Listening. Learning. Leading.®

Research Report
ETS RR–13-36

Investigating the Suitability of Implementing the *e-rater*® Scoring Engine in a Large-Scale English Language Testing Program

Mo Zhang

F. Jay Breyer

Florian Lorenz

December 2013

ETS Research Report Series

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

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

Investigating the Suitability of Implementing the *e-rater*® Scoring Engine in a Large-Scale English Language Testing Program

Mo Zhang, F. Jay Breyer, and Florian Lorenz
Educational Testing Service, Princeton, New Jersey

December 2013

Action Editor: Beata Beigman Klebanov

Reviewers: Brent Bridgeman and Shelby J. Haberman

Abstract

In this research, we investigated the suitability of implementing *e-rater*® automated essay scoring in a high-stakes large-scale English language testing program. We examined the effectiveness of generic scoring and 2 variants of prompt-based scoring approaches. Effectiveness was evaluated on a number of dimensions, including agreement between the automated and the human score and relations with criterion variables. Results showed that the sample size was generally not sufficient for prompt-specific scoring. For the generic scoring model, automated scores agreed with human raters as strongly as, or more strongly than, human raters agreed with one another for more than 97% of the prompts. The impact of substituting e-rater for the second human rater made no practically important impact on test takers' scores at both the item and total test score levels. However, neither automated scoring models nor human raters performed invariantly across all prompts or across different test countries/territories. Further investigation indicated homogeneity in the examinee population, possibly nested within test countries/territories as one potential cause of this lack of invariance. Among other limitations, findings may not be generalizable beyond the examinee population investigated in this study.

Key words: model development, e-rater, automated essay scoring, impact analysis

Acknowledgments

We would like to express our gratitude for the guidance that Shelby Haberman and David Williamson provided throughout this study. Our thanks also go to Brent Bridgeman, Neil Dorans, Yigal Attali, and Catherine Trapani for discussion and technical advice and to Elizabeth Park and Susan Hines for furthering our understanding of issues in test administration. Finally, we greatly appreciate the feedback we received from the editor and the reviewers of this report.

Table of Contents

List of Tables

Page

With *e-rater*® automated essay scoring becoming operational for such ETS testing programs as the *Graduate Record Examinations*® (*GRE*®) revised General Test and the *TOEFL iBT*® test (ETS, 2013a, 2013b), its suitability for another ETS-administered English language testing program has been proposed. This testing program is designed to comprehensively assess nonnative speakers' English proficiency in four major areas: reading, listening, speaking, and writing. Annually, more than 14,000 institutions from more than 150 countries throughout the world use the examination results to decide on qualified candidates with sufficient English communication skills.

The writing assessment portion in the testing program measures test takers' communicative writing proficiencies in workplace settings, including the ability to convey information, ask questions, provide instructions, state narratives, as well as the ability to express opinions on problems and issues in a logical and cohesive manner.

The essay item in the writing assessment is similar to the independent writing task in the TOEFL iBT test and the issue task in the GRE General Test, for which research has shown successful implementations of e-rater (e.g., Ramineni, Trapani, Williamson, Davey, & Bridgeman, 2012a, 2012b). Therefore, we conducted a focused investigation on the feasibility of implementing e-rater operationally to score the essay item in the writing assessment of this language-testing program. This report documents our research procedure and findings.

Background

Description of the Writing Assessment

The writing assessment consists of three item types, with a total of eight individual items. In the first five items, the test takers are asked to describe a picture scenario, in one sentence, that contains two words provided in the prompt. In Items 6 and 7, test takers are asked to respond to an e-mail by following specific instructions (e.g., provide a fact and ask a related question). Last, test takers have 30 minutes to write an essay, with a minimum of 300 words, to support their opinion(s) on a general issue. Even though the intended use of all items is to assess writing proficiencies, each of the three item types stresses somewhat different characteristics in writing (Table 1).

Table 1

Aspects of Evaluation for Different Item Types in the Writing Assessment

Item	Item format	Evaluation aspects
1–5	Describing a scenario in a picture	Grammar; relevance of the sentence to the picture.
6–7	Responding to an e-mail request	Quality and variety of the sentence structure; vocabulary; organization.
8	Essay	Grammar; vocabulary; organization; whether the opinion is supported by reasons and/or examples.

E-rater Automated Essay Scoring[1]

The e-rater automated essay scoring system is used in a wide range of testing programs to grade their essay component. Currently, e-rater is implemented in more than a dozen assessments inside and outside ETS. Moreover, e-rater has become a sole score producer in a number of low- and medium-stakes assessments, such as the *Criterion*® Online Writing Evaluation (ETS, 2013c) and *TOEFL*® Practice Online (ETS, 2010). Since its prototype was introduced in the late 1990s, a number of scientific research studies about e-rater have been published in research reports, journal articles, books, and book chapters in psychometrics and education, in addition to a number of professional conference presentations and publications in educational assessment and computer science (e.g., Attali & Burstein, 2006; Beigman Klebanov & Flor, 2013; Bennett & Bejar, 1998; Haberman, 2011; Williamson, 2013; Williamson, Mislevy, & Bejar, 2006).

With respect to the psychometric research, most emphasis and effort have been placed on the investigation of automated essay score validity. Evidence in score validity is one of the most important indicators of whether e-rater can be integrated into an assessment. Previous validity-related research can be categorized into two general classes: examination of the construct representation of the e-rater scores and development and improvement of the approaches to calibrating e-rater scoring models.

In general, empirical research has supported the construct relevance of the e-rater scores. The features extracted by e-rater are indicators of meaningful dimensions of effective writing (Attali, 2007; Attali & Powers, 2009; Appendix A gives a complete list of features extracted by e-rater). Previous research found that e-rater produced scores that (a) correlate with human

ratings as strongly as human ratings agree with each other (e.g., Attali & Powers, 2009) and (b) modestly correlate with external variables that measure the same construct (e.g., Bridgeman & Trapani, 2011).

The operational e-rater scoring model is constructed by regressing human ratings on the features, which results in a multiple linear regression that can be applied to generate scores that a human rater would assign to a given essay. Although continuous research efforts have been allocated to develop new types of scoring models, with an intention to enhance the automated score validity (e.g., Ben-Simon & Bennett, 2007; Zhang, Williamson, Breyer, & Trapani, 2012), to date, only two types of models have been established for operational practice in large-scale assessments: the *generic* model (G model) and the *prompt-specific* model (PS model).

A G model is built on a group of essay prompts, where the group is defined by the similarity of the writing task posed to the examinee. As a result, all essays in the group have the same scoring algorithm. With G model scoring, substantive scoring consistency across essays and prompts can be achieved. The PS model is, as its name suggests, built on individual prompts. In PS model scoring, both the intercept and feature weights in the regression are customized for a prompt. PS models also differ from G models in that they often employ all 11 e-rater features, although the G models do not use the two content-related features.

Traditionally, e-rater models are built on a subset of the essays and evaluated on the remaining essays. Researchers had also proposed to construct a scoring model in a jackknife or *n*-fold cross-validation fashion (Haberman & Sinharay, 2008).[2] This way, not only the sample size for model calibration is largely increased, but the variation in the entire data set can also be reflected in both model calibration and evaluation. In this study, we call this prompt-based model type the *PS (press)* model, in part because the regular model evaluation procedures are no longer applicable. Instead, indices derived from predicted residual sums of squares (PRESS) statistics are used for evaluation (described in the Model Development Procedure section).

Automated Scoring Model Development

Research Question

We conducted this exploratory study with one main research question in mind: Can e-rater be used to score the essay item either alone or in tandem with human raters? If so, which type of scoring model should be applied? We examined and compared three types of scoring models: a G model serving as a baseline, and two variants of the PS model, one traditional

denoted as *PS (traditional)*, and the other, *PS (press).* The processes for model construction and model evaluation criteria are given in the Model Development Procedure section.

Instrument

We used a data set of 63 prompts administered by the testing program from January 2010 to July 2011. Several demographic variables were retained, including country/territory (where the test was taken) and the test taker's native language. Although the test country/territory was recorded for every examinee, the language background was self-reported. Additionally, testtakers' scores on the other writing tasks, as well as their speaking scores if available, were used in the evaluation.

Two operational human ratings were available for each essay response.[3] A human rating could range from 0 to 5, with 0 indicating off-topic or other aberrant responses. The average of the two human ratings was considered as a test taker's final raw essay score. However, in cases where the two ratings were apart by more than 1 point, an adjudication rater was brought in. If the three scores were adjacent to one another, the adjudication rater's score became the final score. If there appeared to be an outlier among the three scores, the average of the two adjacent ones became the final raw score. If none of the three scores were adjacent to one another, the adjudication rater's score became the final score. The above adjudication rules were also applicable to the other seven writing items, except that only a subset of the submissions in those items was graded by two randomly assigned human raters. (See Appendix B for the scoring rubrics.)

Test takers' final raw scores on the writing assessment were a weighted total of all eight items. Specifically, the essay item was given the highest weight and the first five items were given the lowest weight. For practical purposes, the final raw writing scores were subsequently converted to scale scores and a corresponding writing proficiency level for reporting purpose. A scaled score could range from 0 to 200, in increments of 10.

The quality of human ratings was examined using the aforementioned data set (N = 32,835). Further, we investigated the human rating quality using another data set from the same testing program that was not available when the project started, in order to obtain more information. This additional data set consisted of 123 prompts and was collected between May 2011 and April 2012 (N = 47,804).

The overall interrater agreement is given in Table 2, for which agreement is indicated by the Pearson correlation coefficient, quadratic-weighted kappa, standardized mean score difference between the two operational ratings, exact percentage agreement, and adjacent percentage agreement.

Although the examination of both data sets suggested little concern over human rating quality on the population basis, the interhuman agreement level according to the quadratic-weighted kappa metric was, at times, fairly low on the prompt basis. Specifically, in the 2010–2011 data set, only 18 of 63 prompts had a quadratic-weighted kappa value equal to or greater than 0.70, and 23 prompts had a quadratic-weighted kappa value lower than 0.65. The number of essay responses to the prompts (termed as *prompt size* hereafter) ranged from 285 to 899. In the 2011–2012 data set, of the 86 prompts (of the total 123 prompts) that had a prompt size greater than 100, 31 prompts had a quadratic-weighted kappa value equal to or greater than 0.70, while 37 prompts had a quadratic-weighted kappa value lower than 0.65.[4] (See Appendices C and D for detailed documentation.)

Table 2

Overall Interhuman Agreement

Data set	*N*	Human1		Human2		Pearson correlation	Quadratic-weighted kappa	Abs. std. dif.	% agree	% adj. agree
		Mean	SD	Mean	SD					
Jan. 2010 Jul. 2011	32,835	2.99	0.75	2.99	0.74	0.76	0.76	0.00	69.8	99.3
May 2011–Apr. 2012	47,804	3.00	0.74	2.99	0.74	0.71	0.71	0.01	70.5	99.4

Note. Abs. std. dif. = absolute standardized mean score difference; % agree = exact percentage agreement; % adj. agree = 1-point adjacent percentage agreement.

Finally, e-rater identifies essays that are not appropriate for automated scoring by advisory flags (Appendix E provides a complete list of the current e-rater advisory flags). Because we had no prior knowledge of the essay characteristics, we activated all advisories and excluded flagged essays from subsequent analysis. Additionally, essays with a human score of 0 were also viewed as inappropriate for automated scoring and therefore were excluded from data analyses. It is worth noting that the 32,835 essays used for model calibration or model evaluation were flag-free and legitimate responses.

Participants

Of all 32,835 test takers, 17,523 (53%) took the test in South Korea; 7,076 (22%) in Japan; and 4,288 (13%) in India. The remaining 12% of test takers took the test in 33 other countries/territories, including China, Great Britain, and Taiwan as the three next-largest test countries/territories.[5]

A prospective test taker may take the writing assessment via two different programs. One program (Program-U) generally consists of the test takers who sign up for the test on an individual basis and the other program (Program-K) generally consists of cohorts of test takers from the same institution. In the analyses, we did not separate the data collected from these two administration approaches because (a) Program-U test takers constituted a small portion of the population (i.e., 11%), and (b) a comparison between the test takers in the two programs suggested little material differences on their linguistic ability (Table 3).

Table 3

Comparing Test taker Characteristics Between Programs K and U

Program	N	Human1 (essay)	Human2 (essay)	Rpt. writing score	Rpt. speaking score
K	29,124	Mean = 3.0 SD = 0.8	Mean = 3.0 SD = 0.8	Mean = 146.2 SD = 29.2	Mean = 132.0 SD = 32.0
U	3,711	Mean = 2.9 SD = 0.7	Mean = 2.9 SD = 0.7	Mean = 144.3 SD = 27.2	Mean = 124.5 SD = 31.6

Note. Rpt. = reported.

Model Development Procedure

Generic Model Calibration and Evaluation

A generic model was calibrated on 10,000 essays, with 200 each randomly selected from 50 of 63 prompts that had 400 or more responses. A multiple linear regression model was calibrated by regressing the mean of the two human ratings on nine feature variables (excluding the content feature). The scoring model was further calibrated so that the mean and standard deviation of the resulting automated scores matched the mean and standard deviation of the human ratings in the model-building data set.

We evaluated the G model on the remaining essays that were not selected for model calibration. Evaluation was conducted based on the agreement between automated and human

scores and on the correlational strength of the automated scores with relevant external measures of linguistic ability (i.e., scores on the other writing items and on the speaking assessment).

We adopted the following criteria proposed by Williamson, Xi, and Breyer (2012) to indicate a good model performance for the overall population and on a prompt basis: the correlation coefficient and quadratic-weighted kappa between human and e-rater scores should be equal to or greater than 0.70; the standardized mean score difference between human and e-rater scores (i.e., difference between mean human and mean e-rater scores standardized by the pooled standard deviation of human ratings and e-rater scores) should be equal to or less than 0.15; and the degradation in correlation and weighed kappa from interhuman to e-rater–human should be equal to or less than 0.10.

The model was also evaluated based on those indices for individual test country/territory in order to ensure the fairness of resulting automated scores across different demographic groups. The criteria for correlation coefficient, quadratic-weighted kappa, and degradation remained, but a more stringent criterion for standardized mean score difference was applied, with a preferred value of no more than 0.10 at subgroup level (Williamson et al., 2012).

We also reported percentage agreement as an additional index to provide more information in model evaluation. Nonetheless, this index is less important in decision making due to its limitations of not taking chance into account and of scale dependency.

Unlike human ratings that are awarded as integer values, e-rater scores are continuous values. In computing the correlation coefficients and standardized mean score differences, e-rater scores that were out of range of the scoring rubric were truncated to 0.5001 or 6.4999. After truncation, the e-rater scores were further normally rounded to integers for computing the quadratic-weighted kappa and agreement percentages.

Finally, the correlational strength between the e-rater scores and test takers' scores on other measures of linguistic ability (i.e., Writing Items 1–7, reported speaking score) was examined to provide evidence for the external-relations aspect of validity. This strength was further compared with the correlational strength of human ratings with the same variables.

PS (Traditional) Model Calibration and Evaluation

Each prompt was randomly divided into halves for model building and evaluation, respectively. Due to our concern about small model calibration sample size (which could result in unreliable scoring models), the PS (traditional) models were only constructed and examined

for the 50 prompts with more than 400 examinees. All 11 features were included for model calibration. As in the G model, the mean of the two human ratings was used as a predicted variable. PS (traditional) models were evaluated based on the agreement of the resulting e-rater scores with human ratings using the same agreement indices described previously for the generic scoring approach.

Because the majority of the prompts used in this study were only administered in two or three countries/territories (an issue to be fully elaborated in the Discussion section), coupled with the small overall remaining sampling size in model evaluation, PS (traditional) models were not evaluated on a test country/territory basis.

PS (Press) Model Calibration and Evaluation

PS (press) models were built for each of the 63 prompts separately applying a leave-one-out sampling approach. Models were evaluated using indices derived from the PRESS statistic (Kutner, Nachtsheim, Neter, & Li, 2005, p. 360). Those indices differ from traditional model evaluation indices (e.g., correlation coefficient) in the sense that they attempt to estimate a test taker's true score on a prompt (either from a human or e-rater). Although true score itself is not observable, it can be interpreted as the mean of the observed scores from an infinite number of repeated administrations of the same test (Crocker & Algina, 2006, p. 109). The scoring-model evaluation indices derived from the PRESS statistic function under this true-score framework. It is worth noting that although using deleted residuals for scoring-model evaluation is not uncommon, there is little discussion in the literature of this use for PRESS-statistic evaluation indices. (See Guilford & Fruchter, 1973; Haberman & Sinharay, 2008; and Weisberg, 1985, for exceptions.)

In this study, we used a total of four PRESS-derived indices. The first index was termed the *inflation index (II)*. It was used to determine whether the sample size was adequate for model calibration by taking into account the number of estimated parameters in a regression model. A second index was used to indicate the extent to which e-rater scores can replace human scores in predicting a test taker's true score on a prompt; in other words, the index evaluates the effectiveness of the automated scores relative to human ratings in estimating a true score. We termed this index the *value index (VI)*. For example, if a prompt calls for a content-based response that the automated scoring engine cannot effectively evaluate, this index may value the automated scores less than human ratings. A third index, also derived from the PRESS statistic, is termed the

prediction index (PI). PI was used to address the capability of the e-rater scoring models in predicting a test taker's true score on a prompt. The judgments based on VI and PI should be considered relative to the quality of the human ratings. Traditional agreement indices (e.g., correlation coefficient) are not comparable to the PRESS-derived indices without true-score adjustment. Hence, a fourth index, the *human-rating quality index (HQI)* was used in conjunction with the other three to evaluate the quality of human ratings in predicting a test taker's true score. See Appendix F for the computation of the four PS (press) model evaluation indices.

Table 4 gives the evaluation guidelines for all four PS (press) indices. Values not meeting the stated satisfactory levels would be identified, for which the corresponding PS (press) model would warrant further examination.

Table 4

Thresholds for PRESS Statistic Derived Indices

Index	Functional purpose	Satisfactory level
Human-rating quality index (HQI)	Evaluate the capability of human ratings in predicting true scores	≥ 0.80
Value index (VI)	Evaluate the effectiveness of e-rater scores in replacing human ratings	≥ 1.00
Prediction index (PI)	Evaluate the capability of e-rater scores in predicting true scores	≥ 0.70
Inflation index (II)	Evaluate the adequacy of the sample size for PS (press) model calibration	≤ 0.04

Note. Formulas are given in Appendix F.

Results of Model Development

Generic Model Performance

Human–e-rater agreement. Table 5 presents the model performance on several selected prompts representing different scenarios (see Appendix G for a full account of the results). This table gives the distribution moments (i.e., mean and standard deviation) of the two operational human ratings and resulting e-rater scores, the agreement of Human1 with Human2, the agreement of Human1 with e-rater (i.e., standardized mean score difference, kappa, quadratic-weighted kappa, percentage agreement, 1-point adjacent percentage agreement, and correlation coefficient), as well as the degradation in correlation coefficient and quadratic-weighted kappa from Human1/Human2 to Human1/e-rater.

Table 5

Generic Model Evaluation on Selected Prompts

Prompt	*N*	Human1		Human2		Human1/Human2 agreement statistics						e-rater			Human1/e-rater agreement statistics					Degradation (h1/e-rater–h1/h2)	
		Mean	SD	Mean	SD	Std. diff.	Kappa	Wtd. kappa	% agree	% adj agree	Corr.	Mean	SD	Std. dff.	Kappa	Wtd. kappa	% agree	% adj agree	Corr.	Wtd. kappa	Corr.
C4052	489	2.99	0.79	2.99	0.79	-0.01	0.46	0.71	66.05	99.39	0.71	2.93	0.78	-0.09	0.50	0.74	67.69	99.18	0.80	0.03	0.09
C4084	309	2.47	0.93	2.55	0.98	0.08	0.52	0.78	66.02	98.06	0.79	2.72	0.89	0.28[a]	0.35	0.62[a]	54.05	93.85	0.67[a]	-0.16[a]	-0.12[a]
C4086	329	3.21	0.70	3.16	0.65	-0.07	0.46	0.67[a]	70.52	99.70	0.67[a]	3.10	0.69	-0.16[a]	0.42	0.64[a]	65.35	98.48	0.68[a]	-0.03	0.01
C5923	318	3.36	0.67	3.25	0.66	-0.15[a]	0.53	0.68[a]	73.58	99.37	0.69[a]	3.28	0.62	-0.12	0.49	0.66[a]	70.75	99.69	0.71	-0.02	0.02
E1577	454	2.82	0.63	2.75	0.64	-0.11	0.50	0.67[a]	74.23	99.78	0.68[a]	2.74	0.73	-0.11	0.50	0.71	72.47	99.78	0.77	0.04	0.09
E4365	201	2.69	0.66	2.70	0.62	0.01	0.59	0.74	78.61	100.00	0.74	2.46	0.78	-0.33[a]	0.34	0.63[a]	60.70	100.00	0.76	-0.11[a]	0.02
Unweighted average (across all 63 prompts)	362.5	3.00	0.70	3.00	0.69	-0.01	0.46	0.66[a]	69.89	99.32	0.66[a]	3.00	0.68	-0.01	0.42	0.65[a]	66.19	99.29	0.71	-0.01	0.05

Note. Full account of the results for generic model evaluation is given in Appendix G. Std. diff. = standardized difference on mean scores; Wtd. kappa = quadratic-weighted kappa; corr. = correlation coefficient; % agree = Exact percentage agreement; % adj agree = 1-point adjacent percentage agreement.

[a]Shaded values did not meet the evaluation thresholds.

The overall performance of the generic scoring model in terms of Human1/e-rater agreement achieved a quadratic-weighted kappa value of 0.65 and a correlation coefficient of 0.71, with negligible degradation on quadratic-weighted kappa (-0.01) and an improvement on correlation coefficient (0.05) from human/human agreement (bottom row in Table 5).

However, the generic scoring model yielded mixed performance results across prompts. Strictly speaking, the model performance was not quite satisfactory for a number of individual prompts, with only eight prompts meeting all evaluation thresholds (e.g., C4052). Eighteen of 63 prompts had a human/e-rater standardized mean score difference greater than 0.15, indicating a large distributional difference between the two scores. Nearly half of the prompts (i.e., 24 of 63) had a human/e-rater correlation coefficient value lower than 0.70, and nearly three quarters of the prompts (i.e., 46 of 63) had a human/e-rater quadratic-weighted kappa value lower than 0.70.

Despite the low Human1/e-rater agreement for a large number of prompts, the degradation in both weighted kappa and correlation coefficient was minimal, except for two prompts (C4084 on both indicators and E4365 on quadratic-weighted kappa only). Furthermore, e-rater tended to correlate with Human1 more strongly than Human1 correlated with Human2, indicated by 57 prompts receiving a higher value for correlation coefficient in human/e-rater than in Human1/Human2 and 23 prompts for quadratic-weighted kappa (e.g., E1577). Finally, in most cases where e-rater scores did not correlate with human ratings well, the interhuman agreement was also fairly low (e.g., C4086).

Of note is that one prompt (C5923, of all 63 prompts) showed a greater than 0.15 standardized mean score difference between the two operational human ratings, although it is fairly uncommon to find such large distributional discrepancies between two randomly assigned human raters in operation.

Model Performance on Population Groups

The G model performance on 16 test country/territory groups (that had more than 100 essays in the cross evaluation data set) is given in Table 6. This table can be read in a similar way as Table 5. The interhuman correlation coefficient and quadratic-weighted kappa were low for a number of country/territory groups, including the two large ones, Korea and India. However, the distributional differences between the two human ratings (indicated by standardized mean score difference) were small for all groups but the United States. Therefore, it

is suspected that the low correlation coefficient and quadratic-weighted kappa were subject to a lack of variability in the human rating (see SD columns under Human1 and Human2).

In terms of correlation coefficients, the resulting e-rater scores correlated with human ratings more strongly than human ratings correlated with one another for all country/territory groups but China. There is a slight degradation indicated by quadratic-weighted kappa, though again, only China failed to meet the evaluation threshold (i.e., 0.1). We also noted that there are discrepancies between the distributions of human ratings and e-rater scores for a few country/territory groups. For example, e-rater tended to provide higher scores than humans for test takers from China, Taiwan, France, and Indonesia, while human raters tended to provide higher scores for test takers from Great Britain, Kazakhstan, the Philippines, and the United States. It also appeared that when interhuman agreement was low, the corresponding e-rater/human agreement was also low. This phenomenon was not a surprise, given that e-rater was produced to emulate human ratings, whose quality is likely to be bounded by the quality of the human ratings.

External Relations

Evidence for the external-relations aspect of validity was collected using the same model evaluation data set. Table 7 gives correlations of Human1 and e-rater scores with several linguistic measures, including test takers' operational raw scores on the essay item as well as scores on the Describing a Picture writing tasks, on the Responding to an E-mail writing task, on a weighted total score of the first seven writing items, on reported scaled writing scores, and on reported scaled speaking scores. Of note is that Human1 ratings partially contributed to the operational essay raw scores and reported writing scores.

Table 6

Generic Model Evaluation by Test Country/Territory

Country/ territory	*N*	Human1		Human2		Human1/Human2 agreement statistics					e-rater			Human1/e-rater agreement statistics					Degradation (h1/e-rater–h1/h2)	
		Mean	SD	Mean	SD	Std. diff.	Kappa	Wtd. kappa	% agree	% adj agree	Corr.	Mean	SD	Std. diff.	Kappa	Wtd. kappa	% agree	% adj agree	Corr.	Wtd. kappa	Corr.
KOR	12,311	2.92	0.67	2.92	0.67	-0.01	0.48	0.68[a]	72.30	99.51	0.68[a]	2.93	0.70	0.01	0.43	0.66[a]	67.27	99.45	0.72	-0.02	0.04
JPN	4,689	2.77	0.72	2.78	0.70	0.00	0.49	0.70	70.95	99.51	0.70	2.71	0.73	-0.09	0.47	0.69[a]	68.01	99.32	0.75	-0.01	0.05
IND	3,045	3.67	0.67	3.68	0.65	0.01	0.30	0.48[a]	59.15	98.49	0.48[a]	3.68	0.43	0.01	0.30	0.43[a]	61.51	99.21	0.53[a]	-0.05	0.05
CHN	606	2.93	0.65	2.90	0.61	-0.05	0.44	0.66[a]	74.09	99.50	0.66[a]	3.04	0.50	0.18[a]	0.36	0.48[a]	70.79	97.19	0.53[a]	-0.18[a]	-0.13[a]
GBR	400	2.88	0.77	2.85	0.76	-0.05	0.44	0.68[a]	65.75	99.00	0.68[a]	2.75	0.76	-0.17[a]	0.37	0.65[a]	59.75	98.75	0.73	-0.03	0.05
TWN	313	3.06	0.70	3.15	0.70	0.12	0.56	0.73	76.68	99.36	0.73	3.19	0.63	0.19[a]	0.40	0.64[a]	66.77	99.36	0.73	-0.09	0.00
FRA	238	2.79	0.77	2.80	0.70	0.02	0.39	0.65[a]	64.29	99.16	0.65[a]	2.91	0.62	0.18[a]	0.41	0.63[a]	66.39	98.74	0.75	-0.02	0.10
KAZ	225	2.51	0.88	2.48	0.90	-0.03	0.46	0.72	63.56	98.22	0.72	2.33	1.02	-0.18[a]	0.43	0.76	59.11	99.56	0.82	0.04	0.10
PHL	201	3.01	0.98	2.98	0.97	-0.04	0.47	0.78	62.19	99.00	0.78	2.83	0.94	-0.19[a]	0.47	0.77	61.69	98.01	0.81	-0.01	0.03
PAK	120	2.19	0.77	2.15	0.81	-0.05	0.52	0.72	68.33	99.17	0.72	2.21	0.86	0.02	0.50	0.74	66.67	100.00	0.73	0.02	0.01
IDN	118	3.08	0.71	3.07	0.57	-0.03	0.28	0.54[a]	62.71	100.00	0.56[a]	3.20	0.52	0.19[a]	0.50	0.61[a]	72.88	99.15	0.69[a]	0.07	0.13
USA	109	3.76	0.83	3.57	0.88	-0.23[a]	0.43	0.55[a]	63.30	94.50	0.56[a]	3.51	0.55	-0.36[a]	0.33	0.59[a]	59.63	98.17	0.73	0.04	0.17

Note. Std.diff. = standardized difference on mean scores; Wtd. kappa = quadratic-weighted kappa; Corr. = correlation coefficient; % agree = exact percentage agreement; % adj agree = 1-point adjacent percentage agreement; KOR = Korea; JPN = Japan; IND = India; CHN = China; GBR = Great Britain; TWN = Taiwan; FRA = France; KAZ = Kazakhstan; PHL = the Philippines; PAK = Pakistan; IDN = Indonesia; USA = United States.

[a]Shaded values did not meet the evaluation thresholds.

Table 7

Correlation Matrix for External-Relations Aspect of Validity

	N	e-rater	Human1	Item 8 operational raw score	Sum of Items 1–5	Sum of Items 6 and 7	Weighted sum of Items 1–7	Reported writing score	Reported speaking score
E-rater	22,835	1	0.75	0.80	0.46[a]	0.55[a]	0.61[a]	0.81	0.74[a]
Human1	22,835	-	1	0.92	0.40[a]	0.47[a]	0.52[a]	0.84	0.65[a]
Item 8 operational raw score	22,835	-	-	1	0.44	0.50	0.56	0.91	0.70
Sum of Items 1– 5	22,835	-	-	-	1	0.41	0.80	0.57	0.43
Sum of Items 6 and 7	22,835	-	-	-	-	1	0.87	0.80	0.52
Weighted sum of Items 1–7	22,835	-	-	-	-	-	1	0.83	0.57
Reported writing score	22,835	-	-	-	-	-	-	1	0.72
Reported speaking score	21,908	-	-	-	-	-	-	-	1

[a] Shaded values indicate that e-rater correlates with external linguistic measures more strongly than Human1 does.

E-rater correlated consistently higher than Human1 did with the other seven writing items as well as test takers' speaking scores, all of which can be treated as external measures of English linguistic ability (see shaded values in Table 7). E-rater did not correlate as highly as humans did with the raw and final scaled writing scores, which is partly due to the fact that both criterion variables were computed based on the human ratings.

PS (Traditional) Model Performance

PS (traditional) models were evaluated on the remaining data in a single prompt. For direct comparison purpose, Table 8 shows the evaluation results for the same set of prompts as for the generic scoring model (shown in Table 5) and Appendix H gives the results for all 50 prompts. Note that sampling variation led to variations in the interhuman agreement, although the differences were generally small on both human ratings distributions and interhuman agreements.

Speaking of the six prompts shown in the table only, e-rater scores resulting from PS (traditional) models correlated with humans ratings slightly more strongly than those derived from the G model. All but two values (i.e., standardized mean score difference on C5923 and adjacent percentage agreement on C4086) were more preferable with the PS (traditional) model than with the G model. The number of indices that failed to meet evaluation thresholds was also smaller for the PS (traditional) model. Finally, none of the six prompts violated the degradation threshold, implying that the quality of the e-rater scores is comparable to a human rating.

Speaking of all 50 prompts that were examined for PS (traditional) scoring, only one prompt (i.e., C5923 in Table 8) showed substantively large distributional discrepancies between e-rater scores and human ratings. All prompts met the evaluation threshold regarding degradation. Nonetheless, similarly to the evaluation results of the G model, 32 of 50 prompts had a quadratic-weighted kappa value lower than 0.70 and 15 prompts failed to meet the evaluation threshold for correlation coefficient.

Table 8

PS (Traditional) Model Evaluation on Selected Prompts

Prompt	*N*	Human1		Human2		Human1/Human2 agreement statistics						e-rater		Human1/e-rater agreement statistics						Degradation (H1/e-rater–H1/H2)	
		Mean	SD	Mean	SD	Std. diff.	Kappa	Wtd. kappa	% agree	% adj agree	Corr.	Mean	SD	Std. diff.	Kappa	Wtd. kappa	% agree	% adj agree	Corr.	Wtd. kappa	Corr.
C4052	340	3.00	0.81	2.97	0.79	-0.03	0.47	0.72	66.18	99.41	0.72	3.02	0.81	0.02	0.54	0.78	69.12	99.71	0.81	0.06	0.09
C4084	255	2.43	0.94	2.57	0.98	0.15[a]	0.46	0.76	61.96	97.65	0.77	2.46	0.92	0.04	0.38	0.69[a]	56.08	96.47	0.73	-0.07	-0.04
C4086	268	3.16	0.73	3.08	0.67	-0.12	0.42	0.67[a]	67.54	100	0.68[a]	3.18	0.68	0.02	0.48	0.64[a]	69.03	98.13	0.68	-0.03	0.00
C5923	258	3.31	0.69	3.25	0.68	-0.08	0.53	0.70	73.26	99.61	0.70	3.20	0.64	-0.16[a]	0.54	0.72	73.26	100	0.75	0.02	0.05
E1577	331	2.80	0.65	2.71	0.66	-0.14	0.50	0.70	74.02	100	0.70	2.75	0.64	-0.08	0.52	0.71	75.23	100	0.76	0.01	0.06
E4365	202	2.70	0.64	2.72	0.63	0.03	0.56	0.72	77.23	100	0.72	2.71	0.63	0.02	0.52	0.69[a]	73.76	99.50	0.75	-0.03	0.03
Unweighted average (across all 50 prompts)	283.8	2.99	0.71	2.99	0.70	-0.01	0.46	0.66[a]	69.56	99.38	0.67[a]	2.99	0.69	0.00	0.45	0.67[a]	67.71	99.53	0.72	0.01	0.05

Note. Full account of the results for generic model evaluation is given in Appendix H; Std. diff. = standardized difference on mean scores; Wtd. kappa = quadratic-weighted kappa; Corr. = correlation coefficient; % agree = exact percentage agreement; % adj agree = 1-point adjacent percentage agreement.

[a] Shaded values did not meet the evaluation thresholds.

In comparison with the G model, the PS (traditional) model produced e-rater scores that more strongly agreed with human ratings for 33 of 50 prompts as indicated by quadratic-weighted kappa, for 27 prompts as indicated by correlation coefficient, for 32 prompts as indicated by exact percentage agreement, and for 27 prompts as indicated by adjacent percentage agreement. (See Appendix H for detailed documentation of results regarding PS[traditional] models.)

PS (Press) Model Performance

Our concern with regard to the small prompt size was confirmed via PS (press) model analysis. Fifty-three of 63 prompts did not meet the evaluation threshold for inflation index (II), indicating that they did not have sufficient prompt size for prompt-based model calibration. Most prompts were on the borderline, with an II value of 0.05 to 0.07. The average magnitude of II was 0.06 across all prompts.

Results showed that PS (press) models were able to generate e-rater scores that valued equivalently to one or more human ratings for all 63 prompts, with a mean VI index value of 2.15 across 63 prompts. All prompts met the evaluation criterion in terms of VI, for which value should be at least 1.0 to indicate a satisfactory model performance. This result, in part, indicated that e-rater was more reliable, in many cases far more reliable, than human raters.

The capability of the e-rater scores in predicting a test taker's true score was acceptable for all but four prompts. The average value for the prediction index (PI) was 0.79, with the lowest being 0.52 and highest being 0.93.

However, findings regarding VI and PI can be most appropriately interpreted as relative to the quality of human ratings. Similar to findings regarding human rating quality in G and PS (traditional) model analyses, nearly half of the prompts (i.e., 30 of 63) did not show adequate human rating quality as measured by HQI.

Table 9 presents the results of PS (press) model evaluation on the same selected prompt set as for G and PS (traditional) model evaluation (see Appendix I for full account of the results).

Table 9

PS (Press) Model Evaluation on Selected Prompts

Prompt	*N*	Human-rating quality index	Value index	Inflation index	Prediction index
C4052	689	0.85	2.72	0.04	0.88
C4084	509	0.87	0.69[a]	0.03	0.70
C4086	529	0.81	1.12	0.04	0.71
C5923	518	0.80	1.79	0.05[a]	0.78
E1577	654	0.82	3.01	0.05[a]	0.87
E4365	401	0.82	1.88	0.07[a]	0.82

Note. See Table 4 for evaluation criteria and see Appendix I for full account of the results.

[a] Shaded values did not meet the evaluation criteria.

Conclusions About Model Development and Next Steps

Generic scoring has an operational strength that is absent for prompt-based scoring. That is, a generic scoring model can be applied to grade prompts that are not included in the calibration sample. This scalability puts generic scoring at an advantage in large-scale assessments that have a relatively high prompt turnover, such as the writing assessment in the testing program investigated in this study. Empirical results from this study, coupled with the recommendations from previous research (e.g., Attali & Burstein, 2006), indicated that a small number of responses per prompt could lead to unreliable prompt-based scoring models. Additionally, the evidence for the generic scoring model suggested that e-rater agreed with human raters on the same level as human raters agreed with one another, with negligible degradation. Therefore, even though the generic model performance was not satisfactory for all prompts, we moved forward with the generic model. Further discussion on the results of model development is provided in the Discussion section.

As a next step, it is critical to determine the impact of using e-rater automated scoring on test takers' essay scores and their final writing scores. The next section provides research details in conducting the impact analysis for the generic scoring model.[6]

Impact Analysis on Automated Scoring Implementation

Purpose and Procedure of Impact Analysis

The impact analysis provides the extent of changes in scores on item and form levels for overall and subgroup populations after automated scoring is implemented. Evidence collected from impact analysis can serve to strengthen or weaken the evidence from the empirical model evaluation. For example, if replacing one human rater with e-rater has limited impact on a candidate's final grade, it can be further verified that e-rater scores entail the same value as at least one human rating.

There are several common practical approaches to implementing an automated scoring system in an assessment, including using automated scoring in conjunction with human raters (Zhang, 2013). When automated scoring contributes to the final composite scores, developers and testing programs can choose different weighting strategies of human and automated scores based on their confidence in each scoring method and the strengths of empirical evidence.

In this study, we examined the most commonly used implementation method involving substituting an e-rater score for one of the two human ratings. Under the assumption that an automated score can replace a human rating, we simulated the final composite scores (simulated scores) as the unweighted average of one human rating and one automated score, that is, $(h1 + erater)/2$.

We evaluated the correlational strength of the simulated scores with relevant external measures (i.e., test takers' scores on the other writing items and their speaking scores), and compared them with the correlational strength of human ratings with the same variables.

Subsequently, the amount of changes in scores on the task and form levels after using simulated scores was examined. To ensure fairness, the impact of e-rater application on the different demographic groups was also investigated.

One other step in the impact analysis is to establish an *adjudication threshold*. Adjudication threshold refers to the largest tolerable discrepancy between automated and human scores in operation. When the discrepancy exceeds a predetermined threshold, additional graders are called in to reconcile it. In this sense, having a threshold is a quality-control mechanism for both human and e-rater scoring. In this report, we present results based on thresholds of 1.0 and 1.5. Both thresholds were chosen to be consistent with the current e-rater implementation approach in the TOEFL iBT program, in which e-rater scores contribute to the final grade.

Data Set for Impact Analysis

The complete data set of 32,835 responses was used to gather evidence for the external-relations aspect of validity and the changes in operational scores. The e-rater scores were produced from the generic scoring model. Simulated scores were computed using precise e-rater scores that were truncated to within the scale of the scoring rubric.[7]

Results of the Impact Analysis

Association With External Variables

Table 10 presents the correlation coefficients for simulated scores and human ratings with external measures. Simulated scores correlated consistently more strongly than human ratings did with the other writing or linguistic measures, suggesting support for the external-relations aspect of validity.

Table 10

Correlation Coefficients of Simulated and Human Scores With External Measures

Scores	With sum of Items 1 to 5	With sum of Items 6 and 7	With weighted sum of Items 1 to 5 and 2 x (Items 6 and 7)	Reported speaking score
Simulated	0.46	0.55	0.60	0.74
Human	0.44	0.51	0.51	0.70

Note. Simulated scores were calculated as the unweighted average of one human rating and one automated score. Human scores were the unweighted average of the two human ratings.

Impact on Item and Form Levels

In general, the use of e-rater suggested minimal impact on test takers' essay scores, final raw writing scores, and final scaled writing scores.

Table 11 shows that the agreement between human and simulated scores for both 1.0 and 1.5 adjudication thresholds at the item level. The agreement level between the simulated scores and human ratings was fairly high, partly because simulated scores included the influence of one human rating. However, it was encouraging to find that replacing one human rating with an automated score would result in nearly identical final scores to those based on two human ratings. Additionally, little difference was found between the two adjudication thresholds.

Table 11

Association Between Simulated e-rater and Human Scores on the Item Level

	Association between simulated e-rater and human scores				
	Using precise e-rater scores		Using integer e-rater scores		
Adjudication threshold	Pearson correlation coefficient	Standardized mean score difference	Quadratic-weighted kappa	Exact % agreement	Adjacent % agreement
1.0	0.95	0	0.94	70.4	99.9
1.5	0.95	0	0.93	68.7	99.9

Less than 1% of the test takers' essay raw scores would change by more than 0.5 point with e-rater implementation after applying a threshold of either 1.0 or 1.5 (Table 12). Table 13 shows that, on the total writing raw score level, approximately 0.5% and 0.7% of test takers' final writing raw scores would change by more than 2 points for thresholds of 1.0 and 1.5, respectively, with e-rater implementation. Finally, around 3.8% and 4.6% of test takers' final scaled writing scores would change by more than 10 points with a threshold of 1.0 and 1.5, respectively, after e-rater implementation (Table 14).

Table 12

Changes in Percentage on Raw Item Scores With e-rater Implementation

Adjudication threshold	Reported essay score minus simulated essay score (%)								
	≤ -2	≤ -1.5	≤ -1	≤ -0.5	0	≤ 0.5	≤ 1	≤ 1.5	> 1.5
1.0	0.1	0.1	0.3	15.6	67.3	16.5	0.2	0	0
1.5	0.1	0.1	0.3	15.4	68.0	16.0	0.2	0	0

Table 13

Changes in Percentage on Final Raw Writing Scores With e-rater Implementation

Adjudication threshold	Final raw writing score minus simulated writing score (%)								
	≤ -8	≤ -6	≤ -4	≤ -2	0	≤ 2	≤ 4	≤ 6	> 6
1.0	0	0	0.1	0.6	54.0	44.8	0.4	0	0
1.5	0	0	0.1	0.7	54.1	44.5	0.6	0	0

Table 14

Changes in Percentage on Final Scaled Writing Scores With e-rater Implementation

Adjudication threshold	Reported final scaled score minus simulated scaled essay score (%)								
	-40	-30	-20	-10	0	10	20	30	40
1.0	0.1	0.1	1.7	25.5	49.6	21.0	1.9	0	0
1.5	0.1	0.1	2.1	24.8	50.3	20.2	2.3	0	0

Impact on Individual Test Country/Territory Groups

The impact of using e-rater on individual test country/territory groups is also minimal. High correlation coefficient and quadratic-weighted kappa were found between simulated scores and human ratings for both adjudication thresholds of 1.0 and 1.5 for all country/territory groups. The quadratic-weighted kappa ranged from 0.85 to 0.95 for a threshold of 1.0 and ranged from 0.82 to 0.93 for a threshold of 1.5 across population groups. The correlation coefficient spanned from 0.88 to 0.97 for a threshold of 1.0 and ranged from 0.87 to 0.96 for a threshold of 1.5.

The only index value that raised concern was the standardized mean score difference for Canada (see Appendix K). This finding was likely due to the low variance in human and simulated scores, and/or the lack of examinee population. Table 15 shows the results for the six largest population groups and Appendix K provides a full account of the results.

Table 15

Impact Analysis Results for Large Test Country/Territory Population Groups

Adjudication threshold	Test country/ territory	*N*	Operational essay score		Operational essay score by simulation score (normally rounded to integer)					Operational essay score by simulation score (unrounded)			
					Simulated score		Agreement index			Simulated score		Agreement index	
			Mean	SD	Mean	SD	Wtd. kappa	% agree	% adj. agree	Mean	SD	Std. dif.	Corr.
1.0	KOR	18,288	2.91	0.64	2.93	0.68	0.92	72.22	99.92	2.92	0.66	0.02	0.93
	JPN	7,505	2.73	0.70	2.75	0.74	0.93	71.90	99.96	2.71	0.72	-0.04	0.94
	IND	4,522	3.66	0.57	3.69	0.64	0.85	59.55	99.93	3.66	0.51	0.00	0.88
	CHN	968	2.90	0.61	2.94	0.63	0.91	74.48	99.90	2.93	0.60	0.05	0.93
	GBR	601	2.81	0.75	2.86	0.79	0.93	68.39	100.00	2.77	0.76	-0.04	0.95
	TWN	493	3.15	0.67	3.14	0.72	0.92	74.44	99.80	3.16	0.65	0.02	0.94

Adjudication threshold	Test country/ territory	*N*	Operational essay score		Operational essay score by simulation score (normally rounded to integer)					Operational essay score by simulation score (unrounded)			
					Simulated score		Agreement index			Simulated score		Agreement index	
			Mean	SD	Mean	SD	Wtd. kappa	% agree	% adj. agree	Mean	SD	Std. dif.	Corr.
1.5	KOR	18,288	2.91	0.64	2.91	0.69	0.90	70.74	99.90	2.92	0.66	0.02	0.92
	JPN	7,505	2.73	0.70	2.73	0.75	0.91	70.42	99.91	2.70	0.72	-0.04	0.93
	IND	4,522	3.66	0.57	3.67	0.61	0.82	57.74	99.91	3.66	0.50	0.00	0.87
	CHN	968	2.90	0.61	2.94	0.62	0.88	72.93	99.59	2.94	0.58	0.07	0.91
	GBR	601	2.81	0.75	2.82	0.78	0.91	65.89	100.00	2.76	0.76	-0.06	0.94
	TWN	493	3.15	0.67	3.15	0.71	0.91	72.82	99.80	3.16	0.64	0.03	0.93

Note. Simulated scores were computed as (e-rater + Human1)/2; Wtd. kappa = quadratic weighted kappa; % agree = exact percentage agreement; % adj. agree = 1-point adjacent percentage agreement; Std. dif. = standardized mean score difference; Corr. = correlation coefficient; KOR = Korea; JPN = Japan; IND = India; CHN = China; GBR = Great Britain; TWN = Taiwan.

Discussion

In this study, we explored the possibility of engaging e-rater to grade an essay item in a writing assessment in a large-scale English language testing program. We focused on three types of automated scoring approaches: the generic model, prompt-specific models built with a traditional split-in-half fashion (PS [traditional]), and prompt-specific models built with a leave-one-out fashion (PS [press]).

Each modeling method was examined based on e-rater's agreement with human ratings. Agreement was indicated by standardized difference on mean scores, correlation coefficient, quadratic-weighted kappa, percentage agreement, as well as the degradation in correlation coefficient and quadratic-weight kappa that resulted from e-rater application for G and PS (traditional) model evaluation and was indicated by three PRESS statistic derived indices for PS (press) model evaluation.

In the impact analysis, we simulated test takers' writing scores as if e-rater was to replace one of the two human graders. Automated scores were generated by the G model identified in the

model development stage. We used the equal-weighting method for automated and human scores to produce the simulated scores, which aligns with e-rater operational practice in a comparable large-scale, international language assessment administered at ETS. We examined the association of the simulated scores with human ratings and with external variables, as well as the prospected changes in test takers' scores on both item and form levels.

Discussion on Model Development

For all three modeling approaches, e-rater scores achieved an overall comparable—or even higher—level of agreement with human ratings, compared with the agreement between two human ratings. Previous research yielded similar findings; that is, e-rater tends to perform as well as, or better than, humans in grading general writing prompts, such as the independent writing task in the TOEFL iBT test and the issue writing task in the GRE General Test (e.g., Ramineni et al., 2012a and Ramineni et al., 2012b). The essay item in the writing assessment investigated in this study falls into this category of general writing prompts.

Results of this study also showed that, for the population as a whole, the PS (traditional) models performed better than the G model, producing a 0.01 magnitude increase in correlation coefficient and a 0.02 magnitude increase in quadratic-weighted kappa (see the bottom rows in Tables 5 and 8).[8] For individual prompts, the G model performed slightly worse than PS (traditional) models on all indices and noticeably worse on standardized mean score difference. Table 16 gives the number of prompts that failed to meet the model evaluation criteria identified in the method section.

Table 16

Counts of Prompts That Did Not Meet Evaluation Thresholds

Model	Agreement indices for human and e-rater scores				Degradation in	
	Correlation coefficient	Weighted kappa	Standardized difference	Correlation & kappa	Weighted kappa	Correlation coefficient
G	17	35	14	17	2	1
PS (traditional)	15	32	1	15	0	0

Note. To make the counts comparable, all counts are based on 50 prompts evaluated for both generic and PS (traditional) scoring. G = generic; PS = prompt-specific.

PS (traditional) scoring is likely to produce e-rater scores that associate with human ratings more strongly than G scoring, for at least two reasons. First, the PS (traditional) models apply more text features than the G models (i.e., the inclusion of the two vocabulary-based topic-specific features), which help account for more variance in the human ratings and, as a result, generate higher quality automated scores; although here, quality only refers to the extent of agreement between automated and human scores. Second, unlike the G model, prompt customized models are not affected by the potential (sometimes large) fluctuation in prompt difficulty across a number of prompts. In cases where prompts differ dramatically in their characteristics (e.g., difficulty, topical area that may draw longer or shorter responses), a prompt-specific scoring approach is likely to be more preferable than generic scoring. When prompts are similar in their characteristics, the advantages of PS (traditional) scoring over generic scoring can be diminished.

Our subsequent investigation of the prompt effect was complicated by the fact that, for test security purposes, prompts were not designed to be randomly administered across test countries/territories in the program used in this study. As a result, many prompts were only responded to by test takers from a limited number of countries. Nonetheless, we were able to extract two data sets that contained several prompts administered in more than three test countries/territories, which allowed us to examine the test country/territory and prompt effects using analysis of variance (ANOVA). One data set had three prompts and 14 test countries/territories and the other had six prompts and three test countries/territories. We used the mean of two human ratings as the response variable. Results showed significant test country/territory effect (i.e., $F = 21.79$, $df = 13$, $p < 0.01$ and $F = 307.72$, $df = 2$, $p < 0.01$, respectively, for the two data sets) and inconsistent prompt effect (i.e., $F = 2.85$, $df = 2$, $p = 0.06$ and $F = 25.76$, $df = 5$, $p < 0.01$, respectively, for the two data sets). Of note is that results from both data sets revealed a small effect size for prompt effect (i.e., $\eta^2 = 0.02$ and 0.06) and a large effect size for test country/territory effect ($\eta^2 = 0.43$ and 0.23; see Appendix J for detailed documentation of variance components).

Why did the PS (traditional) model greatly outperform the G model in terms of standardized mean score difference on prompt level? This result could be due to the significantly large test country/territory effect discussed above. As mentioned previously, the prompts were not randomly administered throughout the global administrations and, as a result, a majority of

the prompts were only answered by examinees from a subset of test countries/territories. So, in a sense, the test country/territory was nested within the prompt. Table 6 shows that the ability distribution across test country/territory can vary drastically (e.g., the mean of Human1 equals to 3.69 for test takers from India, whereas the mean equals to 2.77 for test takers from Japan). To the extent that the G model is not customized for individual prompts that mostly were only administered in particular countries/territories, it is likely that the resulting automated score distribution on a prompt level would not emulate human rating distribution well. It is also possible that this result is due to prompt effect, because we were only able to examine nine of 63 prompts and the small prompt effect discovered via ANOVA may not be generalizable to all essay prompts administered in this testing program examined in this study. Unfortunately, the limitation of our data set due to the nonrandomization in prompt allocation would not allow for further investigation of prompt or country/territory effects.

As for PS (press) model evaluation, although a different set of evaluation indices was used because of the way the models were constructed, those indices cover similar aspects of model evaluation to the traditional metrics. An important finding derived from PS (press) model evaluation was that the prompt sizes were generally too small to justify reliable prompt-based scoring. Previous research recommended 500 as a minimal calibration sample size for prompt-specific scoring in operational practice for consequential uses (Attali & Burstein, 2006). In this study, only 54% of the prompts entailed a prompt size greater than 500. If using PS (traditional) scoring (which further split the prompt size into halves), none of the prompts would have a calibration sample size greater than 500.

Another finding from this study is that the level of human/e-rater agreement was not consistent across different prompts or test country/territory groups. However, taking results from the G model evaluation as an example, it appeared that the human/human agreement varied across prompts and across population groups as well (see Tables 5, 6, and 8), and the agreement between the resulting e-rater scores and human ratings varied accordingly. This phenomenon is most likely due to the circularity effect of predicting human ratings in automated scoring; that is, in prediction-based scoring approach, the quality of the automated scores is constrained by the quality of the human ratings. Similar findings were reported in Ramineni et al. (2012a, 2012b).

A more fundamental question is: Why did human/human agreement vary greatly across prompts? As noticed, the interhuman agreement on many prompts or certain population groups

was fairly low. In fact, like phenomena also occurred in other large-scale, international English language assessments, for example, TOEFL iBT (Bridgeman, Trapani, & Attali, 2012) and the GRE Revised General Test (Bejar, Joe, Feng, Zhang, & Sands, 2013). It is not entirely clear why human raters seem to be less able to score essays submitted by test takers bearing certain demographic backgrounds. However, a plausible reason is that raters do not have the same understanding on certain writing features common within a demographic group. As a result, different raters judge the essay quality along different scales. As an example, essays submitted by test takers in the United States tend to have a smaller amount of shell language than non-U.S. submissions (Bejar, Van Winkle, Madnani, Lewis, & Steier, 2013). Some raters may penalize the use of shell language more, thereby compromising the interrater agreement. In the testing program used in this study, test country/territory, to a large extent, is nested within prompts, which might have led to the uneven interrater agreement levels across prompts. However, one should not rule out a possibility that such variations on human agreement could be due to prompt effect. Using GRE argument prompts, Joe, Park, Brantley, Lapp, and Leusner (2012) found that some prompts demanded more cognitive resources from human raters and, as a result, tended to receive a lower interhuman agreement.

A related issue to the lack of invariance of interhuman agreement across prompts and test countries/territories is that the prompts could be administered to nonrandom test takers in the testing program. This issue is more likely to occur in tests administered in Program-K (where the test takers are from the same company, for example) than in Program-U (where the test takers are independent from one another, as in many large-scale, high-stakes assessments). So, in Program-K, a three-level nesting structure is potentially formed (Figure 1).

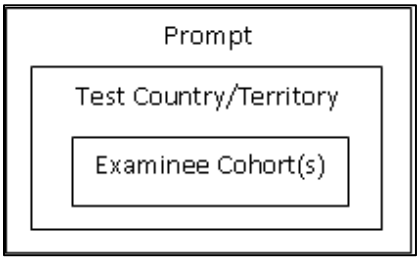

***Figure 1*. Nesting among examinee, test country/territory, and prompt.**

The homogeneity of the test taker population could manifest itself as small variations found in human ratings on the prompt level. Because human ratings are treated as a predicted dependent variable in the multiple linear regression-based scoring approach, a small variance in the human rating would reduce the effectiveness of the scoring model. As mentioned previously, more than half of the prompts were only given in one or two test countries/territories (mainly Korea and Japan), and only nine prompts were administered in more than two countries/territories. Furthermore, it is also likely that a prompt was only administered to an examinee cohort from one company in tests administered under Program-K. Therefore, the small human rating standard deviation reported in Appendices C and D on the prompt level could be largely due to the prompt administration design in the testing program. Although Program-K was the most popular in Korea, the largest test taker population in our data set, the small variation in human ratings for the Korean population (i.e., mean = 2.92, SD = 0.67, shown in Table 6) seemed to have noticeable impact on the overall automated scoring model performance (e-rater/human quadratic-weighted kappa = 0.65 from Table 5).

Discussion on Impact Analysis

For the impact analysis, we chose the G model to be the operational scoring model primarily for the following two reasons. One, the evidence gained from model development suggested potential inappropriateness to construct prompt-based scoring models due to a lack of prompt size. Two, because prompts were not randomly assigned that linked to the nesting issue discussed above, the generalizability of prompt-specific models is questionable.

Impact analysis results revealed that the simulated writing scores with e-rater (G model) implementation correlated highly with human ratings, which implied that automated scores can take the place of one of the two human ratings. This finding further verified the empirical evidence collected during automated scoring model development that e-rater agreed with human rater as strongly as human raters agreed with one another.

Simulated scores also entailed stronger association with external language-based measures than human ratings did. This result could be largely due to the absolute scoring consistency introduced by automated scoring, in that automated scoring applies the same rule across all essays whereas human rating potentially suffers from a variety of human errors, including inconsistency errors (Zhang, 2013). Similar results were reported in Bridgeman and

Trapani (2011) for the TOEFL iBT program, where the authors found that a hybrid score from human and automated scoring has stronger external relations with other linguistic variables.

No material difference was observed between the two adjudication thresholds. This, in part, indicates that it is rare to have automated scores and human ratings differ by more than 1 score point, otherwise such discrepancies would have been reflected in a higher percentage of score change for Threshold 1.5.

Finally, given the evidence in model evaluation that suggested close comparability between e-rater scores and human ratings, it was not a surprise to find that the implementation of automated scoring had minimal impact on test takers' scores on both the task and form levels.

Limitations

The findings of this research are restricted by the following limitations:

First, the findings may not be applicable beyond the essay prompts, writing genre, testing program type, and test taker population composition similar to the ones examined in this study.

Second, findings are not generalizable to essays that were flagged by any type(s) of advisory flags available in e-rater. Those essays were excluded from the model calibration or evaluation analysis. More research is needed to determine the effectiveness of e-rater advisory flags in the writing assessment for this particular testing program.

Three, test takers' scores on the other writing items and the speaking test were the only external variables available in this study. Although providing reasonably sufficient evidence in the external relation aspect of the validity for both human and automated scores, use of different external variables may yield different results.

Four, due to the small prompt sizes, the prompt-specific scoring approaches (i.e., traditional or press) may not have been evaluated reliably for a number of prompts. A greater prompt size may offer similar results or produce competing evidence.

Five, the findings are bounded by the characteristics of the human ratings in the writing assessment in this study. The lack of variance in human ratings, possibly due to the nature of the writing program and administration design of the essay prompts, had impacts on the linear regression-based automated scoring model calibration. Therefore, depending on the intended use of the test scores, e-rater scores produced by emulating human ratings may not effectively differentiate test takers on their writing proficiency.

Recommendations for Additional Research

This research investigated the suitability of applying e-rater automated scoring to grade the essay item in a writing assessment from a language-testing program. Derived from the findings of this study, we suggest the following research topics that can help collect further evidence to support the operational implementation of e-rater in this particular writing assessment.

First, it will be necessary to further evaluate the effectiveness of the generic scoring model on more data sets. The population composition in the writing assessment in this particular testing program can change frequently. Therefore the model constructed on the current population may not be applicable to a new population. Research can start with the 2011–2012 data set (which was not processed by the e-rater scoring system during this research) and advise alternative modeling approaches if needed.

Second, small human-rating variability was observed in a number of prompts and test countries/territories. Additionally, interhuman agreement was found to be varying across prompts and country/territory groups. Researchers are, therefore, advised to consult and collaborate with testing program administrators and to conduct further studies to understand the relationship of these human ratings issues with (a) writing prompt administration design and (b) the test taker population composition. Results of this line of research may lead to revised prompt administration design that is more suitable for automated essay scoring than the current design.

Third, researchers should investigate the effectiveness of e-rater advisory flags in identifying problematic essays in the writing program investigated in this study. E-rater is generally more efficient in identifying certain problems in writing, such as repetitive use of language, than humans are. Essays that are excessively long or short tend to receive invalid automated scores from regression models that award high weights to features correlated with length. On the other hand, humans are much more capable than the machine of judging the concept relevance of an essay. Ideally and statistically speaking, the agreement between human and e-rater should be low for flagged essays as an indication of the effectiveness of e-rater flagging.

References

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

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

Attali, Y., & Powers, D. (2009). Validity of scores for a developmental writing scale based on automated scoring. *Educational and Psychological Measurement*, *69*(6), 978–993.

Beigman Klebanov, B., & Flor, M. (2013). Word association profiles and their use for automated scoring of essays. *Proceedings of the 51st annual meeting of the Association for Computational Linguistics* (pp. 1148–1158). Retrieved from http://aclweb.org/anthology//P/P13/P13-1113.pdf

Bejar, I. I., Joe, J., Feng, G., Zhang, M., & Sands, A. (2013). *Peeking into the black box: A study of rater cognition by means of eye tracking.* Manuscript submitted for publication.

Bejar, I. I., Van Winkle, W., Madnani, N., Lewis, W., & Steier, M. (2013). *Length of textual response as a construct-irrelevant response strategy: The case of shell language* (Research Report No. RR-13-07). Princeton, NJ: Educational Testing Service.

Bennett, R. E., & Bejar, I. I. (1998). Validity and automated scoring: It's not only the scoring. *Educational Measurement: Issues and Practice*, *17*, 9–17.

Ben-Simon, A., & Bennett, R. E. (2007). Toward more substantively meaningful automated essay scoring. *The Journal of Technology, Learning, and Assessment*, *6*(1).

Bridgeman, B., & Trapani, C. (2011, April). *The question of validity of automated essay scores and differentially valued evidence.* Paper presented at the annual meeting of the American Educational Research Association and the National Council of Measurement in Education, New Orleans, LA.

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

Crocker, L., & Algina, J. (2006). *Introduction to classical and modern test theory*. Mason, OH: Wadsworth.

ETS. (2010). *Frequently asked questions about TOEFL practice online*. Retrieved from http://www.ets.org/s/toefl/pdf/toefl_tpo_faq.pdf

ETS. (2013a). *How the test is scored*. Retrieved from http://www.ets.org/gre/revised_general/scores/how/

ETS. (2013b). *Understanding your TOEFL iBT test scores*. Retrieved from http://www.ets.org/toefl/ibt/scores/understand

ETS. (2013c). *Criterion online writing evaluation*. Retrieved from http://www.ets.org/Media/Products/Criterion/topics/topics.htm

Guilford, J. P., & Fruchter, B. (1973). *Fundamental statistics in psychology and education* (5th ed.). New York, NY: McGraw-Hill.

Haberman, S. J. (2011). *Use of e-rater in scoring the TOEFL iBT writing test* (Research Report No. RR-11-25). Princeton, NJ: Educational Testing Service.

Haberman, S. J., & Sinharay, S. (2008). *Sample-size requirements for automated essay scoring* (Research Report No. RR-08-32). Princeton, NJ: Educational Testing Service.

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

Joe, J., Park, Y. S., Brantley, W., Lapp, M., & Leusner, D. (2012, April). *Examining the effect of prompt complexity on rater behavior: A mixed-methods study of GRE Analytical Writing Measure Argument prompts*. Paper presented at the annual meeting of the National Council on Measurement in Education, Vancouver, Canada.

Kutner, M. H., Nachtsheim, C. J., Neter, J., & Li, W. (2005). *Applied linear statistical models* (5th ed.). Burr Ridge, IL: Mc-Graw-Hill/Irwin.

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

Ramineni, C., Trapani, C., Williamson, D., Davey, T., & Bridgeman, B. (2012b). *Evaluation of e-rater® for the TOEFL® independent and integrated prompts* (Research Report No. RR-12-06). Princeton, NJ: Educational Testing Service.

Weisberg, S. (1985). *Applied linear regression* (2nd ed.). New York, NY: John Wiley.

Williamson, D. M. (2013). The conceptual and scientific basis for automated scoring of performance items. In R. W. Lissitz & H. Jiao (Eds.), *Computers and their impact on state assessments* (pp. 157–193). Charlotte, NC: Information Age Publishing.

Williamson, D. M., Mislevy, R. J., & Bejar, I. I. (Eds.). (2006). *Automated scoring of complex tasks in computer-based testing*. Mahwah, NJ: Erlbaum.

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

Zhang, M. (2013). Contrasting automated and human scoring of essays. *R & D Connections*, *21*. Princeton, NJ: Educational Testing Service.

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

Notes

[1] E-rater Engine 11 was used in this study.

[2] Suppose a prompt has *n* responses, (*n* - 1) essays were used for model building, and this process was iterated *n* times for each prompt.

[3] In the testing program used in this study, selected prospective raters are first trained to use a scoring rubric and grade the writing items via the ETS online scoring network. Subsequently, raters must pass a certification test composed of a set of real responses in order to grade operationally. Finally, raters must pass a calibration test each time they log into the scoring system prior to operational grading.

[4] Further investigation revealed that a restriction of range in the human rating on the prompt level might have contributed to the low interhuman agreement. Homogeneity in writing ability of the test taker population (due to the prompt administration design) might have caused the low spread in human rating distribution. This issue is discussed later in this paper in the Discussion section.

[5] The population composition in the testing program may change drastically over time. Therefore the results may not be generalizable beyond the examinee population used in this study.

[6] It is worth noting that, in the research process, we examined several other modeling methods that are not common in operational e-rater deployment, including cumulative logistic regression (used in Haberman & Sinharay, 2010) and equal-weight scoring (used in Attali, 2007). However, although each of those methods has its own merits, we did not find noticeable improvements in terms of the e-rater–human association. Coupled with a lack of supporting evidence in the existing literature for these uncommon model calibration approaches, we concluded that the generic model developed in this investigation was our best candidate to be the scoring model for operational impact analysis.

[7] The truncation method was the same as that used in producing quadratic-weighted kappa.

[8] Results reported were based on 63 prompts for the G model and 50 prompts for the PS (traditional) model.

List of Appendices

Page

Appendix A

Features in the e-rater Automated Essay Scoring System

Model	Feature	Descriptions
G and PS	Grammar	Errors in pronouns, run-ons, missing possessives, etc.
	Mechanics	Errors in capitalization, punctuation, commas, hyphens, etc.
	Style	Errors in repetition of words, inappropriate words, etc.
	Usage	Errors in missing/wrong articles, nonstandard verbs, etc.
	Collocation & preposition	Correct choice and usage of juxtaposition of words
	Organization	Presentation of ideas or discourse elements in order
	Development	Logical connection between elements in an essay
	Word choice	Word frequency measured by standard frequency index (SFI)
	Word length	Average word length in an essay
PS only	Score point value	The score point with the highest cosine correlation of the to-be-scored essay to the training corpus
	Cosine correlation value	The similarity of words used in the to-be-scored essay in comparison with the highest score point in the training corpus (which is usually 6)

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

Appendix B

Human Scoring Rubrics

Human Scoring Rubric for Items 1–5

Score	Response description
3	The response consists of ONE sentence that: • has no grammatical errors; AND • contains forms of both key words used appropriately; AND • is consistent with the picture.
2	The response consists of one or more sentences that: • have one or more grammatical errors that do not obscure the meaning; AND • contain BOTH key words, but they may not be in the same sentence and the form of the word(s) may not be accurate; AND • are consistent with the picture.
1	The response: • has errors that interfere with meaning; OR • omits one or both key words; OR • is not consistent with the picture.
0	The response is blank, written in a foreign language, or consists of keystroke characters.

Human Scoring Rubric for Items 6–7

Score	Response description
4	The response e-mail effectively addresses all the tasks in the prompt, using multiple sentences that clearly convey the information, instructions, questions, etc., required by the prompt. • the response uses organizational logic or appropriate connecting words or both to create coherence among sentences • the tone and register of the response is appropriate for the intended audience • a few isolated errors in grammar or usage may be present but do not obscure the writer's meaning.

Score	Response description
3	The response e-mail is mostly successful but falls short in addressing one of the tasks required by the prompt. • the response may omit, respond unsuccessfully, or respond incompletely to ONE of the required tasks • the response uses organizational logic or appropriate connecting words in at least part of the response • the response shows some awareness of audience • noticeable errors in grammar and usage may be present; ONE sentence may contain errors that obscure meaning.
2	The response e-mail is marked by several weaknesses: • the response may address only ONE of the required tasks or may unsuccessfully or incompletely address TWO OR THREE of the required tasks • connections between ideas may be missing or obscure • the response may show little awareness of audience • errors in grammar and usage may obscure meaning in MORE THAN ONE sentence.
1	The response e-mail is seriously flawed and conveys little or no information, instructions, questions, etc., required by the prompt. • the response addresses NONE of the required tasks, although it may include some content relevant to stimulus • connections between ideas are missing or obscure • the tone or register may be inappropriate for the audience • frequent errors in grammar and usage obscure the writer's meaning most of the time.
0	A response at this level merely copies words from the prompt or stimulus, rejects the topic, is otherwise not connected to the topic, is written in a language other than English, consists of keystroke characters that convey no meaning, or is blank.

Human Scoring Rubric for Item 8

Score	Response description
5	A response 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.
4	A response 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.
3	A response 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.

Score	Response description
2	A response 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.
1	A response 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.
0	A response 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.

Appendix C

Interhuman Agreement for Data Set January 2011–July 2011

Prompt	*N*	Human1		Human2		Human1–Human2 agreement				
		Mean	SD	Mean	SD	Wtd. kappa	Corr.	Std.diff	% agree	% adj agree
Overall	32,835	2.99	0.75	2.99	0.74	0.76	.76	0.00	69.8	99.3
C4152	899	3.14	0.80	3.22	0.84	0.72	.73	-0.09	65.7	99.0
E9333	736	3.16	0.62	3.21	0.67	0.59	.59	-0.08	67.0	99.6
E9007	718	2.93	0.73	2.91	0.69	0.68	.68	0.03	69.6	99.4
C4055	713	2.90	0.88	2.83	0.88	0.76	.76	0.08	65.2	99.0
E5351	706	3.01	0.63	3.05	0.62	0.60	.60	-0.07	69.4	99.7
E9575	704	3.13	0.65	3.12	0.64	0.60	.60	0.01	70.0	99.1
C4052	689	3.00	0.80	3.00	0.80	0.73	.73	0.01	67.2	99.4
E2955	688	2.87	0.73	2.82	0.71	0.70	.70	0.07	72.1	98.7
E5335	681	2.68	0.64	2.71	0.66	0.69	.69	-0.05	73.9	99.9
C4056	673	2.85	0.83	2.81	0.80	0.76	.76	0.05	69.5	99.4
E4369	668	3.17	0.69	3.18	0.67	0.63	.63	-0.02	66.3	99.9
E1577	654	2.80	0.63	2.73	0.66	0.70	.70	0.11	75.1	99.8
E4450	638	3.05	0.60	3.09	0.60	0.65	.65	-0.07	74.3	100
C4154	637	2.88	0.71	2.93	0.69	0.69	.69	-0.07	70.8	99.5
C5863	630	3.27	1.01	3.27	0.98	0.82	.82	0.00	67.6	99.2
E4368	626	2.86	0.65	2.77	0.66	0.67	.67	0.14	72.0	99.7
E4603	625	2.87	0.56	2.86	0.53	0.60	.60	0.03	77.1	99.7
C4057	619	2.75	0.77	2.73	0.77	0.69	.69	0.03	66.7	98.7
C8958	608	2.86	0.61	2.87	0.58	0.63	.63	-0.01	74.3	99.8
E4367	598	2.79	0.74	2.76	0.77	0.70	.70	0.04	66.4	99.7
E5292	592	2.78	0.60	2.80	0.62	0.66	.66	-0.02	75.0	99.8
E9008	591	2.65	0.61	2.66	0.61	0.63	.63	-0.02	74.3	99.5
E4590	576	2.96	0.74	2.97	0.74	0.68	.68	-0.01	66.5	99.3
C5922	558	2.79	0.67	2.80	0.62	0.66	.66	-0.01	71.5	100
E9011	548	3.06	0.64	3.08	0.66	0.66	.66	-0.03	72.8	99.5
E9014	545	3.16	0.69	3.09	0.70	0.63	.64	0.10	68.1	98.9
E2904	544	2.96	0.63	3.02	0.71	0.62	.63	-0.09	68.0	99.1
E9574	542	2.84	0.75	2.84	0.72	0.77	.77	-0.01	75.3	99.8
E7916	540	2.95	0.67	2.96	0.70	0.72	.72	-0.01	75.6	99.3
E4586	538	3.08	0.68	3.06	0.71	0.69	.69	0.03	72.9	99.1
C4086	529	3.17	0.72	3.13	0.68	0.69	.69	0.05	69.8	99.8
E9577	527	2.90	0.83	2.90	0.83	0.74	.74	0.00	65.8	99.4
C5923	518	3.33	0.68	3.25	0.66	0.67	.67	0.12	71.6	99.4
C4084	509	2.42	0.94	2.52	0.98	0.77	.78	-0.10	65.2	97.6
E9580	498	2.92	0.64	2.94	0.62	0.65	.65	-0.03	73.9	99.4
E5342	493	2.95	0.59	2.95	0.62	0.59	.59	0.00	70.2	99.8
E9571	490	2.84	0.64	2.83	0.63	0.69	.69	0.01	75.1	100
E4370	489	3.25	0.68	3.13	0.60	0.59	.60	0.19	68.7	99.0
C5752	465	3.48	0.68	3.50	0.70	0.49	.49	-0.04	57.4	98.3
E5314	461	2.71	0.64	2.70	0.66	0.67	.67	0.02	74.6	99.1
E9572	450	3.08	0.73	3.08	0.67	0.68	.68	-0.01	69.1	99.8
C6232	447	3.61	0.72	3.66	0.71	0.57	.57	-0.07	57.9	99.1
E9330	445	2.97	0.67	2.96	0.73	0.73	.74	0.01	74.2	100
C5926	441	2.94	0.58	2.94	0.57	0.59	.59	0.00	73.5	99.8
C6237	434	3.68	0.71	3.68	0.67	0.39	.39	0.00	56.2	94.9
E7915	432	3.13	0.67	3.12	0.64	0.67	.67	0.00	72.0	100
C6367	418	3.13	0.91	3.06	0.86	0.82	.82	0.08	71.8	99.8
E7913	417	3.10	0.71	3.05	0.69	0.68	.68	0.07	69.3	99.8

Prompt	*N*	Human1		Human2		Human1–Human2 agreement				
		Mean	SD	Mean	SD	Wtd. kappa	Corr.	Std.diff	% agree	% adj agree
C6368	414	3.19	0.75	3.14	0.67	0.68	.69	0.07	68.8	99.8
E4365	401	2.70	0.64	2.71	0.62	0.70	.70	-0.02	76.3	100
E4376	398	2.96	0.81	2.98	0.78	0.62	.62	-0.03	64.1	96.7
E4375	392	2.61	0.66	2.61	0.65	0.70	.70	0.00	75.3	99.7
E9005	382	2.79	0.59	2.66	0.58	0.62	.64	0.23	74.3	99.7
C5859	381	3.08	0.82	3.13	0.83	0.73	.73	-0.06	64.3	99.7
C5862	367	3.33	0.76	3.38	0.73	0.65	.65	-0.08	63.8	98.9
C6373	359	3.78	0.64	3.68	0.63	0.37	.37	0.15	54.9	98.1
C9144	347	2.99	0.59	3.01	0.54	0.61	.61	-0.05	75.8	99.7
C8596	328	2.97	0.55	2.99	0.59	0.60	.60	-0.04	75.0	99.7
C7734	327	3.18	0.98	3.07	0.93	0.70	.70	0.12	59.6	97.2
C8592	310	2.78	0.62	2.67	0.63	0.63	.64	0.18	71.0	100
C6239	301	3.73	0.71	3.73	0.72	0.59	.59	0.00	63.5	98.3
C8987	296	2.72	0.63	2.73	0.62	0.59	.59	-0.03	68.6	100
E9322	285	2.70	0.59	2.74	0.57	0.65	.65	-0.07	76.5	100

Note. Std.diff. = standardized difference on mean scores; Wtd. kappa = quadratic-weighted kappa; Corr. = correlation coefficient; % agree = exact percentage agreement; % adj agree = 1-point adjacent percentage agreement.

Appendix D

Interhuman Agreement for Data Set May 2011–April 2012

Prompt	*N*	Human1 Mean	Human1 SD	Human2 Mean	Human2 SD	Human1–Human2 agreement Wtd. kappa	Corr.	Std.diff	% agree	% adj agree
Overall	17,804	3.00	0.74	2.99	0.74	0.71	.71	0.01	70.5	99.4
C4057	2,044	3.08	0.70	3.03	0.69	0.66	.66	0.07	68.7	99.3
C4056	1,887	3.05	0.80	3.01	0.80	0.74	.74	0.05	68.6	99.6
C6371	1,748	3.08	0.74	3.05	0.73	0.72	.72	0.03	72.3	99.3
C5852	1,642	3.21	0.77	3.20	0.78	0.69	.69	0.02	65.7	99.0
C4055	1,617	2.89	0.86	2.88	0.92	0.78	.78	0.02	67.9	99.1
C5754	1,596	3.23	0.74	3.16	0.70	0.65	.66	0.10	65.8	99.2
C5757	1,400	3.04	0.79	2.98	0.78	0.75	.75	0.07	70.7	99.5
C4059	1,395	3.04	0.81	3.07	0.76	0.74	.74	-0.03	69.8	99.5
C5756	1,380	3.05	0.76	3.05	0.78	0.72	.72	0.00	67.6	99.7
C5759	1,160	3.00	0.77	2.98	0.81	0.72	.72	0.02	67.3	99.3
E5847	968	2.98	0.58	2.96	0.58	0.62	.62	0.04	75.0	99.9
C6232	960	3.27	0.69	3.31	0.65	0.58	.58	-0.05	65.5	99.0
C5752	884	3.39	0.61	3.37	0.60	0.52	.52	0.03	66.3	99.5
E2956	865	3.02	0.58	2.99	0.61	0.58	.58	0.05	72.4	99.3
C6235	806	2.84	0.50	2.80	0.52	0.58	.58	0.09	78.4	99.9
E4455	750	3.01	0.57	3.00	0.57	0.71	.71	0.03	80.9	100
C4052	722	3.49	0.62	3.50	0.66	0.56	.56	-0.01	64.7	99.9
C4087	706	3.51	0.64	3.43	0.60	0.47	.48	0.13	61.6	99.2
E4452	664	2.89	0.61	2.98	0.63	0.62	.63	-0.15	72.7	99.2
C5922	659	3.60	0.66	3.51	0.65	0.46	.46	0.13	57.8	98.5
C6373	658	2.83	0.63	2.81	0.62	0.68	.68	0.03	76.1	99.7
E9569	651	3.04	0.61	3.09	0.62	0.57	.57	-0.09	70.7	99.1
C6236	635	3.54	0.59	3.53	0.59	0.46	.46	0.02	65.0	98.9
C4086	627	3.54	0.62	3.53	0.62	0.44	.44	0.02	60.8	98.7
C4154	603	3.42	0.67	3.40	0.66	0.55	.55	0.03	65.0	98.5
E1549	603	2.88	0.69	2.87	0.65	0.64	.64	0.01	69.2	99.5
E5442	599	2.68	0.60	2.69	0.63	0.69	.69	-0.01	77.0	99.8
E5436	589	2.90	0.73	2.88	0.71	0.75	.75	0.03	74.5	99.8
E5604	567	3.07	0.58	3.07	0.58	0.56	.56	0.00	71.4	99.8
E7908	555	3.04	0.62	3.03	0.65	0.68	.68	0.03	75.1	99.6
E1583	553	2.99	0.58	3.01	0.57	0.57	.57	-0.03	74.0	99.5

Prompt	*N*	Human1		Human2		Human1–Human2 agreement				
		Mean	SD	Mean	SD	Wtd. kappa	Corr.	Std.diff	% agree	% adj agree
E9288	529	2.84	0.69	2.85	0.71	0.71	.71	-0.03	74.1	99.1
E5849	517	2.97	0.59	2.94	0.55	0.67	.68	0.05	78.9	100
E5342	514	2.94	0.61	2.94	0.63	0.60	.60	0.00	69.8	99.8
E5848	509	2.78	0.75	2.78	0.83	0.64	.64	-0.01	65.6	96.3
E2900	491	2.90	0.58	2.88	0.57	0.62	.62	0.03	74.9	100
E4451	484	3.05	0.59	3.04	0.58	0.71	.71	0.02	80.2	100
E4587	482	2.87	0.56	2.89	0.59	0.66	.66	-0.04	78.2	99.8
E4435	475	3.10	0.66	3.07	0.66	0.61	.61	0.05	68.6	99.2
C5857	451	2.78	0.58	2.75	0.62	0.65	.66	0.06	75.2	100
E4366	430	2.56	0.67	2.56	0.69	0.68	.68	0.01	72.3	99.3
E9264	427	3.04	0.62	3.01	0.68	0.65	.65	0.04	71.7	99.5
E5851	424	2.75	0.53	2.75	0.56	0.61	.61	0.02	76.4	100
E9251	422	2.97	0.65	3.01	0.55	0.54	.55	-0.07	68.7	99.3
E4418	403	2.92	0.62	2.90	0.55	0.62	.63	0.04	75.7	99.5
E5851	424	2.75	0.53	2.75	0.56	0.61	.61	0.02	76.4	100
E9251	422	2.97	0.65	3.01	0.55	0.54	.55	-0.07	68.7	99.3
E4418	403	2.92	0.62	2.90	0.55	0.62	.63	0.04	75.7	99.5
E9339	402	2.99	0.65	2.92	0.65	0.67	.68	0.11	73.9	99.5
E2905	394	2.66	0.67	2.64	0.68	0.75	.75	0.02	77.7	99.7
E4583	393	2.77	0.69	2.73	0.72	0.69	.69	0.05	71.2	99.5
E5846	387	2.96	0.68	2.89	0.73	0.64	.65	0.11	66.9	99.0
C5858	372	2.67	0.64	2.68	0.61	0.70	.70	-0.02	76.9	100
C6239	371	3.40	0.62	3.43	0.64	0.56	.56	-0.04	64.7	100
E9009	364	2.34	0.78	2.52	0.75	0.69	.71	-0.24	64.3	99.5
E5321	360	2.99	0.56	2.94	0.56	0.62	.62	0.09	76.7	99.7
C5853	342	2.53	0.63	2.52	0.65	0.72	.72	0.01	76.6	100
C9146	337	2.70	0.58	2.76	0.53	0.60	.60	-0.12	76.0	99.7
C4084	334	2.74	0.83	2.84	0.81	0.75	.75	-0.12	70.4	98.5
E4591	320	2.60	0.64	2.62	0.65	0.70	.70	-0.02	74.7	100
C4153	260	3.56	0.61	3.54	0.62	0.53	.53	0.03	67.3	99.2
C5923	251	3.51	0.60	3.55	0.57	0.52	.52	-0.07	67.3	100
E4422	243	2.41	0.75	2.35	0.71	0.75	.75	0.08	73.7	100
C4152	220	2.87	0.71	2.98	0.71	0.70	.70	-0.15	70.5	99.5
C5751	220	3.65	0.63	3.70	0.57	0.44	.45	-0.07	60.5	100
C4121	218	2.44	0.72	2.55	0.69	0.65	.66	-0.16	66.5	99.5

Prompt	*N*	Human1		Human2		Human1–Human2 agreement				
		Mean	SD	Mean	SD	Wtd. kappa	Corr.	Std.diff	% agree	% adj agree
E4599	218	2.57	0.68	2.46	0.73	0.75	.76	0.16	75.2	100
E2903	217	2.52	0.81	2.53	0.76	0.75	.75	-0.01	68.7	100
C6238	212	2.47	0.74	2.46	0.79	0.73	.73	0.01	69.8	99.5
E2959	208	2.48	0.67	2.58	0.66	0.70	.71	-0.15	75.0	99.5
E9013	193	2.39	0.66	2.35	0.75	0.59	.60	0.06	59.6	100
C8491	191	2.92	0.66	2.90	0.65	0.58	.58	0.04	70.2	97.9
C6374	186	2.98	0.51	3.04	0.54	0.68	.69	-0.11	82.3	100
C9144	182	2.17	0.88	2.15	1.01	0.81	.81	0.02	68.7	98.9
E9579	180	2.73	0.63	2.82	0.59	0.74	.74	-0.14	80.0	100
E4585	179	2.50	0.75	2.49	0.74	0.79	.79	0.01	78.2	99.4
C4085	171	2.51	0.74	2.58	0.76	0.79	.79	-0.09	80.1	98.8
C6370	169	2.46	0.72	2.48	0.72	0.65	.65	-0.02	67.5	98.8
E2892	169	2.62	0.81	2.60	0.87	0.79	.79	0.01	74.0	98.8
E2954	149	2.32	0.80	2.24	0.82	0.81	.81	0.09	75.2	100
E9328	126	2.61	0.62	2.55	0.70	0.84	.85	0.10	85.7	100
C5926	124	2.99	0.50	3.08	0.52	0.46	.47	-0.17	78.2	99.2
E7916	118	3.13	0.71	3.19	0.70	0.67	.68	-0.10	67.8	100
C5758	114	2.77	0.55	2.83	0.56	0.59	.59	-0.11	77.2	99.1
E2957	112	2.34	0.75	2.49	0.71	0.73	.75	-0.21	70.5	100
E5347	109	2.82	0.56	2.88	0.54	0.53	.54	-0.12	71.6	100
C9148	108	3.06	0.58	3.11	0.67	0.47	.48	-0.07	66.7	97.2
E7913	103	3.47	0.64	3.33	0.62	0.44	.45	0.22	61.2	98.1
E7907	101	2.39	0.82	2.45	0.75	0.74	.75	-0.08	68.3	100
C8489	93	3.44	0.76	3.18	0.74	0.60	.63	0.35	55.9	98.9
E9005	88	2.67	0.72	2.31	0.65	0.51	.59	0.53	51.1	98.9
E2958	78	2.74	0.55	2.81	0.54	0.54	.54	-0.12	73.1	100
C8600	65	3.26	0.62	3.17	0.55	0.64	.65	0.16	75.4	100
E9329	65	3.03	0.88	3.09	0.91	0.85	.85	-0.07	75.4	100
E1560	65	2.80	0.51	2.86	0.53	0.71	.71	-0.12	84.6	100
C8608	64	3.22	0.72	3.23	0.73	0.62	.62	-0.02	60.9	100
E9330	64	3.27	0.60	3.25	0.67	0.54	.55	0.02	64.1	100
E5335	62	2.60	0.69	2.61	0.55	0.52	.53	-0.03	62.9	100
E4374	62	2.92	0.61	2.74	0.57	0.62	.65	0.30	72.6	100
E9570	55	2.73	0.71	2.80	0.62	0.71	.72	-0.11	74.5	100
E2960	54	2.24	0.87	2.17	0.84	0.85	.85	0.09	77.8	100

Prompt	*N*	Human1		Human2		Human1–Human2 agreement				
		Mean	SD	Mean	SD	Wtd. kappa	Corr.	Std.diff	% agree	% adj agree
C5860	46	1.89	0.88	1.91	1.09	0.88	.90	-0.02	76.1	100
C6368	44	2.89	0.39	2.93	0.33	0.65	.66	-0.13	90.9	100
C8584	39	3.44	0.85	3.41	0.68	0.71	.73	0.03	66.7	100
C8494	38	3.03	0.49	3.11	0.45	0.58	.59	-0.17	81.6	100
E4444	38	2.87	0.47	2.97	0.49	0.44	.45	-0.22	73.7	100
C8493	37	2.54	0.69	2.51	0.69	0.91	.91	0.04	91.9	100
C9150	34	2.79	0.48	2.91	0.38	0.37	.40	-0.27	76.5	100
C8587	31	3.19	0.60	3.13	0.67	0.59	.60	0.10	67.7	100

Note. Only prompts with more than 30 examinees are included in this table. SD = standard deviation; Std. diff. = standardized difference on mean scores; Wtd. kappa = quadratic-weighted kappa; Corr. = correlation coefficient; % agree = exact percentage agreement; % adj agree = 1-point adjacent percentage agreement.

Appendix E

Advisory Flags in e-rater Automated Essay Scoring System

Flag	Cause	Descriptions
#2	Reuse of language	Compared to other essays written on this topic, the essay contains more reuse of language, a possible indication that it contains sentences or paragraphs that are repeated.
#4	Key concept	Compared to other essays written on this topic, the essay shows less development of the key concepts on this topic.
#8	Not relevant	The essay might not be relevant to the assigned topic.
#16	Restatement	The essay appears to be a restatement of the topic with few additional concepts.
#32	Not resemblance	The essay does not resemble others that have been written on this topic, a possible indication that it is about something else or is not relevant to the issues the topic raises.
#64	Too brief	The essay is too brief to evaluate.
#128	Excessive length	The essay is longer than essays that can be accurately scored and must be within the word limit to receive a score.
#256	Unidentifiable organizational elements	The essay could not be scored because some of its organizational elements could not be identified.
#512	Excessive number of problems	The essay could not be scored because too many problems in grammar, usage, mechanics, and style were identified.
#1024	Unexpected topic	The essay appears to be on a subject that is different from the assigned topic.
#2048	Nonessay	The text submitted does not appear to be an essay.

Appendix F

Computation Procedures for PRESS and PRESS Statistics Based Evaluation Indices

Index	Formula	Description
PRESS	PRESS = $\sum_{i=1}^{n}(Y_i - \hat{Y}_{i(i)})^2$	Y_i refers to the observed score on the *i*th essay (i.e., the average of two human scores); $\hat{Y}_{i(i)}$ refers to the predicted score for the *i*th essay that resulted from the model built without the *i*th essay included.
Inflation index	II = (a - c)/(c - b)	a = PRESS/*N* b = $(H1 - H2)^2/4$ c = [a + (MSE)(*p*)(*N*)]/2 d = $(MSE)(p)/(1 - R^2)(N - 1)$ PRESS refers to the value from the ordinary PRESS statistics; MSE refers to the mean squared error of the e-rater scoring model; *p* refers to the number of parameters including intercept; *N* refers to the sampling size used for model building; H1 and H2 refer to the double human ratings. R^2 refers to the variance percentage explained by the e-rater scoring model.
Value index	VI = 2b {1- [(MSE) - b] /d}/(c - b)	
Prediction index	PI = 1 - (c - b)/{[d+d/2*N*(*N* - 1)] - b}	
Human quality index	HQI = 1 - [b (1 – b / d)]/(d - b)	

Appendix G

Generic Model Evaluation by Prompt

		Human1		Human2		Human1/Human2 agreement statistics						E-rater				Human1/e-rater agreement statistics				Degradation (H1/e-rater minus H1/H2)	
Prompt	*N*	Mean	SD	Mean	SD	Std. diff.	Kappa	Wtd. kappa	% agree	% adj agree	Corr.	Mean	SD	Std. diff.	Kappa	Wtd. kappa	% agree	% adj agree	Corr.	Wtd. kappa	Corr.
C4052	489	2.99	0.79	2.99	0.79	-0.01	0.46	0.71	66.05	99.39	.71	2.93	0.78	-0.09	0.50	0.74	67.69	99.18	.80	0.03	.09
C4055	513	2.89	0.90	2.81	0.89	-0.09	0.49	0.77	65.69	99.03	.77	2.89	0.88	0.00	0.48	0.74	64.72	98.05	.78	-0.03	.01
C4056	473	2.82	0.82	2.79	0.78	-0.04	0.52	0.75	69.98	99.37	.75	2.76	0.82	-0.08	0.46	0.75	65.96	99.79	.80	0.00	.05
C4057	419	2.72	0.76	2.72	0.78	0.01	0.40	0.67[a]	63.72	99.05	.67[a]	2.63	0.81	-0.11	0.36	0.66[a]	59.90	99.28	.74	-0.01	.07
C4084	309	2.47	0.93	2.55	0.98	0.08	0.52	0.78	66.02	98.06	.79	2.72	0.89	0.28[a]	0.35	0.62[a]	54.05	93.85	.67[a]	-0.16[a]	-.12[a]
C4086	329	3.21	0.70	3.16	0.65	-0.07	0.46	0.67[a]	70.52	99.70	.67[a]	3.10	0.69	-0.16[a]	0.42	0.64[a]	65.35	98.48	.68[a]	-0.03	.01
C4152	699	3.15	0.81	3.21	0.83	0.06	0.47	0.73	66.24	99.00	.73	3.14	0.75	-0.02	0.46	0.70	65.81	98.86	.76	-0.03	.03
C4154	437	2.91	0.72	2.96	0.67	0.08	0.52	0.69[a]	72.31	99.31	.69[a]	2.93	0.75	0.03	0.50	0.72	68.88	99.54	.77	0.03	.08
C5922	358	2.78	0.65	2.78	0.61	-0.01	0.42	0.62[a]	70.11	100.00	.62[a]	2.68	0.74	-0.15	0.40	0.64[a]	63.97	99.72	.72	0.02	.10
C5923	318	3.36	0.67	3.25	0.66	-0.15[a]	0.53	0.68[a]	73.58	99.37	.69[a]	3.28	0.62	-0.12	0.49	0.66[a]	70.75	99.69	.71	-0.02	.02
C5926	241	2.90	0.57	2.92	0.53	0.02	0.45	0.59[a]	75.93	99.59	.59[a]	2.79	0.57	-0.21[a]	0.41	0.62[a]	70.95	100.00	.68[a]	0.03	.09
C5752	265	3.49	0.67	3.55	0.69	0.08	0.34	0.52[a]	60.75	98.11	.52[a]	3.47	0.54	-0.04	0.32	0.51[a]	60.38	99.25	.58[a]	-0.01	.06
C5863	430	3.28	1.01	3.29	0.98	0.01	0.52	0.82	66.28	99.53	.82	3.26	0.88	-0.03	0.47	0.79	63.72	99.30	.83	-0.03	.01
C6232	247	3.57	0.70	3.66	0.73	0.12	0.34	0.59[a]	59.51	99.60	.60[a]	3.62	0.48	0.08	0.30	0.50[a]	60.32	99.19	.56[a]	-0.09	-.04
C6237	234	3.65	0.73	3.68	0.65	0.04	0.25	0.40[a]	55.56	95.73	.41[a]	3.68	0.41	0.04	0.30	0.46[a]	61.11	98.29	.57[a]	0.06	.16
C6367	218	3.12	0.92	3.06	0.89	-0.06	0.57	0.82	70.64	100.00	.82	3.06	0.97	-0.06	0.51	0.80	65.60	99.54	.85	-0.02	.03
C6368	214	3.17	0.76	3.11	0.68	-0.08	0.45	0.68[a]	66.82	100.00	.69[a]	3.24	0.68	0.10	0.50	0.72	68.69	100.00	.76	0.04	.07
C8958	408	2.87	0.59	2.85	0.56	-0.02	0.52	0.64[a]	76.47	99.75	.64[a]	2.85	0.65	-0.02	0.41	0.62[a]	67.89	99.75	.68[a]	-0.02	.04
E4603	425	2.87	0.57	2.85	0.53	-0.03	0.45	0.60[a]	77.18	99.53	.60[a]	2.87	0.66	0.00	0.43	0.64[a]	71.53	99.76	.71	0.04	.11
E9330	245	3.02	0.69	3.01	0.75	-0.01	0.57	0.76	75.10	100.00	.76	3.02	0.69	0.00	0.51	0.73	71.43	100.00	.79	-0.03	.03
E9333	536	3.15	0.65	3.23	0.68	0.11	0.37	0.58[a]	64.74	99.44	.59[a]	3.03	0.57	-0.21[a]	0.44	0.62[a]	69.22	99.63	.70	0.04	.11

		Human1		Human2		Human1/Human2 agreement statistics						E-rater				Human1/e-rater agreement statistics				Degradation (H1/e-rater minus H1/H2)	
Prompt	*N*	Mean	SD	Mean	SD	Std. diff.	Kappa	Wtd. kappa	% agree	% adj agree	Corr.	Mean	SD	Std. diff.	Kappa	Wtd. kappa	% agree	% adj agree	Corr.	Wtd. kappa	Corr.
E1577	454	2.82	0.63	2.75	0.64	-0.11	0.50	0.67[a]	74.23	99.78	.68[a]	2.74	0.73	-0.11	0.50	0.71	72.47	99.78	.77	0.04	.09
E5292	392	2.74	0.61	2.78	0.63	0.07	0.50	0.67[a]	75.77	99.74	.67[a]	2.75	0.68	0.02	0.41	0.64[a]	68.11	99.74	.71	-0.03	.04
E5314	261	2.66	0.66	2.64	0.67	-0.02	0.46	0.65[a]	71.65	99.23	.65[a]	2.71	0.67	0.08	0.44	0.65[a]	70.88	98.85	.69a	0.00	.04
E5335	481	2.69	0.62	2.73	0.66	0.06	0.48	0.67[a]	73.60	99.79	.67[a]	2.68	0.65	-0.02	0.44	0.62[a]	69.65	99.17	.69[a]	-0.05	.02
E5342	293	2.94	0.60	2.94	0.61	0.01	0.42	0.60[a]	72.01	99.66	.60[a]	3.05	0.56	0.20[a]	0.42	0.59[a]	68.94	99.66	.65[a]	-0.01	.05
E5351	506	2.99	0.64	3.02	0.65	0.03	0.42	0.64[a]	70.16	100.00	.64[a]	3.04	0.61	0.07	0.45	0.66[a]	70.36	100.00	.71	0.02	.07
E9007	518	2.92	0.73	2.91	0.69	0.00	0.44	0.68[a]	69.50	99.42	.68[a]	2.93	0.75	0.02	0.49	0.72	69.31	99.61	.75	0.04	.07
E9008	391	2.65	0.60	2.67	0.61	0.03	0.52	0.64[a]	75.45	99.49	.64[a]	2.56	0.70	-0.15[a]	0.41	0.63[a]	66.50	100.00	.70[a]	-0.01	.06
E9011	348	3.06	0.66	3.09	0.69	0.04	0.48	0.68[a]	72.13	99.43	.68[a]	3.14	0.64	0.11	0.47	0.67[a]	69.25	99.71	.74	-0.01	.06
E9014	345	3.21	0.67	3.13	0.71	-0.11	0.48	0.66[a]	71.01	98.84	.66[a]	3.18	0.56	-0.05	0.51	0.67[a]	73.33	99.71	.70	0.01	.04
E4365	201	2.69	0.66	2.70	0.62	0.01	0.59	0.74	78.61	100.00	.74	2.46	0.78	-0.33[a]	0.34	0.63[a]	60.70	100.00	.76	-0.11[a]	.02
E4367	398	2.80	0.75	2.78	0.77	-0.02	0.46	0.71	67.59	99.50	.71	2.62	0.82	-0.23[a]	0.49	0.74	66.58	99.25	.79	0.03	.08
E4368	426	2.85	0.67	2.76	0.68	-0.14	0.45	0.66[a]	70.66	99.53	.67[a]	2.73	0.77	-0.16[a]	0.45	0.69[a]	68.08	99.53	.75	0.03	.08
E4369	468	3.16	0.71	3.19	0.67	0.05	0.35	0.62[a]	64.53	99.79	.62[a]	3.29	0.59	0.20[a]	0.34	0.60[a]	61.11	99.79	.70[a]	-0.02	.08
E4370	289	3.19	0.67	3.10	0.55	-0.15	0.37	0.57[a]	68.51	99.65	.59[a]	3.10	0.54	-0.14	0.33	0.56[a]	64.71	99.65	.66[a]	-0.01	.07
E4450	438	3.05	0.60	3.07	0.56	0.03	0.45	0.60[a]	73.52	100.00	.61[a]	3.01	0.63	-0.08	0.53	0.69[a]	73.74	99.77	.72	0.09	.11
E7913	217	3.12	0.75	3.06	0.74	-0.07	0.52	0.72	70.97	99.54	.73	3.15	0.67	0.04	0.45	0.68[a]	67.28	99.08	.72	-0.04	-.01
E7915	232	3.12	0.69	3.08	0.63	-0.05	0.49	0.67[a]	71.55	100.00	.68[a]	3.14	0.64	0.04	0.39	0.64[a]	64.22	100.00	.72	-0.03	.04
E7916	340	2.94	0.65	2.97	0.69	0.05	0.56	0.71	77.35	98.82	.71	3.03	0.64	0.14	0.48	0.68[a]	71.47	100.00	.76	-0.03	.05
E9571	290	2.87	0.63	2.85	0.60	-0.03	0.52	0.69[a]	76.21	100.00	.69[a]	2.88	0.66	0.02	0.41	0.61[a]	68.28	98.97	.67[a]	-0.08	-.02
E9572	250	3.10	0.73	3.05	0.67	-0.07	0.48	0.70[a]	70.00	100.00	.70[a]	3.16	0.67	0.08	0.49	0.71	68.00	100.00	.77	0.01	.07
E9574	342	2.82	0.71	2.83	0.69	0.02	0.53	0.73	74.85	99.71	.73	2.87	0.73	0.08	0.57	0.77	74.56	100.00	.81	0.04	.08
E9575	504	3.13	0.67	3.11	0.66	-0.04	0.42	0.62[a]	69.25	99.01	.62[a]	3.15	0.57	0.02	0.45	0.66	70.24	100.00	.70	0.04	.08
E9577	327	2.90	0.82	2.90	0.85	0.00	0.48	0.75	66.36	99.39	.75	2.95	0.75	0.06	0.44	0.72	63.00	99.39	.77	-0.03	.02
E9580	298	2.93	0.62	2.95	0.58	0.03	0.46	0.61[a]	73.83	99.33	.61[a]	2.98	0.67	0.09	0.33	0.59	63.76	99.66	.66[a]	-0.02	.05

		Human1		Human2		Human1/Human2 agreement statistics						E-rater				Human1/e-rater agreement statistics				Degradation (H1/e-rater minus H1/H2)	
Prompt	*N*	Mean	SD	Mean	SD	Std. diff.	Kappa	Wtd. kappa	% agree	% adj agree	Corr.	Mean	SD	Std. diff.	Kappa	Wtd. kappa	% agree	% adj agree	Corr.	Wtd. kappa	Corr.
E4586	338	3.07	0.69	3.04	0.65	-0.05	0.56	0.70	76.33	99.11	.70	3.24	0.61	0.26[a]	0.45	0.66[a]	67.46	99.70	.74	-0.04	.04
E4590	376	2.94	0.72	2.96	0.74	0.02	0.42	0.67[a]	67.02	99.47	.67[a]	3.15	0.58	0.31[a]	0.39	0.61[a]	64.63	98.94	.72	-0.06	.05
E2904	344	2.96	0.62	3.05	0.69	0.13	0.40	0.63[a]	69.48	99.42	.64[a]	3.15	0.61	0.31[a]	0.37	0.57[a]	63.66	99.13	.69[a]	-0.06	.05
E2955	488	2.89	0.74	2.82	0.72	-0.09	0.48	0.68[a]	71.11	98.36	.68[a]	3.04	0.68	0.22[a]	0.37	0.63[a]	62.50	97.95	.68[a]	-0.05	0
C7734	327	3.18	0.98	3.07	0.93	-0.12	0.43	0.70[a]	59.63	97.25	.70	3.11	0.76	-0.09	0.35	0.67[a]	55.66	97.86	.75	-0.03	.05
C5859	381	3.08	0.82	3.13	0.83	0.06	0.45	0.73	64.30	99.74	.73	3.02	0.83	-0.07	0.47	0.75	65.09	99.21	.80	0.02	.07
C5862	367	3.33	0.76	3.38	0.73	0.08	0.41	0.65[a]	63.76	98.91	.65[a]	3.36	0.69	0.05	0.49	0.70	68.39	99.46	.74	0.05	.09
C6239	301	3.73	0.71	3.73	0.72	0.00	0.40	0.59[a]	63.46	98.34	.59[a]	3.75	0.47	0.03	0.32	0.51[a]	62.79	99.67	.64[a]	-0.08	.05
C6373	359	3.78	0.64	3.68	0.63	-0.15[a]	0.21	0.37[a]	54.87	98.05	.37[a]	3.74	0.43	-0.07	0.24	0.37[a]	60.72	99.44	.45[a]	0.00	.08
C9144	347	2.99	0.59	3.01	0.54	0.05	0.42	0.61[a]	75.79	99.71	.61[a]	2.87	0.60	-0.19	0.31	0.56[a]	65.99	99.42	.63[a]	-0.05	.02
C8592	310	2.78	0.62	2.67	0.63	-0.18[a]	0.46	0.63[a]	70.97	100.00	.64[a]	2.72	0.66	-0.09	0.41	0.58[a]	65.81	99.35	.67[a]	-0.05	.03
C8596	328	2.97	0.55	2.99	0.59	0.04	0.43	0.60[a]	75.00	99.70	.60[a]	2.94	0.64	-0.05	0.38	0.61[a]	69.82	99.70	.68[a]	0.01	.08
C8987	296	2.72	0.63	2.73	0.62	0.03	0.41	0.59[a]	68.58	100.00	.59[a]	2.56	0.77	-0.22	0.33	0.60[a]	60.14	98.99	.68	0.01	.09
E9322	285	2.70	0.59	2.74	0.57	0.07	0.50	0.65[a]	76.49	100.00	.65[a]	2.58	0.72	-0.19	0.39	0.63[a]	65.61	100.00	.71	-0.02	.06
E9005	382	2.79	0.59	2.66	0.58	-0.23[a]	0.47	0.62[a]	74.35	99.74	.64[a]	2.65	0.71	-0.21	0.38	0.63[a]	64.92	100.00	.74	0.01	.10
E4375	392	2.61	0.66	2.61	0.65	0.00	0.56	0.70	75.26	99.74	.70	2.66	0.75	0.06	0.43	0.64[a]	64.54	98.72	.73	-0.06	.03
E4376	398	2.96	0.81	2.98	0.78	0.03	0.44	0.62[a]	64.07	96.73	.62[a]	3.01	0.63	0.06	0.35	0.52[a]	60.05	95.48	.61[a]	-0.10	-.01
Average	362.5	3.00	0.70	3.00	0.69	-0.01	0.46	0.66[a]	69.89	99.32	.66[a]	3.00	0.68	-0.01	0.42	0.65[a]	66.19	99.29	.71	-0.01	.05

*Note***.** Std. diff. = standardized difference on mean scores; Wtd. kappa = quadratic-weighted kappa; Corr. = correlation coefficient; % agree = exact percentage agreement; % adj agree = 1-point adjacent percentage agreement.

[a]Shaded values did not meet the evaluation thresholds.

Appendix H

Prompt-Specific (Traditional) Model Evaluation by Prompt

		Human1		Human2		Human1/ Human2 agreement statistics						e-rater		Human1/e-rater agreement statistics						Degradation (H1/e-rater minus h1/h2)	
Prompt	N	Mean	SD	Mean	SD	Std. diff.	Kappa	Wtd. kappa	% agree	% adj agree	Corr.	Mean	SD	Std. diff.	Kappa	Wtd. kappa	% agree	% adj agree	Corr.	Wtd. kappa	Corr.
C4052	340	3.00	0.81	2.97	0.79	-0.03	0.47	0.72	66.18	99.41	.72	3.02	0.81	0.02	0.54	0.78	69.12	99.71	.81	0.06	.09
C4055	359	2.90	0.88	2.83	0.89	-0.08	0.49	0.77	65.74	99.44	.77	2.86	0.86	-0.05	0.48	0.76	64.35	99.16	.80	-0.01	.03
C4056	339	2.83	0.84	2.81	0.81	-0.03	0.51	0.75	68.73	99.12	.75	2.83	0.87	-0.01	0.52	0.78	68.44	99.71	.82	0.03	.07
C4057	308	2.75	0.77	2.74	0.75	-0.02	0.48	0.70	68.51	99.03	.70	2.74	0.75	-0.01	0.46	0.71	65.91	99.35	.76	0.01	.06
C4084	255	2.43	0.94	2.57	0.98	0.15[a]	0.46	0.76	61.96	97.65	.77	2.46	0.92	0.04	0.38	0.69[a]	56.08	96.47	.73	-0.07	-.04
C4086	268	3.16	0.73	3.08	0.67	-0.12	0.42	0.67[a]	67.54	100.00	.68[a]	3.18	0.68	0.02	0.48	0.64[a]	69.03	98.13	.68[a]	-0.03	.00
C4152	449	3.14	0.82	3.24	0.87	0.12	0.48	0.75	66.15	99.33	.76	3.13	0.87	-0.02	0.48	0.76	65.26	100.00	.79	0.01	.03
C4154	316	2.89	0.71	2.93	0.66	0.05	0.50	0.68[a]	71.84	99.37	.68[a]	2.91	0.70	0.02	0.53	0.74	71.52	100.00	.79	0.06	.11
C5922	278	2.80	0.66	2.78	0.62	-0.04	0.47	0.67[a]	72.30	100.00	.67[a]	2.80	0.67	-0.01	0.44	0.66[a]	68.35	100.00	.74	-0.01	.07
C5923	258	3.31	0.69	3.25	0.68	-0.08	0.53	0.70	73.26	99.61	.70	3.20	0.64	-0.16[a]	0.54	0.72	73.26	100.00	.75	0.02	.05
C5926	224	2.93	0.60	2.93	0.64	0.00	0.45	0.62[a]	71.88	99.55	.62[a]	2.94	0.56	0.02	0.45	0.61[a]	70.98	99.55	.68[a]	-0.01	.06
C5752	240	3.47	0.69	3.52	0.68	0.07	0.23	0.45[a]	54.58	97.92	.45[a]	3.47	0.75	0.00	0.28	0.56[a]	56.25	98.33	.58[a]	0.11	.13
C5863	316	3.26	1.03	3.28	1.00	0.01	0.53	0.83	67.09	99.37	.83	3.26	1.02	0.00	0.51	0.83	66.46	99.68	.86	0.00	.03
C6232	226	3.62	0.73	3.65	0.69	0.06	0.34	0.56[a]	60.18	98.67	.57[a]	3.60	0.74	-0.02	0.23	0.52[a]	51.77	97.79	.55[a]	-0.04	-.02
C6237	216	3.68	0.71	3.68	0.68	-0.01	0.23	0.38[a]	53.70	95.83	.38[a]	3.67	0.70	-0.01	0.24	0.48[a]	54.17	98.15	.51[a]	0.10	.13
C6367	208	3.13	0.91	3.02	0.84	-0.12	0.55	0.79	69.23	99.52	.80	3.06	0.88	-0.07	0.53	0.80	67.79	100.00	.85	0.01	.05
C6368	209	3.19	0.75	3.16	0.64	-0.05	0.50	0.70	70.81	100.00	.71	3.15	0.74	-0.06	0.44	0.65[a]	65.07	98.56	.72	-0.05	.01
C8958	306	2.86	0.62	2.86	0.60	0.00	0.48	0.63[a]	73.53	99.67	.63[a]	2.85	0.60	-0.01	0.46	0.65[a]	71.90	100.00	.71	0.02	.08
E4603	320	2.87	0.57	2.86	0.52	-0.03	0.45	0.59[a]	77.19	99.38	.59[a]	2.87	0.59	0.00	0.44	0.64[a]	74.69	100.00	.72	0.05	.13
E9330	224	2.96	0.69	2.98	0.76	0.03	0.53	0.74	72.77	100.00	.74	3.00	0.67	0.06	0.47	0.70[a]	68.75	100.00	.77	-0.04	.03
E9333	367	3.17	0.63	3.19	0.65	0.04	0.39	0.58[a]	67.03	99.46	.58[a]	3.24	0.60	0.12	0.47	0.64[a]	69.75	99.73	.70[a]	0.06	.12
E1577	331	2.80	0.65	2.71	0.66	-0.14	0.50	0.70[a]	74.02	100.00	.70	2.75	0.64	-0.08	0.52	0.71	75.23	100.00	.76	0.01	.06
E5292	301	2.79	0.61	2.80	0.63	0.02	0.49	0.66[a]	75.08	99.67	.66[a]	2.82	0.56	0.06	0.45	0.64[a]	72.43	100.00	.69[a]	-0.02	.03
E5314	234	2.72	0.65	2.72	0.67	-0.01	0.48	0.65[a]	73.50	98.72	.65[a]	2.70	0.63	-0.04	0.47	0.66[a]	73.08	99.57	.70	0.01	.05
E5335	341	2.70	0.63	2.72	0.65	0.03	0.46	0.65[a]	72.14	99.71	.65[a]	2.72	0.59	0.03	0.42	0.61[a]	70.38	99.71	.71	-0.04	.06
E5342	244	2.96	0.60	2.95	0.61	-0.03	0.38	0.59[a]	70.49	100.00	.59[a]	2.98	0.56	0.03	0.33	0.53[a]	64.75	99.59	.59[a]	-0.06	.00

		Human1		Human2		Human1/ Human2 agreement statistics						e-rater		Human1/e-rater agreement statistics						Degradation (H1/e-rater minus h1/h2)	
Prompt	N	Mean	SD	Mean	SD	Std. diff.	Kappa	Wtd. kappa	% agree	% adj agree	Corr.	Mean	SD	Std. diff.	Kappa	Wtd. kappa	% agree	% adj agree	Corr.	Wtd. kappa	Corr.
E5351	359	3.00	0.65	3.03	0.64	0.04	0.40	0.61[a]	69.36	99.44	.61[a]	2.98	0.61	-0.04	0.38	0.62[a]	67.97	99.72	.69[a]	0.01	.08
E9007	357	2.94	0.73	2.92	0.68	-0.02	0.46	0.69[a]	70.59	99.44	.69[a]	2.94	0.72	0.00	0.54	0.74	72.27	99.72	.78	0.05	.09
E9008	297	2.66	0.60	2.67	0.61	0.02	0.50	0.64[a]	74.41	99.66	.64[a]	2.68	0.63	0.02	0.47	0.65[a]	71.72	100.00	.71	0.01	.07
E9011	277	3.08	0.64	3.10	0.65	0.04	0.49	0.64[a]	72.92	99.28	.64[a]	3.13	0.61	0.08	0.43	0.65[a]	68.59	100.00	.73	0.01	.09
E9014	274	3.17	0.68	3.14	0.68	-0.05	0.42	0.62[a]	68.61	98.91	.62[a]	3.15	0.66	-0.02	0.41	0.63[a]	66.06	99.27	.67[a]	0.01	.05
E4365	202	2.70	0.64	2.72	0.63	0.03	0.56	0.72	77.23	100.00	.72	2.71	0.63	0.02	0.52	0.69[a]	73.76	99.50	.75	-0.03	.03
E4367	298	2.80	0.75	2.79	0.78	-0.01	0.41	0.68[a]	63.76	99.66	.68[a]	2.76	0.78	-0.05	0.50	0.74	68.46	100.00	.80	0.06	.12
E4368	316	2.86	0.65	2.79	0.69	-0.11	0.49	0.70	73.10	100.00	.71	2.78	0.65	-0.13	0.54	0.71	76.27	99.37	.75	0.01	.04
E4369	334	3.16	0.69	3.20	0.66	0.07	0.35	0.62[a]	64.97	100.00	.62[a]	3.19	0.64	0.04	0.36	0.61[a]	63.47	99.70	.68[a]	-0.01	.06
E4370	254	3.25	0.69	3.12	0.62	-0.20[a]	0.41	0.60[a]	68.50	98.82	.61[a]	3.20	0.66	-0.07	0.31	0.59[a]	59.06	100.00	.68[a]	-0.01	.07
E4450	318	3.06	0.61	3.11	0.60	0.07	0.46	0.63[a]	72.96	100.00	.63[a]	3.08	0.57	0.03	0.51	0.67[a]	73.27	100.00	.72	0.04	.09
E7913	208	3.11	0.70	3.09	0.70	-0.03	0.47	0.68[a]	69.23	100.00	.68[a]	3.08	0.71	-0.04	0.47	0.70	67.79	100.00	.73	0.02	.05
E7915	216	3.13	0.68	3.13	0.62	0.00	0.47	0.66[a]	71.30	100.00	.66[a]	3.15	0.61	0.03	0.51	0.71	72.22	100.00	.76	0.05	.10
E7916	272	2.95	0.68	2.96	0.70	0.02	0.54	0.73	74.63	100.00	.73	2.93	0.74	-0.03	0.53	0.74	72.43	100.00	.78	0.01	.05
E9571	248	2.85	0.65	2.85	0.67	0.00	0.50	0.70	74.19	100.00	.70	2.87	0.66	0.03	0.41	0.65[a]	66.94	100.00	.72	-0.05	.02
E9572	226	3.07	0.74	3.05	0.67	-0.03	0.38	0.64[a]	64.16	100.00	.64[a]	3.12	0.68	0.07	0.44	0.69[a]	64.60	100.00	.73	0.05	.09
E9574	274	2.84	0.76	2.82	0.72	-0.02	0.54	0.76	74.09	100.00	.76	2.85	0.76	0.01	0.55	0.78	72.26	100.00	.80	0.02	.04
E9575	349	3.12	0.65	3.11	0.69	-0.01	0.47	0.66[a]	71.92	99.14	.66[a]	3.13	0.63	0.02	0.46	0.67[a]	70.77	100.00	.71	0.01	.05
E9577	265	2.89	0.83	2.83	0.80	-0.06	0.46	0.73	65.66	99.62	.74	2.91	0.84	0.03	0.49	0.76	66.04	99.62	.78	0.03	.04
E9580	251	2.92	0.66	2.96	0.63	0.06	0.52	0.69[a]	76.49	99.20	.69[a]	2.90	0.59	-0.03	0.40	0.62[a]	70.12	99.60	.70[a]	-0.07	.01
E4586	271	3.07	0.70	3.05	0.72	-0.03	0.47	0.66[a]	70.48	98.52	.66[a]	3.07	0.73	0.00	0.44	0.67[a]	66.05	99.26	.75	0.01	.09
E4590	288	2.98	0.73	2.99	0.74	0.01	0.35	0.63[a]	61.81	99.31	.63[a]	2.99	0.70	0.01	0.32	0.61[a]	59.72	99.31	.70[a]	-0.02	.07
E2904	273	2.93	0.66	3.00	0.75	0.09	0.37	0.61[a]	66.30	98.53	.62[a]	2.97	0.70	0.06	0.45	0.66[a]	69.60	98.90	.69[a]	0.05	.07
E2955	342	2.88	0.72	2.83	0.73	-0.06	0.51	0.71	72.51	98.83	.71	2.83	0.73	-0.06	0.36	0.65[a]	62.57	99.42	.72	-0.06	.01
Average	283.8	2.99	0.71	2.99	0.70	-0.01	0.46	0.66[a]	69.56	99.38	.67[a]	2.99	0.69	0.00	0.45	0.67[a]	67.71	99.53	.72	0.01	.05

Note. Std. diff. = standardized difference on mean scores; Wtd. kappa = quadratic-weighted kappa; Corr. = correlation coefficient; % agree = exact percentage agreement; % adj agree = 1-point adjacent percentage agreement.
[a]Shaded values did not meet the evaluation thresholds.

Appendix I

Prompt-Specific (Press) Model Evaluation

Prompt	*N*	Human quality index	Value index	Inflation index	Prediction index
C7734	327	0.82	1.72	0.08[a]	0.80
C4052	689	0.85	2.72	0.04	0.88
C4055	713	0.86	1.75	0.04	0.84
C4056	673	0.86	2.31	0.04	0.88
C4057	619	0.81	2.84	0.05[a]	0.86
C4084	509	0.87	0.69[a]	0.03	0.70
C4086	529	0.81	1.12	0.04	0.71
C4152	899	0.84	1.58	0.03	0.81
C4154	637	0.82	2.48	0.05[a]	0.85
C5922	558	0.79[a]	2.06	0.05[a]	0.80
C5923	518	0.80	1.79	0.05[a]	0.78
C5926	441	0.74[a]	2.09	0.07[a]	0.75
C5752	465	0.66[a]	2.00	0.06[a]	0.66[a]
C5859	381	0.85	1.84	0.08[a]	0.84
C5862	367	0.78[a]	2.50	0.08[a]	0.82
C5863	630	0.90	1.44	0.04	0.87
C6232	447	0.72[a]	1.10	0.05[a]	0.59[a]
C6237	434	0.56[a]	4.53	0.12[a]	0.74
C6239	301	0.75 [a]	1.65	0.09[a]	0.71
C6367	418	0.90	1.38	0.05[a]	0.86
C6368	414	0.81	1.57	0.06[a]	0.77
C6373	359	0.54[a]	1.86	0.09[a]	0.52[a]
C9144	347	0.76[a]	1.54	0.08[a]	0.71
C8592	310	0.77[a]	1.64	0.08[a]	0.74
C8596	328	0.75[a]	2.11	0.09[a]	0.76
C8958	608	0.77[a]	2.18	0.05[a]	0.79
C8987	296	0.74[a]	2.46	0.12[a]	0.78
E4603	625	0.75[a]	3.50	0.07[a]	0.84
E9322	285	0.79[a]	2.28	0.11[a]	0.81
E9330	445	0.85	1.58	0.06[a]	0.82
E9333	736	0.74[a]	2.91	0.05[a]	0.81
E1577	654	0.82	3.01	0.05[a]	0.87
E5292	592	0.79[a]	2.05	0.05[a]	0.80
E5314	461	0.80	1.09	0.05[a]	0.69[a]
E5335	681	0.81	1.85	0.04	0.80
E5342	493	0.74[a]	1.69	0.06[a]	0.71
E5351	706	0.75[a]	2.79	0.05[a]	0.81
E9005	382	0.76[a]	7.68	0.18[a]	0.93
E9007	718	0.81	3.40	0.05[a]	0.88
E9008	591	0.78[a]	1.97	0.05[a]	0.77
E9011	548	0.80	2.22	0.05[a]	0.81
E9014	545	0.78[a]	2.30	0.05[a]	0.80
E4365	401	0.82	1.88	0.07[a]	0.82
E4367	598	0.82	4.12	0.07[a]	0.90
E4368	626	0.80[a]	2.81	0.06[a]	0.85
E4369	668	0.77[a]	2.61	0.05[a]	0.82
E4370	489	0.74 [a]	1.95	0.06[a]	0.73

Prompt	*N*	Human quality index	Value index	Inflation index	Prediction index
E4375	392	0.83	1.25	0.06[a]	0.75
E4376	398	0.76[a]	1.46	0.06[a]	0.70
E4450	638	0.78[a]	2.59	0.05[a]	0.83
E7913	417	0.81	1.58	0.06[a]	0.77
E7915	432	0.80	2.25	0.07[a]	0.82
E7916	540	0.84	1.45	0.05[a]	0.79
E9571	490	0.82	1.05	0.05[a]	0.70
E9572	450	0.81	1.90	0.06[a]	0.80
E9574	542	0.87	1.71	0.05[a]	0.85
E9575	704	0.75[a]	2.94	0.05[a]	0.82
E9577	527	0.85	1.41	0.04	0.80
E9580	498	0.79[a]	1.52	0.05[a]	0.74
E4586	538	0.82	1.73	0.05[a]	0.79
E4590	576	0.81	2.05	0.05[a]	0.81
E2904	544	0.76[a]	2.22	0.05[a]	0.78
E2955	688	0.82	1.51	0.04	0.78

[a]Shaded values did not meet the evaluation thresholds.

Appendix J

Variance Component of ANOVA Analysis on Prompt and Test Country/Territory Effects

Data Set 1:

Source	*df*	*F*	η^2	*p*
Prompt	2	2.85	0.02	0.0580
Test country/territory	13	21.79	0.43	<.0001
Prompt x Test Country/territory	26	2.08	0.13	0.0011
Within-group error	1,774	(0.47)		

Data Set 2:

Source	*df*	*F*	η^2	*p*
Prompt	5	25.76	0.06	<.0001
Test country/territory	2	307.72	0.23	<.0001
Prompt x Test Country/territory	10	6.98	0.03	<.0001
Within-group error	3,567	(0.37)		

Appendix K

Impact Analysis Results for Test Country/Territory Population Groups

Test country/ territory	*N*	Item 8		Item 8 by simulation score (normally rounded to integer)					Item 8 by simulation score (unrounded)			
				Simulated score		Agreement			Simulated score		Agreement	
		Mean	SD	Mean	SD	Quadratic-weighted kappa	Exact percentage agreement	Adjacent percentage agreement	Mean	SD	Standardized mean score difference	Correlation coefficient
Adjudication threshold: 1.0												
KOR	18,288	2.91	0.64	2.93	0.68	0.92	72.22	99.92	2.92	0.66	0.02	0.93
JPN	7,505	2.73	0.70	2.75	0.74	0.93	71.90	99.96	2.71	0.72	-0.04	0.94
IND	4,522	3.66	0.57	3.69	0.64	0.85	59.55	99.93	3.66	0.51	0.00	0.88
CHN	968	2.90	0.61	2.94	0.63	0.91	74.48	99.90	2.93	0.60	0.05	0.93
GBR	601	2.81	0.75	2.86	0.79	0.93	68.39	100.00	2.77	0.76	-0.04	0.95
TWN	493	3.15	0.67	3.14	0.72	0.92	74.44	99.80	3.16	0.65	0.02	0.94
PHL	346	2.92	0.93	2.98	0.97	0.95	64.16	100.00	2.89	0.95	-0.03	0.97
KAZ	335	2.43	0.89	2.48	0.94	0.93	67.16	99.70	2.39	0.94	-0.05	0.95
FRA	299	2.79	0.67	2.81	0.74	0.91	68.23	100.00	2.82	0.66	0.05	0.94
PAK	232	1.95	0.74	2.04	0.75	0.92	66.81	100.00	1.97	0.77	0.03	0.95
IDN	185	3.09	0.62	3.13	0.72	0.89	63.78	100.00	3.14	0.64	0.07	0.93
USA	128	3.64	0.77	3.69	0.82	0.87	64.06	99.22	3.59	0.68	-0.06	0.88
CAN	106	3.14	0.48	3.21	0.55	0.87	73.58	73.58	3.20	0.45	0.13	0.88
Adjudication threshold: 1.5												
KOR	18,288	2.91	0.64	2.91	0.69	0.90	70.74	99.90	2.92	0.66	0.02	0.92
JPN	7,505	2.73	0.70	2.73	0.75	0.91	70.42	99.91	2.70	0.72	-0.04	0.93
IND	4,522	3.66	0.57	3.67	0.61	0.82	57.74	99.91	3.66	0.50	0.00	0.87
CHN	968	2.90	0.61	2.94	0.62	0.88	72.93	99.59	2.94	0.58	0.07	0.91
GBR	601	2.81	0.75	2.82	0.78	0.91	65.89	100.00	2.76	0.76	-0.06	0.94
TWN	493	3.15	0.67	3.15	0.71	0.91	72.82	99.80	3.16	0.64	0.03	0.93
PHL	346	2.92	0.93	2.93	0.95	0.93	60.98	100.00	2.88	0.93	-0.05	0.96
KAZ	335	2.43	0.89	2.41	0.98	0.91	63.58	99.70	2.37	0.96	-0.06	0.94
FRA	299	2.79	0.67	2.81	0.73	0.90	67.22	100.00	2.83	0.66	0.05	0.94
PAK	232	1.95	0.74	2.01	0.79	0.87	61.21	99.57	1.97	0.78	0.02	0.93
IDN	185	3.09	0.62	3.10	0.70	0.89	63.24	100.00	3.13	0.64	0.07	0.93
USA	128	3.64	0.77	3.63	0.77	0.85	63.28	99.22	3.58	0.67	-0.07	0.87
CAN	106	3.14	0.48	3.20	0.51	0.84	72.64	100.00	3.21	0.44	0.14	0.87

Note. KOR = Korea; JPN = Japan; IND = India; CHN = China; GBR = Great Britain; TWN = Taiwan; PHL = The Philippines; KAZ = Kazakhstan; FRA = France; PAK = Pakistan; IDN = Indonesia; USA = United States; CAN = Canada.