Listening. Learning. Leading.®

Research Report
ETS RR–14-24

A Study of the Use of the *e-rater*® Scoring Engine for the Analytical Writing Measure of the *GRE*® revised General Test

F. Jay Breyer
Yigal Attali
David M. Williamson
Laura Ridolfi-McCulla
Chaitanya Ramineni
Matthew Duchnowski
April Harris

December 2014

ETS Research Report Series

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

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

ETS Research Report Series ISSN 2330-8516

RESEARCH REPORT

A Study of the Use of the *e-rater*® Scoring Engine for the Analytical Writing Measure of the *GRE*® revised General Test

F. Jay Breyer, Yigal Attali, David M. Williamson, Laura Ridolfi-McCulla, Chaitanya Ramineni, Matthew Duchnowski, & April Harris

Educational Testing Service, Princeton, NJ

In this research, we investigated the feasibility of implementing the *e-rater*® scoring engine as a check score in place of all-human scoring for the *Graduate Record Examinations*® (*GRE*®) revised General Test (rGRE) Analytical Writing measure. This report provides the scientific basis for the use of e-rater as a check score in operational practice. We proceeded with the investigation in four phases. In phase I, for both *argument* and *issue* prompts, we investigated the quality of human scoring consistency across individual prompts, as well as two groups of prompts organized into sets. The sets were composed of prompts with separate focused questions (i.e., *variants*) that must be addressed by the writer in the process of responding to the topic of the prompt. There are also groups of variants of prompts (i.e., grouped for scoring purposes by similar variants). Results showed adequate human scoring quality for model building and evaluation. In phase II, we investigated eight different e-rater model variations each for argument and issue essays including prompt-specific; variant-specific; variant-group – specific; and generic models both with and without content features at the rating level, at the task score level, and at the writing score level. Results showed the generic model was a valued alternative to the prompt-specific, variant-specific, and variant-group – specific models, with and without the content features. In phase III, we evaluated the e-rater models on a recently tested group from the spring of 2012 (between March 18, 2012, to June 18, 2012) following the introduction of scoring benchmarks. Results confirmed the feasibility of using a generic model at the rating and task score level and at the writing score level, demonstrating reliable cross-task correlations, as well as divergent and convergent validity. In phase IV of the study, we purposely introduced a bias to simulate the effects of training the model on a potentially less able group of test takers in the spring of 2012. Results showed that use of the check-score model increased the need for adjudications between 5% and 8%, yet the increase in bias actually increased the agreement of the scores at the analytical writing score level with all-human scoring.

Keywords Automated essay scoring; e-rater; rGRE analytical writing; check scoring

doi:10.1002/ets2.12022

Previous *GRE*® Analytical Writing research by Ramineni, Trapani, Williamson, Davey, and Bridgeman (2012) showed adequate performance of the *e-rater*® scoring engine as a check score with an adjudication threshold of ±0.5. A check-score implementation model consists of the first human score as the sole score, if confirmed by e-rater, that is, when the unrounded e-rater and the first human are within ±0.5 points of each other. Otherwise, additional humans are required to score the essay. If an additional human is required to score the essay, in the case of a discrepancy larger than ±0.5, then a second human score is required. In any case, however, the e-rater score does not contribute to the analytical writing score.

The GRE Analytical Writing measure consists of two tasks: (a) analyze an argument and (b) analyze an issue. The argument task requires the test taker to evaluate a stated argument presented in the prompt by developing supporting evidence and reasoning. The issue task requires the test taker to develop and support a position on an issue provided in the prompt.

Previous work by Ramineni et al. (2012) was based on more than 750,000 operational responses on 113 issue prompts and 139 argument prompts that were administered between September 2006 and September 2007. These researchers built and evaluated prompt-specific, generic, and generic with prompt-specific intercept e-rater scoring models for the argument and issue scores by prompt, test center country, gender, and ethnic subgroups, and they evaluated correlations with other GRE section scores and simulated GRE Analytical Writing measure scores under each model scenario and various adjudication threshold levels. Results showed that the issue task was best scored with a generic e-rater scoring

Corresponding author: F. J. Breyer, E-mail: fbreyer@ets.org

model with a unique operational prompt-specific intercept and that the argument task required a prompt-specific e-rater scoring model for each argument prompt. Scoring was thus implemented with these recommended models using e-rater as a check score.

Beginning in August 2011, a new GRE revised General Test (rGRE) was launched, and the use of e-rater was discontinued due to changes in the rGRE Analytical Writing measure. These changes necessitated a new study exploring the application of e-rater to the rGRE Analytical Writing prompts. This study provides the scientific basis for implementing the use of e-rater version 12.1 as a check score for the rGRE Analytical Writing measure. Further, the current operational model is different from the e-rater model evaluated in this report.

The GRE Revised General Test

The rGRE Analytical Writing measure is designed to measure critical thinking and analytical writing skills that provide evidence of writing to express and support complex ideas clearly and effectively (Educational Testing Service, 2013)—skills that are necessary for graduate and business school success. The new rGRE Analytical Writing prompts are different from the argument and issue prompts in the original GRE Analytical Writing measure in that they ask a focused question that addresses a specific aspect of the prompt. These focused questions within the argument and issue prompts, known as *variants*, are an added feature of the rGRE Analytical Writing measure. For example, in one variant, the test taker is asked to describe the unstated assumptions in an argument and, in a second variant, the test taker is asked to evaluate a recommendation in the argument; prompts on the same topic can have more than one variant. Variants are organized into *variant groups* for human scoring purposes. The variants are grouped by the assessment developers for ease of scoring, given that the variants within a group do not require scorer recalibration, whereas scoring variants from different variant groups do require recalibration (Joe, Balani, Napoli, & Chen, 2013). An rGRE Analytical Writing test taker receives one variant argument prompt and one variant issue prompt in the rGRE Analytical Writing measure. Given the past successful use of automated scoring for the GRE Analytical Writing measure and the value of e-rater as a check score, investigation of e-rater scoring the rGRE Analytical Writing prompts was undertaken.

Study Purpose and Research Questions

The purpose of this study was to evaluate the feasibility of using e-rater (Burstein, Tetreault, & Madnani, 2013) as a *check-score* model (i.e., to confirm the first human score or request additional human scores) for the rGRE issue and argument Analytical Writing prompts. Specifically, in this study, we investigated whether the use of automated scoring for operational scoring of the rGRE Analytical Writing measure could lower the number of human raters needed, effectively reducing program costs while maintaining score quality and fairness, thus ensuring fast and consistent score production for the large number of test takers worldwide who each year take the test. In this study, we investigated the following research questions that we associate with the different phases of this research project:

1. Are the different prompts, prompt variants, and prompt variant groups similar in scoring consistency among human scorers?
2. Can a generic (G) model function well for all prompts as a check score?
    a. Is there a benefit of using a prompt-specific (PS) model, a variant-specific (VS) model, or a variant-group–specific (VG) model as a check score compared to a G model?
    b. Is there a benefit to using the content features in each of the four models (i.e., G, PS, VS, and VG) in terms of evaluation results?
3. If a generic model is used as a check score, what effect will the writing proficiency of the reference group, on which that model was trained and evaluated, have in terms of human-machine score separation and general evaluation acceptance criteria (see Williamson, Xi, & Breyer, 2012)?
4. What is the effect of using generic models trained on one test-taker cohort (i.e., those tested from March 18, 2012, to June 18, 2012) that may be of lower ability, and cross-evaluating them on another sample that may be of higher ability (i.e., those tested from August 1, 2011, to March 17, 2012)?
5. What is the effect at the human-machine rating level of using a biased automated essay scoring (AES) model (i.e., a model that produces consistently lower scores compared to human scores)?

6. What is the effect on the task score of using an automated scoring model that is biased?
7. What is the effect on the writing score if the automated score is biased?

Scoring the GRE Revised General Test Analytical Writing Tasks

In the human scoring process for the rGRE Analytical Writing measure, written essay responses are distributed to trained raters who assign a score to each essay using a 6-point holistic scoring rubric; these rubrics are provided in Appendix A. The scoring rubric reflects the quality of an essay in response to the specific variant of the argument or issue task. The test taker must address the specific variant in order to receive a score in the upper half of the 6-point holistic scale.

Each essay receives a score from two trained raters, using the 6-point holistic scoring rubric. In holistic scoring, readers are trained to assign scores based on the overall quality of an essay in response to the assigned variant in the task. If the two scores differ by more than 1 point on the scale, the discrepancy is adjudicated by a third GRE rater. Otherwise, the scores on each essay are averaged. The final scores on the two essay tasks are then averaged and rounded to the nearest half-point interval on the 0–6 score scale (e.g., 0, 0.5, 1.0, 1.5, 2.0, 2.5 … 5.0, 5.5, and 6.0). Scores of 0 are assigned to responses that are blank or off-topic or that do not address the prompt (i.e., *bad faith* responses).

After scores for the issue and argument prompts have been obtained separately, the two final scores from the two essay prompts are averaged and rounded *up* to the nearest half-point interval. This results in a single writing *task* score, which is reported for the test taker's performance on the Analytical Writing measure. If the test taker wrote an essay for only one of the two tasks, that test taker receives a score of 0 on the task for which no response is provided, whereas if a test taker did not write to either of the two tasks, an NS (no score) is reported for the Analytical Writing measure.

Sample 1: Initial Cohort

Test takers who took the rGRE examination from August 1, 2011, to March 17, 2012, are referred to as the *initial cohort* in this report. Those test takers tested between August 1, 2011, and November 8, 2011, were scored on a delayed basis using the rGRE Analytical Writing rubrics to permit adequate vetting of the scores and benchmarking of the scale points. These rubrics, addressing the variant focused questions, and benchmarked to the six-point scale, were used to report operational scores up to and including those test takers tested on March 17, 2012.

Sample 2: Retrained Cohort

In the latter part of March of 2012, the human raters were retrained using the revised GRE Analytical Writing rubrics after there were sufficient examples of benchmark scores referencing each of the 6-score points for each task. The scoring rubrics themselves did not change, but rather the changes were in the benchmark papers used in the training and calibration efforts. The raters were subsequently evaluated for quality improvement and retention purposes. A second test-taker cohort was scored by the retrained human scorers starting March 18, 2012, to June 18, 2012, when data from the retrained set of test takers, referred to as the *retrained data set*, were gathered for model building and evaluation purposes.

Sample 3: Trend Cohort

Following retraining, we randomly selected a third sample of 1,800 test takers who were originally tested during the November 9, 2011, and November 19, 2011, computer-based testing windows and rescored following retraining. This third group of test takers serves as a bridge between the writing ability of the retrained group from the spring of 2012 (i.e., thought to be low) and the writing proficiency of the initial cohort from the fall and winter of 2011–2012 (i.e., thought to be high).

Automated Scoring with e-rater

Educational Testing Service (ETS) uses an AES system, e-rater, that employs natural language processing (NLP) tools to extract evidence of writing proficiency from electronic text that is used to generate a number of *microfeatures* that are combined into *macrofeatures* (Attali & Burstein, 2006; Burstein et al., 2013).

Features

Macrofeatures Used in Generic Models

The set of features derived in e-rater v2 (Attali & Burstein, 2006) enabled use of a single scoring model across multiple prompts, referred to as a generic (G) model (Attali, Bridgeman, & Trapani, 2010b). G models are calibrated on a group of related prompts from one task, typically 10 or more prompts. A nonnegative least squares regression model is developed across all prompts so that the resultant model is the best fit for predicting human scores for all the prompts in that task, taken as a whole. As such, a common set of macrofeature weights and a single intercept are used for all prompts regardless of the particular prompt in the set of prompts for a task (i.e., argument or issue). Prior to this research, G models have not taken into account the content of the essay and address only word- and sentence-level conventions, fluency of expression, word choice, organization, and development. Content features related to the vocabulary usage are prompt specific and, therefore, are not included in the regression for most applications to date. The generic modeling approach has the advantage of requiring smaller sample sizes per prompt for model training and evaluation, while at the same time providing the opportunity to place new prompts into operational service without having to rebuild and evaluate a new model.

The microfeatures that are extracted from the text are related to writing aspects rather than the content discussed in the essay. The e-rater AES system employed in this research (i.e., e-rater version 12.1) produced 10 macrofeatures based on the microfeatures that are outlined in bold in the figure presented in Appendix B. The 10 macrofeature (and linguistic examples) adapted from Burstein, Tetreault, and Madnani (2013) are:

1. Grammar (e.g., subject – verb agreement)
2. Usage (e.g., then vs. than)
3. Mechanics (spelling and capitalization)
4. Style (e.g., repetitive phrases and passive voice)
5. Organization (e.g., thesis statement, main points, supporting details, conclusions)
6. Development (e.g., main points precede details)
7. Positive features:
    a. Correct preposition usage (the probability of using the correct preposition in a phrase)
    b. Good collocation use (i.e., collocations occur when two contiguous words appear together more often in language use than other pairs of words, such as the pairing of *tall* trees and *high* mountains as opposed to *high* trees and *tall* mountains)
    c. Sentence variety (i.e., the ability to use correct phrasing and a variety of grammatical structures)
8. Lexical complexity with average word length (i.e., the use of vocabulary with different counts of letters)
9. Lexical complexity with sophistication of word choice (i.e., the use of sophisticated vocabulary)
10. Differential word usage (e.g., this feature weights words given an empirical association with high-scoring or low-scoring essays). It should be noted that this feature is not employed in current e-rater models but was evaluated and used when e-rater was first implemented in August 2012 as a check score.

The macrofeatures are combinations of fine-grained microfeatures, shown as smaller connecting nodes in Appendix B, as well as other microfeatures not shown here. (Appendix C contains a detailed glossary of microfeature names and associated descriptions that are employed in e-rater engine v12.1.) These features are used to derive a prediction of the human score in a nonnegative least squares multiple-linear regression model where the human scores are regressed on the macrofeatures extracted by the scoring engine.

Macrofeatures Used in Prompt-Specific Models

PS models are custom-built models for each prompt in the question pool. They are designed to provide the best fit models for the particular prompt in question, with both the macrofeature weights and the intercept customized for the human score distribution used to calibrate the prompt model. Prompt-specific models typically incorporate the two PS vocabulary-related content macrofeatures into the scoring. However, in this study, we investigated the inclusion of these two content macrofeatures, content vector analysis (CVA) to determine if they add value to the human-machine agreement in any scoring model (i.e., G, PS, VS, VG).

Implementation/Score Usage Formats

Currently, ETS uses one of three score usage formats when deploying e-rater in operational testing programs. For low-stakes tests, such as practice tests, automated scores can be used by themselves (i.e., as the sole reported score). For high-stakes tests, two deployment formats are in use: (a) use as a check on the human score and (b) use as contributing to part of the score in some ratio of human score and automated score. Typically, e-rater and human scores are equally weighted when used as a contributory score. For this study, we investigated the use of e-rater in a check score usage format as it had been implemented in past research (Ramineni et al., 2012).

Development of e-rater Scoring Models

Developing e-rater scoring models is typically a two-stage process: (a) model training/building and (b) model evaluation. Data are split into a randomly selected model-building set and 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. A properly constituted set of training essays includes a random sample of responses that must have been written on the computer and should be representative of the population for which e-rater scores are intended for use. Prior to model building, the selected essay set is subjected to advisory flag analyses.

Advisory Flag Analysis

A number of advisory flags (acting as filters) have been established that indicate when a specific essay is inappropriate for automated scoring. Each advisory flag marks a different problem that would result in an essay being identified as inappropriate for automated scoring. Essays that consist of too few words (i.e., < 25) and too few sentences (i.e., < 2) or that are excessively long (i.e., more than 1,000 words or 65,536 characters) are automatically issued a fatal advisory; experience has shown such essays receive erroneous automated scores. Whenever new models are trained and evaluated, the use of other fatal flags for an assessment is analyzed by comparing when e-rater classifies an essay with a fatal advisory flag other than too brief or too long versus when a human rater assigns a 0 score. All advisories are evaluated individually as well as combined. That is, individual advisories for which e-rater is found to effectively (on par with humans) identify essays that are inappropriate for automated scoring are combined sequentially and subjected to a similar evaluation against human markings. This process of advisory flag analyses helps determine which group of advisories aid e-rater in effectively screening for inappropriate essays and should be included as part of the operational e-rater framework for an assessment. Subjecting the sample of essays to advisory flagging prior to model building improves the quality of model building by filtering out the inappropriate essays from going into the model build phase.

Advisory flags for e-rater are coded depending on the type of issue(s) identified. Table 1 lists the names, a brief description, and binary codes for all the advisory flags. An essay can be flagged for single or multiple issues. For instance, if an essay contains repetition of words, the flag will be set to 2 (reuse of language). However, if an essay contains repetition of words and is not relevant to the assigned topic, the flag will be set to 10, that is, 2 (reuse of language) + 8 (not relevant). Flags 64 (too brief) and higher result in the engine assigning a score of 0, while the other flags are provided as warnings. Advisory flags that force a score of 0 are referred to as fatal advisories in this report. Thus, there are fatal and nonfatal advisories.

If no fatal advisory flags that would preclude automated scoring have been issued and no human scores of 0 have been assigned, then essays are sent to the e-rater program in operational practice. In this study, only essays with nonzero scores and without *any* advisory flags, either fatal or nonfatal, are included in the randomly selected model build set. In model building, e-rater extracts evidence in the form of macrofeatures, including grammar, usage, mechanics, organization, development, and others (see Appendix B for the features extracted by e-rater v12.1). After the macrofeature values are derived from the essay text, the weights for the features are determined using a *multiple-linear regression* (MLR) procedure. These regression weights can then be applied to additional essays to produce a predicted score based on the calibrated feature weights.

Because the feature weights are selected and estimated to maximize agreement with human scores, any evaluation based on the training sample will tend to overstate a scoring model's performance. However, a more appropriate measure of performance can be obtained by applying the model to the independent evaluation sample. The evaluation data set is

Table 1 Advisory Flag Code, Name, and Description

Flag code	Fatal vs. nonfatal	Flag name	Flag description
2	Nonfatal	Reuse of language	Compared to other essays written on this topic, the essay contains more reuse of language, a possible indication that it contains sentences or paragraphs that are repeated.
4	Nonfatal	Key concepts	Compared to other essays written on this topic, the essay shows less development of the key concepts on this topic.
8	Nonfatal	Not relevant	The essay might not be relevant to the assigned topic.
16	Nonfatal	Restatement	The essay appears to be a restatement of the topic with few additional concepts.
32	Nonfatal	No resemblance	The essay does not resemble others that have been written on this topic, a possible indication that it is about something else or is not relevant to the issues the topic raises.
64	Fatal	Too brief	The essay is too brief to evaluate.
128	Fatal	Excessive length	The essay is longer than essays that can be accurately scored and must be within the word limit to receive a score.
256	Fatal	Unidentifiable organizational elements	The essay could not be scored because some of its organizational elements could not be identified.
512	Fatal	Excessive number of problems	The essay could not be scored because too many problems in grammar, usage, mechanics, and style were identified.
1024[a]	N/A	Unexpected topic	The essay appears to be on a subject that is different from the assigned topic.
2048[a]	N/A	Non-essay	The text submitted does not appear to be an essay.

[a]Not applicable to the GRE program.

randomly selected as well. Subsequently, the feature scores and weights are applied to samples of essays in the evaluation set to produce an overall e-rater score and validate the model performance. In general, model performance will appear slightly degraded in this sample in comparison to the training sample. Models are evaluated and recommended for operational use if the results of automated scoring are comparable in agreement between two human raters.

Evaluation Criteria

Following development of the automated scoring models and production of AES scores, standard evaluation criteria (see Williamson et al., 2012) are used to assess the overall quality of the scoring models and to compare the prompt-specific models with the generic models on the independent evaluation data set as a general practice. (See Appendix D for threshold flagging criteria for standard evaluations that we used in this study.) This independent evaluation process provides an independent examination of the usefulness of the AES models as applied to another comparable test-taker sample and thus provides evidence of the validity of the scoring model. The evaluation criteria are as follows.

Construct Evaluation

Does the construct representation of e-rater reflect the relative contributions of related cognitive processing components that are critical for the measured writing construct? Automated scoring capabilities are designed with the conceptualization for which tasks they are useful. Thus, the initial evaluation involves the examination of the intent and goals of the tasks, the kinds of skills and knowledge the tasks are intended to elicit, and the scoring rubrics. This gathered construct evidence is then paired with the constructs measured by the AES model. For the revised GRE Analytical Writing measure, this process of construct-relevance appraisal involved examination of the scoring rubrics, discussion with assessment developers of the goals of the assessment tasks, discussion of the score reporting goals, and review of the score claims.

Agreement of Human Scores

Does the agreement between human and machine scores meet minimum threshold guidelines? Agreement indices calculated during the evaluation process typically fall into descriptive measures that can be either scale dependent or partially scale independent. Exact percentage agreement rates are considered scale dependent because higher degrees of agreement are easier to attain for shorter scales (e.g., 1 – 3 points) compared to longer scales (e.g., 1 – 6 points). Thus, more scale-independent measures of agreement that use product-moment correlations and quadratic-weighted kappa (QWK) are employed for human-machine agreement quantification purposes with a threshold guideline value of 0.70 in this research. This value is selected because it captures the common variance of close to 50% that is shared between human and machine scores. It is useful to note that correlations are calculated on unrounded machine score values, while QWKs are calculated using rounded machine scores. The two indices, correlation and QWK, will agree only if the marginal distributions are the same and the correlations between the rounded and unrounded scores with humans are the same (Fleiss & Cohen, 1973).

Degradation

Is the difference between automated-human score agreement and human-human agreement below a predefined threshold? Recognition of the inherent relationship between the consistency of double-human scoring and the consistency of automated scoring with human scoring has resulted in another criterion of performance in relationship: degradation. Specifically, the degradation criterion requires that the automated-human scoring agreement cannot be more than 0.10 *lower*, in either weighted kappa or correlation, than the human-human agreement. This criterion prevents circumstances in which automated scoring may reach the 0.70 threshold but still be notably deficient in comparison with human scoring. It should be noted that, in practice, we have observed cases in which the automated-human agreement for a particular task has been slightly less than the 0.70 performance threshold, but very close to a borderline performance for human scoring (e.g. an automated-human weighted kappa of 0.68 and a human-human kappa of 0.71), and have approved such models for operational use on the basis of being highly similar to human scoring and consistent with the purpose of the assessment within which they are used. Similarly, it is relatively common to observe automated-human absolute agreements that are *higher* than the human – human agreements for tasks that primarily target control of language, fluency, and vocabulary in addition to requiring additional evidence, such as GRE issue and *TOEFL*® independent tasks (Attali, 2009; Bridgeman, Trapani, & Attali, 2009). As a result of this observation, we have ensured that positive *degradation* indicates higher machine – human agreement than human – human agreement. Conversely, negative degradation indicates a decrease in the human – machine agreement compared to the human – human agreement.

Standardized Mean Score Difference

Is the standardized mean score difference below a predefined threshold? Another criterion for association of automated scores with human scores is that the standardized difference (using the standard deviation of the distribution of human scores) between the human scores and the automated scores cannot exceed 0.15.[1] This measure ensures that the mean of the automated scores is near that of the human scores, and it is effective in determining if a systematic bias exists between automated and human scoring. If there is solely a single prompt, obtaining a 0 standardized difference is trivial with a generic model; a potential problem exists when a generic model is used across multiple prompts.

Association With Other Scores

What is the association of the AES with other scores? The evaluation of automated scoring against the performance of human scoring is a typical criterion. High-quality human scoring is thought to be the best alternative to automated scoring (or vice versa) and is the basis for building the statistical models for scoring within the automated systems. However, the problems and concerns with human scoring are well documented and represent a range of potential pitfalls, including halo effects, fatigue, tendency to overlook details, and problems with consistency of scoring across time. Today, automated scores are thought to be complementary *and* different from human scores (Attali, 2013). As a result, it is relevant to

investigate not just the consistency with human scores, but to also evaluate the patterns of relationship of automated scores, compared to their human counterparts, with external criteria (Attali et al., 2010; Petersen 1997; Powers, Burstein, Chodorow, Fowles, & Kukich, 2002; Weigle 2010). Such an expectation is certainly not new, as it is simply an extension of a classic question of validity (see Campbell & Fiske, 1959), but with the focus on not just establishing a relationship between a score and some independent criterion, but on interpretation of potential differences in this relationship as indicative of relative merits or weaknesses of automated and human scoring.

These independent variables may be scores on other sections of the same test or external variables that measure similar (i.e., related) or different (i.e., divergent) constructs. If human and automated scores reflect similar constructs, they are expected to relate to other measures of similar or distinct constructs in similar ways:

- *Within test relationships.* Are automated scores related to scores on other sections of the test in ways that converge and diverge similar to human scores?
- *Relationship at the task type and reported score level.* Are the relationships between the essay score and external measures similar at the task type and reported score level?

Subgroup Differences

Is it fair to use a machine score as a check score in place of a second human score and only use the first human score for different subgroups if there is substantial separation between human and machine scores? For this subgroup evaluation, we reduce the flagging guideline threshold from 0.15 to 0.10 for standardized difference in order to identify patterns of systematic differences in the distributions of automated and human scores at the task and reported score level.

Operational Impact Analysis

What is the impact of using an automated score as a check on the first human score instead of consistently using a second human score? The final analysis in the evaluation of implementing automated scoring is to determine the possible impact on the total scores by simulating its use for the rGRE Analytical Writing measure. The evaluation simulates the use of e-rater as a check score with one human rater comparing that scenario to the standard double-human scoring. In these simulation analyses that use data from the evaluation samples, the first human score is evaluated against the unrounded e-rater score. The scoring simulation process is similar to what was described before in the scoring section, with the exception that the first human score is compared initially to an e-rater score instead of a second human score. If the e-rater score and the first human score are within a specified adjudication threshold (e.g., 0.5, 1.0, or 1.5), then the first human score is the final task score. If the human – machine score is distant beyond the specified adjudication threshold value, the second human score (i.e., the third rating) is compared with the first. Scoring proceeds as described earlier using the human scores only. The primary interest in such an analysis is what the difference in writing scores using AES as a check score would be versus traditional operational double-human scoring.

Agreement Thresholds and Human Adjudication

What is the impact of using e-rater as a check score with different adjudication thresholds (e.g., 0.5, 1.0, or 1.5) compared to standard double-human scoring? Most operational testing programs with writing prompts that employ double-human scoring establish adjudication rules that are consistent from one task to the next. The GRE program has used ± 1 point for human adjudication purposes when double-human scoring the Analytical Writing measure in that if the first human and second human are more than 1 score point apart from each other, additional raters are employed. For the purpose of this research project, the rGRE program wanted to replicate the use of e-rater as a check score for adjudication by a second human as originally described by Ramineni et al. (2012). Thus, the adjudication threshold investigated was that e-rater and the first human score were to be within ± 0.5 score point; when this threshold was breached, a second human adjudicator would be employed and an average computed between the two human scores. When the first human score and e-rater were within ± 0.5 score points from each other, the first human score would count as the sole score for that task.

Method

Data

Initial Cohort, Retrained, and Trend Samples

Table 2 shows information regarding the three test-taker samples used in this study. The first sample of test takers, the initial cohort, was selected at random from the initial cohort period spanning from August 1, 2011, to March 17, 2012. The second group of test takers, the retrained group, was randomly selected from a period spanning from March 18, 2012, to June 18, 2012. The first two groups were sampled so that at least 1,000 test takers were selected from each prompt for the model-build group, with the remaining test takers chosen for the evaluation group. These two model-build samples were further reduced by eliminating any test takers with advisories and human scores of 0; the remaining test takers were used to build AES models in the initial cohort and retrained periods. A third group of test takers, referred to as the trend sample, was randomly selected at the test-taker level from two window administrations taking place between November 9, 2011, and November 19, 2012. All test-taker samples selected were from the rGRE Analytical Writing operational testing program; no test takers have 0 scores or fatal advisory flags. These test-taker samples were administered both tasks in a computer-delivered format, were double-human scored, and had adequate sample sizes for model building and evaluation.

Prompts Used in this Study

A separate set of 76 argument and 76 issue prompts was used for the initial cohort and the retrained test-taker samples. A subset of 53 issue and 48 argument prompts was used in the trend test-taker data set; this subset of essays is an incidental sample of prompts that was administered to the trend test-taker cohort. For the e-rater model build and evaluations, e-rater engine v12.1 was employed. This version of e-rater consisted of 10 generic features and two additional content features that were only used in the building and evaluation of prompt-specific and generic models during the initial-cohort phase. See Breyer, Ramineni, Duchnowski, Harris, and Ridolfi (2012) for the e-rater engine upgrade report describing the evaluation of e-rater version 12.1 used in this study.

Construct Evaluation

The writing construct of the GRE was evaluated against the construct represented by e-rater as part of a previous study (Quinlan, Higgins, & Wolff, 2009). In this study, the analytic scoring features from e-rater were mapped to the six-trait scoring model (Culham, 2003) that focuses on the writing dimensions of ideas and content, organization, voice, word choice, sentence fluency, and conventions. The two GRE Analytical Writing tasks require the test taker (a) to evaluate a given argument by developing supportive evidence and reasoning and (b) to develop and support a position on an issue presented in the prompt. The GRE Analytical Writing measure requires two well-organized, well-focused analyses, each containing a logical connection of ideas among relevant reasons, along with supportive applicable examples. Responses are also evaluated for the clarity and effectiveness of prose, as well as the degree to which they show control of standard written English.

Table 2 Test-Taker Samples Used in This Research

Task	Sample	Time period	Dates	Test-taker count	Prompt count
Argument	Model-build	Initial cohort	Aug 1, 2011 – Mar 17, 2012	72,917	76
	evaluation			298,126	76
Issue	Model-build	Initial cohort	Aug 1, 2011 – Mar 17, 2012	75,197	76
	evaluation			307,177	76
Argument	Model-build	Retrained	Mar 18, 2012 – Jun 18, 2012	71,434	76
	evaluation			56,830	76
Issue	Model-build	Retrained	Mar 18, 2012 – Jun 18, 2012	74,273	76
	evaluation			58,768	76
Argument	Trend	Initial cohort	Nov 9, 2011 – Nov 19, 2011	1,790	48
Issue	Trend	Initial cohort	Nov 9, 2011 – Nov 19, 2011	1,790	53

These two macrofeatures of organization and development measure the number and average length of discourse units (i.e., functionally related segments of text) in each essay; these two features correlate strongly with the essay length. Further, the rGRE Analytical Writing tasks require fluent and precise expression of ideas using vocabulary and sentence variety. These traits are represented in e-rater by a variety of microfeatures that assess sentence-level errors (e.g., run-on sentences and fragments) and grammatical errors (e.g., subject–verb agreement), as well as the frequency with which the words in an essay are commonly used. The GRE rubric includes an evaluation of test takers' abilities to demonstrate facility with conventions (i.e., grammar, usage, and mechanics required for adequate language control) of standard written English. This language control trait, in particular, is well represented in e-rater by a large selection of microfeatures that measure errors and rule violations in grammar, usage, mechanics, and style.

Critics of AES point out that some obvious language convention errors are not captured, while other minor language control issues may be overly counted. This perception of the potential inaccuracy of AES systems in evaluating language control may be due to two issues: (a) what kinds of errors contribute to lowering a score and (b) the trade-off between falsely detecting something as incorrect versus missing something that is actually incorrect. In the first instance, regarding what kinds of errors serve to lower a score, it is the variety of specific kinds of language control errors, not solely the number, that lower an essay score. In the case of language conventions in AES, syntax rule errors, such as subject–verb agreement (i.e., *the wife go shopping*), incorrect verb forms (i.e., *we have shop for*), and preposition–object errors (i.e., *these shopping trip*) have been observed to have a more deleterious effect on scores than the misuse of homophones (i.e., *their, there*, and *they're*; Leacock & Chodorow, 2003). In the second case, regarding false detection, the tension between the false positives and false negatives is actually a balancing act (Gamon, Chodorow, Leacock, & Tetreault, 2013). On one hand, if the threshold is set above some value below which errors occur, some actual errors will likely be missed, contributing to a false negative result and perception. On the other hand, lowering the threshold too much may increase the number of actual language control errors that are detected, but the number of false positives will increase along with that increase in detection. The decision of where to place the threshold cutoff depends on the seriousness of making each of these errors, where one error may be considered more serious than another, or they may be considered equally serious.

Operational Evaluations

Phase I: Evaluating the Human Scoring Consistency of the Different Prompts and Variants

The first phase of the implementation project involved evaluation of the human scoring consistency among the different essay prompt variants. Prompt *variants* are the focused questions that were an addition to this revised GRE Analytical Writing measure. These focused questions were found Attali, Bridgeman, and Trapani (2010a) to be similar in levels of agreement with regular prompts in a GRE research section administered in the winter of 2009. As noted above, our evaluation included the human operational scoring of 76 argument and 76 issue prompts consisting of the number and variety of variant and variant groups as listed in Table 3 for the argument and issue tasks.

Table 3 Argument and Issue Prompt Counts by Variant Group and Variant

Task type	Variant group	Prompt count	Variant	Prompt count
Argument	1	53	1-Specific evidence	24
			1-Unstated assumptions	29
	2	20	2-Evaluate advice	2
			2-Evaluate a prediction	3
			2-Evaluate a recommendation	10
			2-Evaluate a rec/predict result	5
	3	3	3-Alternative explanations	3
	Total	76		76
Issue	1	48	1-Generalization	32
			1-Recommendation	11
			1-Recommended policy position	5
	2	16	2-Position with counter arguments	11
			2-Two competing positions	5
	3	12	3-Claim with reason	12
	Total	76		76

For each prompt, prompt variant, and variant group, we counted the number of test takers and produced counts (*N*), means, and standard deviations for human 1 and human 2 ratings. We also examined human interrater agreement statistics, including percentage of exact agreement, percentage of exact and adjacent agreement, QWK, product-moment correlation, and standardized difference using the evaluation sample from the initial cohort period. Then, we examined these statistics, presenting them to the automated scoring Technical Advisory Committee (TAC) for approval to proceed with the implementation research. The TAC consists of a group of senior ETS research staff members with considerable experience in technical issues regarding automated scoring. This consultation provided an independent opinion that the quality of human scoring was adequate for AES model building and evaluation.

Phase II: Building and Evaluating PS, VS, and VG Versus G Models on the Initial Cohort Data

The second phase of the implementation research involved building and evaluating agreement statistics for eight types of AES models. We built and evaluated the following eight model types:

1. Prompt-specific AES models using only the generic feature scores (i.e., 10 features, PS-10 model)
2. Prompt-specific models using all 12 feature scores, including the two prompt-specific content feature scores (i.e., PS-12 model)
3. Generic AES models using only the 10 generic feature scores (i.e., excluding the two prompt-specific content feature scores, G-10 model)
4. Generic AES model using all 12 feature scores (i.e., G-12 model)
5. Variant-specific models using only the generic feature scores (i.e., 10 features, VS-10 model)
6. Variant-specific models using all 12 feature scores, including the two prompt-specific content feature scores (i.e., VS-12 model)
7. Variant-group – specific models using only the generic feature scores (i.e., 10 features, VG-10 model)
8. Variant-group – specific models using all 12 feature scores including the two prompt-specific content feature scores (i.e., VG-12 model)

The difference between the 12-feature and the 10-feature models is the inclusion of the two content features in the model (CVA). Generic models can use the two content features trained on essays written to each individual prompt, albeit with the same common weights for all prompts; PS models have weights individualized for essays written in response to each prompt. Table 4 contains descriptions of the important features of the design used in this research study, including the test-taker samples, the AES models evaluated, the levels of those evaluations, the prompt or task type, numbers of prompts, and the level of the evaluation sample.

First, we randomly assigned test takers to either a model-build (MB) or a cross-evaluation (XV) sample group by prompt. We selected only those test takers without any advisories for model building. We took all remaining test takers

Table 4 Overview of Design

Design factor	Factor description	Level descriptions	Number of levels
Samples	Types of respondent samples used in analyses	Initial cohort, retrained, trend	3
AES models	Type of scoring model evaluated per argument/issue task	PS-10, PS-12, G-10, G-12, VS-10, VS-12, VG-10, VG-12	8
Levels of analysis	Level at which an analysis was conducted	1. Rating score 2. Task score 3. Writing score	3
Prompt/task type	Type of task given to test taker	Issue, argument	2
Number of prompts	76 argument 76 issue	N/A	76 76
Level of evaluation sample	Sample breakdown at which evaluations are conducted	Entire sample, subgroup	2

Note. AES = automated essay scoring; G = generic; PS = prompt-specific; VS = variant-specific; VG = variant-group – specific.

and included them in the evaluation group. For a subset of this evaluation group, we matched test takers' argument and issue tasks so that we had both tasks for a specific test taker. This process permitted the evaluation of Level 1 analyses at the rating level, Level 2 analyses at the task level, and Level 3 analyses at the writing score level.

Level 1 Analyses

For each prompt, prompt variant, and variant group, we counted the number of test takers in the evaluation samples without advisories and produced means and standard deviations for human and AES ratings. We also examined human–machine interrater agreement statistics including percentage of exact agreement, percentage of exact and adjacent agreement, QWK, product-moment correlation, and standardized difference using the evaluation sample from the initial cohort period. We examined the standard evaluation rating output at the prompt level, at the variant level, and at the variant group level. We also evaluated the summary of the ratings for each task. We examined the ratings for specific subgroups of test takers, including gender, ethnicity (for U.S. test takers), and test-center country.

Level 2 Analyses

We examined the impact of automated scoring for subgroups at the task level for argument and for issue, evaluating U.S. ethnic/racial groups and test center country (i.e., China, Japan, Taiwan, India, and Hong Kong). We examined cross-task correlations at the task score level for those test takers in the evaluation group, who were matched with their argument and issue performances.

Level 3 Analyses

We examined the simulated results of the different adjudication thresholds of 0.5, 1.0, and 1.5, comparing the check-score model to the contributory model. We also examined, at the writing score level for the G-10 feature model, the impact of score changes, including examining score distributions and the number and percentage of score distribution changes. We evaluated validity at the score level for Campbell and Fiske's (1959) convergent validity—correlations with verbal scores with writing and discriminant validity—correlations with math score and essay length on the other tasks. As a final step in this process, we examined these statistics, presenting them to the automated scoring TAC for approval to proceed with the implementation research.

Phase III: Building and Evaluating G Models on the Retrained Data

Using more recent test-taker data that were scored following human rater training with updated benchmark papers (i.e., the retrained test-taker group), we built only those models approved and recommended by the TAC: the G-10 model for the argument task and the G-10 model for the issue task. We performed these tasks using a randomly selected model-building data set from the period of March 18, 2012, to June 18, 2012. We sampled test takers at the prompt level. Then we selected a subset of those test takers matched for the evaluation data set. This procedure permitted the evaluation of Level 1, 2, and 3 analyses reflective of the rating, task, and writing score, respectively.

Level 1 Analyses

We reviewed the standard evaluations (i.e., standardized difference, QWK, correlation, and degradation) for the selected models on the retrained test-taker data set comparing the scores of the human raters to e-rater. We performed these evaluations at the prompt level regardless of variant or variant group, since we observed no meaningful differences at the variant and variant group levels. We counted the number of flagged prompts, tabulating the number of violations of the guideline threshold levels separately for argument and issue ratings.

Subgroup analyses were then conducted, using these data sets, scored following retraining, exploring gender and ethnicity impact at the rating level for each task. We evaluated Asia 1 (i.e., test centers located in mainland China, Hong Kong, Korea, and Taiwan); gender; U.S. racial/ethnicity subgroups; and test takers from India, China, and Taiwan for score separation between human and e-rater, for the test-taker group scored following retraining.

Level 2 Analyses

For each task, we evaluated the cross-task correlations and measures with external variables using e-rater as a check score. We also evaluated the simulated task score using e-rater as a check score for argument and issue at different adjudication thresholds of 0.5, 1.0, and 1.5, comparing those results within 0.5-point intervals to all-human double scoring and reporting the percentage in agreement.

Level 3 Analyses

We computed correlations at the writing score level between external and internal measures of convergent and divergent validity. We evaluated the difference between the simulated score using e-rater as a check score for human scoring at the analytical writing score level. We compared the check-score model to the current all-human scoring model for the matched set of test takers to determine the quantity and percentage of discrepant cases beyond ± 0.25 score points. We then presented the results to the TAC for a recommendation on the adequacy of the analyses for evaluation of automated scoring as a check score (i.e., quality control) using the G model.

Phase IV: Building the G Models on the Retrained Data and Cross-Evaluating Them on the Initial Cohort Data and Evaluating Trend Data for Bias

As a final step in the process of evaluating the argument and issue G-10 feature models built on the retrained data, we evaluated them on the initial cohort data. This additional analysis was conducted because of a concern that the test-taker cohort tested between the middle of March and the middle of June was of lower proficiency compared to the motivated group of test takers from the fall[2] and winter of 2011. For these analyses, we evaluated the standard agreement indices at the individual prompt level and at the aggregated prompt level for argument and issue, comparing the human–human ratings to the human–machine ratings. We also simulated the results of applying the generic models built on the retrained data to the initial cohort test-taker samples, simulating the results and determining the effects at the task level and at the writing score level.

We evaluated the trend test takers tested in November 2011 at the height of the admissions cycle for the effects of bias — that is, a systematic underscoring by e-rater, where e-rater provides consistently lower scores compared to human scoring — given that the two generic models for argument and issue were trained on mid-March to mid-June test takers. This March-to-June test-taker cohort is thought to be either not as motivated as a fall group (i.e., November) or not as high in proficiency. We examined the effects of the application of the retrained models at the task level for the trend group and at the writing score level, biasing each model by adjusting the intercept by 0 (i.e., no bias), -0.1 (i.e., scoring with a -0.1 bias), and -0.2. We then compared the effect of these purposely biased e-rater scores to human scores in a check-score model with adjudication thresholds of 0.5, 1.0, and 1.5.

When we report results, we place the detailed prompt-by-prompt level information in the appendices for phases I, II, and III. For phase IV, we present adjudication thresholds other than 0.5 in the appendices.

Results

Phase I Results: Evaluating the Human Scoring Consistency of the Different Prompts and Variants

Human Scoring of Variants

Table 5 shows human–human agreement statistics by variant for the argument task, and Table 6 shows similar information for the issue task. These data are based on the initial cohort of examines. Cells with bold font in these two tables show agreement indices that failed to meet the guideline thresholds. For argument prompts, the raters disagree on the second group of variants that require the writer to *evaluate* something. For the issue variants, it appears that the raters have some disagreement over *positions* that test takers are asked to take. While these variants show common variants where the human scorers fail to meet the agreement thresholds, the QWK and the correlations are close to the guideline threshold. More importantly, the standardized differences are consistently near 0, implying that the disagreements are perhaps at select score points, but *on average* the scores from the two human raters are in agreement.

Table 5 Phase I Agreement of Human Scores on Argument Variants

		Human 1 by Human 2								
		Human 1		Human 2		Statistic				
Variant	*N*	*M*	*SD*	*M*	*SD*	Std diff	Wtd kappa	% agree	% adj agree	*r*
1-Specific evidence	97,584	3.28	0.81	3.28	0.81	0.00	0.70	64	99	0.70
1-Unstated assumptions	96,918	3.19	0.83	3.19	0.82	−0.01	0.71	64	99	0.71
2-Evaluate advice	12,090	3.31	0.81	3.33	0.81	0.02	0.71	65	99	0.71
2-Evaluate a prediction	21,432	3.22	0.80	3.23	0.80	0.00	**0.69**[a]	64	99	**0.69**[a]
2-Evaluate a recommendation	35,711	3.25	0.81	3.25	0.81	−0.01	**0.67**[a]	62	98	**0.67**[a]
2-Evaluate a recommendation predicted result	25,831	3.28	0.80	3.29	0.80	0.01	**0.69**[a]	64	99	**0.69**[a]
3-Alternative explanations	9,419	3.28	0.81	3.27	0.82	−0.01	0.71	64	99	0.71
Average	42,712	3.26	0.81	3.26	0.81	0.00	0.70	64	99	0.70

Note. Std diff = standardized difference; Wtd = weighted; adj = adjacent. [a]Agreement indices that failed to meet the guideline thresholds (also indicated by boldface).

Table 6 Phase I Agreement of Human Scores on Issue Variants

		Human 1 by Human 2								
		Human 1		Human 2		Statistic				
Variant	*N*	*M*	*SD*	*M*	*SD*	Std diff	Wtd kappa	% agree	% adj agree	*r*
1-Generalization	113,356	3.22	0.79	3.22	0.79	0.00	0.72	67	99	0.72
1-Recommendation	47,903	3.25	0.77	3.25	0.77	−0.01	0.72	68	99	0.72
1-Recommended policy position	32,069	3.30	0.76	3.30	0.75	0.00	**0.69**[a]	67	99	**0.69**[a]
2-Position with counterarguments	46,800	3.11	0.72	3.10	0.72	0.00	**0.67**[a]	68	99	**0.67**[a]
2-Two competing positions	33,291	3.10	0.77	3.10	0.77	0.01	**0.70**[a]	66	99	**0.70**[a]
3-Claim with reason	35,236	3.15	0.79	3.15	0.78	0.00	0.72	67	99	0.72
Average	51,443	3.19	0.77	3.19	0.76	0.00	0.70	67	99	0.70

Note. Std diff = standardized difference; Wtd = weighted; adj = adjacent. [a]Agreement indices that failed to meet the guideline thresholds (also indicated by boldface).

Human Scoring of Variant Groups

Table 7 shows human–human agreement statistics by *variant group* for the argument task, and Table 8 shows similar information by *variant group* for the issue task. The variant groups that fail to meet the agreement guideline threshold of 0.7 for QWK and correlation for argument and issue mainly consist of the variants that failed to meet agreement thresholds for variants. It is noteworthy that the standardized mean score differences are near 0 for both argument and issue for these variant groups, confirming a similar observation in the scoring of variants. It appears that humans disagree in the argument task at the rating level in scoring test takers who are *evaluating* the development of an argument. Further, human scorers disagree in scoring test takers who are tasked with evaluating a *position* on an issue.

Table 7 Phase I Agreement of Human Scores Within Argument Variant Groups

		Human 1 by Human 2								
		Human 1		Human 2		Statistic				
Variant group	*N*	*M*	*SD*	*M*	*SD*	Std diff	Wtd kappa	% agree	% adj agree	*r*
Group 1	194,502	3.24	0.82	3.23	0.82	−0.01	0.70	64	99	0.70
Group 2	95,064	3.26	0.81	3.26	0.81	0.00	**0.69**[a]	64	99	**0.69**[a]
Group 3	9,419	3.28	0.81	3.27	0.82	−0.01	0.71	64	99	0.71
Average	99,662	3.26	0.81	3.25	0.81	0.00	0.70	64	99	0.70

Note. Std diff = standardized difference; Wtd = weighted; adj = adjacent. [a]Agreement indices that failed to meet the guideline thresholds (also indicated by boldface).

Table 8 Phase I Agreement of Human Scores Within Issue Variant Groups

		Human 1 by Human 2								
		Human 1		Human 2		Statistic				
Variant group	*N*	*M*	*SD*	*M*	*SD*	Std diff	Wtd kappa	% agree	% adj agree	*r*
Group 1	193,328	3.24	0.78	3.24	0.78	−0.01	0.71	67	99	0.71
Group 2	80,091	3.10	0.74	3.10	0.74	0.00	**0.68**[a]	67	99	**0.68**[a]
Group 3	35,236	3.15	0.79	3.15	0.78	0.00	0.72	67	99	0.72
Average	102,885	3.16	0.77	3.16	0.77	0.00	0.70	67	99	0.70

Note. Std diff = standardized difference; Wtd = weighted; adj = adjacent. [a]Agreement indices that failed to meet the guideline thresholds (also indicated by boldface).

Table 9 Phase II Distribution of Advisory Flags in Argument and Issue Prompts

Prompt	Total	No advisory	% total	Nonfatal advisory	% total	Fatal advisory	% total	Either advisory	% total
Argument	75,314	72,917	96.82	2,280	3.03	117	0.16	2,397	3.18
Issue	75,903	75,197	99.07	580	0.76	126	0.17	706	0.93

Appendix E shows human–human agreement for argument (Table E1) and issue (Table E2) tasks on a prompt-by-prompt basis. Comparison of the list of argument human–human agreement statistics on a prompt-by-prompt basis in Table E1 with a similar list of issue human–human agreement statistics in Table E2 shows many more prompts in the argument task violate the QWK and correlation guideline thresholds compared to the issue task. This means that humans have difficulty agreeing on argument prompts compared to issue prompts on a prompt-by-prompt basis. Yet, on average, human raters show very little difference in that not one argument or issue prompt is flagged for the standardized difference statistic.

Phase II Results: Building and Evaluating PS, VS, and VG Versus G Models on the Initial Cohort Data

Summaries of advisories are presented for the argument and issue prompts used in phase II of the study in Table 9. As a proportion of the overall sample, these advisory counts are small, but we present them in Table 9 because we excluded these individuals from both model building and all subsequent evaluations.

Appendix F shows the counts and percentage of respondents with fatal and nonfatal advisory flags that were attained by using the initial cohort samples for argument and issue essays on a prompt-by-prompt basis. Table F1 provides fatal and nonfatal information for the argument task on a prompt-by-prompt basis and Table F2 provides similar information for the issue task. Examination of Tables F1 and F2 shows that, on average, argument prompts have between 3% and 4% advisories, while issue prompts have between 0 and 1% advisories.

Level 1 Analyses

Table 10 shows summary average descriptive statistics and average agreement statistics across prompts for rounded and unrounded e-rater scores with human ratings for eight different models for the argument ratings. The rounded scores are used for the QWK statistics, and the unrounded are used for the standardized mean score differences and the correlations.

Table 10 shows that the mean score of e-rater regardless of AES model is 0.01 less than the human mean for the argument prompts. The standardized differences are near zero, but they do reflect the slight bias. QWKs are all acceptable, correlations (r) on unrounded e-rater scores are all very similar and acceptable, and degradation is near 0 for QWK and well within acceptability for r. Given these results, the use of a generic model without the use of the CVA features is most likely a preferred choice because it gives the program flexibility in adding and removing prompts without having to rebuild the model. Examination of Table 10 shows that the standardized difference between human and unrounded e-rater scores is near 0 for all eight models and QWKs (weighted kappa) are all at or above threshold. Further, degradation statistics for both weighted kappa and correlation are positive, indicating that the agreement of e-rater with human scoring was consistently higher than agreement between the two human raters.

Table 10 Phase II Agreement of Human and e-rater Scores on Argument Prompts: Summary of Eight Models

					Human 1 by e-rater					Degradation[a]	
			Human 1		e-rater (rounded)		Statistics			Wtd kappa	*r*
Model	Type	*N*	*M*	*SD*	*M*	*SD*	Std diff (unrnd)	Wtd kappa	*r*(unrnd)	h1er_rnd-h1h2	h1er-h1h2
Generic	G-10	3,934	3.25	0.81	3.24	0.84	−0.01	0.70	0.75	0.01	0.06
	G-12	3,934	3.25	0.81	3.24	0.84	−0.01	0.70	0.75	0.01	0.06
Prompt-specific	PS-10	3,934	3.25	0.81	3.24	0.86	−0.01	0.71	0.75	0.02	0.06
	PS-12	3,934	3.25	0.81	3.24	0.86	−0.01	0.71	0.76	0.02	0.06
Variant-specific	VS-10	3,934	3.25	0.81	3.24	0.84	−0.01	0.70	0.75	0.01	0.06
	VS-12	3,934	3.25	0.81	3.24	0.84	−0.01	0.70	0.75	0.01	0.06
Variant-group –specific	VG-10	3,934	3.25	0.81	3.24	0.84	−0.01	0.70	0.75	0.01	0.06
	VG-12	3,934	3.25	0.81	3.24	0.84	−0.01	0.70	0.75	0.01	0.06

Note. h1er = human 1 and e-rater; h1h2 = human 1 and human 2; rnd = rounded; unrnd = unrounded; Wtd = weighted; G = generic; PS = prompt-specific; VS = variant-specific; VG = variant-group – specific. [a]Positive values of degradation indicate higher human – machine agreement than human – human agreement; negative values of degradation indicate a reduction in agreement.

Table 11 shows similar information for the issue prompts as that displayed in Table 10 for the argument prompts for the eight different models evaluated.

The human means and e-rater means are close to each other for the issue task and do not show the consistency of one rater being slightly higher than the other across the various models, as was observed for the argument task. Further, examination of Table 11 shows that standardized difference between human and unrounded e-rater scores are near 0 for all eight models with the largest (–0.02) observed for the VS model without the two content variables. Also, weighted kappa for the different models are all at or above threshold, with the lowest (0.74) observed for the G model *with* the two content variables. Similar results are seen for the Pearson correlation, with the lowest (0.75) for the G model without the content features. Further, degradation statistics for both weighted kappa and correlation are all positive, indicating that the agreement of e-rater with human scoring was consistently higher than the agreement between the two human raters.

Both Tables 10 and 11 show good overall average agreement in human ratings with e-rater, with little improvement for the PS, VS, and VG models compared to the G models in the QWK and the correlations for either task. Further, the addition of the two content features to the different models did little to increase the agreement for either task. It is noteworthy that the standardized mean score differences are not different on average and are near 0 on these initial cohort test takers, nor are the means and standard deviations of e-rater meaningfully different among the eight models within

Table 11 Phase II Agreement of Human and e-rater Scores on Issue Prompts: Summary of Eight Models

					Human 1 by e-rater					Degradation[a]	
			Human 1		e-rater (rounded)		Statistic			Wtd kappa	*r*
Model	Type	*N*	*M*	*SD*	*M*	*SD*	Std diff (unrnd)	Wtd kappa	*r*(unrnd)	h1er_rnd-h1h2	h1er-h1h2
Generic	G-10	4,061	3.20	0.77	3.20	0.80	0.00	0.74	0.79	0.04	0.09
	G-12	4,061	3.20	0.77	3.20	0.80	0.00	0.74	0.79	0.04	0.09
Prompt-specific	PS-10	4,061	3.20	0.77	3.19	0.81	−0.01	0.74	0.79	0.04	0.09
	PS-12	4,061	3.20	0.77	3.19	0.81	−0.01	0.74	0.79	0.04	0.09
Variant-specific	VS-10	4,061	3.20	0.77	3.19	0.81	−0.02	0.74	0.79	0.04	0.09
	VS-12	4,061	3.20	0.77	3.19	0.81	−0.02	0.74	0.79	0.04	0.09
Variant-group – specific	VG-10	4,061	3.20	0.77	3.19	0.80	−0.01	0.74	0.79	0.09	0.09
	VG-12	4,061	3.20	0.77	3.19	0.80	−0.01	0.74	0.79	0.04	0.09

Note. h1er = human 1 and e-rater; h1h2 = human 1 and human 2; rnd = rounded; unrnd = unrounded; wtd = weighted; G = generic; PS = prompt-specific; VS = variant-specific; VG = variant-group – specific. [a]Positive values of degradation indicate higher human – machine agreement than human – human agreement; negative values of degradation indicate a reduction in agreement.

each task of argument and issue. Further, degradation statistics are all within guideline threshold limits, indicating that e-rater is performing well given the human rating comparisons.

Appendix G contains the prompt level evaluation listing of these agreement statistics for the G-10, G-12, PS-10, and PS-12 models for argument and issue on a prompt-by-prompt basis. The G-10 feature and G-12 feature models are presented in Tables G1 and G2 for the individual argument prompts; Tables G3 and G4 provide similar information on a prompt-by-prompt basis for the prompt-specific models for the argument task. Comparison of G models with and without CVA shows many similarities in examining the same prompt. Comparison of PS-10 and PS-12 models to the G-10 and the G-12 models for the same prompts show fewer standardized difference violations for the PS-10 and PS-12 models. Comparison to the PS-10 and the PS-12 models shows no improvement because of adding two prompt-specific content features. It appears that the incorporation of the two prompt-specific content macrofeatures adds little improvement for human – machine agreement.

Appendix G also provides Tables G5 and G6, showing the prompt level listing of the G-10 feature and G-12 feature model evaluations for the issue task; Tables G7 and G8 provide the prompt-by-prompt agreement evaluations for the 10-feature and 12-feature PS models. Examination of the issue prompts shows that there are few violations of the guideline thresholds for G-10, G-12, PS-10, and PS-12 models. There are very few threshold violations for either PS-10 or PS-12 model in the issue task, mainly because they are modeled specifically to each prompt. In the past (Ramineni et al., 2012), GRE used a PS-12 model for argument and something akin to a G-12 model, specifically a generic model with a prompt-specific intercept for the issue prompts. As a check score, TAC recommended that we pursue the G-10 models for both argument and issue prompts. We have not included the prompt-by-prompt listing of the VS and VG agreement evaluations in Appendix G since they are very similar to the G and PS models.

Additionally, we evaluated the number of violations (i.e., flags) of the guideline thresholds for human – machine agreement statistics for these eight models for argument and eight models for issue ratings. Tables 12 and 13 show the counts of the number of guideline threshold violations for each of the different models for argument and issue prompts, respectively, which total 76 each.

Examination of these two tables reveals that, for argument prompts, overall weighted kappa is the most prevalent flag, with 27 violations for the G-10 model and 25 for the G-12 model, and that standardized difference is the most prevalent flag for issue, with 15 prompt violations for the G-10 model and 16 violations for the G-12 model. Also, the argument task has more prompts violating guideline thresholds than the issue task. The models with the lowest numbers of flags are the PS model without the two content features for argument and the two PS models both with and without the two content features for issue. Analyses from this point forward will concentrate on the G and PS models only, since there are little differences seen in the VS and VG models compared to the PS and G models.

Table 14 shows rating results assessing score separation between human and e-rater for different major subgroups for the argument task for four of the models evaluated on the initial cohort group, including the 10- and 12-feature G and the 10- and 12-feature PS models only. The major subgroups presented are Asia 1 (i.e., China, Hong Kong, Korea, and Taiwan); gender; U.S. racial/ethnic groups; and test takers tested in the countries of India, China, Canada, Korea, Taiwan, and Hong Kong. These subgroups are some of the largest and have, in the past (see Ramineni et al., 2012), exhibited large human – machine differences.

Table 12 Phase II Counts of Guideline Threshold Violations for Eight Models for the Argument Prompts

			Flag counts			
Model	Type	*N*	Std diff(unrnd)	Wtd kappa	Degradation wtd kappa	Flag exists
Generic	G-10	3,934	15	27	0	32
	G-12	3,934	17	25	0	34
Prompt-specific	PS-10	3,934	0	16	0	16
	PS-12	3,934	0	17	0	17
Variant-specific	VS-10	3,934	12	28	0	30
	VS-12	3,934	14	26	0	30
Variant-group – specific	VG-10	3,934	14	26	0	30
	VG-12	3,934	18	26	0	33

Note. unrnd = unrounded; wtd = weighted; G = generic; PS = prompt-specific; VS = variant-specific; VG = variant-group – specific.

Table 13 Phase II Counts of Guideline Threshold Violations for Eight Models for the Issue Prompts

			Flag counts			
Model	Type	*N*	Std diff(unrnd)	Wtd kappa	Degradation wtd kappa	Flag exists
Generic	G-10	4, 061	15	6	0	18
	G-12	4, 061	16	6	0	19
Prompt-specific	PS-10	4, 061	0	5	0	5
	PS-12	4, 061	0	5	0	5
Variant-specific	VS-10	4, 061	4	6	0	8
	VS-12	4, 061	4	6	0	8
Variant-group – specific	VG-10	4, 061	11	6	0	15
	VG-12	4, 061	11	6	0	15

Note. unrnd = unrounded; wtd = weighted; G = generic; PS = prompt-specific; VS = variant-specific; VG = variant-group – specific.

Table 14 Phase II Subgroup Differences Measured by Standardized Mean Score Differences for the Argument Prompts

			Prompt-specific		Generic	
Subgroups		*N*	PS-10 Std diff(unrnd)	PS-12 Std diff(unrnd)	G-10 Std diff(unrnd)	G-12 Std diff(unrnd)
Asia 1[a]		21,266	**−0.18**[b]	**−0.15**[b]	**−0.28**[b]	**−0.26**[b]
Gender	Female	149,439	0.00	0.00	0.01	0.01
	Male	113,776	−0.03	−0.03	−0.03	−0.03
Race	American Indian or Alaskan Native	1,132	0.07	0.07	0.06	0.05
	Asian or Asian American	11,545	0.07	0.07	0.08	0.08
	Black or African American	16,275	−0.07	−0.08	−0.06	−0.07
	Mexican, Mexican American, or Chicano	5,314	0.03	0.02	0.03	0.03
	Puerto Rican	1,466	0.04	0.04	0.07	0.06
	Other Hispanic, Latino, or Latin American	6,849	0.03	0.03	0.04	0.04
	White (non-Hispanic)	13,7095	0.07	0.06	0.06	0.06
Country	India	18,963	**−0.11**[b]	−0.10	−0.05	−0.05
	China	17,658	**−0.19**[b]	**−0.16**[b]	**−0.31**[b]	**−0.28**[b]
	Canada	4,290	0.04	0.04	0.04	0.04
	Korea	1,583	−0.08	−0.06	**−0.18**[b]	**−0.16**[b]
	Taiwan	1,248	**−0.31**[b]	**−0.30**[b]	**−0.33**[b]	**−0.33**[b]
	Hong Kong	777	−0.02	0.00	−0.03	−0.03

Note. unrnd = unrounded; G = generic; PS = prompt-specific. [a]Test takers from mainland China, Hong Kong, Taiwan, and Korea. [b] Agreement indices that failed to meet the guideline thresholds (also indicated by boldface).

Separation between human and e-rater scores occurs consistently regardless of model used for Asia 1 and test takers tested in Taiwan test centers. Score separation between human and e-rater is less evident (i.e., above the guideline threshold of |0.10| for subgroups) for test takers tested in India, China, and Korea for the argument essays when using either of the G models. We note the differences for test takers from China and Taiwan are both negative, indicating lower e-rater scores compared to human scores. These results are contrary to previous research findings for argument prompts. Bridgeman, Trapani, and Attali (2012) found positive differences for test takers from China (standardized difference = 0.38) on argument prompts, indicating higher e-rater than human scores. Ramineni, Williamson, and Weng (2010) found positive differences for test takers from China (standardized difference = 0.56) and negative differences for test takers from Taiwan (standardized difference = – 0.22). This previous research was conducted using GRE prompts before variants were introduced and with an AES model that did not use differential word use (DWU) as a macrofeature. The scoring engines used in these prior research studies did not use the engine used in this research (i.e., e-rater engine v12.1). It should be noted that DWU is no longer used operationally.

Table 15 shows similar information assessing score separation between human and e-rater for different major subgroups for the issue tasks summarized across all prompts for four of the models evaluated on the initial cohort group, including the 10- and 12-feature G and the 10- and 12-feature PS models.

The results in Table 15 show few score separation differences for the issue prompts between human and e-rater ratings. Asia 1, gender, and U.S. racial/ethnic groups show no standardized mean score differences that violate the guideline threshold of |0.10| for subgroups. For test takers grouped by test center country, only Taiwan shows a threshold violation for the issue essays in score separation between human and e-rater ratings. Again, these results are contrary to previous research regarding score separation between e-rater and human scorers on issue prompts. Ramineni et al. (2012) and Ramineni et al. (2010) found positive differences, indicating higher e-rater scores than human scores for test takers from China (standardized difference = 0.60) and near 0 differences, but nonetheless negative for test takers from Taiwan (standardized difference = −0.06) for issue prompts. Bridgeman et al. (2012) found positive differences for test takers from both China (standardized difference = 0.60) and Taiwan (standardized difference = 0.12) for issue prompts. Again, these previous results used GRE prompts prior to the introduction of variants and on an earlier version of the e-rater scoring engine that did not contain the DWU macrofeature.

Level 2 Analyses

Table 16 shows the results of the cross-task correlations with both convergent (scores from the GRE Verbal Reasoning measure and the alternate task of argument with issue and vice versa) and divergent (i.e., scores from the GRE Quantitative Reasoning measure) measures for the G model without the CVA features under the different adjudication threshold level of 1.5, 1.0, and 0.5. The argument and issue task scores are the result of operational human – human scoring. The argument and issue score under the different adjudication thresholds are the simulated task scores of what would occur when e-rater was used as a check score with a G-10 AES model.

Examination of Table 16 shows that argument and issue correlate 0.16 and 0.25, respectively, with the GRE Quantitative Reasoning measure score, but much higher with the GRE Verbal Reasoning measure score with issue (0.55) and argument (0.64) for the G-10 model. Further, the cross-task correlations between the issue and argument task is 0.66. Also, note that the correlations of the argument and issue tasks with verbal using e-rater as a check score in the G-10 model has the highest correlations with the external verbal score at an adjudication threshold of 0.5 compared to the other adjudication thresholds. Also examining the task score rows of argument and issue and comparing those correlations to the argument and issue at the different adjudication thresholds shows no degradation in cross-task correlations at the 0.5 threshold, but

Table 15 Phase II Subgroup Differences Measured by Standardized Mean Score Differences for the Issue Prompts

			Prompt-specific		Generic	
Subgroups		*N*	PS-10 Std diff(unrnd)	PS-12 Std diff(unrnd)	PS-10 Std diff(unrnd)	PS-12 Std diff(unrnd)
Asia 1[a]		21,658	−0.03	−0.03	−0.02	−0.03
Gender	Female	153,611	0.00	0.00	0.01	0.01
	Male	117,460	−0.03	−0.03	−0.02	−0.02
Race	American Indian or Alaskan Native	1,144	0.01	0.01	0.02	0.02
	Asian or Asian American	11,798	0.07	0.07	0.07	0.08
	Black or African American	16,940	−0.06	−0.07	−0.05	−0.05
	Mexican, Mexican American, or Chicano	5,466	0.01	0.01	0.01	0.02
	Puerto Rican	1,439	0.00	0.00	0.01	0.01
	Other Hispanic, Latino, or Latin American	7,108	0.00	0.00	0.00	0.00
	White (non-Hispanic)	140,353	0.03	0.03	0.04	0.04
Country	India	19,992	−0.07	−0.07	−0.04	−0.05
	China	17,986	−0.01	−0.01	−0.01	−0.02
	Canada	4,396	0.01	0.01	0.01	0.02
	Korea	1,575	−0.05	−0.04	−0.05	−0.05
	Taiwan	1,296	**−0.23**[b]	**−0.23**[b]	**−0.21**[b]	**−0.22**[b]
	Hong Kong	801	0.04	0.04	0.04	0.04

[a]Test takers from mainland China, Hong Kong, Taiwan, and Korea. [b]Agreement indices that failed to meet the guideline thresholds (also indicated by boldface).

Table 16 Phase II Cross-Task Correlations Under Multiple Check-Score Thresholds: G-10 Model

		External		Words		Task score		1.5		1.0		0.5	
		Quantitative	Verbal	Issue	Argument	Issue	Argument	Issue	Argument	Issue	Argument	Issue	Argument
External	Quantitative	1.00	0.36	0.18	0.23	0.16	0.25	0.15	0.24	0.15	0.24	0.16	0.25
	Verbal	0.36	1.00	0.37	0.44	0.59	0.64	0.55	0.60	0.56	0.61	0.58	0.63
Words	Issue			1.00	0.78	0.74	0.55	0.69	0.52	0.71	0.54	0.75	0.56
	Argument			0.78	1.00	0.65	0.71	0.61	0.67	0.62	0.69	0.65	0.72
Task score	Issue					1.00	0.66	0.93	0.62	0.93	0.64	0.96	0.66
	Argument					0.66	1.00	0.61	0.93	0.63	0.94	0.65	0.96
1.5	Issue							1.00	0.58	0.99	0.59	0.96	0.62
	Argument							0.58	1.00	0.59	0.99	0.62	0.96
1.0	Issue									1.00	0.61	0.97	0.63
	Argument									0.61	1.00	0.63	0.97
0.5	Issue											1.00	0.66
	Argument											0.66	1.00

Note. $N = 251{,}987$.

some slight degradation compared to the other adjudication thresholds of 1.0 and 1.5. Finally, the reader should note that the correlation of the argument or the issue task under the 0.5 adjudication threshold with its double-human score is near 1 at 0.96.

Table 17 shows the cross-task correlations and the correlations of each task with external measures using the PS-12 e-rater model. The task score of issue and argument are the operational scores from the double-human scoring, while the argument and issue scores under the different adjudication thresholds of 1.5, 1.0, and 0.5 are the simulated argument and issue task scores when the PS-12 model was used as a check score for those writing tasks.

Table 17 shows the divergent validity of the argument and issue prompts with quantitative is low, but the correlation of the number of words (Words) in each of the argument and issue essays is high (i.e., 0.78). The correlation of the cross-task convergent correlations of argument and issue with the GRE verbal score (i.e., 0.64 and 0.59) are much higher than either is with the GRE quantitative score (i.e., 0.25 and 0.16). The cross-task correlations for argument and issue are moderate as well at 0.66. Under the three different adjudication thresholds, the cross-task correlations are highest at the 0.5 adjudication level at 0.66 and 0.65, indicating that each task measures something different.

Examination of Table 17 shows that argument and issue correlate 0.16 and 0.25, respectively, with the GRE quantitative score, but much higher with the GRE verbal score with issue (0.59) and argument (0.64) when a PS-12 model is used. Further, the cross-task correlation between the issue and argument task is 0.66. Also, note that the correlations of the argument and issue tasks with verbal using e-rater as a check score in the PS-12 model has the highest correlations with the external verbal score at an adjudication threshold of 0.5 compared to the other adjudication thresholds. Also, examining

Table 17 Phase II Cross-Task Correlations Under Multiple Check-Score Thresholds: PS-12 Model

		External		Words		Task score		1.5		1.0		0.5	
		Quantitative	Verbal	Issue	Argument	Issue	Argument	Issue	Argument	Issue	Argument	Issue	Argument
External	Quantitative	1.00	0.36	0.18	0.23	0.16	0.25	0.15	0.24	0.15	0.24	0.16	0.25
	Verbal	0.36	1.00	0.37	0.44	0.59	0.64	0.55	0.60	0.56	0.61	0.58	0.64
Words	Issue			1.00	0.78	0.74	0.55	0.69	0.52	0.71	0.54	0.74	0.56
	Argument			0.78	1.00	0.65	0.71	0.61	0.66	0.62	0.69	0.66	0.71
Task score	Issue					1.00	0.66	0.93	0.62	0.93	0.64	0.96	0.66
	Argument					0.66	1.00	0.61	0.93	0.63	0.94	0.66	0.96
1.5	Issue							1.00	0.58	0.99	0.59	0.96	0.62
	Argument							0.58	1.00	0.59	0.99	0.62	0.96
1.0	Issue									1.00	0.61	0.97	0.63
	Argument									0.61	1.00	0.63	0.97
0.5	Issue											1.00	0.66
	Argument											0.66	1.00

Note. $N = 251{,}987$.

Table 18 Phase II Percentage of Score Differences at the Task Level Between the Use of e-rater as a Check Score Versus the All-Human Double Scoring for the Argument Task

Model	Threshold	≤ −1.5	−1	−0.5	0	0.5	1	≥ 1.5
Generic (G-10)	0.5	0	0.13	7.87	83.48	8.21	0.27	0.04
	1.0	0.01	0.15	14.25	70.48	14.68	0.33	0.09
	1.5	0.06	0.24	16.27	66.36	16.61	0.32	0.13
Prompt-specific (PS-12)	0.5	0	0.14	7.92	83.36	8.27	0.28	0.03
	1.0	0.01	0.16	14.38	70.23	14.79	0.34	0.08
	1.5	0.06	0.25	16.33	66.28	16.63	0.32	0.13

Note. $N = 308{,}450$. All candidate records are drawn from August 1, 2011, through March 17, 2012.

the task score rows of argument and issue and comparing those correlations to the argument and issue at the different adjudication thresholds shows little degradation in cross-task correlation at the 0.5 threshold, but some slight degradation compared to the other adjudication thresholds of 1.0 and 1.5. Finally, the reader should note that the correlation of the argument or the issue task under the 0.5 adjudication threshold with its double-human score is near 1 at 0.96. These results are similar to what was observed with the G-10 model.

Comparison of the correlations in Table 16 with the correlations in Table 17 shows that argument and issue correlate low with the GRE Quantitative Reasoning measure score, but much higher with the GRE Verbal Reasoning measure score with little discernible difference between the G-10 and the PS-12 models. Further, the cross-task correlation between the issue and argument task is 0.66. Also, note that the correlations of the argument and issue tasks with verbal using e-rater as a check score in the G-10 model has the highest correlations with the external verbal score (at an adjudication threshold of 0.5) compared to the other adjudication thresholds. Examining the task score rows of argument and issue and comparing those correlations to the argument and issue at the different adjudication thresholds shows no degradation in cross-task correlation at the 0.5 threshold, but some slight degradation compared to the other adjudication thresholds of 1.0 and 1.5. Finally, the reader should note that the correlation of the argument or the issue task under the 0.5 adjudication threshold with its double-human score is near 1 at 0.96 and that the PS-12 model shows no improvement over the G-10 model.

Tables 18 and 19 show score differences at the task level between the use of e-rater as a check score versus the all-human double-scoring model under the two main e-rater models (i.e., the G and PS models) used as a check score under consideration for each of the different adjudication thresholds (i.e., 0.5, 1.0, and 1.5). Referring to Table 18 regardless of the model or the adjudication threshold used, the percentage of check scores that differ from the all-human score by more than ±0.5 score points is less than 1% for the argument task.

Table 19 shows that over 99% of all task scores for argument and issue are within ±0.5 score point of the score obtained under the all-human double-scoring process. Less than 1% of all scores are different from what would be obtained using e-rater as a check score versus all-human double scoring. Referring to Table 19, regardless of the model used, the percentage of check scores that differ from the all-human score by more than ±0.5 score points is less than 1% for the issue task. This is also true regardless of the adjudication threshold.

Table 19 Phase II Percentage of Score Differences at the Task Level Between the Use of e-rater as a Check Score Versus All-Human Double Scoring the Issue Prompts

Model	Threshold	≤ −1.5	−1	−0.5	0	0.5	1	≥ 1.5
Generic (G-10)	0.5	0	0.06	7.55	84.16	8.1	0.13	0.01
	1.0	0	0.09	14.16	70.95	14.6	0.17	0.03
	1.5	0.02	0.16	15.63	68.03	15.92	0.18	0.05
Prompt-specific (PS-12)	0.5	0	0.06	7.59	84.19	8.04	0.12	0.01
	1.0	0	0.1	14.26	70.82	14.62	0.17	0.03
	1.5	0.02	0.17	15.65	68.01	15.92	0.18	0.05

Note. $N = 310{,}302$. All candidate records are drawn from August 1, 2011, through March 17, 2012.

Table 20 Phase II Convergent and Divergent Correlations at the Writing Score Level for the Generic Model

	Quantitative	Verbal	Average words	Writing raw	1.5	1.0	0.5
Quantitative	1.00	0.36	0.22	0.23	0.22	0.22	0.22
Verbal		1.00	0.43	0.67	0.65	0.65	0.67
Average words			1.00	0.77	0.74	0.75	0.78
Writing raw				1.00	0.95	0.96	0.98
1.5					1.00	0.99	0.98
1						1.00	0.98
0.5							1.00

Note. $N = 251{,}987$.

Level 3 Analyses

Table 20 shows the correlations at the writing score level for the use of e-rater as a check score using the G-10 model with external and internal measures. The writing raw score is the operational all-human score, whereas the correlations under the 1.5, 1.0, and 0.5 adjudication levels are the simulated checked scores under those different adjudication thresholds given the use of e-rater using the G-10 model.

The correlation for the average number of words in the essays is lower for quantitative than for verbal, and the writing score correlates lower with quantitative scores than it does with verbal scores, indicating appropriate divergent and convergent validity. Also in Table 20 are the correlations showing the effects of the use of the check-score model with the different adjudication thresholds under consideration. The correlations at the check-score threshold value of 0.5 are slightly higher with the GRE verbal score than they are for any other adjudication threshold. Further, the correlations with the total writing score using the adjudication threshold of 0.5 show little difference in correlation (0.98) from the writing score with itself.

Table 21 shows the correlations at the writing score level for the use of e-rater as a check score using the PS-12 model. As noted earlier, the correlations under the different adjudication thresholds of 1.5, 1.0, and 0.5 are the simulated check scores using e-rater with a PS-12 model.

There are few differences worth noting between the use of the G-10 model and the PS-12 model in correlations with external and internal measures at the writing score level, as can be seen by comparing Tables 20 and 21, respectively. The correlation of the checked writing score under adjudications of 0.5 has the highest correlation with the writing raw score for both G-10 and PS-12 models. The remaining differences in correlations between the G-10 and the PS-12 models are negligible. Table 22 shows the effects of using the G e-rater model as a check score versus all-human double scoring at the analytical writing score level. This table also shows the effects of using the PS e-rater model as a check score at the analytical writing score level compared to all-human double scoring at the different adjudication thresholds of 0.5, 1.0, and 1.5 points.

Phase III Results: Building and Evaluating G Models on the Retrained and Crossover Data

Recall that we built and evaluated only those models approved and recommended by the TAC—the G-10 model for the argument task and the G-10 model for the issue task using test-taker data that were scored following human rater

Table 21 Convergent and Divergent Correlations at the Writing Score Level for the Prompt-Specific Model

	Quantitative	Verbal	Average words	Writing raw	1.5	1.0	0.5
Quantitative	1.00	0.36	0.22	0.23	0.22	0.22	0.23
Verbal		1.00	0.43	0.67	0.65	0.65	0.67
Average words			1.00	0.77	0.74	0.75	0.78
Writing raw				1.00	0.95	0.96	0.98
1.5					1.00	0.99	0.98
1						1.00	0.98
0.5							1.00

Note. $N = 251{,}987$.

Table 22 Phase II Percentage of Score Differences at the Writing Score Level Between the Use of e-rater as a Check Score Versus All-Human Double Scoring for the Analytical Writing Score

Model l	Threshold	≤ −0.75	−0.5	−0.25	0	0.25	0.5	≥ 0.75
Generic (G-10)	0.5	0.01	0.84	12.96	71.48	13.48	1.13	0.1
	1.0	0.05	2.23	20	54.58	20.32	2.62	0.2
	1.5	0.13	2.86	21.32	50.75	21.61	3.04	0.27
Prompt-specific (PS-12)	0.5	0.01	0.83	13.07	71.37	13.47	1.15	0.09
	1.0	0.05	2.26	20.11	54.37	20.36	2.64	0.2
	1.5	0.15	2.88	21.34	50.71	21.6	3.04	0.28

Note. $N = 251{,}987$. All candidate records are drawn from August 1, 2011, through March 17, 2012.

Table 23 Phase III Distribution of Advisory Flags in Argument and Issue Prompts

Prompt	No advisory	Nonfatal advisory flag	Fatal advisory flag	Flag sum	% flag	Row sum
Argument	71,449	5,826	3,277	9,103	11.30	80,552
Issue	74,303	1,745	2,554	4,299	5.47	78,602

retraining with updated benchmark papers (i.e., the retrained test-taker group). Table 23 shows the average counts of advisory flag types that resulted from using the retrained samples for argument and issue essays.

Level 1 Analyses

Table 24 shows the aggregate summary of the standard evaluations (i.e., standardized mean score difference, QWK, correlation, and degradation) for the argument and issue G-10 models on the retrained test-taker data set. We performed these evaluations at the prompt level as opposed to the VS and VG levels, because we observed no meaningful differences for the variants or variant groups examining agreement at the rating level for argument and issue.

Table 24 shows that the G-10 model meets guideline thresholds for all agreement statistics at the aggregate level across all prompts for both argument and issue tasks. However, comparison of the argument and issue task shows that the argument task has consistently lower human – machine agreement statistics compared to the issue task. This is supportive evidence of the missing e-rater feature that is able to extract argumentation.

Appendix H shows the prompt-by-prompt evaluations between human – human and human – machine ratings agreement for the test takers from the retrained group. Table H1 in Appendix H shows the evaluation results for each prompt for the 10-feature generic e-rater model for the argument task. Table H2 in Appendix H shows the evaluation results for each prompt for the 10-feature generic e-rater model for the issue task.

Table 25 shows a summary of the number of flagged prompts that are individually presented in Appendix H. These are the counts of prompts within each task that violate the guideline threshold levels separately for argument and issue ratings for the G-10 models built using the retrained test-taker samples and evaluated on the same data set.

Table 25 provides additional evidence that the argument task on a prompt-by-prompt basis has more threshold guideline violations than the issue task. There is almost twice as many threshold violations for the argument task than for

Table 24 Phase III Average Agreement of Human Scores on Argument and Issue Prompts: G-10 Model

		H1 by H2									H1 by e-rater (rounded to integers)					H1 by e-rater (unrounded)			
		H1		H2		Statistic					e-rater		Statistic			e-rater		Statistic	
Prompt	*N*	*M*	*SD*	*M*	*SD*	Std diff	Wtd kappa	% agree	% adj agree	*r*	*M*	*SD*	Wtd kappa	% agree	% adj agree	*M*	*SD*	Std diff	*r*
Argument	748	3.37	0.94	3.36	0.94	−0.01	0.75	63	98	0.75	3.36	0.95	0.73	59	98	3.36	0.92	−0.01	0.77
Issue	773	3.36	0.87	3.35	0.87	−0.01	0.75	65	99	0.75	3.36	0.89	0.77	67	99	3.35	0.85	0.00	0.82

Note. N is average across all the prompts. adj = adjacent; H1 = Human 1; H2 = Human 2; Std diff = standardized difference; wtd = weighted.

Table 25 Phase III Performance of G-10 Model Flag Counts Across All 76 Prompts

	Argument				Issue			
	Std diff(unrnd)	Wtd kappa	Deg.wtd kappa	Flag exists	Std diff (unrnd)	Wtd kappa	Deg.wtd kappa	Flag exists
Retrained	21	12	1	25	11	5	0	13

Note. Std diff = standardized difference; unrnd = unrounded; wtd = weighted; deg = degradation.

the issue task, again providing supportive evidence of the missing feature that measures the argumentation aspect of the writing construct.

Table I1 in Appendix I shows the subgroup differences for the argument prompts and Table I2 shows the subgroup differences for the issue prompts using the G-10 model as a check score. Standardized mean score differences are within thresholds for subgroups of gender, U.S. racial and ethnic groups, regardless of task, with sample sizes greater than 200. The exception is the American Indian and Alaskan Native subgroup, who responded to the argument prompts. However, for the argument task, test takers from China and Taiwan show large score separation between human and e-rater for the test-taker group scored following retraining in opposite directions, that is, e-rater scores are higher for Chinese test takers compared to human raters, and e-rater scores for Taiwan test takers are lower compared to human raters. This result for test takers from Taiwan and China confirms previous research (Bridgeman et al., 2009; Ramineni et al., 2010), but are contrary to our Phase II results that occurred prior to retraining (see Table 14). The retraining of human raters, a different examinee cohort (i.e., initial vs. retrained), and a new AES model with the DWU macrofeature are three aspects that may have contributed to this observed difference in subgroup performance.

This opposite direction for test takers from Taiwan and China is not evident in the issue task results presented in Table I2. Subgroup performances for China (standardized difference = 0.35) are positive and near 0 for Taiwan (standardized difference = 0.04) for test takers from the retrained group on issue prompts. This finding may be the result again of the absence of a feature that adequately represents argumentation evidence (Bejar, Flor, Futagi, & Ramineni, in press; Burstein, Beigman-Klebanov, Madnani, & Somasundaran, 2013; Deane, Williams, Weng, & Trapani, 2013) that is not required by the issue task, but is required by the argument task.

Level 2 analyses. Table 26 shows the cross-task correlations and correlations with external variables using e-rater as a check score. As noted previously, the task score of issue and argument are the operational scores from the double-human scoring, while the argument and issue scores under the different adjudication thresholds of 1.5, 1.0, and 0.5 are the simulated argument and issue task scores if the G-10 model was used as a check score for those writing tasks.

Table 26 shows that the correlations of the two tasks with the GRE Quantitative Reasoning measure scores are uniformly low while the correlations with verbal are higher, showing evidence of convergent and divergent validity. The correlations of the argument and issue tasks with the number of words show that longer essays receive higher scores. The

Table 26 Phase III Cross-Task Correlations Under Multiple Check-Score Thresholds: G-10 Model

		External		Words		Task score		1.5		1.0		0.5	
		Quantitative	Verbal	Issue	Argument	Issue	Argument	Issue	Argument	Issue	Argument	Issue	Argument
External	Quantitative	1.00	0.32	0.25	0.29	0.08	0.20	0.08	0.19	0.09	0.19	0.09	0.20
	Verbal	0.32	1.00	0.37	0.45	0.60	0.64	0.56	0.61	0.57	0.62	0.59	0.64
Words	Issue			1.00	0.80	0.72	0.55	0.68	0.53	0.70	0.54	0.73	0.56
	Argument			0.80	1.00	0.63	0.69	0.60	0.67	0.62	0.69	0.64	0.71
Task score	Issue					1.00	0.69	0.94	0.66	0.95	0.68	0.97	0.70
	Argument					0.69	1.00	0.65	0.94	0.67	0.95	0.69	0.97
1.5	Issue							1.00	0.62	0.99	0.64	0.97	0.66
	Argument							0.62	1.00	0.64	0.99	0.66	0.97
1.0	Issue									1.00	0.65	0.98	0.67
	Argument									0.65	1.00	0.67	0.98
0.5	Issue											1.00	0.69
	Argument											0.69	1.00

Note. $N = 47{,}676$.

Table 27 Phase III Percentage of Score Differences at the Task Level Between the Use of e-rater as a Check Score Versus All-Human Double Scoring for the Argument Task

Model	Threshold	≤ −1.5	−1	−0.5	0	0.5	1	≥ 1.5
Generic (no CVA)	0.5	0.01	0.18	7.83	83.49	8.00	0.40	0.10
	1.0	0.04	0.18	13.65	71.28	14.28	0.41	0.18
	1.5	0.07	0.25	15.98	66.33	16.77	0.37	0.23

Note. $N = 59{,}660$. CVA = content vector analysis.

Table 28 Phase III Percentages of Scores Using e-rater as a Check Score That Differ From an All-Human Double Scoring for the Issue Task

Model	Threshold	≤ −1.5	−1	−0.5	0	0.5	1	≥ 1.5
Generic (no CVA)	0.5	0	0.08	7.55	83.52	8.52	0.29	0.02
	1.0	0.01	0.14	14.08	70.07	15.35	0.3	0.06
	1.5	0.03	0.22	16.01	66.42	16.92	0.28	0.13

Note. $N = 59{,}544$. CVA = content vector analysis.

cross-task correlation of 0.69 shows that the tasks are reliable estimates of each other, accounting for slightly less than half the variance. Also, the correlations of issue and argument under the different adjudication thresholds show the highest correlations with adjudications of 0.5 with verbal, but also the cross-task correlations are the highest at the adjudication threshold of 0.5. This is further supportive evidence that the adjudication threshold of 0.5 is most appropriate for both tasks in the rGRE Analytical Writing measure.

Table 27 shows the simulated task score using e-rater as a check score for argument, and Table 28 shows the same information for issue at different adjudication thresholds of 0.5, 1.0, and 1.5 comparing those results with all-human double scoring.

There are few differences in Table 27 beyond ±0.5 for the different adjudication thresholds under consideration compared to all-human scoring when using G-10 model for either the argument or issue task. Recall that when computing the rGRE Analytical Writing score, the task scores are rounded to 0.5 intervals; these results show less than 1% of all test takers are outside the ±0.5 allowable difference.

Similar results are shown for the issue task in Table 28 in that less than 1% of all test takers are outside the allowable difference of ±0.5 regardless of the adjudication threshold.

These results show that using the G-10 model as a check score for either task at the adjudication level of 0.5 shows little difference compared to all-human scoring. Further, there is little to differentiate the different adjudication thresholds given these results, with the exception that the smallest deviation from all-human scoring occurs when using the 0.5 adjudication threshold in a check-score application.

Level 3 analyses. Table 29 presents the correlations at the analytical writing score level between external and internal measures of convergent and divergent validity. Recall that the correlations under the different adjudication thresholds of 1.5, 1.0, and 0.5 are the simulated check scores using e-rater with a PS-12 model.

Table 29 Phase III Convergent and Divergent Correlations at the Writing Score Level for the Test Takers Scored by Retrained Raters

	Quantitative	Verbal	Average words	Writing raw	1.5	1.0	0.5
Quantitative	1.00	0.32	0.29	0.15	0.15	0.16	0.16
Verbal		1.00	0.43	0.67	0.65	0.65	0.67
Average words			1.00	0.74	0.72	0.74	0.75
Writing raw				1.00	0.96	0.97	0.98
1.5					1.00	0.99	0.98
1						1.00	0.99
0.5							1.00

Note. $N = 47{,}676$.

Table 30 Phase III Percentages of Scores Using e-rater as a Check Score That Differ From All-Human Scoring for the Analytical Writing Score

Model	Threshold	≤ −0.75	−0.5	−0.25	0	0.25	0.5	≥ 0.75
Generic (no CVA)	0.5	0.03	0.88	12.77	70.99	13.76	1.39	0.18
	1.0	0.08	2.22	19.45	54.42	20.68	2.77	0.36
	1.5	0.19	2.88	21.11	49.96	22.1	3.26	0.5

Note. $N = 47{,}676$. All candidates from March 18, 2012, through June 18, 2012. CVA = content vector analysis.

Table 29 shows the correlations at the writing score level are highest with the external GRE Verbal Reasoning measure at the 0.5 adjudication threshold and are almost 1 at that level with the all-human writing score. These results are consistent with what was observed in Phase II.

Table 30 shows the differences between the simulated score using e-rater as a check score for human scoring at the analytical writing score level comparing the check-score model to the current all-human scoring model for the matched set of test takers from the retrained period. Recall that the argument and issue rGRE Analytical Writing tasks are averaged and rounded to the nearest 0.25 score point. The results show few differences between the simulated scores using e-rater as a check score versus all-human double scoring extending beyond an acceptable ±0.25 difference.

Phase IV Results: Building the G Models on the Retrained Data and Cross-Evaluating Them on the Initial Cohort Data and Evaluating Trend Data for Bias

Level 1 Analysis

Table 31 shows the aggregate results of the G models built on the test-taker samples scored by the retrained raters, but evaluated on the initial cohort test takers for both argument and issue tasks.

Examination of Table 31 shows that use of the G-10 AES model meets all threshold criteria at the rater level. Recall that there was a concern that the models built using the test takers from the retrained period were less able and that would manifest itself in the standardized mean score differences at the human–machine rating level when applied to the test takers in the fall when motivation is thought to be high to enter graduate study. Further recall that this analysis provides a separate crossover evaluation examining the extent of any seasonality that might occur as a result of building a model on a test-taker sample selected from a specific time period (such as from the retrained sample) and applying that model on a different test-taker cohort. The results in Table 31 show that, on average, the human–machine agreement is adequate. While there is a consistent negative standardized mean score difference, it does not rise to the level of caution.

Level 2 Analysis

Table 32 shows the number of guideline threshold advisory violations for the different models and time periods for comparison purposes for human–machine agreement at the rater level.

Results in Table 32 show the largest number of violations do occur for the model developed on the test-taker group scored by the retrained raters and evaluated on the crossover group from the initial cohort period.

Table 31 Phase IV Average Agreement of Human Scores on Argument and Issue Prompts: Generic Model

		Human 1 by Human 2									Human 1 by e-rater (rounded to integers)							Human 1 by e-rater (unrounded)			
		Human 1		Human 2		Statistic					e-rater		Statistic					e-rater		Statistic	
Prompt	*N*	*M*	*SD*	*M*	*SD*	Std diff	Wtd kappa	% agree	% adj agree	*r*	*M*	*SD*	Std diff	Wtd kappa	% agree	% adj agree	*r*	*M*	*SD*	Std diff	*r*
Argument	2,242	3.47	1.04	3.46	1.05	−0.01	0.79	62	98	0.79	3.38	1.04	−0.08	0.78	58	98	0.79	3.38	1.04	−0.08	0.81
Issue	2,302	3.47	1.01	3.48	1.00	0.01	0.80	64	99	0.80	3.36	0.98	−0.11	0.82	65	100	0.82	3.35	0.95	−0.13	0.86

Note. adj = adjacent; Std diff = standardized difference; wtd = weighted.

Table 32 Phase IV Performance of G-10 Model Flag Counts for Argument and Issue Across 76 Prompts

	Argument				Issue			
	Std diff (unrnd)	Wtd kappa	Deg. wtd kappa	Flag exists	Std diff (unrnd)	Wtd kappa	Deg. wtd kappa	Flag exists
Jump start	15	27	0	32	15	6	0	18
Retrained	21	12	1	25	11	5	0	13
Crossover[a]	60	56	0	68	48	16	0	48

Note. Std diff = standardized difference; unrnd = unrounded; wtd = weighted; deg = degradation. [a]Model build taken from retrained period but cross-validated against initial cohort candidates.

Table 33 Effects of Biased e-rater Scores for the Argument Essays at the Adjudication Threshold of 0.5 on the Trend Test Takers

	Bias					
	Baseline		−0.1		−0.2	
Classification of simulated score	Frequency	Percentage	Frequency	Percentage	Frequency	Percentage
No problem (no H2 invoked)	1,036	57.88	1,012	56.54	947	52.91
H2 invoked and in threshold	734	41.01	761	42.51	822	45.92
High outlier	14	0.78	13	0.73	14	0.78
Low outlier	6	0.34	4	0.22	7	0.39
Double-outlier	0	0	0	0	0	0
Missing e-rater/Flagged e-rater	0	0	0	0	0	0
Task score						
0	0	0	0	0	0	0
0.5	0	0	0	0	0	0
1	11	0.61	11	0.61	11	0.61
1.5	9	0.5	9	0.5	9	0.5
2	163	9.11	165	9.22	165	9.22
2.5	76	4.25	79	4.41	81	4.53
3	578	32.29	589	32.91	596	33.3
3.5	164	9.16	159	8.88	169	9.44
4	548	30.61	546	30.5	531	29.66
4.5	85	4.75	80	4.47	86	4.8
5	127	7.09	122	6.82	112	6.26
5.5	18	1.01	19	1.06	19	1.06
6	11	0.61	11	0.61	11	0.61

Note. H2 = Human 2.

Tables 33 and 34 show the results of purposely adding a bias of 0 (i.e., the baseline of no bias), a bias of −0.1, or −0.2 to the G-10 e-rater model and assessing the effect that bias has on the number of second human raters for argument and issue prompts, respectively, when the adjudication level is 0.5.

Recall that these data were computed using the trend test-taker cohort tested at the height of the graduate school admission period (November 2011) and are thought to be quite motivated. The bias is negative in order to mimic the effect of training the e-rater model on a less able group (i.e., those tested in between March 18, 2012, and June 18, 2012) that may yield artificially lower e-rater scores than if it was trained on a more motivated test-taker cohort from the fall of 2011. We evaluated each of these biased e-rater scores at the additional adjudication thresholds of 1.0 and 1.5. We present these results of the effect of biased e-rater scores in Appendix J in Tables J1 and J2 for argument with adjudication thresholds of 1.0 and 1.5. Tables J3 and J4 display the results for the issue prompts with adjudication thresholds of 1.0 and 1.5, respectively.

Level 3 Analysis

Table 35 shows the effects, at the analytical writing score level, of the number of additional adjudications at each of the bias levels of no bias, −0.1, and −0.2 bias for the adjudication level of 0.5.

Table 34 Effects of Biased e-rater Scores for the Issue Essays at the Adjudication Threshold of 0.5 on the Trend Test Takers

	Bias					
	Baseline		−0.1		−0.2	
Classification of simulated score	Frequency	Percentage	Frequency	Percentage	Frequency	Percentage
No problem (no H2 invoked)	1,190	66.48	1,148	64.13	1,067	59.61
H2 invoked and in threshold	600	33.52	640	35.75	719	40.17
High outlier	0	0	2	0.11	3	0.17
Low outlier	0	0	0	0	1	0.06
Double-outlier	0	0	0	0	0	0
Missing e-rater/Flagged e-rater	0	0	0	0	0	0
Task score						
0	0	0	0	0	0	0
0.5	0	0	0	0	0	0
1	11	0.61	12	0.67	12	0.67
1.5	7	0.39	8	0.45	8	0.45
2	138	7.71	147	8.21	153	8.55
2.5	78	4.36	75	4.19	77	4.3
3	589	32.91	599	33.46	613	34.25
3.5	159	8.88	154	8.6	148	8.27
4	606	33.85	602	33.63	590	32.96
4.5	76	4.25	77	4.3	76	4.25
5	108	6.03	97	5.42	93	5.2
5.5	10	0.56	11	0.61	12	0.67
6	8	0.45	8	0.45	8	0.45

Note. H2 = Human 2.

Table 35 Effects of Biased e-rater Scores for the Analytical Writing Score at the Adjudication Threshold of 0.5

	Score difference group									
Bias	≤ −1.5	−1	−0.5	−0.25	0	0.25	0.5	1	≥ 1.5	Total
Baseline	0	0	22	217	877	595	79	0	0	1,790
−0.1	0	0	29	257	890	555	59	0	0	1,790
−0.2	0	0	33	282	898	530	47	0	0	1,790

The results presented in Table 35 show somewhat surprising increases in agreement with all-human scoring at the adjudication threshold of 0.5; as the bias is increased, the agreement of using e-rater as a check score with all-human scoring increases as well. Note that since the final analytical writing scores are rounded to the nearest 0.25 score point, the numbers of examines who are within ±0.25 score points increases as the bias is increased. This means that, as the bias increases, the difference between using the G-10 models as a check score makes less of a difference compared to all-human scoring, most likely because the number of adjudications increases.

Appendix K shows somewhat similar results of 0 score differences increasing modestly at the analytical writing score level with the threshold at 1.0 (Table K1) and 1.5 (Table K2). It appears that the bigger the adjudication threshold is, the larger allowable difference between e-rater and the first human score contributes to the observation of 0 differences at the writing score level.

Discussion

Phase I Implications: Evaluating the Human Scoring Consistency of the Different Prompts and Variants

In Phase I of the study, we verified that the different prompts, prompt variants, and variant groups were scored consistently by human raters. The small standardized mean score differences for prompts, variants, and variant groups show that, for the most part, human scoring differences at the rating level are small. These human agreement results lead us to explore the use of e-rater for scoring the analytical writing tasks.

Phase II Implications: Building and Evaluating PS, VS, and VG Versus G Models on the Initial Cohort

In Phase II of the study, we compared agreement of eight different AES models at the rating level with human scores. Four of the models were built for each prompt, variant, variant group, and generic applications. Each of these models incorporated two content features or not (i.e., CVA). All agreement statistics for each of these different models met or exceeded guideline threshold limits at the aggregate summary level for both argument and issue essays. Consistent with past research on automated scoring (Ramineni et al., 2012), the models with the fewest flags are the prompt-specific models. These results are not surprising, since each PS model is custom tailored to each group of essays written to the specific prompt. What is most interesting about these results is that the two content features added very little to the reduction in flags of guideline threshold violations; those expected differences were not noticeable, most likely due to the incorporation of the generic differential word use (DWU) feature (Attali, 2011). DWU already takes into consideration the kinds of words associated with high and low scores, thus making the use of the CVA features redundant.

Group differences showing score separation between human and AES ratings occur consistently for test takers tested in Asia1, China, and Taiwan test centers for the argument prompt. The interesting finding is that, contrary to previous findings, Taiwan and China display AES separation from human scoring in the same direction. For essays written to the issue prompts only, those test takers tested in Taiwan test centers continue with large human – machine score separation. This departure from the Ramineni et al. (2010) and Ramineni et al. (2012) observation may be due to the use of DWU in the models. We believe these subgroup differences are unique to this study, since the current version of e-rater no longer uses the DWU macrofeature.

At the task score level, cross-task correlations support some commonality between the issue and argument scores ($r_{ai} = 0.66$), as well as both divergent and convergent validity in that essay scores on each task correlate low with quantitative scores on the GRE General Test and higher with the verbal score. The differences observed at the task score level between the PS model and the G model versus all-human scoring for argument and issue prompts is very small, leading us to recommend the use of the generic model with a 0.5-point adjudication threshold consistent with past practice (i.e., Ramineni et al., 2012).

At the writing score level, there are few differences between the PS and G models in divergent and convergent correlation, leading us to recommend the use of a G model without CVA for use as a check-score implementation. Also, there is very little difference at the analytical writing score level between all-human and the use of e-rater as a check score, accounting for 97.9% agreement at ±0.25 score points between all-human and the e-rater moderated check-score model at the 0.5 adjudication threshold. These results again show the difference at the test score level between the PS and G models is negligible at the adjudication threshold of 0.5.

Phase III Implications: Building and Evaluating G Models on the Retrained and Crossover Data

Our discovery that essays written to the argument prompts have more agreement threshold violations than essays written to the issue prompts is not surprising, since e-rater v12.1 does not have a feature that can measure argumentation in writing (Burstein et al., 2013; Deane et al., 2013). Also, the correlations with external variables including verbal scores are higher for argument than for issue essays, possibly indicating more verbal skills are required to produce a cogent argument than is required by the issue prompt. Subgroup differences for argument prompts in the retrained test-taker samples are consistent with past observations in that e-rater scores are higher for test takers tested in China, but lower for those tested in Taiwan in relation to human scores (Ramineni et al., 2010). Subgroup differences for issue prompts when using e-rater as a check score with the G model were fewer, except for test takers tested in China.

At the task score level, there are few differences between the use of e-rater as a check score and the all-human double-scoring model. For the argument prompts, 99.3% of all scores using e-rater as a check score are within ±0.25 score points of all-human scoring; for the issue prompts, the percentage within ±0.25 score points is 99.6%.

At the analytical writing score level, the difference between the use of e-rater as a check of the human score and all-human scoring is within ±0.25 score points 97.5% of the time when using an adjudication level of 0.5. Further, the analytical writing score is convergent with the verbal GRE score at $r_{wv} = 0.67$ and maintains that high correlation at the adjudication level of 0.5; the correlation of the writing score with the GRE Quantitative Reasoning measure score is suitably low, 0.15, supporting the use of e-rater as a check score from a convergent and divergent validity perspective.

Phase IV Implications: Building the G Models on the Retrained Data and Cross-Evaluating Them on the Initial Cohort and Evaluating Trend Data for Bias

Using the G models on the trend data show that even with a large bias, the use case of deploying e-rater as a check score makes little difference in the task score, as well as in the final analytical writing score. Even though the number of adjudications increases as the bias increases, the percentage of those additional adjudications is small. For the argument task, the increase is less than 5% at an adjudication threshold of 0.5 score points and less than 3% for the issue task. At the analytical writing score level, the percentage of scores within ±0.25 score points actually *increases* the agreement with all-human scoring because there are more adjudications and, thus, more double-human scoring. The discrepancies with all-human scoring are reduced when more bias is introduced to each task at the analytical writing score level, since we are comparing all-human scoring to an increase in double-human scoring as a result of the increase in adjudications at each task level. The increase in agreement is small (about 1.25% at the adjudication threshold of 0.5).

Limitations of the Study

In this study, we examined the use of AES models as a check score as a sole objective. Often that limited choice is based on perceived vulnerabilities that indeed do exist. These vulnerabilities can include, but are not limited to:

- Scoring unusual or gamed responses intended to produce higher automated scores compared to scores assigned by humans that could threaten the scoring model predictions (Bejar et al., in press; Powers et al., 2002);
- Scoring test takers whose first language is not English in a manner that is widely disparate from human scores (Bridgeman et al., 2012; Ramineni et al., 2010); and
- Scoring excessively long or short essay responses, or responses that are in a foreign language, or that are far off topic, hence posing threats to score validity (Higgins, 2013; Higgins, Burstein, & Attali, 2006).

Today, we consider AES as having the capability to provide complementary construct-relevant aspects to human scoring, as opposed to a replacement of human scoring (Attali, 2013). In the future, we think the program should explore the use of AES as a contributing score that should include the exploration of differential weights of AES to contribute to the GRE Analytical Writing score, as well as the capability to predict external criteria such as first year of graduate school writing performances. Future research should explore the use of alternate covariates to provide a more comprehensive detailed analysis of the construct. Such research would detail the relationship of different AES features to important aspects of writing that could provide a sound basis for future use of AES as a contributory score.

We caution against comparing the human–human mean scores among the variants and variant groups, since the assignment of prompts is not random, nor are the test takers selected at random. Each variant and variant group is packaged and deployed in a computer-based testing window. Further, test takers self-select when to take the test.

Another limitation of this study was the restricted sample used to calibrate the recommended models that fused the timing of the rater retraining with the 2-month spring 2012 cohort. This cohort took the GRE General Test after many graduate programs had closed the application process for the autumn of 2012; thus, they may have not been as motivated as desired. Future research should employ a more representative sample for model building and evaluations.

Also, we note that the engine used in this research was e-rater v12.1, which employed the use of DWU as a macrofeature. As noted earlier, DWU is a generic content-based feature assigning words used in an essay with weights associated with high and low scores, similar to what the two PS-12 and G-12 content features do on a prompt-by-prompt basis. Some of our results, such as the finding that the use of G-12 and PS-12 models were just as effective as G-10 and PS-10 models at the task score and analytical writing score level, may not hold true should DWU not be included in an AES model. Also, this research did not compare engine v12.1 with any other engine version.

Conclusion

e-rater's use as a check on a human score was investigated as an alternate approach to a contributory score. Under the check-score approach, the first human score was checked for agreement with the e-rater score within an empirically established range, beyond which a second human score was required. The first human score became the final score for the task, unless a second human rating was required. Various agreement thresholds were evaluated under the check-score model to minimize differences across the subgroups. A discrepancy threshold of half-a-point between the automated and the

human score was selected for e-rater to yield performance similar to double-human scoring, but with significant savings in second human ratings. Also, since we built the e-rater models used in this study on a select range of test takers tested between March 18 and June 18, 2012, that were thought to be of low proficiency or perhaps unmotivated, we evaluated the effect of a bias where e-rater would return scores that are lower compared to human raters. Given the nature of the check score approach, we found very little bias effect at the task score level and less so at the analytical writing score level.

General Recommendations

Our first recommendation is that the use of e-rater should be implemented as a check score using G models for argument and issue prompts with a 0.5 adjudication threshold in operational scoring of the GRE Analytical Writing measure. As part of ongoing efforts, it will be critical to monitor and evaluate e-rater performance in operation from time to time, owing to the anticipated changes in the overall test format, test-taker and human-rater characteristics, and human scoring trends over time, as well as new feature developments and enhancements in the e-rater engine. We will investigate the differences in e-rater and human scores observed for some subgroups in this evaluation to better understand their source and origin. We will also investigate the contribution of e-rater to score validity when combined with human scoring in a weighted contributory model. Thus, we recommend the program continue to collect a reliability sample of 5% at the test-taker level so that we can continue to monitor the performance of the e-rater models on continued operational samples and support the program by providing additional research on the use of automated scoring.

Notes

1 Formula for standardized difference of the mean or the effect size also referred to as Cohen's *d*: $d = \frac{\left[\overline{X}_{AS} - \overline{X}_{H}\right]}{\sqrt{\frac{SD_{AS}^2 + SD_{H}^2}{2}}}$, where $\overline{X}_{AS}$ is the mean of the automated score, $\overline{X}_{H}$ is the mean of the human score, SD_{AS}^2 is the variance of the automated score, and SD_{H}^2 is the variance of the human score.

2 *Fall* refers to data collected sometime during August through December 2011.

References

Attali, Y. (2009, April). *Evaluating automated scoring for operational use in consequential language assessment—the ETS experience*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Diego, CA.

Attali, Y. (2011). *A differential word use measure for content analysis in automated essay scoring* (Research Report No. RR-11-36). Princeton, NJ: Educational Testing Service.

Attali, Y. (2013). *Reliability-based feature weighting for automated essay scoring*. Manuscript submitted for publication.

Attali, Y., Bridgeman, B., & Trapani, C. (2010a). Performance of a generic approach in automated essay scoring. *Journal of Technology, Learning, and Assessment*, *10*(3), 1–16.

Attali, Y., Bridgeman, B., & Trapani, C. (2010b). Performance of a generic approach in automated essay scoring. *Journal of Technology, Learning, and Assessment*, *10*(3), 1–16. Retrieved from http://www.editlib.org/p/106317

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

Bejar, I. I., Flor, M., Futagi, Y., & Ramineni, C. (in press). Effect of a construct-irrelevant response strategy (CIRS) on automated scoring of writing. *Assessing Writing*.

Breyer, F. J., Ramineni, C., Duchnowski, M., Harris, A., & Ridolfi, L. (2012). *E-rater engine upgrade from v.11.1 Linux to 12.1 Linux* (Statistical Report No. SR-2013-012). Princeton, NJ: Educational Testing Service.

Bridgeman, B., Trapani, C., & Attali, Y. (2009, April). *Considering fairness and validity in evaluating automated scoring*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Diego, CA.

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*, 27–40.

Burstein, J., Beigman-Klebanov, B., Madnani, N., & Somasundaran, S. (2013). *Investigations of the structure of argumentation, opinion, and stance, and test-taker writing quality*. Unpublished manuscript, Educational Testing Service, Princeton, NJ.

Burstein, J., Tetreault, J., & Madnani, N. (2013). The e-rater automated essay scoring system. In M. D. Shermis & J. Burstein (Eds.), *Handbook of automated essay evaluation: Current applications and new directions* (pp. 55–67). New York, NY: Routledge Academic.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, *56*, 81–105.

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

Deane, P., Williams, F., Weng, V., & Trapani, C. (2013). Automated essay scoring in innovative assessments of writing from sources. *Journal of Writing Assessment*, *6*(1), 1 – 16.

Educational Testing Service. (2013). *About the GRE revised General Test*. Princeton, NJ: Educational Testing Service.

Fleiss, J. L., & Cohen, J. (1973). The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and Psychological Measurement*, *33*, 613 – 619.

Gamon, M., Chodorow, M., Leacock, C., & Tetreault, J. (2013). Grammatical error detection in automatic essay scoring and feedback. In M. Shermis & J. Burstein (Eds.), *Handbook of automated essay evaluation* (pp. 251 – 266). New York, NY: Routledge Academic.

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

Higgins, D., Burstein, J., & Attali, Y. (2006). Identifying off-topic student essays without topic-specific training data. *Natural Language Engineering*, *12*, 145 – 159.

Joe, J., Balani, M., Napoli, A., & Chen, J. (2013). *An investigation of differences in GRE Analytical Writing measure operational scoring quality and productivity between test-section level and variant-group level calibration*. Manuscript in preparation.

Leacock, C. & Chodorow, M. (2003). Automated grammar error detection. In M. D. Shermis & J. C. Burstein (Eds.), *Automated essay scoring: A cross-disciplinary perspective* (pp. 195 – 207). Mahwah, NJ: Lawrence Erlbaum Associates.

Petersen, N. S. (1997, March). *Automated scoring of writing essays: Can such scores be valid?* Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, IL.

Powers, D., Burstein, J. C., Chodorow, M., Fowles, M. E., & Kukich, K. (2002). Stumping e-rater: Challenging the validity of automated essay scoring. *Computers in Human Behavior*, *18*, 103 – 134.

Quinlin, T., Higgins, D., & Wolff, S. (2009). *Evaluating the construct coverage of the e-rater scoring engine* (Research Report No. RR-09-01). Princeton, NJ: Educational Testing Service.

Ramineni, C., Trapani, C. S., Williamson, D. M., Davey, T., & Bridgeman, B. (2012). *Evaluation of e-rater for the GRE issue and argument prompts* (Research Report No. RR-12-02). Princeton, NJ: Educational Testing Service.

Ramineni, C., Williamson, D. M., & Weng, V. (2010, April). *Understanding mean score differences between e-rater and humans for demographic-based groups in GRE*. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.

Weigle, S. C. (2010). Validation of automated scores of TOEFL iBT tasks against non-test indicators of writing ability. *Language Testing*, *27*, 335 – 353.

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

Appendix A

rGRE Analytical Writing Measure Scoring Guide

Score	GRE scoring guide (argument)	GRE scoring guide (issue)
6	In addressing the specific task directions, a 6 response presents a cogent, well-articulated examination of the argument and conveys meaning skillfully. A typical response in this category: • Clearly identifies aspects of the argument relevant to the assigned task and examines them insightfully • Develops ideas cogently, organizes them logically, and connects them with clear transitions • Provides compelling and thorough support for its main points • Conveys ideas fluently and precisely, using effective vocabulary and sentence variety • Demonstrates superior facility with the conventions of standard written English (i.e., grammar, usage, and mechanics) but may have minor errors	In addressing the specific task directions, a 6 response presents a cogent, well-articulated analysis of the issue and conveys meaning skillfully. A typical response in this category: • Articulates a clear and insightful position on the issue in accordance with the assigned task • Develops the position fully with compelling reasons and/or persuasive examples • Sustains a well-focused, well-organized analysis, connecting ideas logically • Conveys ideas fluently and precisely, using effective vocabulary and sentence variety • Demonstrates superior facility with the conventions of standard written English (i.e., grammar, usage, and mechanics) but may have minor errors

Appendix A: Continued

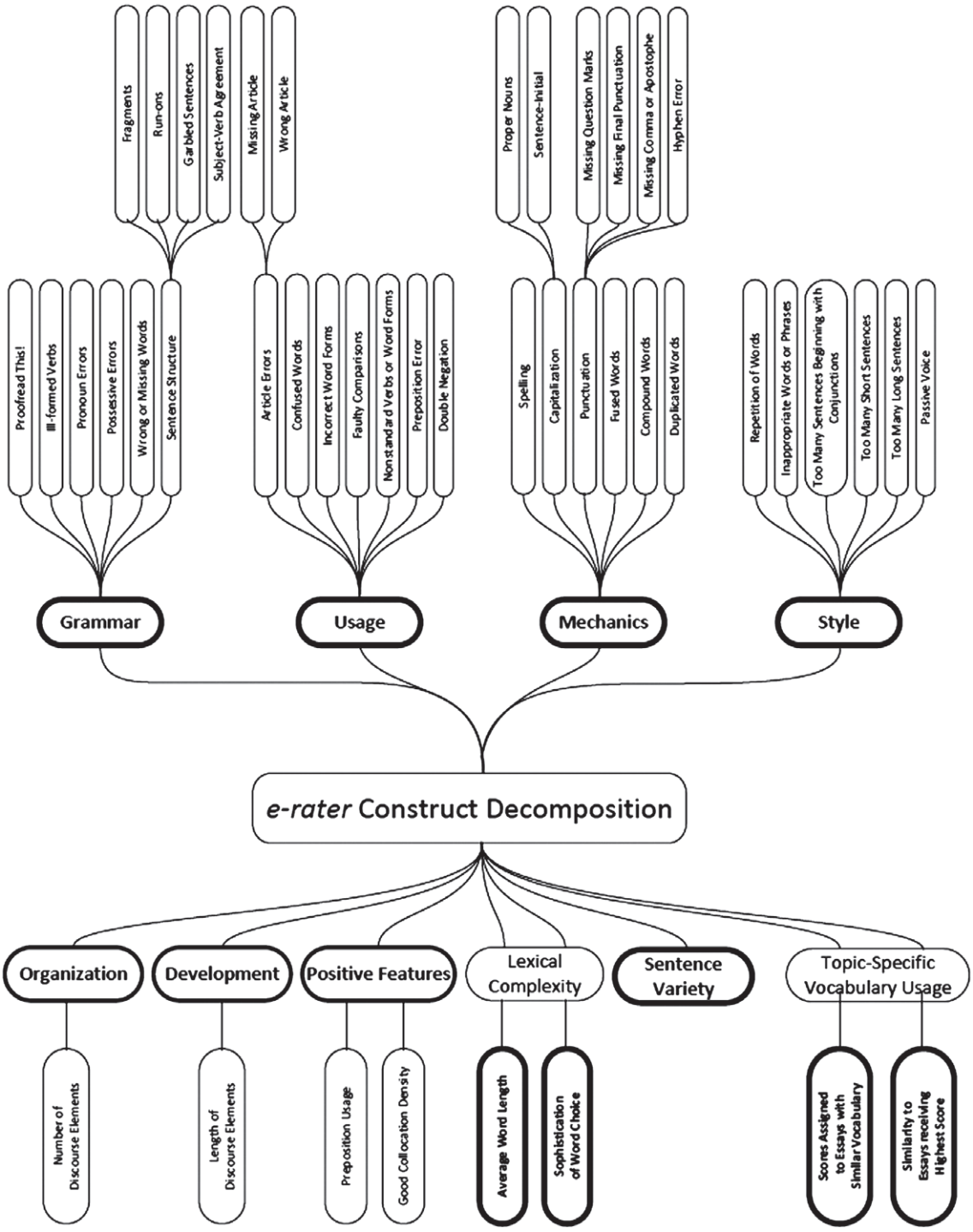
Score	GRE scoring guide (argument)	GRE scoring guide (issue)
5	In addressing the specific task directions, a 5 response presents a generally thoughtful, well-developed examination of the argument and conveys meaning clearly. A typical response in this category: • Clearly identifies aspects of the argument relevant to the assigned task and examines them in a generally perceptive way • Develops ideas clearly, organizes them logically, and connects them with appropriate transitions • Offers generally thoughtful and thorough support for its main points • Conveys ideas clearly and well, using appropriate vocabulary and sentence variety • Demonstrates facility with the conventions of standard written English but may have minor errors	In addressing the specific task directions, a 5 response presents a generally thoughtful, well-developed analysis of the issue and conveys meaning clearly. A typical response in this category: • Presents a clear and well-considered position on the issue in accordance with the assigned task • Develops the position with logically sound reasons and/or well-chosen examples • Is focused and generally well organized, connecting ideas appropriately • Conveys ideas clearly and well, using appropriate vocabulary and sentence variety • Demonstrates facility with the conventions of standard written English but may have minor errors
4	In addressing the specific task directions, a 4 response presents a competent examination of the argument and conveys meaning with acceptable clarity. A typical response in this category: • Identifies and examines aspects of the argument relevant to the assigned task but may also discuss some extraneous points • Develops and organizes ideas satisfactorily but may not connect them with transitions • Supports its main points adequately but may be uneven in its support • Demonstrates sufficient control of language to convey ideas with acceptable clarity • Generally demonstrates control of the conventions of standard written English but may have some errors	In addressing the specific task directions, a 4 response presents a competent analysis of the issue and conveys meaning with acceptable clarity. A typical response in this category: • Presents a clear position on the issue in accordance with the assigned task • Develops the position with relevant reasons and/or examples • Is adequately focused and organized • Demonstrates sufficient control of language to convey ideas with acceptable clarity • Generally demonstrates control of the conventions of standard written English but may have some errors
3	A 3 response demonstrates some competence in addressing the specific task directions, in examining the argument and in conveying meaning, but is obviously flawed. A typical response in this category exhibits ONE OR MORE of the following characteristics: • Does not identify or examine most of the aspects of the argument relevant to the assigned task, although some relevant examination of the argument is present • Mainly discusses 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 its main points • As problems in language and sentence structure that result in a lack of clarity • Contains occasional major errors or frequent minor errors in grammar, usage, or mechanics that can interfere with meaning	A 3 response demonstrates some competence in addressing the specific task directions, in analyzing the issue and in conveying meaning, but is obviously flawed. A typical response in this category exhibits ONE OR MORE of the following characteristics: • Is vague or limited in addressing the specific task directions and/or in presenting or developing a position on the issue • Is weak in the use of relevant reasons or examples or relies largely on unsupported claims • Is limited in focus and/or organization has problems in language and sentence structure that result in a lack of clarity • Contains occasional major errors or frequent minor errors in grammar, usage, or mechanics that can interfere with meaning

Appendix A: Continued

Score	GRE scoring guide (argument)	GRE scoring guide (issue)
2	A 2 response largely disregards the specific task directions and/or demonstrates serious weaknesses in analytical writing. A typical response in this category exhibits ONE OR MORE of the following characteristics: • Does not present an examination based on logical analysis, but may instead present the writer's own views on the subject • Does not follow the directions for the assigned task • Does not develop ideas, or is poorly organized and illogical • Provides little, if any, relevant or reasonable support for its main points • Has serious problems in language and sentence structure that frequently interfere with meaning • Contains serious errors in grammar, usage, or mechanics that frequently obscure meaning	A 2 response largely disregards the specific task directions and/or demonstrates serious weaknesses in analytical writing. A typical response in this category exhibits ONE OR MORE of the following characteristics: • Is unclear or seriously limited in addressing the specific task directions and/or in presenting or developing a position on the issue • Provides few, if any, relevant reasons or examples in support of its claims • Is poorly focused and/or poorly organized • Has serious problems in language and sentence structure that frequently interfere with meaning • Contains serious errors in grammar, usage, or mechanics that frequently obscure meaning
1	A 1 response demonstrates fundamental deficiencies in analytical writing. A typical response in this category exhibits ONE OR MORE of the following characteristics: • Provides little or no evidence of understanding the argument • Provides little evidence of the ability to develop an organized response (e.g., is disorganized and/or extremely brief) • Has severe problems in language and sentence structure that persistently interfere with meaning • Contains pervasive errors in grammar, usage, or mechanics that result in incoherence	A 1 response demonstrates fundamental deficiencies in analytical writing. A typical response in this category exhibits ONE OR MORE of the following characteristics: Provides little or no evidence of understanding the issue Provides little evidence of the ability to develop an organized response (e.g., is disorganized and/or extremely brief) Has severe problems in language and sentence structure that persistently interfere with meaning Contains pervasive errors in grammar, usage, or mechanics that result in incoherence
0	Off topic (i.e., provides no evidence of an attempt to respond to the assigned topic), is in a foreign language, merely copies the topic, consists of only keystroke characters, or is illegible or nonverbal.	Off topic (i.e., provides no evidence of an attempt to respond to the assigned topic), is in a foreign language, merely copies the topic, consists of only keystroke characters, or is illegible or nonverbal.
NS	Blank	Blank

Appendix B

Diagram of 12.1 Features



Appendix C
Glossary of e-rater Macrofeatures and Microfeatures

Feature	Name of microfeature	Brief description	Example
Grammar	Fragment	A sentence-like string of words that does not contain a tensed verb or that is lacking an independent clause	"And the school too."
Grammar	Run-on sentence	A sentence-like string of words that contains two or more clauses without a conjunction	"Students deserve more respect they are young adults."
Grammar	Garbled sentence	A sentence-like string of words that contains five or more errors, or that has an error-to-word ratio > 0.1, or that is unparsable by the Santa module, which organizes words	"And except unusual exception, most children can be ease with their parents not the their teachers."
Grammar	Subject–verb agreement	A singular noun with a plural verb or a plural noun with a singular verb	"A uniform represent the school."
Grammar	Ill-formed verb	A mismatch between the tense of a verb and the local syntactic environment; also, use of *of* for *have*, as in *could of*	"We need the freedom to chose what we want to wear."
Grammar	Pronoun error	An objective case pronoun where nominative pronoun is required, or vice versa	"Us students want to express ourselves."
Grammar	Possessive error	A plural noun where a possessive noun should be; usually the result of omitting an apostrophe	"They stayed at my parents house."
Grammar	Wrong or missing word	An ungrammatical sequence of words that is usually the result of a typographical error or of an omission of a word	"The went to their teacher with a complaint."
Grammar	Proofread this!	An error which is difficult to analyze; often the result of multiple, adjacent errors	"They had many wrong science knowledge."
Usage	Wrong article (Method 1)	A singular determiner with a plural noun or a plural determiner with a singular noun; use of *an* instead of *a*, or vice versa	"I wrote in these book. He ate a orange."
Usage	Articles (wrong, missing, extraneous)	Use of *a* when *the* is required, or vice versa	We had **the good time at the party. (Wrong article)
Usage	Articles (wrong, missing, extraneous)	An article where none should be used or a missing article where one is required	I think it is good for me to share **room with others. (Missing article) I think that mostly people succeed because of **the hard work. (Extraneous article)
Usage	Confused words	Confusion of homophones, words that sound alike or nearly alike	Those young soldiers had to **loose their innocence and grow up. (lose) **Its your chance to show them that you are an independent person. (It's) Parents should give **there children curfews. (their) I think that mostly people succeed because of **the hard work. (Extraneous article)
Usage	Wrong word form	A verb used in place of a noun	"The choose is not an easy one."
Usage	Faulty comparison	Use of *more* with a comparative adjective or *most* with a superlative adjective	"This is a more better solution."

Appendix C: Continued

Feature	Name of microfeature	Brief description	Example
Usage	Preposition error	Use of incorrect preposition, omitting a preposition, or using an extraneous one	Their knowledge **on physics were very important. (of) The teenager was driving **in a high speed when he approached the curve. (at) Thank you for your consideration **to this matter. (of, in)
Usage	Nonstandard verb or word form	Nonword: Various nonwords commonly used in oral language	Nonwords: gonna, kinda, dont, cant, gotta, wont, sorta, shoulda, woulda, oughtta, wanna, hafta
Usage	Double negation	Instances of *not* or its contracted form *n't* followed by negatives such as no, nowhere	"The counselor doesn't have no vacations."
Mechanics	Spelling	A group of letters not conforming to known orthographic pattern	—
Mechanics	Failure to capitalize proper noun	Compares words to lists of pronouns that should be capitalized (e.g., names of countries, capital cities, male and female proper nouns, and religious holidays)	—
Mechanics	Initial caps	Missing initial capital letter in a sentence	—
Mechanics	Missing question mark	An unpunctuated interrogative	—
Mechanics	Missing final punctuation	A sentence lacking a period	—
Mechanics	Missing comma or apostrophe	Detects missing commas or apostrophes	Apostrophe: arent, cant, couldnt, didnt, doesnt, dont, hadnt, hasnt, havent, im, isnt, ive, shouldnt, someones, somebodys, wasnt, werent, wont, wouldnt, youre, thats, theyre, theyve, theres, todays, whats, wifes, lifes, anybodys, anyones, everybodys, everyones, childrens
Mechanics	Hyphen error	Missing hyphen in number constructions, certain noun compounds, and modifying expressions preceding a noun	"He fell into a three foot hole. They slipped past the otherwise engaged sentinel."
Mechanics	Fused word	Fused: An error consisting of two words merged together	"It means alot to me." Fused: alot, dresscode, eachother, everytime, otherhand, highschool, notime, infact, inorder, phonecall, schoollife, somethings, no one
Mechanics	Compound word	Detects errors consisting of two words that should be one	—
Mechanics	Duplicate	Two adjacent identical words or two articles, pronouns, modals, etc.	"I want to to go … They tried to help us them."
Style	Repetition of words	Excessive repetition of words	—
Style	Inappropriate word or phrase	Inappropriate words, various expletives	—
Style	And, and, and	Too many sentences beginning with coordinate conjunction	—
Style	Too many short sentences	More than four short sentences, less than seven words	—
Style	Too many long sentences	More than four long sentences, more than 55 words	—
Style	Passive voice	By-passives: the number of times there occur sentences containing BE + past participle verb form, followed somewhere later in the sentence by the word *by*.	"The sandwich was eaten by the girl."

Appendix C: Continued

Feature	Name of microfeature	Brief description	Example
Organization	Number of discourse elements	Provides a measure of development, as a function of the number of discourse elements	—
Development	Content development	Provides a measure of average length of discourse elements	—
Prompt-specific vocabulary usage	Score-group of essays to which target essay is most closely related.	Compares* essay to essay-groups 6, 5, 4, etc., and assigns score closest relationship (max cosine). *Cosine of weighted frequency.	—
Prompt-specific vocabulary usage	Similarity of essay's vocabulary to vocabulary of essays with score	Compares* essay to essay-group score 6. *Cosine of weighted frequency vectors.	—
Lexical complexity	Sophistication of word choice	Calculates median average word frequency, based on Lexile corpus	—
Lexical complexity	Word length	The mean average number of characters within words	—
Positive features	Preposition usage	The mean probability of the writer's prepositions	—
Positive features	Good collocation density	The number of good collocations over the total number of words	—
Positive features	Sentence variety	The writer's ability to choose the correct phrasing and employ a variety of different grammatical structures as necessary	—
Differential word use (DWU)	N/A	DWU calculates the difference in the relative frequency of a word in high-scored versus low-scored essays. A positive value indicates the essay is using vocabulary more typical of high-scoring essays, and vice versa.	—

Appendix D

Flagging Criterion and Conditions

Flagging criterion	Flagging condition
Quadratic-weighted kappa between e-rater score and human score	Quadratic-weighted kappa less than 0.7 Correlation less than 0.7
Pearson correlation between e-rater score and human score	Standardized difference greater than 0.15 in absolute value
Standardized difference between e-rater score and human score	0.15 is absolute value
Notable reduction in quadratic-weighted kappa or correlation from human/human to e-rater/ human	Decline in quadratic-weighted kappa or correlation of greater than 0.10
Standardized difference between e-rater score and human score within a subgroup of concern	Standardized difference greater than 0.10 in absolute value

Note. All the threshold values are checked to 4 decimal values for flagging.

Appendix E

Human–Human Agreement for Argument and Issue Prompts at Phase I

Table E1 Phase I Agreement Among Human Raters on Argument Prompts

		Human 1 by Human 2								
		Human 1		Human 2		Statistic				
Prompt	*N*	*M*	*SD*	*M*	*SD*	Std diff	Wtd kappa	% agree	% adj agree	*R*
VC048186	4,890	3.29	0.78	3.28	0.79	−0.02	0.71	66	99	0.71
VC048263	5,948	3.31	0.78	3.32	0.79	0.01	**0.70**[a]	66	99	**0.70**[a]
VC048268	4,708	3.29	0.83	3.30	0.82	0.01	0.70	64	99	0.70
VC048328	6,018	3.28	0.81	3.27	0.80	−0.02	0.71	65	99	0.71
VC048389	4,041	3.22	0.81	3.23	0.80	0.01	**0.69**[a]	64	99	**0.69**[a]
VC048408	4,838	3.36	0.86	3.34	0.85	−0.02	0.72	63	99	0.72
VC069377	3,747	3.28	0.81	3.26	0.80	−0.02	0.71	65	99	0.71
VC069384	5,564	3.30	0.79	3.31	0.79	0.02	**0.69**[a]	65	99	**0.69**[a]
VC069394	5,131	3.38	0.82	3.36	0.80	−0.01	0.70	64	99	0.70
VC069396	4,869	3.22	0.86	3.21	0.85	−0.01	0.72	64	99	0.72
VC084832	4,246	3.31	0.85	3.34	0.85	0.03	0.72	64	99	0.72
VC084840	6,954	3.32	0.77	3.30	0.78	−0.02	**0.67**[a]	63	99	**0.67**[a]
VC084849	3,561	3.34	0.79	3.35	0.79	0.01	**0.68**[a]	64	99	**0.69**[a]
VC086531	895	3.06	0.81	3.03	0.82	−0.04	**0.67**[a]	61	98	**0.67**[a]
VC101052	5,179	3.20	0.82	3.23	0.83	0.03	0.72	65	99	0.72
VC101056	4,611	3.29	0.81	3.28	0.81	−0.02	**0.69**[a]	63	99	**0.69**[a]
VC101542	4,133	3.28	0.81	3.30	0.79	0.01	0.71	65	99	0.71
VC140314	418	3.11	0.79	3.12	0.81	0.01	0.75	70	99	0.75
VC249418	3,753	3.29	0.79	3.28	0.79	−0.01	**0.68**[a]	63	99	**0.68**[a]
VC251464	2,234	3.27	0.78	3.25	0.78	−0.02	**0.68**[a]	63	99	**0.68**[a]
VC251477	5,549	3.28	0.80	3.25	0.80	−0.03	**0.68**[a]	62	99	**0.68**[a]
VC251575	1,561	3.20	0.83	3.22	0.84	0.02	**0.67**[a]	60	98	**0.67**[a]
VC251577	1,824	3.42	0.82	3.42	0.82	0.00	0.71	66	99	0.71
VC390618	2,912	3.09	0.84	3.10	0.83	0.00	**0.70**[a]	63	98	**0.70**[a]
VC048246	3,982	3.29	0.86	3.30	0.83	0.00	0.73	65	99	0.73
VC048273	5,326	3.31	0.85	3.31	0.85	0.00	0.72	64	99	0.72
VC048352	2,751	3.16	0.78	3.16	0.78	0.01	**0.66**[a]	62	99	**0.66**[a]

Table E1: Continued

		Human 1 by Human 2								
		Human 1		Human 2		Statistic				
Prompt	*N*	*M*	*SD*	*M*	*SD*	Std diff	Wtd kappa	% agree	% adj agree	*R*
VC048390	2,463	3.12	0.85	3.13	0.83	0.01	0.73	65	99	0.73
VC048411	5,386	3.18	0.82	3.18	0.82	0.00	**0.69**[a]	63	98	**0.69**[a]
VC069378	3,391	3.32	0.78	3.31	0.79	0.00	**0.68**[a]	63	99	**0.68**[a]
VC069382	2,599	3.32	0.85	3.34	0.85	0.02	0.72	63	99	0.72
VC069400	4,464	3.02	0.77	3.00	0.78	−0.03	**0.66**[a]	64	98	**0.66**[a]
VC084829	7,802	3.09	0.80	3.09	0.80	0.00	**0.67**[a]	62	99	**0.67**[a]
VC084835	6,145	3.10	0.81	3.06	0.80	−0.05	**0.69**[a]	65	99	**0.69**[a]
VC084851	4,323	3.21	0.79	3.20	0.77	−0.02	**0.68**[a]	64	99	**0.68**[a]
VC084853	3,591	3.22	0.82	3.20	0.81	−0.02	0.72	66	99	0.72
VC086526	3,265	3.15	0.82	3.16	0.80	0.01	0.70	65	99	0.70
VC093524	2,772	3.22	0.86	3.20	0.87	−0.02	0.75	65	99	0.75
VC093532	1,388	3.16	0.88	3.15	0.87	−0.01	0.71	63	98	0.71
VC101021	4,177	3.25	0.82	3.24	0.81	−0.01	0.71	65	99	0.71
VC101037	2,869	3.12	0.83	3.13	0.83	0.01	0.70	64	99	0.70
VC101537	1,390	3.17	0.85	3.18	0.84	0.02	**0.68**[a]	61	98	**0.68**[a]
VC101541	5,879	3.11	0.76	3.10	0.76	−0.02	**0.68**[a]	66	99	**0.68**[a]
VC207455	2,360	3.30	0.83	3.29	0.84	−0.01	0.70	64	98	0.70
VC207640	2,920	3.33	0.87	3.31	0.86	−0.02	0.71	63	98	0.71
VC209497	2,656	3.30	0.85	3.28	0.85	−0.02	0.73	64	99	0.73
VC248469	1,601	3.04	0.83	3.07	0.82	0.04	0.74	67	99	0.74
VC250603	1,704	3.38	0.83	3.37	0.84	−0.01	**0.68**[a]	62	98	**0.68**[a]
VC251468	4,158	3.29	0.80	3.29	0.81	−0.01	0.71	66	99	0.71
VC251474	1,673	3.22	0.84	3.19	0.84	−0.03	**0.70**[a]	64	98	**0.70**[a]
VC251573	1,398	3.04	0.87	3.06	0.88	0.03	0.72	63	98	0.72
VC390606	1,512	3.19	0.83	3.21	0.85	0.02	0.71	63	99	0.71
VC462771	2,973	3.14	0.83	3.11	0.82	−0.03	0.71	65	99	0.71
VC101540	7,724	3.31	0.80	3.32	0.81	0.02	**0.70**[a]	65	99	**0.70**[a]
VC250595	4,366	3.31	0.83	3.34	0.82	0.03	0.72	64	99	0.72
VC101018	7,004	3.15	0.81	3.14	0.81	−0.01	**0.67**[a]	62	98	**0.67**[a]
VC251475	3,608	3.24	0.83	3.26	0.84	0.02	0.71	64	99	0.71
VC390614	10,820	3.26	0.79	3.27	0.78	0.01	**0.69**[a]	66	99	**0.69**[a]
VC101050	2,769	3.26	0.83	3.26	0.83	0.00	**0.69**[a]	63	98	**0.69**[a]
VC177590	4,506	3.29	0.79	3.28	0.79	−0.01	**0.64**[a]	62	98	**0.64**[a]
VC248460	3,911	3.30	0.81	3.31	0.83	0.01	**0.70**[a]	64	99	**0.70**[a]
VC248479	3,285	3.19	0.80	3.20	0.79	0.01	**0.67** [a]	63	98	**0.67** [a]
VC248488	2,185	3.19	0.82	3.18	0.83	0.00	**0.65** [a]	60	98	**0.65** [a]
VC250589	2,796	3.23	0.82	3.20	0.79	−0.03	**0.69** [a]	65	99	**0.70** [a]
VC251576	5,695	3.27	0.82	3.24	0.82	−0.04	**0.68** [a]	62	98	**0.68** [a]
VC390640	4,394	3.19	0.80	3.20	0.81	0.01	**0.63** [a]	61	97	**0.63** [a]
VC462770	3,891	3.25	0.82	3.25	0.82	0.00	**0.69** [a]	63	99	**0.69** [a]
VE096305	2,279	3.36	0.82	3.33	0.82	−0.03	**0.68** [a]	62	98	**0.68** [a]
VC069380	4,914	3.32	0.79	3.34	0.78	0.02	**0.66** [a]	63	99	**0.66** [a]
VC084843	4,052	3.25	0.78	3.25	0.77	0.01	**0.68** [a]	65	99	**0.68** [a]
VC084846	4,839	3.37	0.78	3.36	0.77	−0.01	**0.70** [a]	66	99	**0.70** [a]
VC101016	6,573	3.16	0.82	3.16	0.83	0.00	**0.70** [a]	63	99	**0.70** [a]
VC101539	5,453	3.34	0.81	3.35	0.80	0.01	**0.70** [a]	65	99	**0.70** [a]
VC140094	2,658	3.29	0.81	3.28	0.81	−0.02	0.70	64	99	0.70
VC209485	3,031	3.24	0.81	3.25	0.82	0.01	0.71	65	99	0.71
VC248473	3,730	3.29	0.82	3.27	0.83	−0.02	0.71	65	99	0.71
Average	3,934	3.24	0.82	3.24	0.82	0.00	**0.70** [a]	64	99	**0.70** [a]

Note. Std diff = standardized difference; wtd = weighted; adj = adjacent.

[a]Agreement indices that failed to meet the guideline thresholds (also indicated by boldface).

Table E2 Phase I Agreement Among Human Raters on Issue Prompts

		Human 1 by Human 2								
		Human 1		Human 2		Statistic				
Prompt	*N*	*M*	*SD*	*M*	*SD*	Std diff	Wtd kappa	% agree	% adj agree	*r*
VC047733	6,030	3.17	0.77	3.18	0.77	0.02	0.72	68	100	0.72
VC047741	5,212	3.16	0.76	3.15	0.76	−0.01	**0.69**[a]	67	99	**0.69**[a]
VC047750	2,693	3.27	0.73	3.30	0.75	0.04	0.70	68	100	0.70
VC047757	3,108	3.17	0.78	3.16	0.78	−0.01	0.73	68	100	0.73
VC047770	4,331	3.18	0.81	3.17	0.80	−0.01	0.74	67	100	0.74
VC047788	6,299	3.31	0.77	3.31	0.78	0.01	0.70	67	99	0.71
VC047792	4,143	3.26	0.79	3.26	0.79	0.00	0.74	68	100	0.74
VC047799	5,201	3.14	0.78	3.15	0.79	0.02	0.71	67	99	0.71
VC047822	6,826	3.26	0.80	3.25	0.81	−0.01	0.71	65	99	0.71
VC048013	3,077	3.16	0.83	3.17	0.80	0.01	0.75	68	99	0.75
VC048019	3,196	3.10	0.79	3.10	0.79	0.00	**0.70**[a]	65	99	**0.70**[a]
VC048027	5,942	3.31	0.77	3.30	0.77	−0.01	0.71	67	100	0.71
VC048031	4,405	3.33	0.79	3.32	0.79	−0.01	0.72	67	99	0.72
VC073155	6,117	3.24	0.78	3.21	0.77	−0.04	0.72	68	99	0.72
VC073160	4,729	3.22	0.78	3.21	0.78	−0.02	0.73	68	100	0.73
VC073163	3,329	3.13	0.82	3.13	0.80	−0.01	0.72	65	99	0.72
VC073164	3,105	3.24	0.76	3.25	0.76	0.02	0.71	69	99	0.71
VC073166	1,440	3.18	0.77	3.16	0.77	−0.04	**0.68**[a]	65	99	**0.68**[a]
VC073168	2,982	3.20	0.80	3.21	0.81	0.01	0.71	66	99	0.71
VC073173	4,437	3.19	0.81	3.19	0.79	0.01	0.71	65	99	0.71
VC073175	1,704	3.10	0.79	3.06	0.78	−0.05	0.72	67	99	0.72
VC073176	1,436	3.17	0.81	3.19	0.81	0.03	0.73	67	99	0.73
VC104275	2,618	3.15	0.80	3.15	0.78	0.00	0.72	68	99	0.72
VC104276	2,720	3.25	0.80	3.26	0.79	0.02	0.72	67	99	0.72
VC104278	4,095	3.32	0.80	3.28	0.80	−0.04	**0.70**[a]	65	99	**0.70**[a]
VC155074	993	3.19	0.80	3.15	0.80	−0.06	0.72	65	99	0.72
VC219551	2,044	3.30	0.76	3.30	0.78	0.00	0.74	71	99	0.74
VC515323	2,553	3.25	0.77	3.23	0.77	−0.03	0.73	69	100	0.73
VC787322	1,853	3.14	0.79	3.13	0.79	−0.01	0.74	68	100	0.74
VC929101	1,934	3.39	0.78	3.38	0.76	−0.02	0.71	68	99	0.71
VC929114	2,458	3.38	0.77	3.37	0.76	−0.01	0.71	67	99	0.71
VC929139	2,346	3.24	0.78	3.24	0.78	0.01	0.72	68	99	0.72
VC048048	3,536	3.28	0.77	3.29	0.77	0.01	0.73	70	100	0.73
VC048051	5,374	3.27	0.73	3.25	0.74	−0.02	**0.68**[a]	68	99	**0.68**[a]
VC048070	3,803	3.24	0.76	3.24	0.77	0.00	0.70	67	99	0.70
VC048077	5,094	3.12	0.78	3.12	0.78	0.00	0.72	67	100	0.72
VC048141	4,149	3.29	0.73	3.30	0.74	0.01	**0.69**[a]	68	100	**0.69**[a]
VC073157	4,605	3.30	0.77	3.29	0.77	−0.02	0.72	69	100	0.72
VC073158	2,961	3.25	0.78	3.22	0.79	−0.03	0.73	68	99	0.73
VC073169	7,645	3.12	0.80	3.10	0.80	−0.02	0.73	67	99	0.73
VC073172	2,499	3.19	0.77	3.19	0.76	0.01	0.73	69	100	0.73
VC104280	3,378	3.39	0.75	3.40	0.74	0.01	0.71	69	100	0.71
VC104281	4,859	3.42	0.76	3.41	0.76	−0.02	**0.70**[a]	67	99	**0.70**[a]
VC084819	6,588	3.34	0.76	3.32	0.75	−0.02	**0.67**[a]	65	99	**0.67**[a]
VC084820	4,555	3.23	0.77	3.25	0.77	0.02	**0.68**[a]	65	99	**0.68**[a]
VC155042	8,352	3.26	0.76	3.25	0.77	−0.01	**0.70**[a]	66	99	**0.70**[a]
VC219591	6,526	3.37	0.72	3.37	0.71	0.01	**0.69**[a]	70	100	**0.69**[a]
VC787354	6,048	3.29	0.77	3.29	0.76	−0.01	**0.70**[a]	67	99	**0.70**[a]
VC084798	6,213	3.13	0.75	3.13	0.77	0.00	0.71	68	100	0.71
VC084799	5,885	2.95	0.78	2.97	0.77	0.03	**0.70**[a]	66	99	**0.70**[a]
VC084804	5,371	3.21	0.74	3.19	0.73	−0.03	**0.66**[a]	66	99	**0.66**[a]
VC104286	7,865	3.18	0.77	3.19	0.78	0.01	0.70	66	99	0.70
VE096407	7,957	3.03	0.76	3.04	0.77	0.01	**0.70**[a]	66	99	**0.70**[a]
VC084555	4,175	3.06	0.74	3.08	0.77	0.02	**0.68**[a]	67	99	**0.68**[a]

Table E2: Continued

		Human 1 by Human 2								
		Human 1		Human 2		Statistic				
Prompt	*N*	*M*	*SD*	*M*	*SD*	Std diff	Wtd kappa	% agree	% adj agree	*r*
VC084754	6,319	3.02	0.73	2.99	0.72	−0.03	**0.67**[a]	67	99	**0.67**[a]
VC104284	6,034	3.06	0.69	3.05	0.68	−0.01	**0.63**[a]	68	99	**0.63**[a]
VC155075	2,204	3.07	0.74	3.09	0.74	0.03	**0.69**[a]	69	99	**0.69**[a]
VC155078	4,367	3.18	0.70	3.18	0.70	0.00	**0.66**[a]	69	99	**0.66**[a]
VC178591	2,389	3.16	0.72	3.15	0.72	−0.01	**0.69**[a]	69	100	**0.69**[a]
VC178595	4,819	3.26	0.70	3.26	0.70	−0.01	**0.65**[a]	68	99	**0.65**[a]
VC178602	4,294	3.11	0.77	3.11	0.75	0.00	**0.70**[a]	67	99	**0.70**[a]
VC515311	6,006	3.08	0.70	3.07	0.70	−0.01	**0.64**[a]	68	99	**0.64**[a]
VC787323	4,327	3.10	0.72	3.10	0.74	0.01	**0.66**[a]	66	99	**0.66**[a]
VC787333	1,866	3.12	0.76	3.16	0.74	0.05	**0.67**[a]	65	99	**0.67**[a]
VC084809	3,686	2.99	0.76	2.97	0.76	−0.01	**0.70**[a]	67	99	**0.70**[a]
VC104290	4,689	3.19	0.78	3.20	0.77	0.02	0.72	68	99	0.72
VC104293	2,971	3.27	0.78	3.27	0.78	0.00	0.72	68	99	0.72
VC104297	2,213	3.00	0.80	3.01	0.79	0.01	**0.68**[a]	64	99	**0.68**[a]
VC104300	4,413	3.24	0.76	3.25	0.79	0.02	0.72	68	99	0.72
VC104302	4,727	3.03	0.78	3.04	0.76	0.01	0.70	67	99	0.70
VC155043	2,531	3.27	0.80	3.26	0.77	−0.01	0.73	67	100	0.73
VC515320	1,957	3.15	0.79	3.14	0.77	−0.02	0.71	67	99	0.71
VC787346	1,691	3.25	0.79	3.28	0.78	0.03	0.72	68	99	0.72
VE096379	2,398	3.11	0.76	3.09	0.75	−0.02	0.72	69	100	0.72
VE096386	1,644	3.18	0.80	3.17	0.79	−0.01	0.74	69	100	0.74
VE096411	2,316	3.17	0.77	3.17	0.77	0.00	0.71	67	100	0.71
Average	4,061	3.20	0.77	3.20	0.77	0.00	0.70	67	99	0.70

Note. Std diff = standardized difference; wtd = weighted; adj = adjacent.

[a]Agreement indices that failed to meet the guideline thresholds (also indicated by boldface).

Appendix F

Advisory Flag Analyses for Issue and Argument Prompts

Table F1 Phase II Distribution of Advisory Flags in Argument Prompts

Prompt	No advisory	Nonfatal advisory flag	Fatal advisory flag	Flag sum	% flag
VC048186	965	35	0	35	4
VC048246	970	30	0	30	3
VC048263	971	30	1	31	3
VC048268	977	23	0	23	2
VC048273	972	31	2	33	3
VC048328	958	42	1	43	4
VC048352	967	34	2	36	4
VC048389	967	34	2	36	4
VC048390	966	33	1	34	3
VC048408	955	43	2	45	5
VC048411	966	33	1	34	3
VC069377	954	46	0	46	5
VC069378	978	20	4	24	2
VC069380	970	30	0	30	3
VC069382	537	12	2	14	3
VC069384	967	30	2	32	3
VC069394	977	22	2	24	2
VC069396	953	47	3	50	5
VC069400	980	20	2	22	2
VC084829	959	40	3	43	4

Table F1: Continued

Prompt	No advisory	Nonfatal advisory flag	Fatal advisory flag	Flag sum	% flag
VC084832	968	30	4	34	3
VC084835	957	42	2	44	4
VC084840	988	12	0	12	1
VC084843	969	32	0	32	3
VC084846	978	21	1	22	2
VC084849	972	28	0	28	3
VC084851	958	40	1	41	4
VC084853	961	40	0	40	4
VC086526	969	29	2	31	3
VC086531	982	16	1	17	2
VC093524	964	35	1	36	4
VC093532	970	25	3	28	3
VC101016	968	32	3	35	3
VC101018	972	26	0	26	3
VC101021	953	47	1	48	5
VC101037	964	31	4	35	4
VC101050	969	29	1	30	3
VC101052	970	28	5	33	3
VC101056	972	26	4	30	3
VC101537	775	23	0	23	3
VC101539	979	20	2	22	2
VC101540	968	29	2	31	3
VC101541	965	36	0	36	4
VC101542	972	25	4	29	3
VC140094	974	27	1	28	3
VC140314	971	26	2	28	3
VC177590	976	23	0	23	2
VC207455	970	30	0	30	3
VC207640	965	32	2	34	3
VC209485	972	27	1	28	3
VC209497	960	40	2	42	4
VC248460	965	35	0	35	4
VC248469	964	34	1	35	4
VC248473	965	32	3	35	4
VC248479	963	35	1	36	4
VC248488	959	41	1	42	4
VC249418	970	30	0	30	3
VC250589	977	22	2	24	2
VC250595	960	39	4	43	4
VC250603	917	31	0	31	3
VC251464	980	21	0	21	2
VC251468	974	25	0	25	3
VC251474	938	33	0	33	3
VC251475	974	26	0	26	3
VC251477	977	21	2	23	2
VC251573	973	26	3	29	3
VC251575	979	20	1	21	2
VC251576	980	21	0	21	2
VC251577	964	36	0	36	4
VC390606	968	27	5	32	3
VC390614	965	34	3	37	4
VC390618	965	35	0	35	4
VC390640	966	35	0	35	3
VC462770	975	24	2	26	3
VC462771	976	24	4	28	3
VE096305	963	31	6	37	4

Table F2 Phase II Distribution of Advisory Flags in Issue Prompts

Prompt	No advisory	Nonfatal advisory flag	Fatal advisory flag	Flag sum	% flag
VC047733	995	4	0	4	0
VC047741	998	1	2	3	0
VC047750	997	3	0	3	0
VC047757	997	0	3	3	0
VC047770	989	12	0	12	1
VC047788	997	0	2	2	0
VC047792	986	13	3	16	2
VC047799	991	9	0	9	1
VC047822	1000	0	3	3	0
VC048013	998	1	2	3	0
VC048019	999	1	1	2	0
VC048027	999	0	1	1	0
VC048031	999	1	0	1	0
VC048048	989	10	1	11	1
VC048051	983	13	3	16	2
VC048070	956	20	0	20	2
VC048077	989	10	0	10	1
VC048141	991	6	3	9	1
VC073155	999	0	1	1	0
VC073157	982	15	2	17	2
VC073158	989	11	0	11	1
VC073160	996	3	1	4	0
VC073163	996	3	0	3	0
VC073164	995	5	7	12	1
VC073166	998	3	0	3	0
VC073168	996	3	1	4	0
VC073169	976	11	1	12	1
VC073172	1001	0	0	0	0
VC073173	995	0	6	6	1
VC073175	997	2	1	3	0
VC073176	990	9	1	10	1
VC084555	997	3	1	4	0
VC084754	978	22	0	22	2
VC084798	994	5	0	5	1
VC084799	993	7	0	7	1
VC084804	988	11	3	14	1
VC084809	988	11	1	12	1
VC084819	994	6	2	8	1
VC084820	988	9	4	13	1
VC104275	992	6	1	7	1
VC104276	998	0	2	2	0
VC104278	998	1	1	2	0
VC104280	991	7	3	10	1
VC104281	990	10	1	11	1
VC104284	969	6	1	7	1
VC104286	981	15	4	19	2
VC104290	996	2	3	5	0
VC104293	986	12	4	16	2
VC104297	994	5	1	6	1
VC104300	985	13	3	16	2
VC104302	991	9	0	9	1
VC155042	988	10	1	11	1
VC155043	974	26	1	27	3
VC155074	992	7	1	8	1
VC155075	986	14	0	14	1
VC155078	944	12	0	12	1

Table F2: Continued

Prompt	No advisory	Nonfatal advisory flag	Fatal advisory flag	Flag sum	% flag
VC178591	978	18	4	22	2
VC178595	984	15	2	17	2
VC178602	995	5	0	5	1
VC219551	988	11	4	15	1
VC219591	990	10	0	10	1
VC515311	991	10	0	10	1
VC515320	991	8	2	10	1
VC515323	995	5	0	5	1
VC787322	992	5	3	8	1
VC787323	987	14	2	16	2
VC787333	996	5	0	5	0
VC787346	995	1	5	6	1
VC787354	998	1	2	3	0
VC929101	988	12	1	13	1
VC929114	994	4	2	6	1
VC929139	996	2	2	4	0
VE096379	978	21	2	23	2
VE096386	989	7	6	13	1
VE096407	949	15	2	17	2
VE096411	985	13	4	17	2

Appendix G

Human and e-rater Agreement for Argument and Issue Prompts With (12 Model) and Without (10 Model) CVA at Phase II

Table G1 Phase II Agreement of Human and e-rater Scores on Argument Prompts: G-10 Model

		Human 1 by e-rater (rounded to integers)									Human 1 by e-rater (unrounded)			
		Human 1		e-rater		Statistic					e-rater		Statistic	
Prompt	*N*	*M*	*SD*	*M*	*SD*	Std diff	Wtd kappa	% agree	% adj agree	*r*	*M*	*SD*	Std diff	*r*
VC048186	4,890	3.29	0.78	3.10	0.86	**−0.24**[a]	0.72	63	99	0.74	3.09	0.82	**−0.26**[a]	0.79
VC048263	5,948	3.31	0.78	3.45	0.83	**0.18**[a]	0.71	65	99	0.72	3.45	0.78	**0.18**[a]	0.77
VC048268	4,708	3.29	0.83	3.37	0.84	0.10	0.71	64	98	0.72	3.37	0.80	0.11	0.75
VC048328	6,018	3.28	0.81	3.37	0.82	0.11	0.71	65	99	0.71	3.36	0.78	0.11	0.76
VC048389	4,041	3.22	0.81	3.34	0.84	0.15	0.71	64	99	0.72	3.34	0.80	0.15	0.76
VC048408	4,838	3.36	0.86	3.62	0.84	**0.30**[a]	**0.69**[a]	60	98	0.72	3.61	0.80	**0.30**[a]	0.76
VC069377	3,747	3.28	0.81	3.26	0.83	−0.02	0.74	67	99	0.74	3.26	0.78	−0.03	0.78
VC069384	5,564	3.30	0.79	3.40	0.84	0.12	0.73	66	99	0.73	3.39	0.81	0.12	0.77
VC069394	5,131	3.38	0.82	3.51	0.84	**0.16**[a]	0.72	64	99	0.73	3.50	0.79	**0.16**[a]	0.77
VC069396	4,869	3.22	0.86	3.27	0.86	0.06	0.72	63	98	0.72	3.27	0.81	0.06	0.76
VC084832	4,246	3.31	0.85	3.39	0.87	0.09	0.73	64	99	0.74	3.39	0.83	0.10	0.78
VC084840	6,954	3.32	0.77	3.30	0.84	−0.02	0.71	64	99	0.71	3.31	0.80	−0.01	0.76
VC084849	3,561	3.34	0.79	3.39	0.82	0.06	0.71	65	99	0.72	3.37	0.78	0.05	0.76
VC086531	895	3.06	0.81	2.97	0.89	−0.11	**0.65**[a]	56	98	**0.66**[a]	2.96	0.84	−0.12	0.71
VC101052	5,179	3.20	0.82	3.26	0.86	0.07	0.71	63	99	0.71	3.25	0.83	0.06	0.75
VC101056	4,611	3.29	0.81	3.36	0.84	0.09	0.71	64	99	0.71	3.35	0.80	0.08	0.75
VC101542	4,133	3.28	0.81	3.25	0.86	−0.04	0.74	66	99	0.74	3.25	0.82	−0.05	0.78
VC140314	4,18	3.11	0.79	2.76	0.85	**−0.43**[a]	**0.67**[a]	54	99	0.73	2.76	0.82	**−0.44**[a]	0.78
VC249418	3,753	3.29	0.79	3.24	0.84	−0.06	0.71	64	99	0.72	3.23	0.80	−0.07	0.76
VC251464	2,234	3.27	0.78	3.34	0.84	0.08	0.70	63	99	0.71	3.34	0.79	0.09	0.75
VC251477	5,549	3.28	0.80	3.34	0.85	0.07	**0.70**[a]	62	99	**0.70**[a]	3.33	0.81	0.07	0.74
VC251575	1,561	3.20	0.83	3.21	0.83	0.01	**0.69**[a]	61	99	**0.69**[a]	3.21	0.79	0.01	0.74
VC251577	1,824	3.42	0.82	3.48	0.81	0.07	0.71	65	99	0.71	3.47	0.78	0.06	0.75

Table G1: Continued

		Human 1 by e-rater (rounded to integers)									Human 1 by e-rater (unrounded)			
		Human 1		e-rater		Statistic					e-rater		Statistic	
Prompt	*N*	*M*	*SD*	*M*	*SD*	Std diff	Wtd kappa	% agree	% adj agree	*r*	*M*	*SD*	Std diff	*r*
VC390618	2,912	3.09	0.84	3.04	0.86	−0.06	**0.70**[a]	60	99	**0.70**[a]	3.05	0.83	−0.06	0.74
VC048246	3,982	3.29	0.86	3.32	0.86	0.03	0.75	66	99	0.75	3.31	0.83	0.02	0.79
VC048273	5,326	3.31	0.85	3.43	0.82	0.13	0.71	64	98	0.71	3.42	0.78	0.14	0.76
VC048352	2,751	3.16	0.78	3.13	0.85	−0.03	0.71	63	99	0.71	3.13	0.80	−0.03	0.76
VC048390	2,463	3.12	0.85	3.15	0.86	0.04	0.73	63	99	0.73	3.14	0.83	0.03	0.77
VC048411	5,386	3.18	0.82	3.09	0.85	−0.11	0.71	62	99	0.71	3.09	0.81	−0.11	0.76
VC069378	3,391	3.32	0.78	3.35	0.82	0.04	0.71	65	99	0.71	3.34	0.78	0.03	0.75
VC069382	2,599	3.32	0.85	3.43	0.85	0.13	**0.69**[a]	60	98	0.70	3.43	0.81	0.13	0.74
VC069400	4,464	3.02	0.77	2.71	0.85	**−0.38**[a]	**0.63**[a]	54	98	**0.68**[a]	2.70	0.81	**−0.40**[a]	0.73
VC084829	7,802	3.09	0.80	3.03	0.85	−0.07	**0.68**[a]	60	99	**0.69**[a]	3.02	0.82	−0.08	0.73
VC084835	6,145	3.10	0.81	2.85	0.83	**−0.31**[a]	**0.62**[a]	53	98	**0.65**[a]	2.84	0.79	**−0.33**[a]	0.70
VC084851	4,323	3.21	0.79	3.17	0.82	−0.06	**0.69**[a]	63	99	**0.69**[a]	3.16	0.77	−0.07	0.73
VC084853	3,591	3.22	0.82	3.15	0.85	−0.08	0.71	63	99	0.72	3.14	0.81	−0.09	0.76
VC086526	3,265	3.15	0.82	3.05	0.84	−0.12	**0.68**[a]	59	99	**0.69**[a]	3.05	0.80	−0.13	0.74
VC093524	2,772	3.22	0.86	3.30	0.85	0.09	0.70	61	98	0.71	3.30	0.81	0.10	0.76
VC093532	1,388	3.16	0.88	3.17	0.87	0.01	0.74	63	99	0.74	3.16	0.84	0.00	0.78
VC101021	4,177	3.25	0.82	3.23	0.83	−0.03	0.71	64	99	0.71	3.22	0.79	−0.03	0.75
VC101037	2,869	3.12	0.83	3.03	0.85	−0.11	**0.70**[a]	60	99	0.70	3.02	0.81	−0.12	0.74
VC101537	1,390	3.17	0.85	3.11	0.83	−0.07	**0.70**[a]	61	99	**0.70**[a]	3.11	0.79	−0.07	0.74
VC101541	5,879	3.11	0.76	2.68	0.81	**−0.54**[a]	**0.59**[a]	48	98	**0.68**[a]	2.68	0.78	**−0.56**[a]	0.73
VC207455	2,360	3.30	0.83	3.27	0.83	−0.04	0.71	63	99	0.71	3.26	0.79	−0.05	0.75
VC207640	2,920	3.33	0.87	3.55	0.83	**0.27**[a]	0.70	62	98	0.73	3.55	0.79	**0.27**[a]	0.76
VC209497	2,656	3.30	0.85	3.38	0.83	0.09	0.72	64	99	0.73	3.37	0.80	0.09	0.77
VC248469	1,601	3.04	0.83	3.11	0.88	0.08	0.74	65	99	0.74	3.10	0.85	0.07	0.77
VC250603	1,704	3.38	0.83	3.39	0.86	0.02	0.71	63	99	0.71	3.40	0.83	0.02	0.76
VC251468	4,158	3.29	0.80	3.40	0.84	0.13	0.70	64	98	0.71	3.40	0.79	0.14	0.75
VC251474	1,673	3.22	0.84	3.31	0.87	0.12	0.71	63	98	0.71	3.30	0.83	0.10	0.76
VC251573	1,398	3.04	0.87	3.06	0.87	0.03	0.70	61	98	0.70	3.06	0.83	0.03	0.74
VC390606	1,512	3.19	0.83	3.31	0.86	0.13	0.72	63	99	0.73	3.30	0.83	0.13	0.76
VC462771	2,973	3.14	0.83	3.06	0.86	−0.08	0.71	63	99	0.71	3.05	0.82	−0.10	0.75
VC101540	7,724	3.31	0.80	3.33	0.84	0.03	0.72	65	99	0.72	3.33	0.79	0.03	0.76
VC250595	4,366	3.31	0.83	3.33	0.86	0.02	0.73	65	99	0.74	3.33	0.82	0.02	0.78
VC101018	7,004	3.15	0.81	2.92	0.84	**−0.28**[a]	**0.67**[a]	57	98	**0.69**[a]	2.92	0.81	**−0.29**[a]	0.74
VC251475	3,608	3.24	0.83	3.24	0.89	0.00	0.74	65	99	0.74	3.23	0.85	−0.01	0.78
VC390614	10,820	3.26	0.79	3.30	0.82	0.05	0.72	66	99	0.72	3.30	0.77	0.05	0.76
VC101050	2,769	3.26	0.83	3.30	0.86	0.05	**0.70**[a]	62	98	**0.70**[a]	3.29	0.82	0.04	0.74
VC177590	4,506	3.29	0.79	3.43	0.83	**0.16**[a]	**0.68**[a]	63	98	**0.69**[a]	3.42	0.78	**0.16**[a]	0.72
VC248460	3,911	3.30	0.81	3.23	0.86	−0.08	0.72	64	99	0.73	3.24	0.83	−0.08	0.77
VC248479	3,285	3.19	0.80	3.14	0.85	−0.06	0.71	63	99	0.72	3.13	0.81	−0.07	0.76
VC248488	2,185	3.19	0.82	3.13	0.82	−0.06	**0.68**[a]	61	99	**0.68**[a]	3.13	0.79	−0.07	0.73
VC250589	2,796	3.23	0.82	3.06	0.84	**−0.20**[a]	**0.69**[a]	60	99	0.71	3.06	0.81	**−0.21**[a]	0.75
VC251576	5,695	3.27	0.82	3.26	0.82	−0.02	**0.69**[a]	63	99	**0.69**[a]	3.25	0.78	−0.03	0.74
VC390640	4,394	3.19	0.80	3.17	0.83	−0.02	**0.65**[a]	60	98	**0.65**[a]	3.16	0.79	−0.04	**0.68**[a]
VC462770	3,891	3.25	0.82	3.19	0.83	−0.08	**0.70**[a]	61	99	**0.70**[a]	3.18	0.79	−0.09	0.74
VE096305	2,279	3.36	0.82	3.37	0.83	0.01	**0.69**[a]	62	99	**0.69**[a]	3.38	0.79	0.02	0.74
VC069380	4,914	3.32	0.79	3.50	0.80	**0.23**[a]	**0.67**[a]	63	98	**0.69**[a]	3.50	0.75	**0.24**[a]	0.74
VC084843	4,052	3.25	0.78	3.28	0.85	0.04	**0.69**[a]	63	99	**0.70**[a]	3.28	0.82	0.04	0.74
VC084846	4,839	3.37	0.78	3.42	0.84	0.06	0.72	66	99	0.72	3.41	0.80	0.05	0.77
VC101016	6,573	3.16	0.82	2.93	0.85	**−0.28**[a]	**0.68**[a]	58	99	0.71	2.93	0.81	**−0.28**[a]	0.75
VC101539	5,453	3.34	0.81	3.47	0.82	**0.16**[a]	0.71	65	99	0.72	3.46	0.77	**0.16**[a]	0.76
VC140094	2,658	3.29	0.81	3.24	0.84	−0.06	0.72	65	99	0.73	3.24	0.80	−0.07	0.77
VC209485	3,031	3.24	0.81	3.21	0.85	−0.04	0.72	65	99	0.72	3.21	0.82	−0.03	0.76
VC248473	3,730	3.29	0.82	3.27	0.87	−0.03	0.74	65	99	0.74	3.26	0.82	−0.04	0.78

Note. Std diff = standardized difference; wtd = weighted; adj = adjacent.

[a]Agreement indices that failed to meet the guideline thresholds (also indicated by boldface).

Table G2 Phase II Agreement of Human and e-rater Scores on Argument Prompts: G-12 Model

		Human 1 by e-rater (rounded to integers)									Human 1 by e-rater (unrounded)			
		Human 1		e-rater		Statistic					e-rater		Statistic	
Prompt	*N*	*M*	*SD*	*M*	*SD*	Std diff	Wtd kappa	% agree	% adj agree	*r*	*M*	*SD*	Std diff	*r*
VC048186	4,890	3.29	0.78	3.09	0.86	**−0.25**[a]	0.72	63	99	0.74	3.08	0.82	**−0.26**[a]	0.79
VC048263	5,948	3.31	0.78	3.45	0.83	**0.17**[a]	0.71	65	99	0.73	3.44	0.78	**0.17**[a]	0.77
VC048268	4,708	3.29	0.83	3.39	0.84	0.13	0.71	64	99	0.72	3.39	0.80	0.13	0.76
VC048328	6,018	3.28	0.81	3.37	0.82	0.11	0.71	65	99	0.71	3.37	0.78	0.11	0.76
VC048389	4,041	3.22	0.81	3.37	0.84	**0.19**[a]	0.70	63	99	0.72	3.37	0.81	**0.19**[a]	0.76
VC048408	4,838	3.36	0.86	3.60	0.85	**0.27**[a]	0.70	61	98	0.73	3.59	0.81	**0.28**[a]	0.76
VC069377	3,747	3.28	0.81	3.17	0.82	−0.14	0.72	65	99	0.73	3.17	0.77	−0.14	0.78
VC069384	5,564	3.30	0.79	3.38	0.85	0.10	0.73	66	99	0.73	3.38	0.81	0.10	0.77
VC069394	5,131	3.38	0.82	3.51	0.84	**0.17**[a]	0.72	64	99	0.73	3.51	0.79	**0.16**[a]	0.78
VC069396	4,869	3.22	0.86	3.23	0.85	0.01	0.72	64	99	0.72	3.23	0.81	0.02	0.76
VC084832	4,246	3.31	0.85	3.36	0.87	0.06	0.74	65	99	0.74	3.36	0.83	0.06	0.78
VC084840	6,954	3.32	0.77	3.31	0.85	−0.02	0.71	64	99	0.71	3.31	0.80	−0.01	0.76
VC084849	3,561	3.34	0.79	3.38	0.83	0.05	0.71	65	99	0.72	3.37	0.79	0.04	0.76
VC086531	895	3.06	0.81	2.96	0.88	−0.12	**0.64**[a]	55	98	**0.65**[a]	2.95	0.84	−0.14	0.71
VC101052	5,179	3.20	0.82	3.27	0.86	0.08	0.71	63	99	0.71	3.27	0.82	0.08	0.75
VC101056	4,611	3.29	0.81	3.34	0.84	0.06	0.71	64	99	0.71	3.33	0.80	0.04	0.75
VC101542	4,133	3.28	0.81	3.24	0.86	−0.05	0.74	66	99	0.74	3.24	0.82	−0.05	0.78
VC140314	418	3.11	0.79	2.78	0.86	**−0.40**[a]	**0.68**[a]	56	99	0.74	2.79	0.83	**−0.40**[a]	0.78
VC249418	3,753	3.29	0.79	3.25	0.84	−0.04	0.72	64	99	0.72	3.25	0.80	−0.05	0.76
VC251464	2,234	3.27	0.78	3.37	0.84	0.13	**0.70**[a]	62	99	0.70	3.38	0.79	0.14	0.75
VC251477	5,549	3.28	0.80	3.36	0.86	0.10	**0.70**[a]	62	99	0.70	3.36	0.82	0.10	0.74
VC251575	1,561	3.20	0.83	3.21	0.83	0.00	**0.70**[a]	61	99	**0.70**[a]	3.20	0.79	−0.01	0.74
VC251577	1,824	3.42	0.82	3.46	0.81	0.05	0.71	65	99	0.71	3.45	0.78	0.04	0.75
VC390618	2,912	3.09	0.84	3.09	0.87	−0.01	0.71	61	99	0.71	3.08	0.83	−0.01	0.74
VC048246	3,982	3.29	0.86	3.33	0.86	0.04	0.75	66	99	0.75	3.32	0.83	0.03	0.79
VC048273	5,326	3.31	0.85	3.40	0.82	0.11	0.71	65	98	0.72	3.40	0.79	0.11	0.76
VC048352	2,751	3.16	0.78	3.15	0.84	0.00	0.71	64	99	0.71	3.16	0.80	0.01	0.76
VC048390	2,463	3.12	0.85	3.15	0.86	0.04	0.73	63	99	0.73	3.15	0.83	0.03	0.77
VC048411	5,386	3.18	0.82	3.10	0.85	−0.10	0.71	63	99	0.71	3.09	0.81	−0.11	0.76
VC069378	3,391	3.32	0.78	3.34	0.82	0.03	0.71	65	99	0.71	3.33	0.78	0.02	0.75
VC069382	2,599	3.32	0.85	3.41	0.85	0.10	**0.70**[a]	60	99	0.70	3.41	0.81	0.11	0.75
VC069400	4,464	3.02	0.77	2.73	0.85	**−0.36**[a]	**0.64**[a]	55	98	**0.68**[a]	2.72	0.81	**−0.38**[a]	0.73
VC084829	7,802	3.09	0.80	3.04	0.85	−0.07	**0.69**[a]	61	99	**0.69**[a]	3.03	0.81	−0.07	0.73
VC084835	6,145	3.10	0.81	2.88	0.83	**−0.27**[a]	**0.64**[a]	55	98	**0.66**[a]	2.87	0.79	**−0.29**[a]	0.71
VC084851	4,323	3.21	0.79	3.19	0.82	−0.03	**0.69**[a]	63	99	**0.69**[a]	3.19	0.77	−0.04	0.73
VC084853	3,591	3.22	0.82	3.14	0.85	−0.09	0.71	63	99	0.72	3.14	0.81	−0.10	0.76
VC086526	3,265	3.15	0.82	3.04	0.85	−0.13	**0.68**[a]	58	99	**0.69**[a]	3.04	0.80	−0.14	0.74
VC093524	2,772	3.22	0.86	3.31	0.85	0.10	0.71	62	98	0.71	3.31	0.82	0.11	0.76
VC093532	1,388	3.16	0.88	3.15	0.87	−0.02	0.74	63	99	0.74	3.14	0.84	−0.02	0.78
VC101021	4,177	3.25	0.82	3.21	0.83	−0.05	0.72	65	99	0.72	3.21	0.79	−0.05	0.76
VC101037	2,869	3.12	0.83	3.02	0.85	−0.12	**0.70**[a]	60	99	0.70	3.01	0.81	−0.13	0.74
VC101537	1,390	3.17	0.85	3.09	0.82	−0.09	**0.70**[a]	61	99	0.70	3.08	0.79	−0.10	0.74
VC101541	5,879	3.11	0.76	2.70	0.81	**−0.51**[a]	**0.60**[a]	50	98	**0.68**[a]	2.70	0.77	**−0.53**[a]	0.73
VC207455	2,360	3.30	0.83	3.25	0.83	−0.06	0.71	63	99	0.71	3.24	0.79	−0.08	0.75
VC207640	2,920	3.33	0.87	3.48	0.82	**0.18**[a]	0.72	64	98	0.73	3.48	0.78	**0.18**[a]	0.76
VC209497	2,656	3.30	0.85	3.39	0.84	0.10	0.72	64	99	0.73	3.39	0.81	0.10	0.77
VC248469	1,601	3.04	0.83	3.10	0.88	0.07	0.74	64	99	0.74	3.09	0.85	0.06	0.77
VC250603	1,704	3.38	0.83	3.38	0.86	0.00	0.72	63	99	0.72	3.39	0.83	0.01	0.76
VC251468	4,158	3.29	0.80	3.43	0.85	**0.17**[a]	0.70	63	98	0.71	3.43	0.80	**0.18**[a]	0.75
VC251474	1,673	3.22	0.84	3.32	0.87	0.12	0.71	62	98	0.71	3.32	0.84	0.12	0.76
VC251573	1,398	3.04	0.87	3.05	0.87	0.01	0.70	61	98	0.70	3.04	0.83	0.01	0.74
VC390606	1,512	3.19	0.83	3.29	0.87	0.11	0.72	63	99	0.73	3.29	0.83	0.12	0.77
VC462771	2,973	3.14	0.83	3.07	0.87	−0.08	0.72	63	99	0.72	3.06	0.83	−0.09	0.75

Table G2: Continued

		Human 1 by e-rater (rounded to integers)									Human 1 by e-rater (unrounded)			
		Human 1		e-rater		Statistic					e-rater		Statistic	
Prompt	*N*	*M*	*SD*	*M*	*SD*	Std diff	Wtd kappa	% agree	% adj agree	*r*	*M*	*SD*	Std diff	*r*
VC101540	7,724	3.31	0.80	3.33	0.84	0.03	0.72	65	99	0.72	3.33	0.79	0.03	0.76
VC250595	4,366	3.31	0.83	3.33	0.86	0.02	0.73	65	99	0.73	3.32	0.82	0.01	0.78
VC101018	7,004	3.15	0.81	2.91	0.84	**−0.29**[a]	**0.66**[a]	57	98	**0.69**[a]	2.91	0.81	**−0.30**[a]	0.74
VC251475	3,608	3.24	0.83	3.26	0.90	0.02	0.74	65	99	0.75	3.26	0.86	0.02	0.78
VC390614	10,820	3.26	0.79	3.30	0.82	0.05	0.72	66	99	0.72	3.30	0.77	0.05	0.76
VC101050	2,769	3.26	0.83	3.30	0.86	0.05	**0.70**[a]	62	98	**0.70**[a]	3.30	0.82	0.05	0.74
VC177590	4,506	3.29	0.79	3.46	0.84	**0.20**[a]	**0.67**[a]	61	98	**0.69**[a]	3.45	0.79	**0.20**[a]	0.73
VC248460	3,911	3.30	0.81	3.23	0.86	−0.09	0.72	64	99	0.72	3.23	0.83	−0.09	0.76
VC248479	3,285	3.19	0.80	3.15	0.84	−0.05	0.71	64	99	0.71	3.15	0.81	−0.05	0.76
VC248488	2,185	3.19	0.82	3.15	0.83	−0.05	**0.69**[a]	62	99	**0.69**[a]	3.15	0.79	−0.05	0.73
VC250589	2,796	3.23	0.82	3.07	0.84	**−0.19**[a]	0.70	61	99	0.71	3.07	0.81	**−0.19**[a]	0.75
VC251576	5,695	3.27	0.82	3.24	0.82	−0.04	**0.69**[a]	62	99	**0.69**[a]	3.24	0.78	−0.05	0.74
VC390640	4,394	3.19	0.80	3.20	0.83	0.01	**0.65**[a]	60	98	**0.65**[a]	3.19	0.79	0.00	**0.69**[a]
VC462770	3,891	3.25	0.82	3.19	0.84	−0.08	**0.70**[a]	61	99	0.70	3.18	0.80	−0.08	0.74
VE096305	2,279	3.36	0.82	3.35	0.84	−0.01	**0.69**[a]	62	99	**0.69**[a]	3.36	0.79	−0.01	0.74
VC069380	4,914	3.32	0.79	3.50	0.79	**0.22**[a]	**0.68**[a]	63	98	**0.69**[a]	3.49	0.75	**0.22**[a]	0.74
VC084843	4,052	3.25	0.78	3.30	0.86	0.07	**0.69**[a]	62	99	**0.70**[a]	3.29	0.83	0.06	0.74
VC084846	4,839	3.37	0.78	3.45	0.84	0.10	0.72	65	99	0.72	3.45	0.81	0.10	0.77
VC101016	6,573	3.16	0.82	2.92	0.85	**−0.29**[a]	**0.67**[a]	58	98	0.70	2.92	0.81	**−0.29**[a]	0.75
VC101539	5,453	3.34	0.81	3.49	0.82	**0.18**[a]	0.71	65	99	0.72	3.49	0.77	**0.19**[a]	0.76
VC140094	2,658	3.29	0.81	3.27	0.84	−0.03	0.72	65	99	0.72	3.26	0.80	−0.04	0.76
VC209485	3,031	3.24	0.81	3.20	0.85	−0.05	0.72	64	99	0.72	3.20	0.82	−0.06	0.76
VC248473	3,730	3.29	0.82	3.23	0.86	−0.08	0.74	65	99	0.74	3.22	0.82	−0.08	0.78

Note. Std diff = standardized difference; wtd = weighted; adj = adjacent.

[a]Agreement indices that failed to meet the guideline thresholds (also indicated by boldface).

Table G3 Phase II Agreement of Human and e-rater Scores on Argument Prompts: PS-10 Model

		Human 1 by e-rater (rounded to integers)									Human 1 by e-rater (unrounded)			
		Human 1		e-rater		Statistic					e-rater		Statistic	
Prompt	*N*	*M*	*SD*	*M*	*SD*	Std diff	Wtd kappa	% agree	% adj agree	*r*	*M*	*SD*	Std diff	*r*
VC048186	4,890	3.29	0.78	3.28	0.83	−0.02	0.75	68	100	0.75	3.27	0.80	−0.03	0.79
VC048263	5,948	3.31	0.78	3.34	0.83	0.04	0.72	66	99	0.72	3.33	0.78	0.03	0.77
VC048268	4,708	3.29	0.83	3.26	0.90	−0.03	0.72	62	99	0.72	3.26	0.87	−0.04	0.76
VC048328	6,018	3.28	0.81	3.23	0.86	−0.06	0.72	63	99	0.72	3.23	0.82	−0.06	0.76
VC048389	4,041	3.22	0.81	3.23	0.87	0.00	0.72	64	99	0.73	3.22	0.82	0.00	0.76
VC048408	4,838	3.36	0.86	3.33	0.92	−0.03	0.72	61	98	0.72	3.33	0.90	−0.04	0.76
VC069377	3,747	3.28	0.81	3.27	0.85	−0.01	0.74	66	99	0.74	3.27	0.82	−0.01	0.78
VC069384	5,564	3.30	0.79	3.31	0.84	0.01	0.73	67	99	0.73	3.31	0.80	0.01	0.77
VC069394	5,131	3.38	0.82	3.38	0.86	0.01	0.73	64	99	0.73	3.38	0.82	0.00	0.77
VC069396	4,869	3.22	0.86	3.20	0.89	−0.02	0.73	62	99	0.73	3.20	0.85	−0.02	0.76
VC084832	4,246	3.31	0.85	3.31	0.89	0.00	0.74	64	99	0.74	3.30	0.86	−0.01	0.78
VC084840	6,954	3.32	0.77	3.32	0.82	0.00	0.70	65	99	0.71	3.32	0.77	0.01	0.76
VC084849	3,561	3.34	0.79	3.34	0.85	0.00	0.71	64	99	0.71	3.33	0.81	−0.01	0.76
VC086531	895	3.06	0.81	3.03	0.86	−0.04	**0.66**[a]	58	98	**0.66**[a]	3.03	0.82	−0.04	0.71
VC101052	5,179	3.20	0.82	3.20	0.86	−0.01	0.71	62	99	0.71	3.19	0.82	−0.02	0.76

Table G3: Continued

		Human 1 by e-rater (rounded to integers)									Human 1 by e-rater (unrounded)			
		Human 1		e-rater		Statistic					e-rater		Statistic	
Prompt	*N*	*M*	*SD*	*M*	*SD*	Std diff	Wtd kappa	% agree	% adj agree	*r*	*M*	*SD*	Std diff	*r*
VC101056	4,611	3.29	0.81	3.30	0.83	0.01	0.71	64	99	0.71	3.29	0.79	0.00	0.75
VC101542	4,133	3.28	0.81	3.30	0.84	0.02	0.74	67	99	0.75	3.30	0.81	0.02	0.79
VC140314	418	3.11	0.79	3.12	0.84	0.01	0.73	66	99	0.73	3.13	0.78	0.02	0.78
VC249418	3,753	3.29	0.79	3.30	0.81	0.02	0.71	65	99	0.71	3.30	0.77	0.01	0.76
VC251464	2,234	3.27	0.78	3.25	0.81	−0.02	0.70	64	99	0.70	3.26	0.76	−0.01	0.75
VC251477	5,549	3.28	0.80	3.24	0.86	−0.05	**0.69**[a]	61	99	**0.70**[a]	3.24	0.83	−0.05	0.74
VC251575	1,561	3.20	0.83	3.24	0.88	0.05	0.71	61	99	0.72	3.23	0.85	0.03	0.75
VC251577	1,824	3.42	0.82	3.38	0.87	−0.04	0.72	63	99	0.72	3.38	0.84	−0.05	0.76
VC390618	2,912	3.09	0.84	3.07	0.90	−0.03	0.70	59	99	0.71	3.06	0.86	−0.04	0.74
VC048246	3,982	3.29	0.86	3.30	0.88	0.01	0.75	65	99	0.75	3.29	0.86	0.00	0.79
VC048273	5,326	3.31	0.85	3.29	0.91	−0.03	0.73	62	99	0.73	3.28	0.89	−0.04	0.77
VC048352	2,751	3.16	0.78	3.16	0.84	0.00	0.71	64	99	0.71	3.16	0.81	0.00	0.76
VC048390	2,463	3.12	0.85	3.10	0.88	−0.02	0.73	62	99	0.73	3.09	0.85	−0.04	0.77
VC048411	5,386	3.18	0.82	3.20	0.87	0.03	0.71	62	99	0.71	3.20	0.84	0.03	0.76
VC069378	3,391	3.32	0.78	3.29	0.84	−0.03	0.71	64	99	0.71	3.29	0.81	−0.03	0.76
VC069382	2,599	3.32	0.85	3.33	0.93	0.01	0.71	59	98	0.71	3.33	0.90	0.02	0.75
VC069400	4,464	3.02	0.77	3.01	0.80	−0.01	**0.68**[a]	64	99	**0.68**[a]	3.01	0.76	−0.02	0.73
VC084829	7,802	3.09	0.80	3.08	0.85	−0.01	**0.69**[a]	61	99	**0.69**[a]	3.08	0.81	−0.01	0.73
VC084835	6,145	3.10	0.81	3.12	0.84	0.03	**0.67**[a]	60	98	**0.67**	3.12	0.80	0.03	0.71
VC084851	4,323	3.21	0.79	3.18	0.83	−0.05	**0.69**[a]	62	99	**0.69**[a]	3.17	0.78	−0.05	0.73
VC084853	3,591	3.22	0.82	3.24	0.87	0.03	0.72	64	99	0.72	3.23	0.83	0.02	0.76
VC086526	3,265	3.15	0.82	3.15	0.85	0.00	**0.69**[a]	61	99	**0.69**[a]	3.14	0.82	−0.01	0.74
VC093524	2,772	3.22	0.86	3.22	0.89	0.01	0.72	61	99	0.72	3.22	0.86	0.00	0.76
VC093532	1,388	3.16	0.88	3.16	0.93	0.00	0.75	62	99	0.75	3.15	0.91	−0.02	0.78
VC101021	4,177	3.25	0.82	3.25	0.87	0.00	0.72	64	99	0.72	3.24	0.83	−0.01	0.76
VC101037	2,869	3.12	0.83	3.12	0.88	−0.01	0.70	60	99	0.70	3.11	0.85	−0.02	0.75
VC101537	1,390	3.17	0.85	3.16	0.86	0.00	0.70	61	99	0.70	3.15	0.83	−0.02	0.74
VC101541	5,879	3.11	0.76	3.12	0.79	0.02	**0.68**[a]	65	99	**0.68**[a]	3.12	0.74	0.02	0.73
VC207455	2,360	3.30	0.83	3.33	0.87	0.03	0.71	62	99	0.71	3.32	0.84	0.03	0.75
VC207640	2,920	3.33	0.87	3.32	0.90	0.00	0.74	63	99	0.74	3.32	0.85	−0.01	0.77
VC209497	2,656	3.30	0.85	3.28	0.88	−0.03	0.73	62	99	0.73	3.28	0.85	−0.03	0.77
VC248469	1,601	3.04	0.83	3.06	0.84	0.02	0.74	65	99	0.74	3.05	0.81	0.01	0.78
VC250603	1,704	3.38	0.83	3.35	0.86	−0.04	0.72	63	99	0.72	3.36	0.84	−0.03	0.76
VC251468	4,158	3.29	0.80	3.28	0.87	−0.02	0.72	64	99	0.72	3.28	0.83	−0.02	0.76
VC251474	1,673	3.22	0.84	3.22	0.88	0.01	0.71	62	99	0.71	3.21	0.84	0.00	0.76
VC251573	1,398	3.04	0.87	3.05	0.92	0.02	0.71	60	98	0.71	3.06	0.90	0.03	0.75
VC390606	1,512	3.19	0.83	3.20	0.86	0.00	0.73	64	99	0.73	3.20	0.83	0.01	0.77
VC462771	2,973	3.14	0.83	3.12	0.85	−0.01	0.71	64	98	0.71	3.12	0.81	−0.02	0.75
VC101540	7,724	3.31	0.80	3.33	0.86	0.03	0.72	64	99	0.72	3.32	0.82	0.02	0.76
VC250595	4,366	3.31	0.83	3.33	0.90	0.02	0.73	64	99	0.74	3.32	0.85	0.01	0.78
VC101018	7,004	3.15	0.81	3.14	0.87	−0.01	**0.69**[a]	61	99	**0.69**[a]	3.14	0.83	−0.02	0.74
VC251475	3,608	3.24	0.83	3.26	0.89	0.03	0.74	64	99	0.74	3.26	0.85	0.02	0.78
VC390614	10,820	3.26	0.79	3.25	0.83	−0.02	0.72	66	99	0.72	3.24	0.79	−0.02	0.76
VC101050	2,769	3.26	0.83	3.26	0.88	0.00	**0.70**[a]	62	98	**0.70**[a]	3.26	0.85	0.00	0.74
VC177590	4,506	3.29	0.79	3.30	0.85	0.01	**0.69**[a]	62	99	**0.69**[a]	3.29	0.82	0.00	0.72
VC248460	3,911	3.30	0.81	3.29	0.87	−0.02	0.73	64	99	0.73	3.28	0.84	−0.03	0.77
VC248479	3,285	3.19	0.80	3.18	0.86	0.00	0.71	63	99	0.71	3.18	0.82	−0.01	0.75
VC248488	2,185	3.19	0.82	3.18	0.85	−0.01	**0.69**[a]	61	99	**0.69**[a]	3.19	0.81	0.00	0.74
VC250589	2,796	3.23	0.82	3.22	0.84	−0.01	0.70	63	99	0.70	3.22	0.80	−0.01	0.75
VC251576	5,695	3.27	0.82	3.21	0.87	−0.07	0.70	61	99	0.70	3.21	0.83	−0.08	0.74

Table G3: Continued

		Human 1 by e-rater (rounded to integers)									Human 1 by e-rater (unrounded)			
		Human 1		e-rater		Statistic					e-rater		Statistic	
Prompt	*N*	*M*	*SD*	*M*	*SD*	Std diff	Wtd kappa	% agree	% adj agree	*r*	*M*	*SD*	Std diff	*r*
VC390640	4,394	3.19	0.80	3.18	0.88	−0.02	**0.65**[a]	58	98	**0.66**[a]	3.17	0.84	−0.03	**0.69**[a]
VC462770	3,891	3.25	0.82	3.24	0.87	−0.02	0.70	61	99	0.71	3.24	0.83	−0.02	0.74
VE096305	2,279	3.36	0.82	3.31	0.82	−0.07	**0.69**[a]	62	99	**0.69**[a]	3.31	0.78	−0.07	0.74
VC069380	4,914	3.32	0.79	3.33	0.82	0.01	**0.70**[a]	64	99	**0.70**[a]	3.33	0.78	0.01	0.74
VC084843	4,052	3.25	0.78	3.25	0.82	0.01	**0.69**[a]	63	99	**0.69**[a]	3.25	0.78	0.00	0.74
VC084846	4,839	3.37	0.78	3.34	0.83	−0.03	0.72	66	99	0.73	3.34	0.79	−0.04	0.77
VC101016	6,573	3.16	0.82	3.12	0.88	−0.04	0.72	63	99	0.72	3.12	0.83	−0.04	0.75
VC101539	5,453	3.34	0.81	3.33	0.87	−0.02	0.72	65	99	0.73	3.31	0.84	−0.03	0.76
VC140094	2,658	3.29	0.81	3.33	0.86	0.04	0.73	65	99	0.73	3.32	0.81	0.03	0.77
VC209485	3,031	3.24	0.81	3.27	0.87	0.03	0.72	65	99	0.72	3.26	0.83	0.02	0.76
VC248473	3,730	3.29	0.82	3.29	0.88	0.01	0.74	65	99	0.75	3.28	0.84	−0.01	0.78

Note. Std diff = standardized difference; wtd = weighted; adj = adjacent.

[a]Agreement indices that failed to meet the guideline thresholds (also indicated by boldface).

Table G4 Phase II Agreement of Human and e-rater Scores on Argument Prompts: PS-12 Model

		Human 1 by e-rater (rounded to integers)									Human 1 by e-rater (unrounded)			
		Human 1		e-rater		Statistic					e-rater		Statistic	
Prompt	*N*	*M*	*SD*	*M*	*SD*	Std diff	Wtd kappa	% agree	% adj agree	*r*	*M*	*SD*	Std diff	*r*
VC048186	4,890	3.29	0.78	3.28	0.83	−0.02	0.75	68	100	0.75	3.27	0.80	−0.03	0.79
VC048263	5,948	3.31	0.78	3.34	0.83	0.04	0.72	66	99	0.73	3.33	0.78	0.03	0.77
VC048268	4,708	3.29	0.83	3.26	0.90	−0.04	0.72	62	99	0.72	3.26	0.87	−0.04	0.76
VC048328	6,018	3.28	0.81	3.23	0.86	−0.06	0.72	63	99	0.72	3.23	0.83	−0.06	0.76
VC048389	4,041	3.22	0.81	3.22	0.87	0.00	0.73	65	99	0.73	3.22	0.82	0.00	0.76
VC048408	4,838	3.36	0.86	3.34	0.93	−0.03	0.72	61	98	0.72	3.33	0.90	−0.04	0.76
VC069377	3,747	3.28	0.81	3.27	0.85	−0.01	0.74	66	99	0.74	3.26	0.82	−0.01	0.78
VC069384	5,564	3.30	0.79	3.31	0.84	0.01	0.73	66	99	0.73	3.31	0.80	0.01	0.77
VC069394	5,131	3.38	0.82	3.38	0.86	0.01	0.73	65	99	0.74	3.38	0.82	0.00	0.77
VC069396	4,869	3.22	0.86	3.19	0.89	−0.03	0.73	62	99	0.73	3.20	0.85	−0.03	0.77
VC084832	4,246	3.31	0.85	3.31	0.89	0.00	0.74	64	99	0.74	3.30	0.86	−0.01	0.78
VC084840	6,954	3.32	0.77	3.32	0.82	0.00	0.70	65	99	0.70	3.32	0.77	0.01	0.76
VC084849	3,561	3.34	0.79	3.34	0.85	0.00	0.71	64	99	0.71	3.33	0.81	−0.01	0.76
VC086531	895	3.06	0.81	3.03	0.87	−0.05	**0.67**[a]	59	98	**0.67**[a]	3.03	0.82	−0.04	0.72
VC101052	5,179	3.20	0.82	3.20	0.86	−0.01	0.71	62	99	0.72	3.19	0.82	−0.02	0.76
VC101056	4,611	3.29	0.81	3.30	0.83	0.01	0.71	64	99	0.71	3.29	0.79	0.00	0.76
VC101542	4,133	3.28	0.81	3.30	0.84	0.02	0.75	67	99	0.75	3.30	0.81	0.02	0.79
VC140314	418	3.11	0.79	3.14	0.84	0.03	0.73	67	99	0.73	3.13	0.78	0.02	0.78
VC249418	3,753	3.29	0.79	3.30	0.82	0.01	0.72	65	99	0.72	3.30	0.78	0.01	0.76
VC251464	2,234	3.27	0.78	3.25	0.81	−0.02	0.70	64	99	0.70	3.26	0.76	−0.01	0.75
VC251477	5,549	3.28	0.80	3.24	0.86	−0.04	**0.70**[a]	61	99	**0.70**[a]	3.24	0.82	−0.05	0.74
VC251575	1,561	3.20	0.83	3.25	0.89	0.05	0.71	60	99	0.72	3.23	0.85	0.03	0.75
VC251577	1,824	3.42	0.82	3.38	0.88	−0.05	0.72	63	99	0.73	3.38	0.84	−0.05	0.76
VC390618	2,912	3.09	0.84	3.07	0.90	−0.03	0.72	60	99	0.72	3.06	0.87	−0.04	0.75
VC048246	3,982	3.29	0.86	3.30	0.89	0.01	0.75	65	99	0.75	3.29	0.86	0.00	0.79
VC048273	5,326	3.31	0.85	3.28	0.91	−0.04	0.73	61	99	0.73	3.28	0.89	−0.04	0.77

Table G4: Continued

		Human 1 by e-rater (rounded to integers)									Human 1 by e-rater (unrounded)			
		Human 1		e-rater		Statistic					e-rater		Statistic	
Prompt	*N*	*M*	*SD*	*M*	*SD*	Std diff	Wtd kappa	% agree	% adj agree	*r*	*M*	*SD*	Std diff	*r*
VC048352	2,751	3.16	0.78	3.15	0.85	−0.01	0.72	64	99	0.72	3.16	0.81	0.00	0.76
VC048390	2,463	3.12	0.85	3.10	0.88	−0.03	0.73	62	99	0.73	3.09	0.85	−0.04	0.77
VC048411	5,386	3.18	0.82	3.20	0.87	0.03	0.71	62	99	0.72	3.21	0.84	0.03	0.76
VC069378	3,391	3.32	0.78	3.29	0.84	−0.03	0.71	64	99	0.71	3.29	0.81	−0.03	0.76
VC069382	2,599	3.32	0.85	3.33	0.93	0.01	0.71	59	98	0.71	3.33	0.90	0.02	0.75
VC069400	4,464	3.02	0.77	3.01	0.80	−0.02	**0.68**[a]	64	99	**0.68**[a]	3.00	0.76	−0.02	0.73
VC084829	7,802	3.09	0.80	3.08	0.86	−0.01	**0.69**[a]	61	99	**0.69**[a]	3.08	0.81	−0.01	0.74
VC084835	6,145	3.10	0.81	3.12	0.84	0.02	**0.68**[a]	61	98	**0.68**[a]	3.12	0.79	0.03	0.72
VC084851	4,323	3.21	0.79	3.18	0.83	−0.05	**0.69**[a]	62	99	**0.69**[a]	3.17	0.78	−0.05	0.73
VC084853	3,591	3.22	0.82	3.24	0.87	0.02	0.72	64	99	0.72	3.23	0.83	0.01	0.76
VC086526	3,265	3.15	0.82	3.15	0.85	0.00	**0.69**[a]	61	99	**0.69**[a]	3.14	0.82	−0.01	0.74
VC093524	2,772	3.22	0.86	3.23	0.89	0.01	0.72	61	99	0.72	3.21	0.86	0.00	0.77
VC093532	1,388	3.16	0.88	3.16	0.93	0.00	0.74	62	99	0.75	3.15	0.91	−0.02	0.78
VC101021	4,177	3.25	0.82	3.26	0.87	0.01	0.72	63	99	0.72	3.25	0.83	−0.01	0.76
VC101037	2,869	3.12	0.83	3.11	0.89	−0.01	0.70	60	99	0.70	3.11	0.85	−0.02	0.75
VC101537	1,390	3.17	0.85	3.17	0.86	0.00	0.71	61	99	0.71	3.15	0.83	−0.02	0.74
VC101541	5,879	3.11	0.76	3.13	0.79	0.03	**0.68**[a]	65	99	**0.68**[a]	3.13	0.74	0.02	0.73
VC207455	2,360	3.30	0.83	3.33	0.87	0.04	0.71	62	99	0.71	3.32	0.84	0.03	0.75
VC207640	2,920	3.33	0.87	3.32	0.90	0.00	0.74	63	99	0.74	3.32	0.85	−0.01	0.77
VC209497	2,656	3.30	0.85	3.28	0.88	−0.02	0.73	62	99	0.73	3.28	0.85	−0.03	0.77
VC248469	1,601	3.04	0.83	3.06	0.84	0.02	0.73	65	99	0.74	3.05	0.81	0.01	0.78
VC250603	1,704	3.38	0.83	3.35	0.86	−0.04	0.72	63	99	0.72	3.35	0.84	−0.03	0.76
VC251468	4,158	3.29	0.80	3.28	0.87	−0.01	0.71	63	99	0.72	3.28	0.83	−0.01	0.76
VC251474	1,673	3.22	0.84	3.22	0.88	0.01	0.72	62	99	0.72	3.22	0.84	0.00	0.76
VC251573	1,398	3.04	0.87	3.05	0.92	0.02	0.71	59	98	0.71	3.05	0.89	0.02	0.75
VC390606	1,512	3.19	0.83	3.21	0.87	0.02	0.73	64	99	0.74	3.21	0.83	0.01	0.77
VC462771	2,973	3.14	0.83	3.12	0.85	−0.02	0.72	64	99	0.72	3.12	0.81	−0.03	0.75
VC101540	7,724	3.31	0.80	3.33	0.86	0.03	0.72	65	99	0.73	3.32	0.82	0.02	0.77
VC250595	4,366	3.31	0.83	3.33	0.89	0.02	0.73	64	99	0.74	3.32	0.85	0.01	0.78
VC101018	7,004	3.15	0.81	3.14	0.86	−0.01	**0.69**[a]	61	99	**0.69**[a]	3.14	0.83	−0.02	0.74
VC251475	3,608	3.24	0.83	3.26	0.89	0.03	0.74	64	99	0.74	3.26	0.85	0.02	0.78
VC390614	10,820	3.26	0.79	3.25	0.83	−0.02	0.72	66	99	0.72	3.24	0.79	−0.02	0.76
VC101050	2,769	3.26	0.83	3.27	0.88	0.01	**0.70**[a]	62	98	**0.70**[a]	3.26	0.85	0.00	0.74
VC177590	4,506	3.29	0.79	3.30	0.85	0.00	**0.68**[a]	62	99	**0.69**[a]	3.29	0.81	0.00	0.72
VC248460	3,911	3.30	0.81	3.29	0.87	−0.02	0.73	65	99	0.73	3.28	0.84	−0.03	0.77
VC248479	3,285	3.19	0.80	3.19	0.86	0.00	0.71	63	99	0.71	3.18	0.83	−0.01	0.76
VC248488	2,185	3.19	0.82	3.18	0.86	−0.01	**0.70**[a]	62	99	**0.70**[a]	3.19	0.81	0.00	0.74
VC250589	2,796	3.23	0.82	3.22	0.84	−0.01	0.71	63	99	0.71	3.21	0.80	−0.02	0.75
VC251576	5,695	3.27	0.82	3.21	0.87	−0.07	**0.70**[a]	61	99	0.70	3.21	0.83	−0.08	0.74
VC390640	4,394	3.19	0.80	3.18	0.88	−0.01	**0.66**[a]	58	98	**0.66**[a]	3.18	0.85	−0.02	**0.70**
VC462770	3,891	3.25	0.82	3.24	0.87	−0.01	0.70	61	99	0.70	3.24	0.83	−0.01	0.74
VE096305	2,279	3.36	0.82	3.31	0.82	−0.07	**0.69**[a]	62	99	**0.69**[a]	3.31	0.78	−0.07	0.74
VC069380	4,914	3.32	0.79	3.33	0.83	0.01	**0.70**[a]	64	99	**0.70**[a]	3.33	0.78	0.01	0.74
VC084843	4,052	3.25	0.78	3.25	0.83	0.01	**0.70**[a]	63	99	**0.70**[a]	3.25	0.78	0.00	0.74
VC084846	4,839	3.37	0.78	3.34	0.83	−0.03	0.72	66	99	0.73	3.34	0.79	−0.04	0.77
VC101016	6,573	3.16	0.82	3.12	0.88	−0.04	0.72	63	99	0.72	3.12	0.83	−0.04	0.75
VC101539	5,453	3.34	0.81	3.32	0.87	−0.02	0.73	65	99	0.73	3.32	0.84	−0.03	0.76
VC140094	2,658	3.29	0.81	3.33	0.85	0.04	0.73	66	99	0.73	3.32	0.81	0.04	0.77
VC209485	3,031	3.24	0.81	3.27	0.87	0.03	0.72	65	99	0.72	3.26	0.83	0.02	0.76
VC248473	3,730	3.29	0.82	3.29	0.88	0.01	0.74	65	99	0.75	3.28	0.84	−0.01	0.78

Note. Std diff = standardized difference; wtd = weighted; adj = adjacent. [a]Agreement indices that failed to meet the guideline thresholds (also indicated by boldface).

Table G5 Phase II Agreement of Human and e-rater Scores on Issue Prompts: G-10 Model

		Human 1 by e-rater (rounded to integers)									Human 1 by e-rater (unrounded)			
		Human 1		e-rater		Statistic					e-rater		Statistic	
Prompt	*N*	*M*	*SD*	*M*	*SD*	Std diff	Wtd kappa	% agree	% adj agree	*r*	*M*	*SD*	Std diff	*r*
VC047733	6, 030	3.17	0.77	3.11	0.80	−0.07	0.76	72	100	0.76	3.11	0.76	−0.07	0.81
VC047741	5, 212	3.16	0.76	3.15	0.81	−0.02	0.74	69	100	0.74	3.14	0.76	−0.02	0.79
VC047750	2, 693	3.27	0.73	3.28	0.77	0.01	0.73	70	100	0.73	3.27	0.72	0.00	0.79
VC047757	3, 108	3.17	0.78	3.09	0.82	−0.10	0.77	71	100	0.77	3.09	0.78	−0.10	0.81
VC047770	4, 331	3.18	0.81	3.12	0.83	−0.07	0.77	70	100	0.77	3.12	0.79	−0.07	0.82
VC047788	6, 299	3.31	0.77	3.31	0.79	0.01	0.75	70	100	0.75	3.30	0.74	−0.01	0.79
VC047792	4, 143	3.26	0.79	3.20	0.81	−0.07	0.76	70	100	0.76	3.20	0.77	−0.08	0.81
VC047799	5, 201	3.14	0.78	3.02	0.82	−0.15	0.74	68	100	0.75	3.01	0.78	**−0.17**[a]	0.80
VC047822	6, 826	3.26	0.80	3.17	0.81	−0.11	0.75	69	99	0.75	3.16	0.77	−0.13	0.80
VC048013	3, 077	3.16	0.83	3.03	0.84	**−0.16**[a]	0.78	70	100	0.79	3.02	0.80	**−0.17**[a]	0.83
VC048019	3, 196	3.10	0.79	3.05	0.83	−0.07	0.75	69	100	0.76	3.05	0.80	−0.07	0.80
VC048027	5, 942	3.31	0.77	3.27	0.78	−0.05	0.75	71	100	0.75	3.27	0.73	−0.05	0.80
VC048031	4, 405	3.33	0.79	3.38	0.79	0.06	0.76	71	100	0.76	3.37	0.75	0.05	0.81
VC073155	6, 117	3.24	0.78	3.16	0.80	−0.10	0.75	70	100	0.76	3.16	0.75	−0.10	0.80
VC073160	4, 729	3.22	0.78	3.22	0.79	−0.01	0.76	71	100	0.76	3.21	0.74	−0.01	0.81
VC073163	3, 329	3.13	0.82	3.06	0.85	−0.08	0.77	69	100	0.77	3.06	0.81	−0.08	0.81
VC073164	3, 105	3.24	0.76	3.26	0.78	0.03	0.76	71	100	0.76	3.26	0.74	0.03	0.79
VC073166	1, 440	3.18	0.77	3.15	0.82	−0.05	0.76	70	100	0.76	3.15	0.77	−0.04	0.80
VC073168	2, 982	3.20	0.80	3.19	0.82	−0.02	0.76	69	100	0.76	3.19	0.76	−0.01	0.80
VC073173	4, 437	3.19	0.81	3.13	0.84	−0.08	0.77	69	100	0.77	3.12	0.80	−0.09	0.81
VC073175	1, 704	3.10	0.79	2.99	0.81	−0.14	0.75	68	100	0.75	2.99	0.79	−0.15	0.80
VC073176	1, 436	3.17	0.81	3.13	0.82	−0.05	0.75	69	100	0.76	3.11	0.77	−0.07	0.80
VC104275	2, 618	3.15	0.80	3.16	0.82	0.01	0.76	70	100	0.76	3.15	0.77	−0.01	0.81
VC104276	2, 720	3.25	0.80	3.26	0.81	0.02	0.77	71	100	0.77	3.25	0.76	0.01	0.81
VC104278	4, 095	3.32	0.80	3.20	0.79	−0.15	0.74	68	99	0.74	3.19	0.74	**−0.16**[a]	0.79
VC155074	993	3.19	0.80	3.01	0.84	**−0.22**[a]	0.74	65	100	0.76	3.02	0.79	**−0.21**[a]	0.81
VC219551	2, 044	3.30	0.76	3.35	0.83	0.06	0.79	74	100	0.79	3.36	0.78	0.08	0.82
VC515323	2, 553	3.25	0.77	3.13	0.80	**−0.15**[a]	0.75	69	100	0.76	3.13	0.77	−0.15	0.81
VC787322	1, 853	3.14	0.79	3.05	0.84	−0.11	0.79	73	100	0.80	3.03	0.80	−0.13	0.83
VC929101	1, 934	3.39	0.78	3.47	0.78	0.10	0.75	70	100	0.75	3.47	0.74	0.10	0.80
VC929114	2, 458	3.38	0.77	3.41	0.79	0.05	0.76	71	100	0.76	3.41	0.74	0.04	0.81
VC929139	2, 346	3.24	0.78	3.12	0.80	−0.14	0.76	71	100	0.77	3.13	0.77	−0.14	0.81
VC048048	3, 536	3.28	0.77	3.37	0.82	0.11	0.76	70	100	0.77	3.36	0.78	0.10	0.81
VC048051	5, 374	3.27	0.73	3.26	0.78	0.00	0.75	72	100	0.75	3.26	0.73	0.00	0.79
VC048070	3, 803	3.24	0.76	3.27	0.80	0.04	0.75	71	100	0.75	3.28	0.74	0.05	0.79
VC048077	5, 094	3.12	0.78	3.08	0.82	−0.05	0.76	69	100	0.76	3.08	0.79	−0.05	0.80
VC048141	4, 149	3.29	0.73	3.38	0.77	0.12	0.74	71	100	0.75	3.38	0.72	0.12	0.79
VC073157	4, 605	3.30	0.77	3.30	0.80	−0.01	0.76	71	100	0.76	3.29	0.75	−0.01	0.80
VC073158	2, 961	3.25	0.78	3.31	0.80	0.08	0.77	72	100	0.77	3.30	0.75	0.07	0.80
VC073169	7, 645	3.12	0.80	3.14	0.82	0.03	0.76	69	100	0.76	3.14	0.78	0.03	0.80
VC073172	2, 499	3.19	0.77	3.21	0.81	0.03	0.75	71	100	0.76	3.21	0.76	0.03	0.80
VC104280	3, 378	3.39	0.75	3.48	0.79	0.11	0.75	70	100	0.75	3.47	0.74	0.11	0.80
VC104281	4, 859	3.42	0.76	3.54	0.77	**0.16**[a]	0.73	69	100	0.74	3.54	0.72	**0.16**[a]	0.79
VC084819	6, 588	3.34	0.76	3.49	0.76	**0.19**[a]	0.72	68	100	0.73	3.49	0.71	**0.20**[a]	0.78
VC084820	4, 555	3.23	0.77	3.28	0.79	0.06	0.72	67	99	0.72	3.27	0.74	0.05	0.77
VC155042	8, 352	3.26	0.76	3.36	0.79	0.13	0.72	68	99	0.73	3.36	0.74	0.13	0.78
VC219591	6, 526	3.37	0.72	3.48	0.77	**0.15**[a]	0.73	71	100	0.74	3.48	0.72	**0.16**[a]	0.79
VC787354	6, 048	3.29	0.77	3.38	0.78	0.11	0.74	70	100	0.75	3.37	0.73	0.11	0.80
VC084798	6, 213	3.13	0.75	3.15	0.80	0.02	0.75	71	100	0.75	3.14	0.76	0.01	0.80
VC084799	5, 885	2.95	0.78	2.96	0.81	0.01	0.73	68	100	0.73	2.95	0.77	0.00	0.78
VC084804	5, 371	3.21	0.74	3.36	0.80	**0.19**[a]	0.72	67	99	0.73	3.36	0.76	**0.19**[a]	0.77
VC104286	7, 865	3.18	0.77	3.15	0.82	−0.03	0.75	69	100	0.75	3.15	0.78	−0.04	0.79
VE096407	7, 957	3.03	0.76	2.96	0.81	−0.08	0.73	68	100	0.74	2.96	0.77	−0.08	0.78

Table G5: Continued

		Human 1 by e-rater (rounded to integers)									Human 1 by e-rater (unrounded)			
		Human 1		e-rater		Statistic					e-rater		Statistic	
Prompt	*N*	*M*	*SD*	*M*	*SD*	Std diff	Wtd kappa	% agree	% adj agree	*r*	*M*	*SD*	Std diff	*r*
VC084555	4, 175	3.06	0.74	3.11	0.84	0.06	0.70	65	99	0.71	3.10	0.81	0.04	0.76
VC084754	6, 319	3.02	0.73	2.88	0.83	**−0.18**[a]	**0.70**[a]	64	100	0.72	2.87	0.78	**−0.19**[a]	0.77
VC104284	6, 034	3.06	0.69	3.10	0.80	0.06	**0.67**[a]	66	99	**0.68**[a]	3.11	0.74	0.06	0.73
VC155075	2, 204	3.07	0.74	3.06	0.83	−0.01	0.73	69	100	0.74[a]	3.05	0.77	−0.02	0.77
VC155078	4, 367	3.18	0.70	3.24	0.78	0.08	**0.68**[a]	66	99	**0.69**[a]	3.24	0.73	0.08	0.73
VC178591	2, 389	3.16	0.72	3.34	0.79	**0.24**[a]	**0.67**[a]	64	99	**0.70**[a]	3.33	0.75	**0.23**[a]	0.74
VC178595	4, 819	3.26	0.70	3.51	0.78	**0.34**[a]	**0.64**[a]	62	99	**0.68**[a]	3.51	0.74	**0.35**[a]	0.73
VC178602	4, 294	3.11	0.77	3.10	0.82	−0.01	0.73	66	100	0.73	3.09	0.78	−0.02	0.78
VC515311	6, 006	3.08	0.70	3.14	0.80	0.08	**0.69**[a]	66	100	**0.70**[a]	3.14	0.75	0.08	0.75
VC787323	4, 327	3.10	0.72	3.13	0.80	0.05	0.70	67	99	0.71	3.14	0.76	0.05	0.76
VC787333	1, 866	3.12	0.76	3.29	0.82	**0.21**[a]	0.72	65	100	0.73	3.28	0.77	**0.21**[a]	0.78
VC084809	3, 686	2.99	0.76	2.81	0.83	**−0.22**[a]	0.72	66	99	0.75	2.80	0.80	**−0.24**[a]	0.79
VC104290	4, 689	3.19	0.78	3.16	0.83	−0.03	0.76	70	100	0.76	3.15	0.79	−0.05	0.80
VC104293	2, 971	3.27	0.78	3.24	0.83	−0.04	0.78	71	100	0.78	3.23	0.79	−0.05	0.82
VC104297	2, 213	3.00	0.80	2.95	0.84	−0.06	0.74	66	99	0.74	2.94	0.81	−0.08	0.79
VC104300	4, 413	3.24	0.76	3.15	0.81	−0.12	0.75	70	100	0.76	3.14	0.78	−0.13	0.80
VC104302	4, 727	3.03	0.78	2.93	0.82	−0.13	0.74	67	100	0.74	2.92	0.78	−0.15	0.79
VC155043	2, 531	3.27	0.80	3.32	0.81	0.06	0.76	70	100	0.76	3.31	0.77	0.05	0.80
VC515320	1, 957	3.15	0.79	3.07	0.81	−0.10	0.74	69	100	0.75	3.08	0.77	−0.10	0.80
VC787346	1, 691	3.25	0.79	3.30	0.81	0.05	0.77	71	100	0.77	3.29	0.78	0.05	0.82
VE096379	2, 398	3.11	0.76	2.96	0.83	**−0.19**[a]	0.75	69	100	0.77	2.95	0.78	**−0.20**[a]	0.81
VE096386	1, 644	3.18	0.80	3.03	0.84	**−0.18**[a]	0.76	68	100	0.78	3.03	0.81	**−0.18**[a]	0.83
VE096411	2, 316	3.17	0.77	3.12	0.81	−0.07	0.74	69	99	0.74	3.11	0.77	−0.08	0.78

Note. Std diff = standardized difference; wtd = weighted; adj = adjacent.

[a]Agreement indices that failed to meet the guideline thresholds (also indicated by boldface).

Table G6 Phase II Agreement of Human and e-rater Scores on Issue Prompts: G-12 Model

		Human 1 by e-rater (rounded to integers)									Human 1 by e-rater (unrounded)			
		Human 1		e-rater		Statistic					e-rater		Statistic	
Prompt	*N*	*M*	*SD*	*M*	*SD*	Std diff	Wtd kappa	% agree	% adj agree	*r*	*M*	*SD*	Std diff	*r*
VC047733	6,030	3.17	0.77	3.11	0.80	−0.07	0.76	72	100	0.76	3.11	0.76	−0.07	0.81
VC047741	5,212	3.16	0.76	3.14	0.81	−0.02	0.74	69	100	0.74	3.14	0.76	−0.02	0.79
VC047750	2,693	3.27	0.73	3.28	0.77	0.01	0.73	71	100	0.73	3.27	0.72	0.00	0.79
VC047757	3,108	3.17	0.78	3.08	0.81	−0.11	0.77	71	100	0.77	3.09	0.78	−0.10	0.81
VC047770	4,331	3.18	0.81	3.12	0.83	−0.07	0.77	70	100	0.78	3.12	0.79	−0.07	0.82
VC047788	6,299	3.31	0.77	3.31	0.79	0.01	0.75	70	100	0.75	3.30	0.74	−0.01	0.79
VC047792	4,143	3.26	0.79	3.20	0.81	−0.07	0.76	70	100	0.76	3.20	0.77	−0.08	0.81
VC047799	5,201	3.14	0.78	3.02	0.82	**−0.15**[a]	0.74	68	100	0.75	3.01	0.78	**−0.17**[a]	0.80
VC047822	6,826	3.26	0.80	3.16	0.81	−0.11	0.75	69	99	0.75	3.16	0.77	−0.13	0.80
VC048013	3,077	3.16	0.83	3.02	0.84	**−0.16**[a]	0.78	69	100	0.79	3.02	0.80	**−0.17**[a]	0.83
VC048019	3,196	3.10	0.79	3.05	0.84	−0.07	0.75	68	100	0.75	3.05	0.80	−0.07	0.80
VC048027	5,942	3.31	0.77	3.27	0.78	−0.05	0.75	70	100	0.75	3.27	0.73	−0.05	0.80
VC048031	4,405	3.33	0.79	3.38	0.79	0.07	0.76	71	100	0.76	3.38	0.75	0.06	0.81
VC073155	6,117	3.24	0.78	3.16	0.80	−0.10	0.75	70	100	0.76	3.16	0.76	−0.10	0.80
VC073160	4,729	3.22	0.78	3.22	0.79	−0.01	0.76	71	100	0.76	3.22	0.74	−0.01	0.81
VC073163	3,329	3.13	0.82	3.06	0.85	−0.09	0.77	69	100	0.77	3.06	0.81	−0.09	0.81

Table G6: Continued

		Human 1 by e-rater (rounded to integers)									Human 1 by e-rater (unrounded)			
		Human 1		e-rater		Statistic					e-rater		Statistic	
Prompt	*N*	*M*	*SD*	*M*	*SD*	Std diff	Wtd kappa	% agree	% adj agree	*r*	*M*	*SD*	Std diff	*r*
VC073164	3,105	3.24	0.76	3.26	0.78	0.03	0.75	71	100	0.76	3.26	0.74	0.03	0.79
VC073166	1,440	3.18	0.77	3.15	0.82	−0.04	0.76	70	100	0.76	3.16	0.77	−0.04	0.80
VC073168	2,982	3.20	0.80	3.19	0.82	−0.01	0.76	69	100	0.76	3.19	0.76	−0.01	0.80
VC073173	4,437	3.19	0.81	3.12	0.84	−0.08	0.77	69	100	0.77	3.11	0.80	−0.09	0.81
VC073175	1,704	3.10	0.79	2.99	0.82	−0.14	0.75	68	100	0.76	2.98	0.79	−0.15	0.80
VC073176	1,436	3.17	0.81	3.13	0.81	−0.05	0.76	69	100	0.76	3.11	0.77	−0.07	0.80
VC104275	2,618	3.15	0.80	3.16	0.82	0.02	0.76	70	100	0.76	3.15	0.77	0.00	0.81
VC104276	2,720	3.25	0.80	3.26	0.81	0.02	0.77	71	100	0.77	3.25	0.76	0.01	0.81
VC104278	4,095	3.32	0.80	3.20	0.79	−0.15	0.73	68	99	0.74	3.19	0.74	**−0.16**[a]	0.79
VC155074	993	3.19	0.80	3.01	0.84	**−0.22**[a]	0.74	66	100	0.76	3.02	0.79	**−0.21**[a]	0.81
VC219551	2,044	3.30	0.76	3.35	0.83	0.07	0.79	74	100	0.79	3.36	0.78	0.08	0.82
VC515323	2,553	3.25	0.77	3.13	0.80	**−0.15**[a]	0.75	69	100	0.76	3.13	0.76	−0.15	0.81
VC787322	1,853	3.14	0.79	3.04	0.84	−0.12	0.79	72	100	0.80	3.03	0.80	−0.13	0.83
VC929101	1,934	3.39	0.78	3.47	0.78	0.10	0.75	70	100	0.75	3.47	0.74	0.10	0.80
VC929114	2,458	3.38	0.77	3.41	0.78	0.05	0.76	71	100	0.76	3.41	0.74	0.05	0.81
VC929139	2,346	3.24	0.78	3.12	0.80	−0.15	0.76	71	100	0.77	3.12	0.77	−0.15	0.81
VC048048	3,536	3.28	0.77	3.37	0.82	0.11	0.76	70	100	0.77	3.36	0.78	0.10	0.81
VC048051	5,374	3.27	0.73	3.26	0.78	0.00	0.74	72	100	0.75	3.26	0.73	0.00	0.79
VC048070	3,803	3.24	0.76	3.28	0.79	0.04	0.75	71	100	0.75	3.28	0.74	0.05	0.79
VC048077	5,094	3.12	0.78	3.08	0.83	−0.06	0.76	69	100	0.76	3.08	0.79	−0.06	0.80
VC048141	4,149	3.29	0.73	3.38	0.77	0.12	0.74	71	100	0.75	3.38	0.72	0.12	0.79
VC073157	4,605	3.30	0.77	3.30	0.80	−0.01	0.76	71	100	0.76	3.29	0.75	−0.01	0.80
VC073158	2,961	3.25	0.78	3.31	0.80	0.08	0.77	71	100	0.77	3.31	0.75	0.08	0.80
VC073169	7,645	3.12	0.80	3.14	0.82	0.03	0.76	69	100	0.76	3.14	0.78	0.03	0.80
VC073172	2,499	3.19	0.77	3.21	0.81	0.03	0.75	70	100	0.75	3.21	0.76	0.03	0.80
VC104280	3,378	3.39	0.75	3.48	0.79	0.12	0.74	70	100	0.75	3.48	0.74	0.11	0.80
VC104281	4,859	3.42	0.76	3.55	0.77	**0.17**[a]	0.73	69	100	0.74	3.55	0.72	**0.17**[a]	0.79
VC084819	6,588	3.34	0.76	3.49	0.76	**0.20**[a]	0.71	68	100	0.73	3.49	0.71	**0.21**[a]	0.78
VC084820	4,555	3.23	0.77	3.28	0.79	0.06	0.72	67	99	0.72	3.27	0.74	0.05	0.77
VC155042	8,352	3.26	0.76	3.36	0.79	0.13	0.72	68	99	0.73	3.36	0.74	0.13	0.78
VC219591	6,526	3.37	0.72	3.49	0.77	**0.16**[a]	0.73	71	100	0.74	3.49	0.72	**0.16**[a]	0.79
VC787354	6,048	3.29	0.77	3.38	0.78	0.11	0.74	70	100	0.75	3.37	0.73	0.11	0.80
VC084798	6,213	3.13	0.75	3.15	0.80	0.02	0.75	71	100	0.75	3.14	0.76	0.01	0.80
VC084799	5,885	2.95	0.78	2.95	0.81	0.00	0.73	68	100	0.73	2.95	0.77	0.00	0.78
VC084804	5,371	3.21	0.74	3.36	0.80	**0.19**[a]	0.72	67	99	0.73	3.36	0.76	**0.19**[a]	0.77
VC104286	7,865	3.18	0.77	3.15	0.82	−0.03	0.75	69	100	0.75	3.14	0.78	−0.04	0.79
VE096407	7,957	3.03	0.76	2.96	0.81	−0.08	0.74	68	100	0.74	2.96	0.77	−0.09	0.78
VC084555	4,175	3.06	0.74	3.11	0.84	0.06	0.70	66	99	0.71	3.10	0.81	0.04	0.76
VC084754	6,319	3.02	0.73	2.87	0.83	**−0.19**[a]	**0.70**[a]	64	100	0.72	2.87	0.79	**−0.20**[a]	0.77
VC104284	6,034	3.06	0.69	3.10	0.80	0.06	**0.67**[a]	65	99	**0.68**[a]	3.10	0.74	0.06	0.73
VC155075	2,204	3.07	0.74	3.06	0.83	−0.01	0.73	69	100	0.74	3.05	0.77	−0.02	0.77
VC155078	4,367	3.18	0.70	3.24	0.78	0.08	**0.68**[a]	66	99	**0.69**[a]	3.24	0.73	0.07	0.73
VC178591	2,389	3.16	0.72	3.34	0.79	**0.24**[a]	**0.67**[a]	64	99	**0.70**[a]	3.33	0.75	**0.24**[a]	0.74
VC178595	4,819	3.26	0.70	3.51	0.78	**0.34**[a]	**0.64**[a]	62	99	**0.68**[a]	3.52	0.74	**0.36**[a]	0.73
VC178602	4,294	3.11	0.77	3.10	0.82	−0.01	0.72	66	100	0.73	3.09	0.79	−0.02	0.78
VC515311	6,006	3.08	0.70	3.14	0.80	0.08	**0.69**[a]	66	100	**0.70**[a]	3.14	0.75	0.08	0.75
VC787323	4,327	3.10	0.72	3.13	0.80	0.05	0.70	67	99	0.71	3.14	0.76	0.05	0.76
VC787333	1,866	3.12	0.76	3.29	0.82	**0.21**[a]	0.72	66	100	0.74	3.28	0.77	**0.20**[a]	0.78
VC084809	3,686	2.99	0.76	2.81	0.83	**−0.22**[a]	0.73	66	99	0.75	2.80	0.80	**−0.24**[a]	0.79
VC104290	4,689	3.19	0.78	3.16	0.83	−0.03	0.76	70	100	0.76	3.15	0.79	−0.05	0.80
VC104293	2,971	3.27	0.78	3.24	0.83	−0.04	0.78	71	100	0.78	3.23	0.79	−0.05	0.82
VC104297	2,213	3.00	0.80	2.95	0.84	−0.06	0.74	66	99	0.74	2.94	0.81	−0.08	0.79

Table G6: Continued

		Human 1 by e-rater (rounded to integers)									Human 1 by e-rater (unrounded)			
		Human 1		e-rater		Statistic					e-rater		Statistic	
Prompt	*N*	*M*	*SD*	*M*	*SD*	Std diff	Wtd kappa	% agree	% adj agree	*r*	*M*	*SD*	Std diff	*r*
VC104300	4,413	3.24	0.76	3.14	0.81	−0.12	0.75	70	100	0.76	3.14	0.78	−0.14	0.80
VC104302	4,727	3.03	0.78	2.92	0.83	−0.13	0.74	67	100	0.74	2.91	0.79	**−0.15**[a]	0.79
VC155043	2,531	3.27	0.80	3.32	0.81	0.06	0.76	70	100	0.76	3.31	0.76	0.06	0.80
VC515320	1,957	3.15	0.79	3.07	0.81	−0.10	0.74	69	100	0.75	3.08	0.77	−0.10	0.80
VC787346	1,691	3.25	0.79	3.29	0.81	0.05	0.77	71	100	0.77	3.29	0.78	0.05	0.82
VE096379	2,398	3.11	0.76	2.96	0.83	**−0.19**[a]	0.76	69	100	0.77	2.95	0.78	**−0.20**[a]	0.81
VE096386	1,644	3.18	0.80	3.03	0.83	**−0.18**[a]	0.76	68	100	0.77	3.03	0.80	**−0.18**[a]	0.83
VE096411	2,316	3.17	0.77	3.12	0.81	−0.06	0.73	69	99	0.74	3.11	0.77	−0.07	0.78

Note. Std diff = standardized difference; wtd = weighted; adj = adjacent. [a]Agreement indices that failed to meet the guideline thresholds (also indicated by boldface).

Table G7 Phase II Agreement of Human and e-rater Scores on Issue Prompts: PS-10 Model

		Human 1 by e-rater (rounded to integers)									Human 1 by e-rater (unrounded)			
		Human 1		e-rater		Statistic					e-rater		Statistic	
Prompt	*N*	*M*	*SD*	*M*	*SD*	Std diff	Wtd kappa	% agree	% adj agree	*r*	*M*	*SD*	Std diff	*r*
VC047733	6,030	3.17	0.77	3.17	0.82	0.01	0.76	71	100	0.77	3.17	0.78	0.01	0.81
VC047741	5,212	3.16	0.76	3.15	0.81	−0.01	0.74	68	100	0.74	3.15	0.75	−0.02	0.79
VC047750	2,693	3.27	0.73	3.31	0.78	0.06	0.73	70	100	0.73	3.30	0.73	0.04	0.79
VC047757	3,108	3.17	0.78	3.15	0.83	−0.03	0.77	70	100	0.77	3.15	0.80	−0.02	0.81
VC047770	4,331	3.18	0.81	3.17	0.83	−0.01	0.77	71	100	0.77	3.16	0.79	−0.02	0.82
VC047788	6,299	3.31	0.77	3.32	0.82	0.02	0.75	69	100	0.75	3.31	0.77	0.01	0.79
VC047792	4,143	3.26	0.79	3.28	0.82	0.03	0.76	69	100	0.76	3.28	0.78	0.02	0.81
VC047799	5,201	3.14	0.78	3.12	0.83	−0.02	0.75	68	100	0.75	3.11	0.79	−0.03	0.80
VC047822	6,826	3.26	0.80	3.25	0.86	0.00	0.76	69	99	0.77	3.24	0.83	−0.02	0.80
VC048013	3,077	3.16	0.83	3.18	0.85	0.02	0.79	71	100	0.79	3.18	0.82	0.02	0.83
VC048019	3,196	3.10	0.79	3.11	0.84	0.01	0.76	69	99	0.76	3.11	0.80	0.02	0.80
VC048027	5,942	3.31	0.77	3.31	0.80	0.00	0.75	70	100	0.75	3.31	0.75	0.00	0.80
VC048031	4,405	3.33	0.79	3.33	0.86	0.00	0.77	70	100	0.77	3.33	0.82	0.00	0.81
VC073155	6,117	3.24	0.78	3.23	0.80	−0.01	0.76	71	100	0.76	3.22	0.75	−0.02	0.80
VC073160	4,729	3.22	0.78	3.24	0.83	0.02	0.76	70	100	0.76	3.24	0.79	0.01	0.81
VC073163	3,329	3.13	0.82	3.10	0.88	−0.03	0.78	69	100	0.78	3.11	0.84	−0.03	0.81
VC073164	3,105	3.24	0.76	3.25	0.81	0.01	0.76	71	100	0.76	3.24	0.76	0.00	0.80
VC073166	1,440	3.18	0.77	3.13	0.86	−0.06	0.76	69	100	0.76	3.13	0.81	−0.07	0.80
VC073168	2,982	3.20	0.80	3.18	0.86	−0.03	0.76	68	99	0.76	3.18	0.82	−0.03	0.80
VC073173	4,437	3.19	0.81	3.17	0.85	−0.02	0.77	69	100	0.77	3.16	0.81	−0.03	0.81
VC073175	1,704	3.10	0.79	3.05	0.84	−0.07	0.75	68	100	0.76	3.04	0.81	−0.08	0.80
VC073176	1,436	3.17	0.81	3.17	0.82	0.01	0.77	71	100	0.77	3.16	0.78	0.00	0.81
VC104275	2,618	3.15	0.80	3.16	0.83	0.01	0.77	71	100	0.77	3.15	0.79	0.00	0.81
VC104276	2,720	3.25	0.80	3.25	0.85	0.01	0.77	70	100	0.77	3.24	0.81	−0.01	0.81
VC104278	4,095	3.32	0.80	3.30	0.85	−0.02	0.75	67	99	0.75	3.30	0.80	−0.02	0.79
VC155074	993	3.19	0.80	3.11	0.88	−0.09	0.75	66	100	0.76	3.12	0.85	−0.08	0.81
VC219551	2,044	3.30	0.76	3.29	0.84	−0.02	0.79	74	100	0.79	3.28	0.79	−0.02	0.82
VC515323	2,553	3.25	0.77	3.23	0.82	−0.03	0.76	70	100	0.76	3.23	0.78	−0.02	0.81
VC787322	1,853	3.14	0.79	3.15	0.85	0.01	0.79	72	100	0.79	3.14	0.81	0.00	0.83
VC929101	1,934	3.39	0.78	3.37	0.83	−0.02	0.76	69	100	0.76	3.37	0.80	−0.02	0.80

Table G7: Continued

		Human 1 by e-rater (rounded to integers)									Human 1 by e-rater (unrounded)			
		Human 1		e-rater		Statistic					e-rater		Statistic	
Prompt	*N*	*M*	*SD*	*M*	*SD*	Std diff	Wtd kappa	% agree	% adj agree	*r*	*M*	*SD*	Std diff	*r*
VC929114	2,458	3.38	0.77	3.36	0.82	−0.02	0.76	71	100	0.76	3.36	0.78	−0.02	0.81
VC929139	2,346	3.24	0.78	3.24	0.82	0.01	0.77	72	100	0.77	3.24	0.79	0.01	0.81
VC048048	3,536	3.28	0.77	3.27	0.84	−0.01	0.77	71	100	0.77	3.27	0.81	−0.02	0.81
VC048051	5,374	3.27	0.73	3.26	0.80	0.00	0.75	72	100	0.75	3.26	0.75	−0.01	0.79
VC048070	3,803	3.24	0.76	3.23	0.82	−0.02	0.75	69	100	0.75	3.24	0.77	−0.01	0.79
VC048077	5,094	3.12	0.78	3.14	0.81	0.03	0.76	70	100	0.76	3.14	0.77	0.02	0.80
VC048141	4,149	3.29	0.73	3.28	0.80	−0.01	0.75	71	100	0.75	3.28	0.75	−0.02	0.79
VC073157	4,605	3.30	0.77	3.31	0.81	0.01	0.76	71	100	0.76	3.30	0.77	0.00	0.80
VC073158	2,961	3.25	0.78	3.23	0.83	−0.01	0.77	71	100	0.77	3.24	0.79	−0.01	0.80
VC073169	7,645	3.12	0.80	3.11	0.85	−0.01	0.76	69	100	0.76	3.10	0.81	−0.02	0.80
VC073172	2,499	3.19	0.77	3.18	0.82	0.00	0.75	70	99	0.75	3.18	0.78	0.00	0.80
VC104280	3,378	3.39	0.75	3.38	0.79	−0.01	0.75	71	100	0.76	3.38	0.75	−0.02	0.81
VC104281	4,859	3.42	0.76	3.40	0.81	−0.03	0.75	70	100	0.75	3.40	0.76	−0.03	0.79
VC084819	6,588	3.34	0.76	3.34	0.81	0.00	0.74	69	100	0.74	3.34	0.77	0.00	0.78
VC084820	4,555	3.23	0.77	3.24	0.80	0.01	0.72	67	100	0.72	3.23	0.75	0.00	0.77
VC155042	8,352	3.26	0.76	3.27	0.81	0.01	0.73	68	100	0.73	3.26	0.77	0.00	0.78
VC219591	6,526	3.37	0.72	3.36	0.78	−0.01	0.75	72	100	0.75	3.36	0.73	−0.01	0.79
VC787354	6,048	3.29	0.77	3.28	0.79	−0.02	0.74	70	100	0.75	3.28	0.75	−0.02	0.80
VC084798	6,213	3.13	0.75	3.13	0.80	0.00	0.75	71	100	0.75	3.12	0.76	−0.01	0.80
VC084799	5,885	2.95	0.78	2.96	0.81	0.01	0.74	68	100	0.74	2.96	0.77	0.01	0.78
VC084804	5,371	3.21	0.74	3.22	0.78	0.00	0.72	69	100	0.72	3.20	0.74	−0.01	0.77
VC104286	7,865	3.18	0.77	3.18	0.80	0.01	0.75	70	100	0.75	3.18	0.75	0.00	0.79
VE096407	7,957	3.03	0.76	2.99	0.79	−0.05	0.73	69	100	0.74	2.98	0.74	−0.05	0.78
VC084555	4,175	3.06	0.74	3.03	0.81	−0.04	0.71	66	99	0.71	3.02	0.78	−0.06	0.76
VC084754	6,319	3.02	0.73	2.99	0.75	−0.04	0.71	69	100	0.71	2.98	0.71	−0.06	0.77
VC104284	6,034	3.06	0.69	3.05	0.73	−0.01	**0.67**[a]	68	100	**0.67**[a]	3.05	0.67	−0.01	0.73
VC155075	2,204	3.07	0.74	3.06	0.79	−0.01	0.73	70	100	0.73	3.06	0.75	−0.02	0.77
VC155078	4,367	3.18	0.70	3.18	0.75	0.00	**0.68**[a]	68	100	**0.68**[a]	3.18	0.69	0.00	0.73
VC178591	2,389	3.16	0.72	3.13	0.77	−0.05	**0.69**[a]	66	100	**0.69**[a]	3.13	0.71	−0.05	0.74
VC178595	4,819	3.26	0.70	3.23	0.78	−0.04	**0.67**[a]	65	100	**0.67**[a]	3.23	0.74	−0.04	0.72
VC178602	4,294	3.11	0.77	3.11	0.80	0.00	0.73	67	100	0.73	3.11	0.75	0.00	0.78
VC515311	6,006	3.08	0.70	3.08	0.75	0.00	**0.69**[a]	69	100	**0.69**[a]	3.08	0.69	−0.01	0.75
VC787323	4,327	3.10	0.72	3.09	0.78	−0.01	0.71	68	100	0.71	3.08	0.73	−0.02	0.75
VC787333	1,866	3.12	0.76	3.14	0.80	0.02	0.74	70	100	0.74	3.13	0.75	0.01	0.78
VC084809	3,686	2.99	0.76	2.97	0.84	−0.02	0.74	69	99	0.75	2.96	0.81	−0.03	0.79
VC104290	4,689	3.19	0.78	3.17	0.85	−0.02	0.76	69	100	0.76	3.17	0.81	−0.03	0.80
VC104293	2,971	3.27	0.78	3.28	0.84	0.02	0.78	71	100	0.78	3.28	0.80	0.01	0.82
VC104297	2,213	3.00	0.80	2.99	0.83	−0.02	0.73	66	99	0.74	2.98	0.80	−0.03	0.79
VC104300	4,413	3.24	0.76	3.27	0.80	0.04	0.76	71	100	0.76	3.26	0.76	0.03	0.80
VC104302	4,727	3.03	0.78	3.01	0.81	−0.02	0.74	68	100	0.74	3.01	0.77	−0.03	0.79
VC155043	2,531	3.27	0.80	3.28	0.84	0.01	0.76	69	100	0.76	3.27	0.81	0.00	0.80
VC515320	1,957	3.15	0.79	3.15	0.81	−0.01	0.75	69	100	0.75	3.15	0.77	−0.01	0.80
VC787346	1,691	3.25	0.79	3.28	0.82	0.03	0.77	71	100	0.77	3.28	0.78	0.03	0.82
VE096379	2,398	3.11	0.76	3.08	0.80	−0.03	0.77	72	100	0.77	3.08	0.75	−0.03	0.81
VE096386	1,644	3.18	0.80	3.14	0.88	−0.04	0.78	69	100	0.78	3.13	0.84	−0.05	0.83
VE096411	2,316	3.17	0.77	3.14	0.83	−0.04	0.74	68	99	0.74	3.14	0.79	−0.04	0.78

Note. Std diff = standardized difference; wtd = weighted; adj = adjacent.

[a]Agreement indices that failed to meet the guideline thresholds (also indicated by boldface).

Table G8 Phase II Agreement of Human and e-rater Scores on Issue Prompts: PS-12 Model

		Human 1 by e-rater (rounded to integers)									Human 1 by e-rater (unrounded)			
		Human 1		e-rater		Statistic					e-rater		Statistic	
Prompt	*N*	*M*	*SD*	*M*	*SD*	Std diff	Wtd kappa	% agree	% adj agree	*r*	*M*	*SD*	Std diff	*r*
VC047733	6,030	3.17	0.77	3.17	0.82	0.01	0.77	71	100	0.77	3.17	0.78	0.01	0.81
VC047741	5,212	3.16	0.76	3.15	0.80	−0.02	0.73	68	100	0.74	3.15	0.75	−0.02	0.79
VC047750	2,693	3.27	0.73	3.31	0.78	0.06	0.73	70	100	0.73	3.30	0.73	0.04	0.79
VC047757	3,108	3.17	0.78	3.15	0.83	−0.03	0.77	70	100	0.77	3.15	0.80	−0.02	0.81
VC047770	4,331	3.18	0.81	3.17	0.83	−0.01	0.77	70	100	0.77	3.17	0.79	−0.01	0.82
VC047788	6,299	3.31	0.77	3.32	0.82	0.02	0.75	70	100	0.75	3.31	0.77	0.01	0.79
VC047792	4,143	3.26	0.79	3.29	0.82	0.03	0.76	70	100	0.76	3.28	0.78	0.02	0.81
VC047799	5,201	3.14	0.78	3.12	0.83	−0.02	0.75	68	100	0.75	3.11	0.79	−0.03	0.80
VC047822	6,826	3.26	0.80	3.25	0.86	0.00	0.77	69	99	0.77	3.24	0.83	−0.02	0.80
VC048013	3,077	3.16	0.83	3.18	0.85	0.02	0.79	71	100	0.79	3.18	0.82	0.02	0.83
VC048019	3,196	3.10	0.79	3.11	0.84	0.01	0.76	69	99	0.76	3.11	0.80	0.02	0.80
VC048027	5,942	3.31	0.77	3.31	0.80	0.00	0.75	70	100	0.75	3.31	0.76	0.00	0.80
VC048031	4,405	3.33	0.79	3.33	0.86	0.00	0.77	69	100	0.77	3.33	0.82	0.00	0.81
VC073155	6,117	3.24	0.78	3.23	0.80	−0.01	0.76	71	100	0.76	3.22	0.76	−0.02	0.80
VC073160	4,729	3.22	0.78	3.24	0.83	0.02	0.76	70	100	0.76	3.24	0.79	0.01	0.81
VC073163	3,329	3.13	0.82	3.10	0.88	−0.03	0.78	69	100	0.78	3.11	0.84	−0.03	0.81
VC073164	3,105	3.24	0.76	3.25	0.80	0.01	0.76	71	100	0.76	3.24	0.76	0.00	0.80
VC073166	1,440	3.18	0.77	3.13	0.86	−0.06	0.76	69	100	0.76	3.13	0.81	−0.07	0.80
VC073168	2,982	3.20	0.80	3.18	0.86	−0.03	0.76	68	99	0.76	3.18	0.82	−0.03	0.80
VC073173	4,437	3.19	0.81	3.17	0.85	−0.02	0.77	69	100	0.77	3.16	0.81	−0.03	0.81
VC073175	1,704	3.10	0.79	3.05	0.84	−0.06	0.75	68	100	0.76	3.04	0.81	−0.08	0.80
VC073176	1,436	3.17	0.81	3.17	0.82	0.01	0.77	71	100	0.77	3.16	0.78	0.00	0.81
VC104275	2,618	3.15	0.80	3.16	0.83	0.01	0.77	71	100	0.77	3.15	0.79	0.00	0.81
VC104276	2,720	3.25	0.80	3.25	0.85	0.00	0.77	70	100	0.77	3.24	0.81	−0.01	0.81
VC104278	4,095	3.32	0.80	3.30	0.85	−0.02	0.74	67	99	0.75	3.30	0.80	−0.02	0.79
VC155074	993	3.19	0.80	3.11	0.88	−0.09	0.75	66	100	0.76	3.12	0.85	−0.08	0.81
VC219551	2,044	3.30	0.76	3.29	0.84	−0.02	0.79	74	100	0.79	3.28	0.79	−0.02	0.82
VC515323	2,553	3.25	0.77	3.23	0.81	−0.03	0.76	71	100	0.76	3.23	0.77	−0.02	0.81
VC787322	1,853	3.14	0.79	3.15	0.85	0.01	0.79	72	100	0.79	3.14	0.81	0.00	0.83
VC929101	1,934	3.39	0.78	3.37	0.83	−0.02	0.76	69	100	0.76	3.37	0.80	−0.02	0.80
VC929114	2,458	3.38	0.77	3.36	0.82	−0.02	0.76	70	100	0.76	3.36	0.78	−0.02	0.81
VC929139	2,346	3.24	0.78	3.24	0.82	0.01	0.77	72	100	0.77	3.24	0.79	0.01	0.81
VC048048	3,536	3.28	0.77	3.27	0.85	−0.01	0.77	71	100	0.77	3.26	0.81	−0.02	0.81
VC048051	5,374	3.27	0.73	3.26	0.80	−0.01	0.75	72	100	0.75	3.26	0.75	−0.01	0.79
VC048070	3,803	3.24	0.76	3.23	0.82	−0.02	0.75	69	100	0.75	3.24	0.77	−0.01	0.79
VC048077	5,094	3.12	0.78	3.14	0.81	0.03	0.76	70	100	0.76	3.14	0.77	0.02	0.80
VC048141	4,149	3.29	0.73	3.28	0.80	−0.01	0.75	71	100	0.75	3.28	0.75	−0.02	0.79
VC073157	4,605	3.30	0.77	3.31	0.81	0.01	0.76	71	100	0.76	3.30	0.77	0.00	0.80
VC073158	2,961	3.25	0.78	3.24	0.83	−0.01	0.77	71	100	0.77	3.24	0.79	−0.01	0.80
VC073169	7,645	3.12	0.80	3.11	0.85	−0.01	0.76	69	100	0.76	3.10	0.81	−0.02	0.80
VC073172	2,499	3.19	0.77	3.18	0.82	0.00	0.75	70	99	0.75	3.18	0.78	0.00	0.80
VC104280	3,378	3.39	0.75	3.38	0.79	−0.01	0.76	71	100	0.76	3.38	0.75	−0.02	0.81
VC104281	4,859	3.42	0.76	3.40	0.81	−0.03	0.75	70	100	0.75	3.40	0.76	−0.03	0.79
VC084819	6,588	3.34	0.76	3.34	0.81	0.00	0.73	69	100	0.74	3.34	0.76	0.00	0.78
VC084820	4,555	3.23	0.77	3.24	0.80	0.01	0.72	67	100	0.72	3.23	0.75	0.00	0.77
VC155042	8,352	3.26	0.76	3.27	0.81	0.01	0.73	68	100	0.74	3.26	0.77	0.00	0.78
VC219591	6,526	3.37	0.72	3.36	0.78	−0.01	0.75	72	100	0.75	3.36	0.73	−0.01	0.79
VC787354	6,048	3.29	0.77	3.28	0.79	−0.02	0.74	70	100	0.74	3.28	0.75	−0.02	0.80
VC084798	6,213	3.13	0.75	3.13	0.80	0.00	0.75	71	100	0.75	3.12	0.76	−0.01	0.80
VC084799	5,885	2.95	0.78	2.96	0.81	0.01	0.74	68	100	0.74	2.96	0.77	0.01	0.78
VC084804	5,371	3.21	0.74	3.22	0.78	0.00	0.72	69	100	0.72	3.20	0.74	−0.01	0.77
VC104286	7,865	3.18	0.77	3.18	0.80	0.00	0.75	70	100	0.75	3.18	0.75	0.00	0.79
VE096407	7,957	3.03	0.76	2.99	0.79	−0.05	0.73	69	100	0.74	2.99	0.74	−0.05	0.78
VC084555	4,175	3.06	0.74	3.03	0.81	−0.04	0.71	66	99	0.71	3.02	0.78	−0.06	0.76
VC084754	6,319	3.02	0.73	2.98	0.76	−0.05	0.71	69	100	0.72	2.98	0.70	−0.06	0.77
VC104284	6,034	3.06	0.69	3.05	0.73	−0.01	**0.67**[a]	68	100	**0.67**[a]	3.05	0.67	−0.01	0.73
VC155075	2,204	3.07	0.74	3.06	0.79	−0.02	0.72	69	100	0.73	3.06	0.74	−0.02	0.77
VC155078	4,367	3.18	0.70	3.18	0.75	0.00	**0.68**[a]	68	100	**0.68**[a]	3.18	0.69	0.00	0.73
VC178591	2,389	3.16	0.72	3.13	0.77	−0.04	**0.69**[a]	66	100	**0.69**[a]	3.13	0.71	−0.04	0.74
VC178595	4,819	3.26	0.70	3.24	0.78	−0.03	**0.67**[a]	65	100	**0.67**[a]	3.23	0.74	−0.04	0.72

Table G8: Continued

		Human 1 by e-rater (rounded to integers)									Human 1 by e-rater (unrounded)			
		Human 1		e-rater		Statistic					e-rater		Statistic	
Prompt	*N*	*M*	*SD*	*M*	*SD*	Std diff	Wtd kappa	% agree	% adj agree	*r*	*M*	*SD*	Std diff	*r*
VC178602	4,294	3.11	0.77	3.11	0.80	0.00	0.73	67	100	0.73	3.11	0.75	0.00	0.78
VC515311	6,006	3.08	0.70	3.08	0.75	0.00	**0.69**[a]	69	100	**0.70**[a]	3.08	0.69	−0.01	0.75
VC787323	4,327	3.10	0.72	3.09	0.78	−0.01	0.71	68	100	0.71	3.08	0.73	−0.02	0.75
VC787333	1,866	3.12	0.76	3.13	0.79	0.01	0.74	70	100	0.74	3.13	0.75	0.01	0.78
VC084809	3,686	2.99	0.76	2.97	0.84	−0.02	0.74	69	99	0.75	2.96	0.81	−0.03	0.79
VC104290	4,689	3.19	0.78	3.17	0.85	−0.02	0.76	69	100	0.76	3.17	0.81	−0.03	0.80
VC104293	2,971	3.27	0.78	3.28	0.84	0.02	0.78	71	100	0.78	3.28	0.80	0.01	0.82
VC104297	2,213	3.00	0.80	2.99	0.83	−0.02	0.73	66	99	0.74	2.98	0.80	−0.03	0.79
VC104300	4,413	3.24	0.76	3.28	0.80	0.04	0.76	71	100	0.76	3.26	0.76	0.03	0.80
VC104302	4,727	3.03	0.78	3.01	0.81	−0.02	0.74	68	100	0.74	3.01	0.77	−0.03	0.79
VC155043	2,531	3.27	0.80	3.28	0.84	0.01	0.76	69	100	0.76	3.27	0.81	0.00	0.80
VC515320	1,957	3.15	0.79	3.15	0.81	−0.01	0.75	69	100	0.75	3.15	0.77	−0.01	0.80
VC787346	1,691	3.25	0.79	3.28	0.82	0.03	0.77	71	100	0.77	3.28	0.78	0.03	0.82
VE096379	2,398	3.11	0.76	3.08	0.80	−0.03	0.77	72	100	0.77	3.08	0.75	−0.03	0.81
VE096386	1,644	3.18	0.80	3.14	0.88	−0.04	0.78	69	100	0.78	3.14	0.84	−0.05	0.83
VE096411	2,316	3.17	0.77	3.14	0.83	−0.04	0.74	68	99	0.74	3.14	0.79	−0.04	0.78

Note. Std diff = standardized difference; wtd = weighted; adj = adjacent.

[a]Agreement indices that failed to meet the guideline thresholds (also indicated by boldface).

Appendix H

Human–Human and Human–e-rater Agreement for Argument and Issue Prompts at Phase III

Table H1 Phase III Agreement of Human Scores on Argument Prompts: G-10 Model

		Human 1 by Human 2									Human 1 by e-rater (rounded to integers)					Human 1 by e-rater (unrounded)			
		Human 1		Human 2		Statistic					e-rater		Statistic			e-rater		Statistic	
Prompt	*N*	*M*	*SD*	*M*	*SD*	Std diff	Wtd kappa	% agree	% adj agree	*r*	*M*	*SD*	Wtd kappa	% agree	% adj agree	*M*	*SD*	Std diff	*r*
VC048186	668	3.47	0.91	3.47	0.95	−0.01	0.77	62	99	0.77	3.20	0.98	0.75	60	98	3.20	0.96	**−0.29**[a]	0.81
VC048263	561	3.28	0.95	3.25	0.96	−0.03	0.80	68	99	0.80	3.51	1.02	0.75	58	97	3.48	0.99	**0.20**[a]	0.79
VC048268	385	3.61	0.97	3.53	0.98	−0.08	0.73	60	97	0.73	3.61	0.95	0.76	62	98	3.59	0.91	−0.02	0.80
VC048328	1,076	3.12	0.87	3.06	0.85	−0.06	0.73	64	99	0.73	3.22	0.95	0.72	59	98	3.21	0.90	0.11	0.76
VC048389	687	3.39	0.94	3.36	0.93	−0.04	0.77	65	98	0.77	3.49	0.93	0.75	64	97	3.49	0.91	0.11	0.79
VC048408	365	3.42	0.98	3.40	1.02	−0.01	0.80	67	98	0.80	3.67	0.99	0.73	58	96	3.67	0.94	**0.27**[a]	0.78
VC069377	606	3.43	0.87	3.43	0.91	0.01	0.75	64	99	0.75	3.32	0.93	0.74	61	99	3.33	0.91	−0.11	0.78
VC069384	757	3.52	0.91	3.57	0.92	0.06	0.77	63	99	0.77	3.60	0.95	0.78	65	99	3.59	0.92	0.08	0.82
VC069394	2,207	3.12	0.75	3.10	0.74	−0.02	**0.64**[a]	63	99	**0.64**[a]	3.50	0.78	**0.56**[a]	51	97	3.50	0.72	**0.52**[a]	**0.67**[a]
VC069396	2,463	3.10	0.81	3.06	0.81	−0.05	**0.65**[a]	61	98	**0.65**[a]	3.30	0.85	**0.64**[a]	56	98	3.29	0.80	**0.24**[a]	**0.70**[a]
VC084832	794	3.54	1.00	3.55	1.02	0.00	0.78	62	98	0.78	3.63	0.96	0.75	61	97	3.63	0.94	0.09	0.79
VC084840	2,064	3.07	0.73	3.10	0.73	0.04	**0.66**[a]	66	99	**0.66**[a]	3.10	0.85	**0.64**[a]	60	99	3.09	0.81	0.02	**0.69**[a]
VC084849	1,173	3.48	0.97	3.48	0.96	0.00	0.76	61	98	0.76	3.54	0.98	0.77	62	98	3.53	0.95	0.05	0.81
VC086531	413	3.24	0.91	3.27	0.92	0.03	0.75	62	99	0.75	3.20	0.97	0.71	56	97	3.20	0.93	−0.04	0.74
VC101052	359	3.23	0.99	3.21	0.96	−0.03	0.78	63	98	0.78	3.32	1.01	0.76	60	98	3.33	1.01	0.09	0.78
VC101056	332	3.38	0.97	3.42	0.98	0.04	0.77	60	99	0.77	3.52	0.98	0.74	59	98	3.53	0.94	**0.15**[a]	0.78
VC101542	392	3.31	0.99	3.30	0.96	0.00	0.81	67	99	0.81	3.24	0.95	0.77	61	99	3.23	0.93	−0.08	0.82
VC140314	804	3.35	0.87	3.38	0.90	0.03	0.74	63	99	0.74	3.04	0.95	**0.68**[a]	52	97	3.03	0.94	**−0.36**[a]	0.76
VC249418	235	3.31	0.94	3.31	0.92	0.00	0.76	64	99	0.76	3.16	0.96	0.73	60	97	3.19	0.96	−0.13	0.80
VC251464	2,338	3.14	0.80	3.14	0.80	0.01	0.71	66	99	0.71	3.33	0.87	**0.67**[a]	58	99	3.33	0.84	**0.23**[a]	0.73
VC251477	454	3.42	0.98	3.33	0.98	−0.10	0.73	57	97	0.73	3.45	1.04	0.76	57	98	3.43	1.00	0.01	0.79
VC251575	838	3.39	0.95	3.31	0.96	−0.08	0.78	65	99	0.79	3.35	0.93	0.75	61	98	3.35	0.91	−0.04	0.77
VC251577	389	3.51	0.93	3.57	1.01	0.06	0.76	64	97	0.76	3.63	0.96	0.74	60	98	3.62	0.93	0.12	0.79

Table H1: Continued

		Human 1 by Human 2									Human 1 by e-rater (rounded to integers)					Human 1 by e-rater (unrounded)			
		Human 1		Human 2		Statistic					e-rater		Statistic			e-rater		Statistic	
Prompt	*N*	*M*	*SD*	*M*	*SD*	Std diff	Wtd kappa	% agree	% adj agree	*r*	*M*	*SD*	Wtd kappa	% agree	% adj agree	*M*	*SD*	Std diff	*r*
VC390618	345	3.14	0.94	3.16	0.94	0.02	**0.69**[a]	58	97	**0.69**[a]	3.05	0.99	0.72	55	97	3.02	0.95	−0.12	0.75
VC048246	583	3.58	0.93	3.60	0.96	0.02	0.76	62	99	0.76	3.67	0.93	0.75	60	98	3.66	0.91	0.08	0.77
VC048273	515	3.52	1.04	3.54	1.06	0.01	0.79	64	97	0.79	3.58	0.95	0.73	56	97	3.58	0.93	0.06	0.77
VC048352	712	3.31	0.92	3.38	0.93	0.07	0.76	63	98	0.76	3.28	0.96	0.77	63	99	3.28	0.94	−0.03	0.81
VC048390	551	3.30	0.97	3.29	0.95	−0.01	0.77	64	98	0.77	3.31	0.94	0.76	61	99	3.30	0.90	−0.01	0.81
VC048411	652	3.32	0.96	3.31	1.00	−0.01	0.80	64	99	0.80	3.24	0.97	0.75	59	98	3.23	0.96	−0.09	0.80
VC069378	470	3.56	0.96	3.57	0.98	0.01	0.78	64	98	0.78	3.51	0.95	0.73	59	98	3.51	0.90	−0.05	0.76
VC069382	709	3.37	1.01	3.34	1.01	−0.04	0.80	65	98	0.80	3.53	0.95	0.73	59	96	3.54	0.93	**0.17**[a]	0.78
VC069400	405	3.27	0.95	3.26	0.93	−0.01	0.76	62	99	0.76	3.03	1.00	0.72	53	98	3.05	0.98	**−0.22**[a]	0.77
VC084829	804	3.34	1.04	3.31	1.02	−0.02	0.76	56	98	0.76	3.26	0.96	0.74	57	97	3.25	0.94	−0.08	0.77
VC084835	1,828	3.15	0.74	3.12	0.75	−0.04	**0.65**[a]	64	99	**0.65**[a]	2.86	0.84	**0.59**[a]	53	97	2.85	0.79	**−0.39**[a]	**0.67**[a]
VC084851	877	3.32	0.93	3.32	0.90	0.00	0.75	65	98	0.75	3.18	0.93	0.76	62	99	3.17	0.91	**−0.17**[a]	0.80
VC084853	697	3.29	0.98	3.38	1.01	0.09	0.78	63	98	0.78	3.25	1.02	0.78	62	98	3.24	1.00	−0.04	0.81
VC086526	618	3.16	0.94	3.20	0.95	0.04	0.74	61	97	0.74	3.13	0.97	0.71	56	97	3.13	0.96	−0.03	0.74
VC093524	708	3.33	0.94	3.32	0.94	−0.02	0.76	63	98	0.76	3.51	0.96	**0.70**[a]	55	97	3.48	0.92	**0.16**[a]	0.75
VC093532	606	3.46	0.96	3.45	0.96	−0.01	0.77	65	97	0.77	3.54	0.91	0.73	61	97	3.53	0.88	0.08	0.76
VC101021	353	3.44	0.96	3.41	0.94	−0.03	0.79	65	99	0.79	3.33	0.95	0.76	59	99	3.33	0.95	−0.12	0.80
VC101037	916	3.34	0.91	3.34	0.94	−0.01	0.74	65	97	0.74	3.26	0.94	0.70	56	98	3.26	0.89	−0.10	0.74
VC101537	663	3.19	0.86	3.20	0.81	0.01	**0.68**[a]	61	98	**0.68**[a]	3.33	0.87	**0.65**[a]	57	97	3.33	0.86	**0.17**[a]	**0.70**[a]
VC101541	1,177	3.52	0.91	3.50	0.89	−0.02	0.74	62	99	0.74	2.99	0.90	**0.61**[a]	41	94	2.97	0.86	**−0.61**[a]	0.75
VC207455	2,331	3.25	0.83	3.22	0.80	−0.04	**0.67**[a]	63	98	**0.67**[a]	3.24	0.85	**0.65**[a]	57	98	3.23	0.81	−0.02	**0.70**[a]
VC207640	787	3.47	1.01	3.41	1.02	−0.05	0.77	63	97	0.78	3.71	0.92	0.71	56	96	3.71	0.89	**0.26**[a]	0.77
VC209497	868	3.49	0.95	3.49	0.95	0.00	0.76	62	98	0.76	3.57	0.95	0.76	63	98	3.57	0.90	0.09	0.80
VC248469	941	3.35	0.97	3.39	0.96	0.04	0.74	61	97	0.74	3.35	0.94	0.73	59	97	3.35	0.92	0.01	0.76
VC250603	977	3.60	0.99	3.59	0.98	−0.02	0.77	60	98	0.77	3.56	0.98	0.74	57	98	3.56	0.96	−0.05	0.79
VC251468	994	3.43	0.93	3.45	0.92	0.02	0.77	65	99	0.77	3.60	0.95	0.74	62	97	3.61	0.92	**0.20**[a]	0.79
VC251474	606	3.38	1.02	3.41	1.02	0.02	0.80	63	98	0.80	3.48	1.05	0.78	58	99	3.47	1.06	0.08	0.81
VC251573	1,403	3.39	1.01	3.36	1.00	−0.03	0.76	61	97	0.76	3.40	0.95	0.72	57	97	3.39	0.93	−0.01	0.75
VC390606	1,083	3.34	1.01	3.34	1.02	−0.01	0.82	65	99	0.82	3.47	1.00	0.76	60	98	3.46	0.98	0.12	0.81
VC462771	436	3.29	1.00	3.30	0.98	0.01	0.80	63	99	0.80	3.25	0.94	0.75	58	98	3.23	0.92	−0.06	0.77
VC101540	498	3.45	0.90	3.48	0.90	0.04	0.73	62	98	0.73	3.53	0.91	0.74	63	98	3.50	0.90	0.06	0.78
VC250595	1,011	3.49	0.95	3.52	0.98	0.03	0.77	63	98	0.77	3.43	0.96	0.76	60	99	3.44	0.92	−0.05	0.80
VC101018	584	3.30	0.97	3.36	0.92	0.06	0.74	61	98	0.75	3.06	0.97	0.72	54	97	3.04	0.97	**−0.27**[a]	0.78
VC251475	529	3.50	0.93	3.48	0.83	−0.02	0.71	59	99	0.71	3.51	0.95	0.77	63	99	3.51	0.91	0.01	0.79
VC390614	235	3.42	0.90	3.42	0.92	0.00	**0.66**[a]	52	97	**0.66**[a]	3.48	0.94	0.71	57	98	3.48	0.93	0.07	0.75
VC101050	293	3.52	1.00	3.48	0.98	−0.04	0.82	69	99	0.82	3.54	0.97	0.78	64	98	3.51	0.91	−0.01	0.80
VC177590	490	3.38	0.95	3.37	0.98	−0.01	0.73	57	98	0.73	3.47	0.97	0.76	61	99	3.47	0.96	0.10	0.78
VC248460	823	3.40	0.92	3.37	0.94	−0.04	0.77	66	98	0.77	3.31	0.94	0.76	63	98	3.30	0.91	−0.12	0.79
VC248479	308	3.45	0.92	3.37	0.93	−0.09	0.72	58	98	0.72	3.29	0.97	0.73	54	99	3.23	0.93	**−0.24**[a]	0.77
VC248488	798	3.48	0.95	3.47	0.95	−0.01	0.72	59	97	0.72	3.34	0.92	**0.69**[a]	56	97	3.35	0.87	−0.15	0.74
VC250589	548	3.49	0.96	3.44	0.98	−0.05	0.78	66	98	0.78	3.25	0.98	0.72	55	98	3.24	0.94	**−0.26**[a]	0.79
VC251576	362	3.44	1.01	3.41	1.00	−0.02	0.80	65	98	0.80	3.33	1.02	0.74	53	98	3.32	0.97	−0.12	0.77
VC390640	587	3.29	0.96	3.27	0.92	−0.02	0.71	58	97	0.71	3.27	0.93	**0.65**[a]	54	95	3.25	0.90	−0.05	**0.70**[a]
VC462770	307	3.41	0.99	3.43	0.96	0.02	0.76	65	98	0.76	3.32	0.95	0.70	51	98	3.31	0.93	−0.10	0.73
VE096305	545	3.52	0.89	3.54	0.91	0.01	0.78	66	99	0.78	3.56	0.90	0.74	63	99	3.57	0.87	0.05	0.79
VC069380	601	3.41	0.90	3.37	0.92	−0.05	0.75	63	99	0.75	3.53	0.93	0.74	60	98	3.54	0.90	0.14	0.77
VC084843	636	3.45	0.96	3.46	0.91	0.02	0.74	60	98	0.74	3.45	0.97	0.78	63	99	3.46	0.95	0.01	0.80
VC084846	246	3.31	0.89	3.30	0.92	−0.01	0.75	62	99	0.75	3.36	0.89	0.77	66	99	3.36	0.87	0.06	0.82
VC101016	340	3.13	0.97	3.05	0.98	−0.08	0.77	60	99	0.77	2.78	1.01	0.72	48	98	2.77	1.00	**−0.37**[a]	0.81
VC101539	279	3.55	0.95	3.58	1.01	0.04	0.76	59	98	0.76	3.65	0.91	0.77	65	99	3.66	0.87	0.12	0.80
VC140094	480	3.38	0.93	3.33	0.94	−0.06	0.77	65	99	0.77	3.29	0.97	0.76	59	100	3.26	0.94	−0.12	0.79
VC209485	832	3.33	0.94	3.35	0.92	0.02	0.76	62	99	0.76	3.21	0.95	0.76	58	99	3.22	0.93	−0.12	0.79
VC248473	523	3.35	0.93	3.33	0.94	−0.02	0.79	66	99	0.79	3.22	1.00	0.78	62	99	3.23	0.96	−0.12	0.81

Note. Std diff = standardized difference; wtd = weighted; adj = adjacent.

[a] Agreement indices that failed to meet the guideline thresholds (also indicated by boldface).

Table H2 Phase III Agreement of Human Scores on Issue Prompts: G-10 Model

		Human 1 by Human 2									Human 1 by e-rater (rounded to integers)					Human 1 by e-rater (unrounded)			
		Human 1		Human 2		Statistic					e-rater		Statistic			e-rater		Statistic	
Prompt	*N*	*M*	*SD*	*M*	*SD*	Std diff	Wtd kappa	% agree	% adj agree	*r*	*M*	*SD*	Wtd kappa	% agree	% adj agree	*M*	*SD*	Std diff	*r*
VC047733	737	3.34	0.87	3.40	0.86	0.07	0.74	66	99	0.74	3.32	0.87	0.77	67	99	3.32	0.82	−0.03	0.81
VC047741	767	3.32	0.93	3.34	0.89	0.02	0.77	65	99	0.77	3.29	0.88	0.79	69	99	3.27	0.84	−0.06	0.83
VC047750	504	3.32	0.87	3.29	0.87	−0.03	0.78	68	99	0.78	3.30	0.89	0.79	68	100	3.30	0.85	−0.03	0.84
VC047757	454	3.34	0.92	3.28	0.90	−0.07	0.77	66	99	0.78	3.21	0.95	0.81	68	100	3.22	0.93	−0.13	0.85
VC047770	313	3.27	1.00	3.29	0.91	0.02	0.78	66	98	0.78	3.18	0.94	0.77	62	98	3.19	0.89	−0.08	0.82
VC047788	607	3.43	0.86	3.40	0.84	−0.03	0.71	61	99	0.71	3.43	0.85	0.75	65	100	3.42	0.81	−0.01	0.79
VC047792	357	3.37	0.96	3.40	0.94	0.03	0.79	66	99	0.79	3.30	0.98	0.81	66	99	3.31	0.94	−0.07	0.85
VC047799	394	3.16	0.89	3.24	0.88	0.09	0.73	66	98	0.74	3.10	0.94	0.80	70	99	3.10	0.94	−0.08	0.83
VC047822	670	3.25	0.95	3.30	0.97	0.05	0.78	65	99	0.78	3.23	0.93	0.81	68	100	3.22	0.92	−0.03	0.84
VC048013	486	3.49	0.93	3.42	0.90	−0.07	0.78	66	99	0.78	3.37	0.88	0.80	69	99	3.36	0.87	−0.14	0.85
VC048019	473	3.34	0.93	3.35	0.92	0.01	0.81	70	99	0.81	3.30	0.96	0.83	71	100	3.32	0.93	−0.03	0.85
VC048027	665	3.55	0.85	3.58	0.81	0.03	0.73	65	99	0.73	3.62	0.81	0.77	68	100	3.61	0.77	0.07	0.81
VC048031	746	3.59	0.86	3.60	0.90	0.01	0.77	67	99	0.77	3.63	0.87	0.79	69	100	3.62	0.81	0.03	0.82
VC073155	669	3.40	0.83	3.38	0.85	−0.03	0.77	69	100	0.77	3.38	0.88	0.76	66	100	3.36	0.85	−0.06	0.81
VC073160	662	3.26	0.93	3.31	0.91	0.06	0.82	70	100	0.82	3.33	0.88	0.80	70	99	3.33	0.85	0.08	0.84
VC073163	1,097	3.41	0.83	3.38	0.86	−0.04	0.76	68	99	0.76	3.34	0.86	0.76	67	99	3.34	0.83	−0.08	0.81
VC073164	406	3.44	0.90	3.38	0.86	−0.07	0.73	61	99	0.73	3.44	0.88	0.78	67	100	3.46	0.86	0.02	0.83
VC073166	1,169	3.22	0.84	3.24	0.87	0.02	0.70	61	99	0.70	3.30	0.88	0.74	65	99	3.30	0.83	0.09	0.79
VC073168	394	3.24	0.90	3.23	0.86	−0.01	0.76	66	99	0.76	3.29	0.89	0.81	72	99	3.29	0.85	0.06	0.83
VC073173	362	3.50	0.86	3.48	0.90	−0.02	0.77	69	99	0.77	3.39	0.91	0.81	72	99	3.39	0.91	−0.13	0.84
VC073175	749	3.32	0.84	3.28	0.81	−0.05	0.75	68	99	0.76	3.24	0.84	0.77	68	100	3.24	0.82	−0.10	0.82
VC073176	815	3.50	0.92	3.47	0.90	−0.04	0.79	68	99	0.79	3.42	0.90	0.81	69	100	3.41	0.85	−0.10	0.85
VC104275	796	3.33	0.94	3.36	0.91	0.03	0.75	63	98	0.75	3.37	0.89	0.77	65	99	3.37	0.85	0.04	0.82
VC104276	585	3.36	0.92	3.35	0.91	−0.02	0.78	67	99	0.78	3.33	0.90	0.79	67	99	3.31	0.88	−0.06	0.83
VC104278	361	3.28	0.87	3.26	0.95	−0.02	0.75	60	100	0.75	3.22	0.89	0.77	65	99	3.21	0.87	−0.08	0.83
VC155074	1,446	3.34	0.90	3.35	0.91	0.01	0.72	60	98	0.72	3.21	0.89	0.76	64	99	3.21	0.85	−0.15	0.81
VC219551	584	3.63	0.92	3.60	0.91	−0.03	0.80	70	99	0.80	3.63	0.91	0.81	69	100	3.62	0.88	0.00	0.86
VC515323	698	3.33	0.92	3.33	0.91	0.00	0.78	66	99	0.78	3.30	0.94	0.82	70	100	3.28	0.93	−0.06	0.86
VC787322	1,400	3.40	0.91	3.38	0.89	−0.02	0.75	63	99	0.75	3.29	0.89	0.78	66	100	3.28	0.86	−0.13	0.83
VC929101	682	3.16	0.85	3.21	0.82	0.05	0.75	66	99	0.75	3.33	0.84	0.74	63.2	99.7	3.32	0.78	**0.20**[a]	0.79

Table H2: Continued

		Human 1 by Human 2									Human 1 by e-rater (rounded to integers)					Human 1 by e-rater (unrounded)			
		Human 1		Human 2		Statistic					e-rater		Statistic			e-rater		Statistic	
Prompt	*N*	*M*	*SD*	*M*	*SD*	Std diff	Wtd kappa	% agree	% adj agree	*r*	*M*	*SD*	Wtd kappa	% agree	% adj agree	*M*	*SD*	Std diff	*r*
VC929114	606	3.59	0.83	3.55	0.79	−0.05	0.71	64	99	0.71	3.61	0.87	0.80	72.3	99.5	3.59	0.84	0.00	0.83
VC929139	1,231	3.44	0.86	3.49	0.84	0.05	0.75	65	99	0.75	3.32	0.86	0.76	65.8	99.5	3.31	0.84	**−0.15**[a]	0.82
VC048048	2,694	3.27	0.84	3.24	0.84	−0.04	0.72	65	99	0.72	3.40	0.83	0.72	62.4	99.2	3.41	0.78	**0.16**[a]	0.77
VC048051	2,423	3.01	0.77	3.00	0.76	−0.01	**0.68**[a]	66	99	**0.68**[a]	3.12	0.81	**0.66**[a]	61.0	98.7	3.10	0.77	0.12	0.71
VC048070	611	3.33	0.88	3.35	0.88	0.01	0.78	67	100	0.78	3.40	0.91	0.79	67.9	99.5	3.39	0.88	0.06	0.84
VC048077	2,086	3.03	0.79	3.03	0.81	0.00	**0.70**[a]	65	99	**0.70**[a]	3.11	0.82	**0.70**[a]	63.4	99.0	3.11	0.79	0.11	0.74
VC048141	920	3.53	0.85	3.57	0.85	0.04	0.74	67	99	0.74	3.63	0.85	0.77	67.9	99.3	3.62	0.81	0.11	0.81
VC073157	237	3.57	0.87	3.57	0.86	0.00	0.78	68	100	0.78	3.48	0.89	0.79	67.1	100.0	3.50	0.83	−0.08	0.83
VC073158	228	3.66	0.85	3.61	0.90	−0.06	0.76	67	99	0.76	3.69	0.87	0.73	62.7	99.1	3.70	0.78	0.04	0.79
VC073169	2,335	2.98	0.75	2.95	0.78	−0.04	0.70	67	99	0.70	3.15	0.84	**0.68**[a]	61.8	98.9	3.15	0.79	**0.21**[a]	0.74
VC073172	434	3.36	0.83	3.37	0.86	0.01	0.77	68	100	0.77	3.38	0.89	0.78	69.4	99.3	3.36	0.84	0.00	0.83
VC104280	586	3.52	0.84	3.55	0.86	0.03	0.77	67	100	0.77	3.65	0.90	0.76	64.7	99.7	3.63	0.86	0.13	0.82
VC104281	241	3.54	0.89	3.53	0.94	0.00	0.78	66	99	0.78	3.61	0.91	0.82	72.6	99.2	3.60	0.86	0.07	0.84
VC084819	1,995	3.03	0.80	3.02	0.80	−0.01	**0.67**[a]	62	99	**0.67**[a]	3.27	0.79	**0.64**[a]	57.8	98.0	3.26	0.74	**0.30**[a]	0.71
VC084820	1,807	2.93	0.76	2.89	0.76	−0.05	**0.65**[a]	65	98	**0.65**[a]	3.08	0.79	**0.66**[a]	60.3	99.2	3.08	0.73	**0.20**[a]	0.72
VC155042	470	3.42	0.89	3.43	0.91	0.01	0.77	65	99	0.77	3.46	0.92	0.79	65.7	100.0	3.44	0.89	0.02	0.83
VC219591	608	3.64	0.85	3.64	0.88	0.00	0.80	70	100	0.80	3.73	0.85	0.79	70.6	99.8	3.74	0.83	0.12	0.84
VC787354	662	3.41	0.88	3.38	0.90	−0.03	0.72	61	99	0.72	3.44	0.89	0.79	69.2	99.4	3.44	0.85	0.03	0.82
VC084798	465	3.30	0.87	3.33	0.88	0.03	0.76	65	99	0.76	3.32	0.94	0.81	70.1	99.8	3.32	0.89	0.01	0.84
VC084799	890	3.47	0.90	3.43	0.86	−0.04	0.75	63	99	0.75	3.34	0.89	0.78	65.2	99.7	3.34	0.86	**−0.15**[a]	0.83
VC084804	701	3.54	0.86	3.49	0.87	−0.06	0.75	66	99	0.75	3.56	0.92	0.80	70.8	99.1	3.57	0.88	0.03	0.84
VC104286	523	3.36	0.88	3.32	0.88	−0.04	0.72	61	99	0.72	3.38	0.91	0.76	64.6	99.2	3.35	0.88	−0.01	0.79
VE096407	774	3.22	0.88	3.30	0.86	0.09	0.74	67	99	0.74	3.24	0.88	0.74	64.0	98.8	3.25	0.85	0.03	0.77
VC084555	297	3.15	0.91	3.16	0.93	0.01	0.74	63	98	0.74	3.16	0.92	0.77	66.0	98.3	3.14	0.90	−0.01	0.79
VC084754	196	3.24	0.99	3.27	0.96	0.02	0.79	63	99	0.79	3.08	0.97	0.79	60.2	99.5	3.06	0.96	**−0.19**[a]	0.83
VC104284	846	3.30	0.81	3.28	0.82	−0.02	0.73	67	99	0.73	3.30	0.86	0.74	64.5	99.8	3.29	0.83	−0.01	0.78
VC155075	510	3.32	0.83	3.29	0.82	−0.04	0.72	63	100	0.72	3.25	0.90	0.73	62.9	98.8	3.25	0.85	−0.09	0.78
VC155078	974	3.47	0.81	3.51	0.79	0.05	**0.70**[a]	66	99	0.70	3.49	0.81	0.70	64.5	98.8	3.49	0.77	0.03	0.74
VC178591	1,049	3.54	0.85	3.46	0.85	−0.09	0.71	63	99	0.72	3.65	0.85	0.75	65.8	99.2	3.64	0.81	0.12	0.79
VC178595	466	3.48	0.84	3.50	0.84	0.02	0.72	66	99	0.72	3.70	0.88	0.70	60.5	98.1	3.70	0.85	**0.26**[a]	0.78
VC178602	653	3.39	0.88	3.35	0.86	−0.05	0.73	62	99	0.73	3.34	0.92	0.78	65.8	99.5	3.32	0.89	−0.08	0.82

Table H2: Continued

		Human 1 by Human 2									Human 1 by e-rater (rounded to integers)					Human 1 by e-rater (unrounded)			
		Human 1		Human 2		Statistic					e-rater		Statistic			e-rater		Statistic	
Prompt	*N*	*M*	*SD*	*M*	*SD*	Std diff	Wtd kappa	% agree	% adj agree	*r*	*M*	*SD*	Wtd kappa	% agree	% adj agree	*M*	*SD*	Std diff	*r*
VC515311	780	3.25	0.86	3.30	0.87	0.06	0.74	65	99	0.74	3.31	0.92	0.78	65.9	99.5	3.32	0.88	0.08	0.81
VC787323	939	3.34	0.91	3.28	0.92	−0.06	0.79	67	99	0.79	3.25	0.92	0.77	62.9	99.5	3.25	0.90	−0.10	0.83
VC787333	568	3.36	0.89	3.34	0.92	−0.02	0.74	62	99	0.74	3.49	0.91	0.76	65.0	98.6	3.50	0.87	**0.15**[a]	0.80
VC084809	234	3.35	0.84	3.29	0.86	−0.06	0.72	65	98	0.72	3.13	0.89	0.77	66.2	99.6	3.11	0.84	**−0.29**[a]	0.82
VC104290	653	3.41	0.84	3.37	0.87	−0.05	0.77	68	99	0.77	3.34	0.90	0.82	72.9	99.8	3.35	0.88	−0.07	0.84
VC104293	528	3.43	0.88	3.42	0.93	−0.01	0.76	67	98	0.76	3.43	0.93	0.79	67.6	99.4	3.41	0.89	−0.02	0.83
VC104297	1,348	3.21	0.88	3.22	0.88	0.01	0.76	66	99	0.76	3.18	0.92	0.79	68.0	99.5	3.18	0.90	−0.03	0.83
VC104300	810	3.38	0.93	3.35	0.94	−0.03	0.74	58	99	0.74	3.27	0.94	0.79	64.1	99.9	3.26	0.91	−0.13	0.84
VC104302	862	3.43	0.89	3.41	0.88	−0.02	0.78	69	99	0.78	3.32	0.90	0.80	69.4	99.5	3.33	0.88	−0.11	0.84
VC155043	296	3.32	0.92	3.24	0.93	−0.08	0.79	66	100	0.80	3.37	0.93	0.78	63.5	99.7	3.35	0.93	0.03	0.81
VC515320	900	3.40	0.87	3.40	0.88	−0.01	0.74	63	99	0.74	3.32	0.90	0.80	70.0	99.6	3.32	0.86	−0.09	0.83
VC787346	707	3.43	0.90	3.41	0.87	−0.02	0.74	63	99	0.74	3.45	0.88	0.79	68.3	99.4	3.45	0.86	0.02	0.83
VE096379	503	3.15	0.84	3.17	0.89	0.02	0.73	64	99	0.74	3.11	0.92	0.77	65.8	99.4	3.09	0.89	−0.07	0.82
VE096386	480	3.36	0.85	3.38	0.81	0.03	0.72	63	99	0.72	3.31	0.89	0.79	69.6	99.6	3.31	0.85	−0.06	0.82
VE096411	594	3.29	0.86	3.27	0.87	−0.03	0.78	70	99	0.78	3.29	0.86	0.79	69.5	99.8	3.30	0.83	0.00	0.84

Note. Std diff = standardized difference; wtd = weighted; adj = adjacent.

[a] Agreement indices that failed to meet the guideline thresholds (also indicated by boldface).

Appendix I

Subgroup Differences Phase III

Table I1 Phase III Subgroup Differences for Argument Prompts: G-10 Model

			Human 1 by Human 2							Human 1 by e-rater				
			Human 1		Human 2		Statistic			e-rater (rounded)		Statistic		
		N	*M*	*SD*	*M*	*SD*	Std diff	Wtd kappa	*r*	*M*	*SD*	Std diff (unrnd)	Wtd kappa	*r* (unrnd)
Asian	Yes	9,492	2.97	0.64	2.97	0.65	0.00	**0.51**[a]	**0.51**[a]	3.10	0.81	**0.19**[a]	**0.51**[a]	**0.57**[a]
Gender	Female	28,835	3.39	0.91	3.38	0.91	−0.01	0.74	0.74	3.41	0.92	0.02	0.71	0.74
	Male	23,727	3.26	0.93	3.25	0.94	−0.01	0.76	0.76	3.26	0.98	−0.01	0.73	0.76
Race	American Indian or Alaskan Native	182	3.40	0.78	3.46	0.93	0.07	0.71	0.72	3.55	0.80	**0.16**[a]	**0.62**[a]	**0.66**[a]
	Asian or Asian American	1,838	3.61	0.93	3.60	0.92	−0.01	0.75	0.75	3.64	0.90	0.02	0.72	0.75
	Black or African American	3,775	3.00	0.85	2.98	0.84	−0.02	0.72	0.73	2.95	0.94	−0.07	**0.69**[a]	0.72
	Mexican, Mexican American, or Chicano	858	3.34	0.89	3.32	0.87	−0.02	0.74	0.74	3.34	0.87	0.01	**0.67**[a]	0.72
	Puerto Rican	240	3.33	0.82	3.42	0.86	**0.10**[a]	**0.70**[a]	0.70	3.40	0.87	0.06	**0.66**[a]	0.72
	Other Hispanic, Latino, or Latin American	1,351	3.30	0.87	3.30	0.88	−0.01	0.71	0.71	3.35	0.89	0.03	**0.67**[a]	0.70
	White (non-Hispanic)	23,961	3.69	0.88	3.69	0.88	−0.01	0.71	0.71	3.71	0.81	0.02	**0.67**[a]	0.72
Country	India	2,994	2.82	0.85	2.80	0.86	−0.02	0.72	0.72	2.83	0.88	0.00	**0.69**[a]	0.73
	China	8,815	2.96	0.63	2.96	0.63	0.00	**0.49**[a]	**0.49**[a]	3.11	0.80	**0.21**[a]	**0.50**[a]	**0.56**[a]
	Canada	250	3.88	1.06	3.84	1.02	−0.03	0.79	0.79	3.73	1.01	**−0.13**[a]	0.75	0.79
	Korea	368	3.14	0.72	3.17	0.81	0.05	**0.61**[a]	**0.62**[a]	3.25	0.86	0.10	**0.64**[a]	**0.69**[a]
	Taiwan	246	2.73	0.71	2.71	0.64	−0.04	**0.65**[a]	**0.65**[a]	2.52	0.87	**−0.29**[a]	**0.68**[a]	0.74
	Hong Kong	63	3.40	0.66	3.40	0.81	0.00	**0.59**[a]	**0.60**[a]	3.49	0.88	0.09	**0.63**[a]	**0.66**[a]

Note. Std diff = standardized difference; wtd = weighted; adj = adjacent; unrnd = unrounded.
[a]Agreement indices that failed to meet the guideline thresholds (also indicated by boldface).

Table I2 Phase III Subgroup Differences for Issue Prompts: G-10 Model

			Human 1 by Human 2							Human 1 by e-rater				
			Human 1		Human 2		Statistic			e-rater (rounded)		Statistic		
		N	*M*	*SD*	*M*	*SD*	Std diff	Wtd kappa	*r*	*M*	*SD*	Std diff (unrnd)	Wtd kappa	*r* (unrnd)
Asian	Yes	9,722	2.82	0.64	2.81	0.64	−0.03	**0.52**[a]	**0.52**[a]	3.04	0.75	**0.32**[a]	**0.55**[a]	**0.63**[a]
Gender	Female	29,698	3.37	0.85	3.36	0.85	−0.01	0.74	0.74	3.40	0.86	0.03	0.76	0.80
	Male	24,581	3.22	0.89	3.21	0.89	−0.01	0.76	0.76	3.24	0.91	0.02	0.77	0.80
Race	American Indian or Alaskan Native	198	3.46	0.79	3.48	0.79	0.02	**0.67**[a]	**0.67**[a]	3.45	0.83	−0.03	0.73	0.78
	Asian or Asian American	1,882	3.58	0.87	3.57	0.85	−0.02	0.72	0.72	3.62	0.84	0.04	0.76	0.81
	Black or African American	3,982	3.12	0.82	3.11	0.82	−0.01	0.74	0.74	3.08	0.89	−0.05	0.77	0.81
	Mexican, Mexican American, or Chicano	871	3.40	0.79	3.39	0.78	0.00	**0.69**[a]	**0.69**[a]	3.39	0.80	0.01	0.74	0.78
	Puerto Rican	262	3.43	0.83	3.38	0.80	−0.06	0.75	0.75	3.34	0.85	−0.10	0.73	0.79
	Other Hispanic, Latino, or Latin American	1,397	3.38	0.79	3.37	0.80	−0.02	0.72	0.72	3.36	0.83	−0.01	0.76	0.80
	White (non-Hispanic)	24,601	3.67	0.80	3.66	0.79	−0.01	0.71	0.71	3.67	0.77	−0.01	0.74	0.79
Country	India	3,056	2.82	0.82	2.82	0.81	0.00	**0.67**[a]	**0.67**[a]	2.86	0.83	0.03	**0.70**[a]	0.74
	China	8,962	2.81	0.62	2.80	0.63	−0.03	**0.49**[a]	**0.50**[a]	3.04	0.74	**0.35**[a]	**0.53**[a]	**0.61**[a]
	Canada	255	3.72	0.94	3.74	0.89	0.02	0.74	0.74	3.62	0.93	−0.09	0.77	0.81
	Korea	377	3.02	0.78	2.96	0.77	−0.08	**0.67**[a]	**0.68**[a]	3.12	0.84	0.09	0.72	0.76
	Taiwan	288	2.70	0.73	2.72	0.69	0.02	**0.63**[a]	**0.64**[a]	2.75	0.83	0.04	**0.68**[a]	0.72
	Hong Kong	95	3.42	0.83	3.34	0.77	**−0.11**[a]	**0.65**[a]	**0.66**[a]	3.56	0.70	**0.20**[a]	**0.62**[a]	**0.69**[a]

Note. Std diff = standardized difference; wtd = weighted; adj = adjacent; unrnd = unrounded.
[a]Agreement indices that failed to meet the guideline thresholds (also indicated by boldface).

Appendix J

Bias Effects With Adjudication Thresholds of 1.0 and 1.5

Table J1 Effects of Biased e-rater Scores for the Argument Essays at the Adjudication Threshold of 1.0

	Bias					
	Baseline		−0.1		−0.2	
Classification of simulated score	Frequency	Percentage	Frequency	Percentage	Frequency	Percentage
No problem (no H2 invoked)	1,569	87.65	1,554	86.82	1,537	85.87
H2 invoked and in threshold	203	11.34	220	12.29	233	13.02
High outlier	13	0.73	12	0.67	13	0.73
Low outlier	5	0.28	4	0.22	7	0.39
Double-outlier	0	0	0	0	0	0
Missing e-rater/Flagged e-rater	0	0	0	0	0	0
Task score						
0	0	0	0	0	0	0
0.5	0	0	0	0	0	0
1	11	0.62	11	0.62	12	0.67
1.5	5	0.28	9	0.51	8	0.45
2	176	9.89	178	9.99	184	10.33
2.5	30	1.69	29	1.63	28	1.57
3	663	37.27	669	37.56	675	37.9
3.5	48	2.7	51	2.86	49	2.75
4	624	35.08	615	34.53	606	34.03
4.5	18	1.01	25	1.4	32	1.8
5	175	9.84	165	9.26	158	8.87
5.5	14	0.79	15	0.84	15	0.84
6	15	0.84	14	0.79	14	0.79

Note. H2 = Human 2.

Table J2 Effects of Biased e-rater Scores for the Issue Essays at the Adjudication Threshold of 1.0

	Bias					
	Baseline		−0.1		−0.2	
Classification of simulated score	Frequency	Percentage	Frequency	Percentage	Frequency	Percentage
No problem (no H2 invoked)	1,685	94.13	1,667	93.13	1,631	91.12
H2 invoked and in threshold	105	5.87	121	6.76	155	8.66
High outlier	0	0	2	0.11	3	0.17
Low outlier	0	0	0	0	1	0.06
Double-outlier	0	0	0	0	0	0
Missing e-rater/Flagged e-rater	0	0	0	0	0	0
Task score						
0	0	0	0	0	0	0
0.5	0	0	0	0	0	0
1	13	0.73	13	0.73	13	0.73
1.5	3	0.17	3	0.17	3	0.17
2	174	9.72	177	9.89	183	10.22
2.5	25	1.4	23	1.28	22	1.23
3	686	38.32	689	38.49	685	38.27
3.5	22	1.23	26	1.45	33	1.84
4	678	37.88	673	37.6	666	37.21
4.5	15	0.84	18	1.01	30	1.68
5	156	8.72	150	8.38	138	7.71
5.5	7	0.39	8	0.45	9	0.5
6	11	0.61	10	0.56	8	0.45

Note. H2 = Human 2.

Table J3 Effects of Biased e-rater Scores for the Argument Essays at the Adjudication Threshold of 1.5

	Bias					
	Baseline		−0.1		−0.2	
Classification of simulated score	Frequency	Percentage	Frequency	Percentage	Frequency	Percentage
No problem (no H2 invoked)	1,758	98.21	1,761	98.38	1,751	97.82
H2 invoked and in threshold	14	0.78	14	0.78	20	1.12
High outlier	13	0.73	11	0.61	12	0.67
Low outlier	5	0.28	4	0.22	7	0.39
Double-outlier	0	0	0	0	0	0
Missing e-rater/ Flagged e-rater	0	0	0	0	0	0
Task score						
0	0	0	0	0	0	0
0.5	0	0	0	0	0	0
1	12	0.67	13	0.73	13	0.73
1.5	1	0.06	0	0	0	0
2	204	11.4	207	11.56	207	11.56
2.5	8	0.45	6	0.34	6	0.34
3	700	39.11	702	39.22	702	39.22
3.5	3	0.17	4	0.22	4	0.22
4	641	35.81	638	35.64	639	35.7
4.5	0	0	1	0.06	2	0.11
5	190	10.61	189	10.56	187	10.45
5.5	3	0.17	4	0.22	9	0.5
6	28	1.56	26	1.45	21	1.17

Note. H2 = Human 2.

Table J4 Effects of Biased e-rater Scores for the Issue Essays at the Adjudication Threshold of 1.5

	Bias					
	Baseline		−0.1		−0.2	
Classification of simulated score	Frequency	Percentage	Frequency	Percentage	Frequency	Percentage
No problem (no H2 invoked)	1,785	99.72	1,777	99.27	1,771	98.94
H2 invoked and in threshold	5	0.28	11	0.61	15	0.84
High outlier	0	0	2	0.11	3	0.17
Low outlier	0	0	0	0	1	0.06
Double-outlier	0	0	0	0	0	0
Missing e-rater/Flagged e-rater	0	0	0	0	0	0
Task score						
0	0	0	0	0	0	0
0.5	0	0	0	0	0	0
1	15	0.84	15	0.84	15	0.84
1.5	0	0	0	0	0	0
2	187	10.45	187	10.45	187	10.45
2.5	2	0.11	5	0.28	6	0.34
3	709	39.61	706	39.44	705	39.39
3.5	2	0.11	2	0.11	4	0.22
4	684	38.21	686	38.32	685	38.27
4.5	0	0	2	0.11	3	0.17
5	170	9.5	167	9.33	165	9.22
5.5	1	0.06	2	0.11	2	0.11
6	20	1.12	18	1.01	18	1.01

Note. H2 = Human 2.

Appendix K

Additional Adjudications as a Function of Bias on the Analytical Writing Scores

Table K1 Effects of Biased e-rater Scores on the Analytical Writing Score at the Adjudication Threshold of 1.0

	Score difference group									
Bias	≤ −1.5	−1	−0.5	−0.25	0	0.25	0.5	1	≥ 1.5	Total
Baseline	0	0	42	366	843	445	84	9	1	1,790
−0.1	0	1	47	382	842	430	79	8	1	1,790
−0.2	0	2	49	395	846	418	73	6	1	1,790

Table K2 Effects of Biased e-rater Scores on the Analytical Writing Score at the Adjudication Threshold of 1.5

	Score difference group									
Bias	≤ −1.5	−1	−0.5	−0.25	0	0.25	0.5	1	≥ 1.5	Total
Baseline	0	4	52	408	842	403	74	6	1	1,790
−0.1	0	6	54	409	841	401	73	6	0	1,790
−0.2	0	6	54	409	841	403	71	6	0	1,790

Suggested citation:

Breyer, F. J., Attali, Y., Williamson, D.M., Ridolfi-McCulla, L., Ramineni, C., Duchnowski, M., & Harris, A. (2014). *A study of the use of the e-rater® scoring engine for the analytical writing measure of the GRE® revised General Test* (ETS Research Report No. RR-14-24). Princeton, NJ: Educational Testing Service. doi:10.1002/ets2.12022

Action Editor: James Carlson

Reviewers: Douglas Baldwin and Brent Bridgeman