Modern English Education, Vol. 17, No. 1, Spring 2016
http://dx.doi.org/10.18095/meeso.2016.17.1.02

A Comparative Analysis of Inter-rater Reliability in English Essay and Korean-English Translation Tests

Lili Park[*]
Hansei University
Inn-Chull Choi
Korea University

Park, Lili & Choi, Inn-Chull. (2016). A comparative analysis of inter-rater reliability in English essay and Korean-English translation tests. *Modern English Education*, *17*(1), 27-48.

Writing performance assessment is gaining momentum in the field of English education in Korea. Writing performance tests, however, are prone to subjective rating that may have adverse effects on the reliability of scores. Previous research recommended reinforced training and clearer scoring criteria. Rater training, however, is costly and demanding in the case of essay tests which require understanding and application of multiple criteria without much guidance on benchmark responses. This study thus sought to examine translation as an alternative writing task that may induce greater score reliability as well as accurate measurement of writing proficiency. Three raters scored the responses of 20 Korean college students to an English to Korean translation test and an English essay test solely based on the rubrics. The results showed that inter-rater reliability was higher for the translation test than for the essay test. In addition, the agreement between two of the raters reached a statistically significant level for the translation test. The implications of these findings are discussed in relation to the viability of implementing translation tests in secondary schools together with suggestions for further study on ways to strengthen the advantages of both essay and translation tests and make up for their weaknesses.

**[writing proficiency/English writing test/English essay/
Korean-English translation/inter-rater reliability/
쓰기 능력/영어 쓰기 시험/영어 논술/한영 번역/채점자 간 신뢰도]**

[*] First author: Lili Park, Corresponding author: Inn-Chull Choi

I. INTRODUCTION

The presence and importance of writing performance testing is growing in the field of English education in Korea. Although multiple-choice tests are known to facilitate the rating process and provide reliable scores, the scope of their measurement is limited in language testing as they can gauge learners' production skills, including writing ability, only via indirect means. Performance testing, on the other hand, enables direct measurement of students' writing ability and simulates a real-life situation in which they are required to use their actual writing skills. It is thus often viewed as a more valid form of measuring production skills, enabling raters to make more trustworthy inferences about learners' writing ability (K. J. Ahn, 2009; Bachman & Palmer, 1996; Y. Chon & D. K. Shin, 2009; Chung & O'Neil, 1997; I. S. Kim & T. J. Park, 2011; Y. A. Lee & S. H. Choi, 2009). In addition, positive washback effects are anticipated from incorporating writing performance into standardized language tests especially in terms of encouraging more communication-based writing instruction in schools (I. C. Choi & Y. H. Uhm, 2001; J. S. Park & M. B. Lee, 2012). Among various types of writing tasks, essay writing has been used widely for standardized language tests in Korea (I. S. Kim & T. J. Park, 2011).

In spite of their positive aspects and growing usage, writing performance tests have often been associated with issues related to difficulty in task development as well as considerable amount of time, cost and effort required for scoring (I. C. Choi & Y. H. Uhm, 2001; K. A. Jin, 2002; I. S. Kim & T. J. Park, 2011). At the forefront of such problems is considered to be the relatively low inter-rater reliability (I. C. Choi & Y. H. Uhm, 2001; McNamara, 1996) as performance testing relies on subjective judgment of educators or raters who can produce divergent rating results depending on what kind of criteria they have and how they apply them (Y. H. Kim, 2013). For example, some raters may give a higher score to well-organized essays with frequent grammatical mistakes than to poorly-organized ones with greater grammatical accuracy, while others may opt to do the opposite. Low reliability is likely to give rise to concerns for fairness in scores, a factor of great significance in high-stakes assessment in Korea where test results are often used for making important decisions related to college admissions and employment (I. S. Kim & T. J. Park, 2011).

Researchers have explored various ways to address the issue of lack of consistency and accuracy in judgment by human raters, which include more effective rating rubrics (Ahmed & Pollitt, 2011; Brown, 2009; Kane, Crooks, & Cohen, 1999; Lind Pantzare, 2015) and reinforced rater training (Alderson, Clapham, & Wall, 2002; Brown, 2003; Lumley & McNamara, 1995; Weigle, 1994, 1998). Administrators of some language proficiency tests, such as the Internet-based version of the Test of English as a Foreign Language (TOEFL iBT), incorporated an automated scoring system into their rating

process. Nevertheless, such automated system is complementary to human judgment and has yet to replace human raters (Enright & Quinlan, 2010).

Researchers keen on inter-rater reliability in writing assessment sought to identify factors contributing to disagreement in scores, and discovered rater variability and bias patterns related to task type, prompts, rating process and rater background. For instance, Y. S. Shin, Y. K. Jong and Y. H. Kim (2010) found some raters to show varying degrees of severity toward different task types, while other researchers, such as He, Gou, Chien, Chen and Chang (2013), noticed bias patterns in relation to raters' area of study. A significant portion of the suggestions put forward by these studies involve improvements in rater training (Eckes, 2008; Fritz & Ruegg, 2013; He, Gou, Chien, Chen, & Chang, 2013; Schaefer, 2008), which, however, is well-known to be costly for test administrators and time-consuming for raters. Also, even the well-trained and reliable raters were found to exhibit differences in the emphasis they give to each rating component (Lumley, 2002).

Other studies looked into some issues related to writing assessment criteria to find explanation for rater variability. While some researchers highlighted the vagueness of the scoring rubrics (Veerappan & Sulaiman, 2012), some others focused on raters' tendency to attach more importance to certain categories of rating (Eckes, 2008, 2012; Schaefer, 2008). Such categories, in many cases, were related to linguistic features of writing (Eckes, 2012; Guo, Crossley, & McNamara, 2013; He et al., 2013; S. K. Lee, 2013; Y. S. Shin, 2010).

Previous research findings highlight the need for a writing performance task that is less demanding, time-consuming and costly, while being able to allow raters to concentrate on fewer and clearer set of evaluation components, especially those related to linguistic features, for a greater level of agreement on scores. In this regard, the present study sought to find an alternative form of writing performance test that can provide reliable scores based on rubrics with minimal rater training, if any.

Korean to English translation test was chosen for investigation in this study as the presence of a source text, which can serve as a benchmark response, may help the scoring process to become less demanding for raters. In addition, it can enable raters to focus on specific scoring categories, such as flow, translation error and linguistic error (J. A. Kim, C. L. Seong, S. W. Lee, H. J. Chang, & H. Lee, 2002), instead of struggling to consider language use with content, organization and other elements of writing all at the same time. The presence of a benchmark response and a more focused set of criteria may reduce the time and cost associated with rater training and, more importantly, help improve inter-rater reliability in writing performance assessment.

For the purpose of this study, novice raters were asked to score students' responses to a Korean to English translation task solely based on rubrics without any prior training. Their ratings for the translation test were then compared in terms of raters' agreement level with those for an essay test under the same topic. Analysis of the study results were guided by

the following research questions:

1) Do raters show statistically significant level of agreement in their assessment of essays written in English?
2) Do raters show statistically significant level of agreement in their assessment of writings translated from Korean to English?
3) Is the level of inter-rater reliability higher for the Korean to English translation test than for the English essay test?

II. LITERATURE REVIEW

1. Rater Variability in Scoring Criteria

Rater variability has been examined by researchers exploring rater reliability in writing assessment (Eckes, 2008, 2012; Weigle, Boldt, & Valsecchi, 2003). Understanding that rater variability may be a potential factor undermining the reliability of writing assessment, researchers have sought to discover the presence and potential sources of the issue by looking into different elements involved in writing performance evaluation including scoring criteria.

One of the major factors causing rater disagreement was found to stem from the different perspectives on the importance of each assessment category. Eckes (2008) conducted an experiment in which raters were asked to indicate how much importance they attach to various criteria in writing performance on a four-point scale. He found that the raters had significantly different views on the importance of each criterion and were far from evenly distributing their attention among the set of evaluation components in focus. Similar findings presented by Schaefer (2008), unveiled that raters, who scored certain categories of content and/or organization severely, rated the other categories of language use and/or mechanics more leniently, and vice versa.

A later study by Eckes (2012) again showed that raters had different perceptions regarding criterion importance and tended to be more severe with the criteria they deemed less significant and more lenient with those they considered more important. In a study examining the effects of rater background on evaluation of ESL student writing (Weigle et al., 2003), raters from English department were more focused on grammar than other rater groups whereas those from psychology department were more keen on content than others.

2. Linguistic Features in Scoring

Linguistic features, especially those related to grammar and use of vocabulary, are viewed by some researchers as the category exerting a great influence on writing scores despite differences in raters' perceived importance of various criteria. According to a study on predicting human judgment of essay quality by Guo et al. (2013), linguistic features were found to be a significant predictor of scores on both integrated and independent writing tasks of TOEFL iBT. Such findings are supported by researchers like K. A. Jin (2002) who claimed that, among performance assessment factors for Korean high school students, sentence structure and grammar had the greatest impact on overall scores.

Linguistic features were also found to particularly attract raters' attention in their feedback to writers. S. K. Lee (2013) revealed that scorers paid more attention to grammar and vocabulary in providing feedback on written texts despite their belief in the importance of comments on content and organization.

3. Use of Translation

Translation, according to Nida and Taber (1982), "consists in reproducing in the receptor language the closest natural equivalent of the source-language message, first in terms of meaning and secondly in terms of style" (p. 12). It requires a good command of both the source and target languages and an ability to adequately use linguistic knowledge in a specific communication situation.

Linguistic features are one of the key elements in translation assessment. After an extensive study on translation quality evaluation components based on the results of three consecutive translation tests, J. A. Kim et al. (2002) suggested flow, translation error and linguistic error as the three main categories for evaluating translation quality. Compared to those used for essay scoring, the translation criteria put forward by the researchers are fewer in number and more focused, which may be helpful in narrowing down raters' differences in understanding and applying writing assessment criteria.

Previous studies discuss how salient linguistic features are in translation. Upon comparing students' direct writing with their translation, B. S. Shin (1998) concluded that translation induces more syntactic complexity overall and greater error frequency among higher-level students. J. A. Kim and C. J. Uhm (2010) also compared direct writing with translated writing in terms of grammatical and discursive features and found positive effects of L1 on organization and vocabulary use in translated texts.

A number of studies reported how L1 is extensively used by second language learners during L2 writing, providing support for utilizing L1 to L2 translation for L2 writing assessment. Cumming (1989) discovered that French ESL learners use their first language

for content generation and word choice when writing in English. In a similar study, Uzawa and Cumming (1989) found that Anglo-Canadian learners of Japanese used English extensively to strategize their writing in Japanese. Cohen and Brooks-Carson (2001) also obtained similar results with French learners who admitted that they often thought in English when writing in French. In a study on the use of L1 as a writing strategy in L2 writing tasks, Y. R. Kim and H. S. Yoon (2014) identified the fact that students used their L1 in their L2 composition regardless of their proficiency levels, and argued that the use of L1 can ameliorate students' L2 writing. Tavakoli, Ghadiri and Zabihi (2014) suggested, while reporting that a majority of the participants in their study thought in their native language of Persian for an English writing task, instruction on translation in class can help learners plan and organize their writing.

The educational benefits of translation have also been explored by researchers, such as D. W. Cho (2009), who presented the positive influence of translation assignments on students' overall writing proficiency. The most noticeable improvements were made in the accuracy of linguistic features. Other researchers, including J. A. Kim and C. J. Uhm (2010), emphasized the benefits of using L1 in L2 writing. Even machine translation was found to encourage learners to produce longer and higher quality writing (Garcia & Pena, 2011). I. C. Choi (2010) also suggested that learners can have a better understanding of both their L1 and L2 by examining and comparing the two languages' deep structure (via meaning-based translation) than just their surface structure (via literal translation). Translation is not a simple world-for-word substitution task as evidenced by various problems associated with machine translation (Y. S. Park, 2002). It requires understanding of the deep structure of the languages involved for meaningful interpretation, which, according to I. C. Choi (2010), can enable learners to achieve greater proficiency in the two languages.

The previously mentioned studies have looked into the possibilities of translation as a tool for rating the writing proficiency of second language learners. The advantages of translation in measuring linguistic accuracy of writing have been explored in depth along with the specific and focused criteria used for translation assessment and the positive educational benefits of translation instruction. A translation task in a writing test may also help alleviate the complexity involved in evaluating other types of writing tasks, for which examiners are required to check for plagiarism or use of exact phrases from the prompt or the presented reading material (Plakans, 2010). However, not much research has been carried out in relation to the potentials of translation as a reliable form of writing performance evaluation and a practical task in terms of time, cost and effort.

III. METHOD

1. Participants

1) Test-takers

The group of test-takers who took part in the study consisted of 20 students from a university in Gyeonggi-do, most of whom were juniors with only four students in their fourth year. Nineteen of them were majoring in English. Despite their similarities in terms of specialty and period of studies in the undergraduate program, the test-takers were at various levels of English proficiency. Their latest TOEIC scores ranged from 550 to 915 with the mean and standard deviation of 776.70 and 110.15, respectively. With the exception of one student, none of the examinees had taken TOEFL in the past.

2) Raters

Three teachers from the same university were employed as raters since the writing tests for this study simulated the TOEFL writing section, for which responses are rated by a maximum of three raters when scores from two raters differ by two points or more (Enright & Quinlan, 2010). Rater 1 and Rater 2 were the only two instructors teaching English writing classes at the university, while Rater 3 was one of the two professional translators among the translation teachers there as well as the only trained rater for the National English Ability Test (NEAT). Table 1 provides some basic information on the raters.

TABLE 1

Information on Raters

Rater	Teaching Subject	Age	Gender	Nationality	Language	Teaching Experience
Rater 1	Writing	38	Female	American	English & Korean	4 years
Rater 2	Writing	36	Female	American	English	3 years
Rater 3	Translation	35	Female	Korean	English & Korean	5 years

2. Procedure

Both the English essay test and the Korean to English translation test were administered consecutively on the same day. With a view to simulating TOEFL iBT, the participating students were required to perform both of the tasks on a computer and submit their writings online. They were asked not to refer to a dictionary or an online website throughout the test sessions. The essay test was given before the translation test so that the students' responses

to the latter would not influence their performance on the former.

The test-takers were allowed 30 minutes to complete each of the tasks. They were advised to write a minimum of 200 words for the first task and to translate the entire contents of a Korean text for the second one.

3. Instruments

1) Tasks

For the purpose of this study, it was imperative that the essay and translation tests involve a high-quality prompt and source text, respectively, and exhibit similarity in terms of difficulty level, desired length and allotted time. Thus, a sample essay question and response for TOEFL, a well-known high-stakes test familiar to many learners, were used to develop the essay and translation tasks, respectively, used for this study. The essay prompt was chosen from among the sample questions for TOEFL's independent writing task that are available on the ETS website (see Appendix 1).

With regard to the translation task, it was assumed that a better comparison would be made between the two tests if the students were asked to translate a text that is under the same topic as the essay prompt. Therefore, one of the sample responses to the essay question used for this study was selected from a set of level 5 (the highest level) benchmark examinee responses provided on the ETS website. The measure was taken based on the assumption that a successful translation of the response will lead to a perfect score. The text was then translated into Korean, and given to the test-takers to be translated back into English. The chosen sample response was translated into Korean by a professional translator for the sake of accuracy and authenticity of the Korean text. It was then edited and shortened to contain slightly over 200 words (see Appendices 2 and 3). The aim of this measure was to induce the students to produce writings that are similar in length as their essays for which the recommended minimum number of words was also 200.

Using a sample TOEFL essay writing for the translation test had a few advantages. It made it possible to anticipate the length of translation to be produced by the test-takers and the amount of time required to complete the task. It also enabled the original English text to be served as a benchmark response for raters who do not understand Korean.

2) Rubrics and Sample Responses

With respect to the essay test, raters were asked to refer to the independent writing rubrics and sample responses for TOEFL iBT during the scoring process. The sample

responses consisted of model essay writings for each of the score level ranging from 0 to 5 . Both of the materials are readily available on the ETS website. No such materials were provided for the translation test as there is no state certified translation examination with sample responses nor nationally accepted translation rubrics. For the sake of consistency and validity, the integrated writing rubrics for TOEFL iBT were edited and modified to serve as the rubrics for the translation task (see Appendices 4 and 5). The rubrics for integrated writing were more adequate than those for independent writing as the former include criteria on the degree of relevance to the information on reading and listening materials, which could easily be altered into the criteria for the degree of relevance to the source text in the case of the translation test. The translation rubrics were thus modified to contain the criterion on a response's connection to the source text in the place of that to the reading and listening materials as well as the three translation quality assessment components suggested by J. A. Kim et al. (2002), which include flow, translation error and linguistic error.

3) Interview

After all the scores were submitted, each of the raters was interviewed separately in person with regard to their rating behavior (see Appendix 6 for interview questions). Six open-ended questions were posed in relation to the degree of importance they attach to each evaluation category and the level of severity with any of the two tests or criteria. Their comments were entered into a computer on the spot and referred to for analyzing their behavior during assessment. Their comments during the interview are discussed together with the study results in the following section of this paper.

4. Rating

The raters scored all of the test-takers' writings, which consisted of 20 essays and 20 translated texts. For valid results, all the submitted writings were identified only by their assigned number during the rating process, and no information on the expected outcome of the experiment was provided to any of the participants. As regards the translation task, both Korean and English versions of the source text were given to the scorers so that they can refer to the one in either of the two languages they are more comfortable with. All of the raters were asked to score each writing holistically on a scale of 0 to 5 based on the rubrics for each task. The holistic approach was adopted instead of the analytic method as the TOEFL rubrics adopted for the study are intended for holistic assessment. In addition, the experiment itself simulated the TOEFL iBT writing section in many aspects including the use of two different tasks, questions, time limit and desired length. In order to obtain

results purely based on the rubrics, the raters were not asked to take part in any training sessions for either essay test or translation test.

5. Statistical Analysis

The level of agreement between the raters was measured and analyzed via IBM-SPSS Statistics (version 22.0). Pearson's correlation coefficient was used to account for the degree of conformity between the raters in both essay and translation tests.

IV. RESULTS AND DISCUSSION

The descriptive statistics on the results of essay and translation tests did not show any considerable differences in terms of minimum/maximum score, mean score and standard deviation (see Tables 2 and 3). While the minimum score for the translation test was 2 on a scale of 0 to 5 for all raters, 1 was the lowest score for Rater 2 in the essay test. The maximum score was 4 for all raters in both tests. The translation test produced a slightly higher mean score for Rater 1 and Rater 3 and a slightly lower standard deviation mark, but was still very similar to the essay test in both statistical indices.

TABLE 2
Descriptive Statistics for Essay Test

	n	Minimum	Maximum	Mean	*SD*
Rater 1 Essay Score	20	2	4	2.90	0.85
Rater 2 Essay Score	20	1	4	2.95	0.83
Rater 3 Essay Score	20	2	4	2.70	0.73

TABLE 3
Descriptive Statistics for Translation Test

	n	Minimum	Maximum	Mean	*SD*
Rater 1 Translation Score	20	2	4	3.00	0.72
Rater 2 Translation Score	20	2	4	2.95	0.76
Rater 3 Translation Score	20	2	4	2.95	0.69

1. Inter-rater Reliability for Essay Test

Table 4 represents the degree of correspondence among raters in their scores for the essay test. The agreement level was the highest between Rater 1 and Rater 2 and the lowest between Rater 2 and Rater 3. The correlation between Rater 1 and Rater 3 and between Rater 2 and Rater 3 stood at the 0.4 level. Even the highest figure of 0.59, which sits just

below the 0.6 level, is lower than the widely accepted correlation of 0.7, and thus falls short of being considered statistically significant. The results show that the raters failed to reach a statistically significant level of agreement with one another in their assessment of essays.

TABLE 4

Inter-rater Reliability for Essay Test

		Rater 1 Essay Score	Rater 2 Essay Score	Rater 3 Essay Score
Rater 1	Pearson Correlation	1	.59	.45
Essay	*p*		.01	.04
Score	*n*	20	20	20
Rater 2	Pearson Correlation	.59	1	.41
Essay	*p*	.01		.07
Score	*n*	20	20	20
Rater 3	Pearson Correlation	.45	.41	1
Essay	*p*	.04	.07	
Score	*n*	20	20	20

During the interview after the scoring process, the three raters agreed that all of the criteria for the essay test should be given fair amount of attention. Their comments in the later part of the interview, however, revealed that the most valued criterion was different among them with Rater 1 valuing structure and organization, Rater 2 concentrating more on topical relatedness and Rater 3 placing greater importance on organization and linguistic features. Such different views of the raters regarding the focus of their assessment uphold the arguments of researchers, such as Eckes (2008, 2012) and Schaefer (2008), who claimed that scorers hold divergent perceptions on the importance of each evaluation category, which emphasizes the need for discussions on the relative importance of each criterion.

2. Inter-rater Reliability for Translation Test

Table 5 shows that the level of translation score agreement between Rater 2 and Rater 3 was significant at 0.80, which is considerably higher than the widely accepted agreement level of 0.70. However, Rater 1's scores did not correlate sufficiently with those of either Rater 2 or Rater 3. The agreement level between Rater 1 and Rater 2 stood at 0.67, falling slightly short of the desired level of 0.70. The correlation between Rater 1 and Rater 3 was lower at 0.53. The results provide only partial support for the second hypothesis of the present study that the raters will show statistically significant level of agreement in their assessment of Korean to English translation test since the inter-rater reliability was meaningfully high between only two raters.

TABLE 5

Inter-rater Reliability for Translation Test

		Rater 1 Translation Score	Rater 2 Translation Score	Rater 3 Translation Score
Rater 1	Pearson Correlation	1	.67	.53
Translation	*p*		.00	.02
Score	*n*	20	20	20
Rater 2	Pearson Correlation	.67	1	.80**
Translation	*p*	.00		.00
Score	*n*	20	20	20
Rater 3	Pearson Correlation	.53	.80**	1
Translation	*p*	.02	.00	
Score	*n*	20	20	20

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

Rater 1 commented during the interview about how she found out, during the scoring process, that she tended to be more severe with the translation test than with the essay test and thus had to revise her scores a few times before submission. The rater's tendency to be more severe with the translation task may be attributable to the presence of the source material. According to Guo et al. (2013), "raters assess grammatical mistakes more severely when source materials are provided because the content is readily available, unlike in independent essays where writers have to struggle with content and language simultaneously" (p. 233). As suggested by the scholars, Rater 1, who admitted to having valued structure and organization more than other elements when grading essays, remarked that she paid greater attention to errors related to grammar and use of words/expressions during translation assessment.

3. Comparison of Inter-rater Reliability in Essay and Translation Tests

Analysis of the rater agreement indices for the two tests confirm this study's third hypothesis that the inter-rater reliability will be higher for the Korean to English translation test than for the English essay test. The highest level of agreement reached 0.59 for the essay test and 0.80 for the translation test. The former's lowest figure recorded 0.40 whereas the latter's marked 0.53, which is slightly below the former's highest level of 0.59. A comparison between the results of both tests indicates that the correlation between any two raters is considerably higher for the translation test than for the essay test.

V. CONCLUSION AND IMPLICATIONS

The present study explored the potentials of Korean to English translation test as a

reliable writing task for measuring examiners' writing performance. With the focus on inter-rater reliability, ratings on a Korean to English translation test were analyzed and compared with those on an English essay test, which is a more widely used means for judging writing proficiency.

Three research questions were posed in relation to how significant inter-rater reliability is for essay and translation tests, respectively, and whether the latter's figure is higher than the former's. According to the findings of this study, the raters failed to reach an adequate level of agreement for the essay test while only one them did not correlate sufficiently with either of the other two raters in the translation test. The other two raters produced a correlation of 0.80, which is a significantly high level considering the fact that they had no prior training on scoring. The inter-rater reliability between any two raters was higher for the translation task than for the essay task.

The results of this study highlight the advantage of translation test over essay test in terms of inter-rater reliability, which has constituted major problems with performance testing in secondary education. The relatively low inter-rater reliability figure for the essay test may be attributable to the fact that novice raters were recruited without any prior training. This may be interpreted as an indication that scores on an essay test are more dependent on rater training than those on a translation test and that minimal training in translation assessment could bring sufficiently reliable scores in general.

Another advantage of translation test could be found in its greater practicality over its counterpart. As previously mentioned, the presence of a source text and a more focused set of criteria can facilitate the rating process and lessen the time, cost and effort required for training raters. Administering a translation test can also alleviate the concerns related to the fairness of judgment as the source text can be utilized as a benchmark response with which students can compare their responses. Thus, a minimal rater training on scoring rubrics will enable secondary school teachers to utilize translation task as a viable approach to performance testing in English classes.

As aforementioned, translation can lend itself to a robust tool for measuring grammatical and lexical accuracy as well as appropriateness of language use. However, coherence in writing, which is considered to be an important criterion in accordance with the componential model of language ability (Bachman, 1990; Bachman & Palmer, 1996), was difficult to measure for the translation test in this study as students' responses basically followed the logic and organization of the source text. This signifies that translation tests are more adequate for large-scale high-stakes performance assessment that is administered to measure test-takers' English writing skills, rather than their overall writing ability, in a reliable and practical fashion. For measurement of other components, such as coherence, through translation assessment, further study is needed to systematically analyze and compare the rating criteria of essay and translation tests through the use of analytic rubrics

with a larger number of subjects in secondary schools. Such research findings are expected to shed light on exploring ways to develop a writing task that can strengthen the advantages and make up for the weaknesses of both types of tests by enabling reliable and practical assessment and also facilitating accurate measurement of coherence or rhetorical organization in addition to linguistic features.

REFERENCES

Ahmed, A., & Pollitt, A. (2011). Improving marking quality through a taxonomy of mark schemes. *Assessment in Education: Principles, Policy & Practice, 18*(3), 259-278.

Ahn, Kyung-Ja. (2009). Learning to write academic English: A sociocultural perspective on L2 learners' writing portfolio. *Korean Journal of Applied Linguistics, 25*(2), 1-34.

Alderson, J., Clapham, C., & Wall, D. (2002). *Language test construction and evaluation.* Cambridge: Cambridge University Press.

Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford: Oxford University Press.

Bachman, L. F., & Palmer, A. (1996). *Language testing in practice*. Oxford: Oxford University Press.

Brown, A. (2003). Interviewer variation and co-construction of speaking proficiency. *Language Testing, 20,* 1–25.

Brown, G. T. L. (2009). The reliability of essay scores: The necessity of rubrics and moderation. In S. D. L. H. Meyer, H. Anderson, R. Fletcher, P. M. Johnston, & M. Rees (Eds.), *Tertiary assessment and higher education student outcomes: Policy, practice and research* (pp. 40-48). Wellington, New Zealand: Ako Aotearoa.

Cho, Dong-Wan. (2009). A study on a college English writing course centering on translation assignments and student presentations. *Foreign Languages Education, 16*(1), 261-289.

Choi, Inn-Chull. (2010). *Silyong Yeongmunbeob Baeggwasajeon* [Practical English grammar encyclopedia] (2nd ed.). Seoul: SaramIn Publishing Co.

Choi, Inn-Chull, & Uhm, Yeon-Hee. (2001). Comparative analyses of computer-based ratings and human ratings based on performance testing of writing. *Multimedia-Assisted Language Learning, 4*(1), 165-184.

Chon, Yuah Vicky, & Shin, Dong-Kwang. (2009). Collocations in L2 writing and rater's perceived writing proficiency. *Korean Journal of Applied Linguistics, 25*(1), 101-129.

Chung, G., & O'Neil, H. (1997). *Methodological approaches to online scoring of essays.*

(CSE Technical Report 461). Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Testing.

Cohen, A. D., & Brooks-Carson, A. (2001). Research on direct versus translated writing: Students' strategies and their results. *The Modern Language Journal, 85*(2), 169-188.

Cumming, A. (1989). Writing expertise and second language proficiency. *Language Learning, 39,* 81-141.

Eckes, T. (2008). Rater types in writing performance assessments: A classification approach to rater variability. *Language Testing, 25*(2), 155-185.

Eckes, T. (2012). Operational rater types in writing assessment: Linking rater cognition to rater behavior. *Language Assessment Quarterly, 9*(3), 317-334.

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

Fritz E., & Ruegg, R. (2013). Rater sensitivity to lexical accuracy, sophistication and range when assessing writing. *Assessing Writing, 18*(2), 173-181.

Garcia, I., & Pena, M. I. (2011). Machine translation-assisted language learning: Writing for beginners. *Computer Assisted Language Learning: An International Journal, 24*(5), 471-487.

Guo, L., Crossley, S. A., & McNamara, D. S. (2013). Predicting human judgments of essay quality in both integrated and independent second language writing samples: A comparison study. *Assessing Writing, 18*(3), 218-238.

He, T. H., Gou, W. J., Chien, Y. C., Chen, I. S. J., & Chang, S. M. (2013). Multi-faceted Rasch measurement and bias patterns in EFL writing performance assessment. *Psychological Reports, 112*(2), 469-485.

Jin, Kyung-Ae. (2002). Applying standards-based English performance assessment in Korean high schools. *Studies in Modern Grammar, 28,* 231-244.

Kane, M. T., Crooks, T., & Cohen, A. (1999). Validating measures of performance. *Educational Measurement: Issues and Practice, 18*(2), 5-17.

Kim, In-Sook, & Park, Tae-Joon. (2011). Paper-based versus image-based scoring of student essays. *Korean Journal of Applied Linguistics, 27*(2), 231-256.

Kim, Jeong-A, & Uhm, Chul-Joo. (2010). Grammatical and discursive features of Korean EFL learners' direct, translated and back-translated writing. *Linguistic Research, 27*(2), 373-392.

Kim, Jin-Ah, Seong, Cho-Lim, Lee, San-Won, Chang, Hyun-Ju, & Lee, Hyang. (2002). A study on quality assessment criteria in English-Korean translation. *Studies in Foreign Literature, 11,* 85-123.

Kim, Youn-Hee. (2013). The effect of different types of rating scales on ESL writing performance assessment. *Korea Journal of English Language and Linguistics,*

13(3), 465-496.

Kim, Young-Ran, & Yoon, Hyun-Sook. (2014). The use of L1 as a writing strategy in L2 writing tasks. *GEMA Online Journal of Language Studies, 14*(3), 33-50.

Lee, Sang-Ki. (2013). Different rater groups and their feedback on written texts: Where do they focus their attention? *Journal of the Korean English Education Society, 12*(1), 57-76.

Lee, Yo-An, & Choi, Seong-Hee. (2009). Reeavluating advanced speaking proficiency for Korean teachers of English. *Korean Journal of Applied Linguistics, 25*(3), 303-336.

Lind Pantzare, A. (2015). Interrater reliability in large-scale assessments: Can teachers score national tests reliably without external controls? *Practical Assessment, Research & Evaluation, 20*(9). Retrieved from the World Wide Web: http://pareonline.net/getvn.asp?v=20&n=9.

Lumley, T. (2002). Assessment criteria in a large-scale writing test: What do they really mean to the raters? *Language Testing, 19*(3), 246-276.

Lumley, T., & McNamara, T. (1995). Rater characteristics and rater bias: Implications for training. *Language Testing, 12,* 54–71.

McNamara, T. (1996). *Measuring second language performance*. London, UK: Longman.

Nida, E. A., & Taber C. R. (1982). *The theory and practice of translation.* Leiden: Koninklijke Brill NV.

Park, Ji-Seon, & Lee, Moon-Bok. (2012). A report on the teaching and learning of speaking and writing in high school English classrooms: Transiting to NEAT. *Modern English Education, 13*(2), 121-149.

Park, Young-Soon. (2002). The basic theory of interpretation and translation studies. *Modern English Education, 3*(2), 99-134.

Plakans, L. (2010). Independent vs. integrated writing tasks: A comparison of task representation. *TESOL Quarterly, 44*(1), 185-194.

Schaefer, E. (2008). Rater bias patterns in an EFL writing assessment. *Language Testing, 25*(4), 465-493.

Shin, Bong-Soo. (1998). Experimental study on the effects of Korean on English writing: Direct writing versus translation writing. *Studies in Language, 14,* 127-142.

Shin, You-Sun. (2010). A FACETS analysis of rater characteristics and rater bias in measuring L2 writing performance. *English Language & Literature Teaching, 16*(1), 123-142.

Shin, You-Sun, Jong, Young-Kyong, & Kim, Yang-Hee. (2010). Investigating variability in tasks, analytic scoring criteria and rater groups on L2 writing in English. *Foreign Languages Education, 17*(2), 25-57.

Tavakoli, M., Ghadiri, M., & Zabihi, R. (2014). Direct versus translated writing: The effect of translation on learners' second language writing ability. *GEMA Online Journal*

of Language Studies, 14(2), 61-74.

Uzawa, K., & Cumming, A. (1989). Writing strategies in Japanese as a foreign language: Lowering or keeping up the standards. *The Canadian Modern Language Review, 46,* 178-194.

Veerappan, V., & Sulaiman, T. (2012). A review on IELTS writing test, its test results and inter rater reliability. *Theory & Practice in Language Studies, 2*(1), 138-143.

Weigle, S. (1994). Effects of training on raters of ESL compositions. *Language Testing, 11*(1), 197–223.

Weigle, S. (1998). Using FACETS to model rater training effects. *Language Testing, 15,* 263–287.

Weigle, S. C., Boldt, H., & Valsecchi, M. I. (2003). Effects of task and rater background on the evaluation of ESL student writing: A pilot study. *TESOL Quarterly, 37*(2), 345-354.

APPENDIX 1

Essay Task

Do you agree or disagree with the following statement?

A teacher's ability to relate well with students is more important than excellent knowledge of the subject being taught.

Use specific reasons and examples to support your answer (30 minutes; 200 words or more).

APPENDIX 2

Translation Task (Translated Korean Text)

Translate the following text into English (30 minutes).

나는 유치원 이후로 나를 가르쳤던 모든 선생님을 기억하고 있다. 만약에 초등학교 1학년 선생님이 누구였는지 알고 싶은 친구가 있다면, 내게 묻기만 하면 된다. 어린아이였던 내 눈에는 선생님들이 모두 친절하고 이해심 있는 분들로 보였다. 그분들은 거의 모든 학생들과 각각 각별한 사이였고 모든 사람들에게 다정했다. 이것이 내가 그분들을 모두 기억하는 이유이다.

교사의 주된 목표는 학생들에게 우리 교과서에 있는 내용과, 더 중요하게는, 서로를

존중하는 방법을 최선을 다해서 가르치는 것이다. 교사들은 모든 사람과 잘 지내면서 더 나은 삶을 사는 방법을 우리에게 가르쳐준다. 그렇게 하기 위해서는 교사들 스스로가 학생들과 잘 교감할 수 있어야 한다.

교사의 주된 목표는 학생들이 서로에게 존중을 표하는 방법을 최선을 다해서 가르치는 것이므로 교사들은 가르칠 때 여러 가지 접근법을 사용한다. 가르치는 과목에 대한 지식은 부차적이다. 이러한 이유에서, 나는 학생들과 잘 교감할 수 있는 능력에 비해 가르치는 과목에 대한 탁월한 지식은 부차적이라는 것을 자신 있게 주장하는 바이다.

APPENDIX 3

Translation Task (Original English Text)

I remember every teacher that has taught me since I was in Kindergarten. If a friend wants to know who our first grade teacher was in elementary school, all they have to do is ask me. The teachers all looked very kind and understanding in my eyes as a child. They had special relationships with nearly each and every one of the students and were very nice to everyone. That's the reason I remember all of them.

A teacher's primary goal is to teach students the best they can about the things that are in our textbooks and more important, how to show respect for one another. They teach us how to live a better life by getting along with everyone. In order to do that, the teachers themselves have to be able to relate well with students.

A teacher's primary goal is to teach students the best they can about how to show respect for one another, so teachers use different approaches when teaching, and knowledge of the subject being taught is secondary. For these reasons, I claim with confidence that excellent knowledge of the subject being taught is secondary to the teacher's ability to relate well with their students.

APPENDIX 4

English Essay Rubrics

(From Independent Writing Rubrics of TOEFL iBT® Test)

Score	Task Description
5	An essay at this level largely accomplishes all of the following: ■ Effectively addresses the topic and task ■ Is well organized and well developed, using clearly appropriate explanations, exemplifications and/or details ■ Displays unity, progression and coherence ■ Displays consistent facility in the use of language, demonstrating syntactic variety, appropriate word choice and idiomaticity, though it may have minor lexical or grammatical errors

4	An essay at this level largely accomplishes all of the following: ■ Addresses the topic and task well, though some points may not be fully elaborated ■ Is generally well organized and well developed, using appropriate and sufficient explanations, exemplifications and/or details ■ Displays unity, progression and coherence, though it may contain occasional redundancy, digression, or unclear connections ■ Displays facility in the use of language, demonstrating syntactic variety and range of vocabulary, though it will probably have occasional noticeable minor errors in structure, word form or use of idiomatic language that do not interfere with meaning
3	An essay at this level largely accomplishes all of the following: ■ Addresses the topic and task well, though some points may not be fully elaborated ■ Is generally well organized and well developed, using appropriate and sufficient explanations, exemplifications and/or details ■ Displays unity, progression and coherence, though it may contain occasional redundancy, digression, or unclear connections ■ Displays facility in the use of language, demonstrating syntactic variety and range of vocabulary, though it will probably have occasional noticeable minor errors in structure, word form or use of idiomatic language that do not interfere with meaning
2	An essay at this level may reveal one or more of the following weaknesses: ■ Limited development in response to the topic and task ■ Inadequate organization or connection of ideas ■ Inappropriate or insufficient exemplifications, explanations or details to support or illustrate generalizations in response to the task ■ A noticeably inappropriate choice of words or word forms ■ An accumulation of errors in sentence structure and/or usage
1	An essay at this level may reveal one or more of the following weaknesses: ■ Limited development in response to the topic and task ■ Inadequate organization or connection of ideas ■ Inappropriate or insufficient exemplifications, explanations or details to support or illustrate generalizations in response to the task ■ A noticeably inappropriate choice of words or word forms ■ An accumulation of errors in sentence structure and/or usage
0	An essay at this level merely copies words from the topic, rejects the topic, or is otherwise not connected to the topic, is written in a foreign language, consists of keystroke characters, or is blank.

APPENDIX 5

Korean to English Translation Rubrics
(Adapted from Integrated Writing Rubrics of TOEFL iBT® Test)

Score	Task Description
5	A translation at this level successfully conveys the information from the original text and coherently and accurately presents this information. The translation is well organized, and occasional language errors that are present do not result in inaccurate or imprecise presentation of the content.
4	A translation at this level is generally good in conveying the information from the original text and in coherently and accurately presenting this information, but it may have minor omission, inaccuracy, vagueness, or imprecision of some content from the original text. A translation is also scored at this level if it has more frequent or noticeable minor language errors, as long as such usage and grammatical structures do not result in anything more than an occasional lapse of clarity or in the connection of ideas.
3	A translation at this level conveys important information from the original text, but it is marked by one or more of the following: ■ The translation may omit one major key point made in the original text. ■ Some of the points made in the original text may be incomplete, inaccurate, or imprecise. ■ Errors of usage and/or grammar may be more frequent or may result in noticeably vague expressions or obscured meanings in conveying ideas and connections.
2	A translation at this level conveys relevant information from the original text, but is marked by considerable language difficulties or by considerable omission or inaccuracy of ideas from the original text; a translation at this level is marked by one or more of the following: ■ The translation considerably omits or considerably misrepresents some points made in the original text. ■ The translation contains language errors or expressions that largely obscure connections or meaning at key junctures or that would likely obscure understanding of key ideas for a reader not already familiar with the original text.
1	A translation at this level is marked by one or more of the following: ■ The translation provides little or no meaningful or relevant coherent content from the original text. ■ The language level of the translation is so low that it is difficult to derive meaning.
0	A translation at this level rejects the original text, is written in a foreign language, consists of keystroke characters, or is blank.

APPENDIX 6

Interview Questions

1. Do you think all of the criteria for writing assessment should be given an equal weight? Please specify the reasons for your answer.
2. Did you feel that the responses to either of the two tests were easier or more difficult to score? If you did, please specify the form of the test together with some possible reasons.
3. Did you feel that certain criteria were easier or more difficult to measure than others in either or both of the two tests? If you did, please specify the criteria together with some possible reasons
4. Did you tend to be more generous or severe with either of the two tests? If you did, please specify the form of the test together with some possible reasons.
5. Did you tend to be more generous or severe with certain criteria in either or both of the two tests? If you did, please specify the criteria together with some possible reasons.
6. Are there any criteria that attracted more of your attention than others in assessing either or both of the two tests? If there are, please specify the criteria together with some possible reasons.

Examples in: English
Applicable Languages: English
Applicable Levels: Secondary/Tertiary

Lili Park
Dept. of International Language
Hansei University
30 Hanse-ro, Gunpo-si, Gyeonggido
Tel: 031-450-5030
Email: lilipark@hanmail.net

Inn-Chull Choi
Dept. of English Language Education
Korea University
145 Anam-ro, Seongbuk-gu, Seoul, Korea
Tel: 02-3290-2358
Email: icchoi@korea.ac.kr

Received 15 December 2015
Revised 27 January 2016
Accepted 12 February 2016