Asian Social Science
www.ccsenet.org/journal.html
Vol. 5, No. 6
June 2009

The Effect of Three Test Methods on Reading Comprehension: An Experiment

Feng Liu
School of Foreign Languages, Qingdao University of Science and Technology
Qingdao 266061, China
Tel: 86-532-8895-8959 E-mail: liufeng7079@163.com

Abstract

This study carries out an experiment to find out the effect of three test methods—multiple-choice questions, gap filling and short answer questions—on reading comprehension. It is found that the three test methods have a significant effect on reading comprehension, gap filling is the most difficult, while multiple-choice questions and short answer questions are easier. Both the low-proficient and high-proficient students are affected by test method, there is no interaction effect. It is also found that it takes the longest time to complete short answer questions and the shortest time to complete multiple-choice questions.

Keywords: Test method, Language proficiency, Reading comprehension

1. Literature Review

Many researchers have done a lot of studies on the effect of test methods on test performances. Bachman (1990) presents a framework for characterizing the facets of test methods that affect performance on language tests. These facets can be grouped into five sets: (1) the testing environment, (2) the test rubric, (3) the nature of the input the test taker receives, (4) the nature of the expected response, and (5) the relationship between input and response. He argues that test performance is affected by the characteristics of the method used to elicit test performance, and that constructed response types are generally more difficult than selected response types.

Shohamy (1984) found that test methods influenced how readers performed on a test of reading comprehension, and that multiple-choice questions were easier than open questions, and the effect was stronger on low-proficient readers. Wolf (1993) carried out a similar experiment, he also concluded that multiple-choice questions were easier than open-ended questions.

Samson (1983) used multiple-choice questions, open-ended questions, and summary tests in a reading comprehension test. The results showed there was no significant difference among the three test methods, so she concluded that the three test methods all tested the same ability or trait of the subjects. But she did find that multiple-choice questions were the easiest, and summary test the most difficult.

In China, much research has been done on testing reading comprehension. Chen & Cao (1999) argue that short answer questions are more effective than multiple–choice questions in testing reading comprehension.

Liu (1998) used multiple-choice questions, true or false questions and short answer questions in his reading comprehension test. The results showed that test methods affected the subjects' performance on reading comprehension tests, and that high-proficient students were more easily affected than low-proficient students. There were significant differences among the scores elicited by the three different test methods, short answer questions were the most difficult.

However, using the same three test methods in a reading comprehension test, Sun (2001) drew different conclusions. Sun did her test on Grade Two Junior Middle School students (Note that Liu's subjects were Grade Three English majors in university), she found that there were no significant differences among the three test methods.

In China, multiple-choice questions, gap filling, and short answer questions are widely used in many large-scale English tests, such as CET-4, CET-6, TEM-4 and TEM-8. Though the term "cloze" is used in the CET tests, for the reasons presented in Chapter 3, the so-called cloze tests in CET should be labeled gap filling tests. As discussed above, there are no conclusive ideas about the effect of test methods on reading comprehension.

2. Methodology

2.1 Research Questions

(1) Are there any significant differences among the reading comprehension scores measured by the three test methods (i.e. multiple-choice questions, gap filling, and short answer questions)?

(2) Is the effect of test methods affected by language proficiency or is there an interaction effect between test method and language proficiency?

2.2 Subjects

The subjects are 96 Grade-two English majors selected randomly from six classes in the English Department, Qingdao University of Science and Technology (There are seven classes in Grade Three, with about 30 students in each class. One class has taken the pilot test).

The ninety-six students are assigned randomly to three test groups, Test Group One, Test Group Two, and Test Group Three, each group with 32 students. Based on the students' final English exam scores, each test group is divided into two sub-groups: the high-proficient group and the low-proficient group.

The one-way ANOVA of the final English exam scores shows that there are no significant differences among the three test groups ($F=20.75$ $p< .01$).

2.3 Instrument

Research tools

Three different test papers are used in the experiment, all of them contain the same four reading passages. Several days before the formal experiment, a pilot test is done. One class is selected from the seven Grade-three classes, and then the students are divided into three groups and are given the three test papers separately. After the pilot test, modifications and changes are made in Paper Two and Paper Three where necessary. Paper One remains unchanged as all the passages and multiple-choice questions are taken from the original CET-4 test papers.

In Paper One, the four reading passages and the five multiple-choice questions after each passage are taken from CET-4 examination papers. Paper Two contains the same four passages, and there are fifteen blanks in each passage (the blanks are created by deleting the original words selectively). For each blank, there are four choices after the passage. In Paper Three, each of the four passages is followed by five questions or incomplete statements, which are converted from the multiple-choice questions in Paper One.

Variables

Independent variable: test method

Dependent variable: reading test scores

Moderator variable: language proficiency

The statistical package SPSS 11.0 for windows is used to analyze the data in this experiment.

2.4. Procedures

This experiment is done about one week after the TEM-4 test. Paper One is given to Test Group One, Paper Two is given to Test Group Two, and Paper Three to Test Group Three. All the three test groups take the test at the same time.

In Paper one, two marks are given for every correct answer, and zero if wrong, the total score is 40. In Paper Two, one mark is given for every right answer, and zero if wrong, so the total is 60 (raw score), the students' raw scores in Paper Two are converted into the final reported scores by multiplying 2/3. All the scores of Paper Two mentioned below are the converted scores. In Paper Three, a mark from zero to two is given to each answer, depending on the ideas and meaning provided by the students. Grammatical errors and misspellings are not considered as long as they don't affect understanding. The total score is also 40.

According to Wood (1993), the correlation between the marks awarded to the same script by the same examiner on two different occasions is usually greater than that between different markers. Therefore, the students' answers in Paper Three are marked by the same rater (the author in this case) twice. The average score for each item is calculated, which is the student's final score. In marking, the rater writes his marks on a separate piece of paper, not on the answer sheet, to avoid any interference.

3 Results

3.1 Reliability and Validity

Reliability is one of the main criteria when assessing a test. Kuder-Richardson formula 21 and Cronbach's alpha are used here to assess the three test papers. The results show that the three tests are reliable:

Insert Table 1 About Here

Insert Table 2 About Here

The concurrent validity of the three test papers is also computed. The subjects' scores in this experiment are compared with their scores on the reading part in TEM-4. The results show that the three test papers have high concurrent validity:

Insert Table 3 About Here

3.2 The Difficulty of the Test

Table 4 displays the mean scores and standard deviations of the three test papers. It shows that gap filling has the lowest mean score (25.503), i.e. it is the most difficult of the three test methods, while multiple-choice questions and short answer questions are roughly at the same difficulty level (28.938 and 29.287 respectively). At the same time, multiple-choice questions has the highest standard deviation (4.7379), which indicates that the scores are more widely spread.

Insert Table 4 About Here

3.3 The Effect of Test Methods

Table 5 shows that test methods have a significant effect on test scores, there are significant differences among the scores elicited by the three test methods ($F=16.021$ $p<.01$), but it doesn't tell which method differs from which.

Insert Table 5 About Here

Table 6 gives a detailed analysis of the method effect (As there are equal subjects in each test group, Tukey HSD is more appropriate). It can be seen that gap filling is significantly different from the other two methods, while there is no significant difference between multiple-choice questions and short answer questions.

Insert Table 6 About Here

The effect of test methods can be more easily shown in Table 7. In the subset columns, the factor levels that do not have significantly different effects are displayed in the same column. In this case, the first column contains gap filling, and the second column contains multiple-choice questions and short answer questions.

Insert Table 7 About Here

3.4 The Effect of Language Proficiency

As described earlier, each test group is divided into a high-proficient group and a low-proficient group. According to Table 4, in each test group, the mean score of the high-proficient group is higher than that of the low-proficient group. Table 5 shows that the scores are significantly different ($F=47.084$ $p<.01$).

3.5 The Interaction Effect

If we look at Table 8, we find that there might be no interaction effect, as all the scores in each test group follow the same pattern (the mean score of the high-proficient group is higher than that of the low-proficient group by about 4 or 5 marks).

There are two parallel lines in Figure 1, this shows that there is no interaction effect.

Insert Figure 1 About Here

And this is further confirmed in Table 5 ($F=.048$ $p>.05$).

4. Discussions

As can be seen from the results, different test methods affect students' scores in reading comprehension. Gap filling test is the most difficult, while multiple-choice questions and short answer questions are easier. Both the high-proficient and the low-proficient students are affected by the test method, there are no interaction effects.

In this gap filling test, the students are asked to choose the correct answer from four options provided by the test constructor. The mean score of gap filling (25.503) is significantly lower than that of the multiple-choice questions and short answer questions (28.938 and 29.287 respectively). In a gap filling test, the words are not deleted mechanically, but are deleted on the basis of the test constructor's rational and subjective judgement.

Numerous research has been done on cloze tests, in which the deletions are made mechanically, from every fifth to eleventh word. Research has shown that cloze test is a reliable and valid test of reading ability as well as overall language ability. However, many people doubt that it can test reading comprehension at a higher level, since it seems to test reading comprehension only at a sentence level. Many cloze items are not constrained by the context above the sentence, but by the immediately adjacent words or phrases. Some researchers (e.g. Alderson, 1978, cited in Weir, 1990) even claim that cloze is essentially sentence bound.

Gap filling test is not without this problem. The two examples below are taken from the original "cloze" test in CET-4

(January, 2002):

…sitting in the theatre I had to look through the 72 between the two tall heads in front of me. …I just heard the 81 of the popcorn crunching between my teeth…

72 A) crack B) opening C) break D) blank

81 A) tone B) voice C) sound D) rhythm

These two items can be answered without referring to the rest part of the passage as long as the reader understands the two sentences.

Sometimes even if students understand the main ideas of a passage, they are still unable to answer some items which are based on vocabulary differentiation, this may account for the low mean score of the gap filling test in this experiment. One day after the experiment, several students who have taken the gap filling test are asked for an interview. They say that although they can understand the main ideas of the passages, they still find it difficult to complete some of the items. Most of them give the following example to show that even they know the meaning of the sentence they still can't make the correct choice:

…Shrinking land 11 and rising costs for burying and burning rubbish are …

11 A) room B) place C) space D) spot

Multiple-choice questions and short answer questions are considered to be qualitatively different by many language constructors. Many believe that multiple-choice questions test only recognition knowledge. Students may get an item right even if they don't understand it, while in short answer questions, students have to understand the text before they can correctly answer the questions.

In this experiment, the multiple-choice questions and the short answer questions have nearly the same mean scores (28.938 and 29.287 respectively). Contrary to the intuitive thinking, short answer questions are not necessarily harder than multiple-choice questions. This may be due to the fact that in a multiple-choice test, candidates are presented with four choices, but they might have not thought of some of the choices. Candidates have to make a choice among the four options, sometimes this may make the task difficult. In short answer question tests, candidates usually can answer the question correctly if they understand the text.

There is also the factor of guessing in multiple-choice questions, so a high-proficient student might get a higher score than would otherwise in the other two methods (not all the high-proficient students), similarly a low-proficient student might get a lower score than would otherwise in the other two methods (not all the low-proficient students). In this experiment, the highest score (36) and the lowest score (18) all appear in the multiple-choice question test. Due to this fact, the multiple-choice test has a higher standard deviation (4.7379) than gap filling (2.8098) and short answer questions (2.8827).

In this experiment, it is found that it takes the longest time to complete short answer questions and the shortest time to complete multiple-choice questions. In short answer questions, the students first need to understand the text, they spend most of the time reading the text, then they write down their answers, while what the students need to do in multiple-choice questions is to choose one answer, they focus more on the provided options. In making the choice, some test-taking strategies or skills, such as guessing, deduction, also play a part.

This study shows that the three test methods have a significant effect on reading comprehension test scores, both the high-proficient and low-proficient students are affected by this effect. In large-scale tests, multiple-choice questions are widely used because they can be marked reliably and economically, but at least this kind of test lacks face validity. Short answer questions are less widely used in reading comprehension tests, critics say it's hard to achieve high rater reliability and marking is not economical, especially in China with millions of candidates taking a test. But as is shown in this experiment and as is indicated by many others (e.g. Heaton, 1988), high rater reliability is still possible.

But just as Alderson (2000) notes, it is inadequate to measure the understanding of text by only one method, and that objective methods can be supplemented by more subjective evaluation techniques. Good reading tests are likely to employ a number of different methods. Given the monopoly position of multiple-choice questions in reading comprehension tests particularly in China, it is hoped that the understanding of the test method effect will help improve this.

References

Alderson, J. C. (2000). *Assessing reading.* Cambridge: Cambridge University Press.

Alderson, J. C., Clapham, C., & Wall, D. (1995). *Language test construction and evaluation.* Cambridge: Cambridge University Press.

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

Chen, Hua & Cao, Jun. (1999). The Feasibility of Subjective Answering Questions in Testing Reading Comprehension. *Shandong Foreign Language Teaching*. 2, 80-84.

Heaton, J. B. (1988). *Writing English language tests.* Pearson Education Limited. Beijing: Foreign Language Teaching and Research Press. 2000.

Liu, Jianda. (1998). The Effect of Test Methods on Testing Reading. *Foreign Language Teaching and Research*, 2, 48-52.

Samson, D. M. M. (1983). *Rasch and reading. In J. van Weeren (Ed.), Practice and problems in language testing.* Arnhem: CITO.

Shohamy, E. (1984). Does the testing method make a difference? *The case of reading comprehension. Language Testing*, 1 (2), 147-170.

Sun, Xiaoying. (2001). Do the Different Test Methods Have an Effect on the Reading Comprehension Scores? *Primary and Middle School English Teaching and Research*. 3, 26-28.

Weir, C. J. (1990). *Communicative language testing*. London: Prentice Hall.

Wolf, D. F. (1993). A comparison of assessment tasks used to measure FL reading comprehension. *The Modern Language Journal*, 77 (4), 473-489.

Wood. R. (1993). *Assessment and testing: A survey of research.* Cambridge: Cambridge University Press.

Table 1. Internal Consistency

KR-21		Cronbach's alpha
Paper One (multiple-choice questions) .89	Paper Two (gap filling) .81	Paper Three (short answer questions) .83

The intra-rater reliability of Paper Three is also calculated. Table 2 shows that the rating is reliable (r =.948 p<. 01).

Table 2. Intra-rater Reliability

		first rating	second rating
first rating	Pearson Correlation	1	.948**
	Sig. (2-tailed)	.	.000
	N	32	32
second rating	Pearson Correlation	.948**	1
	Sig. (2-tailed)	.000	.
	N	32	32

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

Table 3. Concurrent Validity

Test paper	Pearson validity correlation coefficient
Paper One (multiple-choice questions)	r =.81 p< .01
Paper Two (gap filling)	r =.76 p< .01
Paper Three (short answer questions)	r =.75 p< .01

Table 4. Mean Score and Standard Deviation

Dependent Variable: test scores

test method	language proficiency	Mean	Std. Deviation	N
multiple-choice questions	high	31.125	3.3441	16
	low	26.750	5.0000	16
	Total	28.938	4.7379	32
gap filling	high	27.463	1.6260	16
	low	23.544	2.3415	16
	Total	25.503	2.8098	32
short answer questions	high	31.350	1.2469	16
	low	27.225	2.5583	16
	Total	29.287	2.8827	32
Total	high	29.979	2.8552	48
	low	25.840	3.8141	48
	Total	27.909	3.9445	96

Table 5. Two-way ANOVA

Dependent Variable: test scores

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Corrected Model	691.985[a]	5	138.397	15.844	.000	.468
Intercept	74777.588	1	74777.588	8560.828	.000	.990
METHOD	279.882	2	139.941	16.021	.000	.263
PROFICIE	411.268	1	411.268	47.084	.000	.343
METHOD * PROFICI	.835	2	.418	.048	.953	.001
Error	786.137	90	8.735			
Total	76255.710	96				
Corrected Total	1478.122	95				

a. R Squared = .468 (Adjusted R Squared = .439)

Table 6. Multiple Comparisons

Dependent Variable: test scores

	(I) test method	(J) test method	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
Tukey HSD	multiple-choice questions	gap filling	3.434*	.7389	.000	1.674	5.195
		short answer questions	-.350	.7389	.884	-2.111	1.411
	gap filling	multiple-choice questions	-3.434*	.7389	.000	-5.195	-1.674
		short answer questions	-3.784*	.7389	.000	-5.545	-2.024
	short answer questions	multiple-choice questions	.350	.7389	.884	-1.411	2.111
		gap filling	3.784*	.7389	.000	2.024	5.545
Tamhane	multiple-choice questions	gap filling	3.434*	.9738	.003	1.030	5.839
		short answer questions	-.350	.9804	.979	-2.770	2.070
	gap filling	multiple-choice questions	-3.434*	.9738	.003	-5.839	-1.030
		short answer questions	-3.784*	.7116	.000	-5.531	-2.038
	short answer questions	multiple-choice questions	.350	.9804	.979	-2.070	2.770
		gap filling	3.784*	.7116	.000	2.038	5.531

Based on observed means.

*. The mean difference is significant at the .05 level.

Table 7. Homogenous Subsets

	test method	N	Subset 1	Subset 2
Tukey HSD[a,b]	gap filling	32	25.503	
	multiple-choice questions	32		28.938
	short answer questions	32		29.287
	Sig.		1.000	.884

Means for groups in homogeneous subsets are displayed.
Based on Type III Sum of Squares
The error term is Mean Square(Error) = 8.735.

a. Uses Harmonic Mean Sample Size = 32.000.

b. Alpha = .05.

Table 8. Mean Score Pattern

Dependent Variable: test scores

test method	language proficienc	Mean	Std. Error	95% Confidence Interval Lower Bound	95% Confidence Interval Upper Bound
multiple-choice questions	high	31.125	.739	29.657	32.593
	low	26.750	.739	25.282	28.218
gap filling	high	27.463	.739	25.995	28.930
	low	23.544	.739	22.076	25.012
short answer questior	high	31.350	.739	29.882	32.818
	low	27.225	.739	25.757	28.693

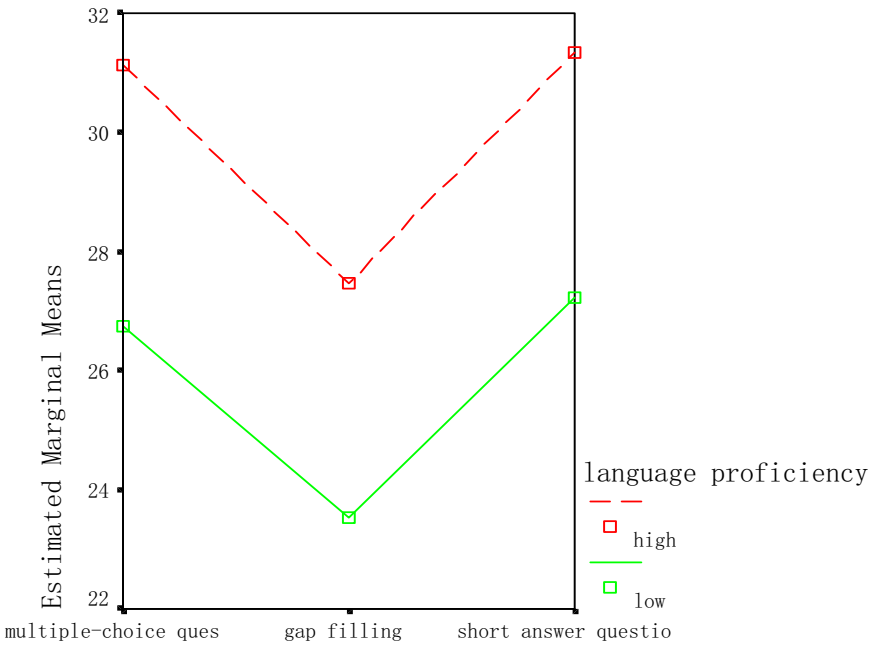

Figure 1. Profile Plot