Measuring the Power of Learning.®

Research Report
ETS RR–16-30

Evaluating the Advisory Flags and Machine Scoring Difficulty in the *e-rater*® Automated Scoring Engine

Mo Zhang

Jing Chen

Chunyi Ruan

December 2016

ETS Research Report Series

Since its 1947 founding, ETS has conducted and disseminated scientific research to support its products and services, and to advance the measurement and education fields. In keeping with these goals, ETS is committed to making its research freely available to the professional community and to the general public. Published accounts of ETS research, including papers in the ETS Research Report series, undergo a formal peer-review process by ETS staff to ensure that they meet established scientific and professional standards. All such ETS-conducted peer reviews are in addition to any reviews that outside organizations may provide as part of their own publication processes. Peer review notwithstanding, the positions expressed in the ETS Research Report series and other published accounts of ETS research are those of the authors and not necessarily those of the Officers and Trustees of Educational Testing Service.

The Daniel Eignor Editorship is named in honor of Dr. Daniel R. Eignor, who from 2001 until 2011 served the Research and Development division as Editor for the ETS Research Report series. The Eignor Editorship has been created to recognize the pivotal leadership role that Dr. Eignor played in the research publication process at ETS.

ETS Research Report Series ISSN 2330-8516

RESEARCH REPORT

Evaluating the Advisory Flags and Machine Scoring Difficulty in the *e-rater*® Automated Scoring Engine

Mo Zhang, Jing Chen, & Chunyi Ruan

Educational Testing Service, Princeton, NJ

Successful detection of unusual responses is critical for using machine scoring in the assessment context. This study evaluated the utility of approaches to detecting unusual responses in automated essay scoring. Two research questions were pursued. One question concerned the performance of various prescreening advisory flags, and the other related to the degree of machine scoring difficulty and whether the size of the human–machine discrepancy could be predicted. The results suggested that some advisory flags operated more consistently across measures and tasks in detecting responses that the machine was likely to score differently from human raters than did other flags, and relatively little scoring difficulty was found for three of the four tasks examined in this study, with the relationship between machine and human scores being reasonably strong. Limitations and future studies are also discussed.

Keywords Automated essay scoring; unusual response; advisory flag; machine scoring difficulty; e-rater

doi:10.1002/ets2.12116

Automated scoring may be thought of as the machine grading of constructed responses that are not amenable to approaches relying on exact matching (such as correspondence with a list of key words; Bennett & Zhang, 2016). These answers are not suitable for exact-matching approaches in that the specific form(s), and/or content of the correct answer(s) is not known in advance. Automated scoring has been employed in multiple content areas including mathematics, science, and the English language arts (e.g., for writing and speaking ability). Regardless of the content area, these scoring methods typically focus on the extraction and aggregation of features from the constructed responses.

In scoring essay responses, which is the subject of this paper, natural language processing methods are commonly used for feature extraction (e.g., grammatical error detection and word association). After feature extraction, evidence is combined, usually by assigning weights to the various response features and then aggregating the weighted feature values. Weights can be set by a panel of experts or derived by regressing the human ratings on the features. The model is then applied to predict the score that a human rater would have given to an unseen essay.

Unusual Responses in Automated Essay Scoring

Among the measurement issues that have not been fully addressed in the field of automated scoring is the detection of unusual responses. In multiple-choice testing, *unusual response* typically refers to an unexpected pattern across answers for an examinee (e.g., incorrect answers to easy questions and correct answers to more difficult questions). A large body of research exists on the detection of unusual response patterns via person-fit statistics (e.g., Karabastos, 2003; Reise & Due, 1991; Rupp, 2013). For automated scoring, however, the focus is on the characteristics of a single but complex item response rather than on a pattern across responses.

For our present purposes, unusual responses are defined to be those answers that are not suitable for machine scoring due to response characteristics that the scoring system cannot accurately handle but that experienced human raters can more often effectively process. Among the characteristics that may lead to an unusual response are off-topic content, foreign language, unnecessary text repetition, random keystrokes, extensive copying or paraphrasing from source materials, prememorized text, unusually creative content (e.g., highly metaphorical), unexpected organization or format (e.g., a poem), or text segments that cannot be processed because the automated scoring system itself is imperfect.

Several aspects of this definition merit comment. First, this definition suggests that these responses result from the interaction between the limitations of the scoring system and the behavior of examinees with respect to a type

Corresponding author: M. Zhang, E-mail: mzhang@ets.org

of assessment task. Such limitations may be particular to an automated scoring system or more generally linked to the state of the art. Second, the definition indicates that one common indicator of unusualness in a response should be a disagreement between machine and human scores because the response has characteristics that more frequently exceed the current capabilities of the machine than those of experienced human raters.[1] Last, the definition makes no assumptions with respect to the examinee's intentions (e.g., purposeful attempts to "game" the system). Such an inference is not necessary for identification and handling of unusual responses.

Whereas there is a large literature on unusual response in multiple-choice testing, there is only limited research for automated essay scoring. In one investigation, Powers, Burstein, Chodorow, Fowles, and Kulich (2001) attempted to trick a machine scoring system by repeating the same paragraphs many times so as to increase text length. In another study, Higgins, Burstein, and Attali (2005) developed a method based on vocabulary patterns to detect off-topic responses. More recently, Chen, Zhang, and Bejar (in press) proposed a method to improve the prediction accuracy of off-topic responses. Finally, several recent publications addressed gameability in automated scoring (e.g., Bejar, VanWinkle, Madnani, Lewis, & Steier, 2013; Higgins & Heilman, 2014). Most of the above investigations were experiments in which scores were compared before and after some manipulation, such as increasing the complexity of the vocabulary and adding shell language that did not necessarily connect to the content.

Detection of Unusual Responses

Whether automated scoring is sensitive to atypical responses and how it processes them affects how the resulting scores can be interpreted and used. Our confidence will understandably be diminished if the scoring system fails to handle those responses effectively.

Unusual responses may be detected at two times in the processing stream. The first occasion is the prescreening stage before scoring occurs. Intentionally or not, individuals may generate answers that are nonsensical or otherwise atypical. Such atypical responses may be blank, have random keystrokes, be off topic, or have unusual linguistic structure. Advisory flags are commonly used to detect such answers during the prescreening stage. In some instances, these responses can still be processed automatically without human involvement. Examples include an empty submission, an essay with language different from the target language, or an essay consisting of a complete copy of the prompt text. In other instances, the unusual response is sent to a human rater for processing, bypassing the automated scoring system entirely.

The second occasion is the post hoc screening stage. Because automated scoring cannot fully measure some of the higher level aspects of the essay-writing construct, most consequential testing programs employ both human and machine scoring methods. Answers that are possibly inappropriate for machine scoring indicated by, for example, low human – machine agreement, can be detected and sent to additional human raters. This supplementary processing is typically triggered by a difference between the machine and human scores of more than a predetermined threshold set by policy decision.

The cost and time required for human scoring have motivated many large-volume testing programs to consider automated scoring as a primary method (e.g., Common Core State Assessments; Educational Testing Service [ETS], 2014a; Partnership for Assessment of Readiness for College and Careers, 2010; SMARTER Balanced Assessment Consortium, 2010). For this scenario to be workable, the efficacy of prescreening and post hoc methods must first be established. For prescreening, the evidence would include confirming that the advisory flags accurately identify responses likely to be inaccurately scored automatically. For post hoc screening, it might include creating means for predicting the chances that a response would have produced a sizeable human – machine disagreement had it been judged by a human rater. Specifically, if machine-scoring difficulty can be accurately predicted, human raters can be involved only if the automated scores were deemed potentially problematic. The effectiveness of this particular approach, however, has not been sufficiently studied.

Prescreening Advisory Flags in *e-rater*®

The automated scoring system used in this study was *e-rater*® v13.1 (Attali & Burstein, 2006). Developed at ETS, e-rater has been used in various testing programs for purposes ranging from classroom assessment to graduate and professional school admissions (ETS, 2014b; ETS, 2014c). In most testing programs, the e-rater scoring model is calibrated through the multiple linear regression of human ratings onto text features such as vocabulary sophistication; essay development and organization; and absence of grammar, mechanics, usage, and style errors.

Table 1 Prescreening Advisory Flags in e-rater

Flag ID	Label	Description
#1	Repetition	May contain too many repetitions of words, phrases, sentences, or text sections
#2	Insufficient development	May not show enough development on topic or concept, or may provide insufficient evidence to support the claims
#3	Off topic	May not be relevant to the assigned topic
#4	Restatement of prompt text	Appears to be a restatement of the prompt text with few additional concepts
#5	Too short	May be too short to be reliably automatically scored
#6	Too long	May be too long to be reliably automatically scored
#7	Unusual organization	May contain unusual organizational elements that cannot be recognized by the automated scoring system
#8	Excessive number of problems	May contain unusually large amount of errors in grammar, mechanics, style, and usage, which may result in unreliable automated scores

The e-rater system uses several prescreening advisory flags to identify responses that the system is likely to misscore. For the present investigation, we analyzed eight advisories available for the essay tasks we examined. Each advisory indicates some questionable aspect of an essay submission (see Table 1). These questionable aspects would be expected to occur in most writing assessment programs and, as a result, comparable prescreening mechanisms have been commonly included in other automated scoring systems (Foltz, Laham, & Landauer, 1999; Page, 2003; Vantage Learning, 2012).

In some of the assessment programs that use e-rater, a type of post hoc screening has also been implemented (i.e., in addition to prescreening advisories like those above). That post hoc screening entails evaluating the discrepancy between the automated score and a human rater's score for each response. When the human – machine discrepancy exceeds a given threshold, a second human rating is solicited. Whereas the specific thresholds employed in operational settings have not been reported, prior research has evaluated thresholds from 0.5 to 1.5 on a 5- or 6-point holistic scoring scale (e.g., Zhang, Breyer, & Lorenz, 2013).

Purpose of This Study

This investigation evaluates the usefulness of approaches for detecting unusual responses as a step toward supporting the use of automated scoring as a primary method. Although various prescreening advisory flags have been integrated into automated essay scoring systems (e.g., Intelligent Essay Assessor, Pearson Education, 2010; IntelliMetric, Vantage Learning, 2012), little research on the utility of those advisory flags has been published. In this investigation, we studied the usefulness of such prescreening flags. In addition, as a precursor to developing a general post hoc screening method, we investigated whether the size of the human – machine discrepancy could be predicted.

Research Questions

We asked two research questions, one concerning the prescreening stage and the other, post hoc screening.

Research Question 1. Are the advisory flags at the prescreening stage useful in detecting responses that the machine is likely to score differently from human raters?

RQ1.1 Is the mean absolute human – machine discrepancy greater for flagged than for nonflagged responses?
RQ1.2 Is human – machine agreement lower for flagged than for nonflagged responses?

Research Question 2. For responses that pass through prescreening, can the size of the human – machine discrepancy be predicted well enough to support an effective postscreening mechanism?

The rationale for posing these research questions is related to supporting the use of automated scoring as a primary method. An answer to the first question will give a measure of the utility of the advisory flags by indicating whether unusual responses (in terms of lower levels of machine – human agreement) can be detected at the prescreening stage for routing to human raters. Note that lower levels of agreement for unusual responses are expected by definition. This expectation is because the purpose of flagging is to indicate a type of response that the machine can only score with higher-than-acceptable uncertainty due to known system limitations relative to well-monitored and carefully trained human raters. As

an example, an essay that contains a well-formulated argument but has many grammatical and mechanical errors would be expected to receive a lower machine score because of the machine's relative inability to judge content and quality of argument.

For essays that get through prescreening, an answer to the second question will suggest whether mechanisms could be created to predict which of those essays would have been likely to produce low human–machine disagreement had they been processed by humans. Accurately identified, such essays could bypass human review entirely, facilitating the sole use of automated scoring.

It is important to note that data relating to the identification of unusual responses is only one piece of evidence needed for evaluating the validity of automated scores for given purposes. Depending upon a test's purpose, other important evidence relates to the automated scoring model (e.g., the construct relevance of the features), generalization (i.e., the degree to which scores on one task associate with scores on other tasks from the universe), external relations (i.e., the degree to which expected relationships with indicators of different and similar constructs are observed), population invariance (i.e., the extent to which scores operate similarly across demographic groups), and impact on learning and teaching practice (Bennett & Zhang, 2016).

Method

Instrument

We used writing responses collected from four essay tasks given in two large-scale, high-stakes testing programs. In one task, examinees were asked to express their opinion on a common issue. In a second task, examines were asked to compose a synthesis of a short article and an audio recording. The score scale for these first two tasks ranged from integer 1 to 5. In a third task, examinees were asked to evaluate an argument by assessing the claims and evidence it provided. Finally, in a fourth task, examinees were asked to construct an argument on a given issue with reasons and examples to support their views. The score scale for these two latter tasks ranged from integer 1 to 6. Included in this study were 71 different prompts for Task 1, 72 prompts for Task 2, 76 prompts for Task 3, and 76 prompts for Task 4.

Data Set

Essay responses were collected between April 2013 and March 2014. The total number of responses was approximately 871,000 for Task 1, 873,000 for Task 2, 516,000 for Task 3, and 520,000 for Task 4. Responses that were flagged accounted for about 5%, 9%, 13%, and 4% of the total sample, respectively for the four tasks. All responses were scored by at least one human rater and e-rater, while a subset was further graded by a second randomly assigned human rater ($n = 40{,}851$ for Task 1, 40,426 for Task 2, 20,355 for Task 3, and 20,153 for Task 4). For Research Question 2, a subset of the total sample was used to examine the extent to which human–machine discrepancy could be predicted. For each task, all responses were included except the ones automatically flagged by the testing programs' prescreening processes.

Data Analyses

Because advisories are intended to detect responses that the machine would not be expected to effectively score, responses with advisory flags should generate lower human–machine agreement than responses triggering no advisory flag. Consequently, for Research Question 1, we compared the means of the absolute differences in human–machine discrepancy between flagged and nonflagged responses separately for each of the advisories using Cohen's *d*. Absolute difference was employed because positive and negative discrepancies can cancel out, hiding large differences between scoring methods.

We next compared the machine–human agreements between the flagged and the nonflagged groups. For this purpose, we used the Pearson correlation coefficient (r), quadratically weighted kappa (QWK), and standardized mean score difference (SMD), with the pooled variance of the machine and human scores as the denominator. The first two statistics denote human–machine agreement at the individual response level, and the last statistic (SMD) reflects distributional differences.

For Research Question 2, a two-step procedure was employed. In the first step, we investigated the extent to which the machine had difficulty scoring responses. Scoring difficulty was evaluated in several ways, each of which employed

the output from cumulative logistic regression of the human ratings on the linguistic features extracted by the machine (Haberman & Sinharay, 2010). First, we evaluated the squared correlation between human scores and the machine scores produced by this regression. A high squared correlation would suggest that the machine had produced scores that tracked human ratings well. Second, we computed the mean squared error (MSE) between machine and human scores. A low MSE would suggest a close correspondence between machine and human scores. Third, we compared the conditional dispersion (CD) of the responses with the MSE:

$$\mathrm{CD} = \sum_{H=1}^{6} [H - E(H|M)]^2 \, P(H|M),$$

where H is the human score, M is the e-rater machine score resulting from the cumulative logistic regression model, $E(H|M)$ is the expected human score given M, and $P(H|M)$ is the probability of H given M. Because CD reflects the estimated expected response dispersion under the model and MSE depicts the observed values (though not assuming the model holds), CD and MSE values should be comparable to one another, with large differences suggesting a lack of consistency in machine scoring. For the purposes of this study, we considered an absolute difference greater than or equal to 0.10 as notable.

Last, for each response, we used the probability produced by the regression for each of the human-score categories. The standard deviation of those probabilities was computed for each response. A response that was difficult for the machine to score would be anticipated to have a very small standard deviation, indicating that the probability of assigning a score category was approximately equal across the range. On a 5-point scale (in Tasks 1 and 2), a response for which the probabilities were equal for all categories would have a standard deviation of 0, whereas a response with a score-category probability of 1 would have a standard deviation of approximately 0.45. This latter response would have a single score category predicted with certainty, implying no scoring difficulty. Similarly, on a 6-point scale (for Tasks 3 and 4), a response with equal probabilities across all score categories would have a standard deviation of 0, and when one of the six score-categories receives a probability of 1, the standard deviation would be approximately 0.41. To summarize results across the data set, the mean and range of the standard deviations were computed, and the distribution was examined.

To evaluate whether the machine had trouble judging responses at different score levels, we computed both MSE within each score level and the correlation of the standard deviation of the probabilities with human scores. For purposes of computing MSE, depending on the scale of the human scores, eight or 10 score levels were created using the machine scores, running from 1 to 5 (for Tasks 1 and 2) or 1 to 6 (for Tasks 3 and 4), in increments of 0.5. The MSE between human and machine scores was computed using both the overall sample and the double human-scored sample. In addition, the MSE between the two human ratings was computed and contrasted with the machine–human MSE. This contrast was made to detect the extent to which scoring difficulty was also manifest in human ratings, which are commonly known to have limitations (e.g., scale shrinkage and inconsistency; Zhang, 2013).

In the second step, a linear regression model was calibrated to predict the size of the absolute discrepancy between human scores and the machine scores resulting from the cumulative logistic regression. The predictors were the e-rater linguistic features, advisory flags not used by the testing program for prescreening, and two more linguistic features. These two features indicated the overlap in vocabulary of the target essay with essays at different score levels. This predictive model was assessed using the Pearson correlation coefficient between the predicted and observed human–machine disagreements. A high correlation would suggest that the size of the disagreement could be predicted and possibly employed as a component in a postscreening process.

The indices described above were computed for the overall sample, as well as for the top five test-center countries/territories based on examinee volume.

Results

Results for Research Question 1

Table 2 shows the results for comparing the mean absolute value of the human–machine discrepancy between flagged and nonflagged response groups. This comparison shows the degree to which human and machine scores disagree at the level of individual responses.[2]

Table 2 Human – Machine Discrepancy Between Flagged and Nonflagged Groups

	Flagged group		Nonflagged group		
Flag ID	*N*	Mean of |Δ| (*SD*)	*N*	Mean of |Δ| (*SD*)	Cohen's *d*
Task 1					
#1	12,051	0.54 (0.42)	821,048	0.50 (0.39)	0.10
#2	1,019	1.17 (0.92)	821,048	0.50 (0.39)	1.71
#3	21,648	0.56 (0.47)	821,048	0.50 (0.39)	0.15
#4	804	0.49 (0.37)	821,048	0.50 (0.39)	−0.03
#5	147	1.34 (1.11)	821,048	0.50 (0.39)	2.15
#6	94	0.97 (0.56)	821,048	0.50 (0.39)	1.21
#7	9,915	0.51 (0.39)	821,048	0.50 (0.39)	0.03
#8	552	1.36 (0.93)	821,048	0.50 (0.39)	2.20
Task 2					
#1	12,404	0.85 (0.66)	798,341	0.78 (0.60)	0.12
#2	689	2.28 (1.50)	798,341	0.78 (0.60)	2.49
#3	50,714	0.93 (0.77)	798,341	0.78 (0.60)	0.25
#4	531	0.80 (0.63)	798,341	0.78 (0.60)	0.03
#5	257	1.95 (1.74)	798,341	0.78 (0.60)	1.95
#7	4,004	0.86 (0.63)	798,341	0.78 (0.60)	0.13
#8	593	2.06 (1.33)	798,341	0.78 (0.60)	2.13
Task 3					
#1	3,575	0.58 (0.47)	448,756	0.52 (0.42)	0.14
#2	2,687	1.16 (0.86)	448,756	0.52 (0.42)	1.51
#4	47,812	0.50 (0.40)	448,756	0.52 (0.42)	−0.05
#5	33	2.12 (1.41)	448,756	0.52 (0.42)	3.81
#6	161	0.52 (0.48)	448,756	0.52 (0.42)	0.00
#7	7,363	0.52 (0.41)	448,756	0.52 (0.42)	0.00
#8	65	1.41 (1.08)	448,756	0.52 (0.42)	2.12
Task 4[a]					
#1	—	—	—	—	—
#2	2,591	0.55 (0.39)	499,573	0.45 (0.35)	0.28
#4	10,243	0.48 (0.36)	499,573	0.45 (0.35)	0.08
#5	52	0.49 (0.21)	499,573	0.45 (0.35)	0.12
#6	360	0.52 (0.45)	499,573	0.45 (0.35)	0.21
#7	7,017	0.44 (0.33)	499,573	0.45 (0.35)	−0.02
#8	127	0.60 (0.41)	499,573	0.45 (0.35)	0.44

Note. |Δ| = absolute value of human score minus machine score; *SD* = standard deviation; #1 = repetition; #2 = insufficient development; #3 = off topic; #4 = restatement; #5 = too short; #6 = too long; #7 = unusual organization; #8 = excessive problems.

As the table indicates, Advisory Flags #2 (insufficient development) and #8 (excessive number of problems) showed practically important, albeit small, effect across all four writing tasks (*d* values greater than 0.20). Advisory Flag #5 (too short) produced such effects for all but Task 4, whereas Advisory Flag #6 (too long) showed effects for only Tasks 1 and 4, and Advisory Flag #3 (off topic) showed an effect for only Task 2. No practically important effects were found for Advisory Flags #1 (repetition), #4 (restatement), and #7 (unusual organization) for any task (*d* value smaller than 0.20).

Table 3 presents three additional agreement statistics between human and machine scores for flagged responses and nonflagged responses. Included in the table are the human – machine SMD, Pearson correlation coefficient (*r*), and QWK.

For the SMD, all flagged groups produced values noticeably greater than the nonflagged groups with few exceptions (e.g., Advisory Flag #4 — restatement of prompt text — in Tasks 1 and 3). Across all four tasks, the largest differences were for Advisory Flags #2 (insufficient development), #5 (too short), and #8 (excessive number of problems), each of which identified responses for which the machine gave a notably lower score on average than did the human raters. These advisories are also the ones that functioned most effectively in terms of *d*.

With respect to the Pearson correlation coefficient, the values for two advisory flags (#5: too short and #8: excessive number of problems) were considerably lower for the flagged groups than for the nonflagged group for all four tasks. Advisory Flag #6 (too long) showed a similar pattern except for Task 2 (where no response was flagged by the advisory). Among the remaining five advisories, smaller differences were apparent for #4 (restatement of prompt text) in Task 1 and

Table 3 Human – Machine Agreements for Flagged and Nonflagged Groups

Flag ID	*N*	QWK	SMD	*r*
Task 1				
#1	12,051	0.65 (0.64, 0.66)	**0.15** (0.14, 0.16)	0.70 (0.69, 0.71)
#2	1,019	**0.58** (0.55, 0.61)	**−0.63** (−0.68, −0.58)	0.83 (0.81, 0.85)
#3	21,648	0.76 (0.75, 0.77)	**−0.19** (−0.20, −0.18)	0.80 (0.80, 0.81)
#4	804	**0.59** (0.54, 0.64)	0.03 (−0.03, 0.09)	**0.64** (0.59, 0.68)
#5	147	**0.05** (0.01, 0.08)	**−1.38** (−1.59, −1.16)	**0.28** (0.12, 0.41)
#6	94	**0.13** (0.06, 0.21)	**1.69** (1.48, 1.90)	**0.51** (0.35, 0.65)
#7	9,915	0.64 (0.63, 0.65)	**−0.21** (−0.22, −0.19)	0.71 (0.70, 0.72)
#8	552	**0.17** (0.13, 0.20)	**−1.46** (−1.55, −1.36)	**0.49** (0.43, 0.55)
Nonflagged group	821,048	0.66 (0.66, 0.66)	0.00 (0.00, 0.01)	0.70 (0.70, 0.70)
Task 2				
#1	12,404	0.59 (0.58, 0.60)	**0.31** (0.30, 0.31)	0.63 (0.62, 0.64)
#2	689	**0.22** (0.19, 0.26)	**−1.43** (−1.51, −1.35)	0.64 (0.59, 0.68)
#3	50,714	0.62 (0.62, 0.63)	**−0.24** (−0.25, −0.23)	0.67 (0.66, 0.67)
#4	531	**0.52** (0.47, 0.57)	**0.33** (0.25, 0.41)	0.58 (0.52, 0.63)
#5	257	**0.13** (0.10, 0.16)	**−1.22** (−1.37, −0.23)	**0.39** (0.28, 0.49)
#6	—	—	—	—
#7	4,004	0.57 (0.55, 0.59)	**−0.09** (−0.12, −0.06)	0.59 (0.57, 0.61)
#8	593	**0.14** (0.11, 0.17)	**−1.66** (−1.75, −1.57)	**0.48** (0.42, 0.54)
Nonflagged group	798,341	0.59 (0.59, 0.60)	0.02 (0.01, 0.02)	0.62 (0.62, 0.62)
Task 3				
#1	3,575	0.72 (0.70, 0.73)	**0.11** (0.09, 0.14)	0.76 (0.75, 0.78)
#2	2,687	**0.44** (0.43, 0.46)	**−0.89** (−0.93, −0.86)	0.78 (0.77, 0.80)
#3	—	—	—	—
#4	47,811	0.72 (0.71, 0.72)	0.03 (0.03, 0.04)	0.76 (0.75, 0.76)
#5	33	**0.05** (0.00, 0.09)	**−1.95** (−2.39, −1.50)	**0.37** (0.03, 0.63)
#6	161	**0.14** (0.00, 0.27)	**0.38** (0.18, 0.57)	**0.29** (0.14, 0.43)
#7	7,363	**0.64** (0.62, 0.65)	0.05 (0.03, 0.07)	**0.69** (0.68, 0.70)
#8	65	**0.16** (0.08, 0.24)	**−1.28** (−1.57, −1.00)	**0.45** (0.24, 0.63)
Nonflagged group	448, 756	0.73 (0.73, 0.73)	−0.02 (−0.02, −0.01)	0.76 (0.76, 0.77)
Task 4				
#1	—	—	—	—
#2	2,591	0.77 (0.75, 0.78)	**−0.40** (−0.43, −0.37)	0.86 (0.85, 0.87)
#3	—	—	—	—
#4	10,243	0.75 (0.74, 0.75)	**−0.16** (−0.17, −0.15)	0.80 (0.79, 0.81)
#5	52	**0.29** (0.26, 0.33)	**−1.48** (−1.85, −1.11)	**0.41** (0.13, 0.65)
#6	360	**0.36** (0.28, 0.44)	**0.33** (0.21, 0.44)	**0.53** (0.45, 0.60)
#7	7,017	0.70 (0.68, 0.71)	**0.10** (0.08, 0.12)	**0.76** (0.75, 0.77)
#8	127	**0.45** (0.33, 0.57)	**−0.91** (−1.01, −0.70)	**0.59** (0.45, 0.72)
Nonflagged group	499,573	0.77 (0.77, 0.77)	−0.04 (−0.04, −0.03)	0.81 (0.76, 0.77)

Note. QWK = quadratically weighed kappa; SMD = standardized mean score difference; *r* = Pearson correlation coefficient; #1 = repetition; #2 = insufficient development; #3 = off topic; #4 = restatement; #5 = too short; #6 = too long; #7 = unusual organization; #8 = excessive problems. Bolded values indicate cases where flagged group is different from nonflagged group by equal to or more than five points and the differences are in the expected direction. Values in the parentheses represent the 95% confidence interval. SMD is computed based on e-rater scores minus human scores, standardized by the pooled variance.

#7 (unusual organization) in Tasks 3 and 4. The three advisories (i.e., #1: repetition, #2: insufficient development, and #3: off topic) had machine – human agreement for the flagged group that was equal to or higher than the nonflagged groups.

Finally, generally similar results were found for the QWK statistic for all but Advisory Flag #2 (insufficient development). For this advisory, the QWK values were considerably lower for the flagged group than for the nonflagged group for Tasks 1 to 3, a result that was not observed in the *r* statistic.

Results for Research Question 2

The second research question concerned whether the size of the human – machine discrepancy for a response could be predicted. This question was addressed through a two-step process, with the first step being an evaluation of the extent

Table 4 Indicators of Machine Scoring Difficulty

	Task 1					Task 2			
Total/subgroup	*N*	R^2	MSE (*SD*)	*r*	Total/subgroup	*N*	R^2	MSE (*SD*)	*r*
Total sample	854,401	0.50	0.34 (0.50)	−0.40	Total sample	854,667	0.40	0.78 (1.06)	−0.07
China	263,940	0.44	0.32 (0.47)	−0.38	China	264,870	0.36	0.78 (1.06)	−0.05
USA	122,522	0.54	0.32 (0.48)	−0.39	USA	121,046	0.42	0.77 (1.06)	−0.21
Korea	66,594	0.49	0.37 (0.59)	−0.38	Korea	67,761	0.38	0.79 (1.09)	0.06
India	57,948	0.46	0.41 (0.55)	−0.40	India	57,544	0.31	0.79 (1.03)	0.33
Japan	45,280	0.57	0.31 (0.47)	−0.37	Japan	45,739	0.42	0.80 (1.09)	−0.29
	Task 3					Task 4			
Total/subgroup	*N*	R^2	MSE (*SD*)	*r*	Total/subgroup	*N*	R^2	MSE (*SD*)	*r*
Total sample	507,547	0.60	0.38 (0.60)	−0.62	Total sample	512,439	0.67	0.28 (0.43)	−0.53
USA	323,011	0.56	0.39 (0.64)	−0.60	USA	325,125	0.64	0.27 (0.41)	−0.55
India	76,511	0.51	0.39 (0.56)	−0.54	India	77,870	0.52	0.32 (0.48)	−0.30
China	47,319	0.34	0.31 (0.45)	−0.44	China	48,089	0.36	0.34 (0.51)	−0.22
Korea	6,125	0.51	0.28 (0.41)	−0.54	Korea	6,252	0.52	0.29 (0.43)	−0.38
Canada	5,441	0.57	0.39 (0.59)	−0.58	Canada	5,560	0.62	0.31 (0.46)	−0.56

Note. R^2 = squared multiple correlation between human and machine scores produced by the cumulative logistic regression; MSE = mean squared error between human and machine scores produced by the cumulative logistic regression; *SD* = standard deviation; *r* = Pearson correlation coefficient between the human scores and the standard deviation of the probabilities for the score categories yielded by the cumulative logistic regression.

to which the machine had difficulty in scoring. This step was undertaken because if little difficulty was encountered, human – machine discrepancy would be rare and hard to predict.

Several indicators of machine-scoring difficulty were examined. The two middle columns in Table 4 show (a) the squared multiple correlation between human scores and the machine scores produced by the cumulative logistic regression (R^2), and (b) the MSE between human and machine scores. These indices are given for the overall sample and for the top five countries/territories based on test-taker volume. For the total sample, the R^2 was 0.50 for Task 1, 0.40 for Task 2, 0.60 for Task 3, and 0.67 for Task 4. Except for Task 2, the R^2 suggests a reasonably strong relationship between machine and human scores. However, clear differences are evident among subgroup populations on this index, suggesting some variation with respect to scoring difficulty. For example, the R^2 ranged from as low as 0.44 to as high as 0.57 in Task 1, from 0.31 to 0.42 in Task 2, from 0.34 to 0.57 in Task 3, and from 0.36 to 0.64 in Task 4. A further examination of the countries/territories revealed that English-native speaking countries tended to have higher levels of R^2 than did non-English speaking countries/territories. In contrast to R^2, relatively little variation was observed for MSE (which is sensitive to differences in scores for individual responses, as opposed to differences in response ordering).

Not shown in Table 4 is a third scoring-difficulty indicator, the standard deviation of the probabilities assigned to each score level by the cumulative logistic regressions. For any given response, this value can range from 0, which reflects the most difficulty in distinguishing among score categories, to approximately 0.45 on a 5-point scale (Tasks 1 and 2) and 0.41 on a 6-point scale (Tasks 3 and 4), which reflects no difficulty. The mean standard deviations of the probabilities were 0.27 ($SD = 0.03$) for Task 1, 0.18 ($SD = 0.04$) for Task 2, 0.25 ($SD = 0.02$) for Task 3, and 0.25 ($SD = 0.03$) for Task 4, with Task 2 showing the most scoring difficulty. Figure 1 shows the distributions of the standard deviations by task. As the figure indicates, for Task 1, 3, and 4, most examinees fell in the upper half of the range of possible values, implying a relative lack of scoring difficulty. In contrast, most cases fell in the lower half of the range in Task 2, indicating some level of machine scoring difficulty.

We also examined scoring difficulty as a function of score level. Two indices were evaluated. One was the Pearson correlation coefficient between the human scores and the standard deviation of the probabilities for the score categories yielded by the cumulative logistic regression. This index is shown in the *r* columns of Table 4.

For all but Task 2, based on the sample as a whole, *r* ranged from −0.62 to −0.40, indicating a moderate relationship between machine-scoring difficulty and score level, such that the higher the score, the greater the difficulty. This index also had negative values for all top five subpopulations, though for some countries (e.g., China, Korea, India), the relationship

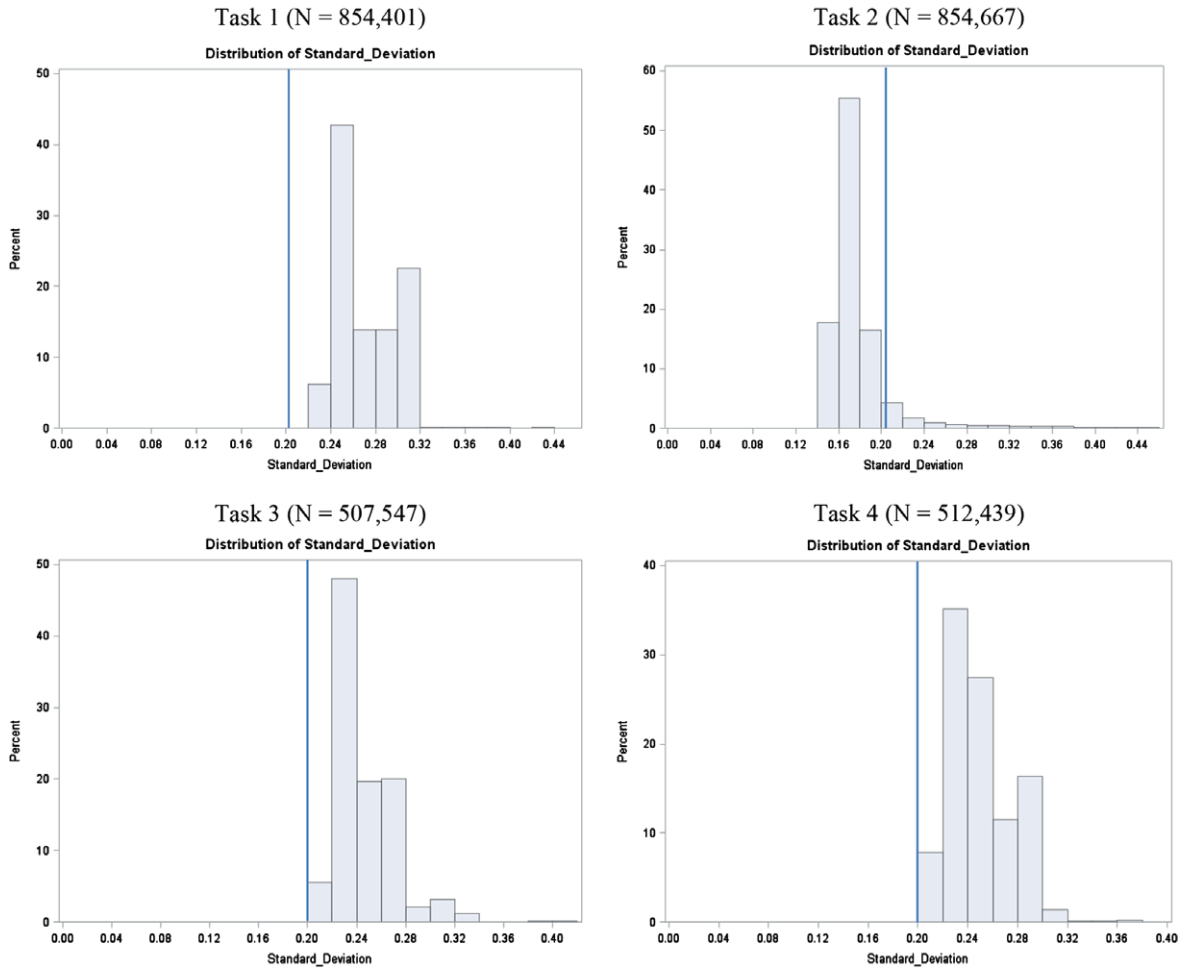

Figure 1 Distribution of the standard deviation of score-level probabilities for the overall sample.
Note. The vertical lines represent midpoints on the scale.

was weaker than for others (e.g., USA and Canada). For Task 2, however, the association of scoring difficulty with score level in the overall population was negligible and varied considerably from one subpopulation to the next.

The second index used to investigate the association of scoring difficulty with level was conditional MSE (based on the machine scores; Table 5), an individual-level indicator of machine–human disagreement. Consistent with the correlational analysis above, the largest MSEs were at the upper end of the scale on the scoring rubrics (i.e., the 3.5-to-4.0 and 4.0-to-4.5 range for Task 1 and the 4.5-to-5.0 and 5.0-to-5.5 ranges for Tasks 3 and 4). For Task 2, however, the largest MSEs occurred around the middle of the score scale—the 2.0-to-2.5, 2.5-to-3.0, and 3.0-to-3.5 ranges. This curvilinear result is in line with the limited correlation between the standard deviation of the probabilities and score level reported above for Task 2.

Finally, we examined the discrepancy between CD and MSE. A notable difference ($\geq$0.10) would suggest a lack of scoring consistency for the cumulative logistic regression at a given score level. Results showed notable discrepancies in the two highest score categories for Task 3 and the one lowest score category in Task 4.

Table 6 shows the MSEs computed by level using the double human-scored sample in each of the tasks as well as the human–machine MSE and the human–human MSE. The latter MSE gives an estimate of the scoring difficulty present in human ratings. Because this MSE can be viewed as a component of the human–machine MSE, a small difference between the two MSEs would suggest that most of the error in the human–machine MSE is attributable to human rating.

For Tasks 1, 3, and 4, this difference was always less than half of the human–machine MSE, suggesting that most of the scoring difficulty can be attributed to unreliability in the human ratings. For Task 2, however, the difference in MSEs is rather large, implying that most of the scoring difficulty is attributable to the machine scoring, a result consistent with

Table 5 Mean Squared Error by Score Level for the Overall Sample

	Task 1				Task 2		
Score level	*N*	MSE	CD	Score level	*N*	MSE	CD
[1.0, 1.5)	6,725	0.18 (0.29)	0.17 (0.10)	[1.0, 1.5)	24,297	0.30 (0.66)	0.32 (0.17)
[1.5, 2.0)	14,048	0.30 (0.41)	0.30 (0.01)	[1.5, 2.0)	47,569	0.65 (0.81)	0.73 (0.08)
[2.0, 2.5)	45,244	0.32 (0.43)	0.32 (0.02)	[2.0, 2.5)	97,544	0.87 (0.95)	0.89 (0.02)
[2.5, 3.0)	150,423	0.28 (0.46)	0.29 (0.02)	[2.5, 3.0)	181,903	0.91 (1.14)	0.90 (0.01)
[3.0, 3.5)	303,175	0.30 (0.49)	0.32 (0.04)	[3.0, 3.5)	254,056	0.82 (1.16)	0.84 (0.02)
[3.5, 4.0)	232,939	0.41 (0.52)	0.40 (0.00)	[3.5, 4.0)	187,485	0.72 (1.02)	0.75 (0.03)
[4.0, 4.5)	90,171	0.43 (0.57)	0.39 (0.01)	[4.0, 4.5)	57,000	0.61 (0.87)	0.61 (0.05)
[4.5, 5.0]	11,676	0.36 (0.64)	0.29 (0.06)	[4.5, 5.0]	4,813	0.47 (0.85)	0.39 (0.07)
	Task 3				Task 4		
Score level	*N*	MSE	CD	Score level	*N*	MSE	CD
[1.0, 1.5)	4,770	0.26 (0.36)	0.16 (0.09)	[1.0, 1.5)	5,373	0.25 (0.33)	0.14 (0.10)
[1.5, 2.0)	13,556	0.26 (0.39)	0.22 (0.03)	[1.5, 2.0)	12,935	0.25 (0.39)	0.20 (0.04)
[2.0, 2.5)	44,827	0.30 (0.38)	0.27 (0.05)	[2.0, 2.5)	38,820	0.28 (0.36)	0.24 (0.04)
[2.5, 3.0)	82,121	0.28 (0.41)	0.33 (0.00)	[2.5, 3.0)	84,006	0.24 (0.39)	0.24 (0.03)
[3.0, 3.5)	125,543	0.34 (0.49)	0.37 (0.03)	[3.0, 3.5)	135,085	0.25 (0.39)	0.26 (0.04)
[3.5, 4.0)	123,341	0.41 (0.64)	0.44 (0.00)	[3.5, 4.0)	121,669	0.29 (0.42)	0.33 (0.01)
[4.0, 4.5)	81,693	0.48 (0.76)	0.46 (0.01)	[4.0, 4.5)	80,190	0.34 (0.51)	0.35 (0.03)
[4.5, 5.0)	27,819	0.56 (0.88)	0.47 (0.01)	[4.5, 5.0)	28,572	0.42 (0.60)	0.40 (0.01)
[5.0, 5.5)	3,810	0.55 (0.86)	0.43 (0.01)	[5.0, 5.5)	5,537	0.43 (0.56)	0.38 (0.01)
[5.5, 6.0]	64	0.46 (1.19)	0.33 (0.03)	[5.5, 6.0]	252	0.31 (0.67)	0.30 (0.04)

Note. Score levels are based on the machine scores. MSE = Mean squared error between human and machine scores produced by the cumulative logistic regression model; *SD* = standard deviation; CD = conditional dispersion. CD is computed as $\sum_{H=1}^{6}[H - E(H|M)]^2 P(H|M)$, where H is the human score, M is the e-rater machine score resulting from the cumulative logistic regressions model, $E(H|M)$ is the expected human score given M, and $P(H|M)$ is the probability of H given M.

findings reported above. In addition, the table shows that human raters also encountered greater difficulty at the higher score levels for Tasks 1, 3, and 4 and greater difficulty toward the middle of the score range in Task 2.

In the second step of addressing Research Question 2, we investigated whether the size of the human – machine discrepancy for a response could be predicted. Table 7 provides the Pearson correlation coefficient between the predicted and observed absolute human – machine discrepancy. The predictive model was based on the machine scoring features, two additional linguistic features, and several advisory flags not used in the prescreening process. For the overall sample as well as for the highest-volume subgroup populations, the prediction accuracy was very limited, particularly for Task 2.

Discussion

The present investigation evaluated the utility of approaches to detecting unusual responses in the automated essay scoring context. Unusual responses were defined as those not appropriate for machine scoring because the responses have characteristics that the scoring system cannot handle. Successful detection of such atypical responses is critical for employing automated scoring as the primary grading method.

Two research questions were addressed. One question concerned the performance of prescreening advisory flags. The other question related to the degree of machine scoring difficulty and whether the size of the human – machine discrepancy could be predicted. The ability to predict such discrepancies can be thought of as a precursor to developing a general post hoc screening method.

For the first research question, we studied the performance of eight prescreening advisory flags in the e-rater automated scoring system. The results suggested that some advisory flags operated far more consistently across measures and tasks in detecting responses that the machine was likely to score differently from human raters than did other flags. For example, Advisory Flag #8 (excessive number of problems) functioned as expected on all four measures (Cohen's *d*, QWK, SMD, and *r*) on every task, and #5 (too short) did the same with the exception of Cohen's *d* on Task 4. In addition, for all tasks on which responses were flagged by Advisory Flag #6 (too long), this advisory functioned as predicted except for

Table 6 Mean Squared Error by Score Level for the Double Human-Scored Sample

	Task 1					Task 2			
Score level	*N*	HM MSE (*SD*)	HH MSE (*SD*)	Diff	Score level	*N*	HM MSE (*SD*)	HH MSE (*SD*)	Diff
[1.0, 1.5)	329	0.18 (0.26)	0.13 (0.44)	0.04	[1.0, 1.5)	1,149	0.30 (0.67)	0.10 (0.47)	0.20
[1.5, 2.0)	648	0.31 (0.40)	0.19 (0.46)	0.11	[1.5, 2.0)	2,277	0.66 (0.87)	0.16 (0.58)	0.50
[2.0, 2.5)	2,192	0.32 (0.43)	0.23 (0.63)	0.09	[2.0, 2.5)	4,669	0.87 (0.91)	0.20 (0.69)	0.67
[2.5, 3.0)	7,270	0.28 (0.44)	0.20 (0.61)	0.08	[2.5, 3.0)	8,553	0.89 (1.14)	0.23 (0.75)	0.66
[3.0, 3.5)	14,395	0.31 (0.50)	0.22 (0.70)	0.09	[3.0, 3.5)	11,846	0.82 (1.16)	0.25 (0.80)	0.57
[3.5, 4.0)	11,115	0.40 (0.53)	0.28 (0.78)	0.13	[3.5, 4.0)	8,873	0.70 (0.98)	0.26 (0.81)	0.44
[4.0, 4.5)	4,338	0.43 (0.56)	0.27 (0.80)	0.16	[4.0, 4.5)	2,848	0.60 (0.88)	0.25 (0.82)	0.35
[4.5, 5.0]	564	0.32 (0.54)	0.21 (0.71)	0.11	[4.5, 5.0]	211	0.45 (0.82)	0.22 (0.84)	0.23
	Task 3					Task 4			
Score level	*N*	HM MSE (*SD*)	HH MSE (*SD*)	Diff	Score level	*N*	HM MSE (*SD*)	HH MSE (*SD*)	Diff
[1.0, 1.5)	173	0.22 (0.26)	0.19 (0.61)	0.03	[1.0, 1.5)	147	0.26 (0.28)	0.16 (0.47)	0.10
[1.5, 2.0)	495	0.25 (0.41)	0.15 (0.51)	0.10	[1.5, 2.0)	422	0.22 (0.38)	0.15 (0.59)	0.07
[2.0, 2.5)	1,694	0.30 (0.37)	0.17 (0.52)	0.13	[2.0, 2.5)	1,510	0.29 (0.37)	0.21 (0.54)	0.08
[2.5, 3.0)	3,198	0.28 (0.41)	0.16 (0.52)	0.12	[2.5, 3.0)	3,140	0.23 (0.40)	0.16 (0.55)	0.07
[3.0, 3.5)	5,046	0.34 (0.50)	0.19 (0.60)	0.16	[3.0, 3.5)	5,244	0.25 (0.39)	0.17 (0.52)	0.08
[3.5, 4.0)	5,036	0.41 (0.66)	0.22 (0.74)	0.19	[3.5, 4.0)	4,921	0.29 (0.41)	0.18 (0.61)	0.11
[4.0, 4.5)	3,356	0.48 (0.77)	0.28 (0.85)	0.21	[4.0, 4.5)	3,311	0.35 (0.52)	0.24 (0.73)	0.11
[4.5, 5.0)	1,158	0.56 (0.89)	0.34 (1.16)	0.22	[4.5, 5.0)	1,193	0.44 (0.60)	0.32 (0.92)	0.12
[5.0, 5.5)	196	0.57 (0.95)	0.30 (0.89)	0.27	[5.0, 5.5)	255	0.44 (0.58)	0.27 (0.79)	0.17
[5.5, 6.0]	3	2.38 (3.81)	1 (1.73)	[a]	[5.5, 6.0]	10	0.22 (0.09)	0.40 (0.42)	[a]

Note. HM = human – machine; HH = human – human; MSE = mean squared error between the two human ratings, and between human and machine scores produced by the cumulative logistic regression model (score levels are based on the machine scores); *SD* = standard deviation; Diff = human – machine MSE minus human – human MSE.

[a]This statistic is not estimable due to sample size.

Task 3 on Cohen's *d*. That is, for these advisories, the flagged responses had noticeably greater absolute machine – human discrepancies on the individual level, greater SMDs on the distributional level, and lower machine – human agreement than nonflagged responses.

Two other advisories showed sensitivity to particular types of machine – human disagreement but not to others. Advisory Flag #2 (insufficient development) was sensitive to machine – human discrepancies on the individual and distributional level (but not to level differences as measured by *r*). Similarly, Advisory Flag #1 (repetition) was sensitive only to distributional discrepancies between machine and human scores.

The remaining advisories (i.e., #3: off topic; #4: restatement of prompt text; and #7: unusual organization) showed no consistent pattern across tasks or indices.

For the second research question, we evaluated the extent to which the machine had difficulty scoring responses and whether the size of the human – machine discrepancy could be predicted accurately enough to support a post hoc screening mechanism. Results showed relatively little scoring difficulty for three of the four tasks, with the relationship between machine and human scores being reasonably strong. For Task 2, the machine had scoring difficulty as indicated by low R^2 and large MSE between human and machine scores. This result might be explained by the fact that this task is substantively different from the other three tasks. In contrast to those tasks, Task 2 requires the examinee to read a text, listen to an audio segment, and write an essay integrating the two sources. The task's heavier dependence on content may make it less amenable to the type of automated scoring approach used here, which draws primarily on features targeting writing fundamentals.

Despite the fact that overall scoring difficulty appeared to be relatively small for three of the four tasks, more difficulty was apparent for the upper than for the other levels of the score scale. This phenomenon was evidenced by a moderate negative correlation of the standard deviation of the score-level probabilities with human scores and by larger MSEs for responses in the upper levels. Interestingly, human raters also showed evidence of greater scoring difficulty at the upper end

Table 7 Correlation Coefficient Between the Predicted and Observed Absolute Human–Machine Discrepancy

	Task 1			Task 2	
Total/subgroup	*N*	*r*	Total/subgroup	*N*	*r*
Total sample	854,401	0.13	Total sample	854,667	0.04
China	263,940	0.14	China	264,870	0.04
USA	122,522	0.12	USA	121,046	0.05
Korea	66,594	0.15	Korea	67,761	0.03
India	57,948	0.08	India	57,544	0.02
Japan	45,280	0.11	Japan	45,739	0.07
	Task 2			Task 4	
Total/subgroup	*N*	*r*	Total/subgroup	*N*	*r*
Total sample	507,544	0.18	Total sample	512,439	0.18
USA	323,008	0.18	USA	325,125	0.18
India	76,511	0.17	India	77,870	0.14
China	47,319	0.17	China	48,089	0.19
Korea	6,125	0.16	Korea	6,252	0.14
Canada	5,441	0.20	Canada	5,560	0.21

Note. r = Pearson correlation coefficient between human and machine scores produced by the cumulative logistic regression model.

of the scale, as indicated by MSE. In fact, a sizable portion of the machine–human MSE might be caused by unreliability among human raters, which in turn may reflect ambiguity in the rubric criteria.

Our attempt to predict human–machine disagreement met with limited success. This outcome was not surprising given the scoring difficulty results. For one, machine scores seemed to correlate highly with human scores overall, leaving relatively little variation in the size of the disagreements. In addition, the small number of examinees who were in the lowest and highest score levels further limited the variation in discrepancy. Last, the level of disagreement did not seem to be similar across score levels, making it harder to predict.

Several limitations of this study should be noted. First, only one human score was employed for most analyses. Because human raters (like machines) are imperfect, additional human raters should make for a more reliable and valid criterion against which to evaluate the flags, scoring difficulty, and the discrepancy-prediction model. Second, we studied advisory flags related to only eight categories of unusual response. Other kinds of unusual response (e.g., responses with many rare or long words) were not investigated because the system employed did not have flags to identify them. Third, some advisories were not triggered by any responses in some tasks, and as a result, the effect of these flags could not be evaluated for those tasks. While it is possible that the test takers were not prone to exhibit those kinds of behaviors, it is also plausible that those advisories were not sensitive to detecting those behaviors in those tasks. Fourth, the effectiveness of the advisory flags in detecting the types of responses that the flag names describe was not evaluated (e.g., that responses flagged as repetitious have that characteristic and that nonflagged responses do not). To the extent that a flag does not accurately identify the types of responses it claims to detect, that flag would not be expected to yield useful information for routing decisions. Thus, one means of improving the advisory flags might be through evaluating this aspect of their detection accuracy. Last, just two assessments and one automated scoring system were evaluated. It is conceivable that differences in examinee population, test content, test purpose, or automated scoring system could produce different results.

This investigation illustrated how the functioning of prescreening flags for detecting unusual responses might be evaluated and the degree to which machine scoring difficulty might be predicted. The methods employed can be used by other investigators in assessing the efficacy of similar flags and the degree of scoring difficulty likely to be found for other systems. That investigation might inform policy judgments about when it is suitable to use specific advisory flags, the criteria for defining atypicality, the subgroup populations, and scale regions where scoring difficulty might be present. Investigators creating automated scoring systems should contemplate including prescreening advisory flags similar to the ones found to function successfully here.

Future research should include qualitative analyses, which might show meaningful patterns in responses that have large human–machine discrepancy. For post hoc screening purposes, those patterns can then be used to model scoring

difficulty. Inclusion of more linguistic features independent of the features used for automated scoring might also increase the prediction of discrepancy. For prescreening purposes, response timing and process data might be useful (Zhang & Deane, 2015). As a simple example, if an examinee submits an essay a short time after the assessment begins (e.g., 30 seconds), that response is unlikely to be legitimate. In any event, given that not all current flags function effectively, that flags don't yet exist to detect some types of unusual responses, and that there remains scoring difficulty for at least one type of task, a stronger collection of prescreening and postscreening methods is likely to be useful in supporting the sole use of automated scoring.

Notes

1 Capabilities of the machine are contingent upon a number of factors including the parser accuracy, the types of text features extracted from the essay, quality of the training data, and the feature aggregation model.
2 Part of the results for Task 4 are also published in Zhang, Chen, and Ruan (2015). Those results are included here by permission of the copyright holder.

References

Attali, Y., & Burstein, J. (2006). Automated essay scoring with e-rater® v. 2. *The Journal of Technology, Learning and Assessment, 4*(3), 2–29.

Bennett, R. E., & Zhang, M. (2016). Validity and automated scoring. In F. Drasgow (Ed.), *Technology and testing: Improving educational and psychological measurement* (pp. 142–173). New York, NY: Taylors & Francis.

Bejar, I. I., VanWinkle, W. H., Madnani, N., Lewis, W., & Steier, M. (2013). *Length of textual response as a construct-irrelevant response strategy: The case of shell language* (Research Report No. RR-13-07). Princeton, NJ: Educational Testing Service. 10.1002/j.2333-8504.2013.tb02314.x

Chen, J., Zhang, M., & Bejar, I. I. (in press). Improving the effectiveness of the pre-screening filtering system of an automated essay scoring engine. *Educational Testing Service Research Report Series.*

Educational Testing Service. (2014a). *Coming together to raise achievement new assessments for the Common Core State Standards*. Retrieved from http://www.k12center.org/rsc/pdf/coming_together_march_2014_rev_1.pdf

Educational Testing Service. (2014b). *Criterion®*. Retrieved from http://www.ets.org/criterion

Educational Testing Service. (2014c). *Understanding your TOEFL iBT® test scores*. Retrieved from http://www.ets.org/toefl/ibt/scores/understand

Foltz, P. W., Laham, D., & Landauer, T. K. (1999). Automated essay scoring: Applications to educational technology. In B. Collis & R. Oliver (Eds.), *Proceedings of world conference on educational multimedia, hypermedia and telecommunications* (pp. 939–944). Chesapeake, VA: Association for the Advancement of Computing in Education.

Haberman, S. J., & Sinharay, S. (2010). The application of the cumulative logistic regression model to automated essay scoring. *Journal of Educational and Behavioral Statistics, 35,* 586–602.

Higgins, D., Burstein, J., & Attali, Y. (2005). Identifying off-topic student essays without topic-specific training data. *Natural Language Engineering, 12,* 145–159.

Higgins, D., & Heilman, M. (2014). Managing what we can measure: Quantifying the susceptibility of automated scoring systems to gaming behavior. *Educational Measurement: Issues and Practice, 33*(4), 36–46.

Karabastos, G. (2003). Comparing the aberrant response detection performance of thirty six person-fit statistics. *Applied Measurement in Education, 16*(4), 277–298.

Page, E. B. (2003). Project Essay Grade: PEG. In M. D. Shermis & J. Burstein (Eds.), *Automated essay scoring: A cross-disciplinary perspective* (pp. 43–54). Mahwah, NJ: Lawrence Erlbaum.

Partnership for Assessment of Readiness for College and Careers. (2010). *The Partnership for Assessment of Readiness for College and Careers (PARCC) application for the Race to the Top comprehensive assessment systems competition*. Retrieved from http://www.parcconline.org/sites/parcc/files/PARCC%20Application%20-%20FINAL.pdf

Pearson Education. (2010). *Intelligent essay assessor™ (IEA) fact sheet*. Retrieved from http://kt.pearsonassessments.com/download/IEA-FactSheet-20100401.pdf

Powers, D., Burstein, J., Chodorow, M., Fowles, M., & Kulich, K. (2001). *Stumping e-rater: Challenging the validity of automated essay scoring* (Research Report No. RR-01-03). Princeton, NJ: Educational Testing Service. 10.1002/j.2333-8504.2001.tb01845.x

Reise, S. P., & Due, A. M. (1991). The influence of test characteristics on the detection of aberrant response patterns. *Applied Psychological Measurement, 15,* 217–226.

Rupp, A. A. (2013). A systematic review of the methodology for person fit research in item response theory: Lessons about generalizability of inferences from the design of simulation studies. *Psychological Test and Assessment Modeling, 55,* 3 – 38.

SMARTER Balanced Assessment Consortium. (2010). *Race to the Top assessment program application for new grants: Comprehensive assessment systems CFDA Number: 84.395B*. Retrieved from http://www.smarterbalanced.org/wordpress/wpcontent/uploads/2011/12/Smarter-Balanced-RttTApplication.pdf

Vantage Learning. (2012). *IntelliMetric®*. Retrieved from http://www.vantagelearning.com/products/intellimetric

Zhang, M. (2013). *Contrasting automated and human scoring of essays* (RDC-21). Princeton, NJ: Educational Testing Service.

Zhang, M., Breyer, F. J., & Lorenz, F. (2013). *Investigating the suitability of implementing the e-rater scoring engine in a large-scale English language testing program* (Research Report No. RR-13-36). Princeton, NJ: Educational Testing Service. 10.1002/j.2333-8504.2013.tb02343.x

Zhang, M., Chen, J., & Ruan, C. (2015). Evaluating the detection of aberrant responses in automated essay scoring. In L. A. van der Ark, W.-C. Wang, S.-M. Chow, D. M. Bolt, & J. A. Douglas (Eds.), *Quantitative psychology research* (pp. 191 – 208). Cham, Switzerland: Springer.

Zhang, M., & Deane, P. (2015). *Process features in writing: Internal structure and incremental value over product features* (Research Report No. RR-15-27). Princeton, NJ: Educational Testing Service. 10.1002/ets2.12075

Suggested citation:

Zhang, M., Chen, J., & Ruan, C. (2016). *Evaluating the advisory flags and machine scoring difficulty in automated essay scoring by* e-rater® *automated scoring engine* (Research Report No. RR-16-30). Princeton, NJ: Educational Testing Service. http://dx.doi.org/10.1002/ets2.12116

Action Editor: Beata Beigman Klebanov

Reviewers: Aoife Cahill and Shelby Haberman