Construct Validity in Human Scoring and Criterion:
What Criterion would(not) measure

Jungyeon Koo

1. Introduction

As the fourth generation revolution has begun, artificial intelligence (AI) has come to be applied in every field. Many students who are native speakers of English use the Automated Writing Evaluation (AWE) system in order to prepare themselves for the essay component of the Scholastic Assessment Test (SAT), a test for college admissions in the United States. In particular, a great amount of attention has been paid to automated essay scoring (AES) in English writing assessment and instruction (Enright & Quilan, 2010; Lee, Gentile, & Kantor, 2010; Weigle, 2010). Educational Testing Service (ETS), which administers the exam, has developed an AWE system called, Criterion. Criterion might be one of the most widely used AES systems in first language (L1) and second language (L2) contexts.

The strength of AES has in its objectivity and consistency in its evaluations and efficiency in terms of time and cost. Because of these strengths, the AES system has been used for scoring performance-based writing tasks, especially high-stake examinations. These systems have an additional upside, one expected of well-designed AES systems: it can

potentially make wider use of constructed or extended response tasks in language assessment (Lee, Gentile, & Kantor, 2008, as cited in Lee, 2016).

One of the important and widely used e-raters, the ETS-developed Criterion, is a web-based system that gives teachers automated scoring and evaluation of students' essays. Criterion provides students with six scaled-scores and feedback on grammar, organization and development, usage, mechanics, and style (Attali, 2004; Burstein, Chodorow, & Leacock, 2004). Some universities in Korea use AES systems like this, with Criterion being the most common. The AES Criterion system has actively been studied to investigate the extent to which its feedback and evaluations agree with human scoring (Stevenson & Phakiti, 2014). These studies show that the feedback cannot help students because it mainly focuses its comments on grammar, mechanics and styles. Moreover, there is little research on the construct dimensions in Criterion.

This is a pilot study aiming to investigate Criterion validity by comparing its evaluations to those of human raters obtained for the TOEFL independent writing tasks, while focusing on construct dimensions. In addition, the main purpose of this research is to determine which essay features are most closely related to each of Criterion's six different analytic dimensions: development, organization, vocabulary, sentence variety/construction, grammar/usage, and mechanics.

2. Literature Review

As mentioned in chapter 1, AES systems were originally created for students who are native speakers of English (L1 students): previous studies have been done mainly on L1 students by the very companies that developed those systems. Since it is only recently that writing teachers began to use AES systems in Korea, few studies have been

conducted on non-L1 students.

2.1. Reliability, Effectiveness, and Agreement between human raters and AES

Lee (2008) touched on the reliability issue by comparing to scores of one AES program, My Access, to those of five human raters and found a strong correlation between them. Also, Park (2011) examined the accuracy of computer scoring of Korean EFL students' essays by comparing agreement rates and correlations, and means between the human raters and the Intellimetric program.

Choi (2011) used Criterion to examine the effectiveness of AES integration types in terms of improving English writing quality and accuracy (the number of grammatical errors). 172 students participated in this research from an ESL program at a U. S. university and an EFL program at a Korean university. The students received writing instruction under three different types of AES integration (NO-AES, optional AES, and integrated AES) in the context of ESL and EFL settings. The results found that the types of AES integration significantly influenced the holistic scores of each writing assignment in terms of writing quality and accuracy. The integrated-AES group received significantly higher scores of the first draft and the final revisions than the optional-AES and the NO-AES groups. Moreover, the integrated-AES group improved holistic scores and reduced errors significantly more than the optional-AES and NO-AES groups. However, the research revealed that the learning environment (ESL vs. EFL) did not influence improvement in the quality and accuracy of students' writing. This research attempted to find AES is an effective instructional means for a formative assessment when it is integrated with writing instruction and process.

High rating agreement was found between AES systems and human rates in independent TOEFL writing tasks (Burstein & Chodorow, 1999;

Chodorow & Burstein, 2004; Powers, Burstein, Chodorow, Fowles & Kukich, 2002).

On the other hand, there are counterarguments against the agreement between AES and human ratings. To be specific, AES scoring is not a perfect substitute for human scoring but can be a useful complement to it (Enright & Quinlan, 2010; Attali, 2013; Cohen, Levi, & Ben-Simon, 2018; Mohsen & Alshahrani, 2019).

Lee, Gentile, and Kantor (2008) examined the distinctness and reliability of analytic rating dimensions and the relationships to holistic scores and e-rater® essay feature variables in the context of the TOEFL computer-based test (CBT) writing evaluation. The results found that 1) all analytic scale scores were not only correlated among themselves but also correlated with the holistic scores; 2) high correlation existed among holistic and analytic scores, which might be caused by the impact of essay length on analytic and holistic scores; 3) there may be some potential for profile scoring based on analytic scores, and some strong relationships were observed between several e-rater® variables and analytic rating. These findings made it possible to compare construct dimensions in human ratings and AES feature variables and to explore how to refine/revise essay feature variables.

In addition, Lee (2016) explored the reliability and validity of AES from different types of e-rater® scoring models in the context of scoring TOEFL independent writing tests. He proposed six different variants of generic and hybrid models from transformed written data for seven types of CBT tasks. His works revealed that similar levels of score agreement were shown between automated and human score pairs and between two human rater scores and that the human rater' scores could be better indicators of test takers' overall ESL language proficiency than the automated ones in terms of Criterion-related validity.

2.2. Feedback by AES

Most research focuses on Criterion feedback (Kim, 2010; Koh, 2017; Lee 2017; Moon & Pae, 2011). Kim (2010) conducted questionnaire surveys with 215 students on their perspectives on Criterion feedback (what Criterion can and cannot provide). Based on the results, she revealed that the students had the most difficulty with feedback on sentence structure and less difficulty with grammar, vocabulary, content, and organization. She concluded that AES feedback can be helpful to the students wishing to improve their writing quality.

Koh (2017) attempted to examine the effects of different application types of automated writing feedback on Korean EFL writing by employing Criterion which provides students with instant feedback when students are writing drafts. Application types were differentiated by the point of the writing stages at which students were able to access to Criterion. Therefore, the participants were classified into non-continuous feedback (NCF)-where students access to the system only once right before they submit their drafts and continuous feedback (CF) group-where students had not restrictions in accessing the system through the overall writing stages. The researcher found that the CF group outperformed the NCF group in the dimension of content and grammar although teachers' feedback primarily focuses on content and organization. This study also showed that the CF group showed a positive attitude toward receiving instant language-related feedback via Criterion than NCF group on grammar and mechanics dimension. However, this study mainly emphasized the effect of Criterion feedback but did not capture the weight of certain dimensions (grammar and mechanics) during feedback.

From different perspectives, Lee (2017) investigated students' perception on AES feedback with questionnaire on Criterion. The survey results showed that a majority of students perceived Criterion feedback as

useful for improving the overall quality of their essay. Among five categories of feedback, students found grammar to be the most useful, followed by mechanics, usage, organization and development, and style, order of usefulness. Students also observed that the use of Criterion contributed to self-directed learning.

Questioning the validity of Kim's (2010) findings, Moon and Pae (2011) also did a questionnaire-based study with university students to investigate short-term effects of automated writing feedback by comparing the university students' essay drafts with their subsequent revisions. The results showed that more than half of the errors were changed successfully from the Criterion feedback. However, they found that students rarely made changes related to the dimensions of organization or development due to the non-specific nature of the Criterion feedback. Since AES systems do not provide individualized feedback at the discourse level, the authors proposed that the students' questionnaires indicated a need for teacher feedback and teacher-to-student conferences to improve the content and organization of their writing.

Also, there are a few articles on correlation between AES scores and teachers' scores in the classroom (James, 2006; Wang & Brown, 2007; Ebyary & Windeatt, 2010; Li, Link, Ma, Yang, & Hegelheimer, 2014). James (2006) attempted to validate the use of Intellimetric scores by examining 60 students writing samples. He observed a low correlation between AES score and instructors' scores. Similarly, Wang and Brown (2007) used writings by Hispanic English-speaking students. The AES program (Intellimetric) scores showed a weak correlation with the instructor scores despite using the same rubric. The researchers proposed that the low correlation might be due to the characteristics of the student population and the writing instruction they received.

In contrast to the research on the Intellimetric program, Ebyary and Windeatt (2010) and Li et al. (2014) employed Criterion for their research.

Ebyary and Windeatt (2010) examined data from 31 experienced instructors and 549 Egyptian trainee EFL instructors. Twenty-four of the trainees received Criterion feedback on two drafts of essays on each of four different topics. Two English language tutors evaluated a representative sample of texts by employing the Criterion scoring scale. The researchers observed a significant inter-rater reliability between the first rater and Criterion and a moderate inter-rater reliability between the second rater and Criterion.

The study conducted by Li et al. (2014) explored the use of holistic scores for classroom purposes in ESL contexts with mixed methods. They examined the correlation between AWE scores and instructors' numeric grades and analytic ratings on two major course assignments. The researchers found low or moderate positive correlations between mixed methods. They examined the correlation between AWE scores and instructors' numeric grades and analytic ratings on two major course assignments. The researchers found low or moderate positive correlations between Criterion scores and the two instructors' grades and analytic rating. This study suggested that Criterion had at least some usefulness and raised questions concerning the use or non-use of automatically generated scores for classroom-based assessment.

In sum, many studies have been conducted that compare Criterion feedback and evaluations given by Criterion with those given by human raters (Stevenson & Phakiti, 2014). On the other hand, there are a few studies that compared AWE holistic scores and human scores in the classroom context and there is inconsistency among correlations between the two scores.

2.3. Research questions

The motivation behind this study is inspired by Moon and Pae's (2011) findings that students scarcely corrected their mistakes on the

dimensions of the organization or development traits due to the non-specific nature of the Criterion feedback. This result is explained by Criterion's lack of detailed, individual feedback on the content and organization. Also, this finding implies that the AES system does not focus on evaluating content and organization as much as other construct dimensions, such as grammar/usage, mechanics, vocabulary, and sentence variety/construction. As mentioned above, most prior research focused primarily on Criterion's student feedback, and students' perceptions thereof, by comparing inter-rater agreement between AES system and human raters. For this reason, a research gap exists in the investigation of the nature of (sub)constructs to be measured in AES scoring. Therefore, the research question (RQ) posed for the current study are as follows:

> RQ 1. Do human raters and Criterion show a high rate of agreement and correlation when scoring an essay?
>
> RQ 2. Are there any relationships between holistic scores and Criterion essay feature variables in the context of the TOEFL *iBT* writing evaluation in terms of
>
> (1) which rating dimensions (among six dimensions[1]) best predict the total score by AES (Automated Essay Scoring)?
> (2) which rating dimensions (among the six dimensions) best predict the total score by human raters?
>
> RQ 3. Which prompt type shows the highest correlation between AES (Automated Essay Scoring) system rating and human rating?

This study has importance in 1) examining which traits (writing

1) Six dimensions consist of development, organization, vocabulary, sentence variety/construction, grammar/usage, and mechanics.

dimensions) in Criterion has the closest association with six dimensions in human rating and the validity by comparing human rating with AES in assessing writings, and 2) in suggesting pedagogical implications in teaching writings to Korean EFL students and in assessing writing through AES by finding the rating scales are different in human rating and AES.

3. Experiment Design and research method

This chapter will describe details on how to gather data, recruit participants, and analyze data to examine the reliability of construct dimensions by human and features by *e-rater*.

3.1. Data Collection

Participant for this study were mostly recruited from an online "seeking jobs" bulletin board in Seoul National University and the remaining participants were recruited from Facebook. They were paid 7500 won per essay. The number of participants chosen was fifty because this research is a test to find patterns and predictability using big data. The participants were college students who specialized in a variety of disciplines, including Spanish, English, linguistics, law, pharmacology, Cyber national defense, business administration, electronic engineering, computer engineering, and economics. They were asked to choose one topic out of five prompts[2)] and given period of 30 minutes to write about it. In addition, participants were also given a survey about their personal history and other information about their English

2) As for the test materials, there are five types of prompts. Topics (independent) are as follows: Prompt 1: Money on Technology, Prompt 2: Change Job or Not, Prompt 3: Learn from Mistakes, Prompt 4: Method of Travel, Prompt 5: Important Plant.

writing. Each consent form was signed by all participants who took the independent TOEFL writing test.

Chodorow and Burstein (2004) and Lee et al. (2008) found that the essay length could be the indicator most associated with writing quality. Therefore, because the purpose of this study is to examine the relationship between the six construct dimensions and the writing score, the essays' word counts should be controlled in order not to overly influenced on the scores given by human raters and the e-rater. Accordingly, the word length in each essay was limited to being 200 and 300 words.

3.2. Analysis Method

The data collected for the current study was analyzed quantitatively and qualitatively. For the quantitative analysis, SPSS 22 was used to compute the agreement rate (correlation) between human raters and Criterion, to predict the best indicators (factors) among six rating constructs adopted from Lee et al. (2008)—see Table 4 in chapter 4—in human and e-rater, and to investigate correlation between a prompt type and total scores measured by human raters and Criterion.

Two human raters evaluated fifty writings. They are English instructors at university and have taught students English for more than 10 years. The two human raters were trained how to measure fifty compositions then asked to assess two sample essays to show the score differences within three points.

For the qualitative analysis, two human raters were interviewed, and all participants who received feedback from Criterion filled out a questionnaire which included questions about the quality of the Criterion feedback (See Appendix 2). However, analyzing questionnaire responses is not main concern in this study. Therefore, the findings will not be shown in detail here. Yet, the results were reaching and providing support for my conclusion.

Concerning Criterion feedback. Criterion provides test-takers with two types of analyses: (a) score analysis and (b) trait analysis in Figure 1 and 2 below. Figure 1 reveals that score analysis is given separately for each construct dimension 1 through 6, not the composite scores Criterion gives its score feedback in terms of 1) word choice, 2) grammar, usage and mechanics, conventions, and 3) organization, development and style such as "Advanced" or "Proficient" with score range ranged from 1 to 6. Lee et al. (2008) suggested six analytic scale construct dimensions and features by e-raters which corresponded to each analytic scale.

The scores determined based on the description, such as advanced or proficient, in three areas:—(1) word choice, (2) grammar, usage and mechanics (conventions), and (3) organization, development and style. Moreover, these areas are assessed as qualitative descriptions, i.e., "advanced" or "proficient," but these descriptions do not indicate the score range directly. Though the same point scale is assigned to two writings, the descriptions of score analysis are different depending on each writer's performance in one's writing. The description of score

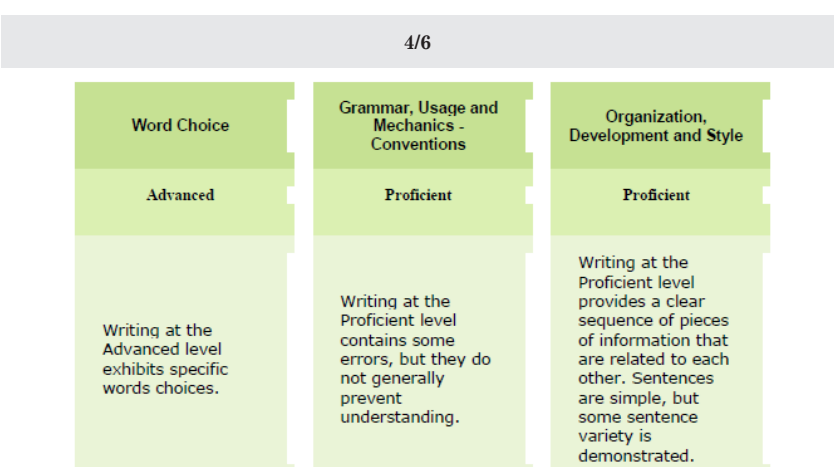


Figure 1. Score Analysis in Criterion

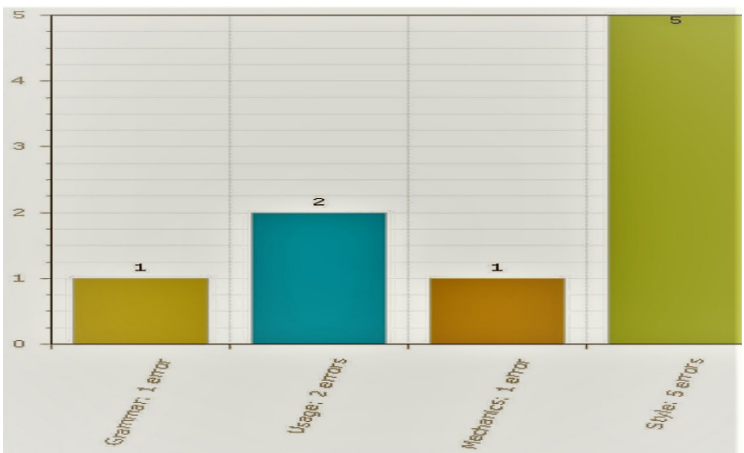


Figure 2. Trait Analysis in Criterion

analysis guide is shown in Appendix 1.

Figure 2 shows the number of errors in each trait feature with graphs. This graph shows the frequencies of errors in usage, mechanics, grammar, and style.

Trait analysis in Figure 3 and 4 also give the test-takers the frequency/existence of introductory parts, theses, main ideas, supporting ideas, and other areas (e.g. transition words and phrases, repetition of words, and so on). Thus, the trait analysis feedback helps the students grasp their errors in each dimension quickly and clearly. However, this feedback does not identify or evaluate their logical flow, the existence of task fulfillment, or the presence of irrelevant sentences (cohesion).

In Figure 2, "Usage" provides writing information on determiner-noun agreement, missing/extra articles, confused words, incorrect word form, faulty comparisons, preposition error, and non-standard word form. "Mechanics" includes frequency errors on spelling, capitalization of proper nouns, missing initial capital letter, missing question mark, missing final punctuation, missing apostrophe, missing comma, hyphen error, fused words, compound words, and extraneous commas.

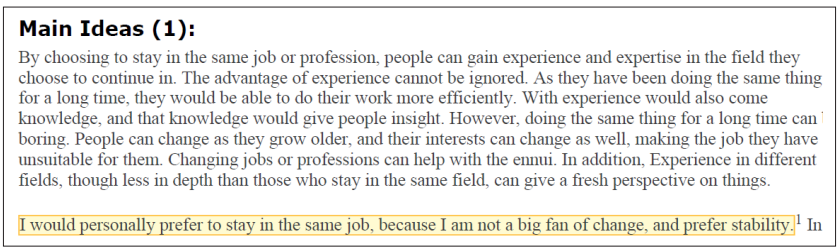


Figure 3. Trait Analysis in Criterion

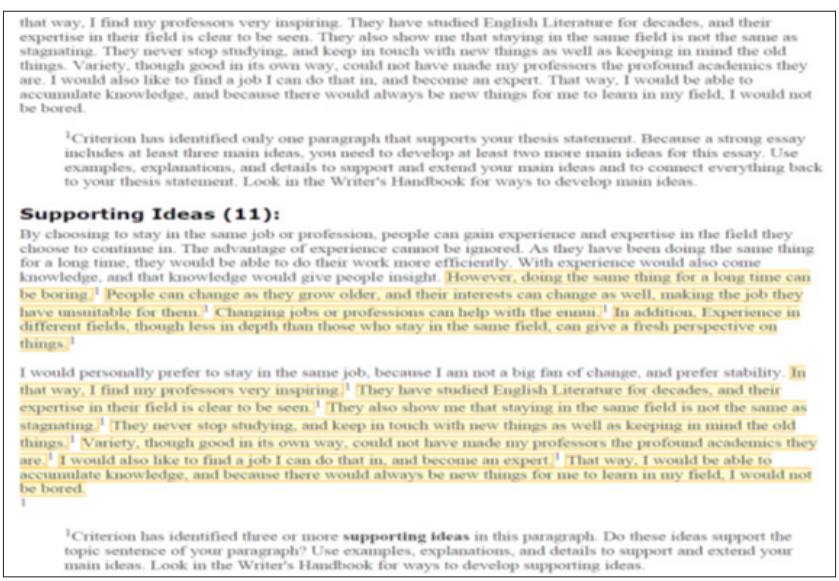


Figure 4. Trait Analysis in Criterion: Supporting Ideas

"Grammar" gives test-takers information on sentence fragments—run-on sentences, garbled sentences, subject-verb agreement, ill-formed verbs, pronoun errors, possessive errors, wrong/missing words, and proofread this![3] errors. Finally, "Style" consists of errors in repetition of words, inappropriate words/phrases, sentences beginning with coordination,

3) Proofread this! is a type of grammatical error that cannot be discern among several types of errors among grammatical/usage/mechanics

short sentences, long sentences, and passive voice.

Concerning the prompt types, the five prompts were given to the participants: 1) money on technology, 2) change job or not, 3) learn from mistakes, 4) method of travel, and 5) important plant (See Appendix 3). Participants could choose one of the five prompts as they wanted to. The type of prompt is argumentative.

4. Results and Discussion

In this section, major findings are summarized and discussed in relation to the aforementioned research questions. The first research question was, "Do human raters and the e-rater show a high rate of agreement and correlation when scoring an essay?" Before responding to the first research question, it should be noted that the inter-rater agreement between two human raters was examined using the obtained Pearson correlation. The value of it is .517** at the 0.01 significance level, which indicates a moderate positive correlation between two human raters.

In addition to the Pearson correlation, Spearman's p (rho) was calibrated because this is a non-parametric test. According to Table 1, this correlation coefficient was .517 at the 0.01 significance level. Returning, then, to the first question, the answer is partially "yes," meaning there was a "moderate" agreement between AES and human raters.

Criterion has a 6-point scale (ranged from 1 to 6) and human raters' score is a continuous scale (ranged from 1 to 30). Thus, each essay was converted into Z-score,[4] to provide a standardized metric for comparison

4) A Z-score is a numerical measurement that describes a value's relationship to the mean of a group of values. It is measured in terms of standard deviations from the mean. If a Z-score is 0, it indicates that the data point's score is identical to the mean score. For example, a Z-score of 1.0 would indicate a value

Table 1. Correlation between human raters and Criterion

		HUMAN	AES
HUMAN	Pearson coefficient	1	.517**
	significance(2tailed)		.000
	N	50	50
AES	Pearson coefficient	.517**	1
	significance(2tailed)	.000	
	N	50	50

**. Correlation is significant at the 0.01 level (2tailed)

between the two modes of scoring.

The value of Spearman's p is .517, which indicates a moderate agreement between human raters and Criterion. This moderate agreement suggests not only that leniency and strictness exist in both the AES and human ratings but also that disagreement emerges among the analytic dimensions in evaluating the same writings. The result for the first research question is thus linked to the second research questions on analytic rating dimensions in human rating and features in Criterion. Moreover, the first finding may posit that there are some dimensions which cannot be captured by Criterion.

The second research questions are two-fold: 1) Which rating dimensions among six do predict the total score by AES?, and 2) Which rating dimensions among six do expect the total score by human raters?

To find a few rating dimensions among six ones which can predict the whole score better, the top three independent variables which has the first, second, and third highest beta value (marked in red in Table below) were chosen by doing linear regression with SPSS 22 based on

---

that is one standard deviation from the mean. It may be positive or negative, with a positive value meaning the score is above the mean and a negative score indicating it is below the mean.

Table 2. The Predictable Construct Dimensions in human rating[5)]

Model		Non-standard coefficient		standard coefficient	t	Significance level	B's 95.0% confidence interval	
		B	SE	beta			minimum	maximum
1	(constant)	1.656	1.160		1.427	.161	-.684	3.996
	DEV	.831	.168	.303	4.955	.000	.493	1.169
	ORG	1.093	.178	.389	6.132	.000	.734	1.453
	VOCA	.746	.328	.145	2.275	.028	.085	1.407
	SENTENCE	.068	.387	.010	.177	.860	-.711	.848
	GRAMMAR	1.169	.254	.242	4.595	.000	.656	1.683
	MEC	1.427	.513	.153	2.782	.008	.392	2.461

a. Dependent Variable: HUMAN

Note: DEV means development, ORG means organization, VOCA means vocabulary, SENTENCE means sentence variety and constructions, GRAMMAR means grammar and usage, and MEC means Mechanics.

the result from Table 2.

The second research questions are two-fold: 1) which rating dimensions among the six best predict the total score by AES? and 2) which rating dimensions among the six best predict the total scores by human raters?

To find which of the six ratings dimensions best predict the total scores, the top three independent variables (those with the first, second, and third highest beta value, respectively) were chosen by performing a linear regression with SPSS 22 based on the result from Table 2. Beta coefficient signifies that each factor shows an explainable power of predicting writing scores.

According to Table 2, beta coefficient of DEV is .303, that of ORG was .389, and that of GRAMMAR & USAGE was .242. This result means that the three writing constructs, development, organization, and grammar

5) For the analysis, linear regression was done and the chart in Table 2 was an excerpt from the whole results.

Table 3. A Model Summary of the relation between construct dimension and total score in human rating

Model Summary[b]				
Model	R	R square	Adjusted R square	Standard Error of the Estimate
1	.961[a]	.923	.912	.90246

a. Predictive: (Constant), MEC, SENTENCE, DEV, GRAMMAR, ORG, VOCA
b. DV: HUMAN

and usage were crucial factors to predict the holistic scores in human rating. In particular, the organization was the most powerful predictor of Criterion overall scores (.389). In addition, development was the second most indicator, and grammar and usage are the third most predictor to predict writing quality by human raters.

Also, the R-squared value showed that this multiple linear regression model in human rating has a strong predictive value because the value of R-squared was .961 (see Table 3). This R-squared value means that variability in outcomes can be largely explained by the regression model.

The second research question concerns relationships to holistic scores and Criterion essay feature variables in the context of the TOEFL *iBT* writing evaluation in terms of rating dimensions in human and Criterion ratings, i.e., which construct dimension in Criterion corresponds to analytic sores in human assessments. Analytic scales in human rating consist of six dimensions: development, organization, vocabulary, sentence variety/construction, grammar/usage, and mechanics (see Table 4).

Each of the six analytic dimensions corresponds to a set of essay features. Development corresponds to “thesis, main ideas, and supporting ideas” in AES. Organization corresponds to “introduction-body-conclusion structures, and the number of transitional words or phrases.” Vocabulary corresponds to “repetition of words.” Sentence variety/construction corresponds to “the use of passive forms, (too) short sentences, and the number of sentences prefaced with coordination.” Grammar/usage

Table 4. Six Analytic Dimensions by human raters and Features by E-rater[6)]

Features by Human Rater	Features by E-rater	S	Items Evaluated
Development	Organization & Development	8	-**Task fulfillment:** Interpretations of prompt -**Appropriateness of Details:** Supporting ideas are relevantly described after main ideas? -**Development:** Extension of Ideas and the length of words (200-300 wds)
Organization		8	-**Organization:** includes Intro, Body, and Conclusion -**Transition Words (TW):** Connectives are adequate? -**Cohesion:** TWs are adequately used to describe a relationship between ideas? Demonstratives and references words are appropriately employed to refer to previous ideas? -**Coherence:** regular use of superstructures and sequential progression
Vocabulary	Lexical sophistication (Type/token ratio, word length, voca level)	4	**-Range of Vocabulary:** repetition of words, levels of words, variety of words
Sentence variety/ construction	Linguistic Accuracy -Grammatical accuracy ratio -Usage accuracy ratio -Mechanical accuracy ratio -Stylistic accuracy ratio	3	-**Syntactic Variety:** Controlled and Varied sentence structures -**Style:** The style of writing is academic and argumentative?
Grammar/ Usage		4	-**Word Choice Errors:** ill-formed verbs, pronoun errors, possessive errors, wrong or missing words, determiner and noun agreement, articles -**Syntax Errors:** run-on sentences, split sentences, subject-verb agreement -Style: Words are academically used?
Mechanics		3	-**Word Form Errors:** spellings -**Capitalization, punctuation, and quotation marks in sentences**, and so on.
Total		30	

6) Each (analytic) construct dimension was given partial points: development and organization had 8 points, vocabulary and grammar had 4 points, and sentence variety/construction and mechanics had 3 points. The different emphasis on each analytic scale is based on the trait coverage and the items in writing analytic scale and corresponding features in Criterion adopted from Lee et al. (2008).

corresponds to grammatical or stylistic errors, such as; "fragments, run-on sentences, garbled sentences, subject-verb agreements, ill-formed verbs, pronoun errors, possessive errors, proofread this! errors (See footnote 6), determiner-noun agreement, missing/extra article, confused words, wrong form of a word, faulty comparisons, preposition errors, inappropriate words/phrases and wrong articles." Finally, mechanics corresponds to "spellings, capitalization of proper nouns, missing initial capital letter, missing question mark, missing final punctuation, missing apostrophe, missing comma/punctuations, hyphen error, extraneous commas, and fused words." Feedback for development and organization are provided as descriptions by marking the positions, i.e., introductory part, conclusion, thesis, main ideas, and supporting ideas. However, feedback for the four scale dimensions (vocabulary, sentence variety/construction, grammar/usage, and mechanics) was given as the number of errors in each dimension in chart.

Another question concerns what rating dimensions among the six in human rating are predictable. According to the findings with respect to RQ 2.1., DEV, ORG and GRAM/USAGE (grammar/usage), the three independent variables show the highest beta value, meaning that these three factors can strongly predict the holistic scores in human-rated essays. Because six scale dimensions are so many, backward elimination was performed in order to measure the predictability of the three scale dimensions (DEV, ORG, and GRAM), hypothesizing that the six analytic scale corresponds to e-rater's feature (variables) in Table 4 (Lee et al., 2008). Thus, based on this result, I ran a linear regression to find which factors among three elements (DEV, ORG, and GRAM) can explain the writing quality in Criterion. Table 5 showed that these three construct dimensions in Criterion are not strong ones which predict writing scores, compared to those in human rating.

According to Table 5, ORG is the strongest factor (.439), though not

Table 5. Beta scores of development, organization, and grammar in AES and human rating

Construct	Beta coefficient in AES (Criterion)	Beta coefficient in human rating
DEV	.086	.220
ORG	.439	.517
GRAM&USAGE	.036	.203

Note: DEV means development, ORG means organization, and GRAM & USAGE means grammar and usage.

as strong as its correlation to human rating (.517). However, DEV (.086) and GRAM/USAGE (.036) constructs in AES are not strong factors in explaining Criterion writing scores; beta coefficient values are much smaller than those in human ratings (.220, .203). This means that AES and human raters weighted the dimensions differently when scoring the essays.

In fact, constructing a linear regression revealed ORG, DEV (development), and GRAM&USAGE (grammar and usage) to be the strongest constructs for predicting Criterion writing scores. These three constructs showed high beta coefficients, though not as high as the top three predictors for human scores.

Criterion's feedback was shown to be incorrect in some 18 cases out of the 50 total essays (essays number 5, 12, 16, 17, 22, 27, 31, 34, 35, 36, 37, 39, 40, 41, 42, 43, 48, and 49). For example, in the description of thesis, main idea, supporting idea, introductory materials (remarks), and conclusion, Criterion could not capture the right position because human coders encoded the information on organization as follows: Theme goes first, then two or three main ideas, and supporting ideas following each main idea. For this reason, if a certain essay did not follow these precise specifications coded in Criterion, the AES system gave a low score and incorrect feedback to the student by detecting and marking the thesis, introductory materials, main idea(s), supporting

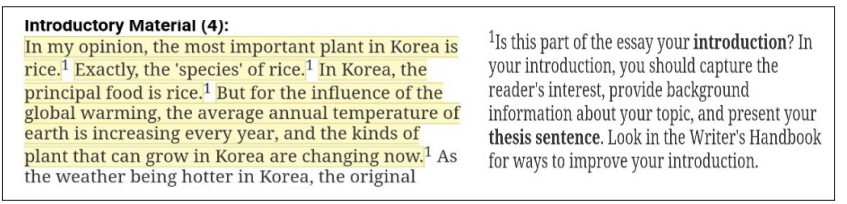


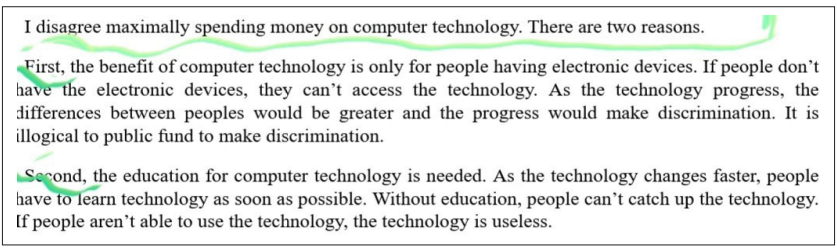
Figure 5. A writing sample with no introduction and Criterion's feedback

idea(s), and conclusion automatically.

Some essays showed inconsistency between AES and human-rater scores. Specifically, nine writings were evaluated favorably by human raters and unfavorably by AES, while seven writings were measured favorably by AES and unfavorably by the human raters. These sixteen discrepancies between the scores, despite there being no difference in and participants' actual writing ability, revealed that human raters and Criterion interpreted writing quality differently.

Examining these discrepancies between Criterion and human scores more closely reveal that essay receiving low scores from AES are not likely to follow the standard organization patterns, such as theme-main idea-supporting ideas. The AES system measured organization based on its trained standard structure whereas human raters gave good scores to the same writings focusing on logical flow without regard for rigid standards of structure. For example, Figure 5 showed that writing started with the main idea and thus Criterion gave feedback that there is not an appropriate introduction. Also, this writing received unfavorable grades compared to human raters' scores.

Additionally, compositions with high scores from AES tended to be lengthy and/or to use connectors precisely and follow the standardized structure, even when the resulting logical flow was odd. For instance, Figure 6 revealed that writer with clear use of connectors received a high score by Criterion. However, human raters gave a low score to this essay



I disagree maximally spending money on computer technology. There are two reasons.

First, the benefit of computer technology is only for people having electronic devices. If people don't have the electronic devices, they can't access the technology. As the technology progress, the differences between peoples would be greater and the progress would make discrimination. It is illogical to public fund to make discrimination.

Second, the education for computer technology is needed. As the technology changes faster, people have to learn technology as soon as possible. Without education, people can't catch up the technology. If people aren't able to use the technology, the technology is useless.

Figure 6. A writing which does not match theme and supporting ideas

because the writer's main idea and supporting ideas did not match well.

In my observation, while Criterion's trait analysis provides students with the number and verbal descriptions of their errors, it had no ability to detect logical flow or task fulfillment despite checking prompt-specific word usage. According to interviews with human raters, human raters put the highest emphasis on development and organization in evaluating writings. These two dimensions were assigned 8 scores (27 per cent of the whole score) according to Table 4.

A few previous studies (Kim, 2010; Moon & Pae, 2011) observed that Criterion feedback was partially wrong and that students did not change their writings in the organization/development trait (Moon & Pae, 2011).

According to Table 6, grammar can predict the overall score in AES by conducting a linear regression. This is untrue for human ratings. This finding implies that description feedback in DEV and ORG is not enough for test-takers to revise their writing effectively and that the overall score does not accurately reflect the quality of the writing along the DEV and ORG dimensions.

The third research question is about which prompt type shows the highest correlation between Criterion and human assessments. After converting writing scores by human raters to standardized Z-scores in order to compare 6-point scales by AES, linear regression was conducted using SPSS. Correlation between prompt type and overall score in AES

Table. 6. A Model Summary of the relation between construct dimension and total score in Criterion

Model Summary[b]					
Model	R	R square	Adjusted R square	Standard Error of the Estimate	Durbin-Watson
1	.501[a]	.251	.202	3.25417	1.799

a. Estimated variables: (Constant), GRAMMAR, DEV, ORG
b. Dependent variables: SCORE_CON

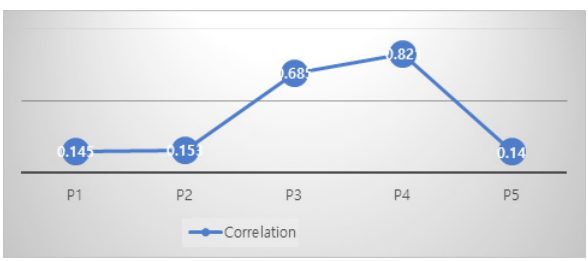


Figure 7. Correlation between prompt type and overall score in AES and human raters

and human raters is observed in Figure 7. Spearman's p coefficient in each prompt was given in a blue circle in the figure.

Figure 7 revealed that prompt 3 (.685) and prompt 4 (.821) showed a strong correlation between human rating and AES. This result means that the writing instructions/questions in prompt 3 and prompt 4 were clear for human raters and AES alike and that the task types in prompt 1 (.145), prompt 2 (.153) and prompt 5 (.14) were vague or difficult to understand both for test-takers and for the AES and human raters. Tiers of difficulty among the five prompts do not exist in principle. However, based on the feedback from the test-takers, some of the prompts were indeed less clear to understand than other were. Further, some of these results might be due to the different emphasis placed on each construct feature/dimension by human raters and the AES.

It is clear that Criterion and human raters showed a high correlation

between their scores for prompts 3 and 4. However, Criterion and two human raters displayed a low correlation between their points for prompts 1, 2, and 5. This result suggests that the prompt type affected the writing quality. According to Shi, Huang, and Lu (2020), prompt type significantly affected the participants' overall continuation writing scores. Most of all, the five writing prompts are issuable (responses should include "Agree" or "Disagree.") and were not intended to vary in difficulty. Thus, it is assumed that types of writing prompts might influence on writing difficulty. This issue should be studied in future research.

5. Conclusions

This study which aims to examine the validity of the AES, Criterion in assessing *iBT* TOEFL independent tasks by comparing its assessments with those given by human raters' and to investigate the validity of construct dimensions in human rating and Criterion. To this end, five different prompts were employed to obtain essay samples from 50 college students in Seoul. The result showed moderate agreement ($p = .517$) between human-rater and Criterion scores.

In addition, by controlling the influence of essay length on evaluating writing quality, development, organization, and grammar emerged as crucial indicators to predict the holistic score in human rating, while organization, mechanics and sentence variety/construction were powerful predictors of overall scores given by Criterion. This result suggests that some sub-features in development (logical flow and task fulfillment) and organization features (the position of thesis, main ideas, supporting ideas, introductory materials, and conclusion) cannot be assessed properly by Criterion and that the e-rater needs to be refined/revised in measuring scores in the development and organization construct dimensions.

Lastly, prompt type showed different correlations with scoring an essay by AES system and human raters. This may reflect that different levels of difficulty existed in comprehending the prompts, although all the writing prompts are issuable. In order to avoid this unintended variable in future studies, prompt type should be controlled for equal measurement.

An interesting discussing point is that the scoring rubric by the human raters in the current study was similarly revised based on Criterion's basic scoring process (Lee, 2016). However, there are some discrepancies between AES 6-point scores and holistic writing scores. It can be assumed that the reason for this is that the respective weight assigned to each subcategory was the same/similar in human raters' scoring rubric, whereas Criterion weights its scores according to 12 features, and the AES software can be influenced by word count. Furthermore, Criterion could not check off-topic sentences or breaks in cohesiveness, though Criterion puts a higher weight on organization (Enright & Quilan, 2010).

The findings have some implications for teaching students process writing and for using AES. Students can use Criterion feedback to revise their writing and are able to receive direct feedback from the AES as well for several times. Therefore, the AES can aid students in developing their process-based writing. That is to say, students can practice self-feedback with the help of AES.

Moreover, teachers can use the AES system to grasp how their students understand and improve their writing by themselves. Technology provides many of the tools necessary to promote learning and knowledge acquisition (Yun, 2014) and even to assess writing and give feedback to EFL learners.

This study has some limitations. First, a larger sample of participants is necessary to detect a pattern/rule in AES scoring. In particular, groups classified based on relative proficiency are required to analyze the Criterion feedback and score. In addition, variety among prompts should

be controlled.

Based on the result in this study, the prompt type showed a correlation with a total score. If research intends to focus on finding the relationship between a certain construct and the total score, the topic should be limited to a single prompt. Finally, if e-rater can be used with the permission of ETS, scoring processes in Criterion will be examined at discourse levels. Thus, it will be possible to find out why feedback on the development and the organization dimensions is weak compared to human scoring.

It is hoped that this study can help enhance construct refinement in Criterion for a stronger agreement between human raters and AES systems.

References

Attali, Y. (2004). Exploring the feedback and revision features of Criterion. Paper presented at the National Council on the measurement in education. San Diego, CA.

Attali, Y. (2013). Validity and Reliability of Automated Essay Scoring. In M. D. Shermiss, & J. Burstein (Eds.), *Handbook of automated essay evaluation: Current applications and new directions* (pp. 181-198). New York, NY: Routledge.

Burstein, J., & Chodorow, M. (1999). Automated Essay Scoring for Nonnative English Speakers. Proceeding ASSESSEVALNLP '99 Proceedings of a Symposium on Computer Mediated Language Assessment and Evaluation in Natural Language Processing, U. S. A., 68-75.

Burstein, J., Chodorow, M., & Leacock, C. (2004). Automated Essay Evaluation: The Criterion Online Writing Service. *AI Magazine 25*(3), 27-36.

Chodorow, M., & Burstein, J. (2004). Beyond Essay Length: Evaluating e-rater®'s Performance on TOEFL Essays. (ETS report, RR-04-04). Prinston, New Jersey, U. S. A. ETS,

Choi, J. H. (2011). Integration of Automated Essay Scoring in the Process of Writing in ESL/EFL Classes. *The Journal of Educational Information and Media, 17*(2), 177-196.

Cohen, Y., Levi, E., & Ben-Simon, A. (2018). Validating human and automated scoring of essays against "True" scores. *Applied measurement in education, 31*(3), 241-250.

Ebyary, K., Windeatt, S. (2010). The Impact of Computer-based Feedback on Students' written work. *International Journal of English Studies, 10*(2), 121-142.

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.

James, C. (2006). Validating a Computerized Scoring system of Assessing Writing and Placing students in composition courses. *Assessing Writing, 11*(3), 167-178.

Kim, T. E. (2010). Reflection on using the Criterion online writing evaluation service. *Multimedia-Assisted Language Learning, 13*(3), 59-83.

Koh, W. Y. (2017). Effective Applications of Automated Writing Feedback in Process-based Writing Instruction. *English Teaching, 72*(3), 91-118.

Lee, S. M. (2008). Exploring the potential of a web-based writing instruction program and AWES: An empirical study using My Access. *Multimedia-Assisted Language Learning, 11*(2), 103-125.

Lee, Y. J. (2017). Students' Perceptions of the Automated Writing Evaluation Feedback in Writing Courses. *Secondary English Education, 10*(4), 143-164.

Lee. Y. W. (2016). Investigating the Feasibility of Generic Scoring Models of E-rater® for TOEFL® *iBT* Independent Writing Tasks. *English Language Teaching, 28*(1), 101-122.

Lee, Y. W., Gentile, C., Kantor, R. (2008). Analytic Scoring of TOEFL® CBT Essays: Scores from Humans and E-rater®. (ETS Report, RR-08-01), Princeton NJ, U. S. A., ETS.

Lee, Y. W., Gentile, C., & Kantor, R. (2010). Toward Automated Multi-trait

Scoring of Essays: Investigating Relationships among Holistic, Analytic, and Text Feature Scores. *Applied Linguistics, 31*, 391-417.

Li, Z., Link, S., Ma, H., Yang, H. J., & Hegelheimer, V. (2014). *The Role of Automated Writing Evaluation Holistic Scores in the ESL classroom. System, 44*, 66-78.

Mohsen, M. A., & Alshahrani, A. (2019). The Effectiveness of Using a Hybrid Mode of Automated Writing Evaluation System on EFL Students' Writing. *Teaching English with Teaching, 19*(1), 118-131.

Moon, Y. I., & Pae, J. K. (2011). Short-term Effects of Automated Writing Feedback and Users' Evaluation of Criterion. *Korean Journal of Applied Linguistics, 27*(4), 125-150.

Park, T. J. (2011). Examining the Accuracy of Computer Scoring of Korean EFL Students' Essays. *The Journal of Linguistic Science. 56*, 53-74.

Powers, D., Burstein, J., Chodorow, M., Fowles, M. E., & Kukich, K. (2002). Comparing the validity of automated and human essay scores. *Journal of Educational Computing Research, 26*(4), 407-425.

Shi, B., Huang, L., & Lu, X. (2020). Effect of prompt type on test-takers' writing performance and writing strategy use in the continuation task. *Language Testing, 37*(3), 361-388.

Stevenson, M., & Phakiti, A. (2014). The effects of computer-generated feedback on the quality of writing. *Assessing Writing, 19*, 51-65.

Wang, J., & Brown, M. S. (2007). Automated Essay Scoring versus Human Scoring: a comparative study. Journal of Technology, *Learning and Assessment, 6*(2), 719-725.

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

Yun, D. H. (2014). How Do On-Line Students Use Technology to Enhance Their Learning? *Studies in Foreign Language Education, 29*(1), 51-67

Appendices

Appendix I. Criterion's Score Guide (Holistic):

Score 6:	A typical essay at this level: • effectively addresses the writing task • is well organized and well developed • uses clearly appropriate details to support a thesis or illustrate ideas • displays consistent facility in the use of language • demonstrates syntactic variety and appropriate word choice, though it may have occasional errors
Score 5:	A typical essay at this level: • may address some parts of the task more effectively than others • is generally well-organized and well-developed • uses details to support a thesis or illustrate idea • displays facility in the use of language • demonstrates some syntactic variety and range of vocabulary, though it will probably have occasional errors
Score 4:	A typical essay at this level: • addresses the writing topic adequately but may slight parts of the task • is adequately organized and developed • uses some details to support a thesis or illustrate an idea • demonstrates adequate but possibly inconsistent facility with syntax and usage • may contain some errors that occasionally obscure meaning
Score 3:	A typical essay at this level may reveal one or more or the following weaknesses: • inadequate organization or development • inappropriate or insufficient details to support or illustrate generalizations • a noticeably inappropriate choice of words or word forms • an accumulation of errors in sentence structure and/or usage
Score 2:	A typical essay at this level is flawed by one or more of the following weaknesses: • serious disorganization or underdevelopment • little or no detail, or irrelevant specifics • serious and frequent errors in sentence structure and usage • serious problems with focus
Score 1:	A typical essay at this level: • may be incoherent • may be undeveloped • may contain severe or persistent writing errors

Appendix II: Writing Prompts

Prompt 1: Money on Technology

Some people think that governments should spend as much money as possible on developing or buying computer technology. Other people disagree and think that this money should be spent on more basic needs. Which one of these opinions do you agree with? Use specific reasons and details to support your answer.

Prompt 2: Change Job or Not

Some people prefer to change jobs or professions during their careers. Others choose to stay in the same job or profession. Discuss the advantages of each choice. Which do you prefer? Use reasons and examples to explain your choice.

Prompt 3: Learn from Mistakes

Do you agree or disagree with the following statement? People always learn from their mistakes. Use specific reasons and details to support your answer.

Prompt 4: Method of Travel

You need to travel from your home to a place 40 miles (64 kilometers) away. Compare the different kinds of transportation you could use. Tell which method of travel you would choose. Give specific reasons for your choice.

Prompt 5: Important Plant

Plants can provide food, shelter, clothing, or medicine. What is one kind of plant that is important to you or to the people in your country? Use specific reasons and details to explain your choice.

Appendix III. Score Rubrics by human raters and features by E-rater (Criterion)

Features by Human Rater	Features by E-rater	Score	Items Evaluated
Development	Organization & Development	8	-**Task fulfillment**: Interpretations of prompt -**Appropriateness of Details**: Supporting ideas are relevantly described after main ideas? -**Development**: Extension of Ideas and the length of words (200-300 wds)
Organization		8	-**Organization**: includes Intro, Body, and Conclusion -**Transition Words (TW)**: Connectives are adequate? -**Cohesion**: TWs are adequately used to describe a relationship between ideas? Demonstratives and references words are appropriately employed to refer to previous ideas? -**Coherence**: regular use of superstructures and sequential progression
Vocabulary	Lexical sophistication (Type/token ratio, word length, voca level)	4	-**Range of Vocabulary**: repetition of words, levels of words, variety of words
Sentence variety/ construction	Linguistic Accuracy -Grammatical accuracy ratio -Usage accuracy ratio -Mechanical accuracy ratio -Stylistic accuracy ratio	3	-**Syntactic Variety**: Controlled and Varied sentence structures -**Style**: The style of writing is academic and argumentative?
Grammar/Usage		4	-**Word Choice Errors**: ill-formed verbs, pronoun errors, possessive errors, wrong or missing words, determiner and noun agreement, articles -**Syntax Errors**: run-on sentences, split sentences, subject-verb agreement -**Style**: Words are academically used?
Mechanics		3	-**Word Form Errors**: spellings -**Capitalization, punctuation, and quotation marks in sentences**, and so on.
Total		30	

Appendix IV. Personal Questionnaire (adopted and revised from Moon & Pae, 2011)

NAME_____

1) Please specify the type of English Standard Exam and the score. _____ (ex. 785, TEPS)
2) Have you ever been to English speaking countries (New Zealand, U.S.A., England, Canada, etc.)
   (If YES, go to 2-1. If NO, go to 3.)
   2-1) Please specify the name of country and the length._____, _____
3) Have you ever taken TOEFL test? (If YES, respond to 3-1 & 3-2) Yes No
   3-1) If yes, how many times have you taken TOEFL?
   3-2) What was the score of TOEFL? _____
4) What is your major? _____
5) How old are you? _____
6) Please specify your gender (Male Female)
7) How many hours do you study English every day? _____
8) Have you ever tried TOEFL writing? _____
9) How do you study TOEFL/English writing? _____

Thank you for your participation.

Appendix V. Questionnaire on Criterion Feedback

1. Can you understand the Criterion feedback (FB) clearly and accurately? (If say "disagree," please explain why).

Strongly disagree ① ② ③ ④ ⑤ Strongly agree

Explain why_____

2. Were you generally satisfied with the automated feedback from Criterion FB? (If say "disagree," please explain why).

   Strongly disagree ① ② ③ ④ ⑤ Strongly agree

3. Criterion FB in GRAMMAR was effective and helpful?

   (If say "disagree," please explain why).

   Strongly disagree ① ② ③ ④ ⑤ Strongly agree

4. Criterion FB in USAGE was effective and helpful?

   (If say "disagree," please explain why).

   Strongly disagree ① ② ③ ④ ⑤ Strongly agree

5. Criterion FB in MECHANICS was effective and helpful?

   (If say "disagree," please explain why).

   Strongly disagree ① ② ③ ④ ⑤ Strongly agree

6. Criterion FB in STYLE was effective and helpful?

   (If say "disagree," please explain why).

   Strongly disagree ① ② ③ ④ ⑤ Strongly agree

7. The Criterion FB in the ORGANIZATION/DEVELOPMENT was effective and helpful? (If say "disagree," please explain why).

   Strongly disagree ① ② ③ ④ ⑤ Strongly agree

8. If the FB was not helpful, please explain why. _____

   _____

9. The quality by Criterion FB was good?

   (If say "disagree," please explain why).

   Strongly disagree ① ② ③ ④ ⑤ Strongly agree

10. The accuracy by Criterion FB was good?

    (If say "disagree," please explain why).

    Strongly disagree ① ② ③ ④ ⑤ Strongly agree

Thank you for your participation.

ABSTRACT

Construct Validity in Human Scoring and Criterion:
What Criterion would(not) measure

Jungyeon Koo

This is a pilot study which aims to examine the reliability of automated essay scoring (AES) and to investigate validity of construct that Criterion would/would not measure. Criterion assessed *iBT* TOEFL independent writing tasks by comparing human raters' evaluation. In particular, the current study explored which essay features were most closely related to each of the six different analytic dimensions for *e-rater* (Criterion). Five types of prompts were employed to assign a writing test to fifty college students in Seoul. The result showed that the agreement between human-rater and Criterion is moderate. In addition, three essay features (development, organization, and grammar) were crucial factors to predict the holistic score in human rating. Grammar, however, was a powerful predictor to tell the whole score in AES, which reflects that development and organization were not evaluated appropriately in Criterion. This result suggests that the feature dimensions in *e-rater* need to be refined/revised in the development and organization construct dimensions. The findings have some implications in teaching students process writing and using AES.

Key Words AES, human-scoring, construct dimension, validity, independent writing