclusively for content. When a word is misspelled, if a rater understands what the intended word is and that word contributes to a correct concept, then the response should receive the same score as it would if the word was not misspelled. Humans are extremely good at recognizing the intended word when a word is misspelled, especially within the context of a sentence or paragraph. People are so good at this that they often are not even aware of the presence of a spelling error. However, computers do not have this talent. When an Automated Scoring system encounters a misspelling, it treats it like a new word—that is, a computer treats *their* and *thier* as two completely unrelated words.
2. Word normalization—conversion of all alphabetic characters to either uppercase or lowercase.
3. Words were often converted to their base forms to reduce the total number of words being analyzed (for example, *run, ran,* and *running* were all represented as *run*). A similar technique that some systems use is *stemming*, which is cruder than morphological analysis and simply removes common suffixes. With stemming, *run* and *running* would be represented differently from *ran*.

For more detailed descriptions of the Automated Scoring systems and their preprocessing steps, see Chapter 3.

The short-text items were scored by five different vendors. Vendor 3 scored all 898 short-text constructed response items. Vendor 8 scored all of the 41 Mathematical Reasoning items. Vendor 2, Vendor 6, and Vendor 9 scored the same random sample of 21 ELA/literacy items and 21

mathematics items. In the Results section we report only the best score for each item.

Procedures

Agreement between the two human raters and between the resolved human score and the scoring engines was measured with QWK.

To determine whether the characteristics can be useful as a predictor of accurate Automated Scoring of items, we conducted an ANOVA analysis. For each characteristic, we performed the ANOVA test to compare the engine-human QWK mean differences. A significant *p*-value would lead to the rejection of the null hypothesis, and would indicate whether engines had more difficulty assigning scores when the feature contained a certain classification.

To determine whether any of the item characteristics could be used to predict accurate human scoring of items, we conducted an ANOVA. For each characteristic, we performed the ANOVA to compare the human-human QWK mean differences. A significant *p*-value would indicate whether a human had more difficulty assigning scores when the feature contained a certain classification.

For each ELA/literacy claim, to determine whether the characteristics could be used to predict accurate human and/or machine scoring of items, we conducted three multiple group *t*-tests on the Claim's Targets and three multiple group *t*-tests on the Claim's Writing Purposes.

In earlier experiments (Leacock, Messineo & Zhang, 2013 and Leacock & Zhang, 2014) our assumption was that human reliability will hold across the characteristics while the engine may not. However, in those experiments we found that engine-human and human-human agreement both tend to correlate with the same item characteristics.

The level of significance was established through a post-hoc comparison using the Tukey HSD (honestly significance difference) to achieve an overall level of significance.

Results

Automated Scoring and handscoring had statistically significant mean QWKs ($p<0.001$) for each content area— but in different directions. For ELA/literacy, the best Automated Scoring engines had a mean QWK of 0.73 compared to 0.70 for handscoring. For mathematics, handscoring had higher agreement— 0.85 mean QWK compared to Automated Scoring's 0.81 mean QWK.

English language arts/literacy

Overall, human and automated scoring showed similar behaviors. Table 6.1 shows the factors that have statistically significant effects for the ELA/literacy items. Both find it easier to score Claim 4 (Performance Task items) than Claim 1 (Reading Comprehension) and Claim 2 (Brief Writes). The best automated engines have an easier time scoring high school responses than they do elementary and middle school responses. Handscoring has a harder time with DOK 3 than DOK 4. Neither showed any effects for Predicted Difficulty. We did not test the number of score points because all of the ELA/literacy items were scored on a 3-point scale.

Chapter 6: Item Characteristics that Correlate with Agreement for Handscoring and Automated Scoring

Table 6.1. ANOVA Results for ELA/literacy

Engine-Human			Human-Human		
Factor	Value	Direction	Factor	Value	Direction
			DOK	DOK 4	Easier
				DOK 3	Harder
Grade level	High school	Easier			
	Middle School, Elementary School	Harder			
Claim	Claim 4	Easier	Claim	Claim 4	Easier
	Claim 1, Claim 2	Harder		Claim 1, Claim 2	Harder

One potential reason for having higher mean QWK, for both Automated Scoring and handscoring, is that Claim 4 (research items in Performance Tasks) have item-specific rubrics with examples of correct answers, whereas the rubrics for Claims 1 (Reading Comprehension) and 2 (Brief Writes) are generic. Although Automated Scoring systems do not understand rubrics in the same way that people do, they do make use of the specific language that is used in the rubrics and example answers.

The success with high school responses may be due, in part, to the length of these responses. The average length of a high school (11th grade) response is 525 characters (including spaces) while that of middle school is 450, and 290 for elementary school. Traditionally, automated scoring of short-text items is held to be more difficult than scoring essays because short-text items tend to have little or no repetition. With essays, an automated system can miss some aspect of the writing in one paragraph but has the chance to identify it elsewhere in the essay. With answers that are a sentence or two long, if the system misses something, it typically does not get another chance to capture that content.

It is not obvious why humans had a higher agreement on DOK 4 (Extended Thinking) than DOK 3 (Strategic Thinking). It may be because, as can be seen in Table 6.7, handscoring had somewhat higher mean QWKs on items of medium difficulty. Thirty-nine percent of the DOK 3 items were predicted to be of medium difficulty as opposed to only 62% in DOK 4. However, the predicted difficulty showed no significant differences for either the top engines or handscoring.

Targets and Writing Purposes within Claims:

Table 6.2 shows the significant effects for the Targets and Writing Purposes for each Claim. The asterisks indicate when the post-hoc test does not show statistically significant differences. For Claim 1 (Reading Comprehension), the best Automated Scoring systems had higher agreement when the reading stimulus was Literary as opposed to Informational—and handscoring showed a similar

trend. Neither showed any affects for the Target—between Central Ideas or Inference.

For Claim 2 (Brief Writes), both the Writing Purpose and the Target were statistically significant for both Automated Scoring and handscoring. For the Automated Scoring best engines, Elaboration was easier to score than Organization-Conclusion which, in turn, was easier to score than Organization-Introduction. Again, length may be an issue here because the average number of characters in an Elaboration response is 600 as compared to Conclusion (410) and Introduction (385). For handscoring, there was no difference between the two Organization Targets. Both found Brief Writes, where the student is adding to a Narrative essay, easier to score than Informational or Argumentative essays.

Finally, for Claim 4 (research items in Performance Tasks), people had higher agreement when the reading passages were narratives than when the stimuli were informational or gave different sides of an argument. The best Automated Scoring systems also showed this as a trend.

Table 6.2. ANOVA Results for ELA/literacy Targets and Writing Purposes (within Claims)

Engine-Human				Human-Human			
Claim	Target/ Purpose	Value	Direction	Claim	Target/ Purpose	Value	Direction
Claim 1	Writing Purpose	Literary	Easier	Claim 1*			
		Informational	Harder				
Claim 2	Target	Elaboration	Easiest	Claim 2	Target	Elaboration	Easier
		Organization-Conclusion	Harder			Organization-Conclusion, Organization-Introduction	Harder
		Organization-Introduction	Hardest				
	Writing Purpose	Narrative	Easier		Writing Purpose	Narrative	Easier
		Informational-Explanatory, Opinion-Argumentative	Harder			Informational-Explanatory, Opinion-Argumentative	Harder
Claim 4*				Claim 4	Target	Evaluate Information	Easier
						Interpret/ Integrate Information, Use evidence	Harder

Note: Post-hoc test does not show statistically significant differences.

Mathematics

Unlike ELA/literacy items, handscoring agreement was statistically significantly higher than the best engine scores for mathematics. There are many potential reasons for why mathematics items were a challenge for Automated Scoring systems. Mathematics items tended to elicit a specific complex concept or a chain or reasoning that could be expressed in a large number of ways. Another characteristic of the Smarter Balanced mathematics items was that the student responses tended to be short. The average length of the mathematics items was about 200 characters, as opposed to the average ELA/literacy character count of about 500.

There were no effects, for both humans or engines, on grade level, Depth of Knowledge, and Predicted Difficulty. The best engine had higher QWK for Claim 2 (Problem Solving) than for Claim 4 (Modeling and Data Analysis), but the post-hoc test did not find it statistically significant.

Handscoring did not show any effects between the Claims. In addition, handscoring, but not Automated Scoring, found it more difficult to score 2-point items. However, it could not be determined whether the 4-point items were more similar to the 2- or 3-point items. We were unable to test the mathematics Targets because there were so many (seven) of them with extremely skewed distributions, ranging from 1 to 89.

For both humans and engines, the QWK was significantly higher for the items that were not written specifically as Mathematical Reasoning (MR) items. However, this does not mean that non-MR items are not testing Mathematical Reasoning. As previously mentioned, the MR items were created, in part, to test the limits of Automated Scoring.

Table 6.3. ANOVA Results for Mathematics

Engine-Human			Human-Human		
Factor	Value	Direction	Factor	Value	Direction
Claim*					
Score points*			Score points	3, 4?	Easier
				2, 4?	Harder
Mathematical Reasoning	Yes	Harder	Mathematical Reasoning	Yes	Harder
	No	Easier		No	Easier

Note: Post-hoc test does not show statistically significant differences.

Since all but two of the Mathematical Reasoning items are Target B (Construct, autonomously, chains of reasoning that will justify or refute propositions or conjectures), we conducted a final test comparing Target B MR and Target B non-MR items. Again, both humans and engines had lower QWKs for the MR Items, but not at a statistically significant level. However, engine performance dropped much more than human performance: 4% drop in mean QWK for handscoring and 8% for the best engine—so there is some evidence that the Mathematical Reasoning items indeed pushed the limits of automated scoring.

Discussion

Overall, the mean QWK for mathematics items was much higher than for ELA/literacy. Handscoring had a mean QWK of 0.85 for mathematics and 0.69 for ELA/literacy while the best engines had a mean QWK of 0.81 for mathematics and 0.74 for ELA/literacy.

One potential explanation is the use of generic rubrics. Leacock, Gonzalez, and Connaroe (2013) conducted a study designed to validate assumptions about the requirements for Automated Scoring that demonstrates the effect of rubric clarity on score reliability. The investigators replaced holistic with specific language in analytic rubrics, exhaustively specified allowable content, and sharpened the rules for each score point for a sample of items. Following this revision effort, human-human

agreement rates increased from a mean QWK of 0.63 to 0.94 on a set of seven ELA/literacy items written to comply with the CCSS.

The rubrics for Reading Comprehension and for Brief Writes are not only holistic, but they are also generic —a single rubric can be used to score hundreds of items. The rubrics for the research Performance Tasks are item-specific and include an example answer. We conducted an ANOVA on rubric type and found a statistically significant difference ($p<.001$). The best scoring engines had a mean QWK of 0.78 for item-specific rubrics and 0.73 for holistic, generic rubrics. Handscoring had a mean QWK of 0.74 for item-specific rubrics and 0.67 for generic rubrics. All of the mathematics items also had item-specific rubrics which, again, may help to explain why both human and engine agreement is so much higher for mathematics than for ELA/literacy.

Further analyses needs to be carried out to try to understand why the Writing Purpose showed effects for Reading Comprehension and for Brief Writes. For Reading Comprehension, there was higher agreement when the text was fictional for Automated Scoring—and handscoring showed a similar trend. With Brief Writes, there was also higher agreement for narrative stimuli.

Chapter 6: Item Characteristics that Correlate with Agreement for Handscoring and Automated Scoring

Tables

Table 6.4. ANOVA ELA/literacy and Mathematics

Factor	df	Engine-Human		Human-Human	
		F	p	F	p
Content	1	46.97	< 0.001	412.29	< 0.001

Table 6.5. Mean QWK for Subject Area

Content	N	Engine-Human		Human-Human	
		Mean QWK	SD	Mean QWK	SD
ELA/literacy	657	0.74	0.09	0.69	0.09
Mathematics	218	0.81	0.18	0.85	0.11

Table 6.6. ANOVA for ELA/literacy Factors

Factor	df	Engine-Human		Human-Human	
		F	p	F	p
Quintile	2	0.00	1.00	0.10	0.90
DOK	1	0.75	0.39	4.89	0.03
Grade-level	2	12.17	< 0.001	19.14	< 0.001
Claim	2	10.80	< 0.001	0.38	0.68

Table 6.7. Mean QWK for ELA/literacy factors

Factor	Value	N	Engine-Human		Human-Human	
			Mean QWK	SD	Mean QWK	SD
Quintile	Easy	124	0.73	0.08	0.67	0.09
	Medium	298	0.76	0.10	0.71	0.09
	Hard	235	0.73	0.10	0.67	0.10
DOK	3	502	0.74	0.09	0.68	0.10
	4	155	0.77	0.06	0.73	0.08
Grade-level	Elementary School	226	0.73	0.11	0.69	0.11
	Middle School	216	0.74	0.08	0.69	0.08
	High School	215	0.77	0.07	0.69	0.08
Claim	1	250	0.73	0.10	0.68	0.09
	2	210	0.73	0.09	0.66	0.09
	4	197	0.78	0.07	0.74	0.08

Table 6.8. T-tests for Targets and Writing Purposes within Claims

Claim	Factor	df	Engine-Human		Human-Human	
			F	p	F	p
Claim 1	Target	1	1.16	0.28	2.72	0.14
	Writing Purpose	1	5.87	0.02	4.30	0.04
Claim 2	Target Language	2	45.22	< 0.001	6.39	0.00
	Writing Purpose	2	33.70	< 0.001	20.77	< 0.001
Claim 4	Target	2	0.76	0.47	3.44	0.03
	Writing Purpose	2	1.19	0.31	0.64	0.53

Table 6.9. Mean QWK for Targets and Writing Purpose within Claims

Claim	Target	Writing Purpose	N	Engine-Human		Human-Human	
				Mean QWK	SD	Mean QWK	SD
1: Reading Comprehension	2&9: Central Idea	Informational	38	0.73	0.08	0.65	0.10
	2&9: Central Idea	Literary	68	0.75	0.08	0.68	0.10
	4&11: Inference	Informational	64	0.71	0.12	0.67	0.09
	4&11: Inference	Literary	80	0.74	0.09	0.69	0.09
2: Brief Writes	Elaboration	Informational/ Explanatory	29	0.77	0.07	0.68	0.09
	Elaboration	Narrative	37	0.81	0.06	0.72	0.08
	Elaboration	Opinion/ Argumentative	30	0.74	0.07	0.64	0.08
	Organization–Conclusion	Informational/ Explanatory	17	0.67	0.07	0.59	0.10
	Organization–Conclusion	Narrative	20	0.78	0.05	0.71	0.04
	Organization–Conclusion	Opinion/ Argumentative	17	0.68	0.03	0.62	0.06
	Organization–Introduction	Informational/ Explanatory	17	0.64	0.07	0.59	0.09
	Organization–Introduction	Narrative	19	0.73	0.06	0.68	0.09
	Organization–Introduction	Opinion/ Argumentative	24	0.65	0.06	0.64	0.07
4: Research terms in Performance Tasks	2	Informational/ Explanatory	46	0.78	0.09	0.74	0.10
	2	Narrative	17	0.79	0.08	0.75	0.08
	2	Opinion/ Argumentative	29	0.75	0.06	0.71	0.08

Claim	Target	Writing Purpose	N	Engine-Human		Human-Human	
				Mean QWK	SD	Mean QWK	SD
4: Research terms in Performance Tasks	3	Informational/ Explanatory	16	0.78	0.08	0.75	0.09
	3	Narrative	14	0.78	0.05	0.78	0.07
	3	Opinion/ Argumentative	13	0.80	0.06	0.79	0.07
	4	Informational/ Explanatory	29	0.79	0.05	0.73	0.06
	4	Narrative	11	0.75	0.07	0.73	0.06
	4	Opinion/ Argumentative	22	0.76	0.07	0.73	0.06

Table 6.10. ANOVA for Mathematics Factors

Factor	df	Engine-Human		Human-Human	
		F	p	F	p
Quintile	2	2.49	0.09	2.23	0.11
DOK	1	0.01	0.98	1.07	0.30
Grade-level	2	2.39	0.09	2.53	0.08
Claim	1	3.97	0.05	3.49	0.06
Score Points	2	3.49	0.03	9.87	< 0.001
Mathematical Reasoning	1	14.64	< 0.001	18.04	< 0.001

Table 6.11. Mean QWK for Mathematics Factors

Factor	Value	N	Engine-Human		Human-Human	
			Mean QWK	SD	Mean QWK	SD
Quintile	Easy	35	0.85	0.14	0.86	0.09
	Medium	111	0.79	0.20	0.83	0.13
	Hard	72	0.81	0.15	0.86	0.10
DOK	2	40	0.83	0.18	0.85	0.12
	3	178	0.80	0.17	0.85	0.11
Grade-level	Elementary School	100	0.82	0.17	0.85	0.12
	Middle School	69	0.78	0.18	0.83	0.13
	High School	49	0.81	0.17	0.86	0.09
Claim	3	169	0.81	0.17	0.85	0.11
	4	49	0.78	0.19	0.84	0.12
Score Points	1	44	0.76	0.15	0.79	0.12
	2	149	0.82	0.19	0.86	0.11
	3	25	0.79	0.12	0.85	0.08
Mathematical Reasoning	No	179	0.82	0.17	0.86	0.12
	Yes	39	0.74	0.18	0.81	0.10

Table 6.12. ANOVA for Mathematical Reasoning within Target B

Factor	df	Engine-Human		Human-Human	
		F	p	F	p
Mathematical Reasoning	1	1.05	0.31	0.20	0.66

Table 6.13. Mean QWK for Mathematical Reasoning within Target B

Factor	Value	N	Engine-Human		Human-Human	
			Mean	SD	Mean	SD
Mathematical Reasoning	No	47	0.78	0.21	0.82	0.15
	Yes	39	0.74	0.18	0.81	0.10

Table 6.14. ANOVA for Generic vs Item-specific Rubrics (ELA/literacy)

Factor	df	Engine-Human		Human-Human	
		F	p	F	p
Type of Rubric	1	39.51	< 0.001	90.22	< 0.001

Table 6.15. Mean QWK for ELA/literacy Type of Rubric

Content	N	Engine-Human		Human-Human	
		Mean QWK	SD	Mean QWK	SD
Item-specific Rubric	197	0.78	0.07	0.74	0.08
Generic Rubric	460	0.73	0.09	0.67	0.09

Chapter 7: Generic Scoring Models

The Smarter Balanced essay-length responses are scored on three traits:

A. Organization/Purpose

B. Evidence/Elaboration

C. Conventions

Scores for both trait A and trait B are, for the most part, based on the content of the essay. This is not the case with trait C. The rubric for trait C covers spelling, capitalization, punctuation, grammar/ usage and sentence completeness on a three-point grading scale—without regard to specific content. Trait C has grade-level rubrics for grades 3 through 8 and another for high school. For example, 4th graders are assessed for the ability to use relative pronouns (*who*, *whose*, etc.) while 3rd graders are not.

Statistically-based Automated Scoring systems are data-hungry, and as such require large amounts of data for training and validating the scoring models. Since trait C is not prompt-specific, it makes sense to experiment with developing a non-prompt-specific generic model for scoring.

Purpose

While most Automated Scoring vendors provide generic models for scoring essays, it is generally agreed that models trained on individual prompts give more accurate scores. This is because the model is sensitive to prompt-specific language, as opposed to genre-specific language (e.g., opinion essays or narrative essays). For a trait C model, there is no need to identify specific language—it is the *form*, not the content, of the essay that is being assessed.

Methodology

Many of the features required for generating a writing conventions score can be identified by good-quality spelling and grammar checkers. Other features can be developed based on counts/ratios in the training essays. For example, if an 8th grade essay consists entirely of sentences that are two and three words long, it is unlikely to be assigned a high trait C score.

It is well known that grammar checkers do not catch all errors. Most, if not all, grammar checkers are designed to favor precision over recall as described in Leacock et al. (2014). That is, they would rather miss true errors than flag too many good constructions as being in error. Even so, grammar checkers that have accuracy comparable to that of Microsoft Word have been shown to be effective in both automated essay scoring and in formative feedback to students (Burstein et al., 2013).

There are several open-source grammar checkers currently available, the most familiar probably being LanguageTool. Since it is likely that their quality is lower than commercially available grammar checkers, we evaluated LanguageTool's accuracy and effectiveness. We also evaluated several commercially available grammar checkers and, for the purposes of this study, licensed the one that we found to be the most effective in terms of identifying the most errors with the fewest false positives (identifying a correct construction as being an error). To give a sense of the difference between LanguageTool and the commercial grammar checker that we licensed, on a set of 10 randomly selected essays at various grade levels, LanguageTool accurately identified 19 errors with no false positives while the commercial checker accurately identified 65 errors with five false

positives. Since the commercial grammar checker identified so many more errors than LanguageTool, we were able to develop five features: grammar/usage errors, white space errors, capitalization errors, punctuation errors, and stylistic suggestions. Due to the sparseness of errors identified by LanguageTool, we developed a single grammar/usage feature.

We developed spelling error features using a licensed spell checker. Due to the construction of the scoring engine for one of the generic models we developed, we were unable to add unknown words to the dictionary. In the second generic model, we were able to add unknown words. The language that we added were words that appeared in the reading passages that the spell checker did not recognize—primarily people and place names and technical terms.

Data Source

The models were built based on 16 Smarter Balanced Field Test essay items—eight 6th grade prompts and eight 11th grade prompts. The outcome was evaluated using the quadratic weighted kappa (QWK). We compared the results with the trait C scores of humans and of the Smarter Balanced Field Test Automated Scoring Vendor 1.

Procedures

For each grade (6 and 11), there were eight prompts with approximately 1,500 training responses for each prompt. To evaluate the generic scoring models, 56 sets were created by taking all the possible combinations of five items out of eight per grade, hence, each training set was comprised of the responses for five of the items. The three unseen prompts were tested on each training set using the Logistic Regression model or the Random Forest model. The classification and overall methodology was similar to the Automated Scoring system. We tested three versions of the features.

1. Vendor 1: The Vendor 1 feature set: includes LanguageTool and a spell checker where words were not added to the dictionary.
2. Enhanced Vendor 1: The Vendor 1 features set—but replacing LanguageTool with a proprietary grammar checker and adding unknown words to the dictionary.
3. Combined Generic: All of the features in items 1 and 2 above.

Using the best system, we then ran an additional generic scoring experiment—training on the training data and testing on the held-out test sets.

As a final experiment, we built a single model for each grade, training on the training data from all eight prompts and testing on the held out test sets for the eight prompts for each grade.

Results

Table 7.1 shows the results for the unseen items only. Of the three feature sets, the combined version was always the best. The Enhanced Vendor 1 version increased Vendor 1 Grade 6 results by almost 2% and Grade 11 results by a little over 1%. Combining the feature sets improved both grades by about 0.05%. The results were stronger for 6th grade than for 11th grade.

Table 7.14. Results From Three Feature Sets

Grade 6				Grade 11			
Item ID	Vendor 1 Generic Model	Proprietary Generic Model	Combined Generic Model	Item ID	Vendor 1 Generic Model	Proprietary Generic Model	Combined Generic Model
56561	0.69	0.70	0.72	54155	0.70	0.71	0.72
61076	0.77	0.81	0.81	54163	0.67	0.71	0.70
61462	0.75	0.78	0.78	54223	0.66	0.67	0.67
61651	0.73	0.75	0.74	54729	0.67	0.70	0.70
61827	0.72	0.75	0.76	56388	0.68	0.69	0.70
61969	0.79	0.80	0.81	56396	0.67	0.69	0.69
61971	0.80	0.81	0.81	56398	0.71	0.72	0.73
61977	0.78	0.79	0.80	56543	0.66	0.67	0.68
Average	0.76	0.77	0.78	Average	0.68	0.69	0.70

Table 7.2 shows the difference between the generic models and item-specific models. The results for the generic models only include the unseen prompts. There are, of course, no unseen prompts in the item-specific models. An unexpected result is that, in all but three prompts, the generic scoring model was better than the prompt-specific model. Thus the error and other features found in other items can inform the trait C score in unseen prompts.

Table 7.15. Comparison of Generic and Item-Specific Models

Grade 6				Grade 11			
Item ID	Combined Generic Model (Avg QWK)	Combined Prompt Specific Model (Avg QWK)	Writing Genre	Item ID	Combined Generic Model (Avg QWK)	Combined Prompt Specific Model (Avg QWK)	Writing Genre
56561	0.72	0.66	Opinion/ Argumentative	54155	0.72	0.76	Opinion/ Argumentative
61076	0.81	0.76	Informational/ Explanatory	54163	0.70	0.74	Opinion/ Argumentative
61462	0.78	0.78	Informational/ Explanatory	54223	0.67	0.66	Opinion/ Argumentative
61651	0.74	0.71	Informational/ Explanatory	54729	0.70	0.68	Opinion/ Argumentative
61827	0.76	0.73	Informational/ Explanatory	56388	0.70	0.67	Informational/ Explanatory
61969	0.81	0.79	Narrative	56396	0.69	0.69	Opinion/ Argumentative
61971	0.81	0.82	Narrative	56398	0.73	0.71	Informational/ Explanatory
61977	0.80	0.79	Informational/ Explanatory	56543	0.68	0.66	Informational/ Explanatory
Average	0.78	0.76		Average	0.70	0.70	

For item 61971, in 6th grade, the decline in performance is negligible. However, for 11th grade items 54155 and 54163, the decline of the generic model is about 4%. Again, we need to understand why the results are so much stronger for 6th grade than for 11th grade. We did not have enough items, per grade, to sample writing genres. Fourth grade items had a single Opinion/Argumentative prompt while the 11th grade items had five of them. And both of the items that had strong declines are Opinion/Argumentative prompts.

Table 7.3 compares the results for prompt-specific models, generic models trained on five prompts and tested on three unseen prompts and when the generic model is trained and tested on all eight prompts. For the most part, generic models with unseen items have lower average QWKs than those with no unseen items. For 6th grade, the average difference was one half of a percent while for 11th grade the average difference in the QWK was 1%.

Table 7.16. Comparison of prompt-specific model, training of 5 items (with unseen items) and training on all 8 items.

Grade 6				Grade 11			
Item ID	Combined Prompt Specific Model (Avg QWK)	Combined Generic Model (Avg QWK)	Train on All Prompts (Avg QWK)	Item ID	Combined Prompt Specific Model (Avg QWK)	Combined Generic Model (Avg QWK)	Train on All Prompts (Avg QWK)
56561	0.66	0.72	0.71	54155	0.76	0.72	0.74
61076	0.76	0.81	0.81	54163	0.74	0.70	0.71
61462	0.78	0.78	0.78	54223	0.66	0.67	0.68
61651	0.71	0.74	0.75	54729	0.68	0.70	0.72
61827	0.73	0.76	0.76	56388	0.67	0.70	0.70
61969	0.79	0.81	0.82	56396	0.69	0.69	0.70
61971	0.82	0.81	0.82	56398	0.71	0.73	0.73
61977	0.79	0.80	0.81	56543	0.66	0.68	0.69
Average	0.76	0.78	0.78	Average	0.70	0.70	0.71

However, overall, the generic models showed improvement over item-specific models. These results suggest that we can train a single model for each grade—instead of having to create a model for each item and still achieve better results compared to prompt specific models.

Discussion

From the above results based on QWK we can conclude that it is very likely that generic scoring methods outperform prompt-specific scoring methods, given a specific grade, in estimating the scores for trait C. The probable explanation for this affect is that when combining the training sets in generic scoring, the training model observes many more possible types of grammatical and convention related errors per given score point than it can find in a single item, and computes the classification parameters based on better knowledge of errors than it may encounter when being tested.

In the cases where the training data for all of the items are available per given grade, the results in Table 7.3 also suggests that it is better to train one classifier model on multiple training items to evaluate the test sets compared to the item specific training for the same reasons as mentioned above.

The two generic methods, namely (1) Generic model trained on five items and (2) Generic model trained on all eight items, have advantages over prompt-specific training in saving a significant amount of time for computing the classifier models and parameters, which in turn could be a very

cost-effective solution when training several items in terms of infrastructure expenses while, at the same time, delivering better results.

As collecting training data is expensive, further research is needed to investigate at what point the learning curve flattens, and also to learn at what point there will be diminished returns from increasing the size of the training set. In addition, we need to learn whether the number of different items in the training data affect the results. Further research will also be focused on determining the cause of the differences in the results between 6th and 11th grades, based on the ranks of the independent strongest features in each of the grades. It will also be focused on investigating whether genre plays any significant role in determining the performance trend between generic and prompt-specific scoring.

References

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

Bridgeman, B. (2013). Human ratings and automated essay evaluation. In M. Shermis & J. Burstein (Eds.), *Handbook of automated essay evaluation: Current applications and new directions* (pp. 222–232). New York: Taylor & Francis.

Burstein, J., Tetreault, J., & Madnani, N. (2013). The e-rater Automated Essay Scoring System. In M.D. Shermis & J. Burstein (Eds.), *Handbook of automated essay evaluation: Current applications and new directions* (pp. 55–67). New York: Taylor & Francis.

Clough, P., & Stevenson M. (2011). Developing a corpus of plagiarised short answers. *Language Resources and Evaluation, 45,* 5–24.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20,* 37–46.

Cohen, J. (1968). Multiple regression as a general data-analytic system. *Psychological Bulletin, 70,* 426–443.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.). New York: Springer.

Higgins, D. (2013). *Proposed Smarter Balanced criteria for CR item acceptance related to suitability for automated scoring*. Unpublished manuscript.

Husson, F., Josse, J., & Pagès, J. (2010). *Principal component methods—hierarchical clustering—partitional clustering: Why would we need to choose for visualizing data?* Technical Report, Agrocampus Ouest, France.

Leacock, C., Chodorow, M., Gamon, M., & Tetreault, J. (2014). *Automated grammatical error detection for language learners* (2nd ed.). San Rafael, CA: Morgan & Claypool.

Leacock, C., Gonzalez, E., & Conarroe, M. (2014). *Developing Effective Scoring Rubrics for Automated Short-Response Scoring*. Monterey, CA: CTB/McGraw-Hill.

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., & Zhang, X. (2014, April). *Identifying predictors of machine/human reliability for short-response items.* Paper presented at the annual conference of the National Council on Measurement in Education, Philadelphia, PA.

References

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013). *Distributed representations of words and phrases and their compositionality.* Retrieved December 23, 2014 from http://arxiv.org/abs/1310.4546

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

Shermis, M. D. (2013). *Contrasting state-of-the-art in the machine scoring of short-form constructed responses.* Paper presented at the annual conference of the National Council on Measurement in Education, San Francisco, CA.

Silver, N. C., & Dunlap, W. P. (1987). Averaging correlation coefficients: Should Fisher's z transformation be used? *Psychological Methods, 1*, 178–183.

Trapani, C., Bridgeman, B., & Breyer, J. (2011, April). *Using automated scoring as a trend score: The implications of score separation over time*. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.

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

Wolpert, D. H. (1992). Stacked generalization. *Neural Networks, 5*, 241–259.

Zbontar, J. (2012). *Short answer scoring by stacking.* Retrieved December 23, 2014 from https://kaggle2.blob.core.windows.net/competitions/kaggle/2959/media/jzbontar.pdf

Zhang, M. (2013). *Contrasting automated and human scoring of essays.* ETS R&D Connections, 21. Princeton, NJ: ETS.