Listening. Learning. Leading.®

Research Report

ETS RR-12-06

Evaluation of the *e-rater*® Scoring Engine for the *TOEFL*® Independent and Integrated Prompts

Chaitanya Ramineni

Catherine S. Trapani

David M. Williamson

Tim Davey

Brent Bridgeman

March 2012

Evaluation of the *e-rater*® Scoring Engine for the *TOEFL*® Independent and Integrated Prompts

Chaitanya Ramineni, Catherine S. Trapani, David M. Williamson,
Tim Davey, and Brent Bridgeman
ETS, Princeton, New Jersey

March 2012

Technical Review Editor: Joel Tetreault

Technical Reviewers: Shelby J. Haberman and Yigal Attali

Abstract

Scoring models for the *e-rater*® system were built and evaluated for the *TOEFL*® exam's independent and integrated writing prompts. Prompt-specific and generic scoring models were built, and evaluation statistics, such as weighted kappas, Pearson correlations, standardized differences in mean scores, and correlations with external measures, were examined to evaluate the e-rater model performance against human scores. Performance was also evaluated across different demographic subgroups. Additional analyses were performed to establish appropriate agreement thresholds between human and e-rater scores for unusual essays and the impact of using e-rater on operational scores. Generic e-rater scoring models were recommended for operational use for both independent and integrated writing tasks. The two automated scoring models were recommended for operational use to produce contributory scores within a discrepancy threshold of 1.5 and 1.0 with a human score for independent and integrated prompts respectively.

Key words: e-rater, automated essay scoring, TOEFL, writing, automated scoring models

Acknowledgments

The authors wish to thank Yigal Attali, F. Jay Breyer, Neil Dorans, Marna Golub-Smith, Shelby Haberman, Don Powers, and Cathy Wendler for their assistance in interpretation of the results; Lin Wang, Robert Kantor, Teresa Sanchez-Lazer, Tony Ostrander, and the TOEFL program for providing the data and their expert knowledge of the TOEFL program; and Jill Carey, Vincent Weng, Slava Andreyev, and Waverely VanWinkle for their assistance with the data and analyses.

Table of Contents

List of Tables

The *TOEFL*® exam is a widely administered English language proficiency test with scores accepted by colleges, agencies and other institutions all across the globe. The TOEFL exam has two current formats depending on the location of the test center. Most test takers take the Internet-based version of the TOEFL exam (*TOEFL iBT*® exam), while test centers that do not have Internet access offer the paper-based test (PBT, roughly 4% of the annual volume of test takers). TOEFL iBT assesses all four language skills (reading, listening, speaking, and writing) that are important for effective communication. The writing section has two writing (constructed response [CR]) tasks limited to a total of 50 minutes; one essay is an *integrated* task that requires test takers to read, listen, and then respond in writing by integrating what they have read and heard, and the other is an *independent* task which requires test takers to support an opinion on a topic.

With the trend of increased use of CR items within the last decade, many other high-stakes assessments such as the *GRE*®, *SAT*®, and GMAT exams also currently include CR items in speaking and/or writing sections. These items are believed to measure aspects of a construct that are not adequately addressed through multiple-choice items. However, compared to their multiple-choice counterparts, such items take longer to administer with smaller contributions to reliability per unit time and delay score reporting due to the additional effort and expense typically required to recruit, train, and monitor human raters. Against this backdrop of increasing use of CR items, there is potential value of automated scoring, in which computer algorithms are used to score CR tasks to either augment or replace human scorers.

Automated scoring systems, in particular systems designed to score a particular type of response that is in relatively widespread use across various assessments, purposes, and populations can provide a greater degree of construct representation. Examples of automated scoring systems include essay scoring systems (Shermis & Burstein, 2003), automated scoring of mathematical equations (Singley & Bennett, 1998; Risse, 2007), scoring short written responses for correct answers to prompts (Callear, Jerrams-Smith, & Soh, 2001; Leacock & Chodorow, 2003; Mitchell, Russell, Broomhead, & Aldridge, 2002; Sargeant, Wood, & Anderson, 2004; Sukkarieh & Pulman, 2005), and the automated scoring of spoken responses (Bernstein, De Jong, Pisoni, & Townshend, 2000; Chevalier, 2007; Franco et al., 2000; Xi, Higgins, Zechner, & Williamson, 2008; Zechner & Bejar, 2006). Of these, the domain that has been at the forefront of applications of automated scoring is the traditional essay response, with more than 12 different automated essay evaluation systems available for scoring and/or for performance feedback and improvement of writing quality. The most widely known of these systems include the Knowledge Analysis Technologies (KAT) engine 5 (Landauer, Laham, & Foltz, 2003), the *e-rater*® system (Attali & Burstein, 2006; Burstein, 2003), Project Essay Grade (Page, 1966, 1968, 2003) and IntelliMetric (Rudner, Garcia, & Welch, 2006). Each of these engines targets a generalizable approach of modeling or predicting human scores, yet each takes a somewhat different approach to achieving the desired scoring, both through different statistical methods as well as through different formulations of what features of writing are measured and used in determining the score. An explanation of how these systems work is beyond the scope of this paper, except for e-rater, which will be provided later in the paper.

Automated scoring in general can provide value that approximates some advantages of multiple-choice scoring, including fast scoring, constant availability of scoring, lower per unit costs, reduced coordination efforts for human raters, greater score consistency, a higher degree of tractability of score logic for a given response, and the potential for a degree of performance-specific feedback, that is not feasible under operational human scoring. These advantages, in turn, may facilitate allowing some testing programs and learning environments to make greater use of CR items where such items were previously too onerous to support. However, accompanying such potential advantages is a need to evaluate the cost and effort of developing such systems and the potential for vulnerability in scoring unusual or bad-faith responses inappropriately, to validate the use of such systems, and to critically review the construct that is represented in resultant scores.

E-rater automated scoring models were evaluated in the past for the writing prompts included in an earlier computer-based version of the TOEFL test (referred to as the TOEFL CBT). The TOEFL iBT was introduced in 2006 to replace the TOEFL CBT. Under the TOEFL CBT, examinees were required to write to one brief essay prompt in 30 minutes. Burstein and Chodorow (1999) and Chodorow and Burstein (2004) evaluated e-rater performance and sensitivity to essay length for responses on these prompts using e-rater99 and e-rater01, followed by Attali and Burstein (2006) who used e-rater v2. Attali (2007) evaluated the performance of a single generic scoring model for data from prompts administered under the TOEFL CBT. E-rater v2 was later evaluated for the iBT independent prompts using data from 2006–2007, and generic e-rater scoring models were found to perform satisfactorily against human scores and compared to other e-rater scoring models such as generic with prompt-specific intercepts and prompt-specific models (Attali, 2008; Attali, Bridgeman, & Trapani, 2010). E-rater generic with prompt-specific and prompt-specific models were implemented for operational use (as a quality check score) for GRE issue and argument prompts, respectively (Ramineni, Trapani, Williamson, Davey, & Bridgeman, 2012), helping the program use human raters more effectively. The two TOEFL writing tasks—independent and integrated—are somewhat analogous to the two GRE writing tasks—issue and argument—with one task requiring the examinee to support an opinion and the other reflecting on examinee ability to analyze and present relevant material, although for the integrated task, some specificity of the response is required.

The success of automated scoring models for the TOEFL CBT prompts and the iBT independent prompts as well as for the GRE writing prompts, along with the projected cost and time benefits for operational use, supported further evaluation of e-rater for scoring the two TOEFL writing tasks. Hence, the purpose of this study was to develop and evaluate e-rater automated scoring models for the TOEFL iBT independent and integrated writing prompts. In particular, this study investigated if e-rater scores could successfully replace one of the two human raters in operational scoring of the two TOEFL writing tasks, thereby effectively reducing the program costs and ensuring fast and consistent score turnaround for the large number of test takers, including prospective graduate applicants who take the test throughout the year at several computer-based test centers in the United States, Canada, and many other countries.

Scoring Rules for TOEFL Writing Tasks

Under the human scoring process for the TOEFL writing tasks, the writing samples from the tests were distributed to trained raters who assigned a score to each essay using a 5-point holistic scale. The scale reflects the overall quality of an essay in response to the assigned task. Each essay received scores from two trained raters, and the scores were averaged unless the two scores differed by more than one point, in which case, a third rating was obtained. If the three scores were adjacent to one another, then the third rating was the final score, but if one of the three scores was an outlier, then the average of the two adjacent scores was assigned as the final score. In the rare instance when none of the three scores was adjacent to the other (e.g., 1, 3, 5, the only possible case), a fourth adjudicated rating became the final score. Also, if any rater assigned a score of 0, the response called for adjudication and the adjudicated rating, which may be 0, 1, 2, 3, 4, or, 5, was the final score. The final scores on the two tasks were then added for each examinee to produce the raw score for the writing section, which ranged from 0 to 10 in increments of 0.5 and was converted to a scaled score of 0 to 30. A complete TOEFL scoring guide for the two writing tasks is included in Table A1.

Automated Scoring With the e-rater System

E-rater is a computer program that scores essays primarily on the basis of features that are related to writing quality. The initial version of e-rater (Burstein, Kukich, Wolff, Lu, & Chodorow, 1998) used more than 60 features to assess quality of writing in written assessments. In e-rater v2 (Attali & Burstein, 2006), the features were combined into a smaller set of features intuitively linking to general dimensions of writing quality for scoring. This set of features is constantly refined and enhanced in newer versions of e-rater, with e-rater v11.1 currently in operation. E-rater primarily emphasizes the characterization of writing quality rather than the content discussed in the essay, although some content features are used in the scoring. The e-rater program essentially uses natural language processing (NLP) technology to evaluate a number of characteristics of the essay, including grammar, usage, mechanics, development, and other features. These characteristics of essay quality are used to derive a prediction of the score that a human grader would have provided for the same response.

Features. E-rater currently uses 11 scoring features, with nine representing aspects of writing quality and two representing content. Most of these primary scoring features are composed of a set of subfeatures computed from NLP techniques, and many of these have multiple layers of microfeatures that have cascaded up to produce the subfeature values. An illustration of the construct decomposition of e-rater resulting from this structure is provided in Figure A1, where the features encapsulated in bold are the independent variables in the regression and the other features are an incomplete illustrative listing of subfeatures measuring aspects of writing quality. The scoring features of e-rater are mapped to the 6-trait model (Culham, 2003) commonly used to evaluate writing by teachers as described by Quinlan, Higgins, and Wolff (2009). More information on the microfeatures is available in Ramineni et al. (2012).

Grammar, usage, mechanics, and style together identify over 30 error types, including errors in subject-verb agreement, homophone errors, misspelling, and overuse of vocabulary. These error types are summarized for each feature as proportions of error rates relative to the essay length. Organization and development features are based on automatically identifying sentences in an essay as they correspond to essay-discourse categories: introductory material (background), thesis, main ideas, supporting ideas, and conclusion. For the organization feature, e-rater identifies the number of elements present for each category of discourse in an essay. For the development feature, e-rater computes the average length for all the discourse elements (in words) in an essay. Lexical complexity of the essay is represented by two features. The first is computed through a word frequency index used to obtain a measure of vocabulary level. The second feature computes average word length across all words in the essay and uses this as an index of sophistication of word usage. A new feature indicative of correct use of collocation and preposition use in the essay was the first feature to be included in e-rater version 10.1 to support further development of measures of positive attributes of writing style and ability (Ramineni, Davey, & Weng, 2010).

Two prompt-specific vocabulary usage features relate to content of vocabulary used in the essay. Both features are based on the tendency to use words typical of those used in prior essays. The first feature indicates the score point level to which the essay text is most similar with regard to vocabulary usage. The second analyzes the similarity of essay vocabulary to prior essays with the highest score point on the scale. A revised version of these features include information for all score points in computing the two measures (Attali, 2009).

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 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 entered on the computer and should be representative of the population for which e-rater is intended for use.

Prior to model build, the selected essay set is subjected to advisory flag analyses. Advisory flags act as filters and mark problems, because of which, an essay would be identified as inappropriate for automated scoring. Some examples of these flags are reuse of language, repetition of words, too brief, and so on. The use of these flags for an assessment is evaluated by comparing when e-rater considers an essay inappropriate versus when a human rater considers an essay inappropriate or off topic. Subjecting the sample of essays to advisory flagging prior to model build improves quality of model build by filtering the inappropriate essays from going into the model build phase for e-rater.

If no severe advisory flags that would preclude automated scoring have been issued, the e-rater program uses NLP technology to evaluate a number of characteristics of the essays in the model build set, including grammar, usage, mechanics, and development, among other features. After the feature values are derived, the weights for the features are determined using a multiple regression procedure. These feature weights can then be applied to additional essays to produce a predicted score based on the calibrated feature weights. Because the feature weights are estimated so as 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. 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 with agreement between two human raters.

The regression-based procedure of using NLP-based features to derive the automated score within e-rater lends itself to multiple methods of model construction. The following model types were built for the TOEFL data:

- *Prompt-specific (PS).* These are custom-built models for each prompt in the item pool. They are designed to provide the best fit models for the particular prompt in question, with both the feature weights and the intercept customized for the human score distribution used to calibrate the prompt model. Prompt-specific models incorporate prompt-specific vocabulary-related content features into the scoring.

- *Generic (G).* The smaller set of features derived in e-rater v2 enabled use of a single scoring model, referred to as a generic model, and standards across all prompts of an assessment. Generic models are based upon taking a group of related prompts, typically 10 or more, and calibrating a regression model across all prompts so that the resultant model is the best fit for predicting human scores for all the prompts, taken as a whole. As such, a common set of feature weights and a single intercept are used for all prompts regardless of the particular prompt in the set. Generic models do not take into account the content of the essay and address only writing quality; content features related to the vocabulary usage are prompt-specific and therefore not included in the regression. The generic modeling approach has the advantage of requiring smaller sample sizes per prompt (with enough prompts) and a truly consistent set of scoring criteria regardless of the prompt delivered operationally.

The generic with prompt-specific intercept model is a variant of the generic model and offers a common set of weights for all features with a customized intercept for each prompt.

Evaluation criteria. Once the automated (e-rater) scores for all essays have been calculated, ETS uses guidelines and criteria to assess the quality of the models. Flagging conditions or thresholds are attached to the evaluation statistics to serve as warnings of potential performance problems. However, the flags are used as guides rather than absolute rules when determining if a scoring model is acceptable for operational use. All the performance guidelines are applied to the independent evaluation sample used to validate the scoring models. The results on the evaluation sample independent from the model building sample represent a more generalizable measure of performance that would be more consistent with what would be observed on future data.

Construct evaluation. Automated scoring capabilities, in general, are designed with certain assumptions and limitations regarding the tasks they will score. Therefore, the initial step in any prospective use of automated scoring is the evaluation of fit between the goals and design of the assessment (or other use of automated scoring) and the design of the capability itself. The process includes a comparison of the construct of interest with that represented by the capability and reviews of task design, scoring rubrics, human scoring rules, score reporting goals, and claims and disclosures.

Association with human scores. Absolute agreement of automated scores with human scores has been a longstanding measure of the quality of automated scoring. Although it is common to report absolute agreements as percentages of cases being exact agreements and exact-plus-adjacent agreements, in evaluation of e-rater for assessment, these figures are only reported in statistical analysis reports as conveniences for laypersons rather than as part of acceptance criteria due to scale dependence (values will be expected to be higher by chance on a 4-point scale than on a 6-point scale) and sensitivity to base distributions (tendencies of human scores to use some score points much more frequently than others). Therefore, the agreement of automated scores with their human counterparts is typically evaluated on the basis of quadratic-weighted kappa and Pearson correlations. Specifically, the preferred quadratic-weighted kappa value between automated and human scoring is 0.70 (rounded normally). This value was derived on the conceptual basis that it represents the tipping point at which signal outweighs noise in the prediction. The identical threshold of 0.70 has been adopted for Pearson correlations. It should be noted that the results from quadratic-weighted kappa and Pearson correlations are not identical as kappa is computed on the basis of values of e-rater that are rounded normally to the nearest scale score point while the correlation is computed on the basis of unrounded values (e-rater scores are provided unrounded so that when multiple prompts are combined for a reported score the precise values can be combined and rounded at the point of scaling rather than rounding prior to summation). It is worthwhile to note that since e-rater is calibrated to empirically optimize the prediction of human scores, the expected performance of e-rater against this criterion is bounded by the performance of human scoring. That is, if the interrater agreement of independent human raters is low, especially below the 0.70 threshold, then automated scoring is disadvantaged in demonstrating this level of performance not because of any particular failing of automated scoring but because of the inherent unreliability of the human scoring upon which it is both modeled and evaluated. Therefore, the interrater agreement among human raters is commonly evaluated as a precursor to automated scoring modeling and evaluation. And, measures for quality of automated scores relative to the quality of the human scores are included in the evaluation framework. Two such measures are described next.

Degradation. Another criterion of performance in relationship with human scores recognizing the inherent relationship between the reliability of human scoring and the performance of automated scoring is degradation. 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 cases are occasionally observed in which the automated-human agreement for a particular prompt 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). Such models have been approved for operational use on the basis of being highly similar to human scoring and consistent with the purpose of the assessment for which they are used. Similarly, it is common to observe automated-human absolute agreements that are *higher* than the human-human agreements for prompts that primarily target writing quality.

Standardized mean score difference. A third criterion for association of automated scores with human scores is that the standardized mean score difference (standardized on the distribution of human scores) between the human scores and the automated scores cannot exceed 0.15. This criterion ensures that the distribution of scores from automated scoring is centered on a point close to what is observed with human scoring in order to avoid problems with differential scaling.

Association with external variables. Problems and concerns with human scoring represent a range of potential pitfalls including halo effects, fatigue, tendency to overlook details, and problems with consistency of scoring across time (Braun, 1988; Daly & Dickson-Markman, 1982; Hales & Tokar, 1975; Hughes & Keeling, 1984; Hughes, Keeling, & Tuck, 1980a, 1980b, 1983; Lunz, Wright, & Linacre, 1990; Spear, 1997; Stalnaker, 1936). Therefore, it is of relevance to investigate more than just the consistency with human scores and to also evaluate the patterns of relationship of automated scores, compared to their human counterparts, with external criteria. Scores on other test sections to examine within-test relationships and external criteria, such as self-reported measures that may be of interest (e.g., grades in English class, academic majors), are some examples that are used for this purpose. It should be noted that the external criteria that are typically available are not a direct external measure of exactly the same construct and hence often pose some problems for interpretation.

Subgroup differences. In evaluating fairness of automated scoring, the question is whether it is fair to subgroups of interest to substitute a human rater with an automated score. Due to lack of a suitable differential item functioning measure for this purpose, two approaches have been proposed and implemented to address measures of fairness for e-rater. The first extends the flagging criterion of standardized mean score differences from the prompt-level analysis discussed above to the evaluation of subgroup differences. A more stringent threshold of performance is adopted, setting the flagging criteria at 0.10, and applied to all subgroups of interest to identify patterns of systematic differences in the distribution of scores between human scoring and automated scoring for subgroups at the reported score level.

The second approach examines differences in the predictive ability of automated scoring by subgroup. This approach consists of two classes of prediction that are likewise related to the guidelines and processes discussed above. The first compares an initial human score and the automated score in their ability to predict the score of a second human rater by subgroup. The second type of prediction compares the automated and human score ability to predict an external variable of interest by subgroup.

Operational impact analysis. The final stage of the evaluation of automated scoring determines predicted impact on the aggregate reported score for the writing section. This impact is evaluated by simulating the score that would result from substituting an automated score for a human score and determining the distribution of changes in reported scores that would result from such a policy. This stage lends an additional opportunity to compare the performance of scoring under the proposed model (automated and human) to that of the traditional model (two human raters). In the empirical comparison, the primary areas of interest are an examination of the rate and degree of raw and scaled score differences resulting from the change, the differences in association of reported scores to other test scores and external criteria, and both of these applied to the level of subgroups of interest. Such an analysis allows for the consideration of issues in scale continuity and other factors that may bear on the decision to implement automated scoring.

Variations in agreement threshold. Alternative thresholds are considered for the definition of discrepancy when evaluating the operational agreement between automated and human scores. In human scoring, it is common practice for most scoring scales in high-stakes programs that use double-human scoring to consider scores that are one point apart (e.g., one rater issuing a 3 and the other a 4) to be in agreement under the interpretation that reasonable judges following the rubric may differ, especially when evaluating a borderline submission. Typically, when two human scores are considered discrepant, an adjudication process occurs in which additional human raters are used and a resolution process is followed to determine the final reported score. These adjudication and resolution processes vary substantially by program and are sometimes conditional on the particular distribution of initial human scores produced. In the implementation of automated scoring with precise values recorded (decimal values), a wider range of options are available for defining agreement, each of which has implications for the extent to which the results of automated scoring influence the final reported scores and therefore the ultimate evaluation of impact under the procedures defined above.

Methods

Independent Prompts

Prompt-specific, generic, and generic with prompt-specific intercept scoring models were first built and evaluated by the automated scoring group in 2007 for 26 iBT independent prompts using e-rater v7.2 on data from October 2006–May 2007. Based on the evaluation criteria, the generic with prompt-specific intercept scoring models were first recommended to the program with a discrepancy threshold (when the difference between e-rater precise value and first human score exceeds the threshold, a second human rater is required) of 2 points between automated and human scores (Williamson et al., 2007). Later, generic with prompt-specific intercept scoring models were graded as inefficient and impractical for implementation considering the design of the TOEFL test administration, which does not allow for pretesting of prompts. Therefore, a re-evaluation was carried out in 2008 on the same data using an upgraded e-rater version striving for improved models/results. Based on the evaluation results from the upgraded e-rater v8.1, generic e-rater scoring model was recommended to the program (Williamson, Trapani, & Weng, 2008). As a result, the generic e-rater scoring model was then approved and accepted by the program to produce a fully contributory score (e-rater score is taken in combination with human score as the reported score for the writing task) in operational scoring of the independent prompts within a discrepancy threshold of 1 point between automated and human scores. However, due to the changing examinee population in newer TOEFL administrations as observed at the end of 2008, another re-evaluation of e-rater was conducted for the independent prompts on more recent data. This re-evaluation was carried out in early 2009 using 38 iBT independent prompts from the year 2008 (the previous data were from 2006–2007). Since the generic with prompt-specific intercept scoring model had already been considered unacceptable for implementation in the preceding evaluations, only prompt-specific and generic e-rater scoring models were built and evaluated on the new data. As a result, the generic e-rater scoring model producing a contributory score was recommended to the program for operational use within a discrepancy threshold of 1.5 between automated and human scores for independent prompts (Williamson et al., 2009a).

Integrated Prompts

Following the evaluation of e-rater for independent prompts, generic and prompt-specific scoring models were evaluated for integrated writing tasks using the same data from 2008 and under e-rater v8.1. As a result, a generic scoring model producing a contributory score was recommended for operational use at a threshold of 1 point between automated and human scores (Williamson et al., 2009b).

The data, methods, and results are reported here for the most recent evaluations of e-rater for both the independent prompts and the integrated prompts that support the most recent recommendations by the automated scoring group for operational use and the program's implementation choices. These evaluations were conducted using more recent data from TOEFL iBT administrations and an upgraded version of e-rater (v8.1) than used in any of the previous studies evaluating use of e-rater for TOEFL.

Data

More than 152,000 operational responses across 38 independent prompts and 38 integrated prompts were drawn from the available test records from January 2008 to October 2008. This resulted in roughly 4,000 essays per prompt. Along with the two human rater scores for each essay, several additional variables were included for analysis—examinee background variables (gender, native language, test center country, and ability level) and other TOEFL section test scores (reading, speaking, and listening).

The quality of the e-rater models estimated and the effective functioning of the models in operational settings depend critically on the nature and quality of the training and evaluation data. Thereby, certain guidelines have been developed by the automated scoring group at ETS that are used to guide the collection and analyses of the data for building and evaluation of automated scoring models (Williamson & Davey, 2007). These include choosing a representative sample, double scored essays in electronic format, and a sufficient number of prompts and minimum sample sizes for model building. For the assumptions not met, there are subsequent implications when interpreting the results. The data provided by the TOEFL program met all the guidelines for automated scoring model building and evaluation.

E-rater v8.1 was used for the evaluations. This version of e-rater had 10 features (excluding the positive measure on the use of collocations and prepositions) and the content features used information only from one score point (unlike the revised content features that derive information from all five score points). Also at the subfeature level, good collocation density and good preposition usage under positive feature and double negation under usage were not present in v8.1. However, it should be noted that during the annual engine upgrade process each year, new models are built and evaluated using the latest e-rater version for all high- and low-stakes assessments using e-rater for operational scoring.

Construct Relevance

The construct of the TOEFL writing assessment was partially evaluated for the independent writing tasks against the construct represented by e-rater as part of a previous study (Quinlan et al., 2009) and can be sufficiently extended to the integrated tasks as well. The two TOEFL writing tasks require test takers to either integrate what they read and hear and respond to it or support an opinion on a topic. The ideas and content in the responses are measured primarily by two e-rater features that use content vector analysis. These features measure topic-specific vocabulary use only, and therefore, the breadth of construct coverage is limited. However, they do a fairly reasonable job of measuring this limited domain. The TOEFL writing assessment demands a well-focused, well-organized analysis representing a logical connection of ideas, which is measured by the organization/development features of e-rater. The organization and development features measure the number and average length of discourse units (i.e., functionally related segments of text) in an essay and are strongly correlated with essay length. In addition, the TOEFL writing tasks elicit fluent and precise expression of ideas using effective vocabulary and sentence variety. These traits are represented in e-rater by a variety of microfeatures that measure sentence-level errors (e.g., run-on sentences and fragments), grammatical errors (e.g., subject-verb agreement), and the frequency with which the words in an essay are commonly used. The TOEFL scoring rubric also emphasizes test takers' ability to demonstrate facility with conventions (i.e., grammar, usage, and mechanics) of standard written English. This 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.

The reviews of task design, scoring rubric, human scoring rules, reporting goals, and claims and disclosures for the assessment were made in conjunction with the TOEFL program as the study progressed.

Model Building and Evaluation

Prompt-specific (PS) and generic (G) scoring models were built and evaluated for the TOEFL independent and integrated data from 2008 using e-rater v 8.1.

Agreement statistics for automated scores with human scores were computed for all e-rater models built and evaluated for the TOEFL data. The best chosen model(s) was then subjected to remaining evaluation criteria of association with external variables, subgroup differences, operational impact analysis, and agreement thresholds for adjudication.

The following section presents the results for each scoring model type developed/evaluated for the two prompt types (independent and integrated). The results for each model are supported with summary tables of performance at the aggregate level in the main text and summary tables of performance at the prompt level in Tables B1–B4.

Results

Advisory Analyses

E-rater has a number of advisory flags to indicate when e-rater is inappropriate for scoring a specific essay response. The use of these flags as effective filters was evaluated following the standard procedures for building and evaluating e-rater scoring models. All advisories were evaluated against human1 (H1) ratings individually and sequentially, and as a result, four flags were identified for use in operational setting: those marking less development of the key concepts than other essays written on the topic, excessive length, brevity, and too many problems (large number of grammar, usage, and mechanics errors).

The use of these rules overall flagged a very small number of cases (less than 1% for independent and just about 1% for integrated) requiring double-human scoring. The majority of flagging that required double human scoring occurred at the lower end of the scale regardless of the prompt type. However, more integrated essays were flagged than independent essays.

Model Build and Evaluation

Prompt-specific and generic scoring models were built for both independent and integrated tasks. There were 10 features in total for e-rater, as described earlier. The two content features related to topic-specific vocabulary usage are included only for PS models. Any features with negative weights were excluded from the final model build. Hence, the feature set for PS models varied from prompt to prompt. The G models for both independent and integrated prompts included all e-rater features (except for the two content features related to topic-specific vocabulary) in the final model build. The sample size was 500 for the model build set for all model types, and the remaining number of responses for each prompt determined the sample size for the evaluation set. The sample size for the evaluation set for each prompt can be found in the tables reporting results for each model at the prompt level in Tables B1–B4.

Agreement With Human Scores

The quality of automated scoring models rests on the characteristics of the human scoring used as the basis for modeling. Evaluation of the differences in raw scores under human/e-rater (H1/e-rater) scoring compared to human/human (H1/H2) scoring was conducted. The raw e-rater scores were produced on a continuous scale under the linear regression model. For computing exact and adjacent agreement percentages and weighted kappa statistics, the raw e-rater scores were first brought into range (truncated) to align with the score scale (1 to 5 for TOEFL writing tasks) and rounded to integers for comparison against the integer human scores. For other agreement statistics, such as Pearson correlation and standardized mean score differences, truncated e-rater scores without rounding were used for comparison with human scores. Tables 1 and 2 show results for quadratic-weighted kappas, Pearson correlations, standardized mean score differences, and degradation of e-rater/human agreement from human/human agreement for independent and integrated prompts respectively. The numbers in the shaded cells in these tables fail to meet the threshold values for that evaluation metric; a summary of the flagging criteria and conditions for evaluating model performance, explained under the evaluation criteria previously, is included in Table A2. It should be noted that all the threshold values are evaluated to four decimal places for flagging purposes; this explains why some two-digit values derived by rounding up are highlighted as not meeting the threshold. The tables reporting results for each model at the prompt level are included in Tables B1–B4.

The operational TOEFL had a correlation of 0.69 and 0.82 for scores by human raters on responses to independent and integrated prompts respectively. The correlation for human scores for the independent prompts was slightly below the set threshold of 0.70 and lower than that for the integrated prompts; the smaller standard deviations for the independent prompts restricted the correlations although the agreement rates for humans were similar for both the independent and the integrated prompts.

Table 1

Agreement With Human Scores for Independent Prompts

	H1 by H2											H1 by e-rater (rounded to integers)						H1 by e-rater (unrounded)				Degradation	
		H1		H2		Stats						e-rater						e-rater		Stats		Wtd kappa	*R*
	N	M	SD	M	SD	Std diff	Kappa	Wtd kappa	% agree	% adj agree	*R*	M	SD	Kappa	Wtd kappa	% agree	% adj agree	M	SD	Std diff	*R.*	H1 by e-rater rounded – H1 by H2	H1 by e-rater unrounded – H1 by H2
Generic	3,483	3.35	0.85	3.35	0.85	0.01	0.39	0.69	60	98	0.69	3.36	0.86	0.39	0.69	59	99	3.36	0.83	0.01	0.74	0.01	0.06
Prompt-specific	3,483	3.35	0.85	3.35	0.85	0.01	0.39	0.69	60	98	0.69	3.36	0.87	0.40	0.70	60	99	3.36	0.84	0.01	0.75	0.02	0.06

Note. N is average across all the prompts. Shaded cells indicate values that fail to meet the thresholds listed in Table A2.

adj = adjacent, H1 = human 1, H2 = human 2, std diff = standardized difference, wtd = weighted.

Table 2

Agreement With Human Scores for Integrated Prompts

	H1 by H2											H1 by e-rater (rounded to integers)						H1 by e-rater (unrounded)				Degradation	
		H1		H2		Stats						e-rater						e-rater		Stats		Wtd kappa	*R*
	N	M	SD	M	SD	Std diff	Kappa	Wtd kappa	% agree	% adj agree	*R*	M	SD	Kappa	Wtd kappa	% agree	% adj agree	M	SD	Std diff	*R.*	H1 by e-rater rounded – H1 by H2	H1 by e-rater unrounded – H1 by H2
Generic	3,316	3.08	1.19	3.07	1.19	0.00	0.48	0.82	60	97	0.82	3.08	1.10	0.20	0.59	39	87	3.07	1.12	-0.01	0.62	-0.23	-0.20
Prompt-specific	3,316	3.08	1.19	3.07	1.19	0.00	0.48	0.82	60	97	0.82	3.08	1.14	0.30	0.70	46	92	3.07	1.15	-0.01	0.73	-0.12	-0.10

Note. N is average across all the prompts. Shaded cells indicate values that fail to meet the thresholds listed in Table A2.

adj = adjacent, H1 = human 1, H2 = human 2, std diff = standardized difference, wtd = weighted.

For the independent prompts, all the e-rater evaluation criteria were sufficiently met for both models at the aggregated level (except for the weighted kappa for human/e-rater under generic model, which was close to the threshold) with improved correlations for e-rater with human score. At the prompt level, however, under the preferred generic model, 23 independent prompts failed to meet the 0.70 threshold for weighted kappa by a relatively small margin, and five prompts exceeded the threshold of absolute value of 0.15 for standardized mean score differences between e-rater and human score (with higher e-rater mean score for three of the five prompts). Under the prompt-specific model, serving as the baseline model, weighted kappa for human and e-rater score was slightly below the threshold value for 16 independent prompts. It should be noted that the weighted kappa and correlation values for the two human scores, which serve as the baseline for human with e-rater agreement statistics, were below the threshold for majority (31 out of 38) of the prompts.

For the integrated prompts, at the aggregate level, the degradation in agreement (lower weighted kappa and correlation values) from H1/H2 to H1/e-rater exceeded the absolute threshold value of 0.10 for both generic and prompt-specific scoring models. In addition, for the prompt-specific model, the threshold for weighted kappa was barely met, while for the generic model both correlation and weighted kappa measured below the desired 0.70 threshold. At the prompt level under generic model, none of the integrated prompts met the threshold criteria for correlation or weighted kappa with weighted kappa as low as 0.52 and correlation as low as 0.57, and 13 prompts were flagged for standardized mean score differences between e-rater and human scores greater than absolute value of 0.15, with largest standardized difference as much as absolute value of 0.25. All the prompts exceeded the allowable threshold of 0.10 for weighted kappa and correlation degradation from H1/H2 to H1/e-rater agreement with maximum absolute values of 0.34 and 0.29 for weighted kappa and correlation degradation respectively. Under the prompt-specific model, 16 prompts did not meet the weighted kappa threshold while 5 prompts failed to meet the correlation threshold for human and e-rater agreement. For 32 prompts, the weighted kappa degradation from H1/H2 to H1/e-rater exceeded the allowable threshold of 0.10 with maximum absolute value of 0.16, while 17 prompts had unacceptable degradation for correlation as high as absolute value of 0.13.

Based on the results for the evaluation criteria at the aggregate and the prompt level, the preferred generic model fared well for independent prompts but required further empirical evidence to support e-rater use for integrated prompts. Hence, the e-rater evaluation for integrated prompts included an evaluation for examinees who retook the TOEFL voluntarily between January and October 2008 (n = 7,894) who were already part of the e-rater study. These additional analyses on the selected retesters' sample allowed simulation and use of scores produced by multiple scoring methods (human and e-rater) on one occasion to predict scores produced by all human scoring at the other occasion. Table 3 shows the correlations between the scores on the two occasions under different simulation models.

Table 3

Association of Human Scoring With Multiple Scoring Methods for Retesters at Two Different Administrations (Time 1 and Time 2)

Simulated scoring methods		Simulated score at Time 1 with all human scores at Time 2	All human score at Time 1 with simulated score at Time 2
Independent	Integrated		
2 human scores	2 human scores	0.72	0.72
e-rater & 1 human	2 human scores	0.73	0.72
e-rater & 1 human	1 human	0.71	0.71
e-rater & 1 human	e-rater (G) & 1 human	0.73	0.73
e-rater & 1 human	e-rater (PS) & 1 human	0.74	0.74

Note. G = generic, PS = prompt-specific.

The results suggest that the use of generic e-rater models for integrated prompts was on par with all human scoring. Hence, the empirical evidence produced from retester analyses was considered adequate to further investigate the use of generic models for operational use for integrated prompts. The subsequent analyses use e-rater scores produced from generic scoring models for both independent and integrated prompts.

Association With External Measures

E-rater and human scores were correlated with external measures, such as scores on other test sections (TOEFL reading, listening, and speaking sections) and the total scaled score with and without writing. Table 4 reports the association of e-rater scores (rounded integer values from the chosen model for independent and integrated prompts) and human scores at rating level with these external measures.

Table 4

Score Association With Other Measures

	TOEFL reading scaled score	TOEFL listening scaled score	TOEFL speaking scaled score	Total scaled score w/o writing
Independent H1	0.53	0.53	0.58	0.62
e-rater	0.54	0.52	0.55	0.61
Integrated H1	0.62	0.65	0.58	0.71
e-rater	0.58	0.55	0.55	0.64

Note. H1 = human 1.

The correlations with the external variables for e-rater and human scores differed on an average by 0.02 for the independent prompts and the differences did not systematically favor human or e-rater scores. For the integrated prompts, the correlations of e-rater scores with external variables were uniformly lower than those with human scores.

Subgroup Differences

Analyses were conducted to investigate further the degree to which e-rater and human scores differ across subgroups, for example, whether males or females receive higher e-rater scores relative to their human scores or whether test takers from different countries receive different scores from e-rater compared to human scores. In general, if the human scores are accepted as the optimal desired score, standardized mean score differences of 0.05 or less are desirable for subgroups and those between 0.05 and 0.10 in magnitude may be considered acceptable; differences exceeding absolute value of 0.10 present concerns. Differences across subgroups based on gender, native language, test center country, and ability level were examined. The language groups represented by examinee population greater than 1% of the total annual test-taker volume were included for these analyses, and the ability level was defined based on the total scaled score. Tables 5 and 6 show the results for quadratic-weighted kappas, Pearson correlations, standardized mean score differences, and degradation of e-rater/H1 agreement from H1/H2 agreement for the different subgroups with significant mean score differences (greater than absolute value of 0.10) for the independent and integrated prompts respectively. Differences were observed for the test center country of China (as large as 0.25 with greater e-rater scores) and for the Chinese and Arabic language groups (as large as 0.21

with greater e-rater scores for Chinese and 0.19 with lower e-rater scores for Arabic) for both independent and integrated prompts. In addition, the mean score differences were also unacceptably large for Hindi and Spanish language groups for independent prompts and Portuguese and Telugu language groups for integrated prompts. Large differences were also observed on independent prompts for the medium ability group with larger e-rater scores and on integrated prompts for low ability level groups with larger e-rater scores and for high ability level groups with smaller e-rater scores For the independent prompts, except for the Arabic language group, H1/H2 agreement measures (weighted kappa and correlation) were lower than the desired threshold (0.70) and the H1/e-rater agreement was an improvement over the H1/H2 agreement. However, for the integrated prompts, degradation was observed in the H1/e-rater agreement compared to the H1/H2 agreement. For subgroups with small sample sizes (less than 1,000), any differences around or beyond the threshold were not considered for further formal review. Results for subgroups based on gender, other native language groups, and test center countries of interest are included in Tables B5 and B6, for independent and integrated prompts, respectively.

Models for Implementation

Various thresholds for allowable discrepancy levels between e-rater and human scores were examined to maximize cost savings related to the use of a second human grader while ensuring valid e-rater scores with acceptable agreement levels with human scores, correlations on par with external measures, and minimal subgroup differences. The allowable discrepancy threshold between the two human scores on a TOEFL writing task is 1 point. Scores discrepant by more than 1 point (that is, apart by 2 or more points as outlined previously under TOEFL scoring rules) are routed to a third human rater. Since e-rater produces real values—unlike human scores, which are restricted to integer values—scores greater than 1 but less than equal to 1.4999 are rounded down to 1 under normal rounding rules. Hence, adhering to the TOEFL scoring rules, a contributory model at threshold of 1.5 was initially chosen for evaluating the impact of including e-rater in operational scoring for both TOEFL writing tasks. However, the discrepancy threshold was reduced to 1 for the integrated writing task to mitigate subgroup differences. Table 7 reports the correlations of final scores with other measures for the TOEFL writing section simulated under the contributory score model: independent only, 1.5 threshold;

Table 5

Agreement With Human Scores on Independent Prompts for Test Takers From County China; Native Arabic, Chinese, Hindi and Spanish Language Groups; and Test Takers With High, Medium, and Low Ability Levels

Independent	H1 by H2											H1 by e-rater (rounded to integers)						H1 by e-rater (unrounded)				Degradation	
		H1		H2		Stats						e-rater						e-rater		Stats		Wtd kappa	*R*
	N	M	SD	M	SD	Std diff	Kappa	Wtd kappa	% agree	% adj agree	*R*	M	SD	Kappa	Wtd kappa	% agree	% adj agree	M	SD	Std diff	*R*	H1 by e-rater rounded – H1 by H2	H1 by e-rater unrounded – H1 by H2
Test center country																							
China	15,480	3.34	0.77	3.35	0.76	0.00	0.35	0.61	60	98	0.61	3.54	0.79	0.34	0.62	57	98	3.54	0.75	0.25	0.69	0.01	0.07
Native language																							
Arabic	7,751	3.03	0.90	3.04	0.90	0.01	0.41	0.72	60	98	0.72	2.87	0.95	0.40	0.73	58	99	2.85	0.93	-0.19	0.78	0.01	0.06
Chinese	25,268	3.30	0.78	3.31	0.77	0.02	0.36	0.63	60	98	0.63	3.46	0.80	0.36	0.64	58	99	3.46	0.76	0.21	0.70	0.01	0.07
Hindi	2,925	3.85	0.84	3.85	0.83	0.00	0.37	0.65	57	98	0.65	3.69	0.77	0.35	0.64	57	99	3.70	0.74	-0.18	0.70	0.00	0.06
Spanish	9,463	3.42	0.82	3.44	0.83	0.02	0.36	0.65	58	98	0.65	3.34	0.79	0.38	0.66	60	99	3.34	0.74	-0.11	0.71	0.01	0.06
Ability level[a]																							
High	48,133	4.01	0.70	4.01	0.69	0.01	0.27	0.46	55	98	0.46	3.96	0.61	0.27	0.47	58	99	3.96	0.55	-0.07	0.54	0.00	0.08
Med	48,561	3.32	0.62	3.33	0.62	0.01	0.26	0.42	60	98	0.42	3.41	0.65	0.28	0.45	59	99	3.41	0.58	0.15	0.51	0.03	0.09
Low	42,653	2.75	0.66	2.76	0.66	0.01	0.34	0.55	64	99	0.55	2.75	0.79	0.33	0.58	60	99	2.75	0.76	-0.01	0.64	0.03	0.08

Note. Shaded cells indicate values that fail to meet the thresholds listed in Table A2. adj = adjacent, H1 = human 1, H2 = human 2, std diff = standardized difference, wtd = weighted.

[a]Low: total scaled score 0–69; medium: 70–93; high: 94–120.

Table 6

Agreement With Human Scores on Integrated Prompts for Test Takers From Country China; Native Arabic, Chinese, Portuguese, and Telugu Language Groups; and Test Takers With High, Medium, and Low Ability Levels

Independent	H1 by H2										H1 by e-rater (rounded to integers)							H1 by e-rater (unrounded)				Degradation	
		H1		H2		Stats						e-rater						e-rater		Stats		Wtd kappa	*R*
	N	M	SD	M	SD	Std diff	Kappa	Wtd kappa	% agree	% adj agree	*R*	M	SD	Kappa	Wtd kappa	% agree	% adj agree	M	SD	Std diff	*R*	H1 by e-rater rounded – H1 by H2	H1 by e-rater unrounded – H1 by H2
Test center country																							
China	16,912	3.03	1.1	3.04	1.12	0.00	0.46	0.80	59	97	0.80	3.2	1.00	0.19	0.56	40	87	3.19	1.00	0.14	0.59	-0.24	-0.21
Native language																							
Arabic	7,918	2.70	1.21	2.69	1.20	0.00	0.54	0.85	64	97	0.85	2.54	1.15	0.21	0.58	39	84	2.49	1.19	-0.17	0.61	-0.26	-0.24
Chinese	27,370	2.98	1.15	2.98	1.14	0.00	0.47	0.81	60	97	0.81	3.11	1.06	0.19	0.57	39	87	3.11	1.06	0.11	0.59	-0.24	-0.22
Hindi	2,675	3.24	1.16	3.23	1.16	-0.01	0.48	0.82	61	97	0.82	3.14	1.06	0.17	0.58	38	87	3.13	1.05	-0.10	0.61	-0.24	-0.21
Spanish	4,048	3.18	1.15	3.19	1.15	0.00	0.46	0.80	59	96	0.80	3.30	1.01	0.16	0.50	37	85	3.30	1.00	0.11	0.53	-0.30	-0.27
Ability level[a]																							
High	48,485	4.08	0.78	4.07	0.78	-0.01	0.33	0.55	57	96	0.55	3.90	0.78	0.09	0.24	41	90	3.91	0.76	-0.22	0.27	-0.3	-0.28
Med	49,065	3.09	0.84	3.09	0.84	0.00	0.36	0.63	58	97	0.63	3.10	0.89	0.06	0.23	37	85	3.10	0.85	0.00	0.25	-0.4	-0.38
Low	43,653	1.95	0.88	1.95	0.88	0.00	0.51	0.75	67	98	0.75	2.17	0.94	0.14	0.33	40	85	2.11	0.99	0.18	0.36	-0.42	-0.39

Note. Shaded cells indicate values that fail to meet the thresholds listed in Table A2. adj = adjacent, H1 = human 1, H2 = human 2, std diff = standardized difference, wtd = weighted.

[a]Low: total scaled score 0–69; medium: 70–93; high: 94–120.

integrated only, 1 point threshold; and independent at 1.5 and integrated at 1 point combined. Compared to the operational writing score produced using two or more human ratings, the new simulated writing scores show fairly equal association with scores on other TOEFL test sections and the total scores with and without writing. There were no subgroup differences of formal concern under these models (Tables B7–B8).

Table 7

Reported Score Association With Other Measures Under Contributory Score Model for e-rater

	TOEFL reading scaled score	TOEFL listening scaled score	TOEFL speaking scaled score	Total scaled score w/o writing
Operational writing score (all human)	0.69	0.71	0.68	0.80
New simulated writing score (e-rater for independent only, 1.5 pt.)	0.69	0.71	0.67	0.79
New simulated writing score (e-rater for integrated only, 1 pt.)	0.68	0.70	0.68	0.79
(e-rater for independent 1.5 pt. and integrated 1 pt.)	0.69	0.69	0.68	0.79

Impact of Implementation

The rates of agreement and the anticipated number of second human ratings for scores based on all human scoring and scores based on human and e-rater combined were compared. Table 8 presents the rates of agreement and anticipated number of third ratings (adjudication) when using all humans versus when using e-rater with humans. For two human scores, the third rating will be provided by a third human rater when the human scores differ by 2 or more points. For one human and one e-rater score, the third rating will be provided by a second human rater when the human and e-rater scores differ by 1.5 points or more for the independent and 1 point or more for the integrated writing task. Results showed that when using e-rater, only 3% cases for independent and 33% cases for integrated needed more than one human score, which suggests more efficient use of human raters and reduced score turnaround time.

Table 8

Change in Agreement and Adjudication Rates for Independent and Integrated Writing Prompts Using e-rater Contributory Score Model

	Two ratings (no adjudication needed)	Anticipated third ratings (adjudication needed[a])
Independent	*N* (%)	
Operational scoring with all humans, adjudication at 2 points	138,772 (98%)	2,431 (2%)
H1/e-rater, adjudication at 1.5 points	137,517 (97%)	3,686 (3%)
Integrated		
Operational scoring with all humans, adjudication at 2 points	137,024 (97 %)	4,179 (3%)
H1/e-rater, adjudication at 1.0 point	94,952 (67%)	46,251 (33%)

[a] Occasionally more than three ratings are required for a very small percentage (<0.5) of cases and are collapsed in this category.

Conclusion

Prompt-specific and generic scoring models were built and evaluated on TOEFL data from January 2008 to October 2008 using e-rater v8.1. These data comprised over 152,000 essay responses written to 38 independent and integrated prompts. Criteria for evaluation of e-rater scoring models included level of agreement with human scores, degradation in agreement from human scoring, standardized mean score differences between human and automated scoring, and correlations with external variables (such as scores on other TOEFL test sections, total scores with and without writing). Based on the evaluation criteria, generic models were recommended for implementation for operational use for both the independent and integrated prompts, upon predicting scores for a sample of retesters. Performance of the generic models was further evaluated across different demographic subgroups. Results revealed adequate performance at the subgroup level, with a notable exception of discrepancy between e-rater and human scores for examinees from China (three-tenths of an SD higher on independent than the human scores), for examinees from certain native language groups (one-fourth of an SD higher than the human scores for Chinese on independent and two-tenths of an SD lower than the human scores for

Arabic on independent), and for certain ability level groups (two-tenths of an SD lower than the human scores for high ability level groups).

E-rater's use was investigated as a contributory score. Under the contributory score model, e-rater score was checked for agreement with the first human score within an empirically established range, beyond which a second human score was required. The average of the human and e-rater scores became the final score for the essay, unless a second human rating was desired. Various agreement thresholds were evaluated under the contributory score model to minimize differences across the subgroups. Discrepancy thresholds of one-and-a-half point and one point between the automated and the human score were selected for independent and integrated prompts respectively to yield performance as similar as possible to double human scoring, and with significant savings in second human ratings.

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 examinee and human rater characteristics, and human scoring trends over time, as well as new feature developments and enhancements in the e-rater engine. Models are being currently explored that differentially weight the independent and integrated tasks in determining the overall writing score. We will also investigate the differences in e-rater and human scores observed for some subgroups in this evaluation to better understand their source and origin.

References

Attali, Y. (2007). *Construct validity of e-rater in scoring TOEFL essays* (ETS Research Report No. RR-07-21). Princeton, NJ: ETS.

Attali, Y. (2008). *E-rater performance for TOEFL iBT independent essays*. Unpublished manuscript.

Attali, Y. (2009). *Interim summary of analyses related to content scoring of TOEFL integrated essays*. Unpublished manuscript. Princeton, NJ: ETS.

Attali, Y., Bridgeman, B., & Trapani, C. S. (2010). Performance of a generic approach in automated scoring. *Journal of Technology, Learning, and Assessment, 10*(3). Retrieved from http://www.jtla.org.

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

Bernstein, J., De Jong, J., Pisoni, D., & Townshend, B. (2000). Two experiments on automatic scoring of spoken language proficiency. In *Proceedings of InSTIL2000 (Integrating Speech Tech. in Learning;* pp. 57–61). Dundee, Scotland: University of Abertay.

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

Burstein, J. (2003). The e-rater® scoring engine: Automated essay scoring with natural language processing. In M. D. Shermis & J. C. Burstein (Eds.), *Automated essay scoring: A cross-disciplinary perspective* (pp. 113–121). Hillsdale, NJ: Lawrence Erlbaum.

Burstein, J., & Chodorow, M. (1999). Automated essay scoring for nonnative English speakers. In M. Broman Olsen (Ed.), *Computer mediated language assessment and evaluation in natural language processings* (pp. 68–75). Morristown, NJ: Association for Computational Linguistics.

Burstein, J., Kukich, K., Wolff, S., Lu, C., & Chodorow, M. (1998, April). *Computer analysis of essays*. Paper presented at the annual meeting of the National Council on Measurement in Education, Montreal, Canada.

Callear, D., Jerrams-Smith, J., & Soh, V. (2001). CAA of short non-MCQ answers. In *Proceedings of the 5th International CAA Conference* (pp. 55–69). Loughborough, UK: Loughborough University.

Chevalier, S. (2007). Speech interaction with Saybot player, a CALL software to help Chinese learners of English. In *Proceedings of the International Speech Communication Association special interest group on speech and language technology in education* (pp. 37-40). Farmington, PA: International Speech Communication Association.

Chodorow, M., & Burstein, J. B. (2004). *Beyond essay length: Evaluating e-rater®'s performance on TOEFL® essays* (ETS Research Report No. RR-04-04). Princeton, NJ: ETS.

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

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

Franco, H., Abrash, V., Precoda, K., Bratt, H., Rao, R., Butzberger, J., Rossier, R., & Cesari, F. (2000). The SRI EduSpeak™ system: Recognition and pronunciation scoring for language learning. In *Proceedings of InSTILL (Integrating Speech Technology in Language Learning)* (pp. 123–128). Scotland: University of Abertay, Dundee.

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

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

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

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

Hughes, D. C., & Keeling, B. (1984). The use of model essays to reduce context effects in essay scoring. *Journal of Educational Measurement, 21*, 277–81.

Landauer, T. K., Laham, D., & Foltz, P. W. (2003). Automated scoring and annotation of essays with the Intelligent Essay Assessor. In M. D. Shermis & J. C. Burstein (Eds.), *Automated essay scoring: A cross-disciplinary perspective* (pp. 87–112). Hillsdale, NJ: Lawrence Erlbaum.

Leacock, C., & Chodorow, M. (2003). C-rater: Scoring of short-answer questions. *Computers and the Humanities, 37*(4), 389–405.

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

Mitchell, T., Russell, T., Broomhead, P., & Aldridge, N. (2002). Towards robust computerized marking of free-text responses. In *Proceedings of the sixth international computer assisted assessment conference* (pp. 233–249). UK: Loughborough University.

Page, E. B. (1966). The imminence of grading essays by computer. *Phi Delta Kappan, 48,* 238–243.

Page, E. B. (1968). The use of the computer in analyzing student essays. *International Review of Education, 14*(2), 210–225.

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

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

Ramineni, C., Davey, T., & Weng, V. (2010). Statistical evaluation and integration of a new positive feature for e-rater v10.1 (unpublished internal report). Princeton, NJ: ETS.

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

Risse, T. (2007, September). *Testing and assessing mathematical skills by a script based system.* Paper presented at the 10th international conference on interactive computer aided learning, Villach, Austria.

Rudner, L.M., Garcia, V., & Welch, C. (2006). An evaluation of IntelliMetric essay scoring system. *The Journal of Technology, Learning and Assessment*, *4*(4).

Sargeant, J., Wood, M. M., & Anderson, S. M. (2004). A human-computer collaborative approach to the marking of free text answers. In *Proceedings of the 8th international CAA conference* (pp. 361–370). UK: Loughborough University.

Shermis, M. D., & Burstein, J. C. (2003). *Automated essay scoring: A cross-disciplinary perspective*. Hillsdale, NJ: Lawrence Erlbaum.

Singley, M. K., & Bennett, R. E. (1998). *Validation and extension of the mathematical expression response type: Applications of schema theory to automatic scoring and item generation in mathematics* (GRE Board Professional Report No. 93–24P). Princeton, NJ: ETS.

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

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

Sukkarieh, J. Z., & Pulman, S. G. (2005). Information extraction and machine learning: Auto-marking short free text responses to science questions. In *Proceedings of the 12th International Conference on Artificial Intelligence in Education (AIED)* (pp. 629–637). Amsterdam, The Netherlands: IOS Press.

Williamson, D. M. & Davey, T. (2007). *Principles and processes for automated scoring: A summary of current policy, procedures and future work* (ETS Statistical Report No. SR-2009-061). Princeton, NJ: ETS.

Williamson, D. M. W., Attali, Y., Bridgeman, B., Davey, T., Ramineni, C., Trapani, C. S., & Weng, V. (2009, December). *Recommendation to use e-rater for TOEFL integrated prompts.* Unpublished manuscript.

Williamson, D. M. W., Attali, Y., Carey, J., Powers, D., Trapani, C. S., & Vezzu, S. (2007). *Recommendation to use e-rater for TOEFL operational scoring*. Unpublished manuscript.

Williamson, D. M. W., Bridgeman, B., Davey, T., Ramineni, C., Trapani, C. S., & Weng, V. (2009, April). *Recommendation to use e-rater for TOEFL independent prompts*. Unpublished manuscript.

Williamson, D. M. W., Trapani, C. S., & Weng, V. (2008). *Recommendation to shift to generic models.* Unpublished manuscript.

Xi, X., Higgins, D., Zechner, K., & Williamson, D. M. (2008). *Automated scoring of spontaneous speech using SpeechRater v1.0* (ETS Research Report No. RR-08-62). Princeton, NJ: ETS.

Zechner, K., & Bejar, I. (2006). Towards automatic scoring of non-native spontaneous speech. In *Proceedings of the human language technology conference of the North American chapter of the ACL* (pp. 216–223). New York, NY.

Appendix A

Table A1

TOEFL Scoring Guide

Score	TOEFL scoring guide (independent)	TOEFL scoring guide (integrated)
5	An essay at this level largely accomplishes all of the following: • effectively addresses the topic and task • is well organized and well developed, using clearly appropriate explanations, exemplifications, and/or details • displays unity, progression, and coherence • displays consistent facility in the use of language, demonstrating syntactic variety, appropriate word choice, and idiomaticity, though it may have minor lexical or grammatical errors	A response at this level successfully selects the important information from the lecture and coherently and accurately presents this information in relation to the relevant information presented in the reading. The response is well organized, and occasional language errors that are present do not result in inaccurate or imprecise presentation of content or connections.
4	An essay at this level largely accomplishes all of the following: • addresses the topic and task well, though some points may not be fully elaborated • is generally well organized and well developed, using appropriate and sufficient explanations, exemplifications, and/or details • displays unity, progression, and coherence, though it may contain occasional redundancy, digression, or unclear connections • displays facility in the use of language, demonstrating syntactic variety and range of vocabulary, though it will probably have occasional noticeable minor errors in structure, word form, or use of idiomatic language that do not interfere with meaning	A response at this level is generally good in selecting the important information from the lecture and in coherently and accurately presenting this information in relation to the relevant information in the reading, but it may have minor omission, inaccuracy, vagueness, or imprecision of some content from the lecture or in connection to points made in the reading. A response is also scored at this level if it has more frequent or noticeable minor language errors, as long as such usage and grammatical structures do not result in anything more than an occasional lapse of clarity or in the connection of ideas.

Score	TOEFL scoring guide (independent)	TOEFL scoring guide (integrated)
3	An essay at this level is marked by one or more of the following: • addresses the topic and task using somewhat developed explanations, exemplifications, and/or details • displays unity, progression, and coherence, though connection of ideas may be occasionally obscured • may demonstrate inconsistent facility in sentence formation and word choice that may result in lack of clarity and occasionally obscure meaning • may display accurate, but limited range of syntactic structures and vocabulary	A response at this level contains some important information from the lecture and conveys some relevant connection to the reading, but it is marked by one or more of the following: Although the overall response is definitely oriented to the task, it conveys only vague, global, unclear, or somewhat imprecise connection of the points made in the lecture to points made in the reading. The response may omit one major key point made in the lecture. Some key points made in the lecture or the reading, or connections between the two, may be incomplete, inaccurate, or imprecise. Errors of usage and/or grammar may be more frequent or may result in noticeably vague expressions or obscured meanings in conveying ideas and connections.
2	An essay at this level may reveal one or more of the following weaknesses: • limited development in response to the topic and task • inadequate organization or connection of ideas • inappropriate or insufficient exemplifications, explanations, or details to support or illustrate generalizations in response to the task • a noticeably inappropriate choice of words or word forms • an accumulation of errors in sentence structure and/or usage	A response at this level contains some relevant information from the lecture, but is marked by significant language difficulties or by significant omission or inaccuracy of important ideas from the lecture or in the connections between the lecture and the reading; a response at this level is marked by one or more of the following: The response significantly misrepresents or completely omits the overall connection between the lecture and the reading. The response significantly omits or significantly misrepresents important points made in the lecture. The response contains language errors or expressions that largely obscure connections or meaning at key junctures, or that would likely obscure understanding of key ideas for a reader not already familiar with the reading and the lecture.

Score	TOEFL scoring guide (independent)	TOEFL scoring guide (integrated)
1	An essay at this level is seriously flawed by one or more of the following weaknesses: • serious disorganization or underdevelopment • little or no detail, or irrelevant specifics, or questionable responsiveness to the task • serious and frequent errors in sentence structure or usage	A response at this level is marked by one or more of the following: The response provides little or no meaningful or relevant coherent content from the lecture. The language level of the response is so low that it is difficult to derive meaning.
0	An essay at this level merely copies words from the topic, rejects the topic, or is otherwise not connected to the topic, is written in a foreign language, consists of keystroke characters, or is blank.	A response at this level merely copies sentences from the reading, rejects the topic or is otherwise not connects to the topic, is written in a foreign language, consists of keystroke characters, or is blank.

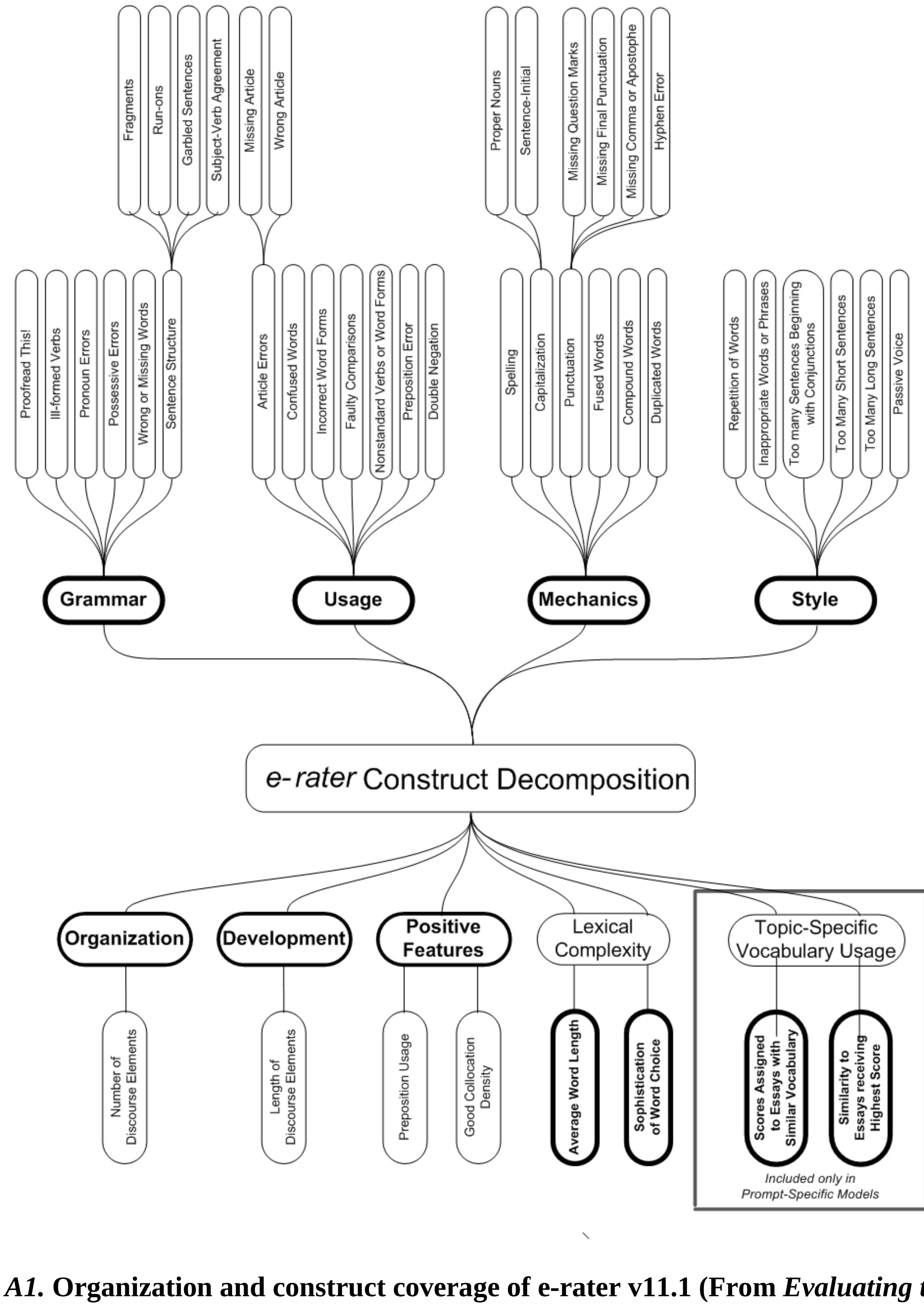

Figure A1. **Organization and construct coverage of e-rater v11.1 (From *Evaluating the Construct Coverage of the e-rater Scoring Engine* (ETS Research Report No. RR-09-01; p. 9), by T. Quinlan, D. Higgins, and S. Wolff, 2009, Princeton, NJ: ETS. Copyright 2009 by Educational Testing Service. Adapted with permission.**

Table A2

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
Pearson correlation between e-rater score and human score	Correlation less than 0.7
Standardized difference between e-rater score and human score	Standardized difference greater than 0.15 in absolute value
Notable reduction in quadratic-weighted kappa or correlation from human/human to automated/human	Decline in quadratic-weighted kappa or correlation of greater than 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 evaluated to four decimal values for flagging.

Appendix B

Table B1

Agreement With Human Scores on Independent Prompts: Generic (G) Model

Independent–G		H1 by H2									H1 by e-rater (rounded to integers)							H1 by e-rater (unrounded)				Degradation	
		H1		H2		Stats					e-rater							e-rater		Stats		Wtd kappa	*R*
Prompt	*N*	M	SD	M	SD	Std diff	Kappa	Wtd kappa	% agree	% adj agree	*R*	M	SD	Kappa	Wtd kappa	% agree	% adj agree	M	SD	Std diff	*R*	H1 by e-rater rounded – H1 by H2	H1 by e-rater unrounded – H1 by H2
VC116880	3,580	3.34	0.88	3.34	0.88	0.01	0.39	0.70	59	98	0.70	3.49	0.87	0.36	0.68	56	98	3.49	0.83	0.18	0.74	-0.02	0.04
VC143964	3,447	3.31	0.86	3.31	0.86	0.00	0.40	0.69	60	98	0.69	3.29	0.90	0.40	0.71	60	99	3.28	0.87	-0.03	0.76	0.02	0.06
VC202818	3,471	3.29	0.85	3.32	0.83	0.03	0.37	0.67	59	98	0.67	3.40	0.84	0.36	0.66	57	98	3.40	0.80	0.13	0.72	0.00	0.05
VC207963	3,498	3.26	0.86	3.28	0.84	0.02	0.38	0.69	59	99	0.69	3.34	0.86	0.39	0.70	59	99	3.33	0.84	0.08	0.75	0.02	0.06
VC213547	3,500	3.28	0.85	3.29	0.86	0.01	0.39	0.68	60	98	0.68	3.31	0.91	0.39	0.70	58	98	3.32	0.88	0.04	0.74	0.02	0.06
VC237267	3,539	3.23	0.86	3.24	0.86	0.00	0.40	0.70	60	98	0.70	3.35	0.89	0.40	0.71	60	98	3.35	0.88	0.14	0.75	0.01	0.05
VC243618	3,478	3.35	0.82	3.38	0.82	0.04	0.39	0.68	61	99	0.68	3.43	0.87	0.38	0.69	59	99	3.43	0.83	0.10	0.74	0.02	0.06
VC251653	3,305	3.32	0.89	3.34	0.90	0.02	0.42	0.72	60	98	0.72	3.20	0.89	0.39	0.71	58	98	3.19	0.87	-0.14	0.75	-0.01	0.03
VC262732	3,501	3.32	0.85	3.32	0.84	0.00	0.41	0.71	62	99	0.71	3.31	0.90	0.42	0.73	61	99	3.31	0.87	-0.02	0.77	0.02	0.06
VC263915	3,510	3.31	0.85	3.30	0.85	-0.02	0.42	0.70	62	98	0.70	3.41	0.89	0.37	0.69	57	99	3.41	0.85	0.11	0.74	-0.01	0.04
VC281990	3,520	3.35	0.89	3.36	0.86	0.02	0.37	0.68	57	98	0.68	3.37	0.87	0.39	0.70	58	99	3.36	0.85	0.02	0.75	0.02	0.07
VC288976	3,505	3.31	0.87	3.33	0.87	0.02	0.35	0.67	57	98	0.67	3.49	0.90	0.39	0.71	58	98	3.49	0.87	0.20	0.76	0.03	0.09
VC298109	3,486	3.36	0.88	3.35	0.89	-0.01	0.42	0.72	61	98	0.72	3.56	0.90	0.38	0.72	57	99	3.56	0.87	0.23	0.77	-0.01	0.05
VC307841	3,253	3.36	0.81	3.36	0.82	-0.01	0.41	0.69	62	99	0.69	3.38	0.84	0.39	0.68	60	99	3.38	0.81	0.02	0.72	-0.01	0.04
VC347396	3,486	3.42	0.86	3.43	0.86	0.00	0.40	0.69	60	98	0.69	3.52	0.85	0.41	0.71	60	99	3.52	0.82	0.12	0.76	0.02	0.07

Independent–G		H1 by H2									H1 by e-rater (rounded to integers)							H1 by e-rater (unrounded)				Degradation	
		H1		H2		Stats						e-rater						e-rater		Stats		Wtd kappa	*R*
Prompt	*N*	M	SD	M	SD	Std diff	Kappa	Wtd kappa	% agree	% adj agree	*R*	M	SD	Kappa	Wtd kappa	% agree	% adj agree	M	SD	Std diff	*R*	H1 by e-rater rounded – H1 by H2	H1 by e-rater unrounded – H1 by H2
VC350994	3,423	3.37	0.85	3.38	0.83	0.00	0.38	0.67	59	98	0.67	3.32	0.83	0.39	0.69	60	99	3.32	0.79	-0.07	0.72	0.02	0.05
VC358929	3,462	3.34	0.89	3.31	0.88	-0.03	0.38	0.69	58	98	0.69	3.36	0.87	0.40	0.71	59	99	3.36	0.83	0.02	0.75	0.02	0.06
VC370506	3,430	3.43	0.85	3.45	0.86	0.03	0.38	0.68	58	98	0.68	3.38	0.84	0.38	0.68	59	99	3.37	0.80	-0.07	0.73	0.00	0.05
VC375512	3,624	3.25	0.85	3.26	0.86	0.01	0.38	0.67	59	98	0.67	3.23	0.88	0.41	0.70	60	98	3.23	0.85	-0.02	0.74	0.03	0.07
VC378148	3,480	3.29	0.84	3.32	0.84	0.04	0.36	0.67	58	98	0.67	3.39	0.83	0.39	0.68	59	99	3.38	0.79	0.12	0.73	0.01	0.06
VC383013	3,595	3.29	0.88	3.31	0.87	0.03	0.39	0.69	59	98	0.69	3.28	0.88	0.38	0.70	58	98	3.27	0.85	-0.03	0.74	0.01	0.05
VC390617	3,462	3.31	0.89	3.30	0.87	-0.02	0.41	0.72	61	98	0.72	3.36	0.92	0.42	0.73	60	99	3.36	0.89	0.05	0.77	0.02	0.06
VC391218	3,467	3.38	0.86	3.38	0.87	0.00	0.38	0.68	58	98	0.68	3.37	0.86	0.41	0.71	60	99	3.37	0.82	-0.01	0.75	0.03	0.06
VC400012	3,387	3.39	0.82	3.39	0.81	0.01	0.36	0.64	58	98	0.64	3.27	0.84	0.38	0.68	59	99	3.27	0.81	-0.14	0.73	0.04	0.08
VC404033	3,416	3.42	0.85	3.41	0.85	-0.01	0.38	0.67	59	98	0.67	3.24	0.83	0.35	0.68	57	99	3.24	0.80	-0.22	0.75	0.00	0.07
VC430895	3,440	3.33	0.83	3.37	0.83	0.05	0.40	0.68	60	99	0.68	3.24	0.83	0.39	0.68	60	99	3.24	0.78	-0.12	0.73	0.00	0.05
VC431370	3,410	3.32	0.84	3.34	0.84	0.03	0.39	0.67	60	98	0.67	3.28	0.83	0.40	0.69	61	99	3.28	0.79	-0.04	0.74	0.02	0.06
VC457348	3,659	3.37	0.84	3.39	0.83	0.02	0.39	0.68	60	98	0.68	3.32	0.89	0.37	0.69	58	99	3.32	0.86	-0.06	0.73	0.01	0.05
VC467512	3,403	3.40	0.86	3.39	0.86	-0.01	0.38	0.68	58	98	0.68	3.35	0.86	0.39	0.70	59	99	3.35	0.82	-0.06	0.75	0.02	0.07
VC467756	3,474	3.40	0.82	3.39	0.84	0.00	0.39	0.67	60	98	0.67	3.28	0.83	0.37	0.67	59	99	3.28	0.78	-0.14	0.72	0.00	0.05
VC506581	3,735	3.35	0.84	3.33	0.84	-0.02	0.40	0.69	61	98	0.69	3.47	0.87	0.37	0.68	58	99	3.47	0.84	0.14	0.74	0.00	0.05
VC508541	3,524	3.26	0.84	3.27	0.84	0.01	0.42	0.69	62	98	0.69	3.26	0.87	0.39	0.70	59	99	3.25	0.84	-0.01	0.74	0.00	0.04
VC515193	3,402	3.41	0.85	3.42	0.86	0.02	0.42	0.70	61	99	0.70	3.38	0.85	0.41	0.70	60	99	3.39	0.81	-0.02	0.75	0.00	0.05

Independent–G		H1 by H2										H1 by e-rater (rounded to integers)						H1 by e-rater (unrounded)				Degradation	
		H1		H2		Stats						e-rater						e-rater		Stats		Wtd kappa	*R*
Prompt	*N*	M	SD	M	SD	Std diff	Kappa	Wtd kappa	% agree	% adj agree	*R*	M	SD	Kappa	Wtd kappa	% agree	% adj agree	M	SD	Std diff	*R*	H1 by e-rater rounded – H1 by H2	H1 by e-rater unrounded – H1 by H2
VC517082	3,490	3.43	0.85	3.42	0.83	-0.01	0.38	0.67	59	98	0.67	3.47	0.81	0.36	0.67	58	99	3.46	0.77	0.04	0.72	0.00	0.05
VC595348	3,442	3.44	0.83	3.43	0.84	0.00	0.40	0.68	61	99	0.68	3.31	0.84	0.38	0.68	59	99	3.31	0.81	-0.16	0.73	0.00	0.04
VC621100	3,439	3.41	0.86	3.43	0.85	0.02	0.40	0.69	60	98	0.69	3.39	0.85	0.39	0.69	59	99	3.39	0.80	-0.03	0.74	0.00	0.04
VC627181	3,664	3.38	0.83	3.40	0.83	0.02	0.42	0.70	62	99	0.70	3.44	0.86	0.40	0.70	60	99	3.43	0.82	0.06	0.75	-0.01	0.04
VC684177	3,440	3.41	0.82	3.42	0.82	0.02	0.37	0.66	60	98	0.66	3.42	0.82	0.38	0.67	59	99	3.41	0.78	0.01	0.73	0.02	0.07
Average	3,483	3.35	0.85	3.35	0.85	0.01	0.39	0.69	60	98	0.69	3.36	0.86	0.39	0.69	59	99	3.36	0.83	0.01	0.74	0.01	0.06

Note. N is average across all the prompts. Shaded cells indicate values that fail to meet the thresholds listed in Table A2.

adj = adjacent, H1 = human 1, H2 = human 2, std diff = standardized difference, wtd = weighted.

Table B2

Agreement With Human Scores on Independent Prompts: Prompt-Specific (PS) Model

Independent–PS		H1 by H2										H1 by e-rater (rounded to integers)						H1 by e-rater (unrounded)				Degradation	
		H1		H2		Stats						e-rater						e-rater		Stats		Wtd kappa	*R*
Prompt	*N*	M	SD	M	SD	Std diff	Kappa	Wtd kappa	% agree	% adj agree	*R*	M	SD	Kappa	Wtd kappa	% agree	% adj agree	M	SD	Std diff	*R*	H1 by e-rater rounded – H1 by H2	H1 by e-rater unrounded – H1 by H2
VC116880	3,580	3.34	0.88	3.34	0.88	0.01	0.39	0.70	59	98	0.70	3.35	0.89	0.37	0.69	57	98	3.36	0.85	0.02	0.74	0.00	0.04
VC143964	3,447	3.31	0.86	3.31	0.86	0.00	0.40	0.69	60	98	0.69	3.30	0.87	0.42	0.71	61	99	3.30	0.84	-0.01	0.76	0.02	0.06
VC202818	3,471	3.29	0.85	3.32	0.83	0.03	0.37	0.67	59	98	0.67	3.32	0.88	0.37	0.69	57	99	3.32	0.84	0.03	0.73	0.02	0.06
VC207963	3,498	3.26	0.86	3.28	0.84	0.02	0.38	0.69	59	99	0.69	3.33	0.84	0.41	0.71	61	99	3.32	0.82	0.06	0.76	0.02	0.07
VC213547	3,500	3.28	0.85	3.29	0.86	0.01	0.39	0.68	60	98	0.68	3.25	0.92	0.40	0.71	59	98	3.26	0.89	-0.02	0.75	0.03	0.07
VC237267	3,539	3.23	0.86	3.24	0.86	0.00	0.40	0.70	60	98	0.70	3.30	0.87	0.41	0.71	61	99	3.30	0.85	0.08	0.76	0.02	0.06
VC243618	3,478	3.35	0.82	3.38	0.82	0.04	0.39	0.68	61	99	0.68	3.35	0.87	0.40	0.70	60	99	3.35	0.83	0.00	0.74	0.02	0.07
VC251653	3,305	3.32	0.89	3.34	0.90	0.02	0.42	0.72	60	98	0.72	3.37	0.95	0.40	0.73	58	99	3.37	0.91	0.05	0.77	0.01	0.04
VC262732	3,501	3.32	0.85	3.32	0.84	0.00	0.41	0.71	62	99	0.71	3.34	0.84	0.42	0.71	61	99	3.34	0.81	0.01	0.76	0.01	0.06
VC263915	3,510	3.31	0.85	3.30	0.85	-0.02	0.42	0.70	62	98	0.70	3.35	0.90	0.39	0.71	58	99	3.34	0.86	0.04	0.75	0.01	0.05
VC281990	3,520	3.35	0.89	3.36	0.86	0.02	0.37	0.68	57	98	0.68	3.34	0.84	0.41	0.71	60	99	3.35	0.81	0.00	0.76	0.03	0.07
VC288976	3,505	3.31	0.87	3.33	0.87	0.02	0.35	0.67	57	98	0.67	3.37	0.94	0.40	0.72	59	98	3.37	0.92	0.07	0.76	0.05	0.09
VC298109	3,486	3.36	0.88	3.35	0.89	-0.01	0.42	0.72	61	98	0.72	3.40	0.88	0.43	0.74	62	99	3.39	0.85	0.04	0.78	0.02	0.06
VC307841	3,253	3.36	0.81	3.36	0.82	-0.01	0.41	0.69	62	99	0.69	3.38	0.79	0.41	0.68	62	99	3.38	0.75	0.02	0.73	0.00	0.05
VC347396	3,486	3.42	0.86	3.43	0.86	0.00	0.40	0.69	60	98	0.69	3.44	0.88	0.42	0.72	60	99	3.44	0.86	0.02	0.76	0.03	0.07
VC350994	3,423	3.37	0.85	3.38	0.83	0.00	0.38	0.67	59	98	0.67	3.37	0.89	0.38	0.69	58	98	3.36	0.85	-0.01	0.73	0.02	0.06

Independent–PS		H1 by H2										H1 by e-rater (rounded to integers)						H1 by e-rater (unrounded)				Degradation	
		H1		H2		Stats						e-rater						e-rater		Stats		Wtd kappa	*R*
Prompt	*N*	M	SD	M	SD	Std diff	Kappa	Wtd kappa	% agree	% adj agree	*R*	M	SD	Kappa	Wtd kappa	% agree	% adj agree	M	SD	Std diff	*R*	H1 by e-rater rounded – H1 by H2	H1 by e-rater unrounded – H1 by H2
VC358929	3,462	3.34	0.89	3.31	0.88	-0.03	0.38	0.69	58	98	0.69	3.35	0.85	0.41	0.71	60	99	3.35	0.82	0.01	0.75	0.02	0.06
VC370506	3,430	3.43	0.85	3.45	0.86	0.03	0.38	0.68	58	98	0.68	3.46	0.85	0.38	0.69	59	99	3.46	0.81	0.04	0.73	0.01	0.05
VC375512	3,624	3.25	0.85	3.26	0.86	0.01	0.38	0.67	59	98	0.67	3.25	0.86	0.40	0.70	60	98	3.25	0.82	0.00	0.74	0.03	0.07
VC378148	3,480	3.29	0.84	3.32	0.84	0.04	0.36	0.67	58	98	0.67	3.28	0.91	0.39	0.69	58	98	3.27	0.88	-0.02	0.73	0.02	0.07
VC383013	3,595	3.29	0.88	3.31	0.87	0.03	0.39	0.69	59	98	0.69	3.30	0.89	0.39	0.70	59	98	3.29	0.86	0.00	0.75	0.01	0.06
VC390617	3,462	3.31	0.89	3.30	0.87	-0.02	0.41	0.72	61	98	0.72	3.33	0.91	0.43	0.74	61	99	3.33	0.89	0.02	0.78	0.03	0.07
VC391218	3,467	3.38	0.86	3.38	0.87	0.00	0.38	0.68	58	98	0.68	3.34	0.85	0.40	0.70	60	99	3.34	0.81	-0.05	0.75	0.02	0.07
VC400012	3,387	3.39	0.82	3.39	0.81	0.01	0.36	0.64	58	98	0.64	3.37	0.86	0.38	0.68	59	99	3.37	0.83	-0.02	0.72	0.04	0.07
VC404033	3,416	3.42	0.85	3.41	0.85	-0.01	0.38	0.67	59	98	0.67	3.38	0.90	0.38	0.71	58	99	3.39	0.87	-0.04	0.75	0.04	0.08
VC430895	3,440	3.33	0.83	3.37	0.83	0.05	0.40	0.68	60	99	0.68	3.34	0.87	0.40	0.70	60	99	3.34	0.84	0.01	0.74	0.01	0.05
VC431370	3,410	3.32	0.84	3.34	0.84	0.03	0.39	0.67	60	98	0.67	3.34	0.86	0.41	0.71	61	99	3.33	0.83	0.02	0.75	0.04	0.07
VC457348	3,659	3.37	0.84	3.39	0.83	0.02	0.39	0.68	60	98	0.68	3.40	0.84	0.40	0.70	60	99	3.40	0.81	0.03	0.74	0.02	0.06
VC467512	3,403	3.40	0.86	3.39	0.86	-0.01	0.38	0.68	58	98	0.68	3.39	0.89	0.40	0.71	60	99	3.39	0.85	-0.01	0.75	0.03	0.07
VC467756	3,474	3.40	0.82	3.39	0.84	0.00	0.39	0.67	60	98	0.67	3.32	0.84	0.38	0.68	59	99	3.33	0.79	-0.09	0.72	0.00	0.05
VC506581	3,735	3.35	0.84	3.33	0.84	-0.02	0.40	0.69	61	98	0.69	3.33	0.90	0.39	0.70	58	99	3.34	0.87	-0.01	0.75	0.02	0.06
VC508541	3,524	3.26	0.84	3.27	0.84	0.01	0.42	0.69	62	98	0.69	3.33	0.84	0.40	0.69	60	99	3.32	0.81	0.08	0.75	0.00	0.05
VC515193	3,402	3.41	0.85	3.42	0.86	0.02	0.42	0.70	61	99	0.70	3.44	0.84	0.43	0.71	61	99	3.45	0.80	0.05	0.76	0.01	0.06
VC517082	3,490	3.43	0.85	3.42	0.83	-0.01	0.38	0.67	59	98	0.67	3.43	0.83	0.38	0.68	59	99	3.44	0.79	0.00	0.72	0.01	0.05

Independent–PS		H1 by H2										H1 by e-rater (rounded to integers)						H1 by e-rater (unrounded)				Degradation	
		H1		H2		Stats						e-rater						e-rater		Stats		Wtd kappa	*R*
Prompt	*N*	M	SD	M	SD	Std diff	Kappa	Wtd kappa	% agree	% adj agree	*R*	M	SD	Kappa	Wtd kappa	% agree	% adj agree	M	SD	Std diff	*R*	H1 by e-rater rounded – H1 by H2	H1 by e-rater unrounded – H1 by H2
VC595348	3,442	3.44	0.83	3.43	0.84	0.00	0.40	0.68	61	99	0.68	3.42	0.86	0.40	0.69	60	99	3.42	0.81	-0.03	0.74	0.01	0.06
VC621100	3,439	3.41	0.86	3.43	0.85	0.02	0.40	0.69	60	98	0.69	3.44	0.86	0.39	0.70	59	99	3.44	0.83	0.03	0.74	0.01	0.05
VC627181	3,664	3.38	0.83	3.40	0.83	0.02	0.42	0.70	62	99	0.70	3.43	0.85	0.43	0.71	62	99	3.42	0.82	0.04	0.76	0.01	0.05
VC684177	3,440	3.41	0.82	3.42	0.82	0.02	0.37	0.66	60	98	0.66	3.42	0.83	0.39	0.69	60	99	3.42	0.79	0.02	0.74	0.03	0.08
Average	3,483	3.35	0.85	3.35	0.85	0.01	0.39	0.69	60	98	0.69	3.36	0.87	0.40	0.70	60	99	3.36	0.84	0.01	0.75	0.02	0.06

Note. *N* is average across all the prompts. Shaded cells indicate values that fail to meet the thresholds listed in Table A2.

adj = adjacent, H1 = human 1, H2 = human 2, std diff = standardized difference, wtd = weighted.

Table B3

Agreement With Human Scores on Integrated Prompts: Generic (G) Model

Independent–G		H1 by H2										H1 by e-rater (rounded to integers)						H1 by e-rater (unrounded)				Degradation	
		H1		H2		Stats						e-rater						e-rater		Stats		Wtd kappa	*R*
Prompt	*N*	M	SD	M	SD	Std diff	Kappa	Wtd kappa	% agree	% adj agree	*R*	M	SD	Kappa	Wtd kappa	% agree	% adj agree	M	SD	Std diff	*R*	H1 by e-rater rounded – H1 by H2	H1 by e-rater unrounded – H1 by H2
VC157528	3,309	2.78	1.26	2.77	1.24	-0.01	0.49	0.83	61	96	0.83	2.93	1.10	0.19	0.55	38	83	2.90	1.11	0.10	0.58	-0.28	-0.26
VC176929	3,408	3.00	1.07	2.98	1.08	-0.01	0.47	0.79	61	97	0.79	3.05	1.11	0.20	0.58	40	88	3.04	1.13	0.04	0.60	-0.21	-0.19
VC189417	3,292	2.89	1.18	2.90	1.18	0.01	0.48	0.82	60	97	0.82	2.94	1.12	0.21	0.58	39	86	2.91	1.14	0.01	0.60	-0.24	-0.21
VC214325	3,347	3.17	1.10	3.18	1.11	0.01	0.45	0.79	59	97	0.79	3.26	1.10	0.23	0.61	42	89	3.25	1.12	0.07	0.63	-0.18	-0.16
VC214330	3,296	2.80	1.21	2.81	1.21	0.01	0.49	0.83	60	97	0.83	2.98	1.14	0.22	0.60	40	86	2.96	1.17	0.13	0.63	-0.23	-0.20
VC229874	3,079	3.10	1.15	3.10	1.15	0.00	0.49	0.81	62	96	0.81	3.01	1.11	0.19	0.57	39	86	2.99	1.15	-0.09	0.60	-0.23	-0.21
VC243308	3,343	2.90	1.24	2.92	1.24	0.02	0.52	0.84	62	97	0.84	3.16	1.13	0.22	0.61	39	86	3.15	1.16	0.21	0.65	-0.23	-0.20
VC243309	3,375	3.22	1.26	3.23	1.27	0.01	0.53	0.86	64	97	0.86	3.16	1.09	0.20	0.62	39	87	3.14	1.10	-0.07	0.65	-0.25	-0.21
VC257573	3,132	3.23	1.21	3.21	1.22	-0.02	0.46	0.82	58	97	0.82	3.04	1.12	0.22	0.61	40	87	3.02	1.14	-0.17	0.65	-0.21	-0.18
VC265780	3,370	3.14	1.21	3.14	1.19	0.00	0.49	0.83	61	97	0.83	2.87	1.08	0.22	0.63	40	88	2.85	1.11	-0.25	0.67	-0.19	-0.15
VC286667	3,332	3.14	1.17	3.13	1.16	-0.01	0.51	0.84	63	98	0.84	3.29	1.12	0.22	0.60	41	86	3.29	1.14	0.13	0.62	-0.24	-0.21
VC315648	3,273	3.12	1.30	3.11	1.29	-0.01	0.49	0.85	60	97	0.85	3.11	1.14	0.22	0.64	40	86	3.09	1.16	-0.03	0.66	-0.21	-0.19
VC315650	3,202	3.07	1.31	3.03	1.30	-0.03	0.52	0.86	62	97	0.86	2.93	1.09	0.20	0.60	37	85	2.91	1.10	-0.13	0.63	-0.26	-0.23
VC315652	3,255	3.23	1.20	3.24	1.23	0.01	0.47	0.83	59	97	0.83	3.35	1.11	0.20	0.62	39	88	3.35	1.11	0.11	0.64	-0.21	-0.19
VC337838	3,285	3.25	1.01	3.24	1.01	0.00	0.41	0.75	57	97	0.75	3.25	1.06	0.19	0.55	41	89	3.25	1.05	0.00	0.58	-0.20	-0.17
VC354913	3,423	3.06	1.15	3.08	1.16	0.01	0.50	0.83	62	98	0.83	2.92	1.13	0.23	0.63	41	89	2.90	1.16	-0.14	0.66	-0.20	-0.18

Independent–G		H1 by H2										H1 by e-rater (rounded to integers)						H1 by e-rater (unrounded)				Degradation	
		H1		H2		Stats						e-rater						e-rater		Stats		Wtd kappa	*R*
Prompt	*N*	M	SD	M	SD	Std diff	Kappa	Wtd kappa	% agree	% adj agree	*R*	M	SD	Kappa	Wtd kappa	% agree	% adj agree	M	SD	Std diff	*R*	H1 by e-rater rounded – H1 by H2	H1 by e-rater unrounded – H1 by H2
VC357067	3,236	3.03	1.28	3.01	1.25	-0.01	0.48	0.84	59	96	0.84	2.71	1.08	0.19	0.60	37	84	2.68	1.10	-0.29	0.65	-0.24	-0.18
VC358128	3,436	3.09	1.17	3.10	1.17	0.01	0.45	0.80	57	96	0.80	3.16	1.15	0.21	0.60	39	87	3.14	1.17	0.05	0.63	-0.20	-0.17
VC364389	3,301	3.01	1.21	3.01	1.22	0.01	0.53	0.85	64	98	0.85	2.73	1.01	0.16	0.52	36	83	2.70	1.03	-0.27	0.57	-0.34	-0.29
VC373909	3,240	2.99	1.26	2.99	1.25	0.00	0.51	0.85	62	97	0.85	3.24	1.12	0.20	0.59	38	85	3.23	1.12	0.20	0.63	-0.26	-0.23
VC373911	3,290	2.97	1.05	2.96	1.04	0.00	0.44	0.77	60	97	0.77	2.94	1.08	0.19	0.56	40	88	2.93	1.09	-0.04	0.58	-0.22	-0.20
VC374076	3,440	3.22	1.10	3.21	1.10	0.00	0.44	0.79	58	97	0.79	3.09	1.12	0.20	0.60	40	88	3.08	1.14	-0.12	0.63	-0.19	-0.16
VC374333	3,510	3.17	1.21	3.16	1.19	0.00	0.46	0.82	59	97	0.82	3.25	1.09	0.19	0.59	39	86	3.25	1.11	0.07	0.61	-0.24	-0.21
VC389573	3,279	3.00	1.20	3.01	1.21	0.00	0.48	0.83	60	97	0.83	3.08	1.13	0.23	0.64	41	89	3.06	1.15	0.05	0.66	-0.20	-0.18
VC389578	3,245	3.15	1.10	3.16	1.10	0.01	0.51	0.82	64	98	0.82	2.99	1.06	0.17	0.55	38	87	2.98	1.06	-0.16	0.58	-0.27	-0.24
VC389592	3,294	3.11	1.25	3.11	1.25	0.01	0.48	0.84	60	97	0.84	3.04	1.12	0.26	0.65	43	88	3.03	1.15	-0.06	0.68	-0.19	-0.17
VC389593	3,238	3.08	1.20	3.09	1.20	0.01	0.42	0.79	55	95	0.79	2.92	1.06	0.16	0.57	36	86	2.91	1.07	-0.16	0.60	-0.22	-0.19
VC399764	3,306	3.23	1.29	3.22	1.29	-0.01	0.52	0.86	62	97	0.86	3.19	1.05	0.18	0.59	36	85	3.19	1.04	-0.03	0.61	-0.28	-0.25
VC400185	3,598	2.98	1.13	2.98	1.12	0.00	0.48	0.81	61	97	0.81	2.82	1.06	0.22	0.61	42	89	2.80	1.08	-0.16	0.65	-0.20	-0.17
VC400187	3,331	3.11	1.18	3.10	1.17	-0.01	0.52	0.84	63	97	0.84	3.19	1.12	0.20	0.60	38	87	3.18	1.13	0.06	0.62	-0.24	-0.21
VC400188	3,204	2.65	1.33	2.65	1.33	0.00	0.56	0.88	66	98	0.88	2.96	1.10	0.16	0.54	34	80	2.95	1.13	0.25	0.59	-0.34	-0.29
VC457888	3,300	3.28	1.11	3.29	1.11	0.01	0.49	0.82	62	97	0.82	3.19	1.09	0.20	0.59	40	88	3.19	1.10	-0.08	0.61	-0.23	-0.21
VC457890	3,265	3.20	1.26	3.18	1.24	-0.02	0.48	0.84	60	97	0.84	3.13	1.13	0.21	0.61	39	85	3.12	1.13	-0.06	0.63	-0.23	-0.21
VC457893	3,508	3.13	0.99	3.13	0.99	-0.01	0.48	0.78	64	98	0.78	3.32	1.11	0.19	0.55	40	87	3.31	1.12	0.17	0.58	-0.23	-0.20

Independent–G		H1 by H2									H1 by e-rater (rounded to integers)							H1 by e-rater (unrounded)				Degradation	
		H1		H2		Stats						e-rater						e-rater		Stats		Wtd kappa	*R*
Prompt	*N*	M	SD	M	SD	Std diff	Kappa	Wtd kappa	% agree	% adj agree	*R*	M	SD	Kappa	Wtd kappa	% agree	% adj agree	M	SD	Std diff	*R*	H1 by e-rater rounded – H1 by H2	H1 by e-rater unrounded – H1 by H2
VC457896	3,367	3.01	1.26	3.02	1.27	0.01	0.50	0.85	61	97	0.85	3.20	1.08	0.20	0.60	38	86	3.19	1.08	0.15	0.63	-0.25	-0.21
VC457898	3,300	3.33	1.16	3.31	1.16	-0.02	0.43	0.80	57	97	0.80	3.19	1.12	0.21	0.61	40	88	3.17	1.13	-0.14	0.64	-0.19	-0.16
VC472548	3,396	3.04	1.16	3.04	1.16	0.01	0.45	0.80	58	96	0.80	3.34	1.14	0.22	0.61	40	87	3.33	1.16	0.25	0.65	-0.19	-0.15
VC503849	3,220	3.03	1.19	3.02	1.20	-0.01	0.46	0.81	58	96	0.81	3.20	1.05	0.18	0.57	38	86	3.19	1.05	0.14	0.60	-0.24	-0.21
Average	3,316	3.08	1.19	3.07	1.19	0.00	0.48	0.82	60	97	0.82	3.08	1.10	0.20	0.59	39	87	3.07	1.12	-0.01	0.62	-0.23	-0.20

Note. *N* is average across all the prompts. Shaded cells indicate values that fail to meet the thresholds listed in Table A2.

adj = adjacent, H1 = human 1, H2 = human 2, std diff = standardized difference, wtd = weighted.

Table B4

Agreement With Human Scores on Integrated Prompts: Prompt-Specific (PS) Model

Independent–PS		H1 by H2										H1 by e-rater (rounded to integers)						H1 by e-rater (unrounded)				Degradation	
		H1		H2		Stats						e-rater						e-rater		Stats		Wtd kappa	*R*
Prompt	*N*	M	SD	M	SD	Std diff	Kappa	Wtd kappa	% agree	% adj agree	*R*	M	SD	Kappa	Wtd kappa	% agree	% adj agree	M	SD	Std diff	*R*	H1 by e-rater rounded – H1 by H2	H1 by e-rater unrounded – H1 by H2
VC157528	3,309	2.78	1.26	2.77	1.24	-0.01	0.49	0.83	61	96	0.83	2.81	1.17	0.30	0.68	46	89	2.78	1.22	-0.01	0.71	-0.15	-0.13
VC176929	3,408	3.00	1.07	2.98	1.08	-0.01	0.47	0.79	61	97	0.79	2.98	1.02	0.25	0.64	45	93	2.97	1.01	-0.02	0.68	-0.15	-0.12
VC189417	3,292	2.89	1.18	2.90	1.18	0.01	0.48	0.82	60	97	0.82	2.85	1.19	0.30	0.70	46	91	2.82	1.23	-0.06	0.72	-0.12	-0.10
VC214325	3,347	3.17	1.10	3.18	1.11	0.01	0.45	0.79	59	97	0.79	3.17	1.06	0.29	0.67	47	93	3.17	1.06	0.00	0.70	-0.12	-0.09
VC214330	3,296	2.80	1.21	2.81	1.21	0.01	0.49	0.83	60	97	0.83	2.84	1.13	0.33	0.72	48	92	2.83	1.16	0.03	0.74	-0.11	-0.10
VC229874	3,079	3.10	1.15	3.10	1.15	0.00	0.49	0.81	62	96	0.81	3.11	1.10	0.32	0.72	49	94	3.10	1.10	0.00	0.75	-0.08	-0.06
VC243308	3,343	2.90	1.24	2.92	1.24	0.02	0.52	0.84	62	97	0.84	2.90	1.17	0.33	0.73	48	92	2.89	1.18	-0.01	0.76	-0.11	-0.09
VC243309	3,375	3.22	1.26	3.23	1.27	0.01	0.53	0.86	64	97	0.86	3.27	1.18	0.29	0.71	45	91	3.26	1.19	0.03	0.73	-0.15	-0.13
VC257573	3,132	3.23	1.21	3.21	1.22	-0.02	0.46	0.82	58	97	0.82	3.18	1.16	0.29	0.71	45	92	3.17	1.17	-0.05	0.73	-0.11	-0.09
VC265780	3,370	3.14	1.21	3.14	1.19	0.00	0.49	0.83	61	97	0.83	3.22	1.15	0.31	0.74	47	94	3.19	1.15	0.05	0.76	-0.09	-0.06
VC286667	3,332	3.14	1.17	3.13	1.16	-0.01	0.51	0.84	63	98	0.84	3.17	1.12	0.31	0.70	48	92	3.16	1.12	0.02	0.72	-0.13	-0.11
VC315648	3,273	3.12	1.30	3.11	1.29	-0.01	0.49	0.85	60	97	0.85	3.11	1.25	0.33	0.75	47	91	3.09	1.28	-0.03	0.77	-0.11	-0.09
VC315650	3,202	3.07	1.31	3.03	1.30	-0.03	0.52	0.86	62	97	0.86	3.05	1.21	0.31	0.74	46	92	3.02	1.22	-0.03	0.77	-0.12	-0.10
VC315652	3,255	3.23	1.20	3.24	1.23	0.01	0.47	0.83	59	97	0.83	3.23	1.18	0.25	0.69	42	91	3.22	1.19	-0.01	0.71	-0.14	-0.12
VC337838	3,285	3.25	1.01	3.24	1.01	0.00	0.41	0.75	57	97	0.75	3.22	1.01	0.25	0.64	46	94	3.22	0.99	-0.03	0.66	-0.11	-0.09
VC354913	3,423	3.06	1.15	3.08	1.16	0.01	0.50	0.83	62	98	0.83	3.06	1.09	0.30	0.71	47	94	3.05	1.08	-0.01	0.74	-0.12	-0.09

Independent–PS		H1 by H2										H1 by e-rater (rounded to integers)						H1 by e-rater (unrounded)				Degradation	
		H1		H2		Stats						e-rater						e-rater		Stats		Wtd kappa	*R*
Prompt	*N*	M	SD	M	SD	Std diff	Kappa	Wtd kappa	% agree	% adj agree	*R*	M	SD	Kappa	Wtd kappa	% agree	% adj agree	M	SD	Std diff	*R*	H1 by e-rater rounded – H1 by H2	H1 by e-rater unrounded – H1 by H2
VC357067	3,236	3.03	1.28	3.01	1.25	-0.01	0.48	0.84	59	96	0.84	3.02	1.22	0.29	0.71	44	90	3.01	1.26	-0.02	0.73	-0.12	-0.10
VC358128	3,436	3.09	1.17	3.10	1.17	0.01	0.45	0.80	57	96	0.80	3.11	1.15	0.29	0.69	46	91	3.09	1.16	0.01	0.71	-0.11	-0.09
VC364389	3,301	3.01	1.21	3.01	1.22	0.01	0.53	0.85	64	98	0.85	3.02	1.14	0.31	0.72	47	93	3.00	1.14	0.00	0.75	-0.13	-0.10
VC373909	3,240	2.99	1.26	2.99	1.25	0.00	0.51	0.85	62	97	0.85	2.96	1.22	0.31	0.73	46	92	2.94	1.24	-0.04	0.75	-0.12	-0.10
VC373911	3,290	2.97	1.05	2.96	1.04	0.00	0.44	0.77	60	97	0.77	3.01	1.01	0.24	0.62	45	92	3.00	1.00	0.03	0.65	-0.16	-0.13
VC374076	3,440	3.22	1.10	3.21	1.10	0.00	0.44	0.79	58	97	0.79	3.21	1.08	0.30	0.70	48	94	3.21	1.08	0.00	0.72	-0.09	-0.07
VC374333	3,510	3.17	1.21	3.16	1.19	0.00	0.46	0.82	59	97	0.82	3.13	1.14	0.26	0.68	43	90	3.13	1.14	-0.03	0.70	-0.15	-0.12
VC389573	3,279	3.00	1.20	3.01	1.21	0.00	0.48	0.83	60	97	0.83	3.10	1.16	0.33	0.73	48	93	3.10	1.17	0.08	0.75	-0.10	-0.08
VC389578	3,245	3.15	1.10	3.16	1.10	0.01	0.51	0.82	64	98	0.82	3.15	1.10	0.28	0.67	47	92	3.15	1.09	-0.01	0.70	-0.15	-0.12
VC389592	3,294	3.11	1.25	3.11	1.25	0.01	0.48	0.84	60	97	0.84	3.23	1.15	0.31	0.73	47	92	3.22	1.18	0.09	0.75	-0.11	-0.09
VC389593	3,238	3.08	1.20	3.09	1.20	0.01	0.42	0.79	55	95	0.79	3.01	1.18	0.29	0.69	45	91	2.98	1.20	-0.08	0.72	-0.09	-0.07
VC399764	3,306	3.23	1.29	3.22	1.29	-0.01	0.52	0.86	62	97	0.86	3.20	1.20	0.32	0.74	47	92	3.20	1.21	-0.02	0.76	-0.12	-0.10
VC400185	3,598	2.98	1.13	2.98	1.12	0.00	0.48	0.81	61	97	0.81	3.01	1.10	0.32	0.73	49	95	2.99	1.11	0.01	0.75	-0.08	-0.06
VC400187	3,331	3.11	1.18	3.10	1.17	-0.01	0.52	0.84	63	97	0.84	3.05	1.14	0.30	0.71	46	92	3.05	1.14	-0.06	0.73	-0.13	-0.10
VC400188	3,204	2.65	1.33	2.65	1.33	0.00	0.56	0.88	66	98	0.88	2.60	1.27	0.37	0.78	50	93	2.57	1.31	-0.06	0.80	-0.10	-0.08
VC457888	3,300	3.28	1.11	3.29	1.11	0.01	0.49	0.82	62	97	0.82	3.35	1.09	0.27	0.68	46	92	3.33	1.09	0.04	0.71	-0.14	-0.11
VC457890	3,265	3.20	1.26	3.18	1.24	-0.02	0.48	0.84	60	97	0.84	3.12	1.19	0.28	0.71	44	90	3.10	1.21	-0.07	0.73	-0.13	-0.11
VC457893	3,508	3.13	0.99	3.13	0.99	-0.01	0.48	0.78	64	98	0.78	3.20	0.98	0.26	0.65	48	95	3.19	0.96	0.06	0.67	-0.13	-0.10

Independent–PS		H1 by H2										H1 by e-rater (rounded to integers)						H1 by e-rater (unrounded)				Degradation	
		H1		H2		Stats						e-rater						e-rater		Stats		Wtd kappa	*R*
Prompt	*N*	M	SD	M	SD	Std diff	Kappa	Wtd kappa	% agree	% adj agree	*R*	M	SD	Kappa	Wtd kappa	% agree	% adj agree	M	SD	Std diff	*R*	H1 by e-rater rounded – H1 by H2	H1 by e-rater unrounded – H1 by H2
VC457896	3,367	3.01	1.26	3.02	1.27	0.01	0.50	0.85	61	97	0.85	3.06	1.20	0.29	0.70	45	89	3.04	1.23	0.02	0.72	-0.15	-0.13
VC457898	3,300	3.33	1.16	3.31	1.16	-0.02	0.43	0.80	57	97	0.80	3.41	1.13	0.32	0.73	49	94	3.40	1.13	0.06	0.75	-0.07	-0.05
VC472548	3,396	3.04	1.16	3.04	1.16	0.01	0.45	0.80	58	96	0.80	2.97	1.11	0.30	0.70	46	92	2.96	1.12	-0.06	0.72	-0.11	-0.08
VC503849	3,220	3.03	1.19	3.02	1.20	-0.01	0.46	0.81	58	96	0.81	3.03	1.17	0.28	0.68	45	90	3.01	1.18	-0.01	0.70	-0.13	-0.10
Average	3,316	3.08	1.19	3.07	1.19	0.00	0.48	0.82	60	97	0.82	3.08	1.14	0.30	0.70	46	92	3.07	1.15	-0.01	0.73	-0.12	-0.10

Note. *N* is average across all the prompts. Shaded cells indicate values that fail to meet the thresholds listed in Table A2.

adj = adjacent, H1 = human 1, H2 = human 2, std diff = standardized difference, wtd = weighted.

Table B5

Subgroup Differences for Independent Prompts: Generic (G) Model

Independent		H1 by H2										H1 by e-rater (rounded to integers)						H1 by e-rater (unrounded)				Degradation	
		H1		H2		Stats						e-rater						e-rater		Stats		Wtd kappa	*R*
Native language	*N*	M	SD	M	SD	Std diff	Kappa	Wtd kappa	% agree	% adj agree	*R*	M	SD	Kappa	Wtd kappa	% agree	% adj agree	M	SD	Std diff	*R*	H1 by e-rater rounded – H1 by H2	H1 by e-rater unrounded – H1 by H2
French	3,982	3.48	0.83	3.49	0.83	0.01	0.36	0.65	58	98	0.65	3.48	0.79	0.39	0.67	60	99	3.48	0.75	0.00	0.72	0.02	0.07
German	4,968	3.85	0.76	3.86	0.76	0.02	0.32	0.57	56	98	0.57	3.83	0.66	0.32	0.56	58	99	3.84	0.61	-0.01	0.63	-0.01	0.06
Italian	2,320	3.31	0.82	3.31	0.82	0.00	0.35	0.64	58	98	0.64	3.32	0.80	0.42	0.69	62	99	3.32	0.75	0.00	0.74	0.05	0.10
Japanese	11,201	3.01	0.83	3.02	0.82	0.01	0.42	0.70	62	99	0.70	3.02	0.90	0.42	0.72	61	99	3.02	0.88	0.01	0.76	0.02	0.07
Korean	26,123	3.25	0.82	3.26	0.83	0.00	0.39	0.68	61	99	0.68	3.34	0.89	0.39	0.70	59	99	3.34	0.86	0.10	0.74	0.02	0.06
Portuguese	2,462	3.44	0.84	3.44	0.83	0.00	0.36	0.65	58	98	0.65	3.39	0.80	0.40	0.68	61	99	3.39	0.76	-0.06	0.73	0.03	0.08
Telugu	3,772	3.38	0.81	3.40	0.81	0.02	0.35	0.61	58	97	0.61	3.41	0.74	0.32	0.59	57	98	3.41	0.68	0.03	0.64	-0.02	0.03
Turkish	3,906	3.31	0.80	3.31	0.79	-0.01	0.37	0.64	60	98	0.64	3.31	0.82	0.36	0.66	59	99	3.30	0.78	-0.01	0.70	0.02	0.06

Note. *N* is average across all the prompts. Shaded cells indicate values that fail to meet the thresholds listed in Table A2.

adj = adjacent, H1 = human 1, H2 = human 2, std diff = standardized difference, wtd = weighted.

Table B6

Subgroup Differences for Integrated Prompts: Generic (G) Model

Integrated		H1 by H2										H1 by e-rater (rounded to integers)						H1 by e-rater (unrounded)				Degradation	
		H1		H2		Stats						e-rater						e-rater		Stats		Wtd kappa	*R*
Native language	*N*	M	SD	M	SD	Std diff	Kappa	Wtd kappa	% agree	% adj agree	*R*	M	SD	Kappa	Wtd kappa	% agree	% adj agree	M	SD	Std diff	*R*	H1 by e-rater rounded – H1 by H2	H1 by e-rater unrounded – H1 by H2
French	4,276	3.18	1.18	3.19	1.18	0.01	0.48	0.82	61	97	0.82	3.17	1.04	0.17	0.56	38	87	3.16	1.04	-0.02	0.59	-0.26	-0.23
German	5,484	3.71	1.01	3.71	1.01	-0.01	0.41	0.73	58	96	0.73	3.70	0.90	0.15	0.48	40	89	3.72	0.88	0.01	0.50	-0.25	-0.23
Italian	2,519	3.10	1.19	3.10	1.17	0.00	0.47	0.83	59	98	0.83	3.11	0.99	0.17	0.53	38	86	3.11	0.97	0.00	0.56	-0.30	-0.27
Japanese	11,789	2.59	1.17	2.59	1.17	0.00	0.53	0.84	64	98	0.84	2.60	1.12	0.23	0.60	41	86	2.55	1.17	-0.04	0.62	-0.25	-0.22
Korean	28,180	3.05	1.18	3.05	1.18	0.00	0.48	0.82	60	97	0.82	3.07	1.11	0.21	0.61	40	88	3.05	1.13	0.00	0.63	-0.21	-0.19
Portuguese	2,675	3.24	1.16	3.23	1.16	-0.01	0.48	0.82	61	97	0.82	3.14	1.06	0.17	0.58	38	87	3.13	1.05	-0.10	0.61	-0.24	-0.21
Telugu	4,048	3.18	1.15	3.19	1.15	0.00	0.46	0.80	59	96	0.80	3.30	1.01	0.16	0.50	37	85	3.30	1.00	0.11	0.53	-0.30	-0.27
Turkish	4,141	2.91	1.16	2.90	1.15	-0.01	0.50	0.82	62	97	0.82	3.03	1.09	0.20	0.57	39	86	3.03	1.09	0.10	0.59	-0.25	-0.23

Note. N is average across all the prompts. Shaded cells indicate values that fail to meet the thresholds listed in Table A2.

adj = adjacent, H1 = human 1, H2 = human 2, std diff = standardized difference, wtd = weighted.

Table B7

Subgroup Differences Under Contributory Score Model for Independent Only at 1.5 Threshold

Independent generic at 1.5		Writing raw operational		Writing raw simulated		Writing raw operational by writing raw simulated						
	N	M	SD	M	SD	Kappa	Wtd kappa	% agree	% adj (< 0.5)	% adj (< 1.0)	Std diff.	*R*
Test center country												
China	16,884	6.39	1.57	6.49	1.58	0.54	0.98	57.88	98.45	99.71	0.06	0.98
Native language												
Arabic	7,864	5.77	1.79	5.69	1.81	0.56	0.98	59.33	98.68	99.83	-0.05	0.99
Chinese	27,286	6.30	1.60	6.38	1.62	0.55	0.98	59.22	98.56	99.72	0.05	0.98
French	4,262	6.69	1.68	6.69	1.66	0.57	0.98	60.35	98.94	99.77	0.00	0.98
German	5,474	7.58	1.39	7.56	1.36	0.56	0.97	60.80	99.10	99.69	-0.01	0.98
Hindi	3,191	7.61	1.60	7.53	1.58	0.53	0.97	57.35	98.46	99.53	-0.05	0.98
Italian	2,506	6.44	1.66	6.45	1.66	0.58	0.98	61.49	98.68	99.76	0.00	0.98
Japanese	11,757	5.64	1.67	5.65	1.70	0.60	0.98	62.83	99.12	99.90	0.00	0.99
Korean	28,069	6.32	1.71	6.36	1.74	0.57	0.98	60.42	98.81	99.82	0.03	0.98
Portuguese	2,651	6.70	1.67	6.68	1.66	0.57	0.98	60.81	98.42	99.55	-0.02	0.98
Spanish	10,026	6.54	1.68	6.50	1.67	0.57	0.98	60.27	98.69	99.79	-0.03	0.98
Telugu	4,000	6.58	1.63	6.58	1.59	0.54	0.98	58.18	98.18	99.65	0.00	0.98
Turkish	4,118	6.23	1.61	6.23	1.62	0.56	0.98	59.81	98.66	99.64	0.00	0.98
Ability level												
High	48,485	8.09	1.02	8.07	0.99	0.52	0.94	58.87	98.52	99.66	-0.03	0.95
Medium	49,065	6.41	0.98	6.45	1.00	0.53	0.94	59.69	98.73	99.76	0.04	0.95
Low	43,653	4.67	1.17	4.67	1.22	0.56	0.96	61.54	98.83	99.86	0.00	0.97

Note. *N* is average across all the prompts. adj = adjacent, H1 = human 1, H2 = human 2, std diff = standardized difference, wtd = weighted.

Table B8

Subgroup Differences Under Contributory Score Model at 1.5 Threshold for Independent and 1-Point Threshold for Integrated

1.5 for independent and 1 pt for integrated		Writing raw operational		Writing raw simulated		Writing raw operational by writing raw simulated						
	N	M	SD	M	SD	Kappa	Wtd kappa	% Agree	% adj (< 0.5)	% adj (< 1.0)	Std diff.	*R*
Test center country												
China	16,884	6.39	1.57	6.52	1.57	0.30	0.94	35.94	84.62	97.97	0.09	0.94
Native language												
Arabic	7,864	5.77	1.79	5.62	1.82	0.31	0.95	36.14	84.32	98.21	-0.08	0.96
Chinese	27,286	6.30	1.60	6.40	1.61	0.31	0.94	36.54	85.19	98.09	0.07	0.95
French	4,262	6.69	1.68	6.65	1.63	0.31	0.95	37.19	86.25	98.47	-0.02	0.95
German	5,474	7.58	1.39	7.54	1.33	0.31	0.92	37.96	87.01	98.43	-0.02	0.93
Hindi	3,191	7.61	1.60	7.48	1.57	0.30	0.94	36.01	84.74	97.93	-0.08	0.94
Italian	2,506	6.44	1.66	6.42	1.60	0.33	0.95	38.55	87.59	98.56	-0.01	0.95
Japanese	11,757	5.64	1.67	5.61	1.72	0.32	0.95	37.42	87.44	98.66	-0.02	0.96
Korean	28,069	6.32	1.71	6.35	1.74	0.31	0.95	36.75	86.07	98.36	0.02	0.95
Portuguese	2,651	6.70	1.67	6.63	1.63	0.33	0.95	39.12	87.51	98.15	-0.04	0.95
Spanish	10,026	6.54	1.68	6.48	1.62	0.33	0.95	38.44	86.68	98.37	-0.04	0.95
Telugu	4,000	6.58	1.63	6.60	1.55	0.30	0.94	36.13	85.43	97.55	0.01	0.94
Turkish	4,118	6.23	1.61	6.25	1.62	0.33	0.94	38.44	86.11	98.28	0.01	0.95
Ability level												
High	48,485	8.09	1.02	8.00	1.00	0.27	0.85	37.14	86.05	98.10	-0.09	0.87
Medium	49,065	6.41	0.98	6.45	1.02	0.26	0.85	36.82	85.47	98.14	0.04	0.86
Low	43,653	4.67	1.17	4.69	1.27	0.28	0.90	36.38	85.88	98.47	0.01	0.91

Note. *N* is average across all the prompts. adj = adjacent, H1 = human 1, H2 = human 2, std diff = standardized difference, wtd = weighted.