How do test methods affect reading comprehension test performance?

Agnes Loch[1]

The paper describes how statistics were employed in language testing research to explore the effect of two test method variables of foreign language reading comprehension tests on test takers' reading comprehension performance. Statistical procedures were applied at three stages of the research: in the validation of the main research instruments, i.e. the reading tests (1), in grouping the participants into four comparable groups (2), and in analysing the participants' test performances on two reading comprehension tests (3).

Statistics and qualitative data analyses show that task type and native language use as test method variables, rarely have a statistically significant affect separately, but may rather exert a joint effect on performance.

Keywords: language testing, test method variables, test method effect

1. Introduction

The research explores the effect of two test method variables of foreign language reading comprehension tests - task type and native language use - on test takers' reading comprehension performance. The aim of the investigation is either to support or to reject the hypothesis that neither task type nor native (L1) or target language (L2) use influences reading comprehension performance significantly if the tasks target the same construct.

As communicative competences cannot be measured in any other way than by observing an individual's language performance, which is supposed to reflect the underlying competences, it is essential to consider all the possible factors that, besides actual reading comprehension ability, may influence performance and test results. Bachman (1991) sets up three categories to classify these contaminating factors: personal attributes (e.g. age, gender, occupation), test method facets (e.g. task type, dictionary use), and random factors (e.g. weather conditions, the test taker's physical or emotional state). Personal attributes and random factors are beyond the control of the examiner whereas method facets can be manipulated to make the assessment procedure and the results valid and reliable.

[1] Loch, Agnes, PhD, senior lecturer, Budapest Business School, Budapest, Hungary

Test method facets are a set of factors that specify the actual method of the assessment procedure. They cover the following categories: testing environment, test rubric, input, expected response, and the relationship between input and response (Bachman 1991). Test method facets can be carefully designed and controlled in order to minimize their distorting effects on an individual's test results. As the present study aims to investigate the effects of test method facets on reading comprehension performance, the crucial facets are the facets of the reading text and the input. The input includes the task and the use of L1 versus L2 in the input (and the expected response).

Several comparative studies have been conducted to investigate the effects of task types and native language use on reading comprehension. In a recent study Liu (2009) compares three task types and claims that gap-filling tasks have a significantly negative effect on test-takers' reading comprehension performance. Other researchers are more cautious in their conclusions. Shohamy's (1984), Wolf's (1993), and Gordon and Hanauer's (1995) studies are particularly remarkable because besides comparing short-answer questions and multiple choice tests, they also examined the effects of native language use. Based on their findings it is clear that the items or questions in the task provide additional information for the reader that may help comprehension. The amount and the quality of this information may substantially differ in the case of different task types. Native language use in the items and the expected response resulted in higher performance in each of these studies. However, due to weaknesses in research methodology the authors could not draw any general conclusions. It is still not explained whether improvement in performance in the studies was due to better understanding the questions, or to using the additional information in the questions to better understand the text itself.

It is worth noting that, although many theoretical works discuss the criteria for validation and reliability in detail (e.g. Bachman 1991, Bárdos 2002, McNamara 2000, Popham 1990), researchers rarely provide any information about the validation methods of the research instruments in their empirical studies.

2. Research questions

The broad research area of the present study is the investigation of how two testing variables – task type and the language of task and task completion – affect reading comprehension performance. The formulation of the exact research questions was based on the taxonomies in the literature (Alderson 2000, Urquhart-Weir 1998) as well as the findings of a teacher's questionnaire (Loch 2007, 2009a) and the statistical results of a Pilot Study including 185 participants. Two task types, short-answer questions (SAQ) and multiple choice items (MC), were selected for comparison in the Main Study. Thus, the main research questions focussing on the mutual relationship of task type and language use, were formulated as follows:

- How does the use of the native language in task rubrics, test items, and task completion influence reading comprehension performance in a short-answer questions test and in a multiple choice test?
- How do short-answer questions and multiple choice items as task types influence reading comprehension test scores when the task and the expected response are in English as the target language and in Hungarian as the native language?

3. Research method

The study compared the participants' performance on two reading comprehension tests including either short-answer questions (SAQ) or multiple choice items (MC): one in English as the target language, and one in Hungarian as the native language. Two sets of texts were selected, and four different tests were developed for each text: a short-answer questions test with rubrics and questions in English and in Hungarian, and a multiple choice test with rubrics and four options in English and in Hungarian. Thus, there were two sets of two texts and eight reading tests altogether.

Four groups of minimum fifty students each were involved in the research. Each group completed two tests. The participants in the same group worked with the same task type in the two tests, with language as the changing variable. Table 1 shows the groups and which versions of the tests they completed.

Table 1. The research matrix

	Group A	Group B	Group C	Group D
Test 1	SAQ test in English	MC test in English	SAQ test in Hungarian	MC test in Hungarian
Test 2	SAQ test in Hungarian	MC test in Hungarian	SAQ test in English	MC test in English

Source: own creation

Statistical analyses were employed at three different stages of the research:

1. in the validation procedure of the main instruments,
2. in forming comparable groups of participants,
3. in analysing data from test results and from questionnaires.

4. Statistical procedures

4.1. Validation of the main instrument

The validity of the tests was ensured in several ways. Besides qualitative methods, the statistical analysis of test results in two pilot studies (involving 185 and 202 students altogether) and correlating test scores with TOEFL[2] scores as a validated third measure (concurrent validity) helped ensure the validity and the reliability of the research instruments.

The data from the tests were processed using SPSS software (Version 11.0). Classical item analysis was carried out to calculate means, standard deviation, item test correlations, and reliability coefficients. The purpose of the analysis was to gain information about the tests as a whole, and to identify items for deletion or modification. Statistical results were expected to help validate the tests and decide which texts and items could be included in the final test booklets for the Main Study.

Poorly performing items were identified and modified after Pilot Study I. Besides modifying the wording of the questions, new items and new distracters were devised when necessary. After administering the tests in the second pilot stage, descriptive statistics and reliability analyses were carried out (Table 2). The results showed that the reliability of the tests increased considerably. Reliability for the SAQ test increased from $\alpha = .7399$ to $\alpha = .8398$, and for the MC test from $\alpha = .4327$ to $\alpha = .6631$ in the case of *Test 2*. In the case of *Test 1*, for the SAQ test it was $\alpha = .8149$, and for the MC test $\alpha = .7012$. The lower reliability coefficients of the MC tests were assumed to be related to the fewer number of items: the first version of both SAQ tests contained 30 items, whereas the MC tests contained 16 items only.

Table 2. The statistical analysis of the tests in Pilot Study II

Test	*M*	Facility value (%)	*SD*	Reliability (Alpha)	Adjusted reliability
Test 1 SAQ -E	21.98	73.2	4.8739	.8149	-
Test 1 MC -E	8.25	51.5	3.1057	.7012	.8148
Test 2 SAQ -E	18.87	62.8	5.6149	.8398	-
Test 2 MC -E	8.45	60.3	2.8559	.6631	.8082

Source: own creation

As reliability increases as items are added (Henning 1987, Csapó 1993), it was assumed that the reliability of the MC tests would increase if the number of items in the tests were increased to a specified length. The Spearman-Brown Proph-

[2] Testing English as a Foreign Language - the most widely accepted English language test developed by ETS (Educational Testing Service) US.

ecy Formula states the relationship between reliability and test length mathematically based on the assumption that the added items are of similar quality to other items in the test. Using the Spearman-Brown Prophecy Formula it was possible to calculate what the reliability of the MC tests would become if they contained the same number of items as the respective SAQ tests. The formula says

$$r_{ttn} = \frac{nr_{tt}}{1+(n-1)r_{tt}}$$

where, r_{ttn} = the reliability of the test when adjusted to n times its original length
r_{tt} = the observed reliability of the test at its present length
n = the number of times the length of the test is to be augmented.

By using the Prophecy Formula, in the case of *Test 1,* the estimated reliability of the MC test version was α = .8148, which corresponded to the respective SAQ test reliability (α = .8149). In the case of *Test 2* the calculated reliability for the MC test was α = .8082, which is also above the .8000 level. Although it was not possible to lengthen the MC tests to that extent, using the Spearman-Brown Formula was still relevant, and its results were reassuring. In an indirect way these results provided information about the items and confirmed their appropriateness for testing purposes.

Besides considering the reliability of the tests, the means and the facility values (calculated from the means) were also considered (Table 2). The analysis of the statistics helped to identify items which were particularly difficult or easy for the pilot population. By deleting problematic items it was possible to set the difficulty (facility value) level of the tests. After deletions, the item number of the SAQ tests was set at 24.

4.2 Forming comparable groups of participants

In order to compare performances on different test versions and draw conclusions on method effects, it was of crucial importance to set up four groups of participants, and to ensure that the groups were equivalent regarding their language proficiency.

Two-hundred and sixty-seven first-year students participated in the Main Study from Budapest Business School. On the basis of their TOEFL tests results (Phillips 1990), the participants were arranged into four groups of comparable language proficiency. As raw scores might not be regarded as interval data, the scores were converted by using the TOEFL Conversion Table. Then, the means and the standard deviations of the four groups were computed (M_A= .439.8, SD =63.6; M_B = 440.7, SD = 61.4; M_C =440.2, SD = 68.9, M_D = 441.8; SD = 63.9), and the means were compared using analysis of variance (ANOVA), which confirmed that there was no significant difference between the group means ($F_{3,234}$ = .168, p = .918), and thus, the groups were comparable. In addition, the participants' ability logits were

computed in a Rasch analysis, and were also compared in an analysis of variance ($F_{3,234}$ = .422, p = .737). The result showed that the groups were highly comparable (Loch 2009b).

4.3 Analysing data from test results

Following the traditional line of Classical Test Theory (CTT), the scores were regarded as interval data and were processed accordingly. For the statistical analyses the Statistical Package for Social Sciences software[3] was used. As the procedures of Item Response Theory (IRT) are recommended for much larger sample sizes, their application was limited and complementary in the present study (Bachman 2004, Baker 1997, Horváth 1997).

The test takers' performances on the eight test versions were compared by using both parametric and non-parametric statistical computations because distribution on one of the eight tests was slightly skewed. The procedures applied are shown in Table 3.

Table 3. Statistical procedures employed in the data analysis

Type of analysis	Non-parametric tests	Parametric tests
Checking for distribution	Chi-square	Chi-square
Comparing means (two data sets)	Wilcoxon test Mann-Whitney U test	Paired-samples *t* test Independent *t* test
Comparing means (more than two data sets)	Kruskall-Wallis test	ANOVA
Relationship between variables	Spearman rank order correlation	Pearson product moment correlation
Relationship among variables		Regression analysis Univariate analysis of variance

Source: own creation

Inferential statistics were run at three levels. First, *Test 1* and *Test 2* versions were compared to see if they were the same difficulty level. Secondly, the English (L2) and the Hungarian (L1) versions of the same tests were compared to check them for language effect. Finally, the short answer question version and the multiple choice version of the same tests were analysed to investigate task type effect. As the four groups were highly comparable concerning language ability, group differences were excluded from the possible reasons for potential differences.

[3] SPSS Inc. (1989-2003). *Statistical Package for Social Sciences* (Versions 11.0, 12.0)

When comparing the difficulty level of the tests, it was found that in the case of the SAQ tests the difference between test means was significant at the $p < .001$ level, with *Test 2* being more difficult for the participants than *Test 1*. However, in the case of the MC tests: in one group it was *Test 1*, whereas in the other group *Test 2* that proved to be significantly more difficult for the students. Questionnaire data and group interviews seemed to suggest that the test takers' insufficient language knowledge did not allow them to choose the correct answer from the only slightly different options provided in the MC items, which resulted in inconsistent test-taking behaviour. In spite of these results, positive correlation was found between the students' scores on the two tests.

Next, the English and Hungarian versions of the same tests were compared to investigate native language use effect on test performance. Both the means and the facility values showed that the Hungarian versions were easier and elicited higher performance (Table 4) although in the case of the SAQ tests the difference did not reach statistical significance at the .05 level.

Table 4. Comparative data about the SAQ and the MC tests

	Test 1 SAQ		Test 2 SAQ	
	in English	in Hungarian	in English	in Hungarian
Group	Group A	Group C	Group C	Group A
N	64	60	62	60
M	13.47	13.67	9.87	10.70
Range	22	18	20	20
SD	4.8500	4.7929	4.8332	4.8198
Variability	23.523	22.972	23.360	23.231
Facility value (p)	.5625	.5688	.4113	.4458
	Test 1 MC		Test 2 MC	
	in English	in Hungarian	in English	in Hungarian
Group	Group B	Group D	Group D	Group B
N	56	66	68	58
M	6.89	8.36	6.57	7.31
Range	12	10	10	9
SD	2.4913	2.4970	2.5934	2.5902
Variance	6.206	6.235	6.726	6.709
Facility value (*p*)	.4308	.5625	.4373	.4908

Source: own creation

In the case of the MC tests, however, the test takers performed significantly better on the Hungarian version. The mean difference between the English and the Hungarian versions of *Test 1* was significant *(t_{120}=-3,245, p = .002)*, and there was a medium effect size (d = .59) (Dancey-Reidy 2004). The adjusted R squared (R^2=

.080) showed that eight percent of the variation in test scores could be explained by the different languages. In the case of *Test 2* the non-parametric Mann-Whiney U test also indicated significant difference (z = -2,054, Asymp. Sig.= 0,40) (Table 5). This indicates that, at least in some cases, the language of the task had a decisive influence on the response.

Table 5. Test of significance in relation to the Mann-Whitney U test statistics of test scores on the L1 and L2 versions of the MC tests

	Test 1	Test 2
Mann-Whitney U	1223.500	.1451.500
Wilcoxon W	2819.500	3662.500
Z	-3.230	-2.054
Asymp. Sig. (2-tailed)	.001	.040

p<.05

Source: own creation

The third comparison focused on investigating the effect of task type. As the English multiple choice version of *Test 2* was slightly positively skewed, in this case the Mann-Whitney U test was applied. Significant difference was found in one case (Table 6): between the means of test scores on the SAQ and MC versions of *Test 1* in English ($t_{113} = 3.800$, $p < .001$), with an effect size $d = .72$. Performance on the SAQ test highly exceeded performance on the MC test. Task type explained 10 percent of the variation in performance ($R^2 = .105$).

Table 6. Results of Independent Samples t test and Mann-Whitney U test to compare means from different test formats

Test	Version	Groups	t	df	Sig. (2-tailed)	Z	Asymp. Sig. (2-tailed)
1	English SAQ + MC	A + B	3.800	113	.000*		
	Hungarian SAQ + MC	C + D	1.288	121	.200		
2	Hungarian SAQ + MC	A + B	1.201	113	.232		
	English SAQ + MC	C + D				-.682	.496

* *p*<.001

Source: own creation

It was also important whether task type and language use exercised any joint effect on the participants' performance. In the case of *Test 1*, the two variables jointly did not show a significant relationship with the test scores although had a significant effect on means separately (Table 7). This fact as well as the findings that the MC format affected performance on *Test 1* negatively and on *Test 2* positively indicate that task type and other factors may interfere.

Table 7. Tests of significance in relation to the Analysis of Variance for the joint effect of task type and language use

Variable	df	F	Significance	Partial Eta squared
Test 1				
Corrected Model	3	6.925	.000	.082
Task type	1	13.204	.000	.053
Language	1	5.183	.024	.022
Task type*language	1	3.382	.067	.014
Total	238			
Test 2				
Corrected Model	3	1.384	.248	.017
Task type	1	1.595	.208	.007
Language	1	2.504	.115	.011
Task type*language	1	.243	.623	.001
Total	238			

Source: own creation

Statistics confirmed the expectations that the participants' reading scores would significantly correlate at the .01 level with their scores on the TOEFL papers on receptive skills (r = .584). However, the overlap between scores was not particularly large, which indicates the distinctiveness of reading skills

5. Summary of findings

In the light of the results, it is obvious that the research hypotheses gained partial verification only. The statistics showed that in most cases no significant difference was found. However, there were exceptions, both in the task type and the language use comparisons, when the mean differences reached the statistically significant level. Due to these mixed results, no general conclusions can be drawn.

Although results were not consistent, some of the findings strongly suggest that test method variables may exert a joint effect with other factors such as text difficulty or test takers' characteristics. As the mixed results

gained about test format effects in this study do not provide a comprehensive conclusion, further research is needed in the area, especially in two directions: how task types influence performance at different levels of proficiency, and how task type effect is related to the conceptual and linguistic difficulty of a reading text.

References

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

Bachman, L. F. 1991: *Fundamental considerations in language testing*. Oxford University Press, Oxford.

Bachman, L. F. 2004: *Statistical analyses for language assessment.* Cambridge University Press, Cambridge.

Baker, R. 1997: *Classical test theory and item response theory. Theory in test analysis* [Special report 2, Language testing up-date]. Department of Linguistics and Modern English Language, Lancaster University, Lancaster.

Bárdos, J. 2002: *Az idegen nyelvi mérés és értékelés elmélete és gyakorlata.* Nemzeti Tankönyvkiadó, Budapest.

Csapó, B. 1993: Tudásszintmérő tesztek. In Falus, I. (ed.): *Bevezetés a pedagógiai kutatás módszereibe.* Keraban Kiadó, Budapest, pp. 277-317.

Dancey, C. P. - Reidy, J. 2004: *Statistics without maths for psychology*. Pearson/Prentice Hall, Harlow, UK.

Gordon, C. M. - Hanauer, D. 1995: The interaction between task and meaning construction in EFL reading comprehension tests. *TESOL Quarterly,* 29, pp. 299-322.

Henning, G. 1987: *A guide to language testing*: *Development, evaluation and research*. Newbury House, Cambridge, MA.

Horváth, Gy. 1997: *A modern tesztmodellek alkalmazása.* Akadémiai Kiadó, Budapest.

Liu, F. 2009: The effect of three test methods on reading comprehension: An experiment. *Asian Social Science, 5,* No. 6. pp. 147-153. Retrieved on 01/05/2010 from http://ccsenet.org/journal/index.php/ass/article/view/2491.

Loch, A. 2007: Task types in assessing reading comprehension. *BGF Tudományos Évkönyv.* BGF, Budapest, pp. 358-362.

Loch, A. 2009a: Native language vs. target language use in language testing. *Acta Linguistica* 8, Vol. 2, University of Matej Bela, Banska Bystrica, Slovakia, pp. 64-70.

Loch, A. 2009b: Minták összehasonlíthatóságának bizonyítása statisztikai eljárásokkal. Paper presented at *New Methods in Applied Linguistics Research Conference*. 26-27. October 2009, Budapest.

McNamara, T. 2000: *Language testing.* Oxford University Press, Oxford.

Phillips, D. 1990: *Longman practice tests for the TOEFL*. Longman, New York.
Popham, W. J. 1990: *Modern educational measurement*. Allyn and Bacon, Boston.
Shohamy, E. 1984: Does the testing method make a difference? The case of reading comprehension. *Language Testing,* 1, pp. 146-169.
Urquhart, A.-Weir, C. J. 1998: *Reading in a second language: Process, product and practice.* Longman. London.
Wolf, D. F. 1993: A Comparison of assessment tasks used to measure FL reading comprehension. *Modern Language Journal*, 77, pp. 473-489.