The Reading Matrix
Vol. 7, No. 1, April 2007

L2 READING COMPREHENSION TEST IN THE PERSIAN CONTEXT: LANGUAGE OF PRESENTATION AS A TEST METHOD FACET

Mohammad Rahimi
mrahimy@gmail.com

Abstract

Test method facet has been considered as an important factor affecting the testee's performance on a test. That is, a test used to assess a particular ability would yield different results when different test methods are used to gauge the same trait. The language of presentation is an aspect of test method conceived of as affecting the performance of the testees on a language test. This study investigated whether presenting the items of an English reading comprehension test in the testees' native language (Persian) would affect their performance on the test. To this end, two versions of an English reading comprehension test--one with items in English (ERC) and the other with items in Persian (TRC)--along with a Persian reading comprehension test (PRC) were given to 193 English majors with different L2 proficiency levels--high, intermediate, and low--so that half of the subjects, as a whole and in each proficiency level, took ERC and the other half, TRC. In addition, all the subjects took PRC, too. The results indicated that the test method, on the whole, did not significantly affect the scores. However, the test method was found to affect the performance of low-proficiency subjects. That is, the low-proficiency group taking TRC outperformed the corresponding group taking ERC.

Introduction

According to Bachman (1990), a testee's performance on any language test is influenced by a large number of factors that must be taken into account in the construction and development of language tests. Accordingly, Bachman (1990) and later Bachman and Palmer (1996) presented a theory of language testing that contained not only different aspects of language ability but also the methods and other factors involved in the measurement of this ability. Bachman (1990: 81) states, "If we are to develop language tests appropriately, for the purposes for which they are intended, we must base them on clear definitions of both the abilities we wish to measure and the means by which we observe and measure these abilities."

One of the factors influencing test performance is test method facets. Test method, according to Bachman (1990: 111), is "the characteristics of methods used to elicit test performance." The methods we use to present the test to the testee, although intended to measure the same ability, might yield quite different results. In other words, an individual's performance on a language test may vary due to the influence of both his language ability and test method facets. Bachman (1990) presents a framework of test method facets consisting of five major categories including testing environment, test rubrics, nature of the input the test taker receives,

nature of the expected response to that input, and the relationship between input and response. A large number of studies have shown how different aspects of test method affect test performance (Bachman and Palmer, 1981; Katz, *et al.*, 1990; Anderson *et al.*, 1991; Perkins and Brutten, 1993; Jafarpur, 2003; Fulcher & Marquez Reiter, 2003, Kobayashi, 2004, to name a few). All these studies have unanimously indicated that the observed performance of a test taker is a representation of both his ability and test method facets. Consequently, the validity of any test must be interpreted with care since the score on the test is not just an indication of the test taker's ability that is purported to be measured. Thus, in order to increase the validity of the test we must try to minimize the effect of test method facets.

Language of presentation is one of the least, yet the most important, researched aspects of test method. This facet refers to the language through which the input, say, the items in a test, is presented to the testee. According to Bachman (1990), a very distinguishing feature of language testing is the fact that the measuring instrument itself is designed on the bases of the language. This makes the task of assessment very challenging for the foreign language professionals, particularly in the case of reading comprehension tests. The reading comprehension test is a case where the language of assessment (L1 vs. L2) can be an important factor affecting the validity of the test (Gordon and Hanauer, 1995; Lee, 1990; Lee and Ballman, 1987; Shohamy, 1984). Shohamy (1984) states that when the items of an L2 reading comprehension test are given in the testee's L1, he performs significantly better than in cases when all the input is presented in the target language. Consequently, his performance on the test may be more indicative of his reading ability and test method might be less involved. As a result, the test could yield more valid results. However, the above hypothesis is yet to be verified and empirically supported before it is adopted.

Literature reveals a few studies pertinent to the language learners' performance on reading comprehension tests with L1 vs. L2 as the language of input. These studies have aimed at investigating the effect of the language of presentation as a test method facet on the performance of the testees.

The leading study in regard to the language of presentation in reading comprehension test is the one conducted by Shohamy (1984). In this study, she investigated two aspects of test method--the test format and the language of presentation. She found significant differences between the performances of the subjects with different levels of proficiency on tests differing in terms of the test format and the language of presentation --multiple-choice or open-ended questions in the target language or the subjects' native language. She found that the language in which the items based on the reading passages were offered made a difference; that is, the items in the testees' L1 made the test easier. Of course, she found that the difference was more remarkable in low proficiency levels. In other words, the performance of the testees with low proficiency was found to be more sensitive to the test method, in general, and to the language of presentation, in particular. The reason, as she mentioned, might be that presentation of the items in Ll could reduce the subjects' anxiety which is more conspicuous among low-level subjects. She further stated that the use of items in L1 makes the test more 'authentic' since students usually translate the items in L2 into their first language. This claim is confirmed by Thomas (1997). In a verbal protocol analysis of a group of subjects taking a multiple-choice reading comprehension test, he found that low-level students frequently made use of their L1 to help them comprehend the passage and the items.

Alderson (2000) states that the items should be put in the testees' first language

since it makes the items easier and we can relate any flaw in the testees' performance to the passage not the questions,

> ... if the language of the questions is harder to understand than the passages themselves, the reader is presented with an additional layer of difficulty and we cannot tell whether poor performance is due to the passage difficulty or to that of the questions. The usual advice to test writers is to ensure that the language of the questions is simple, and certainly easier than the passage. This is often difficult for tests for beginning first-language readers. ... when test-takers share a first language, it might be better to ask questions in that language (Alderson, 2000: 86).

Gordon and Hanauer (1995: 302) remarked, "Having questions written in the L1 would facilitate the test takers' understanding of what is being asked in a particular item. This would decrease the chances of the test taker providing the wrong information due to misunderstanding the comprehension question." They noticed that when test tasks were in the subjects' L1, the information yielded through the test was much greater. Furthermore, Swaffer, Arens, and Byrnes (1991) claimed that a 'feasible' reading test should allow the readers' conceptualization of text meaning in their native language. They further asserted that the use of L1 in L2 reading tests can provide us with a more complete assessment of the learners reading comprehension ability. Lee (1986) found that if the learners' second language reading comprehension was assessed through their native language, they could show their comprehension more clearly. Similarly, Gordon and Hanauer (1995, p.302) stated,

> ... when test takers are presented with comprehension tasks in the L1 or are allowed to answer open ended comprehension questions in their native language, the problem of misunderstanding or not fully understanding tasks is eliminated and they benefit from the opportunity to express the meaning they have constructed without being inhibited by poor reading or writing ability in the L2.

They, however, believed that a major disadvantage of presenting the items in the target language may be the fact that it may become a rich source of knowledge.

Godev *et al.* (2002) investigated the effect of the language of presentation in a Spanish reading comprehension test for the English learners of Spanish. The results showed that when the open-ended items of the reading comprehension test were presented in the native language of the learners and they were allowed to answer in their L1, the test appeared to be a better assessment instrument of their reading ability.

On the other hand, in a study on French learners of English, Donin and Silva (1993) noticed that the language of input (in French) did not have any significant effect on the subjects' performance on their English reading comprehension test.

None the less, some scholars believe that the testees' performance on a reading comprehension test in the L2 is related to their proficiency in their native language reading ability. For instance, Bernhardt and Kamil (1995) claimed that reading performance in a second language is, to a large extent, related to the reading ability in the first language. Markham (1985) studying English learners of German and Lee and Musumeci (1988) studying Italian learners of English found that L1 reading ability of the subjects influenced their second language reading test scores.

Bernhardt and Kamil (1995) found that first language reading ability is a very important variable in second language reading achievement. Carrell (1991), based on the results of her study, claimed that though both first language reading ability and L2 proficiency have significant effects on L2 reading ability, in foreign language situations, L2 proficiency is

more important, whereas in second language contexts first language reading ability accounts for a greater proportion of the variance in L2 reading ability. Furthermore, Aebersold and Field (1998) noticed that the level of fading proficiency in the L1 also appears to be a factor in the learner's development of L2 reading skills.

Significance of the study

Two important points make the present study significant. Firstly, the research in this specific area, i.e., the language of presentation, is very limited and, as mentioned above, more research is needed to establish it as a test method facet. To make the present study more significant, no study has so far been conducted with a standard multiple-choice reading comprehension test with a relatively large number of items. Most of the studies have been conducted with open-ended question and/or with a small number of items.

Secondly, no study of this sort has ever been conducted in an EFL context such as Iran. The very specific feature of language learning in Iran makes the research in this area quite valid and significant. In fact, the EFL learners in Iran have basic problems with the conventional reading comprehension tests due to the unique features of the language learning curriculum in Iran. Since the prevalent teaching method is still the Grammar Translation method, teachers spend most of the class time translating the reading texts into Persian and one can see no instance of teaching the skills and strategies needed for effective reading comprehension. That is why the students usually have problem with multiple choice tests of reading comprehension as they do not know what exactly they are asked to do. Of course the tests, too, are usually based on the factual information directly stated in the text and thus requiring the students to map the items with the texts.

In university, the problem is not usually solved and the same procedure for the teaching of reading comprehension is applied. However, to make the problem more serious, the tests, usually taken from original books intended for EFL/ESL learners, assess higher level skills of reading such as inference making, word guessing, etc. The items, being in English, make the task of taking the tests more difficult. That is why most students often complain that they have problem understanding the items although they might not have much problem understanding the texts. Therefore, it can be hypothesized that giving the items in their native language might make the test more comprehensible for them and the results might be a better and more precise indication of their ability.

The present study, accordingly, seeks the answer to the following questions:

1. *Does it make any difference if the items in an L2 multiple-choice reading comprehension test are given in L1 rather than L2?Does it have different effects on the performance of the testees at different levels of proficiency on a reading comprehension test?*
2. *If the answer to the first question is yes, is it due to the test method effect or the testees' reading ability in the first language?*
3. *If the test method is found to influence the testees' performance on a reading comprehension test, is it better to give the test items in L2 or in L1?*

Method

Subjects

The subjects of the study were 193 English majors, both males and females, from three universities in Fars province, Iran. They were divided into three proficiency levels, i.e., low, intermediate and high based on the scores they obtained on a Test of English as a Foreign Language (TOEFL). The low group consisted of 60 students; the intermediate group, 67; and the high group, 66 students.

Instruments

The instruments of the study were three reading comprehension tests. Two of the tests were in fact two versions of the same test. Both comprised three passages (they were the same in both tests) taken from the ETS (1995) on neutral topics--population growth in Canada, microbe hunters, and deep ocean drilling. The passages were accompanied by 33 multiple-choice items in each test. However, in one of the tests the items, as in the original version, were in English--from now on the test is referred to as ERC (the English reading comprehension version)--whereas in the other, the items were given in Persian (the native language of the testees)--hereafter referred to as TRC (the translated reading comprehension version)--(see the appendix). The items of TRC were the translated version of those of ERC (translation was carried out by the researcher and checked by a colleague). In order to insure the validity of TRC, the translated items along with the original ones were given to two other professors of English to compare. Problematic items in TRC which were believed by both reviewers not to convey the same ideas as the English versions were modified.

The third test was a Persian reading comprehension test constructed by Vatankhah (1991) to measure the reading comprehension skill in Persian—the test is referred to as PRC, hereafter. The test consisted of 16 excerpts on a variety of topics with 40 multiple-choice items.

Procedure

The three tests were given to each of the three groups of proficiency levels in two different sessions; in one session, ERC and TRC were given to the subjects in each proficiency level in such a way that half of the participants received ERC and the other TRC in a random manner. Then, in the second session, all the subjects in each group took PRC. In order to eliminate the order effect, ERC and TRC as well as PRC were given to the subjects in a counterbalanced manner so that one group of the subjects first took ERC and TRC and the next session PRC, and the next group, vice versa.

Data Analysis

Descriptive statistics were calculated for the following cases; for the scores of the two major groups, i.e., all the subjects taking ERC and all those taking TRC; the scores of the participants taking ERC and TRC in each proficiency level--high, intermediate, and low; and the scores of all the participants taking PRC.

An independent t-test was run to see if there was any significant difference between the performances of the two major groups of the subjects--those taking ERC and the ones taking TRC. Three independent t-tests were run to see if there was any significant difference between the scores of the two groups of the subjects, one taking ERC and the other TRC, in each proficiency level to see if there was any difference between their performances.

Another independent t-test was run between the PRC scores of the two groups of low-proficiency subjects --those who had taken ERC and those who had taken ERC. In addition, the correlation coefficients between the ERC/TRC scores of the subjects in the low proficiency level and their PRC scores were calculated. Finally, two one-way tests of ANOVA were run to see if there were any significant differences between the performances of the subjects at the three proficiency levels on ERC and TRC--once with the scores on ERC and next with those on TRC.

Results and discussion

Descriptive statistics

Table 1 shows the descriptive statistics for the overall performance of the two major groups of subjects on ERC and TRC.

Table1: Descriptive statistics for the results of TRC and ERC (N=193)

Statistics	ERC	TRC
K	33	33
X	20.19	20.72
SD	5.01	4.61
Range	20	18
Skewness	0.11	0.13
Kurtosis	-0.82	-0.49
KR-21	0.60	0.53

The mean score for ERC is 20.19 and that of TRC 20.72. The standard deviation for ERC is 5.01 but that of TRC 4.61 indicating less variance among the scores of the participants on TRC. The difference between the range of the scores on ERC and TRC (20 and 18, respectively) shows the fact that the subjects have performed rather more homogeneously on TRC. The distribution of the scores on the two tests is rather normal; with respect to skewness, 0.11 for ERC scores and 0.13 for TRC; with respect to kurtosis, -0.82 for ERC and -0.49 for TRC. The reliability coefficients of the two tests are very close, too; 0.60 for ERC and 0.53 for TRC. All these statistics indicate very minute differences between the results of the two tests.

Table 2 shows the descriptive statistics for the scores of the two groups, i.e., those taking ERC and the ones taking TRC in each proficiency level, i.e., low, intermediate, and high.

Table 2: Descriptive statistics for the results of TRC and ERC in different proficiency levels

	ERC			TRC		
Proficiency	High (n=34)	Int. (n=31)	Low (n=29)	High (n=32)	Int. (n=36)	Low (n=31)
K	33	33	33	33	33	33
X	25	19.36	17.60	24.78	20.47	19.36
SD	3.61	3.18	3.18	3.58	3.24	4.06
Range	14	12	12	13	13	18
Skewness	-0.90	0.50	0.50	-0.10	-0.51	1.02
Kurtosis	0.87	-0.04	-0.04	-0.17	0.22	1.52

According to the table, in the high and the intermediate proficiency levels, the mean

scores obtained on ERC and TRC are very close to each other; in the high group 25 for ERC and 24.78 for TRC; in the intermediate group 19.36 for ERC and 20.47 for TRC. This, however, does not stand true for the low group. That is, the difference is much larger, 17.60 for ERC but 19.36 for TRC. Standard deviations for all the groups, on the other hand, are more or less close, except for the low group in which the standard deviation of ERC is slightly lower than that of TRC (3.18 and 4.06, respectively).This consistency indicates that, irrespective of the level of the subjects and the kind of test they have taken (whether ERC or TRC), the variation among the testees' performances has remained unchanged. Ranges, too, are very close to each other, except for the ranges of the ERC and TRC scores for the low group (12 and 18, respectively). However, looking at the distribution of the scores on TRC in the low group, one can find that the highest score after 29 is 24. Thus, if we remove the extreme score, i.e., 29, the range would be reduced to 13, not so much different than others. As for the distribution of the scores, in only two cases the distribution is not normal--the distribution of the scores of the high group on ERC (skewness=-0.90) and that of the low group on TRC (skewness=1.02). The former shows that ERC has been rather easy for the advanced group, even easier than TRC, and thus presenting the items in the testees' native language has not made the test easier for this group. The latter case indicates that TRC has been rather difficult for the low proficiency testees, even more difficult than ERC, whereas the means of the scores obtained by the two groups of low proficiency subjects indicate a contradictory result. This again was envisaged as being due to the existence of the extreme score (29) in the scores on TRC. Interestingly, the elimination of this score from the distribution resulted in an almost normal distribution (skewness was found to be 0.39). In all other cases, as the table shows, the distributions are rather normal.

Finally, the distribution of the scores for the high group in ERC is peaked (kurtosis=0.87), which indicates a rather homogeneous performance on the part of this group. Similarly, the distribution of the scores of the low group on TRC, too, is peaked but with a higher index (kurtosis=1.52). This indicates that the subjects have performed more homogeneously on TRC. However, like the previous case, the elimination of the extreme score resulted in a very different index, i.e., -0.43, that indicates a more normal distribution. As for the intermediate group, both cases are almost normal (-0.04 for ERC and 0.22 for TRC). All in all, except for the differences in means, in all the cases the performances of all the proficiency groups on both tests are rather similar, except for the performance of the high proficiency group on ERC, which shows a more homogeneous performance than other groups.

Table 3 illustrates the descriptive statistics for the scores of all the subjects on PRC.

Table 3: Descriptive statistics for the results of PRC (N= 193)

Statistics	PRC
K	40
X	22.75
SD	3.15
Range	16
Skewness	-0.11
Kurtosis	-0.22
KR-21	0.35

According to the table, the mean for the scores of all the testees on this test is 22.75. The standard deviation is 3.15. The range is 16. The distribution of the scores is almost normal with respect to both skewness and kurtosis indices. The reliability coefficient is 0.35. Of course, the reliability coefficients reported by Vatankhah 1991 (the source from which the test was taken) are 0.77 (split-half) and 0.54 (KR-20). The low reliability of the test in the present study might be due to the fact that the subjects were more homogeneous than the participants in Vatankhah's study. In fact, the participants of Vatankhah's study were comprised of high school and university students from different levels of study and different majors. However, in the present study, the participants, as mentioned above, consisted of university students studying the same major, and not so much dispersed with regard to the age range. Consequently, there has not been much variance among the testees, resulting in a low reliability coefficient.

Research questions

1. *Does it make any difference if the items in an L2 multiple-choice reading comprehension test are given in L1 rather than L2?Does it have different effects on the performance of the testees at different levels of proficiency on a reading comprehension test?*

In order to see if there was any difference between the overall performance of the two groups of the subjects on the two tests, an independent t-test was run. Table 4 represents the results of the t-test.

Table 4: Results of independent t-test between the scores on ERC (n=94) and TRC (n=99)

Test	X	SD	T-value	P-value
ERC	20.19	5.01	-6.25	0.33 (ns)
TRC	20.72	4.61		

As the table indicates, no significant difference was observed between the scores of the subjects taking ERC and those of the subjects taking TRC (P>0.05). This shows that presentation of the items of the reading comprehension test in the subjects' native language did not significantly affect their performance. The results echo those of Donin and Silva (1993) in that the language of presentation did not affect the testees' performance on a reading comprehension test. However, they contradict the results of Godev *et al.* (2002) and Shohamy (1984).

A likely explanation for the lack of the difference between the scores of the two major groups of the participants would be that the passages and the items were so easy for the subjects that even presenting them in the target language did not pose much difficulty for the participants understanding them. In order to test this hypothesis, the two tests were item analyzed. The results of the item analysis for the two tests presented in Table 5 reject this hypothesis as the item facility indices of the two tests were rather reasonable, i.e., not too easy, not two difficult. An independent t-test was run to see if there was a significant difference between the mean item facility indices of ERC and TRC, on the one hand, and the mean item discrimination indices of the two tests, on the other.

Table 5: Summary item analysis for ERC and TRC

Item statistics	X ERC	Range	X TRC	Range	Difference in means
Item facility	0.57	0.93-.15=0.78	0.58	0.95-.08=0.87	ns
Item discrimination	0.32	0.77-0.00=0.77	0.27	0.77-(-0.15)=0.92	ns

As can be seen in the table, no significant difference can be observed between the overall item facility indices of the two tests (0.57 for ERC and 0.58 for TRC). In fact, as mentioned above, the items of the two tests show a similar level of facility/difficulty. The case for overall item discrimination indices of the two tests is the same, i.e., there is no significant difference between them, although that of ERC is slightly larger.

Of course, since the item facility and item discrimination indices are pertinent to the whole test and not the individual items, we cannot say with certainty that presenting the items in the testees' native language does not affect the testees' performance on the test by reducing the test method effect. Hence, it was hypothesized that this lack of the difference between the results of the two tests could have been due to the particular item types included in the test. In order to see to what extent this hypothesis was true, the items, on the basis of the skill they measured, were divided into six groups (Table 6). Then, the average item facility index of each group of items for both ERC and TRC was measured to see if there was any difference between the item facility indices of the corresponding groups in ERC and TRC.

Table 6: A comparison of the average IF indices of different groups of items measuring different skills in ERC and TRC

Skill	No. of items	No. of items with a higher IF in ERC	No. of items with a higher IF in TRC	Average IF in ERC	Average IF in TRC
Getting the main idea	2	1	1	0.65	0.66
Getting the explicit ideas	9	5	4	0.65	0.54
Guessing the word meaning	10	1	9	0.42	0.55
Understanding the implicit ideas or inferences	6	-	6	0.40	0.56
Distinguishing the reference	5	3	2	0.87	0.66
Understanding the tone of the writer	1	-	1	0.29	0.58

As Table 6 illustrates, three groups of items--those related to *guessing the word meaning*, *understanding the implicit ideas*, and *understanding the toneof the writer*-- showed higher item facility indices in TRC than in ERC; in the case of g*uessing the word meaning,* in 9 out of 10 cases one can see higher item facility indices in TRC than in ERC (the average item facility index for this group of items in TRC was 0.55 as compared to that of ERC, 0.42); the same was

true for the items measuring the testees' *understanding of the implicit ideas*—for all the items the IF indices in TRC were higher than those of ERC (the average being 0.56 for TRC as compared to 0.40 for ERC); and finally in the case of *understanding the tone of the writer* (only one item), the item facility index for TRC was higher than that of ERC (0.58 and 0.29, respectively). On the other hand, in two groups of items, i.e., *distinguishing the reference* and *getting the explicit ideas,* the item facility indices in ERC were higher than those in TRC (the average being 0.65 for ERC and 0.54 for TRC in the case of *getting the explicit ideas* and 0.87 for ERC and 0.66 for TRC in the case of *distinguishing the reference).*

A likely explanation for the higher item facility indices of the first three groups of items in TRC is that providing an answer to these items necessitates more highly complex processes as compared to the other two groups of items. Presenting such items in the target language adds to their complexity for L2 learners. Thus, it can be said that in the case of TRC, the testees' native language has made the items easier for them. None the less, as for the other two groups, since a more surface mapping is required to provide an answer to the items, the testees have had a better performance on ERC than on TRC. Indeed, in such items, a whole sentence/phrase has usually been copied from the text. This has helped the testees to come to the correct answer more easily in the case of ERC items. In other words, presenting such items in the testees' native language would result in a further complication of what they tap. In sum, one can see that for each group of items the item facility index for one of the test is quite different than that of the corresponding items in the other test. Nevertheless, what is of concern here is the average item facility index of all the items in each test and, on the whole, one cannot see a remarkable difference between the average item facility indices of the two tests, since the high and low indices of different items have neutralized each other.

Yet, another reason might be the fact that the composition of the two groups taking the tests may have caused the effect of the test method facet to fade away. As mentioned earlier, a main function of presenting the items in the subjects' native language is reducing anxiety which is more observable in the beginners rather than more proficient language learners. Since each group of the subjects taking any one of the tests consisted of low, intermediate, and high proficiency level students, the performance of the high and intermediate students might have affected the results so that the test method did not show its effect.

In order to see to what extent the above conjecture is true, the scores of the three proficiency groups, each consisting of two groups, one taking ERC and the other TRC, were subjected to independent t-tests. Table 7 illustrates the results of the t-tests.

Table 7: Independent t-test between ERC and TRC scores of the subjects in each proficiency level

Level	Test	X	SD	T-value	P-value
High	ERC	25.84	2.56	3.49	0.08
	TRC	24.78	3.58		
Int.	ERC	19.36	3.18	6.45	0.07
	TRC	20.81	2.87		
Low	ERC	17.90	3.89	5.10	0.00*
	TRC	19.36	3.18		

As shown in Table 7, the results of the t-tests indicate that the difference between the performance of the testees on ERC and TRC is meaningful only in the case of the low-level students. With respect to the high- and intermediate-level students such difference cannot be seen. The results, indicating that the low-level subjects have had a better

performance on TRC than on ERC, lend themselves well to support the claim that the language of presentation, when in the native language of the testees, is more effective in the case of low-level learners since it reduces the anxiety which is more common among such testees. The results obtained here are in line with those of Shohamy (1984) and Alderson (2000).

To sum up, the obtained results so far, provide us with two different answers to the two parts of the first research question. The answer to the first part of the question is "no." That is, it made no significant difference when the items in the L2 multiple-choice reading comprehension test were given in the testees' L1, i.e., Persian. On the other hand, the second part of the question was, to some extent, positively answered. That is, a significant difference was observed between the mean scores of the low group on TRC and ERC. Thus, test method showed its effect only in the case of the low-proficiency subjects.

2. *Is the difference between the two low-proficiency groups, taking ERC and TRC, due to test method effect or the testees' reading ability in the first language?*

In order to answer the above question, first, the correlation coefficient between the PRC scores of the low-proficiency level subjects taking ERC/TRC was calculated. Table 8 presents the results.

Table 8: Correlation between PRC and ERC/TRC scores in the low proficiency level

Tests	Correlation coefficient
ERC & PRC	0.43 ns
TRC & PRC	0.20 ns

As Table 8 illustrates, no significant correlation was found between the performance of the low-proficiency level students on any versions of the English reading comprehension test (ERC or TRC) and their performance on PRC. Thus, the answer to the second research question is negative. That is, no significant relationship was found between the low-proficiency participants' reading comprehension ability in their first and second language. As a result, the difference between the performance of the participants with a low-proficiency level of English taking ERC and TRC might be due to the test method effect, i.e., the language through which the items were presented to the testees, rather than their native language reading comprehension ability. The results contradict those of Bernhardt and Kamil (1995), Markham (1985), and Lee and Musumeci (1988).

Second, an independent t-test was run to see if there was any significant difference between the mean PRC scores of the low-proficiency subjects taking ERC and that of the corresponding group taking TRC. Table 9 reveals the results.

Table 9: Independent t-test between the PRC scores of the low-proficiency group taking ERC and the low-proficiency group taking TRC

PRC scores for the low-proficiency subjects	X	SD	T-value	P-value
ERC group	22.00	2.20	-1.80	0.79 ns
TRC group	23.27	2.35		

According to Table 9, no significant difference can be observed between the PRC mean scores of the two groups of subjects, one taking ERC and the other TRC, in the low group. This can be indicative of the fact that the difference between the scores on ERC and TRC in the low group bears no relation to their Persian reading comprehension ability. Thus, the difference can be attributed to the language of presentation in TRC.

3. *If the test method is found to influence the testees' performance on a reading comprehension test, is it better to give the test items in L1 or L2?*

To answer this question, two one-way tests of ANOVA were run. The first one was run between the scores of the three proficiency groups on ERC. Tables 10 and 11 illustrate the results of the ANOVA and the scheffe test.

Table 10: ANOVA test for ERC scores among the three proficiency levels

Sources	DF	SS	Ms	F
Between groups	2	411.43	205.71	19.41*
Within groups	92	635.64	10.59	
Total	94	1047.07		

P<0.01

Table 11: Scheffe test for ERC scores of the three proficiency levels

X		High	Int.	Low
25	High		*	*
19.36	Int.			*
17.60	Low			

As Tables 10 and 11 illustrate, ERC has appropriately made discrimination among the three groups of the subjects. As expected, the advanced group outperformed the intermediate and the low group. The intermediate group, too, did better than the low group.

Table 12 reveals the results of the second ANOVA test run for the difference among the scores of the three groups of the subjects on TRC.

Table 12: ANOVA test for TRC scores among the three proficiency levels

Sources	DF	SS	Ms	F
Between groups	2	538.64	269.32	20.20*
Within groups	97	826.37	13.32	
Total	99	1365.01		

P<0.01

As Table 12 shows, the result of the ANOVA test is significant. However, the results of the Sheffe test showed the significant difference only between the advanced and the low groups. No significant difference was observed between the performance of the advanced and the intermediate groups or the intermediate and the low groups. In fact, it seems that the distinction between the intermediate and the advanced groups, on the one hand, and the one between the intermediate and the low groups, on the other, has disappeared in the case of TRC.

Drawing on the results of these two tests of ANOVA, one can claim that ERC has made better a discrimination among the different proficiency levels than TRC. In fact, TRC has been rather easy so no fine discrimination has been observed between different proficiency levels. Thus, ERC is preferred over TRC when it is intended to make precise distinctions between the testees' reading comprehension ability in different proficiency levels.

Conclusions

All in all, the results of the study primarily indicated that it makes no difference to present the items of a reading comprehension test in the testees' native language--Persian, here--or the target language--English. In other words, this test method facet did not turn out to influence the performance of the testees. None the less, it was found to affect the test results for elementary language learners. Low-proficiency language learners had a significantly better performance when the items were presented in their native language. This, however, was found not to be pertinent to the testees' reading comprehension ability in their first language.

Furthermore, it was found that L2 reading comprehension test with the items in L2 could make a better discrimination among the subjects with different proficiency levels.

The results of the study further showed that presenting the items of a reading comprehension test in the language learners' first language makes the test easier for low-proficiency level testees by reducing their anxiety. Notwithstanding, this approach is not recommended when the purpose is to make a meticulous discrimination among the testees with different proficiency levels, particularly, when the purpose is making a distinction between the reading comprehension ability of advanced and intermediate examinees.

The above conclusions, however, must be regarded tentative because, first, the English reading comprehension test turned out to be relatively easy even with the items in L2e. This might be the reason why presenting the items in the testees' L1 did not turn out to influence the performance of the testees. A different test would probably yield different results.

Moreover, since the testees did not have the experience of taking a Persian reading comprehension test, they did not do well on it. In other words, the novelty of the test might have affected their performance; otherwise, they would have performed differently on PRC and consequently different results might have been obtained.

.

Pedagogical Implications

A very important implication of the present study is for language testing, particularly when it comes to testing reading comprehension skill. As the results of the present study showed, presenting the L2 reading comprehension test with L1 items reduces the effect of test method facet, and thus, makes the test more valid for the learners with a low level of proficiency. Accordingly, the language teachers can ensure that their assessment of the learners' reading comprehension ability is more accurate and as a result the decisions made based on that are more fair.

Acknowledgements

I would like to express my deepest gratitude to Professor Jafarpur for his invaluable comments throughout the writing of the paper and reading the earlier version of the manuscript. I would also like to thank Dr. Saadat for the precise reading of the paper and providing me with her helpful comments.

References

Aebersold, J.A. & Field, M.L. (1998). *From reader to reading teacher: Issues and Strategies for second language classrooms.* Cambridge: CambridgeUniversity Press.

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

Anderson, N.J., Bachman, L., Perkins, K., & Cohen, A. (1991). An exploratory study into the construct validity of a reading comprehension test: Triangulation of data sources. *Language Testing,* 8, 41-66.

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

Bachman, L.F. & Palmer, A. S. (1981). The construct validation of the FSI oral interview. *Language Learning,* 31, 67-86.

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

Bernhardt, E.B. & Kamil, M.L. (1995). Interpreting relationship between L1 and L2 reading: Consolidating the linguistic threshold and the linguistic interdependence hypotheses. *Applied Linguistics,* 16, 15-34.

Carrel, P.A. (1991). Second language reading: reading ability or language proficiency? *Applied Linguistics,* 12, 159-179.

Donin, J. & Silva, M. (1993). The relationship between first- and second- language reading comprehension of occupation-specific texts. *Language Learning,* 43, 373- 401.

ETS (1995). *TOEFL Practice Tests.* Princeton: Educational Testing Service.

Fulcher, G. & Marquez Reiter, R. (2003). Task difficulty in speaking tests. *Language Testing,* 20, 321-344.

Godev, C.B, Martinez-Gibson, E.A., & Toris, C.C. (2002). Reading comprehension test: L1 versus L2 in open-ended questions. *Foreign Language Annals,*35, 202-21

Gordon, C.M. & Hanauer, D. (1995). The interaction between task and meaning construction in EFL reading comprehension tests. *TESOL Quarterly,* 29, 299-324.

Jafarpur, A. (2003). Is the test constructor a facet? *Language Testing,* 20, 57-87.

Katz, S., Lautenschlager, G., Blackburn, JA., & Harris, F. (1990). Answering reading comprehension items without passages on the SAT. *Psychological Science,* 1, 122-27.

Kobayashi, M. (2004). An investigation of method effects on reading comprehension test performance. *JALT Pan-Sig Proceedings.*

Lee, J.F. (1990). A review of empirical comparisons of non-native reading behaviors across stages of language development . In H. Burmester & P.L. Rounds, Eds., *Variability in second language acquisition, vol 2* (pp. 453-72). Eugene, OR: University of Oregon Press.

Lee, J.F. (1986). Background knowledge and L2 reading. *Modern Language Journal,* 70, 350-354.

Lee, J.F. & Ballman, T.L. (1987). FL learners' ability to recall and rate the important ideas of and expository text. In B. Van Patten, T. Dvorak, & J. Lee, Eds., *Foreign language learning: A research perspective* (pp. 108-18). Cambridge, MA: Newbury House.

Lee, J. & Musumeci, D. (1988). On hierarchies of reading skills and text types. *Modern Language Journal,*72, 173-87.

Markham, P. (1985). The rational deletion cloze and global comprehension in German. *Language Learning,* 35, 423-30.

Perkins, K., & Brutten, S.R. (1993). A model of ESL reading comprehension difficulty. In Huhta, A., Sajavaara, K., and Takala, S., editors, *Language testing: New openings.* Jyväskylä: University of Jyväskylä, 205-18.

Shohamy, E. (1984). Does the testing method make a difference? The case of reading comprehension. *Language Testing,* 1, 199-215.

Mohammad Rahimi,PhD, is assistant professor of TEFL at Department of Foreign Languages and Linguistics at Shiraz University, Iran. His areas of interest are language testing and assessment, research methods, language learning strategies, reading and writing. He is also the educational coordinator at the Language Center of Shiraz University. He may be reached by e-mail at mrahimy@gmail.com