Reliability and Validity of Automated Essay Scores for Independent Writing Tasks: Generic, Hybrid, and Prompt-Specific Models

Lee, Yong-Won

[Abstract]

The current study aims to examine the reliability and validity of automated essay scores from substantively different types of scoring models for e-rater® in the context of scoring TOEFL independent writing tasks. Six different variants of generic and hybrid models were created based on transformed writing data from three different samples of TOEFL® CBT prompts. These generic models (along with prompt-specific models) were used to score a total of 61,089 essays written for seven TOEFL® CBT prompts. The results of data analysis showed that (a) similar levels of score agreement were achieved between automated and human scorer pairs and two human rater pairs, although the automated scoring increased rating consistency across scorers (or scoring models) significantly, and (b) the human rater scores turned out to be somewhat better indicators of test-takers' overall ESL language proficiency than the automated scores, particularly when TOEFL CBT section scores were used as validity criteria. The implications of the findings are discussed in relation to the future use of automated essay scoring in the context of scoring ESL/EFL learners' essays.

Key Words: Automated essay Scoring, E-rater®, Generic scoring Models, Independent writing tasks, ESL (English as a second language) writing assessment.

1. Introduction

In recent years, automated essay scoring (AES) is drawing a great deal of attention in the fields of writing assessment and instruction (Enright & Quinlan 2010; Lee, Gentile, & Kantor 2010; Weigle 2010). Given its consistency and objectivity in scoring and efficiency in cost and time, it is an attractive option for scoring performance-based writing assessment. Another positive impact that is expected of well-designed AES systems is that they can potentially enable a wider use of constructed or extended response tasks in language assessment (Lee et al. 2010). In conjunction with the wide use of AES in writing assessment, another important development is that it is promoted as a means of providing diagnostic (or formative) feedback to learners in the context of learning-oriented second language assessment. Nevertheless, there seems to be only a small body of validation studies evaluating the psychometric quality of the scores computed by AES when it is used in the contexts of EFL/ESL (English as a foreign/second language) writing assessment. Given the current situations, it seems urgently necessary to examine the psychometric quality of automated essay scores produced by the major AES systems when they are used in the EFL/ESL writing contexts, so that teachers, test-takers, and other major stakeholders of writing assessment can make informed judgments and decisions about the future use of AES in writing assessment and instruction.

One particular AES system of interest in this study is a relatively recent version

of e-rater® (version 2.1) that was developed at Educational Testing Service® (ETS®) and has been used to score essays for various large-scale writing assessments (Atalli & Burstein, 2006; Burstein, 2003; Kukich, 2000). One preferred way of adopting e-rater for scoring essays for large-scale writing assessment is to use a single human rating per writing sample and use an automated essay scorer (such as e-rater) as a second reader or as a detection mechanism to flag an unusual rating for human adjudication (Enright & Quinlan 2010; Lee & Kantor 2005). These options are attractive for a testing program because they would be more economical and feasible. As matter of fact, there are a great number of other AES systems currently being used for scoring essays in first and second language writing assessment (for interested readers, see Dikli 2006; Kukich 2000, Rudner & Gagne 2001, and Shermis & Burstein 2003). The psychometric quality of automated scores produced by e-rater has been evaluated primarily by examining the empirical characteristics of automated scores against a criterion of "human rater-assigned scores" (eg, the "rate of score agreement" and "correlation" between human raters and e-rater) (Bennett & Bejar, 1998). It has been shown that an AES system could achieve high score agreement with human raters in many different contexts of large-scale writing assessments including the writing section of the computer-based and internet-based Test of English as a Foreign language™ (TOEFL® CBT and TOEFL® *i*BT, respectively, hence forth) (Chodorow & Burstein 2004; Enright & Quinlan 2010; Powers, Burstein, Chodorow, Fowles & Kukich 2000).

One important issue of interest being explored in the current investigation is the feasibility of using non-prompt-specific (or generic) scoring models of e-rater, in place of prompt-specific models, to score independent writing tasks that are used in TOEFL® family tests, including TOEFL® CBT, TOEFL® *i*BT, and TOEIC® Writing. Until recently, prompt-specific models of AES have been used most widely to score

test-takers' essays. One disadvantage of using a prompt-specific model for a writing task, however, is that the AES model for a particular writing prompt is built and cross-validated based on samples of essays written for that particular prompt (Lee & Bridgeman, 2004). This means that the model building and validation should be done in a very speedy way within a short turn-around time (about 2-3 weeks) allowing for data processing and score reporting for operational testing programs. For this reason, it is usually very challenging to get the model-building process completed and put the validated model in place for essays to be scored in an operational setting in a timely manner. This explains why a prompt-neutral, generic model (built without scored essays for the particular prompt to be scored) is an appealing option for large-scale writing assessment programs, such as TOEFL. In order to justify the use of the generic models for independent writing tasks, though, there should be more accumulated research evidence showing that these generic models are as accurate and valid as the prompt-specific models and human ratings are in the particular context where these generic scoring models are planned to be used.

Another noteworthy aspect of the current study is that the independent writing tasks are the focus of investigation, rather than integrative writing tasks. In fact, TOEFL® *i*BT is made up of four language skill sections (including listening, reading, speaking, and writing), and the writing section of the TOEFL *i*BT consists of two writing tasks (one independent and one integrated writing task); this contrasts with the writing subsection of the TOEFL CBT that has only one independent writing task. However, when it comes to the issue of using e-rater for essay scoring, integrated writing tasks currently pose a greater challenge for automated scoring in general, compared to independent tasks (Lee & Bridgeman, 2004). The scoring of integrated tasks requires the evaluation of content accuracy with respect to the stimulus material (reading, listening texts) associated with specific prompts. This

makes it much more difficult to create a generic scoring model for the integrated tasks due to the "prompt-specific" nature of the content-accuracy scoring. Aside from this, the version of e-rater used in this study does not have automated essay feature variables yet that directly assess this "content accuracy" aspect of essay quality. In contrast, the independent writing tasks are very similar to the TOEFL CBT writing prompts, except for the fact that a 5-point (instead of a 6-point) rating scale is used for TOEFL *i*BT. By design, the current 5-point rating rubric for the independent writing tasks for TOEFL *i*BT was created by collapsing the two lowest score categories of the TOEFL CBT rubric (i.e., 1 and 2) into a single category and combining the verbal descriptors of these two score categories (Lee & Kantor, 2005).

In evaluating the generic scoring models for the independent tasks for TOEFL, it was relatively easier to obtain a large sample of pre-scored TOEFL CBT essays, but it was not possible to obtain pre-scored TOEFL *i*BT essays from a large number of prompts taken by fully representative samples of TOEFL test-takers worldwide that were stable and equivalent across test administrations (Lee & Bridgeman, 2004). Such stable samples of data are required for building and evaluating generic models of e-rater for operational TOEFL *i*BT. One practical issue was that such representative samples could not be obtained until a significant number of years had passed after the initial roll-out of TOEFL *i*BT only to a few selected regions of the world. Considering (a) a close connection and similarity between the TOEFL CBT and TOEFL *i*BT independent writing tasks and (b) the full availability of stable pre-scored essay data accumulated over many years for TOEFL CBT, a decision was made to build generic scoring models based on the transformed TOEFL CBT data and evaluate these models on the essays written for the seven TOEFL CBT prompts first in Phase 1 of the study, the results of which are reported in this paper. The same sets of generic models were also evaluated on the two independent writing

tasks from a TOEFL *i*BT field study in the Phase 2 study, but the results of this Phase 2 study is reported and discussed in a separate paper (Lee, in press).

In addition, it is also worth mentioning that some fundamental psychometric issues about AES-produced scores are examined in the current study (and other related studies) that may go beyond a particular writing assessment context of TOEFL *i*BT. In conjunction with this, one major theoretical concern raised by measurement experts about AES is the "variability of automated scores" (i.e., variability of scores attributable to use of different scoring algorithms of AES for the same tasks) in the context of writing assessment. It is often argued that automated scoring can lead to a reduction in error ascribable to human raters (and also a subsequent increase in reliability), since a single scoring algorithm is consistently used to rate all of the examinee responses to a particular prompt or set of prompts. Nonetheless, some measurement specialists point out that this may not be necessarily true in every testing situation (Bennett, 2004; Brennan, 2000). Because some form of human judgments is required, or used as a basis, for determining a scoring algorithm, there can possibly be multiple, equally viable, mutually complementing (or competing) scoring algorithms for the same writing prompt depending upon what types of expert judgments are adopted or emulated. In this sense, it would be very important to carefully examine the variability of automated scores obtained from substantively different scoring algorithms (or models) from both reliability and validity perspectives.

With these considerations in mind, an attempt was made in this study to create several substantively different scoring algorithms including generic, hybrid, and prompt-specific models and investigate not only their impact on the variability of automated essay scores but also the criterion-related validity of the automated scores produced by these scoring models. More specifically, the current study attempts to

(a) create six different versions of generic scoring models for e-rater using the transformed rating data from 40 TOEFL CBT prompts and (b) investigates the variability and validity of the automated scores computed by these generic models, as compared to the scores obtained from the prompt-specific models and human raters, when these models are used to score essays written for seven TOEFL CBT prompts.

2. Automated Scoring and Model-Building

2.1 Approaches to Model Building in AES

As mentioned previously, prompt-specific models of e-rater have been used to score test-takers' essays for operational testing programs (Chodorow & Burstein, 2004; Powers, Burstein, Chodorow, Fowles & Kukich, 2000). In recent years, there seems to be an increasing demand for non-prompt-specific scoring models for AES. In the context of e-rater-based automated scoring, approaches to building AES models can largely be classified into three major types: (a) prompt-specific; (b) generic; and (c) hybrid (or partially generic) models.

First, in the prompt-specific model, the scoring model for a particular writing task is built and cross-validated based on samples of essays written for that particular prompt to be scored. Due to the short time span allowed for data processing and score reporting, however, it would be very challenging to build a separate scoring model for each prompt in a timely manner in large-scale testing situations.

Second, the generic scoring models require that the same scoring coefficients for all or most of the essay features be applied across multiple prompts. Two different kinds of generic models are considered in this study: (a) a fully generic model with

two prompt-specific vocabulary usage variables dropped from the feature set; and (2) a partially generic (or hybrid) model with the two prompt-specific vocabulary usage variables retained in the feature set. In the fully generic models, some of the essay feature variables that are prompt-specific by definition are dropped from the essay feature set used for model-building and scoring. In the e-rater (v. 2.1), for instance, there are two content-word-vector-based variables that tap into prompt-specific usage of prompt-topic-related vocabulary in test-takers' essays and need to be computed for each prompt to be scored (Attali & Burstein 2006). These two variables were dropped from the essay feature set used for the fully generic models in this study. When these two variables are dropped from the scoring model, however, the scoring models lose some useful information on test-takers'ability to use (or performance in using) important content words related to a specific prompt-topic.

By contrast, these two variables are retained in the feature set for a partially generic (or hybrid) model. In the hybrid models, these two variables are computed based on small samples of essays for a particular prompt to be rated and used for model-building and cross-validation purposes, while the remaining 10 essay-feature variables can use the generic scoring coefficients across prompts to be scored. In this sense, the hybrid models can be viewed as partially-prompt specific and, at the same time, partially generic models.

In addition to the degree of "generic-ness" of the model (ie, fully generic, partially generic, or non-generic), another important factor is introduced in this study that can potentially contribute to the variability of automated scores obtained from different generic models, that is, sampling of prompts to be used in model-building. In the case of the generic-model-based scoring, a group of writing prompts that are used to build the generic scoring model do not usually include an operational writing prompt to be scored by the generic model. Since the set of prompts to be used for

model-building can be sampled from a larger pool of prompts in many ways, it is critical to examine the variability of automated scores due to sampling of prompts for model-building. Rather than focusing only on the number of sampled prompts used for the model building (which may be considered a trivial variation among models), an attempt was made in this study to investigate the impact of a more meaningful aspect of prompt sampling related to types of argument required in the examinees' essays.

Table 1. Schematic Representation of Generic, Hybrid, and Prompt-Specific Scoring Models of E-rater

Model Type	Prompt Type for Model-Building			Total
	Issue	Non-Issue	Issue+non-issue	
Generic	1	1	1	3
Hybrid	1	1	1	3
Subtotal	2	2	2	6
Prompt-Specific	NA	NA	NA	1
	(The same prompt for model-building and scoring)			
Total				7

Table 1 presents a schematic representation of the seven AES models used in this study, which represent the three major approaches to AES and their subtypes associated with different prompt samples used in model-building. In this study, a total of 40 TOEFL CBT prompts were classified by a writing expert into two different types (Issue/Non-issue) of prompts based on prompt content analysis. This was done to maximize the dissimilarity among samples of prompts to be used in the model-building for the generic models.

2.2. Research Questions

Since generic models are built without scored essays for a particular prompt to be scored, such models are a very attractive option for large-scale writing assessments with fast score reporting requirements, such as TOEFL *i*BT. Nevertheless, it is yet to be demonstrated that the generic model is as accurate and valid as the prompt-specific model and human raters in the writing assessment context in which e-rater will be used.

The current program of research was carried out with the following four research questions in mind:

1. How well do the generic models perform when compared to the prompt-specific models and human rater scores in scoring independent writing tasks for TOEFL CBT in terms of score agreement?
2. To what extent do the automated scores from the three different scoring approaches (prompt-specific, generic, and hybrid) vary among themselves in the context of scoring independent writing tasks for TOEFL CBT?
3. What would be the impact of using very different types and numbers of writing prompts (e.g., Issue, Non-issue, and Combined prompts) in the generic and hybrid models on the variability of automated scores computed for independent writing tasks for TOELF CBT?
4. How do the automated scores compare to the human rater scores in terms of their relationships to other criterion measures of ESL language proficiencies and writing abilities?

3. Method

3.1. Data

The main data sets analyzed in this study were: (a) 2 essay scores from human raters, (b) 1 automated essay score from a prompt-specific scoring model, (c) 3 automated essay scores from generic scoring models, and (4) 3 automated essay scores from hybrid scoring models of e-rater® obtained for each of the 61,089 essays written by ESL/EFL learners for seven TOEFL® CBT prompts. Twelve essay feature variables used in e-rater were also obtained for these essays and analyzed in terms of their relationships to the automated and human rater scores. In addition, the scale scores for four sections of TOEFL CBT were used in this study along with the total score (or a weighted composite of the four section scores). These additional scores were used as criterion measures for evaluating the criterion-related validity of the automated and human rater scores.

Table 2. Seven TOEFL CBT Writing Prompts Used for Evaluation

Pr. No.	Prompt	Sample Size
P1	"When people succeed, it is because of hard work. Luck has nothing to do with success." Do you agree or disagree with the quotation above? Use specific reasons and examples to explain your position.	9,676
P2	Many people visit museums when they travel to new places. Why do you think people visit museum? Use specific reasons and examples to support your answer.	9,186
P3	You have been told that dormitory rooms at your university must be shared by two students. Would you rather have the university assign a student to share a room with you, or would you rather choose your own roommate? Use specific reasons and details to	5,651

	explain your answer.	
P4	Do you agree or disagree with the following statement? It is more important for students to study history and literature than it is for them to study science and mathematics. Use specific reasons and examples to support your opinion.	8,578
P5	Some young children spend a great amount of their time practicing sports. Discuss the advantages and disadvantages of this. Use specific reasons and examples to support your answer.	9,791
P6	Do you agree or disagree with the following statement? A persons' childhood years (the time from birth to twelve years of age) are the most important years of a person's life. Use specific reasons and examples to support you answer.	9,702
P7	Do you agree or disagree with the following statement? Children should be required to help with household tasks as soon as they are able to do so.	8,505

A total of seven TOEFL CBT writing prompts were used as evaluation prompts for the scoring models. Table 2 shows the prompt identification number, prompt content, and the total number of essays scored by e-rater for each of these seven TOEFL CBT prompts. Each of the student responses for the TOEFL CBT prompts was double-rated by a pair of trained raters on a 6-point scale. When there was a score discrepancy greater than 1 between the initial two raters, the third rater adjudicated the discrepancy. Before these seven prompts were used for evaluation purposes in this study, the two lowest score points of 1 and 2 were collapsed into a single score point of 1, and the remaining score points were lowered by 1 point to create the 5-point scale (1-5). This was done because this was intended to simulate the independent writing tasks that are used in the Writing Section of TOEFL *i*BT.

3.2. Model Building and Cross-Validation.

A total of six different generic scoring models were built in this study based on essays written for 40 different TOEFL CBT prompts and used to score essays written for a separate set of seven evaluation prompts explained in the previous section. A prompt-specific scoring model was also created for each of the seven prompts. To examine random variability among automated scores computed from different generic scoring models of e-rater, three different samples of prompts and essays were used for both the generic and hybrid models. In general, TOEFL CBT writing prompts can be classified into two major types: Issue (or Persuasive) and Non-issue prompts. A total of 40 prompts were selected from a pool of TOEFL CBT prompts available for research and used to create and validate the generic and hybrid model. Among them, half (20 prompts) were Issue prompts, and another half, Non-issue prompts. For each of these 40 prompts, about 500 essays were randomly selected and used for the model-building process. Three different data samples for model-building were created by (a) using only 20 issue prompts, (b) only 20 non-issue prompts, and (3) all of 40 Issue and Non-issue prompts. Prior to model-building for the generic models, it was necessary to collapse the score points of 1 and 2 into a single score point of 1 and to lower the remaining score points by 1 point for those essays used in the model-building to create the 5-point score scale (1-5) that is consistent with the TOEFL *i*BT.

3.3. Data Analysis

Descriptive statistics. Descriptive statistics were computed for human and automated scores for each of the seven evaluation prompts. Performance of the generic and hybrid models was evaluated by comparing the automated scores from these models with those obtained by human raters and prompt-specific models.

Score agreement rates. For each of the seven evaluation prompts, correlations, score agreement rates, and kappa values were computed for various pairs of human and automated scores (e.g., two initial human ratings, one human rating and one e-rater score). These values were averaged across the seven prompts for TOEFL CBT. The "exact" and "exact plus adjacent" score agreement rates between the two initial ratings were used as a useful criterion against which the score agreement rates for other pairs of e-rater versus human rater scores were compared.

Criterion-Related Validity Analysis. To examine the criterion-related validity of the automated scores, Pearson product-moment correlations were computed between the automated and human rater essay scores for the seven evaluation prompts and a variety of measures of ESL language proficiencies and writing abilities available from the same or related test batteries. For the seven TOEFL CBT prompts, the scale scores for three multiple-choice sections (structure, listening, and reading) and the total scores of TOEFL CBT were used as criterion measures. The correlations were computed between these measures and the automated and human rater essay scores for a prompt being investigated. Then the correlation coefficients were averaged across the seven prompts.

4. Results

4.1. Descriptive Statistics

Means and standard deviations of human and automated scores were computed for each of the seven evaluation prompts. Comparisons of the sore means and standard deviations were made to investigate preliminarily the systematic effects of (a) the scoring mode (human versus automated), (b) types of scoring models in e-rater (prompt-specific versus hybrid versus generic), and (c) prompt samples (Issue versus Non-issue versus Combined) on test-takers' score mean and score variability among the test-takers for each prompt. Table 3 and Figure 1 show the means of two human ratings (Human1, Human2), three automated scores from the three generic scoring models (G1, G2, G3), three automated scores from the three hybrid scoring models (H1, H2, H3), and one automated score from a prompt specific scoring model (PS) for each of the seven TOEFL CBT prompts. In Table 3 and Figure 1, G1, G2, and G3 represent three generic scoring models based on Issue, Non-issue, and Combined prompts, while H1, H2, and H3 represent three hybrid models based on Issue, Non-issue, and Combined prompts, respectively.

As shown in Table 3 and Figure 1, means of two human rater scores (humans 1 and 2) not only were close but also varied in a similar pattern across the prompts. Means of all of the six automated scores from generic and hybrid models also varied together in a similar pattern across prompts, but the automated scores and human rater scores seemed to have different patterns of mean score variation across prompts. For instance, the means of the six automated scores were slightly higher than those of two human raters on Prompt 3, whereas the means of the automated scores were slightly lower than those of the human raters on Prompt 2. In addition, there were

some subtle differences observed among the six automated scores from the generic and hybrid models. Although the mean score difference among the six automated scores varied from prompt to prompt, the score mean from the first hybrid model (H1, based on Issue prompt only) tended to be the lowest across all of the seven prompts among the six automated scores.

Table 3. Means and SDs of Human and Automated Essay Scores for Seven Prompts

Mode	Prompt Number							Average
	P1	P2	P3	P4	P5	P6	P7	
Hm1	2.97 (1.05)	3.02 (1.03)	3.04 (1.01)	3.10 (1.03)	3.10 (1.00)	3.02 (1.05)	3.08 (1.04)	3.05 (1.03)
Hm2	2.96 (1.04)	3.01 (1.02)	3.03 (1.01)	3.12 (1.03)	3.10 (1.02)	3.02 (1.04)	3.10 (1.04)	3.05 (1.03)
G1	2.99 (0.92)	2.97 (0.89)	3.10 (0.92)	3.09 (0.90)	3.08 (0.92)	3.05 (0.93)	3.06 (0.93)	3.05 (0.92)
G2	3.00 (0.93)	2.98 (0.89)	3.11 (0.93)	3.11 (0.91)	3.10 (0.92)	3.07 (0.94)	3.08 (0.94)	3.06 (0.92)
G3	3.00 (0.92)	2.97 (0.88)	3.11(0.92)	3.10 (0.90)	3.09 (0.91)	3.06 (0.93)	3.07 (0.93)	3.06 (0.92)
H1	2.93 (0.91)	2.91 (0.88)	3.09 (0.92)	3.06 (0.90)	3.06 (0.93)	3.02 (0.94)	3.04 (0.94)	3.02 (0.92)
H2	2.97 (0.93)	2.95 (0.90)	3.11 (0.94	3.10 (0.92)	3.09 (0.94)	3.06 (0.95)	3.08 (0.95)	3.05 (0.93)
H3	2.96 (0.92)	2.94 (0.89)	3.11 (0.93)	3.08 (0.91)	3.08 (0.93)	3.04 (0.94)	3.06 (0.94)	3.04 (0.92)
PS	2.94 (0.92)	2.97 (0.87)	3.04 (0.87)	3.07 (0.90)	3.07 (0.89)	3.04 (0.94)	3.05 (0.95)	3.03 (0.91)

Note: The numbers in parentheses are standard deviations (SDs)

Among the seven automated scores, the score means for the prompt-specific model behaved most similarly to those of the two human rater scores across the prompts, but tended to be somewhat lower than those of the two human raters for most of the prompts. Nevertheless, it should be pointed out that the differences among the mean essay scores are very small ones, even negligible ones in a practical sense. If these mean scores are plotted on an unmagnified, 1 to 5 score scale (Y-axis), the mean differences become almost unnoticeable.

Figure 1. Means of Human and Automated Essay Scores for Seven TOEFL CBT Prompts

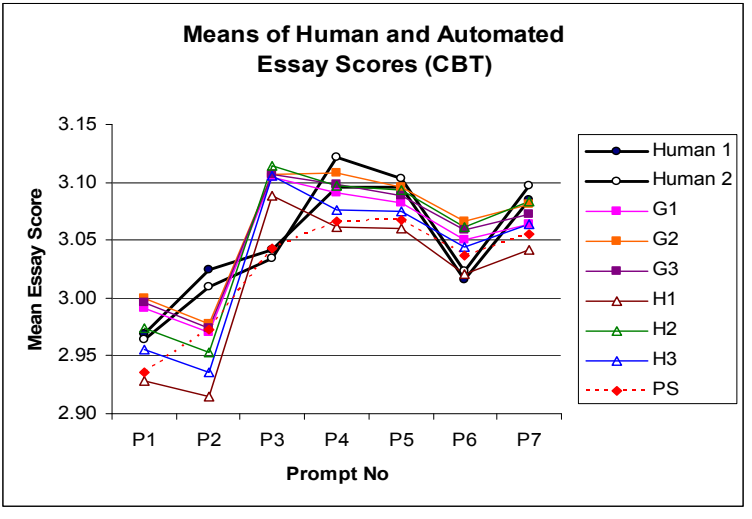

Table 3 also shows the standard deviations of two human ratings, three automated scores from the three generic scoring models, three automated scores from the three hybrid scoring models, and one automated score from a prompt specific scoring model for each of the seven TOEFL CBT prompts. One noteworthy pattern is that

the standard deviations of the automated scores tended to be consistently smaller than those of human rater scores (Human 1, Human 2) across all of the seven prompts, as shown in Figure 2. This means that automated scoring of essays resulted in slightly reduced score variations among test-takers (0.87~0.95), compared to the human ratings (1.00~1.05).

4.2 Averaged Correlations and Rates of Score Agreement

Table 4 displays averaged correlations among the human and automated scores across the seven TOEFL CBT prompts. First of all, the three automated scores from the generic models were extremely highly correlated among themselves (0.99), and so were the three automated scores from the hybrid models (0.98~0.99). Very high correlations were observed between the automated scores from the generic and hybrid models (0.95~0.97). In addition, the automated score from the prompt-specific model was highly correlated to the automated scores from both the generic and hybrid models (0.95~0.96). Nevertheless, the correlations between one of the two human ratings and seven automated scores were moderate (0.79~0.80), which were also similar to the correlation between the two human ratings (0.79). One noteworthy pattern was that the strengths of the correlations between one of the two human ratings and one of the seven automated scores were very similar across pairs. A similar pattern was observed across all of the seven TOEFL CBT prompts.

Table 4. Averaged Correlations among Human and Automated Scores Over the Seven TOEFL CBT Prompts

	Hm1	Hm2	G1	G2	G3	H1	H2	H3	PS
Hm 1	1.00								
Hm 2	0.79	1.00							
G1	0.79	0.80	1.00						
G2	0.79	0.80	0.99	1.00					
G3	0.79	0.80	0.99	0.99	1.00				
H1	0.79	0.79	0.95	0.95	0.95	1.00			
H2	0.80	0.80	0.97	0.97	0.97	0.97	1.00		
H3	0.80	0.79	0.96	0.96	0.96	0.98	0.99	1.00	
PS	0.79	0.80	0.95	0.95	0.95	0.96	0.96	0.96	1.00

Note: Fisher's Z transformation and inverse Fisher's Z were used in computing the mean correlations.

Table 5 presents averaged indices of score agreement (kappa, exact agreement, exact plus adjacent agreement) between the human rater scores and the seven automated scores across the seven TOEFL CBT prompts. Interestingly, the score agreement rates between one of the two human ratings and one of the automated scores were slightly higher than those between the two human ratings overall. For instance, the kappa value for the two human rating pairs was about 0.46, while the kappa values for one of the human ratings and the automated scores ranged from 0.48 to 0.49. Similar patterns were observed for the rates of the exact and exact plus adjacent agreements.

Table 5. Averaged Rate of Score Agreement Between Human and Automated Scores over the Seven TOEFL CBT Prompts.

Scoring Methods Paired		Kappa	Exact	Exact + Adjacent
Human 1	Human 2	0.46	0.60	0.98
	G1	0.48	0.63	0.99
	G2	0.48	0.63	0.99
	G3	0.48	0.63	0.99
Human 1	H1	0.49	0.64	0.99
	H2	0.49	0.64	0.99
	H3	0.49	0.64	0.99
	PS	0.49	0.64	0.99
	G1	0.49	0.64	0.99
	G2	0.49	0.64	0.99
	G3	0.49	0.64	0.99
Human 2	H1	0.49	0.64	0.99
	H2	0.49	0.64	0.99
	H3	0.49	0.64	0.99
	PS	0.49	0.63	0.99

4.3. Correlations Between Human and Automated Essay Scores and TOEFL Section Scores

Table 6 displays the averaged correlations between the automated and human rater scores obtained for TOEFL CBT prompts and the scale scores of other multiple-choice (MC) sections (Structure, Listening, and Reading) and total scores of TOEFL CBT. Overall, both the automated scores and human rater scores were moderately correlated with the MC section scores, suggesting that both modes of scoring may reflect similar aspects of ESL language proficiency.

However, the human rater scores were somewhat more highly correlated to the Structure, Listening, and Reading scores. (i.e., 0. 59 for versus 0.52 for Structure, 0.60 versus 0.53-0.54 for Listening, and 0.55-0.56 versus 0.48-0.49 for Reading).Such a tendency was more salient for the Structure and Written Expression (SWE) section (0.83 versus 0.74) and total scores (0.72 versus 0.64). This was very much expected, because there was a so-called "part-whole overlap" phenomenon involved between the human rater essay scores and SWE section and total scores. To put it another way, the two human ratings (human 1, human 2) for the prompt being investigated contributed to the SWE combined section and total scores.

Table 6. Averaged Correlations Between Human and Automated Essay Scores and Other TOEFL CBT Section and Total Scores Over the CBT Seven Prompts.

	Structure	Structure + Writing*	Listening	Reading	Total*
Hm 1	0.59	0.83	0.60	0.55	0.72
Hm 2	0.59	0.83	0.60	0.56	0.72
G1	0.52	0.74	0.53	0.48	0.64
G2	0.52	0.74	0.53	0.48	0.64
G3	0.52	0.74	0.53	0.48	0.64
H1	0.52	0.74	0.54	0.48	0.64
H2	0.52	0.74	0.53	0.48	0.64
H3	0.52	0.74	0.54	0.48	0.64
PS	0.52	0.74	0.53	0.49	0.64

Note: Fisher's Z transformation and inverse Fisher's Z were used in computing the mean correlations; * indicates that Human 1 and Human 2 scores contributed to the score.

5. Summary and Discussion

The main purpose of the study was to create six different versions of generic scoring models for e-rater using the transformed TOEFL CBT writing data and examine the performance of these generic scoring models in the context of scoring independent writing tasks of TOEFL CBT, with a view to indirectly exploring the feasibility of eventually applying these models for TOEFL *i*BT. In terms of variability and reliability of automated scores, the results of the study have demonstrated that: (a) a high level of score agreement could be achieved between human ratings and automated scores from the generic models, (b) the types of scoring models and sampling of prompts had only a negligible impact on the variability of the automated scores produced by the generic and hybrid scoring models, and (c) the automated scoring in general seemed to have some subtle impact on examinees' score variability across tasks, although it increased rater-related score consistency significantly. In terms of criterion-related validity, the human rater scores turned out to be somewhat better indicators of test-takers' ESL language proficiency than the automated scores, although both the automated and human rater scores seemed to reflect similar aspects of ESL proficiency. However, such a validity difference seemed to disappear to a large extent when the scores from the writing tasks that were very similar to the one being investigated were used as validity criteria. More details of these major findings are discussed next along with their implications and other related issues deserving further investigation.

First, one of the most important findings in this study was that the six automated scores from the generic and hybrid scoring models were extremely highly correlated among themselves and also highly correlated with the automated score from the prompt-specific model. The automated scores from the generic and hybrid models

were also correlated with the human rater scores obtained for the same prompts at the moderate level. When the agreement rates were examined, extremely high levels of score agreement were achieved between the automated scores and human rater scores. Even when the kappa was used as an evaluation index, the agreement rates between one of the automated scores and one of the two human ratings were even higher than those between the two human ratings in the case of TOEFL CBT prompts.

Second, the means and standard deviations of the seven automated scores for the TOEFL CBT prompts were also examined across prompts to investigate if there is any systematic effect of automated scoring on the essay scores across the tasks. The means of the six automated scores from the generic and hybrid models change very similarly across the seven TOEFL prompts, even though there were some very small (and negligible) differences among them in terms of the mean scores.

Another interesting pattern related to the score distribution was that the standard deviations of the automated scores tended to be consistently smaller than those of human rater scores (Human 1, Human 2) across all of the seven TOEFL CBT. This means that automated scoring of essays resulted in slightly reduced score variations among test-takers, compared to the human ratings. One can argue that this difference may be artificial to some extent, because the standard deviations of automated scores from e-rater can be expanded by adjusting distributions of expected scores for the automated scores in the process of developing e-rater scoring model. Nevertheless, one possibility is that expanding the expected score distribution can potentially impact the rates of score agreement between the human and automated scores negatively. Such an adverse impact may need to be investigated in more depth in future studies.

Lastly, the criterion-related validity of automated scores was also investigated by

using the TOEFL CBT section scores. First, both the automated scores and human rater scores were moderately correlated with the TOEFL section scores, suggesting that both modes of scoring may reflect similar aspects of ESL language proficiency. However, the human rater scores were somewhat more highly correlated to the TOEFL CBT section and total scores. When the scores for the seven TOEFL CBT prompts were examined, the human rater scores were somewhat more highly correlated to the Structure, Listening, and Reading scale scores of TOEFL CBT than the automated scores.

6. Conclusions and Avenues for Further Investigation

The current study provided valuable information about the feasibility of using e-rater-based generic scoring models in scoring the essays written for the independent writing tasks of TOEFL CBT when they are scored on the 5-point scale used for TOEFL *i*BT. The study demonstrated that there was minimal variability among automated scores from the generic models and that a high overall score agreement could be achieved between human rater scores and automated scores from the generic scoring models. In addition, sampling of prompts turned out to have only a negligible impact on the performance of the generic and hybrid scoring models. Nothing would keep the TOEFL testing program from further experimenting with the generic scoring models in scoring independent writing tasks for TOEFL *i*BT in the future. It is recommended, however, that the TOEFL program consider carefully several implementation issues before making a decision regarding when and how to use e-rater generic scoring models for operational scoring, which may include but are not

limited to: (a) re-evaluation of the existing generic models on the operational TOEFL *i*BT data; (b) investigating the impact of manipulating relative weights for the essay features variables on the hybrid and generic models.

Acknowledgements

This paper is partly based on research work done while the first author was working at Educational Testing Service (ETS). Some portions of the paper were presented at the annual meeting of National Council on Measurement in Education (NCME) held in Chicago, USA from April 10 through 12, 2007 and at the 30th Annual Language Testing Research Colloquium (LTRC) held in Hangzhou, China, from June 25-28, 2008. I would like to thank Yigal Attali and Chi Lu for building and running scoring models of e-rater used in this project and Rosalie Hirch for proofreading an earlier version of this journal manuscript. Needless to say, the responsibility for any errors that remain is solely mine, and the ideas and opinions expressed in the paper are those of the author, not necessarily of ETS, TOEFL Program or Seoul National University (SNU).

Works Cited

Attali, Yigal, and Jill C. Burstein. "Automated Essay Scoring with E-rater V.2." *Journal of Technology, Learning, and Assessment* 4 (2006). Web. 30 March 2007.

Bennett, Randy E. *Moving the Field Forward: Some Thoughts on Validity and Automated Scoring* (ETS Research Memorandum RM-04-01). Princeton, NJ: ETS, 2004. Print.

_____, and Isaac I. Bejar. "Validity and Automated Scoring: It's not only the Scoring." *Educational Measurement: Issues and Practice* 17 (1998): 9-17. Print.

Brennan, Robert L. "Performance Assessments from the Perspective of Generalizability Theory. *Applied Psychological Measurement* 24 (2000): 339-53. Print.

Burstein, Jill C. "The E-rater® Scoring Engine: Automated Essay Scoring with Natural Language Processing." A*utomated Essay Scoring: A Cross-disciplinary Perspective.* Ed. M.D. Shermis and J.C. Burstein. Mahwah, NJ: Lawrence Erlbaum, 2003. 113-21. Print.

Chodorow, Martin, and Jill Burstein. *Beyond Essay Length: Evaluating E-rater's Performance on TOEFL Essays* (TOEFL Research Report No. 73; ETS RR-04-04). Princeton, NJ: ETS, 2004. Print.

Dikli, Semire. "An Overview of Automated Scoring of Essays." *Journal of Technology, Learning, and Assessment* 5 (2006). Web. 30 March 2007.

Enright, Mary K., and Thomas Quinlan. "Complementing Human Judgment of Essays Written by English Language Learners with E-rater® Scoring." *Language Testing* 27 (2010), 317-34. Print.

Kukich, Karen. "Beyond Automated Essay Scoring." *IEEE Intelligent Systems* September/October 2000 (2000): 22-27. Print.

Lee, Yong-Won. "Investigating the Feasibility of Generic Scoring Models of E-rater for TOEFL *i*BT Independent Writing Tasks." *English Language Teaching* 71. Print.

_____, and Brent Bridgeman. *E-rater Research Agenda and Work Plans for the Next Generation TOEFL*. Internal document. Princeton, NJ: ETS, 2004. Print.

_____, Claudia Gentile, and Robert Kantor. "Toward Automated Multi-trait Scoring of Essays: Investigating Relationships among Holistic, Analytic, and Text Feature Scores. *Applied Linguistics* 31 (2010): 391-417. Print.

_____, and Robert Kantor. *Dependability of New ESL Writing Test Scores: Evaluating Prototype Tasks and Alternative Rating Schemes* (TOEFL Monograph No. MS-31 ETS RR 05-14). Princeton, NJ: ETS, 2005. Print.

Powers, Don E., Jill C. Burstein, Martin Chodorow, Mary E. Fowles, and Karen Kukich. *Comparing the validity of automated and human essay scoring* (GRE No. 98-08). Princeton, NJ: ETS, 2000. Print.

Rudner, Lawrence, and Phill Gagne. "An Overview of Three Approaches to Scoring Written Essays by Computer." *Practical Assessment, Research & Evaluation, 7*(2001). Web. 5 October 2005.

Shermis, Mark D., and C. Jill, J.C. Burstein, eds. *Automated Essay Scoring: A Cross-Disciplinary Perspective.* Mahwah, NJ: Lawrence Erlbaum, 2003. Print.

국문초록

독립적 쓰기과제 에세이 자동채점 점수의 신뢰도 및 타당도: 일반형, 혼합형, 및 과제별 채점모델을 중심으로

이 용 원 (서울대학교)

본 연구는, 토플시험의 독립적 쓰기 과제(independent writing task)의 채점을 염두에 두고 영작문 자동채점시스템인 이레이터(e-rater®)를 사용해 일반형(generic), 혼합형(hybrid), 과제별(prompt-specific) 모형을 포함한 여러 자동화 채점모델을 만들어 보고, 이러한 채점모델들을 적용해 산출된 영어 쓰기 점수의 점수신뢰도와 타당도를 검증해 보는 데 그 목적이 있다. 이를 위해 컴퓨터기반 토플시험(TOEFL CBT) 쓰기과제 은행에서 총 3개의 서로 다른 과제표본을 추출하고 아울러 이 쓰기과제들을 위해 쓰여진 토플 에세이의 변환점수를 사용해서 총 6 개의 일반형 및 혼합형 이레이터 자동채점모델들을 만들었다. 이런 과정을 통해 만들어진 총 6개의 일반형 및 혼합형 채점모형과 과제당 1개씩 별도로 만들어진 채점모델을 총 7개의 토플 쓰기과제들을 위해 작성된 61, 089개의 토플 에세이들을 채점하는 데 사용하였다. 데이터 분석 결과, (a) 비록 에세이 자동채점은 채점자(채점모델) 간 점수 일관성을 증대시키는 효과가 나타났지만 실제 자동채점기 대 인간 채점자 간 점수 일치도와 두 인간채점자 간 점수 일치도는 유사한 수준을 보였고, (b) 인간채점자 점수가 자동채점 점수보다는 수험자의 전반적인 영어숙달도의 좀 더 나은 지표로서 사용될 수 있음이 밝혀졌다. 아울러 본 논문에서는 앞으로 자동채점 기술이 영어를 제2언어 혹은 외국어로 배우는 학습자의 영어 에세이를 채점하는 데 사용될 때 본 연구의 분석결과가 어떤 함의를 가지게 되는지도 논의된다.

주제어 : 에세이 자동채점, 이레이터, 일반형 채점모델, 독립적 쓰기과제, ESL 쓰기 평가

논문접수일: 2015.12.27
심사완료일: 2016.01.21
게재확정일: 2016.02.25

이름: 이용원
소속: 서울대학교 영어영문학과
주소: (08826) 서울특별시 관악구 관악로 1 서울대학교 인문대학 영어영문학과
이메일: ylee01@snu.ac.kr