Improving Reliability Throughout the Automated Scoring Development Process

National Council on Measurement in Education
Vancouver, British Columbia

Peter Foltz, Pearson

Mark Rosenstein, Pearson

Karen Lochbaum, Pearson

Laurie Davis, Pearson

April, 2012

Increasing reliability throughout the automated scoring development process

Automated scoring of written constructed response items has grown rapidly for use in large-scale testing and for formative assessment. The greater availability of online testing platforms makes automated essay scoring (AES) systems increasingly practical to implement and feasible to incorporate into these platforms. These automated essay scoring systems often produce scores more reliably and quickly and at a lower cost than human scoring (see Hearst, 2000; Topol, Olson, & Roeber, 2011; Williamson et al., 2010). As these systems are implemented, it becomes increasingly important to develop methods to ensure that the AES is scoring effectively. Much like in the development and scoring of essays by human graders, steps must be taken to ensure that the AES is providing reliable and valid scores. Throughout the development and implementation of automatically scored writing prompts, there are a number of stages that impact the overall reliability of the scoring of student responses. These stages include:

1) The collection of essays used to train the scoring system.

2) The collection of scores from human raters for the training essays.

3) The creation and testing of algorithms that most reliably detect components of student writing quality and knowledge,

4) The use of methods that detect essays that may be scored less reliably by the automated scoring methods once the system is implemented.

In each of these stages, considerations must be made in order to maximize performance and generalizability to student essay data. For example, decisions must be made early on about how many essays must obtained to be human hand scored and what the desired score distribution should be in order to train the scoring system most effectively.

This paper describes approaches taken to improve the scoring reliability throughout the automated scoring development process. It focuses on four aspects of scoring reliability:

1) the size of the training set needed to provide effective automated scoring

2) the score distribution of the essays in the training set to provide effective automated scoring

3) the effect of human scoring reliability on automated scoring reliability

4) whether the automatic detection of essays that may not be scored as reliably is effective at improving reliability by flagging the essays for human scoring.

The work is part of an ongoing systematic study to determine best approaches to improve overall scoring reliability across item development and implementation. The paper further provides examples of how these approaches were implemented within the context of the Intelligent Essay Assessor (IEA).

Overview of Automated Scoring with the Intelligent Essay Assessor (IEA).

The Intelligent Essay Assessor (IEA) is based on a machine-learning approach in which it is trained to score essays based on the collective wisdom of trained human scorers (Foltz, Landauer & Laham 1999; Landauer, Laham & Foltz 2003). Training the IEA involves first collecting a representative sample of essays that have been scored by human raters. The IEA then extracts features from the essays that measure aspects of student performance such as the student's expression of knowledge and command of linguistic resources. Then, using machine-learning methods, the IEA examines the relationships between the scores provided by the human scorers and the extracted features in order to learn how the different features are weighted and combined to provide a score that models how humans score the essays. The resulting representation is referred to as a "scoring model". This initial stage is critical to the success of AES. Divergence between the training essays and the scoring population of essays or divergence between the human scoring of the training set from that used in production

scoring can impair the reliability of the AES results. Thus, care must be taken at this stage to ensure careful linking of the training set and the scoring process of the training set to the intended population of essays and overall scoring process.

Scoring Features used in IEA

The quality of a student's essay can be characterized by a range of features that measure the student's expression and organization of words and sentences, the student's knowledge of the content for the domain, the quality of the student's reasoning, and the student's skills in language use such as grammar and the mechanics of writing. In developing analyses of such features, the computational measures should extract aspects of student performance that are relevant to the constructs for the competencies of interest (e.g., Williamson et al., 2010). For example, a measure of the type and quality of words used by a student provides an effective and valid measure of a student's lexical sophistication. In contrast, a measure that counts the number of words in an essay, although it can be highly correlated with human scores for essays, does not provide a valid measure. Because a student's performance on an essay typically requires demonstrating combined skills across language expression and knowledge, it is further critical that the scoring features used in the analysis cover the constructs of writing that are being scored. Thus, multiple language features are typically measured and combined to provide a score. The IEA uses a combination of features that measure aspects of the content, lexical sophistication, grammar, mechanics, style, organization, and development within essays.

Evaluating Responses for Scorability

Before scoring any student essay, the IEA analyzes the essay to determine the confidence with which it can be accurately scored . The IEA uses a variety of statistical and probabilistic checks to make this determination based on characteristics of the essays contained in the training set and experience with a variety of both good- and bad-faith essays. Essays that

appear to be off-topic, not English, highly unusual or creative, or flagged for customer specific reasons are typically directed to a human for scoring.

Building a Scoring Model

The IEA is trained to associate the extracted features in each essay to scores that are assigned by human scorers. A regression-based approach is used to determine the optimal set of features and the weights for each of the features to best model the scores for each essay. From these comparisons, a prompt and trait-specific scoring model is derived to predict the scores that the same scorers would assign to any new responses. Based on this scoring model, new essays can be immediately scored by analysis of the features and applying the weights of the scoring model.

Evaluation in Automated Scoring

The performance of a scoring model should be evaluated both in how well the scores match human scoring and how well the scores align with the constructs of interest (e.g., Clauser et al., 2002; Williamson, Xie & Breyer, 2012). The most common benchmark is to assess the reliability of the scoring engine by examining the agreement of the predicted scores to human scorers compared to the agreement between human scorers. Metrics for computing the reliability include correlation, kappa, weighted kappa, and exact and adjacent agreement. Using true scores (e.g., the average of multiple scorers or the consensus score) for the comparison can provide more accurate measures of the IEA's accuracy. However, human agreement is seldom sufficient as a means to evaluate performance. The IEA is often compared against external variables that provide a measure of the validity of the scoring, including comparison of IEA scores with scores from concurrent administrations of tests with a similar construct, agreement with scores from subsequent tests, agreement to scorers with different levels of skill, and tests of scoring across different population subgroups.

Experiment 1: Effects of training set size on scoring performance

Typically, the sample of student responses used for training and evaluating the scoring engine should represent the full range of student responses and scores and be a large enough set to allow the IEA to generalize to the population of expected responses it will score. A general rule of thumb has been that increasing training set size should improve the performance. In the first test, we evaluated the performance of the IEA based on the number of essays used in the training set. A set of four prompts designed for college-age students were used. Each prompt was scored by two independent human raters on an eight point scale and a resolved score was derived by a third rater if there was disagreement From each prompt, 150 students responses were randomly chosen as a test set. From the remaining responses, a number *m* were randomly selected as the training set, where *m* varied from 50 increasing by 50 until the number of responses was exhausted. For each *m,* a scoring model was developed using the training responses and performance measures were made on the test set. In order to get a highly stable estimate of the performance at each m, the process was repeated 30 times and the performance numbers were averaged.

Three performance plots are shown below, figure 1 for the Pearson r between the resolved and IEA predicted scores, figure 2 for exact agreement between human scorer and IEA predicted scores and figure 3 for the adjacent agreement. All three figures show increasing performance as the training set size is increased to about 200 responses at which point it begins to level off. The results indicate that beyond about 200 essays in the training set, there are slight (<5%) improvements, but the majority of the performance benefit can be obtained with 200 essays in the training set. As a comparison, Table 1 shows the human-human correlation and exact agreement for the four prompts.

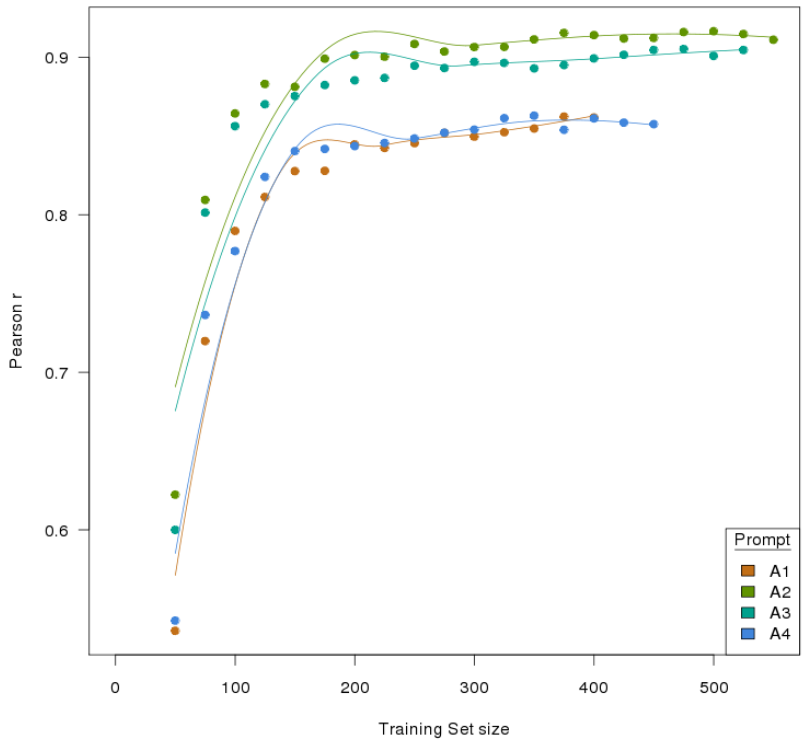

Figure 1. Correlation of predicted vs. resolved scores by training set size

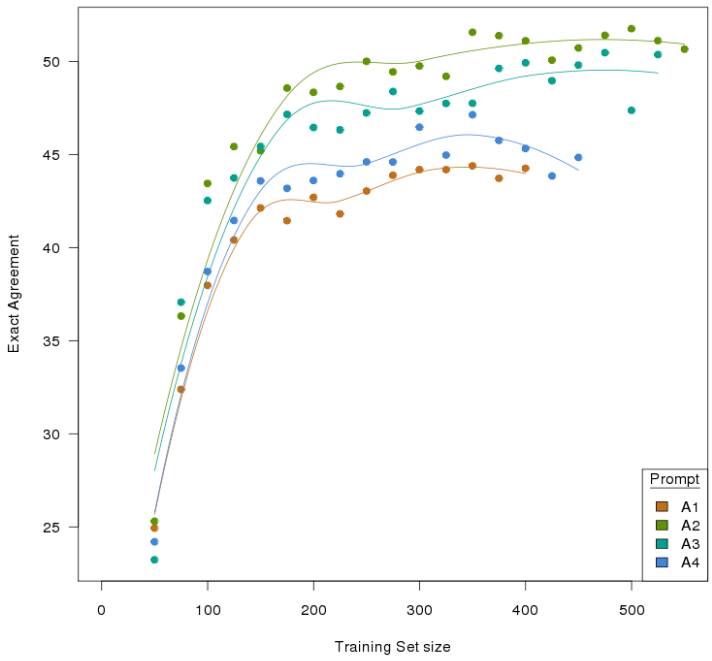

Figure 2. Exact agreement of predicted vs. resolved scores by training set size

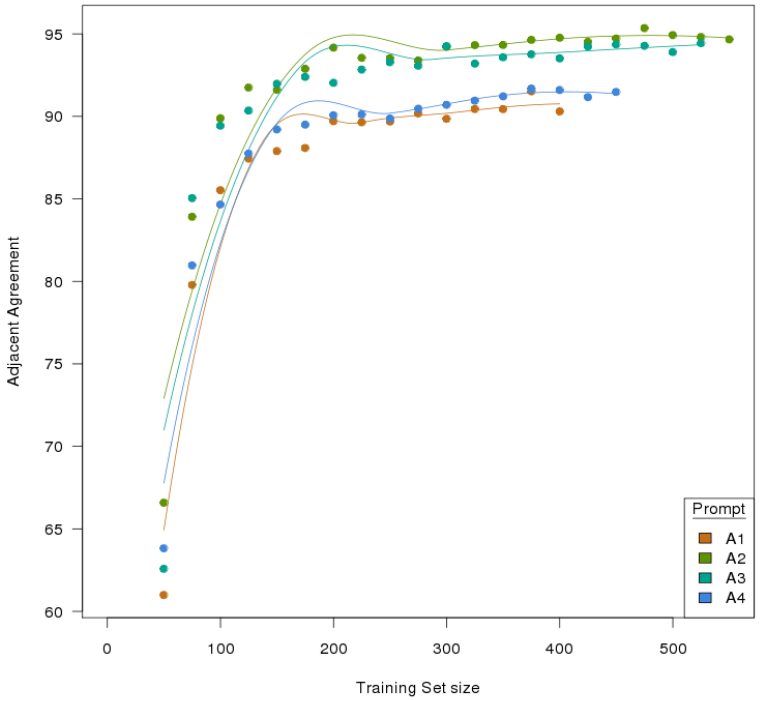

Figure 3. Adjacent agreement of predicted vs. resolved scores by training set size

Table 1. Human-human reliability for the four prompts.

Prompt	Pearson r	Exact agreement	Adjacent agreement
A1	0.88	56.4	94.8
A2	0.92	64.0	95.3
A3	0.91	60.1	95.1
A4	0.90	59.5	94.7

Experiment 2: Effects of the Distribution of Scores on Performance

The above results show that a random sample of training essays can provide an acceptable level of performance for the scoring models. Generally, a random sample will recreate a normal distribution of essays, assuming that the original set had a normal distribution. However, because automated essay scoring systems generalize from the training set, it is

critical to ensure that the set provides a sufficient representation of essays at the different score points. In the second test, we vary the number of training essays in the lowest score points in order to determine how many are needed in order to provide good scoring performance.

The same essay prompts from Experiment 1 were used. To develop the training set, the responses at the lowest score point were removed, because there typically were not enough responses at that score point to perform the experiment. This left a 7 score point item. From the lowest score point 20 responses were randomly chosen and then 130 additional responses were randomly chosen from the remaining score points, yielding a test set of 150 responses. From the remaining responses, a number *k* were randomly selected at the lowest score point, and 300-*k* were selected from the remaining score points, yielding a training set of size 300. *k* was systematically varied from 1,3,5 and then incremented by 5 until the number of lowest score point items was exhausted for each prompt. For each *k*, a model was estimated using the training responses and performance measures were made on the test set. At each *k*, the process was repeated 30 times and the performance numbers were averaged to provide a stable estimate. Figure 4 shows the mean prediction error for just the responses at the lowest score point (absolute value of human resolved score minus IEA predicted score). Figure 5 shows the correlation of the predicted scores for the whole training set as a function of the number of essays in the lowest score point. Overall, the results show that performance maximizes at around 20-40 essays in the lowest score point. There is a slight indication that at much greater numbers of essays in the lowest score point (e.g., 50-70 essays) that performance starts to decrease. This may be due to the fact that the lowest scoring essays are overrepresented in the set relative to the essays at the other score points.

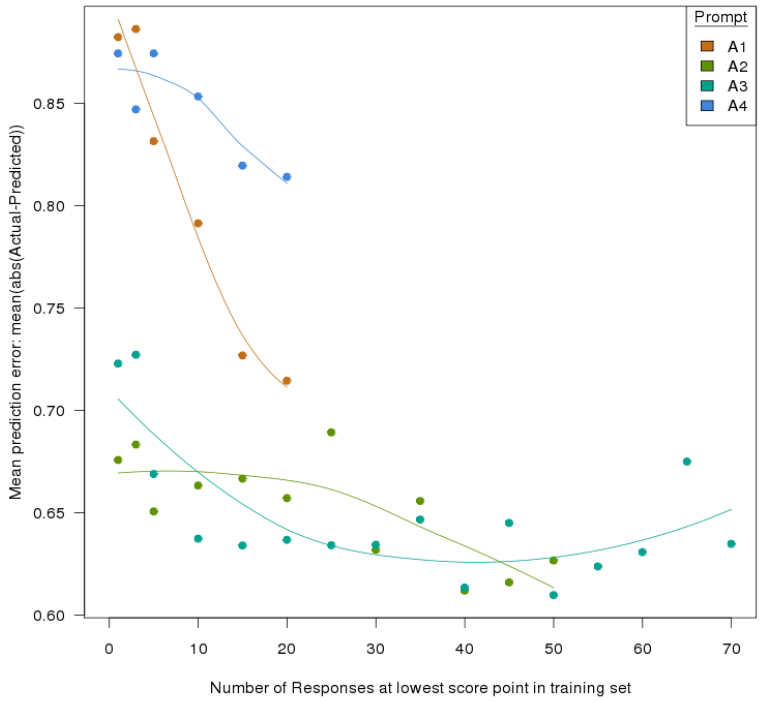

Figure 4. Mean prediction error for responses at the lowest score point based on the number of essays at the lowest score point in the training set.

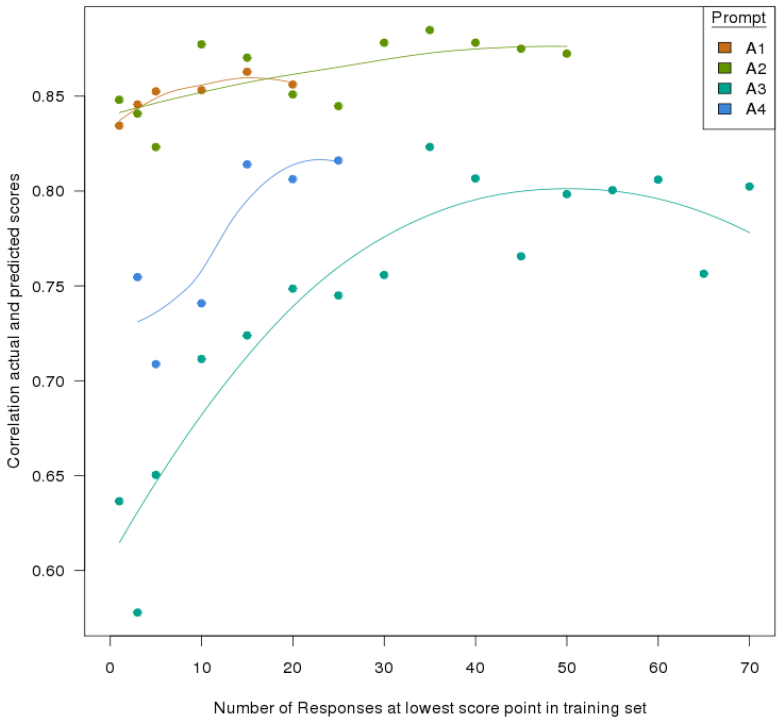

Figure 5. Correlation to resolved scores based on the number of essays at the lowest score point in the training set.

Experiment 3: Effects of human reliability on automated scoring reliability

The IEA is trained to model human scorers through learning to associate human scores on the essays to features in the essays. However, human reliability in essay scoring can vary greatly. This can be due to a number of factors including the type of prompt, the quality and definition of the construct being scored, and the amount of training and experience of the human scorers. If human raters do not agree well on the construct being assessed, it will be evident in more inconsistent scoring and a weaker relationship between the scores given and the features expressed in the essays. We investigated this effect by examining how well the human agreement on different prompts affected the agreement rates of automated scoring engine. We used 87 varied prompts, ranging from fourth grade through high school. All prompts were scored on a six point scale. The human exact agreement rates varied from 43% to 87%. The

prompts were then trained with the IEA and the exact agreement of the IEA score to the human resolved scores were computed for each prompt. Figure 6 shows the correlation of the human-human agreement to the IEA-human agreement. The results show a strong relationship (r=.73) between the agreement rates. As human agreement goes up, so does the IEA's agreement to the human raters. As noted above, there are a variety of factors that may cause agreement rates to be higher or lower. However, the results indicate that overall automated scoring performance improves with better human agreement.

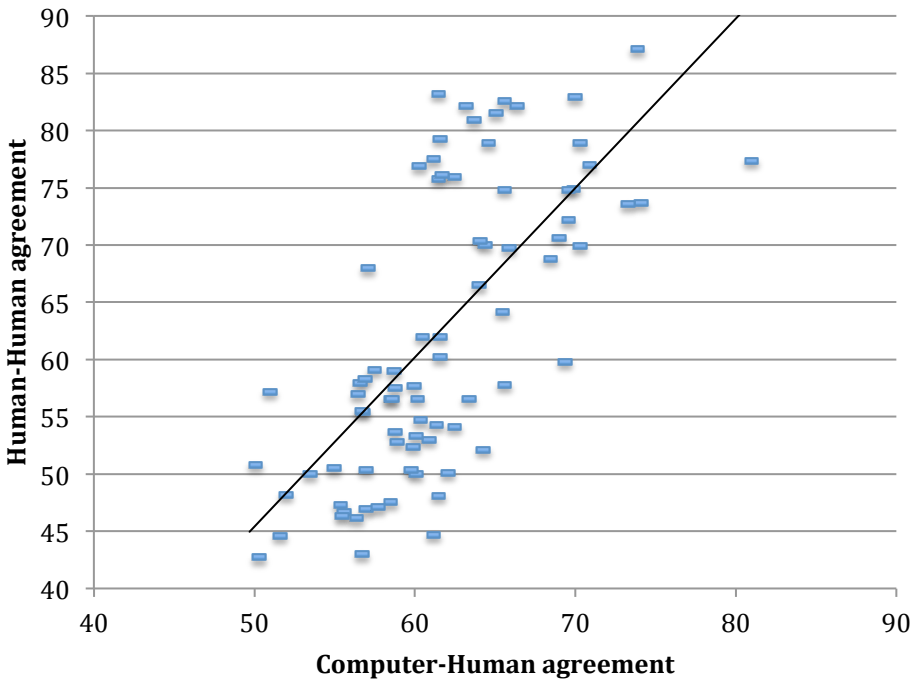

Figure 6. Agreement for human-human vs. human-IEA exact agreement for 87 varied prompts.

Experiment 4: Detection of essays that may not be scored as reliably

The IEA provides the automated capability to detect essays that may not be scored reliably. These essays are typically detected for being off-topic, highly unusual, or containing features that are unlike those that the IEA can assess. Based on a client's needs, these essays can either receive scores and an advisory, or could be passed on to scorers or teachers

for scoring. While detecting "unscorable" essays may increase the number of essays that require human scoring, it does provide a means to ensure that essays that should not be automatically scored can be reviewed.

In order to test whether the essays that were receive "advisories" are those that would likely be score less reliably by the IEA, the system was tested on approximately 24,000 student responses to two essay prompts from an April 2011 online administration for the State of Texas Assessments for Academic Readiness (STAAR™ English I writing assessment (See Davis, Lochbaum, Murphy and Foltz., 2012), Hembry, Davis, Murphy, Lochbaum and Foltz 2012). The IEA flagged 717 essays as having advisories for the first prompt and 630 essays for the second prompt. These numbers exclude essays that were flagged for being too short. All of the essays were scored by the human raters and the IEA. Distributions of the scores given by the IEA and the resolved score for the humans are shown in Table 2. Generally, the IEA's distribution of scores for advisory essays matched the score distribution of the human raters.

Table 2. Score distribution for advisory essays.

	Score point			
	1	2	3	4
Prompt 1 Human	210	201	173	133
Prompt 1 IEA	213	179	220	105
Prompt 2 Human	272	229	90	39
Prompt 2 IEA	200	238	131	61

We then compared the agreement rates of the IEA scores for essays that were deemed as having "advisories" versus those that did not receive advisories ("Good" set). Table 3 shows the exact and adjacent agreement rates between the IEA and the resolved human scores for the advisory and good essay sets. Overall, the results show that scores provided by IEA for the

"advisory" essays have lower agreement rates to the human scorers. The IEA's exact agreement rate is about 2% lower for the advisory essays than the "good" essays. The results indicate that the IEA's detection methods are effective at determining essays that likely will not be scored as accurately.

Table 3. IEA agreement to human scorers for "advisory" essays vs. "Good" essays

Prompt	Advisory type	Exact Agreement	Adjacent Agreement
Prompt 1	Advisory	53.0	96.9
	Good	55.2	97.9
Prompt 2	Advisory	58.9	96.2
	Good	60.7	98.6

Conclusions and implications for implementing reliable automated assessments

The development of automated essay scoring technology requires systematic investigations into how to optimize automated scoring performance throughout development and implementation. Along with providing evidence that automated scoring can score reliably, the results of the above experiments provide additional guidance to help in ensuring reliable scoring.

For the collection of essays used to train the scoring system, the results indicate that scoring performance for the IEA is generally robust when 200 or more essays are provided in the training set. As a general rule of thumb, the results indicate that for general formative and content-based scoring, 200-300 essays should be the minimum to train the scoring engine. For an essay prompt in a high stakes assessment, a sample of about 500 student responses would preferred since even modest gains in agreement with human scoring can be important for improving reliability and defensibility for high stakes classification decisions (i.e. graduate vs. do not graduate).

The results further indicate that the training essays should represented a good (normal) distribution, while ensuring that there are sufficient (e.g., at least a minimum of 10-20) examples at each score point. In order to ensure that the IEA can create a reliable model of the scoring features, the essay set needs to have high human reliability in scoring. Reliable human scoring can be achieved through ensuring that there are clearly defined objectives of what is being tested, a focused topic for the essay prompt and clear and consistent rubric language on how the prompt will be scored. Thus, there must be a well defined construct and consistent scoring rubric for the essays. For training the system, the responses should be 100 percent double-scored by human scorers and also receive resolution scores for non-adjacent agreement. By having scores from multiple human scorers, the IEA can be trained on something closer to the true score (e.g., the average of multiple human raters) rather than the scores of an individual rater. The goal is to have as much, and as accurate, information as possible about how a response should be evaluated.

Finally, once the scoring system is implemented, the results of this paper show that the IEA can be effective at monitoring the performance of the scoring and detecting essays that may not be scored as accurately. Depending on the context and needs of the assessments, these essays can be flagged for human scoring or given advisories.

Evaluation of the performance of a scoring engine should be performed throughout the test development process. In the pilot testing phase of item development, evaluation can performed to determine how amenable items are for automated scoring. Before deployment, finalized scoring models can be evaluated on held-out tests sets to determine generalizability and robustness of the scoring models. During deployment, evaluation of the scoring engine is often performed to ensure that the scoring remains consistent with the goals of the testing. In the case of the IEA being used as the sole scorer, random samples of essays can be chosen for backreads by human scorers as a check on the automated scoring. In the case of the IEA

being used as a second scorer, agreement rates with the other human scorer as well as with resolution scorers can be constantly monitored for performance. In addition, when used as a second scorer, evaluation of the agreement with human scorers can be used to detect drift in the human scorers, scorer consistency and possible lack of homogeneity in the test population in comparison to the training set. In the current paper, we examined several approaches to improving the reliability throughout the development process. In each case, reliability measures (agreement and correlation) were used to determine the effects on improving automated scoring performance. Understanding the factors that can optimize reliability in automated scoring is important to support the choices of testing programs in using automated scoring within formative, summative, and high stakes contexts.

References

Clauser, B. E., Kane, M. T., & Swanson D. B. (2002). Validity issues for performance-based test score with computer-automated scoring systems. *Applied Measurement in Education, 15*(4), 413-432.

Davis, L, Lochbaum, K. E., Murphy, D., & Foltz, P. W. (2012) Automated Scoring for High Stakes K-12 Writing Assessment. Talk presented at the Association for Test Publishers. Palm Springs, CA. February. .

Foltz, P.W., Laham, D., and Landauer, T.K. (1999). The Intelligent Essay Assessor: Applications to Educational Technology. *Interactive Multimedia Electronic Journal of Computer-Enhanced Learning,* 1, 2, http://imej.wfu.edu/articles/1999/2/04/index.asp

Hearst. (2000). The debate on automated essay grading. *Intelligent Systems and Their Applications, 15*(5), 22-37.

Hembry, I., Davis, L.L., Murphy, D., Lochbaum, K., & Foltz, P. (April, 2012). Evaluating an automated essay scoring engine for high stakes consideration: Validity evidence and a generalizability approach for disaggregating error. Paper presented at the annual meeting of the National Council on Measurement in Education, Vancouver, British Columbia.

Landauer, T. K., Laham, D., & Foltz, P. (2003). Automated essay assessment. *Assessment in Education, 10*(3), 295-308.

Topol, B. Olson, J. and Roeber, E. (2011). The cost of new higher quality assessments: A comprehensive analysis of the potential costs for future state assessments. Stanford Center for Opportunity Policy in Education.

Wang, Jinhao, and Michelle Stallone Brown (2007). "Automated Essay Scoring Versus Human Scoring: A Comparative Study", p. 6. Journal of Technology, Learning, and Assessment, 6(2)

Williamson, D. M., Bennett, R., Lazer, S., Bernstein, J., Foltz, P. W., Landauer, T. K., Rubin, D., Way, D., and Sweeney, K. (2010, June). *Automated Scoring for the Assessment of Common Core Standards.*

Williamson, D. M., Xi, X., & Breyer, F. J. (2012). A Framework for Evaluation and Use of Automated Scoring. Educational Measurment: Issues and Practice, 31(1), 2–13.