An exploratory study of the construct measured by automated writing scores across task types and test occasions

Khaled Barkaoui & Johanathan Woodworth
York University

With the growing use of Automated Essay Scoring (AES) to score second language (L2) writing performance, questions are often raised concerning the interpretation of automated scores. Such questions are typically investigated by examining the association between automated scores and human holistic ratings at one point in time. However, this line of research cannot answer questions about whether and how this association varies across tasks and test occasions. This exploratory study addresses this gap by examining the association between automated writing scores and human multiple-trait ratings of essays written by 48 learners of English in response to TOEFL iBT independent and integrated writing tasks on two test occasions, before and after a period of English language study. Each essay (*N*= 192) was scored by *e*-rater and rated by a group of human raters on various writing features. The findings indicated that the associations between *e*-rater scores and human ratings of some writing features varied significantly across task types and tended to be stronger for essays written after, than for essays written before, English language study. We discuss the findings and their implications for future research on automated scoring in L2 writing assessment.

Key words*:* automated essay scoring, human ratings, second language writing, task effects.

Over the last few decades, there has been a steady increase in Automated Essay Scoring (AES) to score second language (L2) writing performance, particularly in large-scale writing tests. AES systems use natural language processing (NLP) techniques to analyze

Email address for correspondence: barkaoui@yorku.ca

the linguistic features of an essay (e.g., grammar, lexis, discourse) and then weigh these features to estimate what score a human rater would assign to the essay (Attali & Burstein, 2006; Shermis et al., 2010). Different AES systems analyze and evaluate different features, but all AES systems are trained on large corpora of human-rated essays to estimate the optimum weights to apply to the linguistic features to predict human ratings (Shermis et al., 2010). This training results in a scoring model or algorithm to predict or emulate human ratings of new essays. The scoring model can be generic or prompt-specific. A generic model uses a general statistical model that applies to all essays regardless of prompt, while a prompt-specific model is configured for essays on specific prompts. Before they are used for operational rating, AES scores are compared to human ratings until an acceptable level of agreement between AES scores and human ratings is achieved (Deane & Quinlan, 2010). A major reason for using AES systems for scoring is that they are more efficient, cost-effective, and consistent than human rating (e.g., Shermis & Hamner, 2013; Weigle, 2013a; Weigle, 2013b). Some writing assessment systems use only AES scores (e.g., the Duolingo English Test), while others combine human and AES scores (e.g., the TOEFL iBT).

The increasing use of AES systems in assessing writing is not without controversy, however, particularly concerning the interpretation of AES scores and, by extension, the validity of the explanation inference of assessments using AES scoring.[2] The explanation inference is based on the warrant that variation in AES scores can be attributed to the target construct of writing ability as defined by the test developers (Chapelle et al., 2008; Knoch & Chapelle, 2018). A key concern is that AES scores have a different interpretation than human ratings because they are generated through different processes and are based on a different definition and operationalization of the writing construct. In terms of processes, human ratings usually involve a recursive process of interpreting a text and

[2] A validity argument applies to an assessment system as a whole, including procedures for eliciting performance (i.e., writing tasks) as well as procedures for evaluating and responding to such performance (e.g., rubrics, rater training). AES is just one component in a writing assessment system. Because AES systems are intended to support, supplement, or replace human raters, the warrants and assumptions underlying the validation of rating processes apply to the validation of the use of AES scores (Clauser et al., 2002; Williamson et al., 2012). See Knoch and Chapelle (2018) for a list of the warrants and assumptions underlying the validation of rating processes.

judging its quality based on a specific set of criteria, expectations, and/or in comparison to other texts (e.g., Barkaoui, 2011; Cumming et al., 2002; Lumley, 2005). In contrast, AES systems usually use a linear weighting of the linguistic features included in their algorithms to compute a score for a given essay (Ramineni & Williamson, 2018). For this reason, in this paper, we use *scores* to refer to marks generated by AES systems and *ratings* to refer to marks generated by human raters.

In addition to involving different processes, AES systems tend to define and operationalize the construct of writing differently than human raters. For example, AES systems may use different criteria and/or weigh the same criteria differently than human raters. Ramineni and Williamson (2018), for example, found that, compared to human raters, *e*-rater (a scoring engine developed and used by Educational Testing Services [ETS]) seems to be less severe on language errors, to overvalue organization and development, and occasionally to undervalue content. Consequently, AES systems and human raters may assign different marks to the same essays and/or assign similar marks for different reasons (e.g., Chen & Cheng, 2008). Additionally, AES scores seem more susceptible to construct underrepresentation and construct-irrelevant variance (Attali, 2007). Construct underrepresentation occurs when a writing assessment does not consider all the dimensions that writing theories and experts agree are important aspects of writing proficiency (Attali, 2007; Ben-Simon & Bennett, 2007; Deane, 2013a).

One of the main critiques of AES is that it only measures a subset of the writing construct. Currently, AES systems cannot measure the meaningfulness of content, argument quality, or the rhetorical effectiveness of writing and can measure other aspects of writing, such as text organization, only indirectly (Deane, 2013a). For example, Chodorow and Burstein (2004) found that human raters were sensitive to writing characteristics that the AES system was not sensitive to. Shermis et al. (2008) found that content plays a relatively minor role in the overall score that Criterion (an AES system based on *e*-rater) assigns to essays, accounting for 1% to 6% of the variance in essay scores. Similar findings have led Ben-Simon and Bennett (2007) to conclude that AES systems tend to provide only partial coverage of the writing construct.

Another issue is that AES scores often suffer from an over-representation of construct-irrelevant features, particularly essay length (e.g., Chodorow & Burstein, 2004; Enright & Quinlan, 2010). Enright and Quinlan (2010), for example, found that *e*-rater scores for essay organization and development were highly correlated with essay length and that essay length accounted for approximately 60% of the variance in *e*-rater scores. Likewise, Jones (2006) found that essay length accounted for 85% of the variance in scores generated by the AES system, IntelliMetric. It should be noted here that text length is one of the strongest predictors of human ratings as well. Crossley and McNamara (2016) reviewed studies indicating that text length is generally the strongest predictor of essay quality and that proficient writers tend to produce longer texts. In L2 writing research, essay length is often used as an indicator of L2 fluency (Woodworth & Barkaoui, 2021), but, as Crossley and McNamara (2016) noted, essay length can be an indicator of text development or "the amount of content that an essay contains" (p. 354). Some AES systems control for essay length statistically so that it does not affect AES scores, ensuring that the AES system measures the quality, rather than the quantity, of writing. One key implication of the findings concerning the effects of essay length is that future studies of the relationships between automated scores and human ratings need to examine and control for the effects of text length.

Literature on the association between AES scores and human ratings

A review of the literature indicated that the typical approach to evaluating the quality of AES scores is to examine the degree of agreement (or association) between AES scores and human ratings of the same writing samples (Attali & Burstein, 2006; Shermis et al., 2010; Williamson et al., 2012). A systematic search of studies comparing AES scores and human ratings of L2 writing published between 2000 and 2020 returned 30 relevant studies. These studies are listed in Appendix A, where the focus, research questions, and research methods of each study are described. In describing these studies, following Hamp-Lyons (1991), we distinguish between two methods of AES scoring, overall and analytic scoring, and two methods of human rating, holistic and multiple-trait rating. As

noted above, AES systems identify and count specific linguistic features in each essay and then use specific algorithms to compute essay scores using these counts. The system then assigns to each essay either multiple analytic scores (i.e., one score for each feature) or one overall score that is based on some linear weighting of all the features in the algorithm (Ramineni & Williamson, 2018). Similarly, human holistic rating involves assigning one mark that reflects the rater's overall judgment of the quality of an essay based on specific criteria. In contrast, multiple-trait rating involves assigning multiple marks to an essay that reflect the rater's judgments of the qualities of the essay's multiple, specific writing features (e.g., organization, language) (Hamp-Lyons, 1991).

Most of the studies we identified have examined the association between AES overall scores and holistic ratings assigned by trained human raters to the same essays. The majority of these studies show that AES overall scores correlate positively and strongly with human holistic ratings (cf. Shermis & Hamner, 2013). *e*-rater scores, for example, tend to correlate highly with holistic human ratings (Attali & Burstein, 2006; Ramineni & Williamson, 2018; Shermis, 2014). Attali and Burstein (2006), for example, reported a correlation of .97 between *e*-rater scores and human holistic ratings of a sample of about 2,000 essays. In another study, Attali et al. (2010) found *e*-rater's agreement with a human rater on the TOEFL iBT writing section to be similar to that of two independent human raters. Other studies reported similar results for other AES systems such as PEG (Shermis et al., 2001; Shermis et al., 2002; Wilson et al., 2016), Bayesian Essay Test Scoring System (Coniam, 2009), MarkIT (Williams & Dreher, 2005), and IntelliMetric (Mikulas & Kern, 2006; Wang & Brown, 2008). Mikulas and Kern (2006) and Rudner et al. (2006), for example, reported correlations of .94 or higher between IntelliMetric scores and human ratings. However, some studies comparing AES scores and human ratings in terms of exact agreement and rank found some differences suggesting that they measure somewhat different constructs (Ben-Simon & Bennett, 2007).

As noted above, AES systems use human ratings to build and optimize statistical models for scoring. As a result, the strength of the association between AES scores and human ratings is an obvious first criterion for evaluating AES scores (Williamson et al., 2012). Although they have their limitations, human ratings are used as the standard because the

response of a reader is an authentic criterion for evaluating written texts. Additionally, a long history of research shows that human ratings are trustworthy, particularly if raters receive adequate training and use specific rating criteria (Hamp-Lyons, 1991; Weigle, 2002). However, the interpretation of the correlations between AES scores and human ratings can be problematic (Bennett & Bejar, 1998). For example, a high correlation between AES scores and human ratings does not necessarily mean that they measure the same construct (Weigle, 2013b). The AES algorithm may consider different aspects of writing than those usually considered by human raters. For example, while AES systems may count the frequency of specific discourse units, a human rater may pay more attention to an essay's overall organization and flow (Weigle, 2013a, p. 43).

As a result of these issues, some researchers have argued that high correlations between AES scores and human ratings are a necessary but not a sufficient condition for the validity of AES score use "because agreement results tell us little about what is measured by automated scores" (Attali, 2007; cf. Shermis & Burstein, 2003). To address this issue, Ben-Simon and Bennett (2007) argued that establishing the validity of the explanation inference of AES scoring requires (a) providing evidence that the features these systems measure adequately cover the dimensions that experts and theories agree are important aspects of writing proficiency and (b) evaluating the relations of automated feature scores to human ratings of the same features (p. 79).

One way some researchers have tried to address the issues associated with correlational studies is by conducting experiments in which the AES system is "tricked." This involves editing or writing essays with specific characteristics and then examining how AES systems score them (e.g., McGee, 2006; Powers et al., 2002). McGee (2006), for example, manipulated specific features of content, style, and mechanics of a set of essays to examine how they impact AES overall scores. McGee found that changing the sentence order and the truth value of propositions did not change AES scores.

Another approach is to examine the relationships between automated scores and human ratings, on the one hand, and non-test indicators of writing ability (e.g., measures of L2 proficiency, grades in writing courses), on the other (e.g., Powers et al., 2015; Weigle,

2013a). Evidence that AES scores achieve similar levels of correlations as human ratings with non-test indicators of writing ability can shed light on whether AES scores measure the same construct. As Williamson et al. (2012) observed, "if human and automated scores reflect similar constructs, they are expected to relate to other measures of similar or distinct constructs in similar ways" (p. 9). Powers et al. (2000) examined the correlations of human ratings and *e*-rater scores with several non-test indicators. They found that *e*-rater scores correlated slightly less strongly with each of the non-test indicators than did human ratings; however, the indicators that related most strongly to human ratings also related most strongly to *e*-rater scores and vice versa. Weigle (2010) found that *e*-rater scores and human ratings of essays on the TOEFL iBT independent writing tasks were highly correlated with the non-test indicators of writing ability and, thus, can be said to be measuring highly similar constructs. Other studies have investigated this question by examining the correlations between automated writing scores and scores on other sections of L2 proficiency tests. Ramineni et al. (2012a), for example, found that the correlations between *e*-rater scores and human ratings for TOEFL iBT writing tasks, on the one hand, and scores on the TOEFL iBT reading, listening, and speaking sections, on the other, were similar for independent prompts. However, for integrated prompts, the correlations of *e*-rater scores with scores on other test sections were uniformly lower than those for human ratings (cf. Lee, 2016). The current study also examines the correlations between *e*-rater scores and scores on other sections of the TOEFL iBT.

Limitations of previous research

Studies on the association between AES scoring and human ratings have three main limitations that the current study aims to address: a tendency to rely mainly on human holistic ratings; few comparisons of AES systems across writing tasks; and a lack of research on whether the association between AES scores and human ratings varies across test occasions. First, most studies we reviewed used overall AES scores and holistic human ratings. Only a few studies have used multiple-trait ratings and analytic AES scores of specific writing features (e.g., McGee, 2006). In all these studies, the primary

focus is on score reliability. Shermis et al. (2002), for example, examined the correlations between AES overall and analytic scores, on the one hand, and human holistic and multiple-trait ratings for five traits (content, creativity, style, mechanics, and organization), on the other. They found that interrater reliability varied significantly across traits and between human ratings and AES scores. Wang and Brown (2008) found that the correlations between AES overall scores and human holistic ratings tended to be lower than those between AES overall scores and human multiple-trait ratings.

Second, few studies have investigated whether AES scores are sensitive to variability in writing performance across task types. Task-related variability in writing scores is often addressed in relation to the generalization inference of writing tests. The generalization inference is based on the warrant that observed scores are accurate "estimates of expected scores that test takers would receive on comparable tasks, test forms, administrations, raters, and rating conditions" (Chapelle et al., 2008, p. 329). The consistency of AES scoring is one of the main reasons for the increasing use of AES systems. Additionally, some studies have found that task type does not seem to affect the strength of the association between AES scores and human ratings (Attali & Powers, 2009; Powers et al., 2000; Ramineni & Williamson, 2018). Attali and Burstein (2006), for example, found the exact agreements between *e*-rater scores and human ratings to be similar for GMAT argument, GMAT issue, and TOEFL writing tasks. However, when AES systems are used to score essays on different tasks or prompts, it is difficult to discern whether they are measuring different writing features across prompts or whether they are simply weighing the same features differently across prompts. As Ben-Simon and Bennett (2007) warned, when AES systems weigh the features differently on different prompts to predict human ratings, such adjustment may essentially constitute a construct redefinition.

Another issue is that, because AES systems cannot assess content or task difficulty, they can perform poorly when assessing different prompts without having a specific scoring model for each prompt (Quinlan et al., 2009). To adjust for prompt difficulty and content knowledge, AES systems can be trained using samples of essays on different prompts to develop a prompt-specific scoring model for each prompt. Higgins et al. (2006) found that topic-specific scoring models performed better than other models for distinguishing

on-topic from off-topic essays. Ramineni et al. (2012a; 2012b) found that the *e*-rater prompt-specific model performed better than the generic model for GRE issue and argument prompts as well as TOEFL iBT writing prompts. Similar findings have been reported for other AES systems (e.g., Mikulas & Kern, 2006, for IntelliMetric).

The third issue concerns the lack of research on whether the association between AES scores and human ratings varies across test occasions. This issue is particularly relevant when using automated scoring to measure changes in L2 writing ability over time and/or in relation to L2 instruction. The issues and questions concerning variability in AES scoring across tasks and prompts apply to comparisons across occasions as well. For example, do AES systems weigh different writing features differently when scoring essays written by the same learners on different test occasions (e.g., before and after a period of L2 study)? Does the association between AES scores and human ratings vary across occasions? We were not able to identify any studies that address these questions. Most of the studies we identified have compared AES scores across proficiency levels at one point in time. Attali and Burstein (2006) examined AES scores across test occasions, but the goal of the study was to estimate the true-score correlation between *e*-rater scores and human ratings rather than to track changes in writing performance after a period of language study. Attali and Powers (2009) examined whether *e*-rater scores could measure writing development in two modes of writing, but the study was cross-sectional as it compared the writings of students at different grade levels at one point in time. Both studies used holistic ratings, which may lack sensitivity to changes in writing performance. For example, test-takers may have different levels of proficiency in different areas of writing (e.g., language, organization), resulting in uneven profiles that could not be captured in a single holistic score (cf. Weigle, 2002). This issue applies to holistic ratings of writing performance across test occasions as well since different test takers may develop different levels of proficiency in different areas of writing after a period of L2 study. Multiple-trait rating may be more sensitive to variability in the proficiency of L2 learners in different aspects of writing across tasks, occasions, and contexts (Barkaoui, 2011).

The present study

To address the limitations outlined above, this exploratory study aimed to examine variation in the relationships between the constructs measured by *e*-rater scores and human multiple-trait ratings across tasks and test occasions. The study is part of a larger study that examined changes in several writing features of the responses of a group of EFL students to the TOEFL iBT independent and integrated writing tasks before and after a period of English language study (Woodworth & Barkaoui, 2021). The goal of the current study is to examine the associations between *e*-rater scores and human ratings of these aspects of writing and whether and how these associations vary across tasks and test occasions. The aspects of writing examined in the study are intended as an operationalization of the construct of L2 writing and the study aimed to evaluate the construct measured by *e*-rater scores by comparing *e*-rater scores to human ratings of these writing aspects. Test occasion in this study is used as a proxy for change in English language proficiency after a period of English language study/instruction to investigate the progression of students' English language proficiency over time. Specifically, by examining the correlations between *e*-rater scores and human ratings of essays written by the same students before and after a period of English study (see below), the study aimed to evaluate whether the construct measured by *e*-rater changes as student proficiency changes. The study addressed the following research questions:

RQ1: What is the relationship between *e*-rater scores and human ratings, on the one hand, and scores on the other sections of the TOEFL iBT, on the other? Do these relationships vary across tasks and occasions (before and after L2 study)?

RQ2: What is the relationship between essay length, on the one hand, and *e*-rater scores and human ratings, on the other? Do these relationships vary across tasks and occasions?

RQ3: What is the relationship between *e*-rater scores and human ratings of important aspects of writing? Do these relationships vary across tasks and occasions?

Method

The scripts and their *e*-rater scores used in this study were obtained from Ling et al. (2014). Ling et al. administered the TOEFL iBT practice test to a group of 90 students in an international high school in China before and after nine months of English language study. All the students were native Chinese speakers. Students took the first TOEFL iBT practice test toward the end of the first year in high school and the second test nine months later. All the students took the two tests voluntarily but were not aware of each test until one or two weeks before the test date. Between the two tests, the participants took the high school English classes required by the general educational guidelines in China, together with other high school classes in Chinese. The English-related coursework was about 15 hours a week on average.

Each participant responded twice to the same TOEFL iBT independent and integrated writing tasks. The independent writing task consisted of writing an argumentative essay about a general topic (30 minutes), while the integrated task consisted of listening to a two-minute lecture and reading a 200-300-word text about the same topic and then writing a summary of both the lecture and the reading (20 minutes). Of the 90 participants, only 48 responded to both writing tasks before and after instruction and were included in this study, resulting in a final corpus of 192 essays. The participants' total TOEFL iBT scores on occasion 1 (before instruction) ranged between 11 and 105 (M= 49.92, SD= 22.38). The low TOEFL iBT scores maybe due to students' low proficiency and/or low motivation given that they completed the test for research purposes.

Automated scores

e-rater scores were computed for the 192 essays as part of Ling et al.'s (2014) study. *e*-rater is an AES system developed at ETS and used in combination with human raters to score the writing sections of the TOEFL iBT. Only the overall *e*-rater scores were used in this study. *e*-rater measures grammar, usage, mechanics, style, organization, development, lexical complexity, and prompt-specific vocabulary usage (Attali, 2013). The *e*-rater overall score ranges from 0 to 6 and is a weighted average of the standardized values of these features and is modelled to predict human scores.

Human ratings

Woodworth and Barkaoui (2021) built and expanded upon the Model of Writing Competence by Connor and Mbaye (2002) and previous research to develop a framework for analyzing and rating L2 essays in terms of six important aspects of writing that are grounded in writing theory and research: grammatical, discourse, sociolinguistic, strategic, content, and source use. The following paragraphs describe the six aspects of writing rated in the study (see Barkaoui and Hadidi, 2021, for copies of the rating scales used in the study).

Grammatical aspects were operationalized in terms of linguistic accuracy, which was measured using a four-point holistic rating scale developed by Chan et al. (2015). The scale evaluates impressionistically the overall impact of errors involving grammar, punctuation, spelling, and word choice on the comprehensibility of the whole essay (cf. Di Gennaro, 2009).

Discourse aspects were operationalized in terms of coherence, cohesion, and text organization. The coherence scale included eight statements (e.g., The ideas in the essay are well-related one to another) from Chiang (1999), while the cohesion rating scale included four statements (e.g., New information is introduced in an appropriate place or manner) from Chiang (2003). Each statement was measured on a four-point scale (strongly disagree to strongly agree). Statement scores were then averaged to obtain one overall cohesion score and one overall coherence score for each essay. In addition, an analytic rating scale, based on the holistic rubric for scoring text organization developed by Kubota (1998), was used to rate the essays in terms of the main idea (i.e., whether the main idea is stated clearly and effectively in the essay), reader orientation (i.e., whether the writer attends to the readability of what they write for the audience) and essay organization (i.e., how well the essay is structured). Each of the three features was rated on a four-point scale. Scores on the three criteria were averaged to obtain one overall text organization score for each essay.

Sociolinguistic aspects were operationalized in terms of register which was measured using a four-point rating scale adapted from one developed by Di Gennaro (2009). The register rating scale evaluated holistically how well each essay represented a formal, written register, one that is appropriate for academic contexts as opposed to informal conversational registers.

Strategic aspects were operationalized using two four-point rating scales of the quality of metadiscourse use in each essay. To evaluate the quality of interpersonal (interactional) metadiscourse or voice, we used a modified version of the holistic scale of voice, with four levels, developed by Zhao (2017). Another four-point, holistic rating scale was developed to rate the quality of the use of interactive (textual) metadiscourse markers. Raters were asked to rate each essay holistically in terms of whether it uses interactive metadiscourse markers judiciously and accurately.

Content: Only essays on the independent task were rated in terms of content, which was operationalized as argument quality. Specifically, each essay was rated in terms of six features of argument quality using a rating scale from Cumming et al. (2005): thesis, claims, data, warrants, recognition of opposition, and response to opposition. Each feature was evaluated on a four-point scale in terms of its clarity, relevance, and completeness. Scores on the six features were averaged to obtain one overall argument quality score for each essay.

Source Use: Only essays on the integrated task were rated in terms of source use. A modified version of Chan et al.'s (2015) rubric for evaluating reading-into-writing skills was used to rate the quality of the source used in each integrated essay. The rubric included three criteria on a four-point scale: understanding source materials, selecting relevant content from source texts, and identifying common themes and links across sources. Scores on the three criteria were averaged to obtain one overall source use score for each essay.

Each essay was rated by two independent raters on each rating scale after extensive rater training and establishing acceptable levels of inter-rater agreement. Eleven raters, who had three to more than 20 years of experience teaching ESL and writing, rated the essays.

Each rater was assigned different subsets of essays for each writing aspect and rating scale to avoid the halo effect. Ratings assigned to the same essay were then compared across raters. When there was a difference greater than one point between two raters' scores for the same essay, the two raters discussed and resolved the discrepancy so that the final difference was no more than one point. The final score for each essay for each writing feature is the average of the scores from the two raters.

Inter-rater agreement was computed in two ways. First, we computed the intra-class correlation (ICC) among the ratings of the four raters, before resolving disagreements, for each rating scale for all the essays in the main study (*N*= 276 essays). As Table 1 shows, all measures have relatively high inter-rater reliability indices, with ICC ranging between .74 (for cohesion ratings) and .94 (for argument quality and source use ratings). Second, we computed the Pearson *r* correlation between the ratings of each pair of raters, before resolving disagreements, for each rating scale for the 192 essays in this study. The results of these analyses, which are consistent with the results in Table 1, are reported in Table 5 below. Finally, a word processing program was used to compute the number of words per essay.

Table 1. Reliability statistics for human ratings of writing features (*N*= 276 essays)

Writing Feature	**Intra-Class Correlation (ICC)**
Accuracy	.76
Coherence	.80
Cohesion	.74
Organization	.82
Register	.86
Metadiscourse	
Voice	.89
Textual	.86
Argument Quality	.94
Source Use	.94

Statistical analyses

To address RQ1, Pearson r correlations between *e*-rater scores and human ratings, on the one hand, and scores on other sections of the TOEFL iBT (i.e., listening, speaking, and

reading), on the other, were examined. To address RQ2, the correlations between essay length, on the one hand, and *e*-rater scores and human ratings, on the other, were computed. Pearson *r* correlations between *e*-rater scores and human ratings of the different aspects of writing were computed to address RQ3. Correlations below .30 are considered low; correlations between .30 and .69, moderate, and correlations of .70 and above, high. To examine variability in the correlation coefficients across tasks and occasions (RQs 1-3), correlation coefficients were compared across tasks and occasions using the interactive calculators developed by Lee and Preacher (2013; Preacher, 2002). These calculators test the equality of two correlation coefficients by first converting each correlation coefficient into a *z*-score using Fisher's *r*-to-*z* transformation and then comparing the two estimates to determine if they differ significantly. These correlation-difference tests rely on asymptotic distribution theory and have been shown to be rather robust in small samples (Lee & Preacher, 2013; Preacher, 2002). Appendix B includes descriptive statistics for all measures in the study by task type and test occasion.

Findings

RQ1: Associations between e-rater scores and the TOEFL iBT section scores

Table 2 reports the correlations between *e*-rater scores and TOEFL iBT reading and listening scores by task and occasion. TOEFL speaking scores were not included because they were estimated using automated scoring too. Most of these correlations are moderate but significant. The correlations of *e*-rater scores with listening scores are slightly higher than those with reading scores for both tasks and both occasions. Additionally, the correlations of *e*-rater with listening and reading scores increased on occasion 2. The correlations increased from .26 for reading and .30 for listening to .40 and .51, respectively, for the integrated task. For the independent task, they increased from .39 and .50 to .42 and .47. None of these increases was statistically significant, however.

Table 2. Correlations between *e*-rater scores and TOEFL iBT section scores by task and occasion

	e-rater scores	
Task	*Occasion 1*	*Occasion 2*
Integrated		
Reading scores	.26	.40**
Listening score	.30*	.51**
Independent		
Reading score	.39**	.42**
Listening score	.50**	.47**

N= 48 students
** P<.05 ** p<.01*

Table 3 reports the correlations between human ratings and TOEFL iBT section scores by task and occasion. Generally, human ratings tended to have slightly stronger correlations with listening scores than they did with reading scores across tasks and occasions, except for the integrated task for occasion 1, as shown in Table 3. Table 3 also shows the following patterns:

- Generally, human ratings of coherence, cohesion, organization, and textual metadiscourse tended to have the highest correlations with scores on the other sections of the TOEFL iBT for both tasks and occasions.
- Human ratings of source use (for the integrated task) and argument quality (for the independent task) correlated highly and significantly with scores on the other TOEFL iBT sections. These correlations tended to be higher on occasion 1 than on occasion 2 for the integrated task and higher on occasion 2 than on occasion 1 for the independent task for reading and listening scores.
- On average, the correlations between human ratings of all writing features and scores on the other TOEFL iBT sections were higher for occasion 1 than for occasion 2 for both tasks.
- The correlations between human ratings and scores on the other TOEFL iBT sections do not seem to vary much across tasks for both occasions except for voice ratings with reading and listening scores.

Table 3. Correlations between human ratings (average of 2 raters) and TOEFL iBT section scores by task and occasion

TOEFL Section	Reading Score		Listening Score	
Occasion	1	2	1	2
Integrated Task				
Accuracy	.27	.24	.26	.29*
Coherence	.62**	.37**	.55**	.43**
Cohesion	.48**	.29*	.40**	.34*
Organization	.50**	.27	.48**	.25
Register	.25	.22	-.01	.36*
Voice	.52**	.04	.49**	.11
Textual	.35*	.27	.47**	.36*
Source Use	.56**	.38**	.64**	.48**
Independent Task				
Accuracy	.25	.25	.26	.33*
Coherence	.26	.27	.48**	.35*
Cohesion	.35*	.14	.48**	.36*
Organization	.20	-.01	.46**	.06
Register	.15	.26	.32*	.29*
Voice	.15	.35*	.29*	.44**
Textual	.39**	.04	.44**	.23
Argument Quality	.25	.32*	.23	.46**

N= 48 students
* P<.05 ** p<.01

RQ2: Associations between essay length and e-rater scores and human ratings

Table 4 reports the correlations between essay length, on the one hand, and *e*-rater scores and human ratings, on the other, by task and occasion. As Table 4 shows, *e*-rater scores were significantly, positively, and strongly correlated (r >.80) with essay length for all occasions and tasks. The correlations between human ratings (average of two raters) and essay length are also positive and high for most writing features, except for accuracy and register ratings. As shown in Table 5, the number of words per essay was strongly correlated with human ratings of voice, source use, organization, coherence, and textual metadiscourse, and moderately correlated with human ratings of cohesion and argument quality. In contrast, accuracy and register ratings had very low correlations with essay

length, particularly for the independent task on occasion 1. These patterns suggest that longer essays tended to receive higher *e*-rater scores as well as higher human ratings except for accuracy and register.

Table 4. Correlations of essay length with human ratings (average of 2 raters) and e-rater scores by task and occasion

	Number of words	
Occasion	*1*	*2*
Integrated Task		
e-rater scores	.84**	.89**
Human ratings		
Accuracy	.12	.25
Coherence	.62**	.73**
Cohesion	.51**	.60**
Organization	.63**	.72**
Register	-.03	.27
Voice	.66**	.45**
Textual	.76**	.68**
Source Use	.71**	.80**
Independent Task		
e-rater scores	.83**	.87**
Human ratings		
Accuracy	.06	.42**
Coherence	.66**	.64**
Cohesion	.48**	.44**
Organization	.69**	.51**
Register	.34*	.34*
Voice	.72**	.55**
Textual	.72**	.62**
Argument Quality	.57**	.59**

N= 48 students
** P<.05 ** p<.01*

To further examine the association between individual human ratings and essay length and the impact of essay length on levels of inter-rater agreement, we conducted three sets of correlational analyses for the 192 essays in this study. First, we computed the correlations between the ratings of each pair of raters for each rating scale. Table 5 reports the lowest, highest, and average correlations between pairs of raters for each rating scale. It shows that the lowest average correlations between pairs of raters were observed for

cohesion (average r =.57) and accuracy (average r =.61), while the highest average correlations were observed for argument quality and source use (average r = .89). These results are consistent with the interclass correlation coefficients reported in Table 1 above. Second, we computed the correlations between the ratings assigned by each rater and essay length for each rating scale. As Table 5 shows, human ratings of accuracy (average r =.15) and register (average r =-.16) have the weakest correlations with essay length, while ratings of source use (average r =.74), voice (average r =.71), organization (average r =.61), argument quality (average r =.60), and coherence (average r =.60) have the strongest correlations with essay length.

Third, we computed the partial correlation between the ratings of each pair of raters while controlling for essay length to examine whether essay length impacted inter-rater agreement for each rating scale. Partial correlation examines the relationship between two variables (e.g., ratings by two human raters) while controlling for the effect of a third variable (e.g., essay length) has on both (Field, 2009). As Table 5 shows, when essay length is controlled for, the average correlations between pairs of human raters decreased slightly, except for the correlations for the ratings of voice, which exhibited a large decline from .77 to .39 when essay length is considered. The inter-rater correlations for ratings of source use and organization also decreased when essay length was controlled for. The inter-rater correlations for register, cohesion, argument quality, and accuracy were the least affected by essay length.

Table 5. Correlations between human ratings and essay length and partial correlations between human ratings

	Correlations between Ratings of Pairs of Human Raters			*Correlations of Individual Human Ratings with Essay Length*			*Partial Correlation between Pairs of Human Ratings*[a]		
Writing Feature	*Lowest*	*Highest*	*Average*	*Lowest*	*Highest*	*Average*	*Lowest*	*Highest*	*Average*
Accuracy	.40	.74	.61	.01	.37**	.15	.12	.74**	.56
Coherence	.80**	.89**	.85	.55**	.66**	.60	.58**	.81**	.75
Cohesion	.19	.78**	.57	.26*	.48**	.37	.32	.73**	.55
Organization	.52**	.91**	.77	.48**	.69**	.61	.11	.86**	.61
Register	.57**	.89**	.74	-.22*	-.07	-.16	.57**	.88**	.74
Voice	.63**	.95**	.77	.62**	.80**	.71	.21	.88**	.39
Textual	.68**	.97**	.79	.41**	.54**	.48	.64**	.95**	.72
Argument Quality	.68*	.95**	.89	.52**	.66**	.60	.65*	.94**	.85
Source Use	.85**	.94**	.89	.66**	.83**	.74	.63	.85**	.74

* *P<.05* ** *p<.01*

[a] Controlling for essay length

RQ3: Association between e-rater scores and human ratings

Table 6 reports the zero-order (Pearson *r*) and partial (controlling for essay length) correlations between human ratings and *e*-rater scores by task and occasion. For the integrated task, *e*-rater scores were significantly and positively correlated with human ratings for all writing features for both occasions, except for accuracy and register on occasion 1. However, when essay length is taken into account, *e*-rater scores correlated significantly only with human ratings of accuracy and source use on occasion 2. For the independent task, *e*-rater scores were significantly and positively correlated with human ratings for all writing features for both occasions, except for accuracy on occasion 1. When essay length is taken into account, *e*-rater scores correlated significantly only with human ratings of coherence, register, and textual metadiscourse on occasion 1 and with accuracy, register, and voice ratings on occasion 2.

Comparisons of the correlations between *e*-rater scores and human ratings across tasks and occasions in Table 6 indicate the following patterns. First, the correlations of *e*-rater scores with human ratings of organization, register, and voice exhibited significant differences across tasks. The correlation between organization ratings and *e*-rater scores for the integrated task (r = .71) was significantly stronger (Z= 1.97, p<.05) than that for the independent task (r =.44) on occasion 2, but not on occasion 1 (r = .44 and .59, respectively). The correlation between register ratings and *e*-rater scores for the independent task (r =.47) was significantly stronger (Z= 1.85, p<.05) than that for the integrated task (r =.12) on occasion 1, but not on occasion 2 (r = .36 and .51, respectively). The correlation between voice ratings and *e*-rater scores for the independent task (r = .67) was significantly higher (Z= 2.06, p <.05) than that for the integrated task (r =.36) on occasion 2, but not on occasion 1 (r = .47 and .64, respectively). When essay length is taken into account, the correlations between *e*-rater scores, on the one hand, and coherence, register and voice ratings, on the other, tended to be higher for the independent task, while the correlation between *e*-rater scores and organization ratings tended to be higher for the integrated task.

Table 6. Zero-order and partial correlations between human ratings and e-rater scores by task and occasion

	e-rater scores			
Occasion	*1*		*2*	
Correlation	*Zero-order*	*Partial*	*Zero-order*	*Partial*
Integrated Task				
Accuracy	.18	.14	.42**	.45**
Coherence	.50**	-.10	.55**	.08
Cohesion	.55**	.27	.64**	.28
Organization	.44**	-.23	.72**	.26
Register	.12	.28	.37*	.28
Voice	.47**	-.24	.36*	-.09
Textual	.67**	.06	.62**	.10
Source Use	.57**	-.10	.85**	.54**
Independent Task				
Accuracy	.16	.20	.52**	.33*
Coherence	.75**	.35*	.65**	.25
Cohesion	.50**	.23	.50**	.27
Organization	.59**	.06	.44**	.05
Register	.47**	.38*	.51**	.42*
Voice	.64**	.16	.67**	.44**
Textual	.76**	.44**	.50**	-.02
Argument	.52**	.14	.59**	.24

N= 48 students
** P<.05 ** p<.01*

Second, the strength of the associations between *e*-rater scores, on the one hand, and human ratings of four writing features (accuracy, coherence, organization, textual metadiscourse, and source use), on the other, exhibited significant differences across occasions. The correlation between accuracy ratings and *e*-rater scores was higher for occasion 2 than occasion 1 for both tasks; the increase (from r = .16 to .51) was significant for the independent task (Z= 1.90, p<.05) but not for the integrated task (from r = .18 to .42). Additionally, the correlation between textual metadiscourse ratings and *e*-rater scores was higher for occasion 1 than occasion 2 for both tasks; the decline (from r = .76 to .50) was significant for the independent task (Z= 2.12, p<.05) but not for the integrated task (from r = .67 to .62). For coherence and source use, there were significant differences for the integrated task but not for the independent task. The correlation between coherence ratings and *e*-rater scores was higher for occasion 2 than occasion 1 for both

tasks; the increase (from r =.50 to .75) was significant for the integrated task (Z= 2.01, p<.05) but not for the independent task (from r = .55 to .64). Finally, the correlation between source use ratings and *e*-rater scores for the integrated task was significantly higher (Z= 2.89, p<.01) on occasion 2 (r =.85) than it was on occasion 1 (r =.57).

When essay length is taken into account, the following patterns emerged. The correlations of *e*-rater scores with human ratings of the following writing features tended to be higher on occasion 1 than they were on occasion 2: coherence and textual metadiscourse for the independent task and voice for the integrated task. The correlations of *e*-rater scores with human ratings of the following writing features tended to be higher on occasion 2 than they were on occasion 1: accuracy for both tasks, source use for the integrated task, and argument quality for the independent task.

Summary and Discussion

This exploratory study aimed to evaluate the construct measured by automated writing scores by examining the correlations between *e*-rater scores and human ratings of 192 essays written by 48 Chinese EFL learners on TOEFL iBT independent and integrated writing tasks before and after a period of English language study. The findings indicated that both *e*-rater scores and human ratings correlated positively with TOEFL iBT reading and listening scores. Both *e*-rater scores and human ratings tended to have slightly stronger correlations with listening scores than they did with reading scores, suggesting that *e*-rater scores and human ratings tap similar aspects of general English language proficiency. Lee (2016) also found that *e*-rater scores tended to correlate the lowest with reading scores. Additionally, the strength of the associations between *e*-rater scores and human ratings, on the one hand, and TOEFL iBT reading and listening scores, on the other, tended to vary across test occasions, but not tasks. While the correlations of *e*-rater scores with TOEFL listening and reading scores increased slightly on occasion 2, those for human ratings tended to be slightly higher for occasion 1. We are not aware of any other studies that have compared the correlations of *e*-rater scores and human ratings with scores on other TOEFL iBT sections across test occasions. Our findings indicated that the relationship between (a) writing ability as measured by either *e*-rater scores or

human ratings and (b) reading and listening abilities does not vary across tasks, suggesting that both *e*-rater and human ratings measure the same aspects of English language proficiency across tasks. However, the variation across test occasions may indicate that changes in listening and reading performance are associated with changes in different aspects of writing and that human ratings and *e*-rater scores differ in terms of their ability to detect such changes in writing performance.

Regardless of test occasion and task type, essay length correlated positively and strongly with *e*-rater scores and human ratings of most writing features, except for ratings of accuracy and register. This finding suggests that human ratings and *e*-rater scores are equally susceptible to essay length effects. However, it is possible that essay length plays different roles in the human rating and *e*-rater scoring processes. For human ratings, essay length effects might be because longer essays usually include more details (Crossley & McNamara, 2016). As the correlation analyses indicated, it was human ratings of writing features most closely related to content (i.e., source use, voice, organization, argument quality, and coherence) that had the strongest correlations with essay length. In contrast, human ratings of features related to language (i.e., accuracy and register) had the lowest correlations with essay length. Additionally, essay length did not seem to have impacted levels of inter-rater agreement much, except for voice ratings. In contrast, *e*-rater does not understand meaning; its scores are based on counting the occurrence of specific features, such as counting the number of words per discourse element to measure essay organization and development, features that are largely influenced by text length (Perelman, 2012). Future experimental studies could shed more light on the effects of essay length on the human rating and *e*-rater scoring processes by manipulating essay length and using think-aloud protocols with raters.

Essay length also affected the strength of the correlations between *e*-rater scores and human ratings. For the integrated task, *e*-rater scores were significantly and positively correlated with human ratings for all writing features, except for accuracy and register on occasion 1. However, when essay length is taken into account, *e*-rater scores correlated significantly only with human ratings of accuracy and source use on occasion 2. For the independent task, *e*-rater scores were significantly and positively correlated with human

ratings for all writing features, except for accuracy on occasion 1. When essay length was taken into account, *e*-rater scores correlated significantly only with human ratings of coherence, register and textual metadiscourse on occasion 1 and with accuracy, register and voice ratings on occasion 2. None of the previous studies on the relationships between automated writing scores and human ratings has examined the moderating effects of essay length on these relationships. Given that essay length may affect the human rating and automated scoring processes differently, it is important that further studies take this variable into account.

Generally, the correlations of *e*-rater scores with human ratings of coherence, organization, register, and voice exhibited significant differences across tasks. Specifically, the correlations between *e*-rater scores, on the one hand, and coherence, register, and voice ratings, on the other, tended to be higher for the independent task, while the correlation between *e*-rater scores and organization ratings tended to be higher for the integrated task. These variations in the strength of the correlations across task types might be because human raters are more sensitive to differences across tasks in coherence, organization, register, and voice; writing features and differences that the version of *e*-rater used in this study is not equipped to detect.

Finally, the strength of the associations between *e*-rater scores, on the one hand, and human ratings of accuracy, coherence, organization, textual metadiscourse, and source use, on the other, exhibited significant differences across occasions. For example, the correlations of *e*-rater scores with human ratings of coherence and textual metadiscourse tended to be stronger on occasion 1 than on occasion 2 for the independent task. In contrast, the correlations of *e*-rater scores with human ratings of accuracy tended to be stronger on occasion 2 than on occasion 1 for both tasks. These variations in correlation strength across test occasions suggest either that human raters and *e*-rater assess these writing features differently and/or that one is more sensitive to changes in some or all of these features than the other. For example, essays may exhibit different types and degrees of change in accuracy after instruction, such as a decrease in the occurrence of linguistic errors with or without improvement in essay comprehensibility. However, while *e*-rater counts the occurrence of language errors, the human raters in this study evaluated only

the impact of language errors on essay comprehensibility. Similarly, essays may exhibit quantitative and/or qualitative changes in coherence, organization, textual metadiscourse use, and source use after instruction, but while human ratings attend to qualitative changes, automated scoring can only attend to changes in the occurrence of linguistic tokens related to these features. However, given the correlational design of the study, these explanations should be treated as hypotheses to be examined in the future, ideally experimental, studies that manipulate changes in specific writing features and then examine whether and how these changes result in changes in human ratings and automated scores.

Implications

Before discussing the implications of the findings, some limitations of the study need to be acknowledged. First, the sample of essays in the study was small, which might have affected the strength of the correlation coefficients. Second, only one task per task type was included; variability between tasks within task type was not examined. Third, the students responded to the same exact tasks before and after instruction. Although this enhances comparability across occasions, it may have introduced some practice effects. Fourth, including only two occasions may not provide a full picture of the changes in writing features across test occasions, particularly if the changes are non-linear. Fifth, we only examined *e*-rater overall scores in this study, although *e*-rater can generate analytic scores too.

Sixth, many of the human ratings in this study involved averaging multiple measures. For example, coherence ratings were based on the average ratings of eight items, while source use ratings were based on the average ratings of three items. Each measure is quite noisy, and averaging them helps address this issue and thus can lead to more reliable ratings, which can, in turn, lead to stronger and more stable correlations between raters and between human ratings and external measures, such as TOEFL iBT section scores. Given the exploratory nature and the small sample of this study, we were not able to examine the effects of combining and averaging human ratings of multiple measures on rater consistency or the relationships between human ratings and *e*-rater scores. Finally, like

most previous studies on the relationship between AES scores and human ratings, this study was correlational. As noted above, correlation studies may not be sufficient for validating AES scoring because they tell us little about what is measured by automated scores (Attali, 2007, p. 2).

Future studies will need to include larger samples of tasks, test-takers, essays, and occasions and to go beyond correlational analyses to examine whether and to what extent AES scores and their relationship with human ratings vary across tasks, test-taker groups, and test occasions. For example, L2 learners from different L1 backgrounds and in different learning contexts (e.g., ESL vs. EFL) may exhibit different profiles, patterns, and rates of L2 writing development over time and/or after L2 instruction. Currently, there is little to no research on the sensitivity of automated scoring to changes in writing ability over time and variability in writing performance and development across test-takers from different backgrounds and contexts. There is a growing interest in using automated scores to assess L2 writing development because AWS scoring is more efficient, reliable, and cost-effective than human ratings. To support such use, evidence is needed that automated scores are sensitive, or at least as sensitive as human ratings, to true changes in writing performance over time and/or after L2 instruction. Research is also needed that examines the sensitivity of automated analytic scores to changes in writing performance and how they compare to human ratings in terms of their sensitivity to differences across tasks and changes over time.

Lastly, one critique of AES scoring that highlights the close link between the explanation and utilization inferences is that, because they do not attend to the social and communicative dimensions of writing, AES systems decontextualize the writing activity (Conference on College Composition and Communication (CCCC), 2014; Herrington & Moran, 2001) which can have a negative effect on the teaching and learning of writing (Blood, 2011). This can undermine the validity of the utilization inference of writing assessments using AES systems. The utilization inference is based on the warrant that test results are useful for making fair decisions about learners and positively impact learning and instruction (Chapelle et al., 2008; Knoch & Chapelle, 2018; Williamson et al., 2012). However, the use of AES scoring may affect teachers' and students' views of, and

approaches to, writing in negative ways. For example, machine scoring can lead students and instructors to only attend to features that are most likely to affect students' scores, such as essay length and lexical sophistication (Blood, 2011; Herrington & Moran, 2001). Given the potential negative effects of the use of AES scoring, Deane (2013b) has argued for avoiding the two extreme positions of the unrestricted use of AES to replace human rating and the complete avoidance of automated methods, calling for combining both approaches instead. It is expected that such an approach can assuage some of the potential negative effects of the use of AES scoring alone. However, there is little to no research on the impact of using AES scoring on learning and teaching L2 writing, particularly on learner L2 writing beliefs, processes, motivation, engagement, and achievement. Such research can take the form of case studies to provide much needed information on the impact of the use of AES scoring on the teaching, learning, and development of L2 writing in specific learning and assessment contexts.

Acknowledgements

We would like to thank the participating raters and to acknowledge that this research was funded by Educational Testing Services (ETS) under a TOEFL Committee of Examiners research grant. ETS does not discount or endorse the methodology, results, implications, or opinions presented by the authors.

References

Attali, Y. (2007). Construct validity of E-rater® in scoring TOEFL® essays. *ETS Research Report Series*, *2007*, i-22.

Attali, Y. (2013). Validity and reliability of automated essay scoring. In M. D. Shermis & J. Burstein (Eds.), *Handbook of automated essay evaluation: Current applications and new directions* (pp. 181-198). Routledge.

Attali, Y., Bridgeman, B., & Trapani, C. (2010). Performance of a generic approach in automated essay scoring. The Journal of Technology, Learning and Assessment, 10(3), 1-17.

Attali, Y., & Burstein, J. (2006). Automated essay scoring with E-Rater® V.2.0. *ETS Research Report Series*, *2004(2)*, i-21.

Attali, Y., & Powers, D. (2008). A developmental writing scale. *ETS Research Report Series*, *2008(1)*, i-59.

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.

Barkaoui, K. (2011). Effects of marking method and rater experience on ESL essay scores and rater performance. *Assessment in Education: Principles, Policy & Practice*, *18(3)*, 279-293.

Barkaoui, K. & Hadidi, A. (2021). *Assessing changes in second Language writing performance*. Routledge.

Ben-Simon, A., & Bennett, R. E. (2007). Toward more substantively meaningful automated essay scoring. *The Journal of Technology, Learning, and Assessment*, *6*, 1-47.

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.

Blood, I. (2011). Automated essay scoring: A literature review. *Studies in Applied Linguistics and TESOL*, *11(2)*, 40-64.

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.

Buzick, H., Oliveri, M. E., Attali, Y., & Flor, M. (2016). Comparing human and automated essay scoring for prospective graduate students with learning disabilities and/or ADHD. *Applied Measurement in Education*, *29(3)*, 161-172.

Chan, S., Inoue, C., & Taylor, L. (2015). Developing rubrics to assess the reading-into-writing skills: A case study. *Assessing Writing*, *26*, 20-37.

Chapelle, C. A., Enright, M. K., & Jamieson, J. (2008). *Building a validity argument for the test of English as a foreign language*. Routledge.

Chen, C. F. E., & Cheng, W. Y. E. C. (2008). Beyond the design of automated writing evaluation: Pedagogical practices and perceived learning effectiveness in EFL writing classes. *Language Learning & Technology*, *12(2)*, 94-112.

Chiang, S. (1999). Assessing grammatical and textual features in L2 writing samples: The case of French as a foreign language. *The Modern Language Journal*, *83(2)*, 219-232.

Chiang, S. (2003). The importance of cohesive conditions to perceptions of writing quality at the early stages of foreign language learning. *System*, *31(4)*, 471-484.

Chodorow, M., & Burstein, J. (2004). Beyond essay length: Evaluating e-rater®'s performance on TOEFL essays. *ETS Research Report Series*, *2004(1)*, i-38.

Clauser, B. E., Kane, M. T., & Swanson, D. B. (2002). Validity issues for performance-based tests scored with computer-automated scoring systems. *Applied Measurement in Education*, *15(4),* 413-432.

Coniam, D. (2009). Experimenting with a computer essay-scoring program based on ESL student writing scripts. *ReCALL: The Journal of EUROCALL*, *21(2)*, 259-279.

Conference on College Composition and Communication (CCCC). (2014). *A position statement on writing assessment*. Retrieved 2020-08-04 from https://cccc.ncte.org/cccc/resources/positions/writingassessment

Connor, U., & Mbaye, A. (2002). Discourse approaches to writing assessment. *Annual Review of Applied Linguistics*, *22*, 263-278.

Crossley, S. A., & McNamara, D. S. (2016). Say more and be more coherent: How text elaboration and cohesion can increase writing quality. *Journal of Writing Research*, *7(3)*, 351-370.

Cumming, A., Kantor, R., & Powers, D. E. (2002). Decision making while rating ESL/EFL writing tasks: A descriptive framework. *The Modern Language Journal*, *86(1)*, 67-96.

Cumming, A., Kantor, R., Baba, K., Erdosy, U., Eouanzoui, K., & James, M. (2005). Differences in written discourse in independent and integrated prototype tasks for next generation TOEFL. *Assessing Writing*, *10(1)*, 5-43.

Deane, P. (2013a). Covering the construct: An approach to automated essay scoring motivated by a socio-cognitive framework for defining literacy. In M. D. Shermis & J. Burstein (Eds.), *Handbook of Automated Essay Evaluation: Current Applications and New Directions* (pp. 298-312). New York: Routledge.

Deane, P. (2013b). On the relation between automated essay scoring and modern views of the writing construct. *Assessing Writing*, *18(1)*, 7-24.

Deane, P., & Quinlan, T. (2010). What automated analyses of corpora can tell us about students' writing skills. *Journal of Writing Research*, *2(2)*, 151-177.

Di Gennaro, K. (2009). Investigating differences in the writing performance of international and Generation 1.5 students. *Language Testing*, *26(4)*, 533-559.

Enright, M. K., & Quinlan, T. (2010). Complementing human judgment of essays written by English language learners with e-rater® scoring. *Language Testing*, *27(3)*, 317-334.

Field, A. (2009). *Discovering statistics using IBM SPSS Statistics,* Sage Publications Limited.

Hamp-Lyons, L. (1991). Scoring procedures for ESL contexts. In L. Hamp-Lyons (Ed.), *Assessing second language writing in academic contexts* (pp. 241-276). Norwood, NJ: Ablex.

Herrington, A., & Moran, C. (2001). What happens when machines read our students' writing. *College English*, *63(4)*, 480-499.

Higgins, D., Burstein, J., & Attali, Y. (2006). Identifying off-topic student essays without topic-specific training data. *Natural Language Engineering*, *12*(2), 145-159.

Huang, S. J. (2014). Automated versus human scoring: A case study in an EFL context. *Electronic Journal of Foreign Language Teaching*, 11, 149-164.

Jones, E. (2006). Accuplacer's essay scoring technology. In P. F. Ericsson & R. Haswell (Eds.), *Machine scoring of student essays* (pp. 93-113). Logan, UT: Utah State University Press.

Kelly, P. A. (2005). General models for automated essay scoring: Exploring an alternative to the status quo. *Journal of Educational Computing Research*, *33(1)*, 101-113.

Knoch, U., & Chapelle, C. A. (2018). Validation of rating processes within an argument-based framework. *Language Testing*, *35(4)*, 477-499.

Kubota, R. (1998). An investigation of Japanese and English L1 essay organization: Differences and similarities. *Canadian modern language review*, 54(4), 475-508.

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

Lee, I. A., & Preacher, K. J. (2013, October). Calculation for the test of the difference between two dependent correlations with no variable in common [Computer software]. Available from http://quantpsy.org

Lee, Y. W. (2016). Investigating the feasibility of generic scoring models of e-rater for TOEFL iBT independent writing tasks. 영어교육연구/*English Language Teaching*, *28*, 101-122.

Ling, G., Powers, D. E., & Adler, R. M. (2014). Do TOEFL iBT® scores reflect improvement in English-language proficiency? Extending the TOEFL iBT validity argument. *ETS Research Report Series*, *2014(1)*, 1-16.

Lumley, T. (2005). *Assessing Second Language Writing*. Peter Lang Pub Incorporated.

McGee, T. (2006). Taking a spin on the Intelligent Essay Assessor. In P. F. Ericsson & R. Haswell (Eds.), *Machine scoring of student essays: Truth and consequences* (pp. 79-92). Logan, UT: Utah State University Press.

Mikulas, C., & Kern, K. (2006). *A comparison of the accuracy of automated essay scoring using prompt-specific and prompt-independent training*. In Annual

Meeting of the American Educational Research Association (AERA), San Francisco, CA, San Francisco, CA., (pp. 1-10).

Nichols, P. (2005). Evidence for the interpretation and use of scores from an automated essay scorer. *PEM Research Report 05*,

Perelman, L. (2012). Construct validity, length, score, and time in holistically graded writing assessments: The case against automated essay scoring (AES). In C. Bazerman, C. Dean, J. Early, K. Lunsford, S. Null, P. Rogers, & A. Stansell (Eds.), *International advances in writing research: Cultures, places, measures* (pp. 121-131). Fort Collins, Colorado/Anderson, SC: WAC Clearinghouse/Parlor Press.

Powers, D. E., Burstein, J. C., Chodorow, M., Fowles, M. E., & Kukich, K. (2000). Comparing the validity of automated and human essay scoring. *ETS Research Report Series*, *2000(2)*, i-23.

Powers, D. E., Burstein, J. C., Chodorow, M., Fowles, M. E., & Kukich, K. (2002). Stumping e-rater: challenging the validity of automated essay scoring. *Computers in Human Behavior*, *18(2)*, 103-134.

Powers, D. E., Escoffery, D. S., & Duchnowski, M. P. (2015). Validating automated essay scoring: A (modest) refinement of the "gold standard". *Applied Measurement in Education*, *28(2),* 130-142.

Preacher, K. J. (2002, May). Calculation for the test of the difference between two independent correlation coefficients [Computer software]. Available from http://quantpsy.org.

Quinlan, T., Higgins, D., & Wolff, S. (2009). Evaluating the construct-coverage of the e-rater® scoring engine. *ETS Research Report Series*, *2009(1),* i-35.

Ramineni, C., & Williamson, D. (2018). Understanding mean score differences between the E-rater® automated scoring engine and humans for demographically based groups in the GRE® general test. *ETS Research Report Series*, *2018(1),* 1-31.

Ramineni, C., Trapani, C. S., Williamson, D. M., Davey, T., & Bridgeman, B. (2012a). Evaluation of the e-rater® Scoring Engine for the TOEFL® Independent and Integrated Prompts. *ETS Research Report Series*, *2012(1),* i-51.

Ramineni, C., Trapani, C. S., Williamson, D. M., Davey, T., & Bridgeman, B. (2012b). Evaluation of e-rater for the GRE issue and argument prompts. *ETS Research Report Series*, *12(02)*, 1-106.

Rudner, L. M., Garcia, V., & Welch, C. (2006). An evaluation of IntelliMetric™ essay scoring system. *The Journal of Technology, Learning and Assessment*, *4(4)*, i-22.

Shermis, M. D. (2014). State-of-the-art automated essay scoring: Competition, results, and future directions from a United States demonstration. *Assessing Writing*, *20*, 53-76.

Shermis, M. D., & Burstein, J. (Eds.). (2003). *Automated essay scoring: A cross-disciplinary perspective*. New Jersey: Mahwah.

Shermis, M. D., & Hamner, B. (2013). Contrasting state-of-the-art automated scoring of essays. In M. D. Shermis & J. Burstein (Eds.), *Handbook of automated essay evaluation: Current applications and new directions* (pp. 313-346). Routledge.

Shermis, M. D., Burstein, J., Higgins, D., & Zechner, K. (2010). Automated essay scoring: Writing assessment and instruction. *International Encyclopedia of Education*, *4(1),* 20-26.

Shermis, M. D., Koch, C. M., Page, E. B., Keith, T. Z., & Harrington, S. (2002). Trait ratings for automated essay grading. *Educational and Psychological Measurement*, *62(1)*, 5-18.

Shermis, M. D., Mzumara, H. R., Olson, J., & Harrington, S. (2001). On-line grading of student essays: PEG goes on the World Wide Web. *Assessment & Evaluation in Higher Education*, *26(3)*, 247-259.

Shermis, M. D., Shneyderman, A., & Attali, Y. (2008). How important is content in the ratings of essay assessments? *Assessment in Education: Principles, Policy & Practice*, *15(1)*, 91-105.

Wang, J., & Brown, M. S. (2007). Automated essay scoring versus human scoring: A comparative study. *Journal of Technology, Learning, and Assessment*, *6(2)*, 1-29.

Wang, J., & Brown, M. S. (2008). Automated essay scoring versus human scoring: A correlational study. *Contemporary Issues in Technology and Teacher Education*, 8(4), 310-325.

Weigle, S. C. (2002). *Assessing writing*. Cambridge University Press.

Weigle, S. C. (2010). Validation of automated scores of TOEFL iBT tasks against non-test indicators of writing ability. *Language Testing*, *27(3),* 335-353.

Weigle, S. C. (2013a). English as a second language writing and automated essay evaluation. In M. D. Shermis & J. Burstein (Eds.), *Handbook of automated essay evaluation: Current applications and new directions* (pp. 58-76). Routledge.

Weigle, S. C. (2013b). English language learners and automated scoring of essays: Critical considerations. *Assessing Writing*, *18(1),* 85-99.

Williams, R., & Dreher, H. (2005). Formative assessment visual feedback in computer graded essays. *Journal of Issues in Informing Science and Information Technology*, *2*, 23-32.

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

Wilson, J., Olinghouse, N. G., McCoach, D. B., Santangelo, T., & Andrada, G. N. (2016). Comparing the accuracy of different scoring methods for identifying sixth graders at risk of failing a state writing assessment. *Assessing Writing*, 27, 11-23.

Woodworth, J. & Barkaoui, K. (2020). Perspectives on using automated writing evaluation systems to provide written corrective feedback in the ESL classroom. *TESL Canada Journal, 37*(2), 234-247.

Zhao, C. G. (2017). Voice in timed L2 argumentative essay writing. *Assessing writing*, *31,* 73-83.

Appendix A: Studies comparing AES scores and human ratings

Study	***Sample***		***AES****			***Human Ratings****		***Main Analyses***		
	N essays*	***N prompts***	***System***	***Scoring***	***G or S***	***Rating***	***Raters per essay***	***Correlation***	***Agreement indices***	***Other***
Powers et al. (2000)	About 6,000	40	e-rater	O	G & S	H	2	x		
Shermis et al. (2001)	617	2	PEG	O	NR	H	6	x		
Powers et al. (2002)	63	4	e-rater	O	S	H	2	x	x	
Shermis et al. (2002)	386	2	PEG	O & A	NR	H & MT	6	x		
Landauer et al. (2003)	3,2962	6	IEA	O & A	S	H	2	x		
Chodorow & Burstein (2004)	About 6,000 from 3 L1 groups.	7	e-rater01, e-rater99	O	S	H	2	x	x	
Kelly (2005)	598	6	e-rater	O	G & S	H	2	x	x	
Nichols (2005)	3,244	5	IEA	O	S	H	2 experts and 2 Readers	x	x	
Williams & Dreher (2005)	20	NR	MarkIt	O	NR	H	1	x		
Attali & Burstein (2006)	More than 25,000	64	e-rater v.2	O & A	Different G & S models	H	2	x	x	
Mikulas & Kern (2006)	100	11	IntelliMetric	O	G & S	H	2	x	x	
Rudner et al. (2006)	10,100	101	IntelliMetric	O	S	H	Randomly selected 1 score from 3 raters	x	x	
Attali (2007)	10,012	2	e-rater v.2	O	optimal and equal weights	H	2	x		Regression
Ben-Simon & Bennett (2007)	2510	2	e-rater	O	4 different scoring approaches	H	2	x		ANOVA
Wang & Brown (2007)	107	NR	IntelliMetric	O	NR	H	2	x		ANOVA
Attali & Powers (2008)[a]	1810	2	e-rater v.2	O	NR	H	2			ANOVA, Factor analysis

Study	Sample		AES			Human Ratings		Main Analyses		
	N essays	N prompts	System	Scoring	G or S	Rating	Raters per essay	Correlation	Agreement indices	Other
Wang & Brown (2008)	214	NR	IntelliMetric	O & A	NR	H & MT	2 groups of 2 (faculty and NES)	x	x	
Attali & Powers (2009)	1,810	2	e-rater v.2	O	NR	H	2			ANOVA
Coniam (2009)	330	3	BETSY	O	S	H	2	x		
Weigle (2010)	772	2	e-rater	O & A	S (TOEFL Topics)	H	2	x		T-tests, ANOVA
Bridgeman et al. (2012)[b]	132,347	38	e-rater	O	S (TOEFL Topics)	H	2	x		
Bridgeman et al. (2012)[b]	630,000	252	e-rater	O	S (GRE Topics)	H	2	x		
Ramineni et al. (2012b)	750,000	252	e-rater v7.2	O	G & S	H	2	x	x	
Ramineni et al. (2012a)	152,000	76	e-rater v8.1	O	G & S	H	2	x	x	
Huang (2014)	103	4	e-rater	O	NR	H	2	x		t-tests
Shermis, (2014)	4,343	8	9 AES systems	O	S	H	Various[c]	x	x	
Buzick et al. (2016)	7,788 with a reference group of 445,000	2	e-rater	O	S (GRE Topics)	H	1-3	x		Mean comparisons
Lee (2016)	598	2	e-rater	O	3 G, 1 S models, & 3 hybrid models	H	2	x	x	
Wilson et al. (2016)	272	NR	PEG	O	S	H	2	x		Regression
Ramineni & Williamson (2018)	215,000	215	e-rater v10.1	O & A	S (GRE Topics)	H	2	x	x	Regression

*N essays refers to the number of essays scored by AES and human raters; it does not include the number of essays used for training the AES system if they are reported.

** AES systems provide overall (O) or analytic (A) scores and use generic (G) or topic-specific (S) models. Human ratings are either holistic (H) or multiple-trait (MT). NR indicates that information was not reported for the modelling set or is not clear from the description

a Only from the human scoring experiment section of the study

b The study reported two different studies: one for TOEFL and another for GRE.

c Different data sets used different number of raters.

Appendix B: Descriptive statistics

Task	*Integrated*				*Independent*			
Occasion	*1*		*2*		*1*		*2*	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
TOEFL iBT Scores								
Total	49.92	22.50	71.35	23.31	49.92	22.50	71.35	23.31
Reading	11.38	8.19	19.96	8.16	11.38	8.19	19.96	8.16
Listening	8.90	7.80	16.40	8.97	8.90	7.80	16.40	8.97
e-rater score	2.25	1.19	2.67	1.26	2.46	0.99	3.15	1.03
Number of Words	160.29	67.68	186.25	53.09	237.08	86.98	295.23	84.50
Human Ratings								
Accuracy	3.04	0.78	3.08	0.66	3.07	0.65	3.10	0.59
Coherence	2.72	0.47	3.15	0.40	2.77	0.47	3.25	0.43
Cohesion	3.40	0.39	3.66	0.39	3.32	0.38	3.59	0.32
Organization	2.73	0.62	3.24	0.43	2.83	0.58	3.30	0.41
Register	3.07	0.68	3.19	0.53	1.75	0.51	2.17	0.65
Metadiscourse								
Voice	1.74	0.64	1.66	0.54	2.75	0.67	3.18	0.70
Textual	2.58	0.79	3.01	0.71	2.04	0.76	2.91	0.70
Argument Quality					2.45	0.54	2.77	0.45
Source Use	1.53	0.76	2.03	0.74				