Reliability of paraphrasing scores: Determining appropriate number of items for a paraphrasing test for Korean EFL learners

Minkyung Kim
Seoul National University

Kim, Minkyung. 2022. Reliability of paraphrasing scores: Determining appropriate number of items for a paraphrasing test for Korean EFL learners. *SNU Working Papers in English Language and Linguistics 18, 35-48.* Paraphrasing is one of the major writing skills that has been discussed in academic writing but have not been tested sufficiently in the context of the second language writing. With great importance, paraphrasing skill is essential enough to be considered as a new material for writing assessment. Thus, a paraphrasing test was developed in Kim's MA thesis (2020) with remarkable reliability and validity. In this study, as an expansion of Kim's MA thesis, further investigation is held in terms of complementing the test. Figuring out the optimal number of the paraphrasing test is the aim of the current study. Various methods of analysis are employed including Cronbach's alpha and Spearman-Brown Prophecy Formula. The score reliability of the composite scores for analytic scoring rubric and holistic total scores in Kim (2020) is .83 and .88. However, slightly lower values are investigated in each section of scoring rubrics. To increase the reliability of the paraphrasing test, more items should be involved in the test rather than five items. The optimal test length for the test is around 20 to reach the targeted reliability, .90. Furthermore, the exact number of the items for each rating dimension was investigated. This study would provide some instructional advice regarding determining the test length to the potential test-designers in the same field. **(Seoul National University)**

Keywords: language assessment, psychometric analysis, writing assessment, spearman-brown prophecy formula, paraphrasing task

1. Instruction

Paraphrasing is one of the major writing skills that has been argued in academic writing but not have been tested thoroughly in the field of the assessment of second language writing. The paraphrasing ability is an essential skill for undergraduate and graduate students as they need to acquire some skills to avoid plagiarism. After learning and practicing the paraphrasing skill, it is well-known that they have to be tested to figure out if they have acquired the skills properly. In this sense, the paraphrasing test from Kim (2020) has been developed. Kim (2020) modified the paraphrasing task and scoring rubrics in Chen et al. (2015) and concluded that this re-modified paraphrasing task has a significant reliability and validity to be considered as a novel material for writing assessment. Moreover, educational and instructional advice for the ESL and EFL field was meaningful.

This current study wanted to take a step further to build some more useful insight regarding this paraphrasing task. The paraphrasing task in Kim (2020) consists of five items with two separate sentences each. Even though this test was significant and valid in terms of reliability and validity, scanning each item thoroughly and finding better combination would be a helpful step.

To develop a paraphrasing task with a more perfection, the work of investigating of the appropriate number of the test item was held. The Spearman-Brown Prophecy formula (Brown, 1910 & Spearman, 1910) was employed to investigate the item reliability. This formula reveals the right number of items which should be positioned in the paraphrasing task and this work was progressed with fining the saturation point of the item reliability. It turned out that the 20 items would be perfect for this paraphrasing task with an item reliability .90.

2. Literature review

When researchers develop tests, they should consider various things such as test time and the number of items properly. Whether a test is reliable or not in terms of the number of items included in a test is one of the most important elements that researchers should ponder seriously. Lee (2005) addressed critical issues regarding score dependability of the new version of the Test of English as a Foreign Language (TOEFL) and figured out the optimal number of tasks in

the New TOEFL where the reliability gets .90. In this study, generalizability theory (G-theory, Cronbach, Gleser, Nanda, & Rajartnam, 1972) was employed because the Speaking section in the new TOEFL involves more than one major random facet such as tasks and raters. Therefore, it needs a multifaceted analysis. G-theory aims to analyze more than one measurement facet at the same time. In Lee (2005), the optimal task length was around 11 to 12 tasks with one or two ratings. As the participants in Lee (2005) were taking 13 speaking tasks for three task types, it is believed that the test takers took high reliable tasks.

de Vet HCW et al. (2017) aims to increase insight into reliability studies by pointing to the assumptions of reliability coefficients, similarities between various coefficients, and the subsequent new applications of reliability coefficients. Spearman-Brown Prophecy Formula was used to figure out the proper test length in the study. Moreover, the formula was also employed for raters instead of items, to predict inter-rater reliability when the number of raters changes. Figure 1 shows the reliability values among two to four raters.

Figuring out the optimal number for the test or the rater would be a useful guidance for the test designers. With applying reliability values in various formula, the results can help develop a better test. Therefore, the current study aims to figure out the optimal number for the paraphrasing test in Kim (2020). As the number of raters in Kim (2020) is fixed for two, the optimal number for

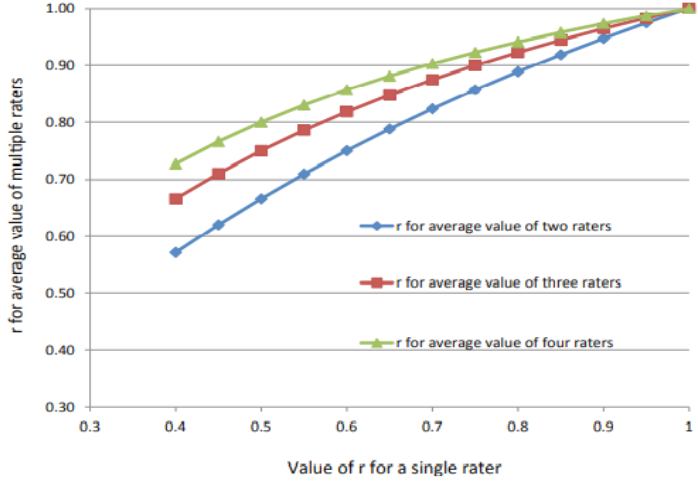


Figure 1. Correspondence between reliability coefficients for single raters and more

raters would be investigated in further steps.

The research questions of the current study are as follows:

1. What is the optimal test length for each section and total in terms of score reliability?
2. What is the exact number of items that requires targeted reliability of .90 for each section?
3. What are the results of reliability when increasing and decreasing the number of items in test from 1 to 30?

3. Methodology

3.1. Participants

A total of 100 Korean EFL learners participated in this study to take the paraphrasing test developed for the current study. Table 1 summarizes the background information of the participants, including the group means of subjects' gender, age, TEPS score, and major. The gender distribution among the test-takers was almost equal between males and females. Fifty-six female and forty-four male students attended the testing session. Since they were university students (graduate and undergraduate), their age varied from early twenties to mid-thirties. The average age was 24.52 years old.

3.2. Paraphrasing test

The paraphrasing task in Kim (2020; Appendix A) is comprised of 5 items with two sentences each. Test-takers should paraphrase two sentences for each item. The total of 10 sentences were presented in the paraphrasing test. The value of reliability was .83 meaning significant in her study but it is still questionable whether it presents the best number of the items.

3.3. Scoring rubrics

In Kim (2020), there were two kinds of scoring rubrics, analytic and holistic (Appendix B), to explore the inter-rater reliability between two raters. The analytic scoring rubric has four rating dimensions: Syntactic Change (SC), Word Change (WC), Semantic Equivalency (SE) and Grammatical &

Mechanical Errors (G&M). Since there were a total of four rating dimensions with a 6-point scale ranging from 0 to 5 per task, the possible score range for participants would be 0-20. Given the fact that there were five paraphrasing tasks used in this study, the participants' total test score could range from 0 to 100. On the other hand, when the holistic rubric was used, there was only one single holistic dimension for rating and the two raters' ratings were averaged on this single dimension. For this reason, the total score from the holistic rubric could vary from 0 to 25.

3.4. Method of analysis

To get the reliability of paraphrasing test, "Cronbach's alpha if deleted" in SPSS was employed. The reliability was obtained in six categories including four rating dimensions, analytic composite score and holistic total score. Table 1 below shows the score reliability.

To figure out the best number for the items in the paraphrasing task, the Spearman-Brown Prophecy Formula (SB Formula, Spear, 1910 & Brown, 1910) was employed and the values were calculated in Excel. The SB formula was originally developed independently by Spearman and Brown published in the same journal in 1910 and nowadays still a topic of interest. The formula expects the reliability of a questionnaire when using the subgroups of items (split in half, thirds, fourths, etc.) to examine internal consistency. The SB formula was used to predict reliability for another number of items assuming the average correlation (mean *r*) remains the same.

In current study, there were two SB formulas employed: the original version (formula 1) and the one for the optimal test length (formula 2).

$$(1)\ r_{SB} = \frac{nr}{1 + (n - 1)r}$$

In the formula 1, *n* is the factor by which the number of items will be multiplied, and r is the reliability (internal consistency) of the test. For example, if the test

Table 1. Score reliability of Cronbach's alpha

	SC	WC	SE	G&M	Comp	H_Total
Reliability	0.7	0.69	0.68	0.72	0.83	0.88

is shortened by a factor 2, *n* will be 0.5, and if the test contains twice as many items, *n* will be 2. This is useful in the developmental phase of a questionnaire when internal consistency appears to be rather low because only a few items are included. In that case, the researchers could add items to improve the reliability of the test (de Vet HCW et al., 2017).

The revised formula for calculating the optimal number of items is shown below.

$$(2)\ n = \frac{r_{SB}\,(1 - r)}{r(1 - r_{SB})}$$

In the revised formula, n means the optimal length and is the reliability calculated by the SB formula. *r* means the item reliability by Cronbach's alpha. With this formula, the optimal number for the paraphrasing test can be investigated.

4. Results

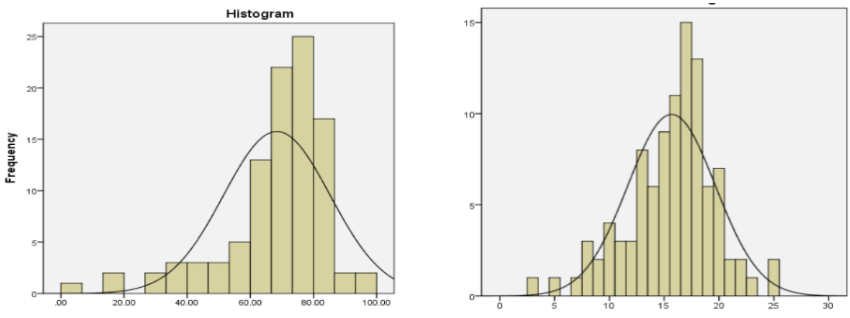
4.1. Descriptive statistics

A total of 100 participants took the paraphrasing test and their scores were rated with two different kinds of rubrics, analytic and holistic ones. Table 2 shows descriptive statistics of the scores from two rubrics. The average score of analytic rubric is 74.5 while that of holistic one is 15.7.

Figure 2 presents the histogram of the test-takers' scores of the paraphrasing test in both analytic and holistic ratings in Kim's MA thesis. The scores used in obtaining the distribution here are a sum of analytic composite scores across tasks and the test-takers' total holistic score computed by adding up all task scores.

Table 2. Descriptive statistics of paraphrasing scores

Rubric	N	Minimum Score	Maximum Score	Mean	SD
Analytic	100	45.5	99.5	74.5	8.9
Holistic	100	10	24.5	15.7	2.7



a. Analytic Composite Rating b. Holistic Total Rating

Figure 2. Histograms of analytic and holistic scores

In Figure 2-a, the distribution curve for the summed analytic composite is slightly more left-skewed, indicating that most test-takers achieved a score between 75 and 85 in total of a hundred. The curve in Figure 2-b suggests a moderate symmetrical distribution which indicates the scores from holistic rating occur at more regular frequencies than those from analytic rating, when evaluated from the perspective of normal distribution.

4.2. Optimal test length

The optimal test length for the paraphrasing test in Kim's MA thesis was

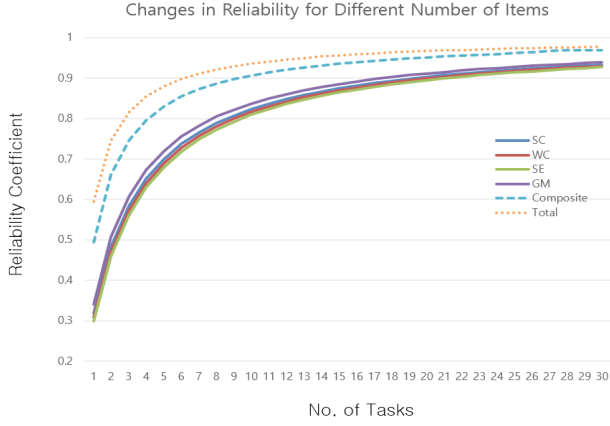


Figure 3. Changes in reliability for different number of items in the paraphrasing test

investigated in Figure 3. With the original five items, the item reliability is .83 for the composite score of analytic scoring rubric. It can be assumed that if the test length increases, the reliability also increases. Around 20 items would be optimal for each rating dimension.

The exact number for the paraphrasing test where each rating criterion gets the reliability .90 is shown in Table 5. The rating dimensions related to paraphrasing skills, SC, WC and SE, require around 19 to 21 items to reach .90 in SB formula. As the paraphrasing test has five items, it needs to get four times more items.

4.3. Exact number for the targeted reliability

The formula 2 in method section is the revised one for calculating the exact number of items for the targeted reliability. In this study, targeted reliability is .90. Once the optimal test length was investigated before, the figures roughly indicated around 20 items for each rating dimension. However, there are slight differences between each rating criterion. The numbers are shown in Table 3 below.

For SC, to reach the targeted reliability, it needs 19.28 items. For WC, it needs 20.21 items, and for SE, 21.17 items. For G&M, it needs the lowest number of items among the analytic scoring rating dimensions, 17.5.

When the scores are combined, the reliability is usually higher than the separate section. Likewise, the number of items needed to get the targeted reliability decreases for composite and the holistic total score compared to the dimensions in analytic scoring rubric. For composite scores, it needs 9.21 items to reach the .90 reliability. As the original paraphrasing test has five items, it needs to be doubled to get around 10 items. For holistic total scores, it is almost similar to the original number of the paraphrasing test. Approximately six items are enough in terms of holistic scoring.

Table 3. Exact no. of items for targeted reliability (.90)

SC	WC	SE	G&M	Comp	H_total
19.28	20.21	21.17	17.5	9.21	6.13

5. Discussion

This present study aims to investigate the optimal test length for the paraphrasing test in Kim's MA thesis. The paraphrasing test originally included five items and the score reliability, value of Cronbach's alpha, was .83 for the composite score of analytic scoring rubric. Even though the score reliability for the analytic composite scores was quite high, the reliability figures for each rating dimension for analytic scoring rubric were around .70.

There is a possibility that reliability can increase if the number of items in the paraphrasing test changes. More than five items would be necessary to obtain higher score reliability. With the Spearman-Brown Prophecy Formula, the optimal number was calculated. It would be ideal to have around 20 items for the paraphrasing test to reach the targeted reliability, .90.

The results revealed some of interesting points that need to be discussed. Firstly, it is remarkable that three rating dimensions related to the paraphrasing skills require similar number of items, which is 20, while the dimension for grammatical and mechanical errors (G&M) needs slightly fewer number of items. The G&M dimension is much more related to general writing skill rather than a paraphrasing skill. Thus, the slight difference in the number of items might be due to the difference in the skills of test-takers.

6. Limitations and future studies

Figuring out the optimal test length is a crucial process for a better test. The paraphrasing test can be further developed with some improvements based on the results in the current study. However, there is still a room for a further investigation regarding the multi-facet analysis. As Kim's study had two raters, the changes on the number of the raters could be another step. Finding the optimal number for raters for the paraphrasing test will also help the test to be a more reliable and valid one.

References

Brown, W. (1910). Some experimental results in the correlation of mental abilities. *British Journal of Psychology*, *3*, 296-322.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The Dependability of Behavioral Measurements: Theory of Generalizability*. New York: John Wiley.

de Vet HCW, Mokkink LB, Mosmuller DG & Terwee C. B. (2017). Spearman-Brown prophecy formula and Cronbach's alpha: Different faces of reliability and opportunities for new applications. *Journal of Clinical Epidemiology*, *85*, 45-49.

Kim, M. K. (2020). Investigating the reliability and validity of scores from a paraphrasing test for Korean EFL learners. Unpublished MA thesis, Seoul National University, Seoul, Korea.

Lee, Y. W. (2005). Dependability of scores for a new ESL speaking test: Evaluating prototype tasks. *Monograph Series*. MS – 28. ETS TOEFL Publications.

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.

Spearman, C. (1910). Correlation calculated from faulty data. *British Journal of Psychology*. *3*, 271-295.

Appendix A

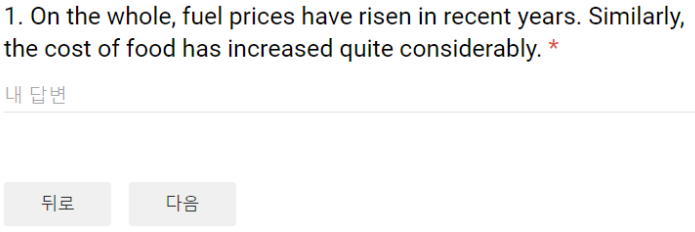


Figure A1. A sample item in the paraphrasing test

Appendix B

Table B1. Analytic Scoring Rubric

	Lv. 5	Lv. 4	Lv. 3	Lv. 2	Lv. 1	Lv. 0
Syntactic Change (Sentence Structure)	The sentences are paraphrased with some syntactic changes on the list for examples included. • Changing the positions of adverbials • Changing the voice of sentences. • Changing the structure using complementizer that.	The sentences are paraphrased with one or more syntactic changes on the list for examples included. • Changing the positions of adverbials • Changing the voice of sentences. • Changing the structure using complementizer that.	The sentences are paraphrased with a syntactic change, but they do not seem appropriate.	The structure of the paraphrased sentences seems to be similar to the original sentences.	The structure of the paraphrased sentences does not seem to be changed.	The participants do not follow the directions or restate the original sentences.
Word Change	The sentences are paraphrased with more than three appropriate phrase changes.	The sentences are paraphrased with one or two appropriate word or phrase changes.	The sentences are paraphrased with an appropriate single word change.	The sentences are paraphrased with inappropriate words or phrases.	The words of the paraphrased sentences seem to be similar to the original sentences.	The participants do not follow the directions or restate the original sentences.
Semantic Equivalency	The paraphrased sentences are semantic equivalents. The original meaning has been retained clearly. It is easy to comprehend the meaning of the paraphrased sentences compared to the original sentences.	The paraphrased sentences are semantic equivalents in general. The original meaning has been retained generally.	The semantic equivalency of the paraphrased sentences seems not to be retained. It is confusing to comprehend the meaning of the paraphrased sentences compared to the original sentences.	The paraphrased sentences do not seem to be semantic equivalents in general.	The paraphrased sentences are not semantic equivalents.	The participants do not follow the directions or restate the original sentences.
Grammatical Error & Mechanical Accuracy	The paraphrased sentences include no grammatical, misspelling, and punctuations errors.	The paraphrased sentences include a few grammatical errors, misspellings, and punctuations errors. .	The paraphrased sentences include some grammatical errors, misspellings, and punctuations errors. .	The paraphrased sentences include a lot of grammatical errors, misspellings, and punctuations errors. .	The paraphrased sentences are grammatically wrong and there are a lot of misspellings and punctuation errors. .	The participants do not follow the directions or restate the original sentences.

Table B2. Holistic Scoring Rubric

Level	Task Description
5 Paraphrased sentences at this level largely accomplish all of the following criteria.	♦ Participants restate **both of original sentences** (replace phrases and rearrange sentence structure) with more than ***three words and parts of structures*** changing along with ***no grammatical and mechanical errors***, and the paraphrased sentences ***remain semantic equivalents***.
4 Paraphrased sentences at this level largely accomplish all of the following criteria.	♦ Participants restate ***one or two parts*** of the original sentences by changing the sentence structures and replace ***one or two phrases or words*** with ***few or no*** grammatical errors, and the phrases ***remain semantic equivalents***.
3 Paraphrased sentences at this level largely accomplish all of the following criteria.	♦ Participants restate ***a single part*** of the structure of the original sentences and replace ***a single word*** with ***few or no*** grammatical and mechanical errors, and the phrases ***still remain semantic equivalents*** but they are generally ***confusing to understand.*** ♦ Participants restate a single part of the structure of the original sentences and replace a single word with some grammatical and mechanical errors, and the phrases ***do not remain semantic equivalents in general.*** Participants restate and replace ***more than three parts and phrases or words*** with more grammatical errors, and the replacement results in inaccuracy, vagueness of semantic equivalency, or imprecision of some content.
2 Paraphrased sentences at this level largely accomplish all of the following criteria.	♦ Participants restate the original sentences but the paraphrased sentences have ***similar sentence structures***, replace the words in the original sentences with ***inappropriate words*** along with a few grammatical and mechanical errors, and the paraphrases sentences ***are not semantic equivalents in general***. ♦ Participants replace ***single words*** appropriately and replace only ***one or two phrases or words*** with few or no grammatical errors but the replacement results in inaccuracy, vagueness, or imprecision of some content.
1 Paraphrased sentences at this level largely accomplish all of the following criteria.	♦ Participants do not change the sentence structure of the original sentences, use similar words to the paraphrased sentences and do not retain the semantic equivalency. ♦ The paraphrased sentences are generally grammatically wrong.
0	♦ Participants do not follow the directions or restate the original sentences.

Minkyung Kim
minkyung.kim0322@snu.ac.kr