Investigating Reliability of Re-modified Scoring Rubrics for EFL Paraphrasing Task

Minkyung Kim
(**Seoul National University**)

Kim, Minkyung. 2018. Investigating Reliability of Re-modified Scoring Rubrics for EFL Paraphrasing Task. *SNU Working Papers in English Linguistics and Language 16, 36-56.* Paraphrasing is one of crucial writing skills not only for native writing but also for L2 writing. Therefore, it gains importance as a source of a writing performance test in EFL writing. However, the types of scoring rubrics for paraphrasing tasks used in previous studies highly vary indicating using their own developed scales. This implies that there is a research gap that defining reliable scoring rubrics for paraphrasing task is necessary. Thus, this current study would like to shed light on investigating reliability of the re-modified scoring scales for the EFL paraphrasing data collected by 20 participants in the proficiency of advanced and intermediate group. Furthermore, as syntactic and word change are some of the key elements considered in paraphrasing, syntactic and lexical complexity of the data were also measured using programs, LCA and L2SCA (Ai & Lu, 2010). The results turned out that there is significant correlation and reliability of the scoring rubrics which are re-designed for this study, but no meaningful conclusions in terms of syntactic and lexical complexity following the results of previous studies (Engber, 1995; Linnarud, 1986 and Lu, 2015). **(Seoul National University)**

Keywords: language assessment, writing assessment, paraphrasing, reliability of scoring rubrics, correlation coefficients, syntactic and lexical complexity.

1. Introduction

Writing, one of main language abilities, is an important area to be assessed. There are a flood of materials designed to test L2 learners' writing ability. However, the types of writing performance tasks are extremely restricted for both EFL students and teachers in Korea (Ji, 2018). Thus, efforts to develop and find various kinds of writing performance test have been carried out. As a noble tool for measuring L2 writing ability, a paraphrasing task has been employed recently. Furthermore, previous studies on paraphrasing tasks indicate that these

materials can be used as a valuable source of assessing L2 writing ability. Likewise, Ji (2018) claimed that paraphrasing tasks are valid as a test for evaluating EFL learners' writing ability.

However, a serious problem concerning paraphrasing tasks has arisen. As a standard criterion for scoring EFL paraphrasing tasks isn't defined yet, the scores are graded depending on their researchers' or teachers' own scales. Each researcher conducting a study on paraphrasing tasks came up with their own scoring rubric. Therefore, a proper evaluation standard should be developed for all EFL paraphrasing data written by L2 students.

Meanwhile, M.-H. Chen et al. (2015) focused on developing a corpus-based paraphrase program for Chinese EFL learners to let them improve their writing skills. While other researches working on paraphrasing tasks, they modified the TOEFL's Integrated Writing Rubrics to grade the paraphrasing data collected from their participants. They edited the scoring rubric from TOEFL test because paraphrasing is an important ability in the writing section of the exam. Unlike other researchers creating their own scoring rubrics for paraphrasing task while adapting some critical elements from previous studies, they took the advantage of the availability of the TOEFL's scoring rubrics. This was highly impressive as the scoring rubrics from TOEFL are accessible and already valid as they are designed by ETS and currently used in the field.

The current study firstly planned to shed light on the modified rubrics from Chen et al. (2015). While in the rating procedure, it turned out that still the modified version of scoring rubric is not clear enough for the hired raters to assess the paraphrasing data from the participants. Thus, further modification was conducted for clearer rating procedure. After the re-modification of the scoring rubrics with key elements that are usually assessed in paraphrasing tasks, the raters finished grading on the data. Based on the scores, investigation of the reliability of the re-modified scoring rubric was held.

In addition to evaluating the reliability, syntactic and lexical complexities were measured to see whether a correlation between the complexity and the rated scores exists or not. The syntactic and lexical complexities on L2 writing has garnered a lot of attention as they show how the performing task is elaborated and varied and the degree of sophistication of the structures used in writings (Lu & Ai, 2015). Thus, the relationship between the syntactic and lexical complexity and paraphrasing task was also investigated. Consequently, the correlation and reliability of the re-modified scoring rubrics show significant results whereas there was no meaningful result on complexity.

2. Literature Review

2.1 Paraphrasing Task

In the field of both native and non-native English writing, paraphrasing is an essential writing skill (Keck, 2006). Moreover, paraphrasing can be considered as an important ability not only for clarifying ideas for essays and improving memory (Reid, Lienemann, & Hagamann, 2013), but also for avoiding plagiarism (Keck, 2006). Therefore, a lot of researches have been conducted on paraphrasing task, and the area has been expanded even to EFL writing. With some limited sources of assessing EFL learners' writing ability, Ji (2018) investigated the validity of paraphrasing task as a new type of test for adequately figuring out Korean L2 learners' writing skills. To test the validity of the paraphrasing tasks, the study associated the paraphrasing tasks with the self-assessments of L2 learners. If the EFL learners' self-assessment scores have correlation with the scores of the paraphrasing tasks, it means that the paraphrasing tasks are valid to verify the L2 learners' language performance. This assumes that the learners' self-assessments are reliable based on the previous studies (Fitzgerald, White, & Gruppen, 2003).

To investigate validity of the paraphrasing task, 364 test takers ranging from grade 7 to university students were collected since the study also planned to figure out which paraphrasing task would be the most adequate one for each grade level. Among the 364 participants, 111 middle school and 169 high school students took the self-assessment task before they filled in the paraphrasing tasks whereas 80 university freshmen's language abilities were scored by TOEIC. The paraphrasing task performed in Ji was composed of three parts: (1) four items for gap-filling paraphrased sentences, (2) three items for partial paraphrasing (3) three items for entire paraphrasing. The division of the task was in the necessity of exploring the most suitable task format for each grade level.

The paraphrasing tasks and scoring rubrics used in Ji (2018) were invented by recruited school teachers. The results have proven that the self-assessed scores and the scores of paraphrasing writing tasks have correlation coefficients implying the validity of the paraphrasing task as a test item. For the results, the middle and high school groups yielded statistically significant yet weak or moderate correlations between their paraphrasing ability and self-assessed English ability. On the other hand, the college students showed a high correlation between their paraphrasing ability and language proficiency measured by TOEIC. This suggests that paraphrasing task can reflect test-takers' English proficiency although the degree of reflection varies among groups implying that paraphrasing task might be more appropriate for the university students who are relatively more familiar with writing essays. Even though there were some different results among the groups, the research found that paraphrasing task has a potential as a valid writing test item.

Despite figuring out the potentiality of paraphrasing task, the study employed its own scoring rubrics invented by some teachers and researchers, which questions the validity and reliability of the scoring rubrics.

2.2 Scoring Rubrics for Paraphrasing Task

In terms of scoring rubrics, unlike other researchers who invented their own rubrics for paraphrasing tests, M. -H. Chen et al. (2015) modified the TOEFL's Integrated Writing Rubrics to be suitable for the task when grading the paraphrases gathered from 64 participants. The scoring rubric is composed of a 10-point grading scale ranging from 0 to 9 within 5 levels. For each level, two points were given to allow the raters some freedom to better paraphrasing performance. The key elements in this rubric were consisted of three parts, single-word replacing, phrase replacing, and sentence restructuring. Furthermore, the most considered factor in grading was not the number of the phrases changed, but the quality of the paraphrasing. Modifying the existing TOEFL's Integrated Writing Rubrics to be scoring rubric for paraphrasing task is a meaningful work as rubrics from TOEFL are already reliable, valid and easily available online.

However, these advantages of the modified rubrics could not help the raters on this current study to grade the paraphrasing task readily. Still the rubrics were vague and inappropriate in the process of scoring the material in this study, re-modification was inevitable.

With the research gap that the reliability and validity of scoring rubrics applied on paraphrasing tasks on previous studies are not defined yet, this present research would like to mainly focus on evaluating the reliability of the re-modified version of the scoring rubrics from Chen et al.

2.3 Syntactic and Lexical Complexity

Not only with the analyzing reliability between raters, but also measuring syntactic and lexical complexity of EFL essays has been conducted on previous studies. Lu & Ai (2015) investigated syntactic complexity in L2 writings from college students with diverse

backgrounds. They compared essays from non-native speakers with those from native speakers collected by corpus of ICLE 2.0 (International Corpus of Learner English) and LOCNESS (Louvain Corpus of Native English Essays). The syntactic complexity was measured with L2 Syntactic Complexity Analyzer (L2SCA), a program developed by Lu (2010). This program analyzes syntactic complexity with 14 indices, such as length of production unit, amount of subordination, amount of coordination, degree of phrasal sophistication, and overall sentence complexity.

In respect of lexical complexity, it can be investigated using Lexical Complexity Analyzer, a tool that allows language teachers and researchers to analyze lexical complexity of written English language samples, using 25 different indices of lexical density, variation and sophistication proposed in the first and second language development literature (Ai & Lu, 2010). Lu (2012) shed light on the relationship of lexical richness of the quality of ESL leaners' oral narratives using this program. The data in this study were selected from the Spoken English Corpus of Chinese Learners (Wen et al, 2005). While measuring lexical complexity of the L2 learners' narratives, the relationship between the raters' judgement and the figures of lexical density was also included in one of the research questions. However, it turned out that there is no significant correlation between them in the data. The previous researches on L2 writing also suggested that the proportion of lexical words in an oral narrative does not appear to have a relationship with the quality of spoken data (Engber, 1995; Linnarud, 1986).

Nevertheless, as the key elements of paraphrasing task is syntactic and word change, investigating correlation between the raters' scores and the figures of syntactic and lexical complexity measured from the paraphrasing data would be meaningful in this present study.

To fill some of research gaps spotted in the previous studies, three research questions are addressed:

(1) Do the scores graded following the re-modified scoring rubric for

paraphrasing task have correlation coefficients?
(2) Does the re-modified scoring rubric for paraphrasing task have inter-rater reliability?
(3) To what extent do the figures of syntactic and lexical complexity of the paraphrasing data from participants have correlation with the scores from raters?

3. Methodology

3.1 Paraphrasing Task

The paraphrasing test items were cited from the sample task presented in M.-H. Chen et al. (2015). Five target items of paraphrasing were composed of two sentences each and provided one by one online. The participants were supposed to write the paraphrased sentence below the target items (See Appendix A). The task was conducted online following the way of TOEFL writing test. Before taking the test, participants should fill out simple survey questions inquiring their gender, age, and scores of official English test such as TEPS, TOEIC, and TOEFL (See Appendix B).

3.2 Participants

The participants were divided into two groups considering their English proficiency. As Ji (2018) indicated that Korean EFL learners with low proficiency in their English feel the paraphrasing task demanding, this present study only recruited 10 participants for the advanced and intermediate level respectively. The mean score and standard derivation of participants' official English test scores are provided in Table 1.

Table 1. Participants

Group	N	Proficiency Level	Official English Test Scores Range	*M*	*SD*
group A	10	Advanced	(> TEPS 800)	886.5	81.38
group B	10	Intermediate	(TEPS 500 - 650)	589.0	65.16

3.3 The Re-modified Scoring Rubrics

Paraphrasing is restating of a sentence such that both sentences would generally be recognized as lexically and syntactically different while remaining semantically equal (McCarthy et al, 2009). The important factors related to assessing paraphrasing are mainly syntactic change, word change, and semantic equivalency. Since the task has validity as a writing performance test, grammatical errors and mechanical accuracy have to be also considered in scoring. With these five key elements, the modified TOEFL's Integrated Writing Rubrics from Chen et al. has been re-modified (See appendix C).

As previous research conducted by Connor and McCagg (1983) demonstrated that L2 writers have a stronger tendency to be reluctant to transform sentence structures, syntactic change has higher value in the rubrics, which means that if appropriate word change occurs without syntactic change, it can't guarantee high scores. For example, even in the same level of 3 in the scales, the score of combining the presence of syntactic change and only single word change would have same value to that of no syntactic change of one or two adequate phrase changes.

The re-modified rubric construction follows the rubrics from Chen et al. which are composed of a 10-point grading scale ranging from 0 to 9 within 5 levels. However, to give some more freedom to the raters and cover various instances, the syntactic change serves as a variable in each level.

In respect of word change, replacing words and phrases indicates

getting different points. Replacing phrases would imply getting more points than just changing some words. Furthermore, as the most important part in paraphrasing is the semantic equivalency, if the rephrased sentences can't remain semantically equal to the target sentences, even though the syntactic and lexical changes are appropriate, they could not get high level.
As the task was conducted online, there were some chances of typing misspellings which allow the raters to deduct some points in mechanical error section.

3.4 Rating Process

To figure out the inter-rater reliability of the re-modified scoring rubrics, two raters were recruited for the current research project. They were supposed to grade five target items consisted of two sentences respectively from 20 participants. Total 200 paraphrased sentences were assessed by both raters. Each item was graded separately and the average score of each question was given the appropriate level later. Therefore, each participant was provided one level at once.
The raters were trained with sample paraphrasing sentences with the modified TOEFL's Integrated Writing Rubrics from Chen et al. However, it turned out that the rubrics were still unclear for the raters to assess the data appropriately. Therefore, the re-designing of the scoring rubric was held. With the re-modified scales, raters could get an agreement with the correlation coefficient of .77 ($p < .001$).

4. Results

The raters graded 200 paraphrasing sentences. The task was composed of five items respectively and the average score of five items indicated the paraphrasing level from 0 to 5 (rater 1: M=3.15, SD=.745; rater 2:

M=3.0, SD=.917). The overall level graded by the re-modified scoring scales showed correlations as well as the scores of each item. Table 2 shows Pearson correlation coefficient of the overall paraphrasing level graded by two raters using the re-modified scoring rubrics. It indicates correlation of .77 ($p<0.001$), suggesting that the scores from two raters are well co-related. Additionally, not only the overall level, but also the scores of each target paraphrasing item graded by raters show correlations in Table 3. The scatterplots, which show the relationship between participants' paraphrasing scores graded in each of the five items of the rubric across both raters, and then the overall scores, are presented in Figure 1.

Table 4 shows correlations of raters' scores for each group, the advanced and intermediate. The correlation coefficient of the scores of advanced group is .745 ($p<0.005$) and that of the intermediate group is .726 ($p<0.005$). This implies that the grades of the advanced group got slightly more consensus than the intermediate group.

Table 2. Correlations (Overall Level)

	Rater 1	Rater 2
Rater 1	1.00	.770**
Rater 2	.770**	1.00

** p < .001

Table 3. Correlations (Scores of Each Item)

Scores correlated between R1 and R2	Correlation coefficient
Q1	.568**
Q2	.653**
Q3	.760**
Q4	.692**
Q5	.693**

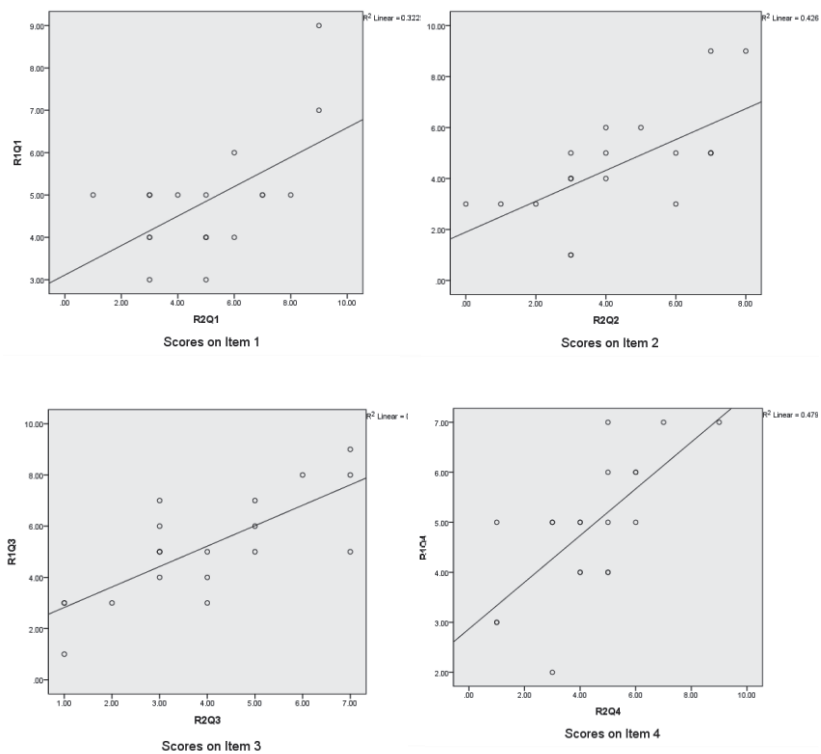
** p < .001

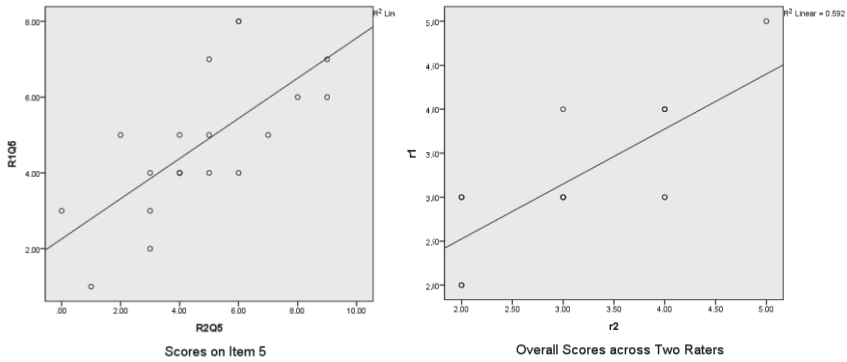
Table 4. Correlations (Scores of Each Group)

Scores of Each group correlated between R1 and R2	Correlation coefficient
Advanced Group	.745*
Intermediate Group	.726*

** p < .005

Figure 1. Correlations of scores on each item across two raters





In the perspective of inter-rater reliability, Cronbach's α of the scores from two raters was estimated. Cronbach's alpha, α (or *coefficient alpha*), developed by Lee Cronbach in 1951, measures reliability, or internal consistency. If the figures measured by Cronbach alpha are above 0.9, it means "*excellent*". Generally the scores in range of 0.9 and 0.8 suggest "*good*", 0.8 and 0.7, "*acceptable*", and 0.7 and 0.6, "*questionable*". Table 5 shows the results of reliability measured by Cronbach alpha, which indicates the scoring rubrics employed in this study have inter-rater reliability to be "*good*" as the Cronbach's α of overall level is .859. Mostly the figures of coefficient alpha from the scoring rubrics are above 0.7 except for Q1 showing reliability of .672.

Table 5. Inter-rater reliability

Inter-rater reliability of scores from two raters	Cronbach's alpha
Q1	.672
Q2	.863
Q3	.788
Q4	.782
Q5	.801
Overall level	.859

Regarding the previous studies on syntactic and lexical complexity and

writing materials of L2 leaners, the correlation between syntactic and lexical complexity and the results of paraphrasing task also have been investigated. Engber (1995) and Linnarud (1986) implied that correlation did not exist between the oral narratives and lexical density measured. Moreover, Lu (2012) also found that there was not a significant correlation between lexical density and scores from raters in oral narratives. To evaluate lexical complexity of paraphrasing data in this study, Lexical Complexity Analyzer (Lu, 2010) was employed. Among 25 indices measuring lexical complexity, lexical density of paraphrasing data was analyzed following previous studies (Lu, 2015). The results from this present study showed similar indication with previous studies in terms of correlation between paraphrasing data and lexical density. Table 6 shows that there was no significant correlation between them.

A slightly different implication with that of lexical complexity was suggested for the results from correlation of syntactic complexity and raters' scores, still there was no significant correlation between them. Table 7 shows the correlation between figures measured for syntactic complexity and scores from two raters. L2 Syntactic Complexity Analyzer (Ai & Lu, 2010) was employed to measure syntactic complexity of paraphrasing data from 20 participants. Among 14 indices, to check the overall sentence complexity, clauses per sentence were measured (Lu & Ai, 2015).

Table 6. Correlations (Lexical Complexity)

	Rater 1	Rater 2	Lexical Complexity
Rater 1	1.00	.770**	.073
Rater 2	.770**	1.00	-.095
Lexical Complexity	.073	-.095	1

** $p < .001$

Table 7. Correlations (Syntactic Complexity)

	Rater 1	Rater 2	Syntactic Complexity
Rater 1	1.00	.770**	.014
Rater 2	.770**	1.00	.224
Syntactic Complexity	.014	.224	1

** $p < .001$

5. Discussions

This present study aims to investigate reliability of the re-modified scoring rubrics for the paraphrasing task for Korean EFL learners. The figures of Cronbach's alpha confirmed that the scores rated by two raters using the same scales are reliable. Aside from this quantitative analysis, a qualitative analysis with the raters was also conducted using a brief survey with a few questions related to the rating process.

Both of the raters said that they well understood the training held before the real rating procedure. However, they still questioned about the clarity of the re-designed scoring rubrics. When giving points, they felt that the data from some participants did not really fit exactly to the provided rubrics in the section of syntactic change and semantic equivalency. Thus, they had to decide the scores with a little subjectivity included. Despite this lack of clarity in the rubrics, the scores could get an agreement and gain significant reliability.

6. Limitations and Conclusions

This study revealed significant and elaborate correlations and reliability

of the re-modified scoring rubrics employed in paraphrasing task. As paraphrasing has been highlighted for being a new source of a writing performance test, it gains its value in the field of EFL writing. However, the absence of a standard in grading paraphrasing data makes raters and teachers feel demanding and allows them to assess the material with their subjectivity. Thus, defining an adequate criterion for scoring EFL paraphrasing task is highly necessary.

With this aim, the study was conducted figuring out correlations and reliability of the re-edited version of scoring rubrics from Chen et al. Although some of meaningful results have been presented, there were a few limitations which should be handled in future studies.

6.1 Items in Paraphrasing Task

The paraphrasing task conducted in the current study was composed of five questions with two sentences each. In other words, participants were supposed to paraphrase two sentences for each question. All of ten target sentences are cited from Chen et al. (2015).

As paraphrasing is a crucial skill for writing essays, writing a whole paragraph would be more suitable for the task. However, replacing the whole paragraph would be highly challenging to participants as they claimed that task of changing two sentences was hard enough. For future study, inventing a paraphrasing task containing adequate number of target sentences would be necessary.

6.2 Scoring Rubrics

In light of the re-modified scoring rubrics, a little more specificity for five key elements, syntactic change, word change, semantic equivalency, grammatical errors, and mechanical accuracy, would be required. The collected paraphrasing data could not properly fit into the levels provided by the rubrics.

Furthermore, the rubrics would need a confirmation of an expert in the field of writing. As for this study, the re-modification of the existing rubrics from Chen et al. was only conducted by the current researcher. If the scoring scales are ensured by professionals in the field, the vagueness would be reduced helping raters assess more easily.

6.3 Syntactic and Lexical Complexity

For measuring syntactic and lexical complexity, the programs used in the previous studies, LCA and L2SCA (Ai & Lu, 2010), were employed. The numbers of indices for each complexity that the program can measure are twenty-five and fourteen, respectively. However, for this current study, only single index among various items was evaluated. The measure was selected following the previous studies analyzing both of complexities. Lexical density of the paraphrasing data was analyzed for lexical complexity, and clause per sentence was measured for syntactic complexity. However, they indicated no significant results. Thus, if all 25 and 14 indices offered by each program were evaluated, the results of participants could be meaningful. Further study would be expected to analyze syntactic and lexical complexity.

Nonetheless, the re-modified scoring rubrics were reliable for grading EFL paraphrasing tasks and there was correlation between the scores of two recruited raters. With some more details added to the rubrics, raters in EFL field would assess the paraphrasing data much more readily. This truly means helping teachers and test administrators in teaching and learning environment. Since the paraphrasing task has its validity as a noble source for a writing performance test, proper and clear scoring rubrics could be useful in the education area.

References

Ai, H., & Lu, X. (2010). A Web-based System for Automatic Measurement of Lexical Complexity. *Paper presented at the 27th Annual Symposium of the Computer-Assisted Language Consortium* (CALICO-10). Amherst, MA. June 8-12.

Chen, M. -H., Huang, S. -T., Chang, J. S., & Liou, H. -C. (2015). Developing a Corpus-based Paraphrase Tool to Improve EFL learners' Writing Skills. *Computer Assisted Language Learning*, 28:1, 22-40.

Connor, U., & McCagg, P. (1983). Cross-cultural Differences and Perceived Quality in Written Paraphrases of English Expository Prose. *Applied Linguistics*, 4(3), 259-268.

Engber, C. A. (1995). The Relationship of Lexical Proficiency to the Quality of ESL Compositions. *Journal of Second Language Writing*, 4, 139–155.

Fitzgerald, J. T., White C. B., & Gruppen L. D. (2003). A Longitudinal Study of Self-Assessment Accuracy. *Medical Education*, 37(7), 645-649.

Ji, N. Y. (2018). Investigation into Validity of Paraphrasing Task as a Writing Performance Test Item for EFL Learners. *Modern English Education*, 19(2), 20-29.

Keck, C. (2006). The Use of Paraphrase in Summary Writing: A Comparison of L1 and L2 Writers. *Journal of Second Language Writing*, 15(4), 261-278

Linnarud, M. (1986). Lexis in Composition: A Performance Analysis of Swedish Learners' Written English. Lund, Sweden: CWK Gleerup.

Lu, X. (2012). The Relationship of Lexical Richness to the Quality of ESL Learners' Oral Narratives. *The Modern Language Journal,* 96(2):190-208.

Lu, X., & Ai, H. (2015). Syntactic Complexity in College-level English Writing: Differences among Writers with Diverse L1 Backgrounds. *Journal of Second Language Writing,* 29, 16-27.

McCarthy, P. M., Guess, R. H. & McNamara, D. S. (2009). The Components of Paraphrase Evaluations, *Behavior Research Methods*, 41(3), 682-690.

Reid, R., Lienemann, T. O., & Hagamann, J. L. (2013). Strategy Instruction for Students with Learning Disabilities (2nd ed.). New York: Guilford Publications

Wen, Q., Wang, L., & Liang, M. (2005). Spoken and Written English Corpus of Chinese Learners. *Beijing: Foreign Language Teaching*

and Research Press.

Appendix A. The Paraphrasing Task conducted Online

Paraphrasing Test (Question no.1 out of 5) (originally in Korean)

Paraphrase the given sentences changing the sentence structure and vocabulary but not changing the number of sentences or meaning.

ex) A great number of people use Mandarin. Therefore, learning Mandarin language becomes a popular hobby.

→ Mandarin is spoken by a substantial number of people. Thus, many people are interested in learning Mandarin.

* Do not use a dictionary and only use vocabulary you already know.

1. On the whole, fuel prices have risen in recent years. Similarly, the cost of food has increased quite considerably.

_____

Appendix B. A Brief Survey for Participants

Survey (originally in Korean)

Gender

□ Male □ Female

Korean age

_____

Official English test score (choose one from TEPS/TOEIC/TOEFL)

ex) TEPS 600, TOEIC 800, etc.

_____

Appendix C

Table A.1. Re-modified Scoring Rubrics for Paraphrasing Task

Level (points)		Syntactic Change (Sentence Structure)	Word Change	Semantic Equivalency	Grammatical Error	Mechanical Accuracy
Level 1	0	restating the target sentence				
		No syntactic change	inappropriate one or two single words change	**semantically not equivalent**	grammatically wrong	misspelling, punctuation errors contained
	1	No syntactic change	inappropriate one or two single words change	**semantically not equivalent**	grammatical sentence	few errors contained
		one of Syntactic change occur (positions of Adverbials, voice (태), using complementizer that)	no word change	**semantically equivalent**	-	-
Level 2	2	No syntactic change	inappropriate single word change	semantically equivalent	some grammatical errors	some errors contained

	3	No syntactic change	appropriate single word change	semantically equivalent	few grammatical errors	few errors contained
		one of Syntactic change occur (positions of Adverbials, voice (태), using complementizer that)	inappropriate single word change	**semantically not equivalent**	-	-
Level 3	4	No syntactic change	appropriate more than two words change	semantically equivalent	some grammatical errors	some errors contained
	5	No syntactic change	appropriate one or two phrases change	semantically equivalent	few grammatical errors	few errors contained
		one of Syntactic change occur (positions of Adverbials, voice (태), using complementizer that)	appropriate single word change	semantically equivalent	few grammatical errors	few errors contained
Level 4	6	one of Syntactic change occur (positions of Adverbials, voice (태), using complementizer that)	appropriate more than two words change	semantically equivalent	some grammatical errors	some errors contained

		some of Syntactic change (positions of Adverbials, or voice (태), using complementizer that)	appropriate single word change	semantically equivalent	some grammatical errors	some errors contained
	7	one of Syntactic change occur (positions of Adverbials, voice (태), using complementizer that)	appropriate one or two phrases change	semantically equivalent	few grammatical errors	few errors contained
Level 5	8	one of Syntactic change occur (positions of Adverbials, voice (태), using complementizer that)	more than three appropriate phrases change	semantically equivalent	few grammatical errors	few errors contained
	9	some of Syntactic change (positions of Adverbials, or voice (태), using complementizer that)	more than three appropriate phrases change	semantically equivalent	no errors	no errors

Minkyung Kim
minkyung.kim0322@snu.ac.kr