THE INFLUENCE OF TEST FORMAT ON LANGUAGE TEST PERFORMANCE

Róbert Majzlík, Petra Laktišová
University of Prešov, Slovakia

Anotácia: Nedávne štúdie dokazujú, že existujú určité faktory, ktoré majú negatívny vplyv na výkonnosť testovaných subjektov. Jedným z nich je vplyv testovacej metódy. Tento článok predstavuje výsledky štúdie, ktorej cieľom bolo zistiť, do akej miery sa výkony študentov stredných škôl 4. ročníka líšia v testovacích úlohách rôznych formátov. Štúdia porovnávala a korelovala výkonnosť študentov pri piatich rôznych úlohách s cieľom zistiť ako sa prostredníctvom rôznych testovacích nástrojov demonštrujú ich znalosti daných gramatických javov. Výsledky naznačujú, že existuje úzka súvislosť medzi formou testovacej úlohy a výkonnosťou študentov.

Kľúčové slová: test, skóre, výkon, forma testu

Abstract: It has been recently acknowledged that there are certain factors that have contaminating influence on test takers' performance. One of these is the influence of the test method. This paper presents the results of a study the aim of which was to find the extent to which the performance of 4th-year secondary school students differs depending on test tasks of various formats. The study compared and correlated students' performance in five different tasks in order to see how their knowledge of certain grammatical features is demonstrated through the various testing instruments. The results indicate that there is a close relationship between the form of a test task and students' performance.

Keywords: test, score, performance, test form

Introduction

In a language test scenario, the test helps the teacher to uncover the potential of a student. Language evaluation is often seen as measurement of a person's communication skills. It is done to find out if a person has communication problems. It is imperative to know at all times that the aim of classroom assessment is to give students insights about their success rather than criticise them for the things they have not caught grasped. According to Bachman and Palmer (Bachman, Palmer 1996: 66), "[a] number of language testers have point out that there is need to adopt tests formats that will elicit test-takers' best performance. It is believed that one approach to achieve this is to develop the characteristics of the test task so as to promote feelings of comfort or safety in test-takers that will in turn facilitate the flexibility of response on their part". The basic focus of this piece of research is thus to look at various tests formats,

analyse their influence on students' performance and make recommendations for teaching practice.

Literature review

Evaluation is of great importance when trying to assess the language performance of students. In today's world, practically everyone gets evaluated. According to Crooks (Crooks 1988), sitting formal written tests alone takes nearly 15 percent of students' time. Students have to cope with different examination types during their study. According to Douglas (Douglas 2014: 2), a test may be viewed as "a measuring device, no different in principle from a ruler, a weighing scale or a thermometer." He further reiterates that "a language test is an instrument for measuring language ability" although he does express the following doubt: "In what sense can we actually measure a concept as abstract as language ability?" Purpura (Purpura 2004: 4) affirms that "there is a glaring lack of information available on how the evaluation of grammatical ability might be carried out, and how the choices we make in the assessment of grammatical ability might affect the observation we make about our students' knowledge of grammar, the decisions we make on their behalf and their ultimate development". Opportunities to identify the possible relationships between form and function should be accorded to all language learners. However, nowadays, assessing grammar is based on an ability to choose between several given options.

The effect of test format on language test performance has been investigated by a huge number of researchers in the last twenty-five years. However, our literature review as regards the influence of test format on language test performance came up with very limited results. Mouti and Ypsilandis (Mouti, Ypsilandis 2014) conducted a study on the method's effect on test-takers' performance and confidence in language tests. With the aim to decide whether different test formats have any effect on the performance and confidence of the language learners, they selected a group of 200 first- and second-year students of Greek nationality and Greek mother tongue, aged 18 to 20. The first group (55 students) completed 2 sets of items, the second group (48 students) completed 6 sets of items, the third group (65 students) completed 2 sets of items and the fourth group (32 students) completed 2 sets of items during the two stages of the research. The result of their study showed that multiple-choice tasks might be considered easier than gap-filling ones as the former provided the higher scores.

Grabowski (Grabowski 2008) conducted another study in which he analysed the test format relevant to the scores of 60 adult EFL students at a tertiary level (Teachers College, Columbia University). Based on the results of his study, he argues that, although different test

formats lead to different results in learners' tests performance, using two different test formats is a recommended method of assessing learner's language competence.

Shohamy (Shohamy 1984) and Wolf (Wolf 1993) suggested that the method used to test students influenced language test performance and that multiple-choice questions were simpler than open-ended questions. This claim was investigated by several researchers. Tsagari (Tsagari 1984) did a piece of research on the above claim empirically and made a comparison between the effects of two test formats on two reading comprehension tests with similar content. The findings from the study showed that the effect of test a format can be a source of variation in scores and may even measure different abilities.

It is clear that there have been various researchers who have attempted to measure the influence of test formats on learners' performance in language tests but there is still some space for further research on how these test formats influence the performance in tests testing specific grammar structures. Despite much research on the influence of test formats on language test performance, it is still essential to carry out further research to determine the degree of influence a test format has on language test performance.

Aim and methodology

Several researchers have analysed the effect of different test types on learners' performance in language tests but most of the studies dealt with older EFL learners and people from different ethnicities. Another vital issue is that, in the present post-communicative age, learners should be presented with numerous opportunities to deeply analyse the interface between form and function of a language. Most teachers consider test formats using discrete-point items popular because of their relatively easy construction and administration in the classroom. On the other hand, test formats like cloze tests may be more difficult to fill in even though they may measure the same feature and using the same linguistic context.

This piece of research was carried out in order to find out whether different test formats testing the same language structure produce comparable/similar results. It was aimed at examining 5 different test formats and their influence on the overall the language test performance of L2 learners of English (B2 level). The present study constitutes a comparative analysis of the effects of different test formats on the performance of Slovak intermediate students of English. It involved both qualitative and quantitative approaches. The quantitative part of the research dealt with 5 measurement techniques (cloze test – banked, cloze test, multiple choice, sentence transformation and word formation), while the qualitative part of the research included analysis of the effectiveness of each test format.

Research questions

This study attempts to answer the following research questions:

1. Do all task types suit all test-takers in the same way?

2. Is there a relationship between task-types/test question formats and language test performance?

Characteristics of the tests

In order to find out to what degree a test method can influence the performance of test takers, we compared and analysed the performance of students in 5 discrete-point language tests of different formats (cloze test – banked (T1), sentence transformation (T2), word formation (T3), cloze test without prompts (T4), and multiple choice(T5)). The tests were designed to measure students' language competence relevant to grammar and vocabulary at the B2 level of the CEFR. The administration time of all 5 tasks was 60 minutes. Each task contained a sample answer written just after the instructions. The instructions and the sample answers were written with a different font style (bold for the instructions, italics for the sample). Answers were marked on separate answer sheets. The test items in tests T1, T3, T4 and T5 aimed to assess grammar as structure through a sentence discrete-point format, i.e. meaning-independent/dependent and context-dependent, the items in test T2 were context-independent and did not require testees to select structures and forms according to context and language use.

Cloze Test

Nowadays, cloze tests can be found in many variants. There are some grammarians who accept only one kind of test as the standard cloze. Under that interpretation, it consists of one or more reading paragraphs with words deliberately deleted according to a fixed ratio. Here, each omitted word is replaced by a gap which testees are required to fill in with a suitable word (Read 2000).

Cloze tests are also widely popular and used for evaluating language components simultaneously. Taylor (Taylor 1953) first used the term cloze-test when he defined this test format as a new practical instrument for determining levels of readability. Cloze tests are popular among teachers, too. They find them rather integrative because, when doing a cloze test, a test taker has to analyse language components simultaneously, which is similar to speaking. They, however, have some limitations. According to Harmer (Harmer 2002: 323), "a cloze test is the removal of every nth word in a text."

Multiple choice test

Multiple-choice (MC) questions are commonly used to measure language competence because they are quick, economical and straightforward to score. When compared to the

standard format of a cloze test, the multiple-choice version of the task makes a cloze test more a measure of reading comprehension rather than focusing also on the writing abilities of the testees, which are necessary when doing a cloze test exercise. A typical Multiple-choice test item consists of two fundamental parts: the stem (a question or a problem to be solved) and a list of possible answers, which usually contains one correct (or "best") answer and a number of incorrect options (distractors).

Sentence Transformation Test

By transformation here, we mean change of form. The sentence transformation test format simply means a test format in which the test taker is required to change the forms of sentences without changing their meaning. A sentence in direct speech can be transformed into indirect speech and vice versa. Similarly, a simple sentence can be converted into a compound or complex sentence; an affirmative sentence could be expressed in a negative form and vice versa.

Word Formation Test

"The subject of word-formation has not until recently received very much attention from descriptive grammarians of English, or from scholars working in the field of general linguistics" (Adams 2016: 3). But the situation is changing, especially in the field of vocabulary learning. With regard to recent research, it has been proved that learners' vocabulary knowledge necessarily includes the knowledge of the derivative forms in a word family (Schmitt, Meara 1997). Moreover, this knowledge does not seem to occur incidentally. Schmitt and Zimmerman (Schmitt, Zimmerman 2002: 145) speak about "a need for more direct attention to the teaching of derivative forms."

Data collection

For the purpose of this study, 30 non-native learners of English were selected, all of them students of the Cirkevná spojená škola in Vranov nad Topľou in their final year of preparation for the national examination in English (Maturita). The test-takers' age, both of male and of female test-takers, ranged from 18 to 19. By the time the test-takers qualified to take the examination, they had passed the Oxford Placement Test (OPT) to prove that the group was a homogeneous sample of all students. During the testing procedure, students were not allowed to use any dictionaries, in paper or online, nor were they allowed to communicate with each other. It should also be noted that test-takers did not receive any special training with regards to the current testing procedure. The items included in the tasks were thus selected primarily based on their previous learning experience and the knowledge of English. Although

Alderson, Clapham and Wall (Alderson, Clapham, Wall 1995) recommend executing piloting testing, in this case there was no piloting examination before the testing procedure.

Data analysis and discussion of the results

The data obtained from the study were subjected to analysis. Microsoft Excel 2010 was used to find out if the test formats used in our study influenced the performance of the test takers.

The table and figures below show the results of measures of a general trend (mean, median and mode) as well as other tools of descriptive statistics. For each test, mean scores were calculated based on which we compared the success rate of the participants taking these tests. Table 1 constitutes a detailed list of results.

Table 1 – *Statistical data on the analysed tests*

	CT-B 10	ST 5	WF 10	CT 10	MC 15
mean	9.11	3.27	6.06	4.44	8.61
median	9	3	7	4	8
minimum	6	0	3	2	5
maximum	10	5	8	7	14
standard deviation	1.13	1.31	1.58	1.24	2.99
skewness	-1.61	-0.92	-0.49	0.244	0.43
overall performance	85.61%	61.56%	60.55%	44.44%	57.40%

As seen in Table 1 above, the lowest mean was reached in sentence transformation test (ST). The highest mean scores were registered in cloze test –banked (CT-B). It can be said that, in the tested group of students, the form of a test had a direct impact on test-takers' performance. Since all test formats tested almost identical grammatical structures, the difference between the highest and lowest scores, though quite big, is related to the test format difference. Multiple choice test formats were very commonly used with our test-takers. Therefore, it can be concluded that they are more accustomed to the form. Moreover, the fact that each test item contains 1 correct answer and 3 incorrect distractors makes it easier to simply

guess the correct answer in case they are not able to answer it. A similar procedure can be applied to the cloze test – bank test type. Here testees have a bank of possible words they can choose from. The fact that the bank usually contains only function words, no lexical words, makes it easier to decide on the correct alternative. One of the limitations of this test type is that the test-taker does not have to have the knowledge of the real use of grammatical tenses; it is enough for him/her to know the form it has. And where the missing word is an auxiliary verb, has e.g. like in the present perfect simple tense, third person, singular, the test taker might not know why this tense is used there. Such knowledge is derived simply from knowledge of the structure rather than real use of the particular tense. On the other hand, cloze tests without a bank of provided options require a more complex understanding of grammatical relationships within the given sentence in a test and thus force the student to produce an answer on his/her own.

Figure 1 – *Overall success rate in all tests (in percentages)*
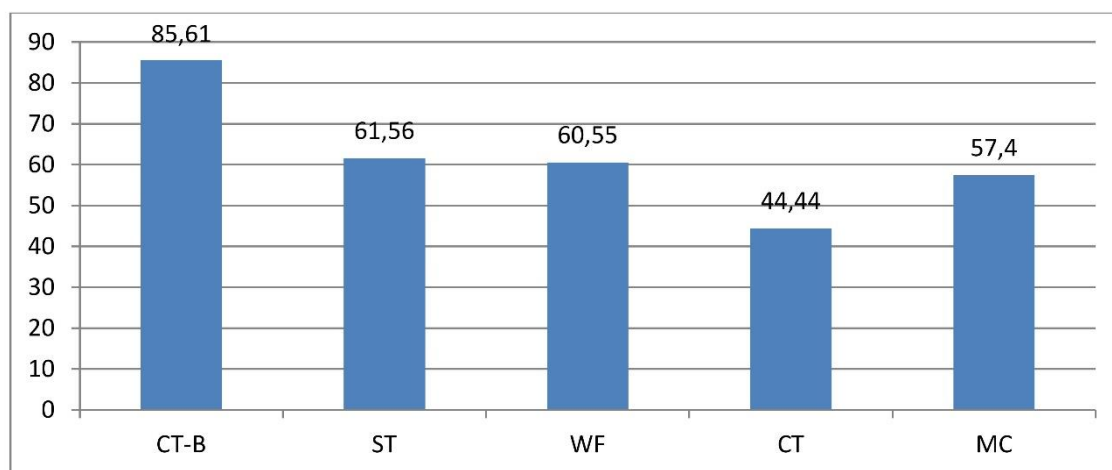
As expected, the test types which include options that test takers can choose from (recognition tasks) produce better results and lead to higher performance than the test types which require some kind of production, e.g. cloze test or sentence transformation. This leads us to the first conclusion, i.e. that test type format is responsible for variations in accuracy and overall performance.

Based on results of the descriptive statistics above, it is clear that although the multiple-choice tasks are generally considered to be easier than the gap-filling exercises, that was not the case in this research procedure. The difference between these five test formats and their scores lies in the fact that multiple choice tests were easier for students because there were a number of options provided in each test item so learners did not have to go to further effort to

create new forms but only choose from the options given. The students could easily identify the constructions that were appropriate after they had become familiar with the relevant context. The analysis of test items indicates that, although similar grammatical structures were presented in test formats, students had problems to understand these contexts correctly.

Conclusion

Our findings revealed that the test type/test format might be a source of variation in scores. Although the findings of this research show that students are more successful in cloze tests rather than multiple choice tests, it is not the aim of this research to recommend using only one test format. Smith and Levin (Smith, Levin 1996: 111) claim that "no single type of assessment can always meet all purposes, in all situations." In normal-referenced situations, the increase in scores that could be attributed to question format may not have a significant impact while, in criterion-referenced situations, the choice of question formats would need to be carefully examined and implemented. Alternatively, any possible variations would need to be supported by equivalent measurements.

It should be mentioned that this study may be subject to some limitations. At first, the number of our sample was rather small and further and more complex research is necessary to replicate the results obtained. Moreover, the fact that grading test types such as sentence transformation and cloze tests are to a certain extent subjective even if scoring criteria are developed (Powell, Gillespie 1990) has to be taken into consideration. Secondly, the assessors of a test always differ in their strictness/leniency (Engelhard 1996), therefore, follow-up research should involve larger data sets and more assessors in order to overcome these limitations.

Finally, there is still some ground to cover for interested researchers who may wish to investigate the impact of other types of test methods such as matching items on other language skills such as writing or listening.

References

Adams 2016: Adams, V. An Introduction to Modern English Word-Formation. - In: English language series, 7th edit. Routledge. [date of entering 10.5.2018]. <https://books.google.sk/books?hl=sk&lr=&id=HoWTDAAAQBAJ&oi=fnd&pg=PP1&dq=word+formation+test&ots=0ymGHDZKdj&sig=PihQEyEgAvwsVRoysZmEr3v641w&redir_esc=y#v=onepage&q=word%20formation%20test&f=false>

Alderson 1995: Alderson, J. Ch., Clapham, C., Wall, D. Language Test Construction and Evaluation. Cambridge: Cambridge University Press, 1995.

Bachman 1996: Bachman, L. F., Palmer A. S. Language testing in practice. Oxford: Oxford university press, 1996.

Crooks 1988: Crooks, T. J. The impact of classroom evaluation practices on students. In: Review of Educational Research 58(4), 1988, 438-481.

Douglas 2014: Douglas, D. Understanding language testing. New York: Routledge, 2014.

Engelhard 1996: Engelhard, G. Clarification to Examining Rater Errors in the Assessment of Written Composition With a Many-Faceted Rasch Model. - Journal of Educational Measurement, 33(1), 1996, 115-116.

Grabowski 2008: Grabowski, K, C. Investigating the Construct Validity of a Performance Test Designed to Measure Grammatical And Pragmatic Knowledge. In: Spaan Fellow Working Papers in Second or Foreign Language Assessment, 6, 2008, 131-179.

Harmer 2002: Harmer, J. The Practice of English Language Teaching. Third Edition. England: Longman Press, 2002.

Mouti 2014: Mouti, A. L. - Ypsilandis, G. S. Method-effect on test-takers' performance and confidence in language tests. - e-TEALS: an e-journal of teacher education and applied language studies, 5/2014, 65-81.

Powell 1990: Powell, J. L., Gillespie, C. Assessment: All Tests Are Not Created Equally. Paper presented at the Annual Meeting of the American Reading Forum, Sarasota, FL, 1990.

Purpura 2004: Purpura, J. E. Assessing Grammar. Cambridge: Cambridge University Press, 2004.

Read 2000: Read, J. Assessing Vocabulary. Cambridge: Cambridge University Press, 2000.

Shohamy 1984: Shohamy, E. Does the Testing Method Make a Difference? The Case of Reading Comprehension. - Language Testing 1.2, 1984, 147-170.

Schmitt 2002: Schmitt, N., Zimmerman, CH. B. Derivative Word Forms: What Do Learners Know? - Tesol Quarterly 36 (2), 2002, 145-171.

Schmitt 1997: Schmitt, N, Meara, P. Researching vocabulary through a word knowledge framework (Word associations and verbal suffixes). In: Studies in Second Language Acquisition 1, 1997, 17-36.

Smith 1996: Smith, M., Levin, J. Coherence, assessment, and challenging content. In: Performance based student assessment: Challenges and possibilities, Chicago, Illinois: The University of Chicago Press, 1996, 104–124.

Taylor 1953: Taylor, W. Cloze procedure: A new tool for measuring readability. - Journalism Quarterly 30. 1953, 414-433.

Tsagari 1984: Tsagari, D. Method Effect on Testing Reading Comprehension: How Far Can We Go? Master Thesis. University of Lancaster, 1984.

Wolf 1993: Wolf, D. A Comparison of Assessment Tasks Used to Measure FL Reading Comprehension. - Modern Language Journal 77. 1993, 473-489.