Reading recall protocols as means of measuring reading comprehension

Young Ye Park
(KAIST)

Park, Young Ye. 1999. Reading recall protocols as means of measuring reading comprehension. *Linguistics*, 7-1, 345-360. The purpose of the present study is to examine reliability and validity of reading recall protocols as means of measuring reading comprehension. Thirty-seven university students at the Korea Advanced Institute of Science and Technology (KAIST) participated as subjects. They were given reading recall protocols and the TOEFL reading subtest as reading measures. The students' recall protocols were assessed by two independent raters. The results demonstrated that there was a strongly high inter-rater reliability coefficient (r=.98), implying that the method was very reliable. The moderately high correlation coefficient(r=.68) between reading recall protocols and the TOEFL reading subtest indicated the existence of concurrent validity of the test. The scoring procedures revealed that there would be some problems in adopting reading recall protocols. The scoring procedures are described in detail, and suggestions are included in the final discussion for further research that should be taken into consideration to employ reading recall protocols more efficiently. **(Korea Advanced Institute of Science and Technology)**

1. Introduction

In general, it has been assumed that productive skills such as speaking and writing are more difficult to measure than receptive skills like reading and listening. This assumption is partly attributable to the fact that productive skills are often measured by subjective scoring methods that are usually unreliable. However, according to

Hughes(1989), receptive skills may also be difficult to assess because they are difficult to observe in overt behavior.

Reading comprehension is often measured through such objective meth ods as multiple-choice tests, true/false tests, cloze tests, and sentence completion tests. The most widely used is probably the multiple-choice test, due to the simplicity of administration and scoring. Since multiple-choice tests are vulnerable to guessing(Connor & Read, 1978), they sometimes lead to misinterpretation of the student's ability to comprehend. Moreover, reading research suggests that memory plays an important part in an interaction between the reader and the text. It is also possible that in multiple-choice tests, students might choose the right answer based on their test-taking strategies(Nevo, 1989) such as differentiating the correct answer from the distracters by looking at answer length. Other tests, such as true-false and cloze tests, are also frequently adopted in assessing students' reading comprehension for their objective scoring methods.

Despite its convenience in administration and scoring, the traditional multiple-choice comprehension test often focuses on isolated lexical items; as a result, it fails to measure students' actual understanding of the text, as pointed out by Been(1975). Similarly, cloze tests are also criticized for their limited ability to measure students' comprehension at the local level (Alderson, 1979; Cohen, 1984; Swaffar, Arens, & Byrnes, 1991).

As the extent to which students can recall the information from the text would explain how it is comprehended(Bernhardt, 1984), it is necessary to assess the student's ability to recall the text. Bernhardt(1991), who claimed that a second language assessment mechanism should provide in-depth information as well as quantifiable data, recommended immediate recall protocols as means of measuring reading comprehension for the following reasons:

> First, recall can show where a lack of grammar is interfering with the communication between text and reader, while not

> focusing a reader's attention on linguistic elements in texts. Second, generating recall data does not influence a reader's understanding of a text. . .a free recall measure provides a purer measure of comprehension, uncomplicated by linguistic performance and tester interference. (p. 200)

Furthermore, as asserted by Stevens(1988), since neither single words nor individual sentences carry the same importance in meaning, the reader's ability to identify the main ideas in the text will reflect his/her ability to comprehend. In a study(Brown & Smiley, 1977) comparing L1 readers' ability to recognize main ideas in a text, more proficient readers tended to recall more information, suggesting that their recall is relevant to their ability to remember the main ideas.

The purpose of the present study is to investigate whether reading recall protocols are a reliable and valid method for measuring students' reading comprehension. It describes the scoring procedures of reading recall protocols; then it attempts to evaluate reading recall protocol methods in terms of some considerations suggested by Henning(1987), for selection or development of an appropriate test. Advantages and disadvantages of the method are also discussed.

2. Methodology

2.1 Subjects

Thirty-seven university students at the Korea Advanced Institute of Science and Technology(KAIST) participated as the subjects in the current study. They were enrolled in an English reading course that the investigator was teaching at the time of this investigation. All but five students were freshmen. There were three female students; the remainder were male. As the students were not placed in the class according to their English proficiency levels, their reading proficiency varied. Since the course emphasized reading, it was appropriate to ask

the students to recall a passage and then summarize it in writing.

2.2 Measures

2.2.1 Reading recall protocols

The students were given the following passage selected from the TOEFL. The TOEFL reading comprehension passage was reprinted with the permission of Educational Testing Service, the copyright owner. The rhetorical type of the passage was description. As the TOEFL was developed to measure the language proficiency of the students intending to study at the university level, the level of text difficulty was not investigated in this study.

The students were encouraged to read the passage as often as they liked. They were given 10 minutes to read the passage several times. They were not permitted to write anything while they read, but they were allowed to underline or circle words or phrases in the passage if it helped them to better understand the passage. After the students finished reading, they were required to put the passage out of sight. They were then asked to write down everything they remembered from the passage. Again, 10 minutes were given to write.

2.2.2 The TOEFL reading comprehension subtest

The reading comprehension section of the Test of English as a Foreign Language(TOEFL)--Form 3KTF12--published by the Educational Testing Service was also used by permission of the publisher to measure the students' reading comprehension. The TOEFL consisted of three sections: Section 1, Listening Comprehension; Section 2, Structure and Written Expression; and Section 3, Vocabulary and Reading Comprehension. Section 3 of the TOEFL consisted of 30 vocabulary items and 30 reading comprehension items. Thirty reading comprehension items from Section 3 were adopted in the present study.

The TOEFL reading comprehension subtest contains five reading passages covering varied topics dealing with science, history, education,

technology, and business. These reading passages were drawn from various sources, but topics that were likely to give unfair advantages to certain cultural groups were eliminated (Peirce, 1992). The length of the shortest passage in this study was approximately 160 words, while the longest passage was approximately 340 words in length.

2.3 Procedures

2.3.1 Raters

The students' recall protocols were first assessed by two independent raters: the investigator of this study and an another Korean instructor who was teaching the same course. For comparison of inter-rater reliability between two pairs of raters, the students' responses were also scored by a second pair of the independent raters. Both of these were native speakers of English who were teaching communication and writing courses at KAIST when the investigation was carried out.

2.3.2 Scoring procedures

A seven-step process was adopted in scoring the recall protocol: 1)choosing a scoring method; 2)breaking the reading passage down into scoring units; 3)weighting the units; 4)creating a scoring template; 5)preliminary scoring; 6)generating a scoring guidelines; and 7)final scoring. In consultation with D. F. Wolf, an experienced second language reading researcher, a simple segmentation plan was chosen, using weighted pausal units proposed by Johnson(1970) for scoring. This procedure is less complicated and easier to follow than the syntactic analysis of the Meyer method(Meyer, 1985), since it retains the original sequence of the passage. Bernhardt(1991) was able to establish a reasonable level of overlap between the scores obtained using both methods on the same data, and recommends the segmentation plan as more efficient.

The passage was segmented into scoring units at the natural pauses when the text was read aloud. Two raters did a first pass individually;

a final version of the pausal units was generated at a conference in which the raters' versions were pooled with a version generated by the experienced researcher mentioned above. The resulting units turned out to be reasonably simple to score, with the exception of six units(out of a total of 58). The problem units contained two or more items of information that were often recalled separately; this necessitated a disjunct scoring category (credit was given for either item as well as for both). This points to the only notable flaw with pausal units; some units contain more information than pauses. A reading recall protocol could be scored on a simple dichotomous basis, or the scoring units could be weighted according to their relative importance(salience to the main idea). Since reading for main ideas was major factor in good comprehension, it made sense to weight the units. Furthermore, although individual schemata are closely related to comprehension of the text, it is also true that "diverse readers will recognize the same macro-features in a text because those features exist independently of individual readings," as Swaffar, Byrnes, and Arens(1991, p. 74) asserted.

For simplicity, the raters decided to use three levels of importance rather than the usual four. First, five graduate students with a good command of English were asked for input. Five different versions of weighting from readers were generated. To settle disagreements among these five (especially with regard to "main ideas"), three fluent bilingual speakers of Korean and English were asked to write a summary of the main ideas in the passage. The final weightings were assigned as follows: 1)where four out of five readers agree, that weight was assigned; 2) where two readers disagreed, the summary protocols were consulted to make the decision.

The passage turned out to be very dense in "main ideas" and not very well organized. Where this was the case, it might have been helpful to use the four-level system(Bernhardt, 1991) to distinguish between first statements and repeats of the main ideas. It is possible that this weighting procedure could also be improved by allowing for

some adjustment to unit segmentation from the data itself during the preliminary scoring phase.

A scoring template for the passage was then generated. Scoring the recall protocols was fairly straightforward using this procedure, since many subjects recalled what they had read in nearly the same sequence as the passage. The sequential scoring template allowed the scorers to check for each item of information as they went along. Errors of spelling, punctuation, and grammar were not penalized.

2.3.3 Final weighting in scoring

Dorothea Dix left home / at an early age /--of her own free
3 2 2
will /--to live with her grandmother. / At fourteen, / Dorothea
1 2
was teaching school / at Worcester, Massachusetts. / A short
3 1
time after she had begun teaching, / she established a school for
2 3
young girls / in her grandparents' home. / Stress was placed
2
on moral character at Dorothea's school, / which she conducted
3
until she was thirty-three. / She was forced to give up teaching/
2 3
at her grandparents' home, / however, / when she became ill./
1 1 2
A few years of inactivity followed. / In 1841/ Dorothea began
2 2
to teach again, / accepting a Sunday school class / in the East
3 2
Cambridge, Massachusetts, / jail. / Here, / she first came upon
1 3 1 3
insane people / locked up together with criminals. / In those
3 2
days / insane people were treated even worse then criminals. /
2
There were only a few asylums / in the entire country./
3 2

Therefore / jails, / poorhouses, / and houses of correction were
1 3 3 3
used to confine the insane. / Dorothea Dix made a careful
3
investigation / of the inhumane treatment / of the insane. /
3 3
It was considered unfeminine/ for a woman to devote herself to
2 2
such work / at this time. / But this did not stop Dorothea Dix /
1 3
in her efforts to provide medical care / for the insane. /
2 2
because of her investigation, / conditions were improved. / More
2 2
than thirty mental institution were founded / or reestablished /
3 3
in the united states / because of her efforts. / Dorothea also
2 2
extended her investigations / to England / and to other parts of
3 2 2
Europe./ During the Civil War, / Dorothea served as
2
superintedent of women hospital nurses / in the Union army. /
3 2
When the war was over, / she returned to her work / of
2 2
improving conditions / for insane people.
2 2

2.3.4 Inter-rater reliability

In order to measure inter-rater reliability, the students' protocols were scored independently by two pairs of raters. The first pair were the primary scorers who generated the scoring template. The second consisted of two native speakers of English, as previously described. The investigator demonstrated to the raters both the scoring procedures and use of the scoring template.

Inter-rater reliability was calculated on the first and second pairs of raters. The Pearson correlations between the two independent raters in the first pair was .98, and for the second pair .94. These results

suggest that reading recall protocols were reliable in measuring students' reading comprehension ability. Inter-rater reliability estimates are shown in Table 1.

Table 1. Inter-rater reliability of reading recall protocols--Correlation matrix

	1st Pair of Raters		2nd Pair of Raters	
	Rater 1	Rater 2	Rater 1	Rater 2
Rater 1	1.00	.98	1.00	.94
Rater 2	.98	1.00	.94	1.00

2.3.5 Correlation with the TOEFL

Correlation between the reading recall protocols and the TOEFL was measured. The results indicate that the students who scored higher on the reading recall protocols also performed better on the TOEFL reading subtest, as evidenced by the significant positive correlation coefficient ($r=.68$, $p<.01$).

3. Discussion

The only serious problem in scoring had to do with ambiguous inferences, often caused by ambiguity in the text itself. For instance, the sentence 'she was forced to give up teaching at her grandparents' home, however, when she became ill.' was one such ambiguous unit. Some students construed this as her grandparents forcing her to give up teaching when she became ill. It is possible to understand the sentence that way, although most of the students recalled that it was her illness that forced her to give up teaching. Credit was given for either form.

Another sentence, 'During the Civil war, Dorothea served as superintendent of women hospital nurses.' also created ambiguity because it was not clear what she was in charge of. Was she in charge of all the nurses at one hospital, or all women nurses at all hospitals,

or just women nurses at one women's hospital? The students provided multiple versions of this unit. Thus, it was decided to discard this unit in the scoring procedure because it was virtually impossible to score it reliably.

There were two pausal units that contain more than one frequently recalled item. In the pausal unit, 'an East Cambridge, Massachusetts, jail,' some students recalled the locale without the jail; many recalled the jail without the locale. Credit was given for either, but no extra credit was given for recalling both. In the other unit, 'more than thirty mental institutions were founded,' some students recalled the number but not the type of institution; others recalled that mental hospitals or insane asylums were started, but without the number. Credit was given for either, but again, no extra credit was given for recalling both in this unit.

The raters were, however, able to resolve most of these problems in the scoring conference after trying the template on the first ten subjects. During this conference, the raters briefly discussed each scoring unit and wrote down examples of acceptable phrases. The problems with disjunctive categories were also resolved, as mentioned above. The raters finished scoring quite efficiently with the guidelines at hand.

The biggest advantage of this scoring procedure is that it is simple and easy to use after the raters worked together to generate the scoring template. The procedure can also be used with any passage in almost any context. Moreover, unlike multiple-choice items in which reading comprehension measurement can be contaminated by guessing, recall protocols discourage guessing and measure reading comprehension in an objective manner. Casanave(1988) noted that recall protocols are "relatively easy to score and use as a basis for inferences about comprehension" (p. 284). Thus, the claim that it is unrealistic for classroom teachers to score recall protocols(Wells, 1986) could be in part criticized.

On the other hand, disadvantages of the scoring procedure include the

amount of time needed to segment and weight the passage. The entire process of segmentation and weighting took five to six hours, including the time spent soliciting volunteer readers and summarizers. Apparently, the scoring procedure also requires some training and experience on the part of the raters.

Assigning weights to meaning units in a passage quickly reveals the rhetorical strengths and/or weaknesses of the text. It may be helpful to try segmenting and weighting a passage before using it in a recall procedure. In a related project, Vann and Schmidt(1993) recommend choosing a passage for its clarity of rhetorical organization to save time in the scoring procedure.

It is also a problem that there is no way to separate memory effects on subjects' comprehension. One of the main criticisms regarding recall protocols is that they only foster students' local comprehension because their comprehension processes tend to focus on the details of the text rather than on the main ideas(Wolf, 1991).

4. Suggestions for further research

There are some criteria with which one can make a decision in evaluating appropriateness of tests. According to Henning(1987, pp. 10-13), ten essential criteria can be used: test validity, test difficulty, test reliability, test applicability, test relevance, test replicability, test interpretability, test economy, test availability, and test acceptability. In order to adopt reading recall protocols appropriately in measuring reading comprehension, it is necessary to take these criteria into consideration. Any criteria that could not be directly applicable to reading recall protocols are excluded in the following discussion.

4.1 Test validity

A test is said to be valid when it fulfills its intended purpose in

measurement. In relation to reliability, Henning(1987, p. 89-90) points out that "it is possible for a test to be reliable without being valid for a specific ^purpose, but it is not possible for a test to be valid without first being reliable." In empirical methods of determining the validity of a test, it has been found that an increase in reliability will result in increase in validity, thus suggesting that validity is closely related to the reliability of a test.

There are several kinds of validity to be considered in developing an appropriate test. In the present study, however, it is beyond its intended purpose to examine all types of validity. As a reading recall protocol was administered to students taking a reading course, it was likely that the content of the test was sufficiently representative, and therefore ensured face validity. The moderately high correlation coefficient ($r=.68$ $p<.01$) between the reading recall protocol and the TOEFL reading subtest indicated the existence of concurrent validity. The significantly high inter-rater reliability coefficient($r=.98$ & $r=.94$) also implied the construct validity of reading recall protocols.

Considering that some reading specialists(Lee & Ballman, 1987; Lee, Ballman & Wolf, 1988) criticize reading recall protocols as tests of the micro-level recall aspect of comprehension rather than macro-level global comprehension, more empirical data should be collected and mathematically analyzed in order to ensure other types of validity in reading recall protocols.

4.2 Test difficulty

The difficulty level of reading passages and the familiarity of vocabulary would usually determine the appropriateness of reading recall protocols. The length of the reading passage might be an additional indicator of the test difficulty. As Cha's(1995) study found that students' reading performance was better on longer texts than on shorter passages, the length of the reading text has a significant influence on reading comprehension. It is, then, difficult to formulate

how long the passage should be to make reading recall protocols an adequate level of testing, though it is suggested that the text be "sufficiently long to ensure the involvement of macro processes in comprehension"(Kintsch & van Dijk, 1978, p. 376). According to Henning(1987), either critical inspection or piloting of the test would make it possible to come up with some estimate of the extent to which the text length is appropriate to the test.

4.3 Test reliability

Higher test reliability can be established when there is relatively less measurement error. In reading recall protocols, measurement error can be decreased by ensuring consistency of estimates on the part of the raters. It is, however, likely that raters making subjective estimates tend to be inconsistent in judgment because their judgment is often influenced by rater fatigue, the quality of examinee handwriting, and the personal relationship with examinees when their names are revealed(Henning, 1987). To minimize such influence in the scoring process, two raters should be called upon and make the estimates independently. If there is any large discrepancy between the raters, it is necessary for the raters to repeat the estimates to ensure high inter-rater reliability. High inter-rater reliability can also be obtained by providing detailed scoring guidelines and selecting raters with same levels of experience in subjective scoring.

4.4 Test applicability and replicability

In terms of test applicability, teachers or test administrators can use reading recall protocols appropriately in any reading class to measure students' reading comprehension, because the nature of the test is suitable. It is unlikely that the format and features of the test would be unfamiliar to the students in reading classes. In order for the test to be replicable, the teacher or the test administrator could employ the same

reading text over time. Once the scoring templates are developed, it is possible to replicate the tests in equivalent class levels and to compare students/classes from one administration to another.

4.5 Test relevance

The teacher or test administrator can increase the relevance of the test by considering the characteristics of the examinees. It would be relatively easier to develop the test with a highly adequate degree of relevance if the examinees have homogeneous characteristics, because the domain reflected in the test could be closely related to a particular group of students taught with the same objectives. For instance, when the examinees primarily consist of students studying science, it is desirable to select a reading text dealing with science-related information. In such cases, the students will find the test of reading recall protocols more relevant to the test domain as well as to the teaching objectives. If reading recall protocols are administered to a group of students with different interests in reading, the reading text should be selected based on the analysis of common denominators of their interests to ensure the relevance of the test .

5. Conclusion

There are certainly various ways of measuring reading comprehension, and perhaps they all have some limitations. Though reading recall protocols have been used widely as means of measuring reading comprehension, they have been criticized for their tendency to measure primarily bottom-up comprehension and to become a test of memory rather than of comprehension. In terms of test economy, they have also been criticized for their scoring procedures being too time-consuming.

The present study attempted to examine the reliability and validity of reading recall protocols that have been rarely investigated in second language reading research. The results of the study revealed that

reading recall protocols were likely to ensure content validity, concurrent validity and some construct validity. Moreover, it would not be as time-consuming as has been suggested if an appropriate scoring template is developed. The process of developing scoring templates might be less difficult for the teachers or scorers if the appropriate training is provided.

Compared to other reading comprehension tests that are often used for their convenience in administration and scoring, reading recall protocols have some disadvantages, as discussed previously. To employ them efficiently in measuring reading comprehension, further research should focus on verifying their validity with more empirical data. It is also necessary to develop other measures that can be used with reading recall protocols while compensating for their limitations.

References

Alderson, J. C. (1979). The cloze procedure and proficiency in English as a foreign language. *TESOL Quarterly*, 13, 219-227.

Been, S. (1975). Reading in the foreign langauge program. *TESOL Quarterly*, 9, 233-242.

Brown, A. L., & Smiley, S. S. (1977). Rating the importance of structural units of prose passages: A problem of metacognitive development. *Child Development*, 48, 1-8.

Cha, K-A. (1995). The effect of text length in the diagnosis of reading comprehension. *Journal of The Applied Linguistics Association of Korea*, 8, 249-275.

Cohen, A. D. (1984). On taking language tests: What the students report. *Language Testing*, 1, 70-81.

Connor, U., & Read, C. (1978). Passage dependency in ESL reading comprehension tests. *Language Learning*, 28, 149-157.

Lee, J. F., & Ballman, T. L. (1987). FL learners' ability to recall and rate the important ideas of an expository text. In B. BanPatten, T. Dvorak, & J. F. Lee (Eds.), *Foreign language elarning: A reasearch perspective* (pp. 108-118). Lowley, MA: Newbury House.

Lee, J. F., & Ballman, T. L., Wolf, F. (1987). Accounting for early stage foreign langauge learners' recall of passage information. A paper presented at the 1987 Conference of American Association of Teachers of Spanish and Portuguese, Los Angeles, August.

Henning, G. (1987). *A guide to language testing.* Boston, Massachusetts: Heinle & Heinle.

Hughes, A. (1989). *Testing for langauge teachers.* Cambridge: Cambridge University Press.

Kim, S-A. (1994). Korean EFL readers'ability to identify and recall the important information in L2 text, *English Teaching,* 49, 151-169.

Kintsch, W. & van Dijk, T. A. (1978). Toward a model of text comprehension and production. *Psychological Review,* 85, 363-394.

Nevo, N. (1989). Test-taking strategies on a multiple-choice test of reading comprehension. *Language Testing,* 6, 199-215.

Stevens, R. J. (1998). Effects of strategy training on the identification of the main idea of expository passages. *Journal of Educational Psychology,* 80(1), 21-26.

Swaffar, J. K., Arens, K. M., & Byrnes, H. (1991). *Reading for meaning: An integrated approach to language learning.* Engleswood Cliffs, NJ: Prentice-Hall.

Wells, D. R. (1986). The assessment of foreign language reading comprehension: Refining the task. *Die Unterrichtspraxis,* 19, 178-184.

School of Humanities and Natural Sciences
Korea Advanced Institute of Science and Technology
373-1 Kusong-dong, Yusong-gu, Taejon 305-701
E-mail: yypark@sorak.kaist.ac.kr
Tel: +82-42-869-4628
Fax: +82-42-869-2380