ETS GRE

Assess Ability. Predict Performance.

ETS GRE® Board Research Report

ETS GRE®–18-01

ETS RR–18-12

Understanding Mean Score Differences Between the *e-rater*® Automated Scoring Engine and Humans for Demographically Based Groups in the *GRE*® General Test

Chaitanya Ramineni

David Williamson

December 2018

The report presents the findings of a research project funded by and carried out under the auspices of the Graduate Record Examinations Board.

****

Researchers are encouraged to express freely their professional judgment. Therefore, points of view or opinions stated in Graduate Record Examinations Board reports do not necessarily represent official Graduate Record Examinations Board position or policy.

The Graduate Record Examinations and ETS are dedicated to the principle of equal opportunity, and their programs, services, and employment policies are guided by that principle.

****

As part of its educational and social mission and in fulfilling the organization's non-profit Charter and Bylaws, ETS has and continues to learn from and also to lead research that furthers educational and measurement research to advance quality and equity in education and assessment for all users of the organization's products and services.

GRE-ETS

PO Box 6000

Princeton, NJ 08541-6000

USA

To obtain more information about GRE
programs and services, use one of the following:
Phone: 1-866-473-4373
(U.S., U.S. Territories*, and Canada)
1-609-771-7670
(all other locations)
Web site: www.gre.org

*America Samoa, Guam, Puerto Rico, and US Virgin Islands

RESEARCH REPORT

Understanding Mean Score Differences Between the *e-rater*® Automated Scoring Engine and Humans for Demographically Based Groups in the *GRE*® General Test

Chaitanya Ramineni & David Williamson

Educational Testing Service, Princeton, NJ

Notable mean score differences for the *e-rater*® automated scoring engine and for humans for essays from certain demographic groups were observed for the *GRE*® General Test in use before the major revision of 2012, called rGRE. The use of e-rater as a check-score model with discrepancy thresholds prevented an adverse impact on the examinee score at the item or test level. Despite this control, there remains a need to understand the root causes of these demographically based score differences and to identify potential mechanisms for avoiding future instances of discrepancy. In this study, we used a combination of statistical methods and human review to propose hypotheses about the root cause of score differences and whether such discrepancies reflect inadequacies of e-rater, human scoring, or both. The human rating process was found to be influenced strongly by the scale structure and did not fully correspond to the e-rater scoring mechanism. The human raters appeared to be using conditional logic and a rule-based approach to their scoring, while e-rater uses linear weighting of all the features. These analyses have implications for future research and operational policies for the scoring of the rGRE.

Keywords Automated scoring; essay scoring; *GRE*® writing; subgroup differences; shell text; CART

doi:10.1002/ets2.12192

The *e-rater*® automated scoring engine (Attali & Burstein, 2006), developed at Educational Testing Service (ETS), is used for scoring responses to the constructed-response items in various assessments used for making high-stakes decisions, such as the *GRE*® General Test[1] (Ramineni, Trapani, Williamson, Davey, & Bridgeman, 2012a) that was in use until it was replaced by the revised GRE in 2012 and the *TOEFL*® test (Ramineni, Trapani, Williamson, Davey, & Bridgeman, 2012b) and in low-stakes practice environments such as the *CRITERION*® online writing evaluation service (Ramineni, 2012), the *TOEFL PRACTICE ONLINE TPO*® practice tests, and the *SAT*® online test. The e-rater is a computer program that scores essays primarily on the basis of features that are related to writing quality. The program uses natural language processing (NLP) technology, a branch of artificial intelligence, to evaluate a number of aspects of writing proficiency that a computer can identify, tabulate, and aggregate to act as substitute measures for human scores, for example, grammar, usage, mechanics, and development. This set of features is constantly refined and enhanced, and the e-rater scoring system is upgraded annually. The current e-rater scoring system uses 11 features, with nine representing aspects of writing quality (grammar, usage, mechanics, style, organization, development, word choice, average word length, and good preposition and collocation usage) and two representing content or use of prompt-specific vocabulary. Most of these primary scoring features are composed of a set of subfeatures computed from NLP techniques, and many of these have multiple layers of microfeatures that have cascaded up to produce the subfeature values. An illustration of the construct decomposition of e-rater resulting from this structure is provided in Figure 1, where the features encapsulated in bold are the independent variables in the regression and the other features are an incomplete illustrative listing of subfeatures measuring aspects of writing quality.

Developing e-rater scoring models is typically a two-stage process: model training/building and model evaluation. Human scores are used as the criterion variable for training and evaluating the e-rater scoring models. The quality of the e-rater models and the effective functioning of the models in an operational environment depend in part on the nature and quality of the training and evaluation data. Therefore ETS uses certain guidelines to guide the model building and evaluation of automated scoring models (Williamson, Xi, & Breyer, 2012). Data are split into a model-building set and

Corresponding author: C. Ramineni, E-mail: chaitanya.ramineni@gmail.com

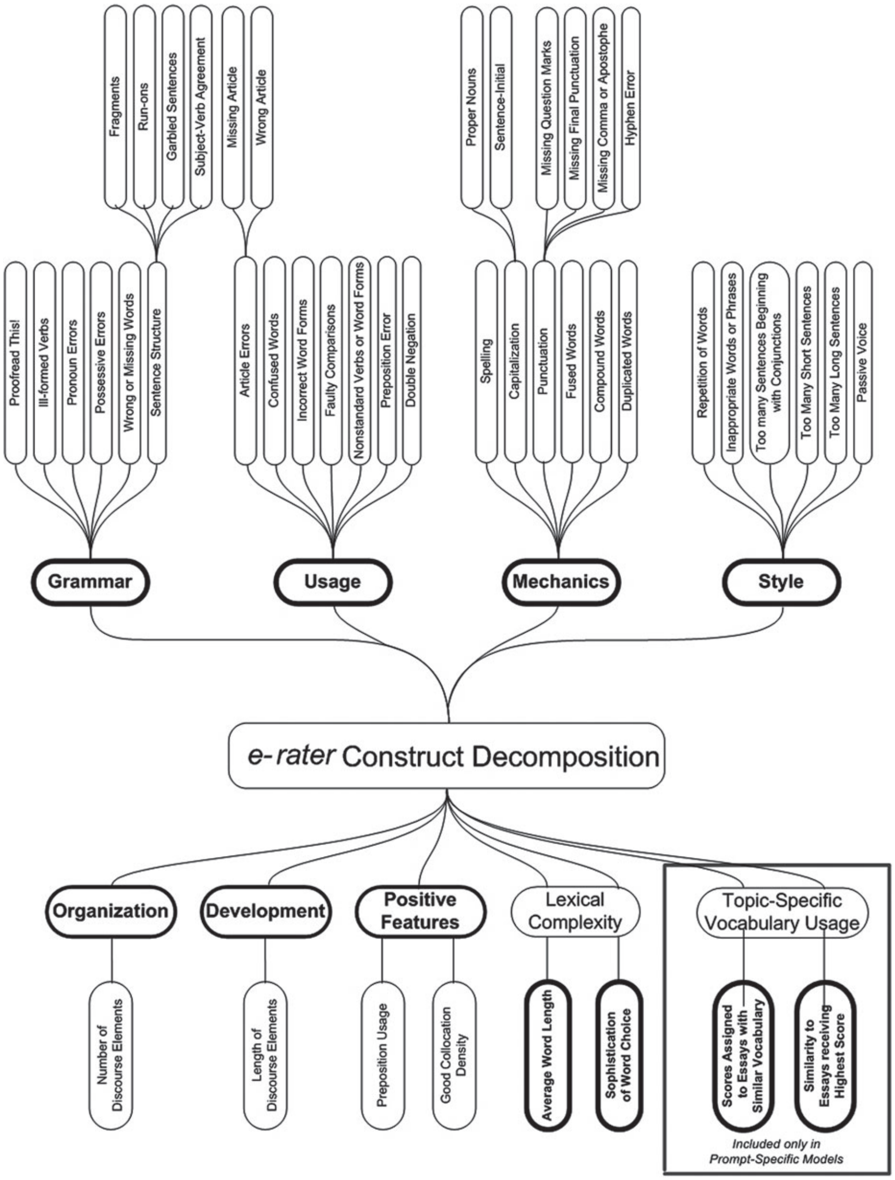

Figure 1 Organization and construct coverage of e-rater Version 10.1. Features in bold are independent variables in the regression; other features are an incomplete illustrative listing of subfeatures measuring aspects of writing quality. Adapted from "Evaluating the Construct Coverage of the e-rater Scoring Engine" (Research Report No. RR-09-01), by T. Quinlan, D. Higgins, and S. Wolff, 2009, Princeton, NJ: ETS, p. 9. Copyright 2009 by Educational Testing Service. Reprinted with permission.

an evaluation set. Training/building of an e-rater model is a fully automated process, given a properly constituted set of training essays in the model-building set. Such a set of training essays consists of a random sample of responses that must have been entered on the computer and should be representative of the population for which e-rater is intended for use. An advisory flag analysis is conducted on the data prior to model build, which serves as a filtering mechanism to remove essays that are inappropriate for automated scoring. Each advisory flag indicates a different kind of problem, such as extensive repetition of words, a topic that is not relevant to the assigned topic, or an essay that is too brief. This advisory flagging improves the quality of the model build by filtering the inappropriate essays from the model build process.

The e-rater program then evaluates characteristics of the essays, such as grammar, usage, mechanics, and development, in the model build set. After these feature values are computed, the weights for the features are determined using a multiple regression procedure with human scores for the essays as the criterion variable. These feature weights can then be applied to additional essays to produce a predicted score based on the calibrated feature weights.

Table 1 Flagging Criteria and Conditions for e-rater Evaluation

Flagging criterion	Flagging condition
Quadratic-weighted Kappa between e-rater score and human score	Quadratic-weighted kappa less than .70
Pearson correlation between e-rater score and human score	Correlation less than .70
Standardized difference between e-rater score and human score	Standardized difference greater than .15 in absolute value
Notable reduction in quadratic-weighted Kappa or correlation from human – human to automated – human	Decline in quadratic-weighted kappa or correlation of greater than .10
Standardized difference between e-rater score and human score within a subgroup of concern	Standardized difference greater than .10 in absolute value

Note. All the threshold values are evaluated to four decimal values for flagging.

The evaluation of e-rater quality is based on the performance of e-rater on an evaluation set of essays that is different from the model build set. Typical e-rater evaluation consists of the following steps: evaluation of the alignment between the construct targeted by the assessment task and the e-rater scoring capability; measures of association with human scores (weighted kappa and Pearson correlations, degradation statistics for change in human – human to human – e-rater agreement, and standardized mean score differences between human and e-rater); association with external variables (such as other test section scores); analysis of subgroup differences (flagged by standardized mean score differences greater than .10 between human and e-rater); and operational impact analysis under varying agreement threshold levels for human and e-rater scores. Table 1 displays a summary of the flagging criteria and conditions for evaluating e-rater model performance. The numbers/statistics that fail to meet the threshold values for an evaluation metric are highlighted in the tables reporting the evaluation results.

The regression-based procedure of e-rater model building lends itself to multiple methods of model construction. Prompt-specific scoring models are custom-built models for each prompt in the item pool. They provide the best-fit models for the particular prompt in question, as both the regression weights and the intercept are customized for the human score distribution used to calibrate the prompt model. Prompt-specific models incorporate prompt-specific vocabulary-related content features into the scoring. An alternate approach, referred to as a generic model, is based on calibration on a group of related prompts, typically 10 or more, and calibrating a regression model across all such prompts so that the resultant model is the best fit for predicting human scores for all the prompts, taken as a whole. As such, a common set of feature weights and a single intercept are used for all prompts regardless of the particular prompt in the set. Generic models do not take into account the content of the essay, because content features related to vocabulary usage are prompt specific, so that generic models address only writing quality. The generic modeling approach has the advantage of requiring smaller sample sizes per prompt (with enough prompts) and a truly consistent set of scoring criteria regardless of the prompt delivered operationally. The generic with prompt-specific intercept model is a variant of the generic model and offers a common set of weights for all features, with a customized intercept for each prompt.

e-rater Scoring for GRE Writing

The old GRE Writing section consisted of two writing tasks: The issue prompts required examinees to provide perspective on a given issue, while the argument prompts required examinees to critique the logic of an argument. The 6-point scoring rubric common to the two tasks is included in Table 2. The lower half of the rubric focuses mainly on language control and language errors; moving up the scale, the emphasis shifts to development of ideas, original contribution, content, and cohesion.

In the e-rater evaluation for GRE in 2007 (Ramineni et al., 2012a), approximately 750,000 operational responses across 113 issue prompts and 139 argument prompts were sampled, representing all responses between September 2006 and September 2007. Each essay included two human rater scores, plus additional human scores if the two were discrepant and required adjudication, as well as several demographic variables (gender, ethnicity, test center country, undergraduate overall and major grade point average, English as best language) and other GRE section test scores. Evaluation of performance for e-rater Version 7.2 resulted in the deployment of generic e-rater models with prompt-specific intercepts for the issue prompt and prompt-specific scoring models for the argument prompt. These were implemented as a *check-score*, in

Table 2 GRE Scoring Guide for Human Raters

Score	GRE scoring guide (issue)	GRE scoring guide (argument)
6	A 6 paper presents a cogent, well-articulated analysis of the complexities of the issue and conveys meaning skillfully. A typical paper in this category ● presents an insightful position on the issue; ● develops the position with compelling reasons and/or persuasive examples; ● sustains a well-focused, well-organized analysis, connecting ideas logically; ● expresses ideas fluently and precisely, using effective vocabulary and sentence variety; and ● demonstrates facility with the conventions (i.e., grammar, usage, and mechanics) of standard written English but may have minor errors.	A 6 paper presents a cogent, well-articulated critique of the argument and conveys meaning skillfully. A typical paper in this category ● clearly identifies important features of the argument and analyzes them insightfully; ● develops ideas cogently, organizes them logically, and connects them with clear transitions; ● effectively supports the main points of the critique; ● demonstrates control of language, including appropriate word choice and sentence variety; and ● demonstrates facility with the conventions (i.e., grammar, usage, and mechanics) of standard written English but may have minor errors.
5	A 5 paper presents a generally thoughtful, well-developed analysis of the complexities of the issue and conveys meaning clearly. A typical paper in this category ● presents a well-considered position on the issue; ● develops the position with logically sound reasons and/or well-chosen examples; ● maintains focus and is generally well organized, connecting ideas appropriately; ● expresses ideas clearly and well, using appropriate vocabulary and sentence variety; and ● demonstrates facility with the conventions of standard written English but may have minor errors.	A 5 paper presents a generally thoughtful, well-developed critique of the argument and conveys meaning clearly. A typical paper in this category ● clearly identifies important features of the argument and analyzes them in a generally perceptive way; ● develops ideas clearly, organizes them logically, and connects them with appropriate transitions; ● sensibly supports the main points of the critique; ● demonstrates control of language, including appropriate word choice and sentence variety; ● demonstrates facility with the conventions of standard written English but may have minor errors.
4	A 4 paper presents a competent analysis of the issue and conveys meaning adequately. A typical paper in this category ● presents a clear position on the issue; ● develops the position on the issue with relevant reasons and/or examples; ● is adequately focused and organized; ● expresses ideas with reasonable clarity; and ● generally demonstrates control of the conventions of standard written English but may have some errors	A 4 paper presents a competent critique of the argument and conveys meaning adequately. A typical paper in this category ● identifies and analyzes important features of the argument; ● develops and organizes ideas satisfactorily but may not connect them with transitions; ● supports the main points of the critique; ● demonstrates sufficient control of language to express ideas with reasonable clarity; and ● generally demonstrates control of the conventions of standard written English but may have some errors
3	A 3 paper demonstrates some competence in its analysis of the issue and in conveying meaning but is obviously flawed. A typical paper in this category exhibits one or more of the following characteristics: ● is vague or limited in presenting or developing a position on the issue; ● is weak in the use of relevant reasons or examples; ● is poorly focused and/or poorly organized ● presents problems in language and sentence structure that result in a lack of clarity; and/or ● contains occasional major errors or frequent minor errors in grammar, usage, or mechanics that can interfere with meaning.	A 3 paper demonstrates some competence in its critique of the argument and in conveying meaning but is obviously flawed. A typical paper in this category exhibits one or more of the following characteristics: ● does not identify or analyze most of the important features of the argument, although some analysis of the argument is present; ● mainly analyzes tangential or irrelevant matters, or reasons poorly; ● is limited in the logical development and organization of ideas; ● offers support of little relevance and value for points of the critique; ● lacks clarity in expressing ideas; and/or ● contains occasional major errors or frequent minor errors in grammar, usage, or mechanics that can interfere with meaning.

Table 2 Continued

Score	GRE scoring guide (issue)	GRE scoring guide (argument)
2	A 2 paper demonstrates serious weaknesses in analytical writing. A typical paper in this category exhibits one or more of the following characteristics: ● is unclear or seriously limited in presenting or developing a position on the issue; ● provides few, if any, relevant reasons or examples; ● is unfocused and/or disorganized; ● presents serious problems in the use of language and sentence structure that frequently interfere with meaning; and/or ● contains serious errors in grammar, usage, or mechanics that frequently obscure meaning.	A 2 paper demonstrates serious weaknesses in analytical writing. A typical paper in this category exhibits one or more of the following characteristics: ● does not present a critique based on logical analysis but may instead present the writer's own views on the subject; ● does not develop ideas or is disorganized and illogical; ● provides little, if any, relevant or reasonable support; ● has serious problems in the use of language and in sentence structure that frequently interfere with meaning; and/or ● contains serious errors in grammar, usage, or mechanics that frequently obscure meaning.
1	A 1 paper demonstrates fundamental deficiencies in analytical writing. A typical paper in this category exhibits one or more of the following characteristics: ● provides little or no evidence of the ability to understand and analyze the issue; ● provides little or no evidence of the ability to develop an organized response; ● presents severe problems in language and sentence structure that persistently interfere with meaning; and/or ● contains pervasive errors in grammar, usage, or mechanics that result in incoherence.	A 1 paper demonstrates fundamental deficiencies in analytical writing. A typical paper in this category exhibits one or more of the following characteristics: ● provides little or no evidence of the ability to understand and analyze the argument; ● provides little or no evidence of the ability to develop an organized response; ● has severe problems in language and sentence structure that persistently interfere with meaning; and/or ● contains pervasive errors in grammar, usage, or mechanics that result in incoherence.
0	A 0 paper is demonstrates no understanding of analytical writing. A typical paper in this category exhibits one or more of the following characteristics: ● off-topic (i.e., provides no evidence of an attempt to respond to the assigned topic); ● in a foreign language, merely copies the topic; ● consists of only keystroke characters; and/or ● is illegible or nonverbal.	A 0 paper is demonstrates no understanding of analytical writing. A typical paper in this category exhibits one or more of the following characteristics: ● off-topic (i.e., provides no evidence of an attempt to respond to the assigned topic); ● in a foreign language, merely copies the topic; ● consists of only keystroke characters; and/or ● is illegible or nonverbal.
Not scored	Blank	Blank

which the automated score was used to "check" the human score and, if they were in agreement, the human score was used as the final essay score. If the two scores were discrepant, additional human raters were used to determine the score for the essay.

While e-rater is trained to maximize the prediction of human scores, and most performance guidelines were met for most e-rater models, notable differences between machine and human scores have been observed for certain demographic or language subgroups (Attali, 2008; Attali, Bridgeman, & Trapani, 2007; Bridgeman, Trapani, & Attali, 2012; Burstein & Chodorow, 1999; Chodorow & Burstein, 2004). Burstein and Chodorow (1999) and Chodorow and Burstein (2004) found differences related to language groups in the evaluation of essays from the *TWE*® test using earlier versions of e-rater. The Bridgeman et al. (2012) and Attali et al. (2007) studies reported differences related to test center country and ethnicity for the GRE general test. Similar differences were observed during more recent e-rater evaluations for the GRE general test (Ramineni et al., 2012a). Notable differences were observed between e-rater and human scores (standardized mean score differences > .10) for African American examinees and for test takers from China. E-rater scores were lower than human scores for the African American subgroup for argument tasks and higher than human scores for the subgroup of China for both issue and argument tasks. Such differences, if systematic in nature, may have potential impacts on fairness of scores for examinees. The investigations by Attali et al. (2007) and Bridgeman et al. (2012) revealed differences in the lengths of the essays for different demographically based groups and suggested that e-rater may be placing a greater value on long essays, resulting in the differences between e-rater and human scores for test takers from China. During operational scoring procedures, human scoring experts have also identified the use of shell text (memorized well-formed text) to inflate essay length without necessarily advancing the claim or evidence as a prominent test-taking strategy in the China subgroup.

Motivation for the Study

Although the use of both human and automated scores, with adjudication of discrepant scores, controls the impact of any systematic differences in the operational use of e-rater for assessments used in making high-stakes decisions, there remains a need to understand the root causes of these differences. The initial investigations by Attali et al. (2007) and Bridgeman et al. (2012) suggested essay length as a potential source. Alternatively, the human raters may be evaluating features not captured by e-rater. Also, there may be additional test-taking strategies, such as use of shell text, contributing to the sources of differences between human and e-rater scores. Discrepancies between e-rater and human scores do not necessarily indicate e-rater bias, as there is not sufficient evidence that the human score used as the gold standard is necessarily a better indicator of writing ability than the e-rater score (Bennett & Bejar, 1997). Research studies have identified problems and concerns with human scoring of essays that represent a range of potential pitfalls, including halo effects, fatigue, tendency to overlook details, and problems with consistency of scoring across time (Braun, 1988; Daly & Dickson-Markman, 1982; Hales & Tokar, 1975; Hughes & Keeling, 1984; Hughes, Keeling, & Tuck, 1980a, 1980b, 1983; Lunz, Wright, & Linacre, 1990; Spear, 1997; Stalnaker, 1936). A better understanding of these differences may provide insight into whether such differences are a potential failing of e-rater scoring models, of human scoring practices, or both to some degree. Therefore the current study targeted three research questions:

1. Are there characteristics of writing (e-rater features) consistently associated with these demographically based differences?
2. Are these differences an artifact of the modeling procedures for e-rater?
3. Can we gain insight into the consistency and root causes of differences through expert review of discrepant cases and/or other empirical methods?

Methods

The findings reported by previous studies on subgroup differences for GRE (Attali et al., 2007; Bridgeman et al., 2012) used e-rater scores produced from Version 2 (Attali & Burstein, 2006) and data from 2006 to 2007. For this study, we used data for the GRE General Test from 2009 to 2010 and e-rater scores produced by Version 10.1. The new versions of e-rater include new features or subfeatures and/or computationally based or statistical modifications of existing features or subfeatures. The major changes for e-rater Version 10.1 included the introduction of a new positive feature—a combined measure of good preposition and collocation use (Ramineni, Davey, & Weng, 2010)—and the content features were revised to include information for all score points in computing the two measures (Attali, 2009). The e-rater Version 10.1 therefore used 11 score features (10 features in earlier versions), with nine representing aspects of writing quality and two representing content. Also, whereas the earlier studies focused on empirical investigation of the differences, we used a combination of empirical methods and qualitative review of GRE data to develop and test hypotheses about the root cause of score discrepancies between e-rater and human raters.

Data

The data used for this study were sampled from examinee responses to the GRE issue and argument prompts from July 2009 to February 2010, spanning 101 issue prompts and 114 argument prompts. One thousand responses were randomly sampled per prompt, resulting in a total of 215,000 responses, and operational human and e-rater Version 10.1 scores for these responses were used for the study. Because the e-rater models were implemented for both issue and argument writing tasks in 2007, thus eliminating the need for double human scores on all data, only limited data (roughly 40%, discrepant cases by default) had double scores available, and these were only for cases in which the original human and e-rater scores were discrepant by more than half a point on the rating scale. Therefore, a single human score was used as the criterion variable in place of the average human score that served as the criterion variable for models built in 2007.

Procedures

The scoring models originally approved and implemented in 2007 were rebuilt using more recent data and under e-rater Version 10.1 to account for the changing examinee population as observed under the routine operational work for the assessment. These models were built using the single human score available as the criterion variable. It should be noted that the availability of single human scores only and the absence of a reliability sample in these more recent data prevented direct evaluation of the quality of human scores used for building the models. The new models were also evaluated at the subgroup level to identify the subgroups of concern (mean score differences between e-rater and humans beyond an acceptable threshold level) for both issue and argument writing tasks based on the e-rater Version 10.1 evaluations of the new data.

For the first research question (Are there characteristics of writing [e-rater features] consistently associated with these demographically based differences?), as in Bridgeman et al. (2012), we analyzed the e-rater feature scores for each of these subgroups and writing tasks to examine the differences in writing styles and characteristics for each subgroup. In addition, we also examined the weights associated with the features across models based on different groups to further understand the differences between the subgroups' writing styles. All the data (215,000 responses) were used for these analyses.

For the second research question (Are these differences an artifact of the modeling procedures for e-rater?), we evaluated e-rater model performance at the overall and the subgroup levels using (a) logistic regression, (b) unit-weighted modeling, and (c) classification and regression tree (CART) modeling. The choice of each model was informed by prior research and was based on the supporting rationale for its potential feasibility for the e-rater scoring engine, as described briefly in the text that follows. Again, all the data (215,000 responses) were used for these analyses.

Logistic Regression

The feature scores are included in a linear regression model to predict the criterion variable, that is, the human essay score. The linear regression method produces e-rater scores on a continuous score scale, subsequently requiring multiple truncations and rounding procedures to arrive at an integer score on a 6-point scale for GRE writing tasks. A cumulative logit model that assumes that the response is categorical has therefore been suggested previously as more appropriate for estimating the categorical human scores (Feng, Dorans, Patsula, & Kaplan, 2003; Haberman, 2007). These studies were designed to study individual prompts and compared the results for the cumulative logit model to the then operational e-rater models, showing some advantage to using the cumulative logit model over the linear regression model. More recently, the logit models were empirically investigated in greater depth and compared to the linear regression model for larger groups of prompts under different e-rater model choices, that is, generic, generic with prompt-specific intercepts, and prompt-specific models (Haberman & Sinharay, 2010). The study found that the average mean squared error was substantially lower for the cumulative logit model compared to the linear regression model and recommended that it is preferable to replace the ordinary regression analysis with a cumulative logit model for automated essay scoring. A study of the cumulative logit model for TOEFL prompts showed that the use of a logistic regression model did not mitigate the mean score differences between human and e-rater scores for the native language subgroups (Sinharay, Davey, Weng, & Ramineni, 2009). However, an evaluation of e-rater performance under a cumulative logit regression model versus linear regression model using current operational procedures and performance criteria has not yet been conducted. In addition, whether replacing the linear regression model with the logistic regression model would eliminate or mitigate the human and e-rater score differences for the subgroups for GRE has also not been fully investigated. We therefore chose to evaluate the performance of a logistic regression model for e-rater using these data for GRE at both the overall and the subgroup levels.

Unit-Weighted Regression

The distribution of weights across e-rater features can widely differ. According to Attali (2007), the organization and development features are most highly weighted (roughly 30% each) under the linear regression model predicting the human score, and their sum is highly correlated with essay length. Therefore the examinees or subgroups with lower scores on these features can be anticipated to receive lower e-rater scores in general as a result. Conversely, these responses may still be assigned higher scores by human raters for their strength in content and other features and characteristics of writing that received lower weights in e-rater scoring and were therefore subsumed by the two highly weighted features. Similarly, examinees and/or subgroups who write longer essays with poor language control and lack of argumentation or relevant content can potentially receive higher scores from e-rater on account of the two highly weighted features, thus leading to discrepancies, with lower scores more likely to be assigned by human raters.

Previous research has shown that equal-weight linear models can perform as well as optimal models under a set of general conditions (Dawes & Corrigan, 1974; Dorans & Drasgow, 1978; Wainer, 1976). The research has supported the use of equal or unit weights for ease of estimation and robustness. Attali (2007) compared e-rater scores produced by a linear regression model in which optimal weights are derived for variables or features empirically by regressing them on a human score against e-rater scores produced by a linear regression model where, forgoing the prediction of human scores, all features are assigned equal weights. Attali reported that both sets of e-rater scores were comparable in alternate-form reliability and replication of human scores, with equal weights reducing the correlation with essay length. Harik, Baldwin, and Clauser (2013) found that unit-weighted models performed at par with the linear regression model that used empirically derived weights to predict human scores, as well as other methods, when comparing different strategies for scoring the computer-based case simulations currently used to assess physicians' patient-management skills as part of the Step 3 United States Medical Licensing Examination. Unit weighting is advocated strongly in situations where population changes from time to time (Lawshe & Schucker, 1959; Trattner, 1963; Wesman & Bennett, as cited in Dawes & Corrigan, 1974), something that is often seen with both the GRE and the TOEFL test-taking populations. We therefore decided to evaluate the performance of a unit-weighted model for the data in this study, both at the overall and at the subgroup level.

Classification and Regression Trees

CART (Breiman, Friedman, Olshen, & Stone, 1984) has been used previously in the context of automated scoring by Zechner, Higgins, Xi, and Williamson (2009) for building and evaluating scoring models for the *SpeechRater*SM automated scoring service and by Williamson, Bejar, and Sax (2004) as an automated tool to help subject matter experts (SMEs) evaluate the human and machine score discrepancies. As Williamson et al. (2004) noted, CART has been successfully used in prior research on classification problems in psychometrics (e.g., Sheehan, 1997, as cited in Williamson et al., 2004, for proficiency scaling and diagnostic assessment; Bejar, Yepes-Baraya, & Miller, 1997, as cited in Williamson et al., 2004, for modeling rater cognition; Holland, Ponte, Crane, & Malberg, 1998, as cited in Williamson et al., 2004, for computerized adaptive testing). Similar to the linear regression approach, CART begins with a training data set that consists of observations for which both independent and dependent variables are known (referred to as a supervised learning technique). In this context, the e-rater features were the independent variables and the human scores were the dependent variable. The training set serves as the basis for building the model (or the tree), which is then applied to the cross-validation or the evaluation set to predict the scores (the dependent variable) from the independent variables. We evaluated the automated scores produced using scoring models based on CART against the human scores.

We replicated the three different types of regression models (logistic, unit weighted, and CART) following the current operational procedures of (a) splitting data for model build and evaluation and (b) scaling e-rater scores to match human score distribution and evaluated their performance against the linear regression model using the evaluation criteria described in Table 1. We also evaluated the performance of the various models at the subgroup level for the three subgroups of China, Taiwan, and African American—the subgroups identified as a concern from the original evaluation results. Although the Taiwan and African American subgroups were not identified as a concern in the regular evaluation results for issue prompts, we included them in these evaluations to allow us to analyze and compare results for more than just one subgroup, that is, China, across the different regression models. For all the approaches, the data were split into model build and evaluation sets, where the model build set is used to derive the feature weights, which are then applied to the evaluation set to assess e-rater model performance. For the unit-weighted approach, the data were first standardized before splitting, and then the standardized variables or feature scores were summed to form a composite. This composite was regressed on the human score in the model build set to derive a weight for the composite, which was then applied on the evaluation set to predict human scores. For the purpose of this study, the generic with prompt-specific intercept operational model for issue prompts was simplified to a generic model (without requiring the customized intercept for each prompt). For the argument prompts, prompt-specific models were built under linear and logistic models, but a generic model with content features was evaluated under unit-weighted and CART models. These changes were made to capitalize on the efficiency allowed by generic models, as the same set of weights and intercepts is applicable across all prompts, whereas training and evaluating prompt-specific models or generic models with prompt-specific intercepts can be very complex and time consuming. Also, the operational scaling approach used to match e-rater scores to human score distributions was difficult to replicate under the other modeling procedures. As a result, we used an alternate adjustment method based on the Kelly formula and applied it to all raw e-rater scores from linear, logistic, unit-weighted, and CART models prior to evaluating against human scores.

For the third research question (Can we gain insight into the consistency and root causes of differences through expert review of discrepant cases and/or other empirical methods?), from all the data (215,000 responses), we identified a set of 100 maximally discrepant essays across 10 prompts (10 essays per prompt) for each subgroup of concern for either writing task, with the exception of the Taiwan subgroup for the argument task, for which we included all responses (average of 12) for the 10 prompts, which resulted in slightly more than 100 cases for that subgroup. Thus a total of 323 argument responses (100 each for China and African American and 123 for Taiwan) and 100 issue responses (for China) were subjected to rescoring and review by expert human raters. Five scoring leaders were selected by ETS test developers as expert human raters for their extensive operational scoring experience for GRE. As the raters were selected from the current operational rater pool for GRE with extensive rating experience, the raters conducted all rescoring without any formal training/calibration but were monitored by the test developer during the scoring session. The raters were provided a brief description of the overall purpose of the study during recruitment and hence were aware that they would be rescoring essay responses that had received different operational scores from a human rater and e-rater. However, the operational scores by the human rater and e-rater for these discrepant essays were not released to the new raters until the rescoring of all the selected cases was completed. Scoring was completed for 313 of the argument responses (100 for China, 90 for African American, and 123 for Taiwan) and 90 of the 100 issue responses (for China) in under 5 hours. In the expert review, human raters identified language control as a key threshold variable in determining the score of the essay response. The raters unanimously agreed that essays that fail to exhibit good language control automatically fail to qualify for the upper half of the scale. In contrast, e-rater operationally uses linear weighting of all features.

We further used CART on all selected data for the study (215,000 responses) to identify key threshold variables (e-rater feature scores and their values) that will classify responses into low (1–3) and high (4–6) e-rater score categories. We further used CART to identify e-rater features that will classify the responses with human and e-rater score differences equal to 0.5, equal to 1, and greater than 1 into different categories. From the discussions during expert review, we formulated two additional hypotheses — use of shell text (memorized well-formed text) as a test-taking strategy and difference in e-rater and human approaches to scoring of spelling errors (human raters ignore these errors, whereas e-rater is very sensitive to these errors) as potential sources for e-rater and human score differences. We submitted the cases selected for expert review to a shell detection tool (Madnani, Heilman, Tetreault, & Chodorow, 2012) that identifies the organizational elements in the text. These organizational elements are well-formed text that can be memorized and used by test takers to inflate the length of an essay without necessarily advancing the claims and evidence. We also submitted the cases to a spelling correction module (Flor & Futagi, 2012) to lower the number of spelling errors and therefore the spelling error count/penalty by e-rater. The ultimate goal was to make the treatment of shell text and spelling errors by e-rater similar to that of human raters.

On the basis of the preliminary results reported here from the empirical investigations and the qualitative review procedures, we conclude with a discussion of the findings and potential avenues for further research that will be useful in informing program policies and decisions.

Results

Identifying Demographic Subgroups for the Study

New models were built on data from 2009 to 2010 to account for the change in examinee population. Data are split into training and cross-validation sets. The e-rater models are built on a training set with a single human score as the criterion variable and evaluated on a separate cross-validation set. Table 3 reports a summary of the overall agreement between human and e-rater scores for new models on the validation set, as measured by percentage agreement, quadratic-weighted kappa, and Pearson correlations. Because the models were built and evaluated against single human scores, human–human agreement statistics and degradation statistics (change in agreement from human–human to human–e-rater), which are part of default e-rater evaluation criteria, are not presented in these tables. The human–e-rater agreement statistics under new models were very similar to those observed under the old models reported in Ramineni et al. (2012a).

All the evaluation criteria and thresholds were met at the overall score level. However, the analyses at the subgroup level revealed standardized mean score differences between e-rater and humans that exceeded the flagging threshold of .10 for some demographic subgroups. Tables 4 and 5 report the observed differences for subgroups by ethnicity and test center country for issue and argument prompts, respectively. Results for only a subset of the total test center countries that were identified of interest in the original e-rater evaluation for GRE (Ramineni et al., 2012a) are reported. The highlighted cells are cases that violate the flagging conditions and criteria described in Table 1. Other demographic characteristics, such as gender, undergraduate major, and English as best language, were also examined, but no instances were flagged, and so they are not presented here.

On the basis of the results for the subgroup analyses, test takers from China (4%) were flagged for issue prompts, and test takers from China (4%) and Taiwan (<1%) and African American test takers (6%) were flagged for argument prompts (India was a borderline case, and the sample size for Japan was relatively small to support further investigation). These four subgroups (China for issue; China, Taiwan, and African American for argument) were therefore selected for the study. It is interesting to note that for the issue task, the weighted kappa for human and e-rater agreement for the ethnic group of Whites just missed the threshold of .70 and was the lowest among the other subgroups. Compared to the results from the original evaluation of e-rater for GRE (Ramineni et al., 2012a), the standardized mean score differences were slightly

Table 3 e-rater Evaluation Results for GRE Writing Prompts

		Human1 by Human2		Human1 by e-rater (rounded to integers)						Human1 by e-rater (unrounded)	
Prompt	*N*	Mean	*SD*	Mean	*SD*	Kappa	Wtd. kappa	% agree	% adj. agree	*SD*	Corr.
Issue	531	3.74	0.85	3.74	0.89	0.44	0.73	63	99	0.00	0.77
Argument	525	3.61	0.98	3.61	1.01	0.38	0.74	56	97	0.00	0.77

Note. Wtd. Kappa = weighted kappa; Corr. = Pearson correlation. *N* is the average across all prompts.

Table 4 e-rater Evaluation Results by Test Center Country and by Ethnicity for Issue Prompt

		Human1 by Human2		Human1 by e-rater (rounded to integers)						Human1 by e-rater (unrounded)	
Subgroup	*N*	Mean	*SD*	Mean	*SD*	Kappa	Wtd. kappa	% agree	% adj. agree	*SD*	Corr.
Test center country											
China	4,005	2.96	0.58	3.41	0.76	0.13	0.38	43	96	0.68	0.52
India	7,887	2.99	0.78	3.06	0.87	0.33	0.63	56	98	0.09	0.68
Japan	303	3.18	0.76	3.16	0.89	0.41	0.69	61	99	−0.02	0.76
Korea	1,008	2.81	0.73	2.84	0.92	0.33	0.64	55	98	0.02	0.69
Taiwan	672	2.76	0.72	2.73	0.89	0.34	0.62	58	97	−0.06	0.68
Ethnicity											
White	56,058	4.00	0.75	3.97	0.78	0.44	0.69	65	99	−0.04	0.74
African American	6,263	3.53	0.77	3.46	0.88	0.43	0.71	63	99	−0.09	0.76
Hispanic	5,401	3.72	0.80	3.68	0.88	0.44	0.73	64	99	−0.05	0.77
Asian	8,746	3.50	0.88	3.59	0.93	0.41	0.72	59	99	0.09	0.76
American Indian	771	3.29	0.94	3.28	1.01	0.45	0.77	61	99	−0.03	0.81
Other	4,977	3.62	0.93	3.63	0.96	0.44	0.75	61	99	0.01	0.79

Note. Wtd. Kappa = weighted kappa; Corr = Pearson correlation. The highlighted cells are cases that violate the flagging conditions and criteria described in Table 1.

Table 5 e-rater Evaluation Results by Test Center Country and by Ethnicity for Argument Prompts

		Human1 by Human2		Human1 by e-rater (rounded to integers)						Human1 by e-rater (unrounded)	
Subgroup	*N*	Mean	*SD*	Mean	*SD*	Kappa	Wtd. kappa	% agree	% adj. agree	*SD*	Corr.
Test center country											
China	4,923	3.09	0.65	3.47	0.75	0.15	0.39	45	96	0.56	0.50
India	8,613	2.87	0.86	2.79	0.91	0.35	0.68	56	98	−0.10	0.72
Japan	359	3.05	0.85	2.89	1.01	0.38	0.71	57	98	−0.18	0.78
Korea	1,174	3.03	0.71	3.10	0.87	0.30	0.59	56	98	0.08	0.63
Taiwan	761	2.87	0.65	2.71	0.85	0.28	0.57	55	98	−0.22	0.65
Ethnicity											
White	62,221	3.86	0.94	3.87	0.92	0.39	0.71	57	97	0.02	0.74
African American	6,879	3.19	0.93	3.08	1.05	0.35	0.72	53	97	−0.13	0.76
Hispanic	6,033	3.47	0.96	3.47	1.03	0.37	0.73	55	97	0.00	0.76
Asian	10,183	3.47	0.97	3.53	1.01	0.38	0.73	55	98	0.06	0.76
American Indian	826	3.15	1.02	3.06	1.08	0.44	0.78	59	97	−0.06	0.81
Other	5,458	3.50	1.02	3.50	1.05	0.40	0.76	56	98	0.00	0.79

Note. Wtd. Kappa = weighted kappa; Corr = Pearson correlation. The highlighted cells are cases that violate the flagging conditions and criteria described in Table 1.

Table 6 Mean e-rater Scores, Human Scores, and Score Differences for Overall Sample and Selected Subgroups for Issue and Argument Prompts

		Mean (*SD*)		
Subgroup	*N*	Operational e-rater score	Operational human score	Mean diff. (e-rater, human)
Issue				
Overall	103,151	3.73 (0.86)	3.74 (0.86)	−0.004 (0.58)
China	4,005	3.40 (0.72)	2.96 (0.58)	0.44 (0.64)
Argument				
Overall	115,071	3.60 (0.99)	3.61 (0.99)	−0.002 (0.67)
China	4,923	3.47 (0.71)	3.09 (0.65)	0.37 (0.68)
Taiwan	761	2.70 (0.84)	2.87 (0.65)	−0.17 (0.65)
African American	6,879	3.06 (1.06)	3.19 (0.93)	−0.13 (0.71)

larger for China issue and slightly lower for African American argument but much larger (double) for China argument. Mean score differences were not reported for Taiwan in original evaluations owing to an insufficient sample size. Table 6 reports the mean operational e-rater scores, human scores, and mean score differences for the overall sample and the selected subgroups for the two prompt types.

The mean operational human and e-rater scores were greater on average for the issue prompts than for the argument prompts. The average e-rater and human scores for the China subgroup on issue prompts were lower, but the mean score difference between e-rater and human scores was greater than that for the overall. The large positive difference implied e-rater producing higher scores for essays for test takers from China compared to human raters.

For the argument prompts, average scores for all subgroups were lower than those for the overall sample. Taiwan received the lowest average e-rater and human scores; the average e-rater score for the African American subgroup was lower than that for the China subgroup, but the mean operational human score was slightly greater than that for China. The distance between the human and e-rater scores was positively large for China, implying higher scores by e-rater than by humans, whereas for the other two subgroups, Taiwan and African American, the distance was negative, implying lower scores by e-rater compared to humans for the essays from these subgroups.

Research Question 1

We conducted feature-level analyses for each of these demographic subgroups and also examined and compared the weights for the different features in the e-rater models for different subgroups and for the overall sample. Figures 2–5

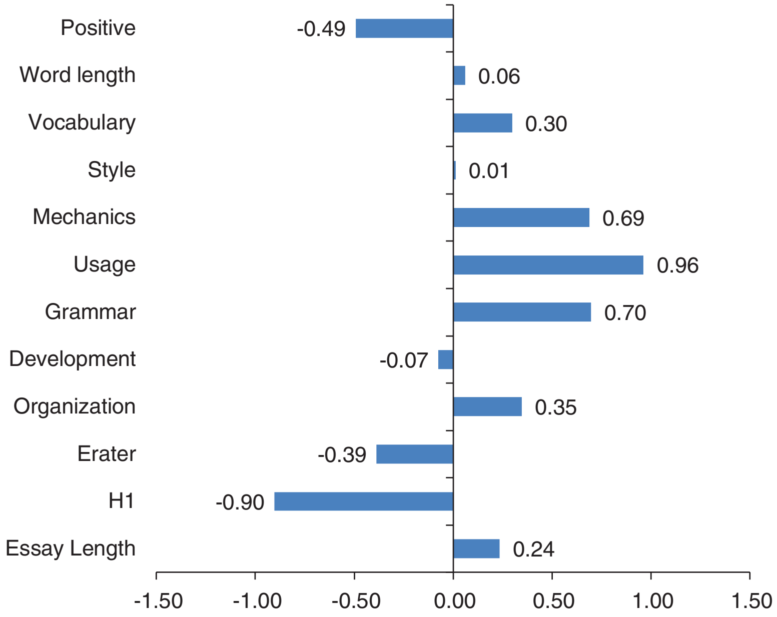

Figure 2 Plot of standardized feature scores, essay lengths, and human and e-rater scores for China issue ($N = 4{,}005$).

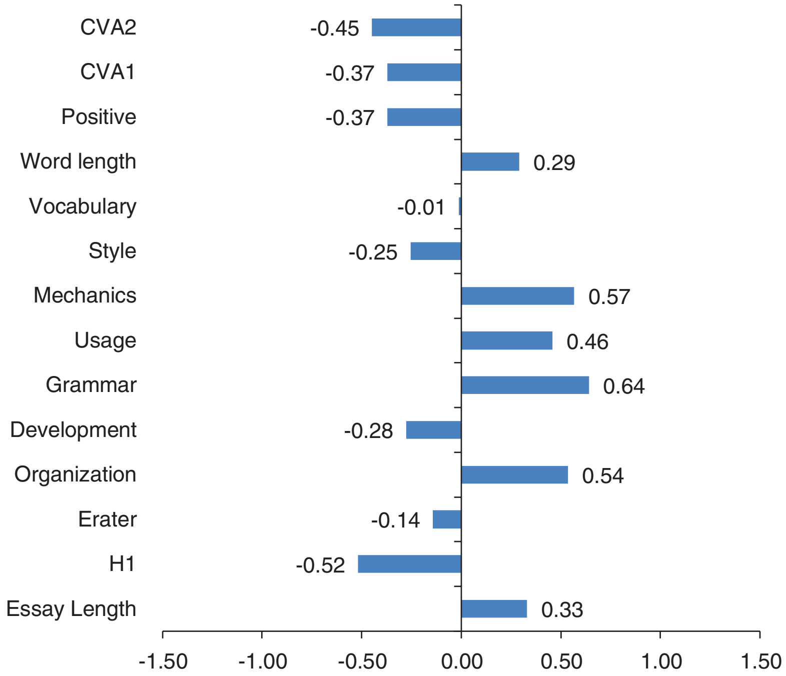

Figure 3 Plot of standardized feature scores, essay lengths, and human and e-rater scores for China argument ($N = 4{,}923$).

present the plots for standardized feature scores, essay lengths, and human (H1) and e-rater holistic scores for China issue, China argument, Taiwan argument, and African American argument, respectively. Positive denotes the positive feature measuring collocation and preposition usage; word length is the average word length over the essay response; vocabulary denotes the sophistication of word choice; style, mechanics, usage, and grammar errors are typical measures of language control; and development and organization are related to the structure of the text where organization is a count of the number of discourse elements and development is measured as the average length of discourse elements. CVA1 and CVA2 are content features measuring similarity of vocabulary of the essay response to sample responses at various score points and are included only for the argument prompt type.

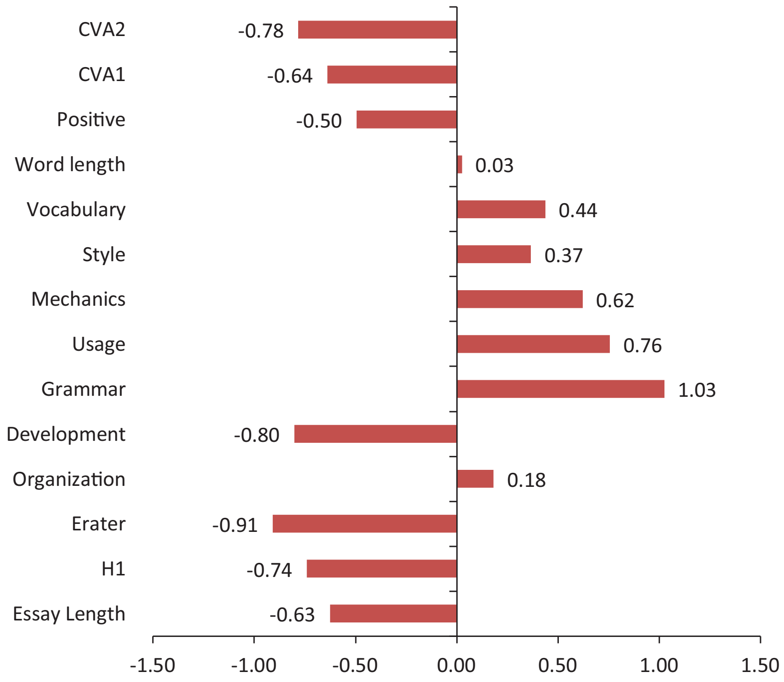

Figure 4 Plot of standardized feature scores, essay lengths, and human and e-rater scores for Taiwan argument ($N = 761$).

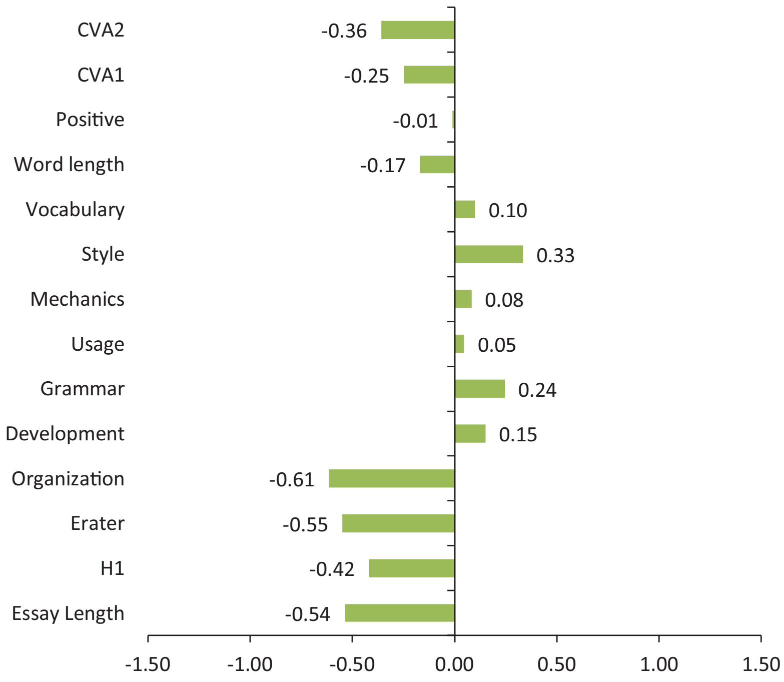

Figure 5 Plot of standardized feature scores, essay lengths, and human and e-rater scores for African American argument ($N = 6,879$).

Feature Values

On the issue task, the test takers from China wrote longer essays on average than the overall population and received lower human and e-rater scores on average than everyone. At the e-rater feature level, these examinees had more language errors (grammar, usage, mechanics) but obtained higher organization scores and scored higher for sophisticated word choice than the overall population.

Similar to the issue task, on the argument task, the test takers from China wrote longer essays on average than the overall population and received lower human and e-rater scores on average than the overall population. At the e-rater feature level, they had more language errors (grammar, usage, mechanics) but obtained higher organization scores and used longer words on average (as opposed to scoring for choosing more sophisticated words, as for issue responses) than the overall population. Furthermore, their scores for development were noticeably lower on average, and so were the average scores for the two content features.

The test takers from country of Taiwan, on the argument task, wrote shorter essays on average than the overall population and received lower human and e-rater scores on average than the overall population. At the e-rater feature level, they had considerably more language errors (grammar, usage, mechanics, and style) and very low development and content scores. Vocabulary or word choice is the only feature on which they scored higher on average than the overall population, with some advantage in their scores for organization.

Similar to the test takers from Taiwan, the African American test takers, on the argument task, wrote shorter essays on average than the overall population and received lower human and e-rater scores on average than the overall population. At the e-rater feature level, they had lower content scores and organization score on average than the overall population. They made more grammar and style errors on average compared to everyone and had only a slight advantage in development and word choice, while the average word length was shorter.

Feature Weights

Following feature-level analyses, we built separate e-rater scoring models for each subgroup[2] as well as for the overall sample and compared the weights for e-rater features across the different models. For these analyses, we built regular generic e-rater models for issue (9 features with fixed intercept and weights across prompts) and generic models, including the content features, for argument (11 features with fixed intercept and weights across prompts) to avoid dealing with more complex operational models (generic with prompt-specific intercept for issue and prompt-specific for argument). Tables 7 and 8 report the weights (converted to percentages for ease of interpretation) for e-rater features under different models based on different group of examinees for issue and argument, respectively. The weights for the features varied widely among the subgroups as well as compared to those for the overall sample. When compared to the model built on the overall sample, the feature weights were most discrepant for the China subgroup, for both issue and argument, with larger weights for grammar and mechanics and smaller weights for organization and development. For issue, the weights for the vocabulary features were also larger in models built on data from China only and Taiwan only, and for argument, the weights for the two content features were lower for the model built on essays from China only. One of the content features (CVA2) was also weighted lower in the model built on argument essays from the Taiwan subgroup only, while the same feature received a larger weight in the model built on the argument essays from the African American subgroup only, compared to the model built on the overall sample.

Table 7 Weights for e-rater Features From Models Built on Subgroup Data Only and on Overall Data: Issue (Generic)

	Standardized estimate percentage			
Subgroup	Taiwan	China	African American	Overall
Parameter				
Grammar	8.0	10.6	4.0	4.8
Usage	6.7	8.8	3.9	11.8
Mechanics	10.9	15.7	5.4	8.2
Style	−3.3	2.5	0.8	0.7
Organization	26.1	17.8	39.7	30.1
Development	27.6	22.3	33.7	29.9
Positive	6.3	6.2	1.2	2.6
Word length	10.5	6.4	7.0	5.7
Vocabulary	7.1	9.8	4.4	6.3
Adj. R^2	0.469	0.282	0.590	0.592

Note. Relative weights are converted to percentages.

Table 8 Weights for e-rater Features From Models Built on Subgroup Data Only and on Overall Data: Argument (Generic With Content Features)

	Standardized estimate percentage			
Subgroup	Taiwan	China	African American	Overall
Parameter				
Grammar	11.5	17.1	2.1	4.0
Usage	5.0	6.5	1.9	5.1
Mechanics	5.7	14.1	0.1	2.4
Style	−0.7	0.3	1.7	0.2
Organization	28.8	19.3	38.2	34.3
Development	28.2	20.3	27.2	27.1
Positive	5.6	5.0	1.7	2.6
Word length	4.0	4.8	2.7	2.7
Vocabulary	−1.2	1.6	−0.3	0.3
CVA1	8.7	4.2	10.5	9.8
CVA2	4.4	6.9	14.1	11.4
Adj. R^2	0.421	0.274	0.561	0.586

Note. Relative weights are converted to percentages. CVA1 and CVA2 are content features measuring similarity of vocabulary of the essay response to sample responses at various score points.

Summary

The test takers from China, who received higher e-rater scores than humans, wrote longer essays on average. Their essays exhibited strength in vocabulary, word length, and organization but weaknesses in language conventions and content for argument, implying the possible use of memorized well-formed text (shell text) to inflate essay lengths. (Whether and what proportion of variance is indeed accounted for by the shell text will need to be investigated empirically once we have developed a feature, quantitative measure of shell text that can be used in the regression model.) The writing of test takers from Taiwan showed weaknesses across all features with the exception of vocabulary, while the African American test takers' writing displayed weaknesses in content and organization of the response. Both Taiwan and African American test takers wrote shorter essays on average. Because the organization and development features, the sum of which is strongly correlated with essay length, are most highly weighted (roughly 30% for each, as reported in Tables 7 and 8) in the linear regression model predicting human score, the subgroups with lower scores on these features can receive lower e-rater scores as a result.

As noted for China, responses showed weaknesses in language conventions and content, which receive relatively lower weights in the model built on the overall sample than in the model built on data from China only. These weaknesses are further subsumed by higher scores on essay length-related features, which are assigned larger weights. Overall, the much lower R^2 (∼.27, .30 lower than the R^2 for the model on overall data) for the models built on essays from China only for issue and argument suggests that a large amount of variance in the observed scores is not being captured or explained by the current set of e-rater features. Analyses of the feature scores and feature weights provided us with some overview of the differences in the writing characteristics that can be associated with the differences observed for each subgroup and how the weights for different features capturing these different characteristics, empirically determined in a model trained on a large pool of data, may influence e-rater scores and result in a poor model fit for some groups. We further investigated if the subgroup differences were an artifact of the modeling procedures.

Research Question 2

Model Evaluations

Tables 9–12, and 13 report the evaluation results for the overall sample and for the three subgroups China, Taiwan, and African American test takers for issue prompts under the different regression models. Although the Taiwan and African American subgroups were not identified as of concern in the regular evaluation results for issue prompts (reported in Table 4), we included them for these analyses to allow us to analyze and compare results for more than just one subgroup, that is, China, across the different regression models. Models were built and evaluated against single human scores;

Table 9 e-rater Evaluation Results for Issue Prompts Under Different Regression Models

		Human1 by e-rater (rounded to integers)								Human1 by e-rater (unrounded)	
Model	*N*	Human1 mean	Human1 *SD*	e-rater mean	e-rater *SD*	Kappa	Wtd. kappa	% agree	% adj. agree	*SD*	Corr.
Linear	107,200	3.74	0.86	3.74	0.92	0.43	0.73	62	99	0.00	0.77
Logistic	103,151	3.74	0.86	3.74	0.91	0.44	0.74	62	99	0.01	0.77
Unit wtd.	103,151	3.74	0.86	3.74	1.02	0.11	0.38	38	86	0.003	0.41
CART	103,151	3.74	0.86	3.66	1.34	0.31	0.62	50	87	0.02	0.70

Note. CART = classification and regression trees; Wtd. Kappa = weighted kappa; Corr. = Pearson correlation. The highlighted cells are cases that violate the flagging conditions and criteria described in Table 1.

Table 10 e-rater Evaluation Results for Subgroups Under Linear Generic Model

		Human1 by e-rater									
Subgroup	*N*	Human1 mean	Human1 *SD*	e-rater mean	e-rater *SD*	Kappa	Wtd. kappa	% agree	% adj. agree	*SD*	Corr.
China	4,005	2.96	0.58	3.41	0.78	0.13	0.38	42	95	0.67	0.52
Taiwan	672	2.76	0.72	2.71	0.91	0.32	0.61	56	97	−0.09	0.68
AA	6,263	3.53	0.77	3.46	0.91	0.42	0.71	62	99	−0.10	0.76

Note. AA = African American; Wtd. Kappa = weighted kappa; Corr. = Pearson correlation. The highlighted cells are cases that violate the flagging conditions and criteria described in Table 1.

Table 11 e-rater Evaluation Results for Subgroups Under Logistic Generic Model

		Human1 by e-rater									
Subgroup	*N*	Human1 mean	Human1 *SD*	e-rater mean	e-rater *SD*	Kappa	Wtd. kappa	% agree	% adj. agree	*SD*	Corr.
China	4,005	2.96	0.58	3.40	0.76	0.14	0.38	45	96	0.65	0.52
Taiwan	672	2.76	0.72	2.73	0.89	0.33	0.61	57	97	−0.07	0.68
AA	6,263	3.53	0.77	3.46	0.89	0.42	0.71	62	99	−0.10	0.76

Note. AA = African American; Wtd. Kappa = weighted kappa; Corr. = Pearson correlation. The highlighted cells are cases that violate the flagging conditions and criteria described in Table 1.

Table 12 e-rater Evaluation Results for Subgroups Under Unit-Weighted Generic Model

		Human1 by e-rater									
Subgroup	*N*	Human1 mean	Human1 *SD*	e-rater mean	e-rater *SD*	Kappa	Wtd. kappa	% agree	% adj. agree	*SD*	Corr.
China	4,005	2.96	0.58	3.59	0.92	0.05	0.21	33	84	0.84	0.32
Taiwan	672	2.76	0.72	3.10	1.07	0.12	0.37	38	86	0.37	0.44
AA	6,263	3.53	0.77	3.44	1.10	0.10	0.36	37	85	−0.10	0.40

Note. AA = African American; Wtd. Kappa = weighted kappa; Corr. = Pearson correlation. The highlighted cells are cases that violate the flagging conditions and criteria described in Table 1.

Table 13 e-rater Evaluation Results for Subgroups Under Classification and Regression Trees Generic Model

		Human1 by e-rater									
Subgroup	*N*	Human1 mean	Human1 *SD*	e-rater mean	e-rater *SD*	Kappa	Wtd. kappa	% agree	% adj. agree	*SD*	Corr.
China	4,005	2.96	0.58	2.91	1.42	0.06	0.26	35	69	−0.01	0.37
Taiwan	672	2.76	0.72	2.32	1.26	0.15	0.41	37	78	−0.45	0.55
AA	6,263	3.53	0.77	3.37	1.30	0.28	0.56	50	86	−0.10	0.66

Note. AA = African American; Wtd. Kappa = weighted kappa; Corr. = Pearson correlation. The highlighted cells are cases that violate the flagging conditions and criteria described in Table 1.

therefore human–human agreement statistics and degradation statistics (change in agreement from human–human to human–e-rater), which are part of default e-rater evaluation criteria, are not presented in these tables.

The evaluation results for the overall sample were comparable under the linear and the logistic regression models, with somewhat lower weighted kappa (less than .70) and correlations under the CART model and very low weighted kappa and correlations (around .40) under the unit-weighted model. For the subgroups, using the linear model as the baseline, the standardized mean score difference between e-rater and humans decreased by .02 for the China and Taiwan subgroups and increased by .004 for the African American subgroup under the logistic model. The differences became larger for China and Taiwan under the unit-weighted model, with the difference reversed in direction for Taiwan. Under the CART model, for China, which was identified as the subgroup of concern, even though the weighted kappa and the correlations were very low, the standardized difference was reduced by .66 and was fully within the acceptable threshold (.10). However, for Taiwan, the standardized difference increased by .38, therefore making it a subgroup of concern. The magnitude and direction of the standardized difference remained more stable for the African American subgroup under different models, with some significant changes for China and Taiwan under certain models. For issue prompts, the CART model was the most successful in bringing the large standardized difference (.65) observed for China under the linear model well below the acceptable threshold (.10). However, there was a negative impact for Taiwan.

Tables 14–17, and 18 report the evaluation results for the overall sample and for the three subgroups of China, Taiwan, and African American test takers for argument prompts under the linear and logistic prompt–specific models and unit-weighted and CART generic models. Logistic regression models could not be computed for four argument prompts and were excluded from the logistic regression analyses (hence there is a small difference in the sample sizes for the logistic regression model results in the tables). Again, the evaluation results for the overall sample were comparable under the linear and the logistic models. The weighted kappa and correlation values were slightly lower under the other two models but above the threshold of .70 for the CART model and below but close to the threshold value for the unit-weighted model. For the subgroups, using the linear model as the baseline, the standardized mean score difference between e-rater and humans was lowered by .03 for the China subgroup and by .013 for the African American subgroup (bringing it under the acceptable threshold value) under the logistic model. For the unit-weighted model, the standardized difference was reduced by .37 for the China subgroup, thereby bringing it much closer to the threshold value of .10, and a .06 reduction for the African American subgroup brought the standardized difference well below the acceptable threshold. There was an increase in the standardized difference for Taiwan by .14. Under the CART model, there was a reduction in the standardized difference for China by .29 but at the same time an increase for the Taiwan and African American subgroups by .25 and .09, respectively. For argument prompts, the unit-weighted model was the most successful in bringing the large standardized difference (.56) for China observed under the linear model close to the acceptable threshold (.19) as well as for bringing the standardized difference for the African American subgroup (−.11) well below the threshold (.05). However, there was a small negative impact for Taiwan.

Score Distributions

The scores for issue tasks produced by the linear model were almost perfectly correlated with scores produced by the logistic model, highly correlated with the scores from the CART model ($r = .85$), and moderately correlated with the scores produced by the unit-weighted model ($r = .52$). For argument, the scores from the linear model were highly correlated with scores from all the other models ($r = .84$ for logistic, $r = .87$ for unit weighted, and $r = .86$ for CART). Tables 19 and 20 report the distribution of operational Human1 and e-rater scores and e-rater scores under the new models for the essay responses on the issue and argument prompts, respectively.

The distributions for e-rater scores were fairly similar under the different modeling procedures with differences for some score categories but with no consistent patterns. Under the CART model, however, many more essays received a score of 6 than under any other model or under human scoring. Also for the CART model for issue, the scores of 2s were converted to 1s upon applying the Kelly formula (for scaling e-rater scores to match human score distributions) and rounding off to the nearest integer value. The unit-weighted model behaved differently for issue and argument prompts, assigning many more 6s compared to the linear model or human raters for responses to issue prompts and very few 6s for argument prompts. Overall, e-rater assigned many more 1s and fewer 6s than human raters for both issue and argument, with the exceptions of the unit-weighted and CART models for issue and the CART model for argument. The cross-task

Table 14 e-rater Evaluation Results for Argument Prompts Under Different Regression Models

		Human1 by e-rater (rounded to integers)								Human1 by e-rater (unrounded)	
Model	*N*	Human1 mean	Human1 *SD*	e-rater mean	e-rater *SD*	Kappa	Wtd. kappa	% agree	% adj. agree	*SD*	Corr.
Linear	115,071	3.61	0.99	3.62	1.01	0.39	0.74	56	97	0.01	0.77
Logistic	111,027	3.61	0.99	3.63	1.01	0.41	0.75	57	98	0.02	0.78
Unit wtd.	115,071	3.60	0.99	3.61	1.01	0.27	0.65	48	94	0.002	0.67
CART	115,071	3.61	0.99	3.64	1.22	0.36	0.71	52	93	0.02	0.72

Note. CART = classification and regression trees; Wtd. Kappa = weighted kappa; Corr. = Pearson correlation. The highlighted cells are cases that violate the flagging conditions and criteria described in Table 1.

Table 15 e-rater Evaluation Results for Subgroups Under Linear Prompt–Specific Model

		Human1 by e-rater									
Subgroup	*N*	Human1 mean	Human1 *SD*	e-rater mean	e-rater *SD*	Kappa	Wtd. kappa	% agree	% adj. agree	*SD*	Corr.
China	4,722	3.09	0.65	3.47	0.75	0.15	0.39	45	96	0.56	0.50
Taiwan	733	2.87	0.65	2.73	0.85	0.28	0.57	55	98	−0.19	0.64
AA	6,658	3.19	0.92	3.09	1.04	0.35	0.71	53	97	−0.11	0.75

Note. AA = African American; Wtd. Kappa = weighted kappa; Corr. = Pearson correlation. The highlighted cells are cases that violate the flagging conditions and criteria described in Table 1.

Table 16 e-rater Evaluation Results for Subgroups Under Logistic Prompt–Specific Model

		Human1 by e-rater									
Subgroup	*N*	Human1 mean	Human1 *SD*	e-rater mean	e-rater *SD*	Kappa	Wtd. kappa	% agree	% adj. agree	*SD*	Corr.
China	4,722	3.09	0.65	3.45	0.74	0.16	0.40	47	96	0.53	0.51
Taiwan	733	2.87	0.65	2.74	0.80	0.32	0.59	59	99	−0.19	0.64
AA	6,658	3.19	0.92	3.11	1.00	0.38	0.72	56	97	−0.10	0.76

Note. AA = African American; Wtd. Kappa = weighted kappa; Corr. = Pearson correlation. The highlighted cells are cases that violate the flagging conditions and criteria described in Table 1.

Table 17 e-rater Evaluation Results for Subgroups Under Unit-Weighted Generic Model

		Human1 by e-rater									
Subgroup	*N*	Human1 mean	Human1 *SD*	e-rater mean	e-rater *SD*	Kappa	Wtd. kappa	% agree	% adj. Agree	*SD*	Corr.
China	4,923	3.09	0.65	3.24	0.84	0.17	0.43	48	96	0.19	0.47
Taiwan	761	2.87	0.65	2.62	0.92	0.21	0.48	48	94	−0.33	0.56
AA	6,879	3.19	0.93	3.25	0.99	0.27	0.62	48	94	0.05	0.64

Note. AA = African American; Wtd. Kappa = weighted kappa; Corr. = Pearson correlation. The highlighted cells are cases that violate the flagging conditions and criteria described in Table 1.

Table 18 e-rater Evaluation Results for Subgroups Under Classification and Regression Trees Generic Model

		Human1 by e-rater									
Subgroup	*N*	Human1 mean	Human1 *SD*	e-rater mean	e-rater *SD*	Kappa	Wtd. kappa	% agree	% adj. agree	*SD*	Corr.
China	4,923	3.09	0.65	3.45	1.17	0.09	0.31	38	81	0.27	0.39
Taiwan	761	2.87	0.65	2.69	0.97	0.17	0.45	48	92	−0.44	0.50
AA	6,879	3.19	0.93	3.10	1.09	0.38	0.69	55	95	−0.20	0.71

Note. AA = African American; Wtd. Kappa = weighted kappa; Corr. = Pearson correlation. The highlighted cells are cases that violate the flagging conditions and criteria described in Table 1.

Table 19 Distribution of Human1 and e-rater Scores for Issue Prompts

Score level	Human1	e-rater (linear)	e-rater (logistic)	e-rater (unit wtd.)	e-rater (CART)
1	592	1,752	1,425	2,361	13,870
2	6,094	7,766	6,790	8,887	0
3	33,628	28,492	29,414	27,330	25,945
4	49,514	48,820	46,135	41,421	37,721
5	15,498	19,808	18,034	20,822	18,843
6	1,874	562	1,353	2,330	6,772

Note. CART = classification and regression trees; unit wtd. = unit weighted.

Table 20 Distribution of Human1 and e-rater Scores for Argument Prompts

Score level	Human1	e-rater (linear)	e-rater (logistic)	e-rater (unit wtd.)	e-rater (CART)
1	904	3,177	5,906	3,624	4,105
2	13,584	12,062	12,835	11,655	15,404
3	38,704	32,504	33,364	31,837	34,436
4	41,525	46,178	41,346	47,111	33,376
5	17,429	19,990	19,718	20,267	19,657
6	2,925	1,160	1,902	577	8,093

Note. CART = classification and regression trees; unit wtd. = unit weighted.

correlations computed on a subset of examinees with scores present on both issue and argument tasks were highest under the linear model ($r = .73$), followed by the logistic model ($r = .62$), the CART model ($r = .56$), and the unit-weighted model ($r = .52$).

Summary

None of the regression modeling procedures improved the e-rater scoring model performance at the overall level when compared to the linear regression modeling procedure. There was varying impact on the mean score differences between e-rater and humans for the subgroups under different models, particularly the unit-weighted and CART models, with the differences falling to well below or close to the acceptable threshold.

We then turned to rescoring and SME review of discrepant cases combined with empirical analyses to further explore and understand the root causes of these differences for the demographic subgroups.

Research Question 3

Rescoring and Review

To further review the discrepancies closely between e-rater and human scores, we identified a set of discrepant essays—323 argument responses (100 each for China and African American and 123 for Taiwan) and 100 issue responses (for China)—and submitted them for rescoring by the five expert raters. Scoring was completed for 313 of the argument responses (100 for China, 90 for African American, and 123 for Taiwan) and 90 of the 100 issue responses (for China). Table 21 reports the mean score differences between operational human and e-rater scores and expert human scores for each of the four subgroups.

The human scores remained consistent for issue and argument for China. However, the differences between e-rater and new human scores were lower for the Taiwan (almost nil) and African American subgroups, while suggesting differences between scores given by operational human raters and expert human raters for these subgroups.

After the completion of rescoring, the expert raters were led in a facilitated discussion and review of essays and their scoring observations. A general review was followed by a discussion of a few maximally discrepant essays from each subgroup, representing both positive (overscored by e-rater) and negative (underscored by e-rater) differences, for which the raters were provided with operational human and e-rater scores as well as the new expert human scores. They evaluated these cases to identify themes and hypothesize reasons for differences between the scores. The raters agreed that all the

Table 21 Mean (*SD*) Score Differences Between Operational Human and e-rater and Expert Human Scores

Subgroup	e-rater vs. operational human	Operational human vs. expert human	e-rater vs. expert human
China issue[a]	1.30 (0.31)	0.00 (0.56)	1.30 (0.56)
China argument[b]	1.31 (0.28)	−0.03 (0.41)	1.28 (0.42)
Taiwan argument[c]	−0.20 (0.61)	0.24 (0.62)	0.04 (0.66)
African American argument[a]	−1.35 (0.42)	0.54 (0.60)	−0.81 (0.74)

[a] $n = 90$. [b] $n = 100$. [c] $n = 123$.

essays encountered for rescoring were at the lower end of the scale (mostly 2s and 3s). The raters quickly became aware of the narrow range when scoring and had to carefully avoid an urge to compare and contrast adjacent papers in an effort to widen their score ranges. The raters felt that the majority of the essays exhibited poor language control; however, some essays with poor language but good content received higher scores (3s) than the rest (examples identified in the African American subgroup). The raters also found the use of shell text along with the text that is part of the writing prompt/question prevalent across the essay responses of test takers from China, particularly for argument, resulting in three-paragraph-long responses but suffering from lack of ideas and poor language. The raters identified instances of the use of generic (non-content-specific) statements, syntactic repetition at the beginning of each paragraph, repetition of ideas around generic statements or prompt text, and lack of cohesion in the essay after the first few memorized sentences as cues to shell text. The raters understood that test takers from China frequently use shell text, but they do not view shell text as problematic or as a negative style of writing. They emphasized that they are trained to be neutral to the use of shell text and that they look for original ideas and content beyond the shell text in the response to determine the appropriate score. In some cases, they expressed the idea that the examinees are able to use shell text cleverly to enhance the structure and framework of their responses without compromising originality, cohesion, and content. The raters emphasized that these essays commonly suffered from poor language use, such as lack of punctuation, incorrect capitalization, idiomatic use, lack of cohesion, and lack of development. There was some variability across prompts, but all responses were relatively short.

The raters pointed out that the rubric with analytically defined score categories helps them to choose the appropriate score and/or decide between adjacent scores. The lower half of the current rubric focuses heavily on language control and language errors, and the emphasis shifts to development of ideas, original contribution, content, and cohesion when moving up the scale (see Table 2 for the actual rubric). The lower half of the rubric uses the language “a typical response exhibits one or more of the (undesirable) characteristics,” with emphasis shifting to “a typical response exhibits” all the (desirable) characteristics moving up the scale. The scoring increasingly becomes more holistic at the upper end of the scale, with raters increasing their focus on the overall descriptor for the score category, for example, “A 6 paper presents a cogent, well-articulated critique of the argument and conveys meaning skillfully.” The raters unanimously agreed that the essays that fail to exhibit good language control automatically fail to qualify for the upper half of the scale. Whereas human raters use language control as a threshold variable, e-rater uses linear weighting of all features.

The raters did not seem to pay particular attention to the five-paragraph traditional essay format and only evaluated for appropriate transitions rather than the number of transitions or number of discourse elements. They agreed that length directly relates to development of the essays and demonstrating effective language control but emphasized that content dominated the length in their evaluation. In terms of specific subfeatures scored by e-rater under grammar, usage, mechanics, and style, raters unanimously expressed that possessive errors, non-American English syntax, and some meaningful fragments in American English were easily ignored as long as they did not inhibit raters’ understanding of the text. Similarly, wrong use of prepositions or articles and faulty comparisons were not perceived as problematic in isolation unless they were frequent and/or recurrent errors affected the aggregate quality of the response. Raters ignored spelling and capitalization errors and considered these as typographical errors rather than mechanical errors, treating the responses as first drafts submitted without access to a spell checker. The raters considered use of passive voice as a writing style and not an error. Under writing style, they reported evaluating parallel structure, sentence variety, use of complex and compound sentences, embedding, and use of subordinating conjunctions (although/however), which are only partially evaluated by e-rater currently. In terms of organization and development, raters felt that strong writers often deviate from the common structure taught in English composition classes and are able to connect the ideas logically using approaches other than the standard paragraph format with introduction followed by main point followed by supporting ideas and conclusion. For example, often the main point may be stated at the end of the response, but the raters are able to comprehend the flow and hence identify the various organizational elements in varying/nonstandard sequence in the text. We do not know the extent to which e-rater can be flexible in evaluating the structure of the text. Raters disagreed with e-rater giving higher weight to organization and development. They felt that language control frequently trumped organization and development, citing examples of three-paragraph-long responses that were scored very low by human raters for poor language and were scored high by e-rater. They reiterated that it would be impossible for a response to cross over to the upper half of the scale (which is where organization and development along with content and analysis of ideas will be considered important) if it failed on language control. They also considered lexical complexity and content as more important factors in evaluating writing. During the discussion of individual essays, raters noted that e-rater appeared to be undervaluing content and analysis of evidence or argument for short responses on the argument prompts (identified from the African American subgroup). Furthermore, they pointed out that most of the coached essays start out by disagreeing with the prompt and speculated that in cases where the writer agreed with the prompt, e-rater was probably not able to understand the writer's agreement with the prompt and misjudged it as repetition of prompt text. As noted previously, the decrease in differences between e-rater and new human scores from expert raters for essays from Taiwan and the African American subgroups suggested differences between scores given by operational human raters and expert human raters for these subgroups. While the current investigation focused on understanding the differences between human scores and e-rater scores solely when reviewing discrepant essays, for the discrepant cases where new expert human scores aligned more with e-rater scores, the expert raters' comments suggested the relatively "less-experience" of some operational human raters in applying the rubric to essays that fall on the borderline for two score categories as a possible source of discrepancies between the two sets of human scores.

Summary

The human scores for the essays from the China subgroup remained unchanged after rescoring for both issue and argument. The score differences were reduced for Taiwan (almost nil) and African American subgroups. In the expert review, several interesting observations were made by expert raters regarding the use of the scoring scale/rubric, the human rating process/procedures, and the e-rater scoring mechanism. The expert raters noted that the discrepant essays belonged to the lower end of the scale and suffered from poor language control. They also identified the presence of shell text in essay responses for argument from the China subgroup. The human rating process was found to be influenced strongly by the scale structure and did not fully correspond to the e-rater scoring mechanism. The human raters appeared to be using conditional logic and a rule-based approach to their scoring, whereas e-rater uses linear weighting of all the features. The raters were treating the scale as two halves conditioned on language control and errors. Within each half, the scoring process was then guided by rules in each score category to determine if the rules were met, followed by a check of the overall descriptor for the score category as a holistic check. E-rater, in contrast, evaluates each essay for all nine or 11 features and assigns weights to each of the features with minimal penalty for language control and maximum credit for organization and development (influenced by essay length). From this human versus e-rater review, it appeared that, relative to human raters, e-rater is less severe on language errors, overvalues organization and development, and occasionally undervalues content.

Classification and Regression Trees

We used CART to analyze the decision rules or the splitting criteria in the classification trees. The purpose was to understand how these rules aligned with the rationale the human raters followed, as discussed under the expert review, and to identify the specific source(s) of discrepancies between the human and the automated scores.

Feature Variables

The trees for this analysis were built on a smaller set of combined features and fewer score categories formed by collapsing adjacent categories. This was done to deal with the complexity and length of rules. The CART method uses a binary recursive partitioning algorithm to grow the classification tree. As per the algorithm, a series of binary partitions are conducted on the training data using the values for each independent variable. These variables are called *splitting variables* and are based on values greater than or less than a particular value on that variable; the observations or cases in the data

Table 22 Description of the Combined Feature Variables

Feature	Description
GUMS: Language control	Grammar, usage, mechanics, style errors (error counts were summed and normalized by document length)
OD: Fluency	Organization, development (combined as log(words))
POS: Good attributes of writing	Correct preposition usage, good collocation usage (probability measure)
VOC: Vocabulary[a]	Word frequency, average word length
CON: Content[a]	Association with sample responses in each score category, association with sample responses in the highest score category (both measured as cosine values)

Note. CON = content; GUMS = language control; OD = fluency; POS = good attributes of writing; VOC = vocabulary.
[a]The highest contributing feature was picked.

set are split into two subsets as the tree progresses. A variable can appear more than one time as a splitting variable in the sequence, and there may be more than one terminal node for the same classification or the value of the dependent variable (e.g., score category). With multiple features (independent variables) of continuous nature and multiple score categories (dependent variables), we observed that the nodes multiplied exponentially and the tree grew very quickly and became unmanageable to work with. We combined the features into five categories (as informed by the Culham, 2003, six-trait model) and therefore used five features for building the tree: language control (error counts for grammar, usage, mechanics, and style were summed and normalized by document length), fluency (organization and development were combined into log of words), good attributes of writing (positive feature was used as existing), vocabulary (feature with higher correlation with human score between word frequency and word length was chosen), and content (feature with higher correlation with human score between association of the response with sample responses in each score category and association of the response with sample responses in the highest score category was chosen). These features are further described in Table 22. Four of the five features, excluding content, were used as a generic model for the issue prompt. We also collapsed the six score categories into two classes: low (1–3) and high (4–6). These modifications reduced the tree to a more modest size and simplified to some extent the evaluation and interpretation of CART decision rules against the rationale the human scorers used.

Decision Rules

CART provides options to generate multiple trees by growing and pruning the tree. There is a misclassification cost associated with each tree, calculated as specified in the initial settings. An optimal tree with minimum cost can therefore be chosen as the best fit by comparing the misclassification costs for the maximal tree (fully grown with no termination criteria) and all the alternate trees (pruned versions). The classification tree with the minimum cost had 71 terminal nodes for issue and 25 terminal nodes for argument. The CART output also provides an order of importance for all the independent variables in relation to the tree. The order of importance of variables for the issue prompt type was as follows:

1. fluency (organization and development combined as log of words; OD)
2. content (included only for argument task; CON)
3. language control (grammar, usage, mechanics, and style combined; GUMS)
4. good attributes of writing (positive feature; POS)
5. vocabulary (word frequency; VOC)

The relative order of importance of these variables remained the same for the argument prompt type, with content included in the order at the second place between fluency and language control. More than one terminal node belonged to each score class, and the splitting rules for sequences that led to three of the purest terminal nodes (with a majority of cases belonging to that score class) are shown for issue and argument tasks in Tables 23 and 24, respectively. We further computed and analyzed the mean and median values for the combined features in relation to these scoring rules for the subgroups of concern.

For the Chinese test takers on the issue task, the average value for the OD feature was 6.22, with a median of 6.26. These values were greater than 5.97, the cutoff value for the first splitting variable, implying that the majority cases would

Table 23 Classification and Regression Trees Decision Rules for Each of the Score Classes (1–2) for Issue Task

Score class	Rule	Rule
1	1	OD ≤ 5.97, OD ≤ 5.79
	2	OD ≤ 5.97, OD > 5.79, GUMS >0.25, POS ≤ 0.61
	3	OD ≤ 5.97, OD > 5.79, GUMS ≤0.25, VOC ≤ 4.95, VOC ≤ 4.62
2	1	OD > 5.97, GUMS ≤0.26, GUMS ≤0.19, OD > 6.13
	2	OD > 5.97, GUMS ≤0.26, GUMS ≤0.19, OD ≤ 6.13, VOC > 4.84
	3	OD > 5.97, GUMS >0.26, POS > 0.59, OD > 6.30, POS > 0.66

Note. CON = content; GUMS = language control; OD = fluency; POS = good attributes of writing; VOC = vocabulary.

Table 24 Classification and Regression Trees Decision Rules for Each of the Score Classes (1–2) for Argument Task

Score class	Rule	Rule
1	1	CON ≤0.36, OD ≤ 5.68
	2	CON ≤0.36, OD > 5.68, CON ≤0.26
	3	CON >0.36, OD ≤ 5.71, OD ≤ 5.50
2	1	CON >0.36, OD > 5.71, CON >0.48
	2	CON >0.36, OD > 5.71, CON ≤0.48, GUMS ≤0.23
	3	CON >0.36, OD > 5.71, CON ≤0.48, GUMS >0.23, POS > 0.64

Note. CON = content; GUMS = language control; OD = fluency; POS = good attributes of writing; VOC = vocabulary.

be classified into Score Class 2 (high-score category). The average value for GUMS errors was 0.29 and for POS and VOC was 0.60 and 4.96, respectively. When compared to the cutoff values for these features, Rule 3 for Score Class 2 would allow majority cases to score highly on the scale. A high score for OD (correlated with essay length) as per the splitting criteria, along with the order of importance of variables (OD > GUMS > POS > VOC), would offset the large number of language control errors (GUMS) in scoring.

For the Chinese test takers on the argument task, the average value for the OD feature was 5.98, with a median of 6.02. These values were greater than 5.71, the cutoff value for OD as the splitting variable, implying that a majority of essay responses from this subgroup would be classified into Score Class 2 (high-score category). The average value for the CON feature was 0.32, for the GUMS errors was 0.29, and for POS and VOC was 0.60 and 5.05, respectively. When compared to the cutoff values for these features, the average values for all features, except OD, were below the cutoff value, but given the order of importance of the splitting variables (OD > CON > GUMS > POS > VOC), a high OD feature score (correlated with essay length) would offset a large number of GUMS errors and low scores on other features. Therefore all the essay responses with such characteristics will be classified into Score Class 2 (high-score category).

For the test takers from Taiwan on the argument task, the average value for the OD feature was 5.62, with a median of 5.65. These values were smaller than 5.71, the cutoff value for OD as the splitting variable. The proximity of the mean and median to the cutoff suggests that the number of essay responses in Score Class 1 (low-score category) may be slightly higher than the number of essay responses classified into Score Class 2 (high-score category). The average value for the CON feature was 0.27, for the GUMS errors was 0.33 (larger than the value for the subgroup of China), and for POS and VOC was 0.59 and 4.98, respectively. When compared to the cutoff values for these features, the average values for all features were below the cutoff values, which further suggest that a majority of the responses for this subgroup will probably be classified into Score Class 1.

For the African American test takers on the argument task, the average value for the OD feature was 5.62, with a median of 5.67. These values, similar to the Taiwan subgroup, were smaller than 5.71, the cutoff value for OD as the splitting variable. However, the proximity of the mean and median to the cutoff suggests that the number of essay responses in Score Class 1 (low-score category) may be only slightly higher than the number of essay responses classified into Score Class 2 (high-score category). The average value for the CON feature was 0.33 (largest among the three subgroups of concern), for the GUMS errors was 0.29 (comparable to the China subgroup), for POS was 0.63 (largest among the three subgroups), and for VOC was 4.93 (smallest among the three subgroups but comparable). When compared to the cutoff values for these features, the average values for all features were below the cutoff values for the high-score category, which further suggests that a majority of the responses for this subgroup will probably be classified into Score Class 1. When compared to the China subgroup, despite the comparable scores for the two subgroups on the CON and GUMS features, the lower score on the OD feature (the most important splitting variable) triggers the classification rules for Score Class 1 (low-score category) and offsets any gains on any other less important features. We further revised the score classes to represent the categories of score differences between e-rater and human scores. There was no clear classification of cases and corresponding decision rules for issue tasks, but for argument tasks, the splitting criteria leading to the purest terminal nodes for the two score categories were primarily based on fluency and content.

Summary

From the expert review, it appeared that relative to human raters, e-rater is less severe on language errors, overvalues organization and development, and occasionally undervalues content. We used CART to test this hypothesis further. We derived e-rater scoring rules in the form of decision rules classifying an essay response into a certain score category and compared them to the human scoring rules reflected in the SME review. Contrary to the human scoring rules, which condition on language control and errors as the first splitting criterion between the low- and high-score categories, e-rater uses organization and development (correlated with essay length) as the most important splitting variable in classifying an essay response into the low-score versus the high-score category. A high score for organization and development combined, therefore, appeared to offset better or poor performance on other features, including language control errors, thus creating a trend of e-rater overscoring or underscoring for demographic subgroups that consistently score high or low on organization and development, respectively.

Spelling Correction

Human Scoring Context

During the qualitative review of the essays with the expert human raters, it was noted that the human raters ignore and do not count or penalize for spelling errors, unless the spelling errors interrupt the flow and/or impede overall understanding of the text. They treat the response as a first draft submitted by examinees without access to a spell checker. E-rater counts all spelling errors. This suggests that some groups of examinees may receive different scores from humans and e-rater due to this difference in scoring practice. To test this hypothesis and to make the treatment of spelling errors by e-rater reflect more closely the treatment of such errors by human raters, we submitted the essays to a spelling correction system and studied the impact of using spelling correction on e-rater scores. The essays selected for human rescoring/review were processed through ConSpel (Flor & Futagi, 2012), a system for detection and automatic correction of nonword (words that do not occur in the dictionary) misspellings for English, and subsequently reprocessed through the e-rater engine to produce new feature scores and e-rater scores for these essays. The spelling correction was conducted to test if the postcorrection e-rater feature scores and the overall e-rater scores would notably differ from the previous scores. Table 25 presents the mean scores for essay length and e-rater features for each subgroup before and after using the spelling detection and correction module.

There was an increase in the mean e-rater scores after spelling correction. At the feature level, as anticipated, there was a slight decrease in the mean number of mechanics errors across all subgroups, as the e-rater mechanics feature includes a count of spelling errors. The average word count increased by a few words for each subgroup. The spelling correction made the most difference in the average word count for the two China subgroups compared to the other two subgroups. There was an increase in the mean feature score for organization (the number of discourse elements) for each subgroup. As the organization and development features are inversely related, a simultaneous decrease in the development feature score (average length of the discourse elements) was also observed for all the subgroups. The average essay length, strongly related to the organization and development features, also increased after spelling correction.

The mean score differences between operational human and e-rater scores and the new e-rater score produced after spelling correction are reported in Table 26 for each of the four subgroups.

When the essay responses were corrected for spelling, the mean score differences between operational human and new e-rater scores for the two China subgroups were slightly greater than the score differences observed previously for e-rater and humans, whereas the mean score differences with new e-rater scores were slightly lowered for the Taiwan and African American subgroups but still exceeded the flagging threshold.

Table 25 Mean (*SD*) for e-rater Holistic Score, Essay Length, and e-rater Feature Scores for the Four Subgroups Before and After Spelling Correction

	China issue[a]		China argument[b]		Taiwan argument[c]		African American argument[a]	
Feature	Before	After	Before	After	Before	After	Before	After
Operational e-rater score	4.02 (0.50)	**4.25** (0.46)	4.07 (0.43)	**4.13** (0.46)	2.68 (0.72)	**2.77** (0.77)	1.90 (1.29)	**2.04** (1.34)
No. words (essay length)	640.16 (119.47)	**653.79** (119.64)	471.97 (75.62)	**479.99** (75.48)	285.80 (70.70)	**291.54** (73.18)	213.63 (121.56)	**217.34** (123.29)
Organization	4.23 (0.20)	**4.46** (0.54)	3.98 (0.19)	**4.27** (0.56)	3.64 (0.31)	**3.93** (0.62)	3.97 (0.54)	**4.08** (0.58)
Development	2.21 (0.18)	**2.00** (0.55)	2.17 (0.15)	**1.89** (0.55)	1.99 (0.33)	**1.71** (0.01)	1.23 (0.65)	**1.14** (0.63)
Grammar errors	0.01 (0.01)	0.01 (0.01)	0.01 (0.07)	0.01 (0.01)	0.02 (0.01)	0.02 (0.01)	0.02 (0.01)	0.02 (0.01)
Usage errors	0.01 (0.01)	0.02 (0.01)	0.01 (0.01)	0.01 (0.01)	0.02 (0.01)	0.02 (0.01)	0.02 (0.01)	0.02 (0.01)
Mechanics errors	0.03 (0.02)	**0.01** (0.01)	0.04 (0.02)	**0.02** (0.01)	0.04 (0.02)	**0.02** (0.01)	0.04 (0.03)	**0.03** (0.02)
Style errors	0.03 (0.02)	0.03 (0.02)	0.02 (0.03)	0.02 (0.02)	0.05 (0.05)	0.06 (0.05)	0.09 (0.08)	0.09 (0.08)
Median word frequency	58.14 (2.34)	57.88 (2.30)	56.08 (1.48)	56.05 (1.47)	57.99 (2.70)	57.75 (2.68)	57.76 (2.36)	57.53 (2.26)
Avg. Word length	5.00 (0.32)	4.89 (0.28)	5.10 (0.21)	5.01 (0.19)	4.98 (0.30)	4.88 (0.28)	4.85 (0.31)	4.76 (0.27)
Positive feature	0.61 (0.08)	0.62 (0.08)	0.61 (0.07)	0.62 (0.07)	0.60 (0.11)	0.61 (0.12)	0.62 (0.15)	0.64 (0.14)
Content feature1[d]			4.38 (0.48)	4.29 (0.50)	3.88 (0.67)	3.86 (0.66)	3.78 (1.03)	3.85 (1.06)
Content feature2[d]			0.20 (0.09)	0.20 (0.12)	0.10 (0.09)	0.10 (0.10)	0.09 (0.13)	0.10 (0.14)

Note. Bold face denotes scores that changed as a result of spelling correction.
[a]*N* = 90. [b]*N* = 100. [c]*N* = 123. [d]Included for the prompt specific model.

Table 26 Mean (*SD*) Score Differences Between Operational Human and e-rater and New e-rater Scores

Subgroup	e-rater vs. operational human	e-rater vs. new e-rater	New e-rater vs. operational human
China issue[a]	1.30 (0.31)	−0.23 (0.12)	1.53 (0.31)
China argument[b]	1.31 (0.28)	−0.06 (0.20)	1.37 (0.32)
Taiwan argument[c]	−0.20 (0.61)	−0.09 (0.20)	−0.11 (0.65)
African American argument[a]	−1.35 (0.42)	−0.14 (0.21)	−1.21 (0.47)

[a]*N* = 90. [b]*N* = 100. [c]*N* = 123.

Summary

Human raters and e-rater treat spelling errors differently in their approach to scoring. Human raters are instructed to ignore spelling errors, whereas e-rater penalizes for spelling errors as mechanics errors. We therefore attempted to modify e-rater to overlook the spelling errors by detecting and correcting such errors prior to scoring. However, the role of spelling errors as a source of the observed human and e-rater score differences was not clear.

Shell Detection

Measuring Shell Text

To evaluate presence of shell text as a possible source of human and e-rater discrepancy, we submitted the set of discrepant essays selected for SME review for processing through the ETS shell detection tool (Madnani et al., 2012). The tool detects organizational elements used to structure an argument or main point as shell. Use of shell text in a construct-irrelevant manner (lack of content or argumentation) simply to extend the length of the essay can lead to e-rater assigning higher scores to such essays than humans would on account of length. The average essay length for this group of essays was 441 words (*SD* = 213), and the average length or amount of shell text was 92 words (*SD* = 78), which equals roughly 21% of the average essay length. The correlation between the two lengths was .42.

Figure 6 displays the percentage of shell text (the amount of the total text identified as shell by the tool) across the 423 essays. A majority of the essays have 10% to 20% shell text as part of the essay. Few essays have shell text above 40%, and only some have 60% or more. Figure 7 displays the percentage of shell text across the essays within each subgroup. The percentage of shell text was relatively less (below 40%) for argument essays from the African American subgroup

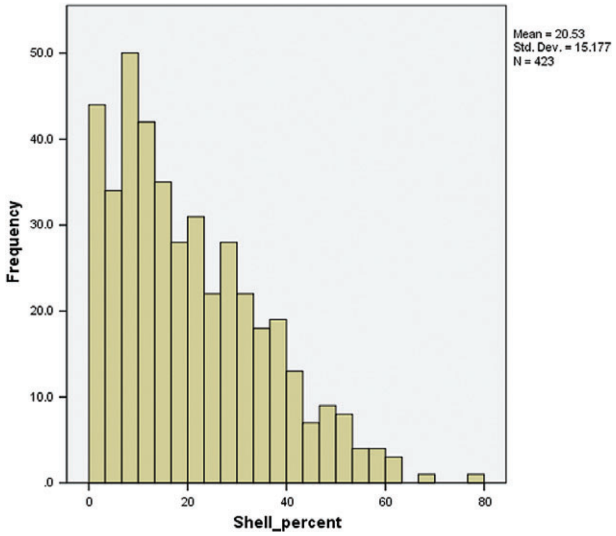

Figure 6 Percentage of shell text in the selected set of essays submitted for SME review.

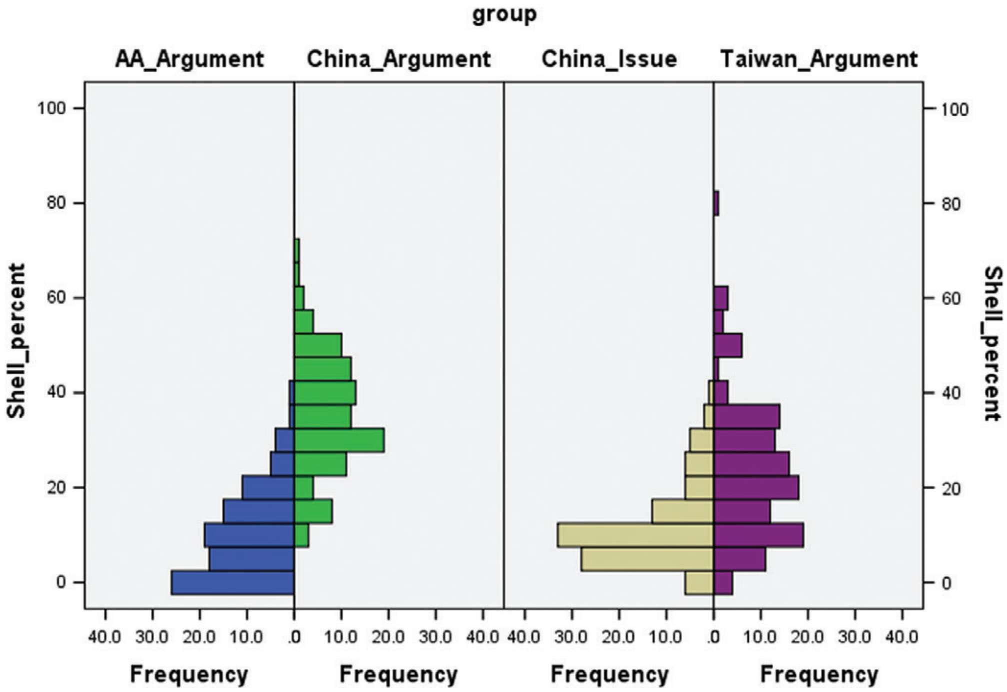

Figure 7 Percentage of shell text in the selected set of essays submitted for SME review by subgroup and prompt type.

and issue essays from the China subgroup. The percentage of shell text was more prevalent in argument essays from the China and Taiwan subgroups, with more than 40% shell text for approximately one third of the essays from China and for approximately one-fourth of the essays from Taiwan, with as much as 80% of the text identified as shell text for one or two essays from Taiwan.

Human Versus e-rater

Figures 8 and 9 display the distribution of percentage of shell text across human and e-rater scores, whereas Figure 10 displays the distribution of percentage of shell text across human and e-rater score differences. When looking at the distribution of amount of shell text across human scores, shell text appeared more prevalent in essays that received scores

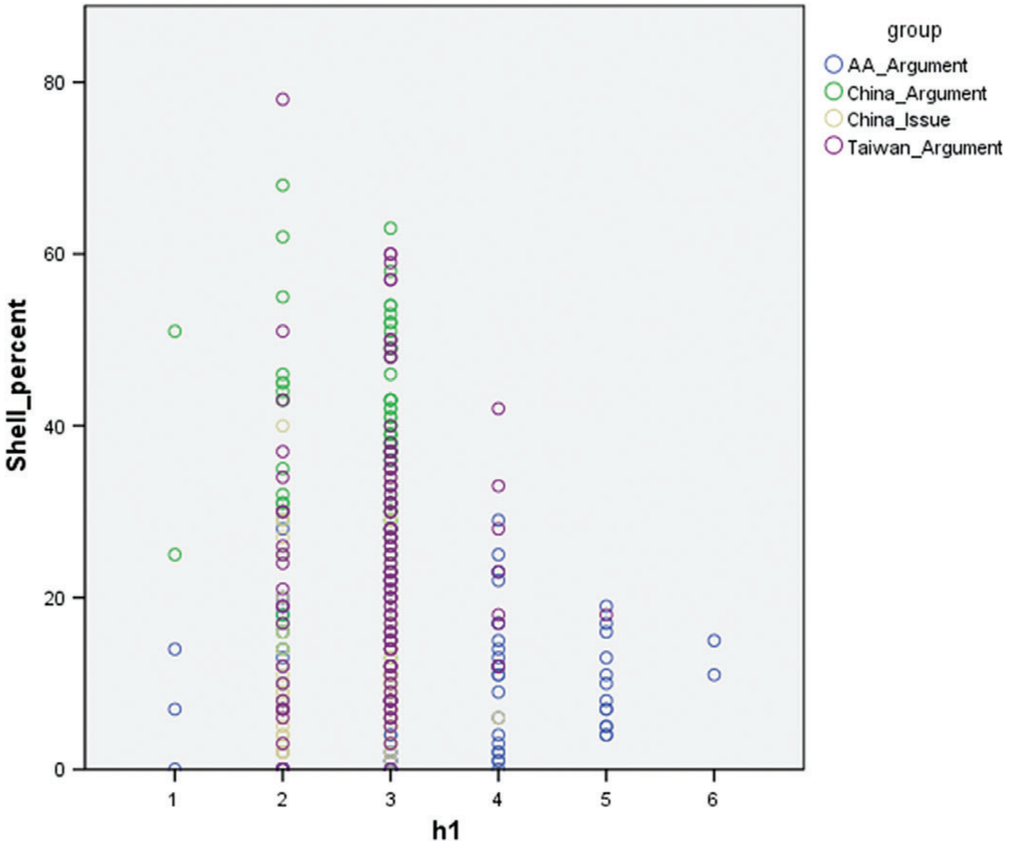

Figure 8 Distribution of percentages of shell text across human scores.

of 2 or 3. The percentage of shell text was limited in the upper half of the scale for this group of essays (40% or below for score 4 and 20% or below for scores 5 and 6). On the other hand, when looking at the distribution of shell text across e-rater scores, the amount of shell text was greater in essays that received a score of 4 (up to 70%) and 5 (up to 55%); e-rater did not assign a score of 6 to any essay in this set. A majority of the essays at these two score levels assigned by e-rater were written by examinees from China for both argument and issue; however, the percentage was higher for argument essays. We also examined the distribution of shell text across the e-rater and human score differences (at the .5 level) and found that the amount of shell text was more prevalent in essays that received higher scores from e-rater than from humans. Once again, the majority of these essays were written by examinees from China for both argument and issue; however, the percentage of shell text was higher for argument essays.

Summary

The results from the shell detection tool reveal some patterns regarding prevalence of shell text. The percentage of shell text was higher in essays at lower score levels from human raters and higher score levels from e-rater. These essays were mostly written by test takers from China, and the amount of shell text was relatively greater in responses for the argument task.

Conclusion

Different demographic groups have different characteristics in their writing. The variations may occur in the length of the response, the structure of the response, vocabulary use, and language use. The weights associated with different features that contribute to the final score can impact the scores of writers with unconventional styles. For example, short responses with simple vocabulary but relevant content may exhibit lower e-rater scores relative to human scores, while the inverse may be true for long responses with sophisticated vocabulary but that are low on content. The impact of other weighting schemes than those investigated in this study, such as feature reliability–based weights (under study by Attali, 2013),

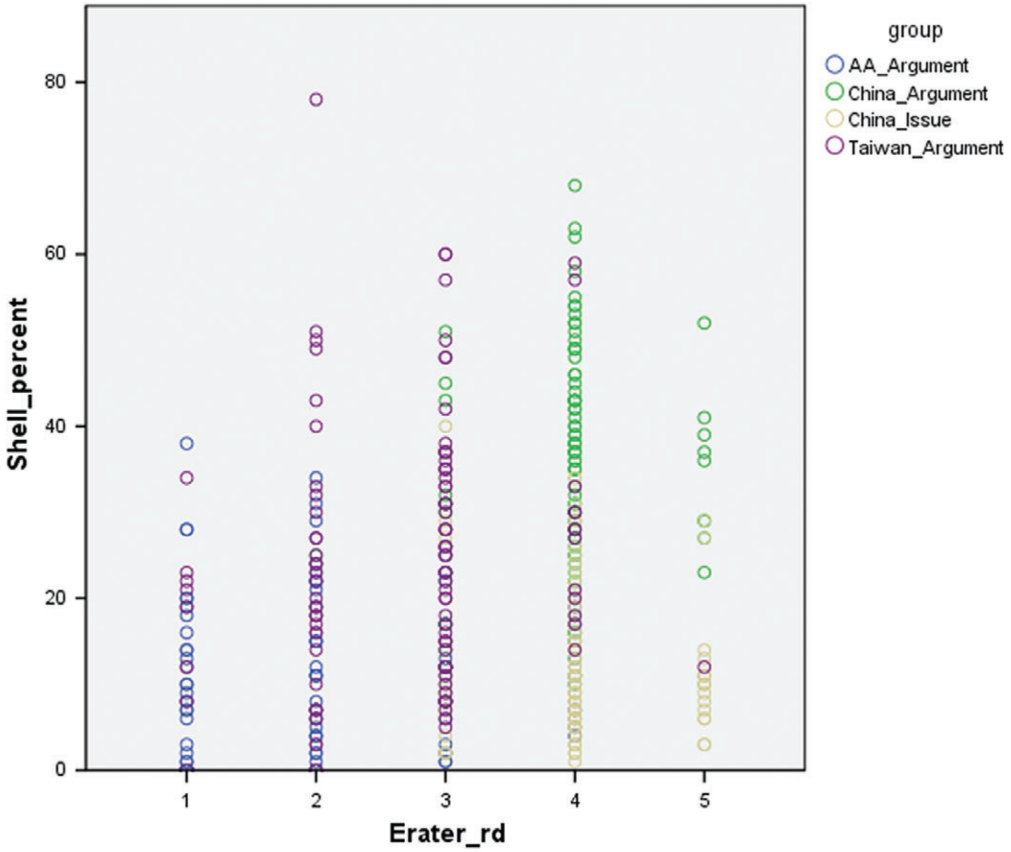

Figure 9 Distribution of percentages of shell text across e-rater scores.

in reducing the undue influence of a few features on the final score may potentially reduce the score differences at the subgroup level. Another potential avenue for investigation is the impact of variations in sample size for the training set as well as for the representation of the different demographic subgroups in the training sample (as studied by Zhang, Williamson, Breyer, & Trapani, 2012) on feature weights and consequent model performance at the subgroup level.

The alternate modeling procedures to linear regression were not effective in improving the overall performance of the scoring model. However, under certain models, such as the unit-weighted model and CART, mean score differences for certain subgroups and task types were reduced substantially. However, the differences were aggravated for other subgroups, thus offering only a partial solution to the problem on hand. Therefore the varying impact and effectiveness of different models across subgroups and task types observed in this study need to be more fully investigated for a potential solution.

Raters made several interesting observations during the qualitative review of these cases regarding the use of the scoring scale/rubric, the human rating process/procedures, and the e-rater scoring mechanism that help identify some differences between the objective e-rater and the holistic human scoring process guided by an analytic scoring scale. The scale structure and layout strongly influence the human rating process and do not fully correspond to the present e-rater scoring mechanism. The human raters appear to be using conditional logic and a rule-based approach to their scoring, whereas e-rater uses linear weighting of all the features. The raters treat the scale as two halves conditioned on language control and errors. Within each half, the scoring process is then guided by going down the bullet points in sequence in each score category to determine if the rules are met, followed by a check of the overall descriptor for the score category as a holistic check. The e-rater, in contrast, evaluates each essay for all 9 or 11 features and assigns weights to each of the features, with minimal penalization for language control and maximum credit for organization and development (influenced by essay length). From this human versus e-rater review, it appears that e-rater is not severe enough on language errors, overvaluing organization and development and occasionally undervaluing content. One approach to investigating the source of deviations in e-rater scoring is by conditioning the scores for organization and development (to avoid undue influence of

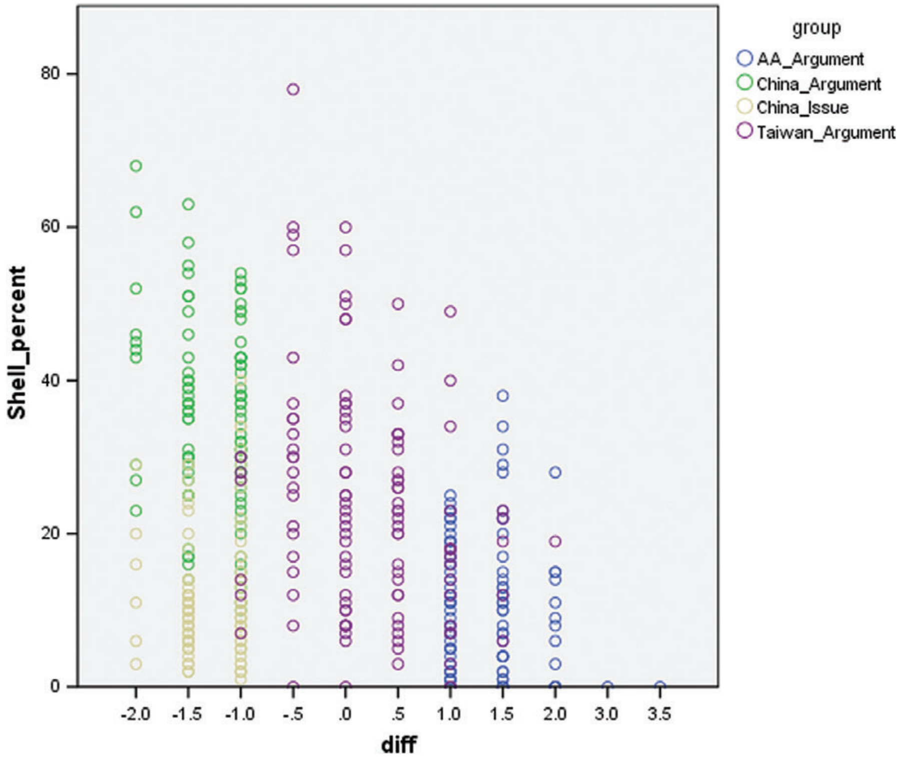

Figure 10 Distribution of percentages of shell text across human and e-rater score differences (differences calculated as h1 − e-rater).

essay length) or the overall score on language errors and content scores, for example, by setting a maximum for organization and development scores or the overall holistic score for responses that fail to meet a certain threshold for content scores or surpass the threshold for acceptable language errors. Such an investigation may lead to a potential solution for reducing the differences between the e-rater and human scores for groups of writers with certain writing styles.

For an objective comparison of human and e-rater scoring processes, we explored the statistical method of CART to observe the various paths or rules leading to each score category for e-rater. The comparison of a few select decision rules (reflective of the e-rater scoring approach) to the approach of human raters suggested that the importance assigned to the features by e-rater and humans is different.

The qualitative review of selected cases with scoring leaders revealed that the GRE program raters are instructed not to count or penalize for spelling errors while scoring the essay responses. The e-rater mechanics feature includes a count of spelling errors as one of the subfeatures contributing the most to the mechanics feature score. To address this anomaly in the human and e-rater scoring processes, the spelling error detection and correction module, which is under development at ETS, was used on a small and selective set of approximately 100 essays from each subgroup in its experimental version. The use of spelling correction did not mitigate all the mean score differences. The impact varied for different subgroups, reducing the differences for some and increasing the differences for others. These analyses were performed on very small sample sizes and therefore warrant caution in interpretation. The use of a spelling detection and correction module should be replicated and analyzed with larger samples to draw more reliable and valid conclusions about the impact of using the system in the e-rater scoring mechanism relative to the human scoring process.

Presently, human raters are trained in scoring to identify and treat any shell text neutrally while scoring the essay response. The e-rater lacks any such training presently and may be overscoring essays with the heavy presence of shell text. The processing of discrepant sets of essays through the shell detection tool confirmed the presence and prevalence of shell text in such essays and for particular demographic subgroups. Integration of a shell detection tool may help e-rater treat such text neutrally while scoring, similarly to human raters, thus improving the prediction of human scores. Also, when analyzing the feature weights, it was observed that the current scoring model and features do not explain a large amount of variance in the scores for the China subgroup. This further suggests that new features, such as percentage of shell text and features measuring more accurately the relevance of content, should be developed and evaluated to help improve the model fit for such groups. A new feature (e.g., for shell score) can be developed and included in the e-rater scoring model to improve prediction of human scores and therefore reduce human and e-rater mean score differences. However, a larger representative sample of essay responses will need to be processed through the shell detection tool to develop and evaluate the performance of such a feature in the scoring model.

The e-rater was not used in scoring during initial rGRE deployment. The e-rater was evaluated for rGRE based on human scores collected over a period of approximately 1 year since the introduction of the revised exam. On the basis of the evaluations, a check-score model was implemented. The evaluations revealed that human and e-rater score differences persisted for some demographic subgroups, such as China. Although e-rater continues to be successfully used in operational implementation without being detrimental to the reliability and validity of writing scores, the topic of this study remains an important area of research to enhance the quality of e-rater scoring models.

Notes

[1] This study is conducted on data from *prior* GRE; a new revised GRE test was launched in August 2011.

[2] This approach would not be feasible in operational environment and was undertaken here purely in research context.

References

Attali, Y. (2007). *Construct validity of e-rater in scoring TOEFL essays* (Research Report No. RR-07-21). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2007.tb02063.x

Attali, Y. (2008). *e-rater evaluation for TOEFL iBT independent essays*. Unpublished report.

Attali, Y. (2009). *Interim summary of analyses related to content scoring of TOEFL integrated essays*. Unpublished report.

Attali, Y. (2013). *Alternative bases for AES*. Unpublished manuscript.

Attali, Y., Bridgeman, B., & Trapani, C. (2007). *e-rater performance for GRE essays*. Unpublished report.

Attali, Y., & Burstein, J. (2006). Automated essay scoring with e-rater v. 2. *Journal of Technology, Learning, and Assessment, 4*(3), 1–30.

Bennett, R., & Bejar, I. (1997). *Validity and automated scoring: It's not only the scoring* (Research Report No. RR-97-13). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.1997.tb01734.x

Braun, H. I. (1988). Understanding scoring reliability: Experiments in calibrating essay readers. *Journal of Educational Statistics, 13*, 1–18.

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). *Classification and regression trees*. Belmont, CA: Wadsworth.

Bridgeman, B., Trapani, C., & Attali, Y. (2012). Comparison of human and machine scoring of essays: Differences by gender, ethnicity, and country. *Applied Measurement in Education, 25*(1), 27–40.

Burstein, J., & Chodorow, M. (1999). Automated essay scoring for nonnative English speakers. In *Proceedings of the Workshop on Computer-Mediated Language Assessment and Evaluation of Natural Language Processing* (pp. 68–75). College Park, MD: Association of Computational Linguistics and International Association of Language Learning Technologies.

Chodorow, M., & Burstein, J. (2004). *Beyond essay length: Evaluating e-rater's performance on TOEFL essays* (TOEFL Research Report No. TOEFL-RR-73). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2004.tb01931.x

Culham, R. (2003). *6 + 1 traits of writing: The complete guide*. New York, NY: Scholastic.

Daly, J. A., & Dickson-Markman, F. (1982). Contrast effects in evaluating essays. *Journal of Educational Measurement, 19*, 309–316.

Dawes, R. M., & Corrigan, B. (1974). Linear models in decision making. *Psychological Bulletin, 81*, 9S–106.

Dorans, N., & Drasgow, F. (1978). Alternative weighting schemes for linear prediction. *Organizational Behavior and Human Performance, 21*, 316–345.

Feng, X., Dorans, N. J., Patsula, L. N., & Kaplan, B. (2003). *Improving the statistical aspects of e-rater R: Exploring alternative feature reduction and combination rules* (Research Report No. RR-03-15). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2003.tb01907.x

Flor, M., & Futagi, Y. (2012, June). On using context for automatic correction of non-word misspellings in student essays. In *Proceedings of the Seventh Workshop on Building Educational Applications using NLP* (pp.105–115). Montreal, CA: Association for Computational Linguistics.

Haberman, S. J. (2007). Electronic essay grading. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics* (Vol. *26*, pp. 205–233). Amsterdam, Netherlands: North-Holland.

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.

Hales, L. W., & Tokar, E. (1975). The effect of the quality of preceding responses on the grades assigned to subsequent responses to an essay question. *Journal of Educational Measurement, 12*, 115–117.

Harik, P., Baldwin, P., & Clauser, B. E. (2013). Comparison of automated scoring methods for a computerized performance assessment of clinical judgment. *Applied Psychological Measurement, 37,* 587 – 597.

Hughes, D. C., & Keeling, B. (1984). The use of model essays to reduce context effects in essay scoring. *Journal of Educational Measurement, 21*, 277 – 281.

Hughes, D. C., Keeling, B., & Tuck, B. F. (1980a). Essay marking and the context problem. *Educational Research, 22*, 147 – 148.

Hughes, D. C., Keeling, B., & Tuck, B. F. (1980b). The influence of context position and scoring method on essay scoring. *Journal of Educational Measurement, 17*, 131 – 135.

Hughes, D. C., Keeling, B., & Tuck, B. F. (1983). The effects of instructions to scorers intended to reduce context effects in essay scoring. *Educational and Psychological Measurement, 43*, 1047 – 1050.

Lunz, M. E., Wright, B. D., & Linacre, J. M. (1990). Measuring the impact of judge severity on examination scores. *Applied Measurement in Education, 3,* 331 – 345.

Madnani, N., Heilman, M., Tetreault, J., & Chodorow, M. (2012, June). Identifying high level organizational elements in argumentative discourse. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 20 – 28). Montreal, CA: Association of Computational Linguistics.

Ramineni, C. (2012). Validating automated essay scoring for online writing placement. *Assessing Writing, 18*(1), 25 – 39.

Ramineni, C., Davey, T., & Weng, V. (2010). *Statistical evaluation and integration of a new positive feature for e-rater v10.1*. Unpublished report.

Ramineni, C., Trapani, C., Williamson, D., Davey, T., & Bridgeman, B. (2012a). *Evaluation of e-rater for the GRE issue and argument prompts* (Research Report No. RR-12-02). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2012.tb02284.x

Ramineni, C., Trapani, C., Williamson, D., Davey, T., & Bridgeman, B. (2012b). *Evaluation of e-rater for the TOEFL independent and integrated prompts* (Research Report No. RR-12-06). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2012.tb02288.x

Sinharay, S., Davey, T., Weng, V., & Ramineni, C. (2009). *e-rater invariance for TOEFL essays*. Unpublished report.

Spear, M. (1997). The influence of contrast effects upon teachers' marks. *Educational Research, 39*, 229 – 233.

Stalnaker, J. M. (1936). The problem of the English examination. *Educational Record, 17*, 41.

Wainer, H. (1976). Estimating coefficients in linear models: It don't make no nevermind. *Psychological Bulletin, 83*, 312 – 317.

Williamson, D. M., Bejar, I., & Sax, A. (2004). *Automated tools for subject matter expert evaluation of automated scoring* (Research Report No. RR-04-14). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2004.tb01941.x

Williamson, D. M., Xi, X., & Breyer, F. J. (2012). A framework for evaluation and use of automated scoring. *Educational Measurement: Issues and Practices, 31*(1), 2 – 13.

Zechner, K., Higgins, D., Xi, X., & Williamson, D. (2009). Automatic scoring of non-native spontaneous speech in tests of spoken English. *Speech Communication, 51*, 883 – 895.

Zhang, M., Williamson, D. M., Breyer, F., & Trapani, C. (2012). Comparison of e-rater automated essay scoring model calibration methods based on distributional targets. *International Journal of Testing, 12*, 345 – 364.

Suggested citation:

Ramineni, C., & Williamson, D. (2018). *Understanding mean score differences between e-rater® automated scoring engine and humans for demographically based groups in the GRE® general test* (GRE Board Research Report No. GRE-18-01). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/ets2.12192

Action Editor: Brent Bridgeman

Reviewers: This report was reviewed by the GRE Technical Advisory Committee and the Research Committee and Diversity, Equity and Inclusion Committee of the GRE Board.