Smarter Balanced Assessment Consortium

Pilot Test: Automated Scoring Research Design

in accordance with Smarter Balanced RFP 17

CTB/McGraw-Hill Education

July 18, 2013

Table of Contents

Table of Tables

Chapter 1: Automated Scoring Plan

The purpose of this document is to overview the research studies that will occur in the area of Automated Scoring for the Smarter Balanced Pilot test. The research plan describes CTB's work to investigate the Automated Scoring of various item types and Automated Scoring engines that are a component of this project. For the purposes of the Smarter Balanced work, CTB will refer to Automated Scoring as a variety of methods, including statistical techniques, natural language processing, and machine learning, used to computationally derive a score for a constructed-response item. Some of these techniques are related to the field of artificial intelligence (AI). Because scoring methods vary across the engines that will be deployed for the Smarter Balanced Assessment Consortium pilot, some of which do not use AI techniques, CTB will use the term Automated Scoring rather than AI in this document. Chapter 1 describes the item types as they are defined in the data files delivered from the American Institutes for Research (AIR) with the purpose of defining and consistently using the same terms when moving from reviews of data to the CTB research plan. The Smarter Balanced Assessment Consortium pilot administration source data stems from two main files: an item metadata file that describes the content domain, claim, target, and standards an item measures and other item attributes (e.g., 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, Chapter 1 describes how items are selected for handscoring, Automated scoring, or validation only studies. Chapter 2 describes the criteria CTB will use to evaluate the functioning of the Automated Scoring of items. Chapters 3–8 provide the proposals for the research studies that will be investigated to enhance Smarter Balanced's efficacy in its use of Automated Scoring.

Item Descriptions

Item metadata was available for all 5,412 pilot items in the test delivery system, 1,494 of which require hand-scoring by CTB and our scoring sub-contractors, either to produce the score of record or to prepare Automated Scoring engine training and validation sets. The remaining 3,918 items were identified as selected-response and simple and complex technology enhanced items (including graphic response items). These technology enhanced items will be sent through rubric validation as per the Scoring Plan. Selected-response and technology enhanced items are scored by AIR's test delivery system.

Table 1.1 shows the constructed-response item type as identified in metadata files for the 1,494 items, a description of the item type, the number of items coded with the item type by content area, and the available Automated Scoring engine developers for each item type. As noted in the table, some of these developers participated in the Automated Student Assessment Prize (ASAP) competition.

The Automated Scoring research in which CTB engages in on behalf of the Smarter Balanced Assessment Consortium will focus primarily on the Natural Language (NL), Short Answer (SA), and Essay type items; however, we will also score and evaluate novel validation responses (responses not seen during rubric validation) for Equation (EQ) type items. Both the NL and SA items use Natural Language Processing Techniques; the distinction has only to do with the presence of a machine rubric specific to the AIR proposition engine for the NL items, and not for the SA items.

Table 1.1. Description of Pilot Constructed-Response Item Types and Scoring Engines

Constructed-Response Type	Description	Number of ELA/literacy Pilot items	Number of MA Pilot items	Available Scoring Engines*
NL	Short constructed-responses, text only. Machine scoring rubric for AIR's proposition scoring engine available.	515	128	AIR-Proposition Measurement Incorporated Luis Tandalla (ASAP 1) Jure Zbontar (ASAP 2) Xavier Conort (ASAP 3) James Jesensky (ASAP 4) Pawel Jankiewicz (ASAP 5) LightSIDE
SA	Short constructed-responses, text only. Machine scoring rubric for AIR's proposition scoring engine unavailable.	47	200	Measurement Incorporated Luis Tandalla (ASAP 1) Jure Zbontar (ASAP 2) Xavier Conort (ASAP 3) James Jesensky (ASAP 4) Pawel Jankiewicz (ASAP 5) LightSIDE
Essay	Extended constructed-responses, text only.	49	0	AIR-Open Source Engine Measurement Incorporated CTB-Bookette LightSIDE Luis Tandalla (ASAP 1) Jure Zbontar (ASAP 2) Xavier Conort (ASAP 3) James Jesensky (ASAP 4) Pawel Jankiewicz (ASAP 5)
EQ	Equation responses. Unlimited possibility for student response.	0	555	AIR-EQS (AIR Equation System)

*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.

Automated Scoring Training and Validation Plan

Response sampling and routing.

To sample and route responses to the scoring clients, CTB assigned each item to one of five cases based upon engine validation and research needs and the number of on-grade responses available. Note that Case 0 is the case for selected-response and technology-enhanced items, for which no human scoring will be conducted.

Descriptions of Cases 1 through 4 and items included in those cases follow, and are represented in Table 1.2:

1. *Case 1. Hand-scoring only*. CTB selected a stratified random sample of 1,800 responses of the available on-grade responses per item for a single human read. Five percent of these responses will be scored with a second read for inter-rater reliability purposes. Items include the following:
    a. NL, SA, and Essay Performance Task items designed for grades 4, 7, and 11. This enabled pre range-finding activities to occur for these grades prior to receipt of the full response data from the test delivery vendor, as in other cases listed below the range-finding data must be a subset of the Automated Scoring training set, which could not be sampled until the full set of responses was received.
    b. Any grade level NL, SA, or Essay items with fewer than 1,500 on-grade responses. For the pilot, CTB set a threshold of 1,500 on-grade responses for the Automated Scoring efforts to allow for 1,000 training responses and 500 validation responses based on CTB's and our sub-contractor's previous experience. Studies during the pilot phase to determine optimal training and validation set size are described in chapter 3 of this document and will inform this threshold for the field test.
2. *Case 2. Validation only*. CTB selected stratified random sample of 500 responses of the available on-grade responses per item. Each response will receive two human reads and adjudication of any non-exact scores by a senior human rater. The score of record will be the senior rater for adjudicated responses or the matched score when the two human scores agree. Scores of record will be compared against engine scores to validate the engine. Items include:
    a. Equation (EQ) items designed for the AIR equation engine with at least 100 responses available for validation. This engine has not yet received a validation against highly validated human scores on novel responses, whereas other technology enhanced scoring methods have and rubric validation will therefore be sufficient.
3. *Case 3. Automated Scoring training and validation*. CTB selected a stratified random sample of 1,800 responses of the available on-grade responses per item. Of these, a random sample of 1,000 responses will be designated as training responses and 500 responses will be designated as validation responses. The training and validation responses will receive two human reads and adjudication of any non-exact scores by a senior human rater. The score of record will be the senior rater for adjudicated responses or the matched score when the two human scores agree. Scores of record will be

compared against engine scores for validation purposes. The remaining 300 responses will be scored by the qualified engines only. Items include:

a. NL, SA, and Essay Performance Task items designed for grades 3, 5, 6, 8, 9, and 10 with at least 1,500 on-grade responses. Choosing these grade levels of performance tasks allowed CTB to receive the full response data from the test delivery platform prior to pulling range-finding and rubric validation samples. This was important as the range-finding and rubric validation samples must be a subset of the Automated Scoring training sample to ensure comparability of AIR's proposition scoring engine, which uses a rubric validation process, with the other Automated Scoring engines, which use AI-type training processes, on the same items. In order to sample the training set, the full response data was required.

b. All non-performance task (CAT) NL and SA items with at least 1,500 on-grade responses. Again, the rubric validation sample for the NL items is the same as the range-finding set, and a subset of the training set, to ensure alignment of human and machine training materials and allow for AIR to perform rubric validation and the other Automated Scoring vendors to train their engines on the same data.

4. *Case 4. Automated Scoring training and* validation, and special studies. CTB selected stratified random sample of 2,500 of the available on-grade responses per item. Of these, a random sample of 1,500 responses will be designated as training responses and a random sample of 1,000 of these responses will be designated as validation responses. The score of record will be the senior rater for adjudicated responses or the matched score when the two human scores agree. Scores of record will be compared against engine scores for validation purposes. Only items for which at least 2,500 on-grade responses are available are eligible for this case as the 2,500 sample size is important to the research studies defined later in this document. The items in this category are of NL, SA, and Essay type. Note that the random selection of these items from the Case 3 items provides for the comparison of engine performance and the study of item characteristics to engine effectiveness. Items include:

a. A random selection of the Case 3 items with at least 2,500 on-grade responses.
   i. 5 essay items
   ii. 2 English language arts (ELA)/literacy NL items
   iii. 3 ELA/literacy SA items
   iv. 2 mathematics NL items
   v. 3 mathematics SA items

Table 1.2. Description of Cases

Case	Hand Scoring	Hand Scoring Set Size	Scoring Engine	Training Set Size	Validation Set Size	Score from Engine Only Size
0	No	NA	No	NA	NA	NA
1	Yes	1,800	No	NA	NA	NA
2	Yes	V set only	Yes	0	500	All Available
3	Yes	T&V set only	Yes	1,000	500	300
4	Yes	T&V set only	Yes	1,500	1,000	NA

Table 1.3 and 1.4 provides the total number of items selected in each of Cases 1, 2, 3, and 4 by item type and performance task membership once the above rules were applied.

Table 1.3. Number of Items of Each Score Type and Assigned Scoring Case, ELA/literacy.

Content Area	Constructed-Response Type	PT	Assigned Case	Number of Items Assigned
ELA/literacy	Essay	Y	1	32
			3	12
			4	5
	NL	N	1	1
			3	372
			4	2
		Y	1	95
			3	45
	SA	N	3	36
			4	2
		Y	1	5
			3	3
			4	1

Table 1.4. Number of Items of Each Score Type and Assigned Scoring Case, Mathematics

Content Area	Constructed-Response Type	PT	Assigned Case	Number of Items Assigned
Mathematics	EQ	N	0	22
			2	309
		Y	2	224
	NL	N	1	2
			3	56
			4	2
		Y	1	53
			3	15
	SA	N	3	28
			4	2
		Y	1	133
			3	36
			4	1

Selection of responses.

The number of on-grade responses for each item varied from less than 500 to more than 12,000. To select the scoring samples reflected in Table 1.2, random samples were drawn from the response data for the item stratified on IEP status, LEP status, and ethnicity, Note that the proportions for each of these strata were based on the proportions in the pilot sample for each item; adjustments were not made to the proportions as the test delivery engine randomized assignment of test form and CTB wanted to ensure the randomization was not disturbed by oversampling in any field.

Selection of items for each engine.

Finally, CTB determined which Automated Scoring engines will receive each item. Total counts are represented in Table 1.5.

- Case 1 items will not go to Automated Scoring engines but will be routed to hand-scoring systems only.
- Case 2 items will go to the AIR-EQS engine.

- Case 3 items
  - All NL items will be routed to the AIR-PROP and MI engines.
    - In ELA/literacy, a subset of 5 of these will also be sent to the ASAP 1, ASAP 2, ASAP 3, ASAP 4, and ASAP 5 engines.
    - In Mathematics, a subset of 3 of these will also be sent to ASAP 1, ASAP 2, ASAP 3, ASAP 4, and ASAP 5 engines.
  - All SA items will be routed to the MI engine.
    - In ELA/literacy, a subset of 5 of these will also be sent to the ASAP 1, ASAP 2, ASAP 3, ASAP 4, and ASAP 5 engines. A further subset of 2 will also be sent to LightSIDE.
    - In Mathematics, a subset of 2 of these will also be sent to the ASAP 1, ASAP 2, ASAP 3, ASAP 4, and ASAP 5 engines. A further subset of 1 will also be sent to LightSIDE.
  - All Essay items will be routed to the AIR-OSE, CTB-Bookette, LightSIDE, and MI engines.
- Case 4 items
  - All NL items will be routed to the AIR-PROP, LightSIDE, MI, ASAP 1, ASAP 2, ASAP 3, ASAP 4, and ASAP 5 engines.
  - All SA items will be routed to the LightSIDE, MI, ASAP 1, ASAP 2, ASAP 3, ASAP 4, and ASAP 5 engines.
  - All Essay items will be routed to the AIR-OSE, CTB-Bookette, Lightside, MI, ASAP 1, ASAP 2, ASAP 3, ASAP 4, and ASAP 5 engines.

Table 1.5. Scoring Engine Clients and Item Counts by Score Type

Content	Score Type	Case	AIR-OSE	AIR-PROP	AIR-GRS	AIR-EQS	CTB	LIGHT SIDE	MI	ASAP 1	ASAP 2	ASAP 3	ASAP 4	ASAP 5
ELA/literacy	Essay	3	12				12	12	12					
		4	5				5	5	5	5	5	5	5	5
	NL	3		417					417	5	5	5	5	5
		4		2				2	2	2	2	2	2	2
	SA	3						2	39	5	5	5	5	5
		4						3	3	3	3	3	3	3
Mathematics	EQ	2				533								
	NL	3		71					71	3	3	3	3	3
		4		2				2	2	2	2	2	2	2
	SA	3						1	64	2	2	2	2	2
		4						3	3	3	3	3	3	3

Chapter 2: Criteria for Automated Scoring Acceptance

The content of an assessment, the conditions of measurement, and the examinee population are the three broad characteristics of an assessment that define the construct represented by a test score (Kolen, 2011). When Automated Scoring is used to score an assessment, in addition to the evaluation of statistical thresholds, these three characteristics must also be documented and inspected as important pieces of validity evidence for the assessment. This is because Automated Scoring is one facet of the conditions of measurement while simultaneously being one facet of test content. Though the items remain the same regardless of whether Automated Scoring or human scoring is used, in the Automated Scoring method, the content assessed depends on the features and statistical techniques a scoring engine uses to model human scores as well as the accuracy of the modeling process (Schneider, Waters, & Wright, 2012). Moreover, the examinee population upon which the engines are trained and evaluated should be shown to represent the population of test takers (Higgins, 2013a).

In Automated Essay Scoring (AES), the human-engine relationship is a central benchmark for evaluating the AES model. As human-human agreement served as benchmarks for reliability of AES scoring, measures used to assess human-human agreement were adopted to evaluate engine-human agreement. The human-engine relationship is intended to evaluate the reliability of scores derived from the AES engine, similar to the way the human-human relationship is interpreted as reliability evidence for human scoring (Attali & Burstein, 2006). The same is true when Automated Scoring is used to score other types of constructed-response items. A second benchmark examines how the engine-human relationship compares to the human-human relationship. Should the engine-human relationship be similar to (i.e., no more than 0.10 below the human weighted kappa value) or exceed that of two humans during the model building validation studies, then the engine is generally accepted for operational work (see Williamson, Xi, and Breyer, 2012, for a discussion). Note that both of these benchmarks focus on scores rather than the processes underlying the production of scores by a human or an engine.

Weighted kappa has recently been the primary focus for Automated Scoring evaluation. Scoring has traditionally been considered acceptable if the human-engine weighted kappa for a prompt is above 0.70 (Condon, 2013; Higgins, 2013a). In the recent Automated Student Assessment Prize public competitions, weighted kappa was used as the sole criterion to rank order public competitor success in modeling engines. It is possible, however, for engines to have high weighted kappas yet not model the human score distributions in terms of means and standard deviations. Therefore, 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 engine quality 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, 2013a). 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. In 2011 CTB adopted the ETS framework (See Table 2.1), with some minor adjustments.

The ETS framework is described in depth by Williamson, Xi, and Breyer (2012), Ramineni and Williamson (2013), and Higgins (2013a); thus, it will not be discussed in this proposal. Interested readers should refer to these publications and manuscripts. Rather, we focus on the minor

adjustments that CTB has made to the ETS framework for our operational Automated Scoring system.

Developers and users of the ETS framework have noted that measuring consistency with human raters in terms of the percent exact agreement is problematic. Exact agreement percentages are related to the number of score points in the rubric and the distribution of responses along the scales. Bridgeman (2013) noted that high agreement between two raters can occur when raters are truncating the rubric score range. When raters are making use of four or more points, he noted a review of this index makes sense. These observations are all correct. CTB has found an engine's weighted kappa 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 (of >0.05 difference) in perfect agreement rates between the human-human and human-engine 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. The purpose for this check is to ensure that the distribution of automatically derived scores is centered with human scoring in order to avoid problems with differential scaling. 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 to flagging the population SMD if it exceeds 0.12 in absolute value. CTB will also provide the direction of this index for the population and subpopulation. All other indices are used as originally specified in the ETS framework. CTB proposes to use the modified ETS framework found in Table 2.1, given that the Smarter Balanced Assessment Consortium goal is to use the automated score as the student's score of record for computer adaptive testing (CAT) purposes if feasible, and because this framework closely aligns to the recommendations made by Higgins (2013a).

Data Sources for Engine Evaluation

CTB will evaluate the results of Automated Scoring for Case 3 and Case 4 items which receive full engine training and validation processes and Case 3 items that are validation only items according to the thresholds set forth in Table 2.1. In addition, CTB will include evaluations of the demographic representativeness of the training and validation samples for each item to the census population of record in the data files. CTB will also evaluate how aligned the features available for use in Automated Scoring engines are to the construct, in keeping with recommendations of Higgins (2013a). In addition CTB will evaluate highly discrepant human-engine papers for subset of items that warrant such an investigation to better understand research findings. Outside of the scope of our study, given the data sources available and our proposal response, are the relationships of automated scores to external measures and to indices based student's reported test scores.

Chapter 2: Criteria for Automated Scoring Acceptance

Table 2.1: Statistical Criteria for the Evaluation of Automated Scoring Engines

Flagging Criterion	Flagging Threshold
Weighted Kappa for engine score and human score	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 weighted Kappa or correlation from human-human to engine-human	Decline in 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

Chapter 3: The Relationship between Sample Size and Engine Reliability

The score a human rater gives a student for a constructed-response item is an approximation of the student's true score on that particular prompt plus rater error plus random error. By removing sources of rater error from the engine training process, it is possible that the accuracy of the engines and the validity of test scores may be increased. For this reason users and creators of AES systems use different methodologies for building training sets, typically using information about a response from more than one human rater (Foltz, Streeter, Lochbaum, & Landauer, 2013; Rich, Schneider, & D'Brot, 2013, Ramineni, Williamson & Weng, 2011). Dikli (2006) hypothesized that more accurate engines could be built by taking the average of hundreds of reads of the same response.

Although averaging hundreds of reads to estimate the true score of an essay is a reliable way to generate the information necessary to train engines and may improve engine accuracy, it is also unfeasible due to time and cost constraints. A more plausible and cost efficient approach to improving the accuracy of scores derived from automated engines may be to add to the number of papers in the training sets. Bejar, Williamson, and Mislevy (2006) noted that the number of performances being observed limits the precision of measurement because each performance "is itself an imperfect indictor of the examinee's proficiency even if it was scored without error (p.63)." Larger training sets may provide better representation regarding what a true score for each point on a rubric represents.

The ideal size of a training set has yet to be established. The Automated Scoring literature has typically shared the lower bound training set size. In CTB's own development efforts, we have found that some short constructed-response items can be successfully trained using 150 valid responses per score point and validated using 100 responses per score point (Leacock, Messineo, & Zhang, 2013). This response count is similar to what was reported for a NAEP study that used ETS' c-raterTM (Sandene, 2005). In the automated essay scoring literature, the reported size of training sets has been lower than what has been reported for short answer constructed-response items.

Researchers have reported that surprisingly few papers are necessary to adequately train AES engines. Elliott (2003) reports that typically 300 or more papers are necessary to build models, and that a minimum of 20 papers are needed at the tails of the scoring rubrics. He did note, however, that models have been built with as few as 50 papers. Similarly, Foltz, Streeter, Lochbaum, and Landauer (2013) report that around 500 papers are needed for high stakes purposes. Rich, Schneider, and D'Brot (2013) wrote that for automated essay scoring a minimum of 60 responses per score point (on a 6-point rubric) was necessary and reported 250 responses are typical for validation sets. With the exception of Rich et al., the reported training set sizes have not been linked to human-engine agreement statistics so it is difficult to ascertain the accuracy of engines trained with small training sets.

Recently, the William and Flora Hewlett Foundation hosted the Phase I and Phase II event of the Automated Student Assessment Prize (ASAP) in which educational vendors and private individuals developed automated essay scoring engines for eight prompts and short constructed-response item scoring engines for 10 prompts (Shermis & Hamner, 2013, Shermis, 2013). Essay training sets provided by ASAP to the study participants ranged from 918–1,805 cases, and short constructed-response item training sets ranged from 1,278–1,799. It is likely that the actual training sets vendors and private individuals used for their internal model building process were somewhat smaller because participants likely created validation sets for their own use. Using a validation set size of 250 responses, we can expect that the typical training set used for model building may have ranged from 668-1,555 responses for essays and 1,028–1,549 for short answer constructed-

responses items. For approximately 50% of the essays, the human-engine weighted kappas exceeded those of the human-human counterparts and for the other 50% of essays, the human-engine weighted kappas were slightly lower than those of two humans (Shermis & Hamner, 2013). For the short answer constructed-response items no human-engine statistics were comparable to the human-human statistics (Shermis, 2013). Given the generally large training sets available in these studies, it does suggest that the size of the training set may not be the only influence on the ability of the engines to score accurately. The engine quality is dependent upon the accuracy of the human-human statistics and perhaps, the attributes of the items themselves (Leacock, Messineo, & Zhang, 2013).

Purpose

The size of the sample for which Automated Scoring engines may be sufficiently trained for high-stakes assessments for both Automated Essay Scoring purposes and short answer constructed-response Automated Scoring purposes as well as the size of the validation set has not been the subject of study in the research literature, therefore, the purpose of this study is to investigate the following questions:

- What is the optimal sample size needed to training an Automated Scoring engine?
- What is the optimal sample size needed to validate an Automated Scoring engine?

Methodology

Data Source

Case 4 items, described earlier in this document, will be the data source for this study. Therefore five randomly selected essay items, five randomly selected ELA/literacy constructed-response items, and five randomly selected mathematics constructed-response items administered as part of the online pilot will be investigated. For each prompt, 2,500 responses will be scored by human readers: 1,500 responses will be scored from a training set and 1,000 responses will be scored for a validation set. Each response will receive two human reads. If the two human scores do not agree, the response will be routed to a senior human reader whose score will become the human score of record.

Procedures

After the initial model building and scoring phase of the pilot administration is complete, CTB and its partners will implement this special study. Using the score of record with condition codes dropped (a condition code is any special code assigned to an essay by human scorers to indicate that an essay cannot be scored according to the scoring rubric) from the sample, CTB and its partners will investigate the mean, standard deviations, and frequency distributions, and the percentage of total papers at each score point for all papers for the total score and at the trait level in the training set and validation set. CTB and its partners will need to randomly subset the essays into training sets (T) and validation sets (V) of varying size to implement the study.

Using the features originally selected for the production scoring, CTB and its partners will retrain the engines with stratified random samples of papers from the training set that range in size from 100 responses to 1,500 responses increasing in size by increments of 100. For each of the training sets, scoring engines will be evaluated using the ETS framework using a stratified random sample of

papers from the validation set that range in size from 100 responses to 1,000 responses increasing in size by increments of 100. Recommendations regarding the optimal training and validation set sizes that inform the field study will be reported to Smarter Balanced Assessment Consortium, as well as a report of the study findings.

Chapter 4: Engine Read and Read-Behind Scenarios

Automated Scoring is quickly increasing its role in educational assessments (Bejar, 2011). In 2009, ETS began using its Automated Essay Scoring system, E-rater, in place of a second human score on the Test of English as a Foreign Language (TOEFL, Trapani, Bridgeman, and Breyer, 2011). Both West Virginia and Utah now use an AES as the student score of record for their large scale summative assessments used for accountability purposes (CTB, 2010; Shermis & Hamner, 2013). The Partnership for Assessment of Readiness of College and Careers, a consortium of 24 states collaborating to develop a common assessment, plans to use a combination of human and Automated Scoring in the consortium's future assessment system. The Smarter Balanced Assessment Consortium has recently requested to include Automated Scoring in their piloting and field testing process.

One benefit of using Automated Scoring, in addition to cost savings (Attali & Burstein, 2006), is that engines can ensure that scores from one student to another do not drift based upon external influences, which may not always be the case when human scores are used. Human scores on a large scale assessment are not always interchangeable with one another due to rater effects such as fatigue effects, halo effects, range restriction, leniency, and severity (Saal, Downey, & Lahey, 1980; Zhang, 2013). These rater effects may also vary during the scoring window as a scoring project proceeds (Myford & Wolfe, 2009). Ramineni and Williamson (2013) reported that admissions or licensure tests often have two human raters score each performance task in an effort to reduce sources of rater effects. This allows scores that are discrepant to be resolved. The same solution is not always feasible in K–12 educational testing. Because states have required cost savings in recent years, many states have moved to a single-rater-to-response scoring model. The quality of a single-rater-to-response scoring model is typically evaluated through read-behinds and check reads. Read-behinds are meant to provide evidence that raters are scoring consistently (or Automated Scoring engines are scoring accurately), and check reads provide evidence raters are applying the scoring rubric to student responses accurately (McClellan, 2010).

Similarly, when Automated Scoring systems are used for a single-rater-to-response scoring model, the quality of the engines are typically evaluated through validation studies, read-behinds, and check reads (Schneider & Osleson, 2013). During the validation study process, the engine-human relationship is used to evaluate the quality of scores derived from the Automated Scoring engine. Read-behinds from humans are meant to provide ongoing evidence that engines are generalizing from a validation study into a production environment, and check reads provide evidence that human and engine scores are comparable to one another. Although this usage of Automated Scoring reduces the number of human reads, which lowers costs and allows student responses to be scored in real time or quickly thereafter, it does not allow sufficient time to "troubleshoot" should a lead psychometrician need to investigate issues related to engine generalizability, newly discovered gaming techniques, or human rater drift concerns. This may be why Zhang (2013) posited that Automated Scoring was not yet sufficiently advanced enough to support its use in the single-rater-to-response mode.

Zhang (2013) wrote that in order for Automated Scoring to be used as the score of record three criteria must be met: (a) features used in scoring should be transparent, (b) validity evidence should warrant the intended use of the scores, and (c) quality control measures (see Bejar, 2011) should be able to detect aberrant responses that may be attempts to "game" the system or that are outside of the boundary of papers used to train and validate a particular system. He also noted that Automated Scoring could be used in conjunction with human scoring in two main ways. First, the score of record

could be based on a linear combination of the human score and engine score. Second, the Automated Scoring could be the read-behind of the human score. Although both of these approaches reduce the number of human reads which lowers costs and saves time, these approaches are not likely viable if the goal is to score student responses in real time or quickly thereafter. It should be noted, however, whether engine scores are used alone or in conjunction with human scores, there will be some portion of responses identified as needing to be routed to humans in compliance with the quality control measures of any Automated Scoring system (Jones & Vickers, 2011).

Bennett (2011) proposed an interesting potential resolution to the dilemmas noted above. He suggested independently training and deploying two engines for each item and routing papers to a human read when the engines were discrepant. This approach would reduce the number of human reads which lowers costs and saves time, may score a significant number of student responses in real time or quickly thereafter; but it still requires some unknown percentage of responses to be routed to human raters.

Purpose

To our knowledge this idea has not yet been investigated in the research literature. Therefore, the purpose of this study is to contrast the scoring accuracy for the score of record for each of the following scenarios to that derived from two humans. The following cases will be investigated:

- Scores derived from two engines in which discrepant scores are routed to an expert rater.
- Scores derived from one engine.
- Scores derived from one engine and one human in which the engine score contributes to the student score of record by averaging the two scores and rounding to the nearest integer when scores are adjacent and in which nonadjacent scores are routed to an expert rater.
- Scores derived from one engine and one human in which the human is the score of record and in which nonadjacent scores are routed to an expert rater ("engine read-behind").
- Scores derived from two humans in which rater 1 is the score of record and in which nonadjacent scores are routed to an expert rater.

Methodology

Data Source

Case 4 items, described earlier in this document, will be the data source for this study. Therefore five randomly selected essay items, five randomly selected ELA/literacy constructed-response items (2 NL and 3 SA), and five randomly selected mathematics constructed-response items (2 NL and 3 SA) administered as part of the online pilot will be investigated.

Procedures

For each of the Case 4 items, CTB will have 1,000 responses in the validation set along with automated scores from each of the vendors who score the particular item type. Each Case 4 essay

will have scores from nine independently developed scoring engines. Each Case 4 ELA/literacy NL item will have scores from eight independently developed scoring engines, and each Case 4 ELA/literacy SA will have scores from seven independently developed scoring engines. Similarly, in mathematics, each Case 4 NL item and SA item will have scores from seven independently developed scoring engines. For the purposes of this study, after evaluating the engine functioning using the ETS framework, CTB will select the two engines it evaluates as the top performing engines to use for the mock-production scoring with two engines. The highest performing engine will be used in the single engine score of record and engine-to-human comparisons.

CTB will use the ETS framework to compare the accuracy of each approach in comparison to what would be observed with two humans. In addition to the agreement statistics, CTB will provide contingency table comparisons showing the effect on student scores under each scoring method for each item so that the Smarter Balanced Assessment Consortium can determine the practical effect each approach would have on the student score of record. We will use the following human score resolution method as the basis of the study. The rater 1 score is the score of record when rater 1 score is adjacent to the rater 2 score. The rater 3 score is the score of record when rater 1 and rater 2 are discrepant. CTB will also analyze highly discrepant responses under each of these scenarios.

Chapter 5: Automatic Identification of Papers Likely to Require Human Scoring

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 scores from engines will deviate from human scores for some responses. Although metrics to evaluate engines describe overall scoring consistency compared to humans (see earlier section on Criteria for Constructed-Response Item Acceptance in Part 2 of this document) such indexes neither identify which responses are likely to receive scores discrepant with human scores nor include a statistical measure of risk associated with use of the engine.

In a single-engine-score-to response scenario it is important to know which responses are sufficiently aberrant such that they are at risk of being discrepant with what a human likely would have given. In such situations, flagged responses should be routed to humans for scoring. At the engine level, it is important to understand the percentage of item responses that would need to be routed for human scoring based upon the item level risk index as well as potential other engine level risk indicators.

Engines may provide a predicted score that is discrepant from a human score for the following two reasons:

1. The response features fall outside of the training set feature space; the response is novel or an outlier. The ability to identify a novel response or an outlier given the highly dimensional feature space of many of the Automated Scoring engines can be numerically challenging and even intractable; therefore, careful selection of an index used to quantify the risk can reduce costs related to routing responses likely to be scored well by engines to hand scoring (reduce false positives) while simultaneously providing a means to identify papers that should be sent to hand scoring (reduce false negatives). The need to identify novel responses or outliers in relation to a feature space used for an automated classification engine is not unique to the field of educational assessment (see, for example, Furusjo, Svenson, Rahmberg, & Andersson, 2004); yet, due to the potential high-stakes use of scores from Automated Scoring engines, is of great importance to the field.
2. The engine model does not fit the training data sufficiently, or was tuned insufficiently in terms of the tradeoff between bias and complexity (overfitting). Some of the machine learning methods deployed for the Smarter Balanced Assessment Consortium can provide response-level residuals, a probability that a predicted score will match the expert assigned score, or other measures.

At the item level, a risk index may help to expand and strengthen evaluation criteria for Automated Scoring engines, informing decision making regarding which of the multiple engines to deploy, whether to include an item in CAT, whether to recommend hand-scoring for an item, or to recommend use cases for which the Automated Scoring of the item is valid (e.g., formative use versus high-stakes use). Multiple techniques may be of value when calculating a risk index, for example, calculating the likelihood of an incorrect categorization or calculating the prediction error (defined by loss and complexity functions) when machine learning techniques are deployed in an engine. To date, CTB is not aware of published research comparing the evaluation of Automated Scoring engines using measures such as these to the evaluation of Automated Scoring engines based on comparisons of human-engine and human-human agreement statistics. In addition, the typical measures of engine performance for engine-human and human-human agreement rely on

rule of thumb thresholds rather than statistical tests for difference in performance (Breyer et al., 2012).

Purpose

The purpose of this study is to investigate possible methods for automatic identification of responses for which automated scores are likely to be discrepant with human scores. This study will investigate techniques to (a) identify the risk that a response to a given item will be or has been discrepantly scored and (b) quantify the risk at the engine level so that the likelihood of a discrepant score being assigned by an engine may be understood.

Methodology

Data Source

Case 4 items, described earlier in this document, will be the data source for this study. Therefore five randomly selected essay items, five randomly selected ELA/literacy constructed-response items (2 NL and 3 SA), and five randomly selected mathematics constructed-response items (2 NL and 3 SA) administered as part of the online pilot will be investigated.

Procedures

The study will include a review of literature related to the following:

- Methods for the calculation of training error and prediction error for the various machine learning techniques used by the Automated Scoring engines when scoring the constructed-response items (see, for example, Hastie, Tibshirani, & Friedman, 2009).
- Outlier identification and novelty identification as it relates to proximity of validation set responses to feature spaces (Hodge & Auston, 2004; Cerioli & Farcomeni, 2001; Chandola, Banjeree, & Kumar, 2009).

Following a review of the literature, a subset of engines and related feature sets will be used as empirical examples for the exploration and demonstration of methods identified as promising during the literature review. Responses flagged using response-level risk indexes will be compared against responses from the validation set receiving discrepant and adjacent scores to determine the potential for added value of using these indexes to flag responses for human scoring. Note that while collateral information may provide an additional avenue for identification of responses likely to be discrepantly scored by automated engines, methods requiring use of collateral information will not be included in this study due to limited collateral information available in this pilot administration. Engine-level risk indexes based on prediction error statistics will be compared against the human-human and human-engine agreement statistics to determine the potential for added value of analysis of engines using these indexes.

Chapter 6: Towards Predicting Whether a Short Constructed-Response Item Can be Scored Using Automated Scoring Engines

Streeter, Bernstein, Foltz, and DeLand (2011) reported that about half to two-thirds of the short-answer science items they investigated could be scored automatically with accuracies similar to humans. That is, one-third to one-half of the items could not be scored using Automated Scoring. This drop-out rate is consistent with CTB's own experience. In order to make Automated Scoring of short answer constructed-response items a viable option, the drop-out rate needs to be reduced significantly. The goal of this study is to improve the Automated Scoring rate of short answer constructed-response items by identifying the characteristics of items that are scored successfully using this technology.

Representatives from ETS, Pearson, and The College Board (Williamson et al., 2010) summarized the desired and challenging characteristics of short answer constructed-response items and their rubrics should Automated Scoring be the goal:

> One challenge associated with such systems is to develop items with definitive correct answers that the Automated Scoring system can verify. If the items call for opinions or other unverifiable discussion, the expected response set becomes less certain and more difficult for the Automated Scoring system to handle. Thus, a variety of factors influence the success of these systems for scoring, including the number of potential concepts that could be generated in a response, the variety of ways in which these concepts might be expressed, and/or the degree to which there is a clear distinction between correct and incorrect *representations of the concept*, among others (p. 3).

Interestingly, Leacock, Messineo, and Zhang (2013) found that many of these same characteristics—definitive answers, a limited number of concepts, etc.—also contribute to human rater reliability. One way to ensure an item has definitive correct answers is to ask that item developers provide analytic rubrics. Leacock, Gonzalez and Conarroe (in process) looked at the effects of using analytic versus holistic rubrics on eight ELA/literacy short answer constructed-response items. In the first round of scoring, the rubrics contained much holistic language. For example, the difference between full- and partial-credit was the difference between "sufficient evidence" and "limited evidence." They revised the holistic components of the rubrics into analytical language that explicitly stated the concepts and the requirements for each score point—including the score of zero. In addition, a sample answer for each score point was created. Based on the revised rubrics, new anchor papers were pulled and used to retrain the raters. With the analytic rubrics, inter-rater reliability, as measured by weighted kappa, rose from an average of 0.66 to an average of 0.93.

Currently, there are two options for determining whether short answer constructed-response items are scorable using Automated Scoring. The first option is to have Automated Scoring experts examine the individual questions and their rubrics and make a guesstimate of each one's scorability based on factors such as those suggested by Williamson et al., (2010). The drawback of this approach is that Automated Scoring potential may be lost based upon a guesstimate. In our experience, some items that experts have declared unscorable via Automated Scoring get very good scoring results. The second option is to apply Automated Scoring methods to all of the items and see which will work based on agreement with a held-out validation set. The drawback of this approach is the expense of generating highly validated human scores for a substantial number of responses to

each item coupled with the expense of building engines that may be unusable. A potential third option is to understand empirically which features affect whether or not an item may be successfully scored using automated techniques, and to train item writers to use these attributes during the item development process.

Purpose

The goal of this research study is to determine whether we can learn to predict, in advance of scoring, whether short answer constructed-response items can be scored automatically based on surface features found in (a) the metadata, (b) the item stem, and (c) the scoring rubric. This research extends the recent work of Leacock et al., (2013). They found that human-engine score differences were significant when the number of possible supporting text-based details from a reading passage was greater than five. Because only 41 of 76 ELA/literacy items in this study had human agreement rates that were acceptable (as measured by a weighted kappa of 0.70 or higher), Leacock and colleagues investigated rater agreement based upon the same surface features as were investigated for the Automated Scoring engines. They found rater agreement declined significantly when (a) there were more than five possible text-based supporting details, (b) there was more than one possible correct answer, and (c) the correct answer required the student to make an inference. Thus, it may be possible to improve the probability of an item being successful with Automated Scoring by improving the quality of the human scores if we can identify those features that correlate with human rater reliability.

CTB will extend the work of Leacock et al (2013) in a number of ways. We will do the following:

1. Add mathematics to the subject area coverage.
2. Increase the sample size of items investigated from 41 to several hundred.
3. Increase the metadata features investigated, listed below. For example, in the previous study, all but a few of the items had a DOK of 2. In this study, there is enough range in DOK to include it as a variable. In the previous study, the only metadata features used were grade and number of score points.
4. Introduce new and modify previous hand-coded features such as those in Section C below.
5. Increase the number of coders: each feature will be coded by two people.
6. Evaluate on both MI's PEG (for all items) and AIR's propositional engine (for a subset of the items)
7. Expand the evaluation criteria from weighted kappa of 0.70 to the ETS framework.

The features that will be included in this study are listed below.

I. Many of the features can be taken directly from the metadata: These may include, but are not limited to:
    a. Depth of Knowledge (DOK)
    b. Primary Claim
    c. Primary Assessment Target

d. Stimulus Type
e. Predicted Item difficulty
f. Achievement level descriptor classification
g. Grade
h. Number of score points
i. Stimulus type
j. Whether the item is part of a Performance task or not.

II. Features that can be semi-automatically extracted from the rubric include, but are not limited to:

Counts of the words and numbers in the exemplar (averaged if more than one exemplar).

III. Features that require human judgment and each value needs to be hand-coded by two persons. These features will include, but will not be limited to:

a. English language arts/literacy and Mathematics:
    i. Analytic or Holistic rubric
    ii. The analytic rubric has holistic elements
    iii. The analytic rubric clearly states the requirements for each score point
    iv. Scoring rule complexity
    v. Is there a definitive correct answer that can be expressed in a limited number of ways? (We need to establish whether coders agree on this possible variable)
    vi. Is there a choice of correct concepts?
    vii. Number of synonyms/paraphrases item writers included for the proposition scoring engine

b. English language arts/literacy only:
    i. Question type: for example, Explain with Supporting Detail, Define a Word or Phrase, Author intent, etc
    ii. Number of possible text-based supporting examples (when delimited in the rubric)—the coder may need to inspect the reading passage as well.
    iii. Answer found in text versus inference
    iv. Whether there is more than one possible correct answer. For example, the student chooses an argument and defends it.

c. Mathematics Only:
    i. Question type: for example, Use and Apply, Analyze/Categorize/Hypothesize, Answer and Explain (Schneider, Huff, Egan, Gaines, and Ferrara, 2013).

Chapter 6: Towards Predicting Whether a Short Constructed-Response Item can be Scored Using Automated Scoring Engines

Methodology

Scoring Engines

We will use eight scoring engines in this research: Measurement Incorporated's (MI) Project Essay Grade (PEG), AIR's propositional scoring engine, LightSIDE and the top five engines that were developed for phase 2 of the ASAP competition (Shermis, 2013). All of the short answer constructed-response items will be scored by MI. A subset of them, for which item writers supplied paraphrases and synonyms in their example answers, will be scored by AIR's proposition engine. Twenty-five will be scored by each of the five ASAP winners and 13 will be scored using LightSIDE (which was developed for essays, not short answer constructed-response items).

Data Source

The data set will comprise manually scored Pilot validation sets for a subset of randomly selected items from 479 ELA/literacy and 140 mathematics Case 3 and Case 4 items. To determine whether the features can predict accurate Automated Scoring of items, we will conduct a multiple regression analysis using features as independent variables and the acceptance or rejection of an item for Automated Scoring as the dependent variable

It has long been established that Automated Scoring accuracy for short answer constructed-response items is improved by automated spelling error correction (Leacock & Chodorow, 2003). Six of the eight engines in this study use spelling correction—but all use different algorithms. We propose to compute the accuracy of the spell correction algorithms for all vendors who agree to send CTB their spell corrected output for two or three items.

Procedures

Once the set of features whose values need to be hand-coded is finalized, CTB researchers will develop a protocol for training the coders. The protocol will include training coders on the subcategories of each feature through the use of training sets and along with independent scoring and discussion of a qualifying set for each feature. Hand coding will be done by CTB content experts in ELA and mathematics.

CTB will hold a three-day workshop to train and qualify coders. The first day will consist of training the coders using the training sets, independent scoring of the qualification set, as well as group discussion centered in clarifying questions. Each feature for each item will first be independently coded by two coders. In cases where the coders disagree, they will work together to resolve their differences. A random selection of items will be periodically re-coded to ensure consistency of protocol over time.

To determine whether the features can predict accurate machine scoring of items, we will conduct a multiple regression analysis using the features as independent variables and the evaluation metric as the dependent variable.

IV. Spell Correction Evaluation. Most of the eight automated scoring engines try to fix spelling errors when they encounter a non-word (one that is not recognized by their dictionary) in a student response. We will evaluate the methods used by the different

engines. First, we need to develop a spell correction evaluation protocol for all non-words in the responses. For example, if the intended word for lefs is the plural noun leaves and the spell corrector changes it to the singular noun, leaf, then it would be considered an appropriate correction. However, correcting it to the verb left would be an error. Thus coders must be trained using the above protocol for this phase of the study as well.

All spell checkers use edit distance (the number of keystrokes it takes to transform a non-word to a recognized word) to generate a list of suggestions. More sophisticated spell checkers prune and reorder the suggestions based on (1) the context of the non-word—those words that surround it, and (2) the likely pronunciation of the non-word.

Code accuracy on 1,000 spelling errors. The same set of the items will be given to both coders and their agreement will be monitored. In cases where the raters disagree, they will work together to resolve their differences. To evaluate the accuracy of each engine, we will compute percentage of spelling errors that were accurately corrected for each engine.

Chapter 7: Detecting Gaming in Automated Scoring Systems

Lochbaum, Rosenstein, Foltz, and Derr (2013) defined gaming in Automated Scoring as a student's "deliberate injection of construct-irrelevant features" into a response in order to create an observed increase in performance. Gaming is one risk of using Automated Scoring, and the ability of an engine or system to detect gaming attempts should be one factor in the evaluation of Automated Scoring system (Higgins, 2013b). Higgins found that depending upon the architecture of the Automated Scoring system, some systems may be more susceptible to certain types of gaming attempts than others.

The detection of certain types of gaming varies in difficulty. Overt techniques such as key banging (where students randomly strike the keyboard creating nonsensical letter patterns) or repeating the same words consecutively are relatively easy to catch through rule based system flags. Other techniques such as incorporating words found in the prompt into a response or substituting words in a response with less frequent, longer synonyms are harder to detect. This is because some gaming techniques are also construct-relevant ways students demonstrate (and are taught to demonstrate) their achievement (Higgins, 2013b). It is not clear whether studies that investigate substituting words in a response with longer synonyms would, in fact, demonstrate the effect of gaming an Automated Scoring system or validate that the system is rewarding construct-relevant achievement.

Increasing response length is a well-known gaming methodology (Powers, Burstein, Chodorow, Fowles, & Kukich, 2001; Higgins, 2013b; Lochenbaum et al., 2013) that has been shown to inflate student scores on both essays and short answer constructed-response items. Students typically increase the length of their response in a variety of ways such as repeating characters, words, phrases, paragraphs, or by padding their responses. Lochenbaum and colleagues defined padding as adding to the length of a response without adding construct-relevant content. Rambling can generally be one category of "padding" that humans may not recognize as a gaming attempt. Humans will, however, typically give a rambling response a lower score. An Automated Scoring system without a padding flag may also not recognize the response as a gaming attempt but will give the response a higher score. Finally, it will not always be clear (although sometimes it is) whether such response is true example of the student's skill or an attempt to game. A studied approach of padding (see Higgins, 2013b) has been to repeat portions of the response, append randomly chosen academic words or groups of larger content words to the end of essays, and combinations of the two. These techniques were found to increase scores in some of the Automated Scoring systems used in the Smarter Balanced Assessment Consortium pilot test.

Another common gaming approach is student use of shell language. Shell language has been defined in the literature in different ways, and its use is not unique to Automated Scoring applications. Bejar, VanWinkle, Madnani, Lewis & Steier (2012) defined shell language specifically within the context of persuasive and argumentative writing (or speaking) as formulaic language used to structure an argument without being specific to the position being described. Therefore, though the arguments may differ, students may use the same general *memorized* language and transition structure to organize their writing. This technique has been found operationally in assessments with high stakes for students interested in pursuing their education at the university level (Ramineni, Williamson, & Weng, 2011; Trapani, Bridgeman & Breyer, 2011)

Powers el al. (2001) found using a coachable shell to be a highly successful strategy to game an Automated Scoring system. This coachable shell technique is also similar in some respects to plagiarism (discussed in a following section of this document). Under the coachable shell

methodology, the student pads their response by repeating paragraphs, but rewording the first sentence of each paragraph slightly, reordering the subsequent sentences, and using synonyms. Thus, the essay appears to be a response to the topic, but it actually repeated paraphrases of a single paragraph, and as a result, it may be more difficult to detect by an Automated Scoring system.

Purpose

Researchers (see Higgins, 2013b) have not widely published the susceptibility of Automated Scoring systems to gaming strategies or the proportion of gamed responses flagged by the Automated Scoring systems. It is highly likely that researchers have been engaging in such endeavors, but they have not released the work publically to prevent circumvention of gaming detection methods. When evaluating an Automated Scoring system, users must understand how (a) accurately an Automated Scoring system flags gaming attempts and routes them to human raters and (b) susceptible an Automated Scoring system is to a particular gaming strategy. A case can be made that should a system correctly flag gaming attempts, it is not susceptible to such a strategy. Therefore, this study seeks to answer the following questions:

1. What proportion of student responses on the pilot administration are coded by human raters as gaming attempts?
2. What proportion of human identified gaming attempts are flagged by each engine?
3. How susceptible are engines to various gaming techniques?

Methodology

Data Source

Validation sets from Case 4 items that were described earlier in this document will be the data source for this study. Therefore 5 randomly selected essay items, 5 randomly selected ELA/literacy constructed-response items (2 NL and 3 SA), and 5 randomly selected mathematics constructed items (2 NL and 3 SA) administered as part of the online pilot will be investigated.

Procedures

In this study, CTB will extend the work of Higgins (2013b) to investigate both short answer constructed-response items in ELA/literacy and mathematics as well as essays. In addition to simulating increased length through the use of repeating a response, randomly adding content words, and randomly adding academic words CTB will add the condition of repeating the response through a paraphrase. In addition, each student response in the validation set will be coded for gaming techniques such that we have an estimate of the attempts to game in the student population and a quantification of the amount of true gaming each automated essay scoring system captures.

Chapter 7: Detecting Gaming in Automated Scoring Systems

Detecting Plagiarism in Student Essays Using Latent Semantic Analysis

Plagiarism, including direct copying and extensive paraphrasing, is increasingly a problem in secondary and post-secondary education. For instance, 58% of high school students admitted to plagiarism in a survey study by McCabe (Meyer, 2010). The increase in plagiarism has been linked to the rise of the internet (Bennett, 2005; Lathrop & Foss, 2000).

Plagiarism potentially increases a student's score on an essay and is therefore a direct threat to the validity of that score. Score validity, the extent to which test scores support subsequence inferences about student ability, is a central concern in educational assessment (AERA, APA, & NCME, 1999). Although the issue of plagiarism may not be specific to automated essay scoring (compared with other forms of gaming, e.g., Powers, Burstein, Chodrow, Fowles & Kukich, 2001), it nevertheless needs serious attention.

Currently, the scoring engines working to support the Smarter Balanced Assessment Consortium do not have extensive features to detect plagiarism. For longer essays, it may be possible to use some of the technologies used in essay scoring to identify papers that are particularly similar to one another. Unusually similar essays, especially when written during the same testing sessions, should be flagged for review.

Given the potential impact of false positives (that is, mistakenly identifying essays as instances of plagiarism), any automated detection can be considered only as a signal of possible plagiarism. Flagged texts should be routed as suspicious to humans for review and policymakers for subsequent investigation and, if necessary, collection of additional pieces of evidence to support claims that a student has plagiarized. Besides the legal aspects (Foster, 2002) an additional factor in the evaluation of potential plagiarism is the fact that students' perceptions of plagiarism (Marshall & Garry, 2005) may not be identical to the perceptions of other stake holders, e.g., academics (Nadelson, 2007). Moreover, careful attention is needed to ensure that a detection algorithm does not differentially flag essays written by students from certain subpopulations (e.g., English learners).

Latent Semantic Analysis

Latent semantic analysis (LSA; Deerwester, Dumais, Furnas, Landauer, & Harshman, 1990) is a mathematical method to compute the semantic similarity of words and texts. Informally, the LSA process consists of the following steps:

1. A matrix is constructed based on the occurrences of terms in documents
2. The dimension of this matrix is reduced to identify semantically similar terms and documents, producing a latent semantic space
3. Documents are represented by vectors in this latent semantic space; the distance between two document vectors provides a measure of semantic similarity

This technique has been used successfully in automated essay scoring (Foltz, Laham, & Landauer, 1999) as well as automated formative feedback on proper of quotations and citations (Britt, Wiemer-Hastings, Larson, & Perfetti, 2004). Moreover, LSA has been proposed as a technique to detect plagiarism (Cosma & Joy, 2012; Češka, 2008), where it may hold promise in detecting rearrangement and paraphrasing, forms of plagiarism that are difficult to detect even by state-of-the-art algorithms (Kakonnen & Mozgovoy, 2010). LSA may have two key advantages over other plagiarism detection techniques (e.g., string matching):

- LSA ignores word order making it less susceptible to paraphrasing
- LSA compares documents at a deeper ("latent") semantic level, making it possible to detect plagiarism of ideas rather than words

Purpose

CTB will conduct an exploratory study to develop and pilot an automated plagiarism detection algorithm based on latent semantic analysis, and provide Smarter Balanced Assessment Consortium with a research report including findings and recommendations on statistical computations that are likely to be successful.

Methodology

After a literature review, a proof-of-concept implementation of a detection algorithm will be developed. The performance of LSA in detecting source-code plagiarism is known depends on different parametric settings (Cosma & Joy, 2012). It is likely that optimal performance in case of student essays likewise requires appropriate parameters. Therefore, several algorithmic features will be evaluated, including the following:

1. Appropriate granularity of documents (essay, paragraph, sentence)
2. Required preprocessing (e.g., stemming, stopword filtering)
3. Effective weighting of terms (e.g., normalized, IDF, or entropy weights)
4. Optimal dimensionality reduction technique (e.g., fixed number, percentage of cumulated values, share of values, fraction; see Kaiser; Kakkonen, Myller, Sutinen, & Timonen, 2008)
5. Best similarity measures (cosine; Pearson, Spearman, Kendall correlation)

A practical problem in the development of plagiarism detection algorithms is the evaluation of their effectiveness in a wide-range of realistic situations. Unfortunately, a large collection of correctly identified, on-topic instances of plagiarized papers is generally unavailable for operational essay prompts. Recently, various corpora have recently been constructed for automated plagiarism detection, featuring natural and artificial plagiarism (Clough & Stevenson, 2011; Potthast et al., 2012).

CTB will evaluate the relevance of existing corpora to plagiarism detection in student essays. Additionally, we will investigate the feasibility of artificially generating a set of plagiarized papers from a training set of essays. For example, plagiarized essays could be generated by replacing words with synonyms or by combining paragraphs of several essays into a new paper. Once a collection of papers has been obtained, the performance of any detection algorithm can be quantified using statistical measures based on current research (Barrón-Cedeño, Potthast, Rosso, & Stein, 2010).

The end result of this exploratory study is a research report, describing the performance of different algorithmic settings and recommending potentially optimal settings for performance.

Chapter 8: Detection of Administration Anomalies

There are often multiple indicators of a security breach or an irregularity in test administration practices. It is prudent to analyze data for evidence of a security breach or test administration irregularity as these incidences may indicate a need for an investigation and subsequent score invalidation. For the Smarter Balanced Assessment Consortium pilot administration, this may mean that particular items must be handled differently than they otherwise might have been during item analysis, calibration and equating exercises. However, given that the pilot administration does not produce scores for students, teachers, or test sites for evaluation purposes, the motivation to inflate scores is much lower than one might see in a high-stakes summative administration. Therefore, CTB expects that little evidence of security breach or administration anomaly, if any, will be found.

Common analyses to date used in summative testing for the purpose of identifying a possible security breach include detection of unusual changes in test scores over time and answer changing behavior. However, for the pilot administration, these techniques are less relevant for the following reasons:

- The pilot test was a one-time administration for a sample of students. There is no history of performance on Smarter Balanced Assessment Consortium items for students or administration sites by which to measure unusual changes in scores.
- The pilot test was administered online using linear-fixed forms. In the test delivery platform, answer changing is largely limited to within the student's online testing session, with the exception of a 20-minute window immediately following the test session. Incidences of teachers or test administrators changing answers are much less likely to occur than on a paper and pencil linear fixed form due to the testing format.
- As this was a pilot administration, items had not yet been exposed in any test administration. Therefore, exposure of the items is much less likely than in scenarios where items are given in multiple administrations, such as when anchor items are used for equating.

Purpose

Common methods of identifying possible security breaches are less likely to provide useful information for the Smarter Balanced Assessment Consortium pilot. Therefore, CTB will perform analysis better suited to context in which the items were administered. In some cases, these analyses are related but more highly constrained than those listed above. CTB seeks to detect whether the following occurred:

1. Anomalies in time of the day a test was taken. A test taken outside of typical school hours may indicate a test administrator completed a student test.
2. Situations where (a) a test is modified after the session ends, (b) modifications include many wrong to right responses, and (c) modifications occur frequently for a given test administrator.
3. Aberrant responses times, indicating students are completing the tests much more quickly than counterparts of the same ability, which may indicate some pre-knowledge. Aberrant responses times may also indicate students taking the test for the sole purpose of memorizing and distributing items. Again, given the stakes associated with the pilot

and the fact that this was the first administration of many of these items, CTB expects to find little evidence of these phenomena.

Methodology

Data Source

Data for Part 1 of this study will be exported from the AIR test delivery platform and delivered to CTB. All tests taken will be included in this study.

Data for Part 2 of this study will be exported from the AIR test delivery platform and delivered to CTB. Only tests on which the test was modified after the session ends will be included in this study.

Data for Part 3 of this study will include information on the amount of time each student spent on each item (response time), the student response to each item, and the item parameters for each item. CTB expects that item parameters will be made available by ETS following the psychometric analyses for the pilot administration. Should the item parameters not be forthcoming, CTB will use the response time and response data for a similar but more limited study.

Procedures

Part 1. CTB will prepare a report of all tests taken outside of typical school hours, defined as 8:30 a.m. to 3:30 p.m. local time.

Part 2. For the set of tests for which modification occurred after the session ended, CTB will calculate the frequency of wrong-to-right answer changes. The procedure is formulated for an analysis with school classrooms as the unit of analysis. In this formulation, classrooms may be intact instructional classrooms, homerooms, or any testing group that can be identified using the name or ID of the teacher responsible for leading the test administration for that group.

In the description of the procedure, $i=1,\ldots,I$ denotes the classes in the state whereas n_i and m_i denote the sample size and mean number of wrong-to-right (WTR) changes for class i, respectively. In addition, μ and σ denote the mean and the standard deviation of the distribution of the number of wrong-to-right answer changes of the population of individual students in the state.

The basic idea underlying the procedure is a statistical test of the null hypothesis (H_0) that the mean number of wrong-to-right changes for the school class constitutes a random sample from the administration distribution of wrong-to-right changes. The hypothesis is tested against the (right-sided) alternative (H_1) that the mean number is too high to be explained by random sampling. Classes for which H_0 has to be rejected are flagged for further scrutiny. According to the central limit theorem, the sampling distribution of m_i is asymptotically normal with mean

$$\text{Mean}(m_i)=\mu$$

and standard deviation

$$\text{SD}(m_i)=\sigma/\sqrt{n_i}\ .$$

The classroom flagging criterion for each classroom is adjusted for the number of test takers in a classroom. This adjustment ensures that the flagging criterion is equally stringent for classrooms with considerably different numbers of test takers. Considering the nested structure (students within schools) and the potential dependencies within schools, we may incorporate the variance component due to schools into the computation of the standard deviation.

In addition, minimizing the probability of false positive (Type I) errors in this statistical test is crucial in this analysis. Flagging classrooms for further scrutiny is typically perceived as suspicion that students or educators have cheated by erasing incorrect answers and replacing them with correct answers.

The statistical procedure is as follows:

> For each class i=1,...,I, calculate $\mu+4\sigma/\sqrt{n_i}$.
>
> Flag the classes for which m_i is larger than the result.

Statistically, the flagging criterion proposed is very conservative. The standard normal table shows that under random sampling the (asymptotic) probability of a sample mean more than four standard deviations above the population mean is less than 0.0001. However, rejection of H_0 only tells us that the observed mean number of wrong-to-right erasures is unlikely to be the result of random sampling. Specifically, it does not necessarily prove any form of cheating by teachers.

The following caveats are always applicable:

1. The normal distribution holds only for large classes; for smaller classes the result is approximate.
2. Rejection of H_0 does not necessarily imply cheating. Alternative explanations are possible.
3. The flagging criterion should thus only be taken as a stimulus to look for additional evidence and find out what really happened in the classroom.

Part 3. CTB will use a procedure described in van der Linden and Guo (2008) to identify aberrances in test administration related to response time. This procedure is intended for use in computer adaptive testing, but it also is relevant to the on-line linear fixed form pilot administration. The procedure may lay the groundwork for future Smarter Balanced Assessment Consortium Computer Adaptive Test anomaly detection.

As noted earlier CTB assumes that item and person parameter estimates for all items and test takers will be made available to CTB from ETS, and in addition, the administered test design is clearly identified in the data and/or is provided by ETS. CTB will then apply the response time model using the time per item data received from AIR. Should CTB not receive item and person parameters,

or sufficient information regarding test design, CTB will conduct a limited analysis to simply examine the data for instances where the time a student takes to complete a test form is significantly less than that of other students taking the same test form with the same number correct score. In either case summary statistics of response times will be calculated and examined.

Response time analysis is proposed for this study as, as pointed out in van der Linden and Guo (2008):

1. Response times are continuous rather than binary, allowing information on the size of aberrances;
2. A response time statistical check on possible aberrance is expected to maintain its power throughout the test even when the difficulties of items and persons are close, a condition which causes residual-based person-fit statistical checks to lose power;
3. The response time model proposed separates the time intensity of an item from the speed of the test taker. It would be very difficult if not impossible for a test taker with pre-knowledge or memorization intent to time responses to match the time intensity (item) parameter and the speed (person) parameter for all items administered.

The response time model to be used for this analysis is as follows.

$$f(t_i;\ \tau_j, \alpha_i, \beta_i) = \frac{\alpha_i}{t_{ij}\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left[\alpha_i\left(\ln t_{ij} - (\beta_i - \tau_j)\right)\right]^2\right\},$$

where τ_i is the speed at which test taker *j* takes the test, β_i is the time intensity of item *i*, and α_i is a discrimination parameter for item *i*. van der Linden (2006) describes estimation procedures for these *parameters*.

CTB will estimate the response time parameters, and then identify aberrant response time patterns. Each log response time will be standardized using a predicted mean and standard deviation given the response times on all other items by the same test taker. A response time to an item will be flagged as aberrant when its standardized residual is more than 1.96 or less than -1.96. Rates of flagging outside of the significance level of the test, along with patterns of aberrances found, will be examined and reported.

In future years, the Smarter Balanced Assessment Consortium may want to consider the use of ANOVA and CUSUM (Egberink, Meijer, Veldkamp, Schakel, & Smid 2010; Van Krimpen-Stoop, & Meijer, 2001) analyses on item response times to detect possibly compromised items from administration to administration.

The following caveats are always applicable:

1. Explanations for aberrant response time patterns exist other than cheating. For example, poor time management, testing interruption.
2. The flagging criterion should only be taken as a signal for policymakers to collect additional evidence about what occurred in the classroom.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Attali, Y. & Burstein, J. (2006). Automated essay scoring with e-rater® v.2. *Journal of Technology, Learning, and Assessment, 4*. Retrieved June 3, 2010, from http://www.jtla.org

Barrón-Cedeño, A., Potthast, M., Rosso, P., & Stein, B. (2010). *Corpus and evaluation measures for automatic plagiarism detection.* Paper presented at the seventh conference on International Language Resources and Evaluation, La Valletta, Malta.

Bejar, I. I. (2011). A validity based approach to quality control and assurance of automated scoring. *Assessment in Education: Principles, Policy, and Practice, 18,* 319–341.

Bejar, I. I., VanWinkle, W., Madnani, N., Lewis, W., & Steier, M. (2013). *Length of textual response as a construct-irreleveant response strategy: The case of shell language.* ETS Research Report RR-13-07. Princeton, NJ: Educational Testing Service.

Bejar, I. I., Williamson, D. M., & Mislevy, R. J. (2006). Human scoring. In D. M. Williamson, R. J. Mislevy, & I. I. Bejar (Eds.), *Automated scoring of complex tasks in computer-based testing* (pp. 49–82). Mahwah, NJ: Lawerence Erlbaum.

Bennett, R. (2005). Factors associated with student plagiarism in a post-1992 university. *Journal of Assessment and Evaluation in Higher Education, 30,* 138–162.

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

Breyer, F.J., Brew, C., Lewis, C., Williams, F., & Blackmore, J. (2012, April). *Using inferential confidence intervals to assess automated scoring agreement with human scoring agreement in short-text tasks*. Paper presented at the National Council of Measurement in Education, Vancouver, Canada.

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.

Britt, M. A., Wiemer-Hastings, P., Larson, A. A., & Perfetti, C. A. (2004). Using intelligent feedback to improve sourcing and integration in students' essays. *International Journal of Artificial Intelligence in Education, 14,* 359–374.

Cerioli, A. & Farcomeni, A. (2001). Error rates for multivariate outlier detection. *Computational Statistics and Data Analysis, 55,* 544–553.

References

Češka, Z. (2008). Plagiarism detection based on singular value decomposition. *Advances in Natural Language Processing, 5221,* 108–119.

Chandola, V. Banjeree, A. & Kumar, V. (2009). Anomaly detection: A survey. *ACM Computing Surveys*, *41,* Article 15.

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

Condon, W. (2013). Large-scale assessment, locally-developed measures, and automated scoring of essays: Fishing for red herrings? *Assessing Writing, 18,* 100–108.

Cosma, G., & Joy, M. (2012). Evaluating the performance of LSA for source-code plagiarism detection. *Informatica, 36,* 409–424.

CTB/McGraw-Hill. (2010). WVDE WESTEST 2 technical report. Monterey, CA: Author.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science, 41,* 391–407.

Dikli, S. (2006). An overview of automated scoring of essays. *Journal of Technology, Learning, and Assessment, 5*. Retrieved November 9, 2011, from http://www.jtla.org.

Egberink, I., Meijer, R., Veldkamp, B., Schakel, L., & Smid, N. (2010). Detection of aberrant item score patters in computerized adaptive testing: An empirical example using the CUSUM. *Personality and Individual Differences, 48*, 921–925.

Foltz, P. W., Laham, D., & Landauer, T. K. (1999). The Intelligent Essay Assessor: Applications to educational technology. *Interactive Multimedia Electronic Journal of Computer-Enhanced Learning, 1*. Retrieved June 2, 2013, from http://imej.wfu.edu/articles/1999/2/04/

Foltz, P. W., Rosenstein, M. & Lochbaum, K. E. (2013, April). *Improving performance of automated scoring through detection of outliers and understanding model instabilities*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, CA.

Foltz, P. W., Streeter, L. A., Lochbaum, K. E., & Landauer, T. (2013). Implementation and applications of the Intelligent Essay Assessor. In. M. Shermis & J. Burstein (Eds.), *Handbook of automated essay evaluation: Current applications and new directions* (pp. 68–88). New York: Taylor & Francis.

Foster, A. (2002, May 17). Plagiarism-detection tool creates legal quandary. *Chronicle of Higher Education,* A37.

Furusjo, E., Svenson, A., Rahmberg, M., & Andersson, M. (2006). The importance of outlier detection and training set selection for reliable environmental QSAR predictions. *Chemosphere, 63,* 99–108.

References

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

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

Higgins, D. (2013b, April). Managing what we can measure: Quantifying the susceptibility of automated scoring systems to gaming behavior. Paper presented at the annual meeting of the National Council on Measurement in Education Conference, San Francisco. CA.

Hodge, V., & Auston, J. (2004). A survey of outlier detection methods. Artificial Intelligence Review, 22, 85–126.

Jones, M. & Vickers, D. (2011). *Considerations for performance scoring when designing and developing the next generation assessments.* Retrieved March 29, 2013, from http://www.pearsonassessments.com/hai/images/tmrs/Performance_Scoring_for_Next_Gen_Assessments.pdf.

Kakkonen, T., & Mozgovoy, M. (2010). Hermetic and web plagiarism detection systems for student essays: An evaluation of the state-of-the-art. *Journal of Educational Computing Research, 42*, 135–139.

Kakkonen, T., Myller, N., Sutinen, E., & Timonen, J. (2008). Comparison of dimension reduction methods for automated essay grading. *Educational Technology & Society, 11,* 275–288.

Kolen, M. J. (2011, April). *Comparability issues associated with assessments for the Common Core State Standards*. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.

Lathrop, A., & Foss, K. (2000). *Student cheating and plagiarism in the Internet era: A wake-up call.* Englewood, CO: Libraries Unlimited.

Leacock C., & Chodorow, M. (2003). C-rater: Automated scoring of short answer questions. *Computers and the Humanities, 37,* 389–405.

Leacock, C., Gonzalez, E., & Conarroe, M. (in process). *Developing effective scoring rubrics for AI short answer scoring.* Manuscript in process.

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.

Lochbaum, K.E., Rosenstein, M., Foltz, P., & Derr, M. A. (2013, April). *Detection of gaming in automated scoring of essays with the IEA.* Paper presented at the annual meeting of the National Council on Measurement in Education Conference, San Francisco, CA.

McClellan, C.A. (2010, April). *Quality Assurance and control in human scoring*. Paper presented at the annual meeting of the National Council on Measurement in Education, Denver, CO.

References

Marshall, S., & Garry, M. (2005, Dember). *How well do students* really *understand plagiarism?* Paper presented at the annual meeting of the Australasian Society for Computers in Learning in Tertiary Education, Brisbane, Australia.Meyer, J. P. (2010, May 27). Students' cheating takes a high tech turn. *Denver Post.*

Mozgovoy, M., Kakkonen, T., & Cosma, G. (2010). Automatic student plagiarism detection: Future perspectives. *Journal of Educational Computing Research, 43,* 511–531.

Myford, C.M. & Wolfe, E.W. (2009). Monitoring rater performance over time: A framework for detecting differential accuracy and differential scale category use. *Journal of Educational Measurement. 46*, 371–389.

Nadelson, S. (2007). Academic misconduct by university students: Faculty perceptions and responses. *Plagiary, 2,* 1-10.

Potthast, M., Gollub, T., Hagen, M., Graßegger, J, Kiesel, J., Michel, M., Oberländer, A., Tippmann, M., Barrón-Cedeño, A., Gupta, P., Rosso, P., & Stein, B. (2012). Overview of the 4th international competition on plagiarism detection. In P. Forner, J. Karlgren, & C. Womser-Hacker (Eds.), *CLEF 2012 Evaluation Labs and Workshop—Working Notes Papers*.

Powers, D., Burstein, J. C., Chodrow, M., Fowles, M. E, & Kukich, K. (2001). *Stumping e-rater: Challenging the validity of automated essay scoring* (GRE Board Professional Rep. No. 98-08BP). Princeton, NJ: Educational Testing Service.

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

Ramineni, C., Williamson, D. M., & Weng, V. (2011, April). *Understanding mean score differences between e-rater® and humans for demographic-based groups in GRE®*. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.

Rich, C., Schneider, M. C., & D'Brot, J. (2013). Applications of automated essay evaluation in West Virginia. In. M. Shermis & J. Burstein (Eds*). Handbook of automated essay evaluation: Current applications and new directions* (pp. 99–123). New York: Taylor & Francis.

Saal, F. E., Downey, R. G., & Lahey, M. A. (1980). Rating the ratings: Assessing the psychometric quality of rating data. *Psychological Bulletin, 88*, 413–428.

Sandene, B., Horkay, N., Bennett, R. E., Allen, N., Braswell, J., Kaplan, B., & Oranje, A. (2005). *Online assessment in mathematics and writing: Reports from the NAEP technology-based assessment project, research and development series.* (NCES 2005-47). U.S. Department of Education, National Center for Education Statistics. Washington, DC: U.S. Government Printing Office.

Schneider, M. C., Waters, B., & Wright, W. (2012, April). *Stability of automated essay scoring engines across time and subgroups when student ability changes.* Paper presented at the annual meeting of the National Council on Measurement in Education, Vancouver, Canada.

References

Schneider, M.C., & Osleson, L. (2013, April). *Evaluating the comparability of engine and human scores over time.* Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, CA.

Shermis, M. D., & Hamner, B. (2013). Contrasting state-of-the-art automated scoring of essays. In M. D. Shermis & J. Burstein (Eds.), *Handbook of automated essay evaluation: Current applications and new directions* (pp. 313–346). New York: Taylor & Francis.

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

Streeter, L., Bernstein, J., Foltz, P., & DeLand, D. *(*2011). *Pearson's automated scoring of writing, speaking, and mathematics.* Retrieved on June 25, 2013, from http://www.pearsonassessments.com/hai/images/tmrs/PearsonsAutomatedScoringofWritingSpeakingandMathematics.pdf

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., Bennett, R. E, Lazer, S., Bernstein, J., Foltz, P. W., Landauer, T. K., Rubin, D. P., Way, W. D., & Sweeney, K. (2010). *Automated scoring for the assessment of Common Core Standards*. Retrieved on December 15, 2010, from http://professionals.collegeboard.com/profdownload/Automated-Scoring-for-the-Assessment-of-Common-Core-Standards.pdf

van der Linden, W. J. (2006). A lognormal response model for response times on test items. *Journal of Educational and Behavioral Statistics, 31,* 282–2004.

van der Linden, W. J., & Guo, F. (2008). Bayesian procedures for identifying aberrant response-time patterns in adaptive testing. *Psychometrika, 73,* 365–384.

van Krimpen-Stoop, E. M. L. A., & Meijer, R. R. (2001). CUSUM-based person fit statistics for adaptive testing. *Journal of Educational and Behavioral Statistics, 26*, 199–218.

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.