The Experimental Study of the Effect of Functional-Variational Factors on the Results of Linguistic Testing

Antonina V. Hryvko[1,*] & Yurii O. Zhuk[1]

[1]Monitoring and Assessment of the Education Quality Department, Institute of Pedagogy, National Academy of Pedagogical Sciences of Ukraine, Kyiv, Ukraine

*Correspondence: Monitoring and Assessment of the Education Quality Department, Institute of Pedagogy, National Academy of Pedagogical Sciences of Ukraine, 52-D Sichovykh Striltsiv Street, 04053, Kyiv, Ukraine. E-mail: av.hryvko@gmail.com

Received: December 2, 2021 Accepted: January 4, 2022 Online Published: January 17, 2022

doi:10.5430/jct.v11n1p45 URL: https://doi.org/10.5430/jct.v11n1p45

Abstract

A feature of the presented study is a comprehensive approach to studying the reliability problem of linguistic testing results due to the several functional and variable factors impact. Contradictions and ambiguous views of scientists on the researched issues determine the relevance of this study. The article highlights the problem of equivalence of diagnostic potential and the complexity of closed and open test tasks. The issues of polymorphism and monomorphism of tests, as well as variability of tests on the sequence of different question forms and their ratio in one test, are revealed. The article authors substantiate the need for in-depth empirical research to further improve the basic quality design principles and improve tests for final testing of high school students. The main purpose of the study was to examine the controlled and uncontrolled factors that affect test results. This article represents a study conducted among the students of the 8th and 9th grades (N=332) and in the subject field of Ukrainian (as a native language). In the first phase of the study the criteria-oriented author's tests and questionnaires were used by the authors as research tools. Data analysis was performed using factor and analysis of variance (and other mathematical and statistical methods), which allowed proving the direct influence of the factor "form of test tasks" on the test results (impact power is up to 35%). The methodology of the study allowed us to determine the combination of forms of tasks that have the lower impact on test results and, accordingly, are characterized by the highest diagnostic reliability. In the second (refining) phase of the study, taking into account the previous results, it was found that the ratio of test questions, as well as their sequence, can also affect the test results and are detected depending on the gender characteristics of the test participants. According to the results of the research, some ideal externally-organized design of Ukrainian language tests for in-school control and classroom-assessment is proposed. It is provided by a combination of the multiple-choice questions and constructed-response questions in equal proportions in descending order of operational complexity of tasks (taking into account gender peculiarities of testing participants). The prospects for continuation and deepening of the research, connected with further study of the optimal empirically valid test structure, are substantiated. In the second (refining) phase of the study, taking into account the previous results, it was found that the ratio of test questions, as well as their sequence, can also affect the test results and are detected depending on the gender characteristics of the test participants. According to the results of the research, some ideal externally-organized design of Ukrainian language tests for in-school control and classroom-assessment is proposed. It is provided by a combination of the multiple-choice questions and constructed-response questions in equal proportions in descending order of operational complexity of tasks.

Keywords: educational testing, secondary school student evaluation, test format, item order, item difficulty, item analysis, scores

1. Introduction

1.1 Problem Statement

The most used pedagogical technologies for testing and evaluating the educational achievements of students today are testing technologies. At the same time, there is a debate in the scientific community about their shortcomings

related to the possibility of objective pedagogical control and further decisions based on education. Such discussions support the relevance of research aimed at studying the design features of test technologies, taking into account the possible impact on the results of their application of several functional and variational factors. The proposed article aims to analyze the factors influencing the results of linguistic testing of students.

1.2 The Rationale for the Relevance of the Study

The most discussed issue regarding the application of test technologies is the problem of choosing the form of test questions. This problem is often associated with the influence of the questions form on the complexity of the test. Thus, the authors of some papers argue that the test with closed-ended questions and the choice of answers from several proposed is easier than the test that contains open-ended questions, which accordingly affects the test results (Melovitz Vasan, DeFouw Holland & Vasan, 2018; Orlov, Ponomareva, Chukajev & Pazuhina, 2017). This opinion led scientists to conclude that it is appropriate to combine different forms of questions in one test. At the same time, some studies cast doubt on the unambiguity of such conclusions, as they prove the equivalence of closed-ended and open-ended questions based on statistical significance and reliability (Marengo, Mic1eli & Settanni, 2016; Mozaffari, Mohammad Alavi & Rezaee, 2017; Schladitz, GroB Ophoff, & Wirtz, 2017) or based on calculations of the correlation between the total scores for tests in different subjects, constructed with and without open-ended questions (CR). It makes it possible to conclude that the test results will not change considerably if you change the form of the proposed questions, as an example, replace CR with MCQs of appropriate complexity (for example, Lissitz, Hou & Slater, 2012). In contrast, other scientists prove the existence of the effect of various factors (question forms, several options in MCQs, the structure of the test as well as the different combinations of these factors) on the psychometric characteristics of the test (AlKhatib, Brazeau, Akour & Alrnuhaissen, 2020). Kastner and Stangl (2011) tried to find out the reason for the ambiguity of scientists' conclusions about the equivalence of diagnostic potential and the complexity of the tests with open-ended questions and multiple-choice questions. The authors concluded that the research results differ since they take into account different factors that may affect the test results (type of closed questions, accepted rules for evaluating questions with constructed answers, test structure) and others). Thus, the results of such studies are incomparable, and therefore it is impossible to draw unambiguous conclusions, which leads to conflicting views of scientists on this issue and actualizes further research related to unresolved issues. They are: measuring the impact of different types of closed and open questions; various options for their combination, ratio and sequence in the test for test results and determine the model of the optimal structure of the test to ensure the objectivity and reliability of test results.

2. Theory

2.1 The Theoretical Basis of the Research Hypothesis

The hypothesis of the influence of variational and functional factors on the test results, which is the basis of this study, is based on the theory of Bachman – Palmer faces developed in the domain of linguistic testing (Bachman & Palmer, 1996). This theory clearly defines the relevance of the study of various aspects of assessing the language achievements of students, in particular the dependence of assessment of factors that constitute following the characteristics of testing tools and conditions of the procedures of its organization. Bachmann-Palmer facet theory is distinguished five main categories: the relationship between answers and questions; profiling of test questions; characteristics of predicted responses; (4) test structure (sections); (5) test environment (Bachman & Palmer, 1996).

In the course of scientific research, based on in-depth theoretical analysis of scientific literature and in the context of the chosen research topic, the authors identified variational and functional factors that may affect test results: (1) controlled - externally organized and meaningful test constructs; (2) uncontrolled - characteristics of respondents: group (in particular, class profile) and personal (cognitive - knowledge and skills, affective - perceptual and emotional assessment of the object of activity - test questions).

The semantic constructs of the test include the optimally reduced reflection of the content of education in the system of test tasks (the content of the test as a whole) and the presentation of the share of the content of the discipline in test form (the content of test tasks). Depending on the purpose of the assessment (for example, general determination of levels of student achievement at a specific stage of learning or assessment of knowledge on a particular topic or grouping of students by orientation or grade of knowledge acquired) determine the content of the test. Although the material of test tasks should be subordinated to the assessing goal of students' mastery of the curriculum or its elements, the selection and formation of test material is one of the most problematic issues, as the content will always be narrower than the overall content. Therefore, it is logical that the forming and composing test tasks process requires the participation of relevant experts.

Predicting the impact of the content of test questions on the test results immediately depends on externally organized test constructs, which are the ways, conditions and means of presenting and designing the content of the test. Externally organized constructs include test specification (its general characteristics - number of test tasks, structure, purpose, defined sequence of tasks, as well as principles and procedures for evaluating results and conducting testing), setting characteristics (time criteria for test tasks, physical and technical conditions of testing), the format of the test tasks themselves. In turn, the format of test tasks means a way to perform certain actions or initiate them, which is implemented by: forms of test tasks, programming the method of processing material, i.e. determining the number of actions required to perform the task, language methods, tools and scope of task formulation. Methods of ordering the elements of the problem are determined by its forms: closed (test tasks to establish the sequence; tasks to establish correspondence between the elements of two or more sets); selective (with the possibility of choosing one, several or all correct answers, with the choice of the most complete or thorough answer, with the choice of true or false statement) and open (constructed with a regulated answer - supplementary tasks, freely designed ie extended answer tasks). The form of the answer is the method of solving the problem, defined by its form. Depending on the needs of taking into account the criteria of validity of the pedagogical measurement, the choice of the appropriate form of the test task is made.

2.2 Analysis of Scientific Works on the Topic of the Research

Random or unmotivated choice of forms of test tasks leads to the difficulty of its understanding by the tested students and distortion of its meaningful expression. Researchers mostly associate the choice of forms of test questions with the parameters of test complexity, which are determined by the results of psychometric. In comparing MC and CR tests developed according to Bloom's taxonomy, Hancock (1994). experimentally proved that these two test task formats measure similar cognitive constructs at the same taxonomic level but at different levels of complexity. At the same time, according to Srivastava, Dhar and Aggarwal (2004) MCQs tap mainly recognition memory but the structure of knowledge is better revealed by free-recall tasks than by recognition tasks, and as Johnson (2010) points out, it is easier to recognize than to remember. At the same time, researchers Kan, Bulut and Boylan (2019) in their scientific works substantiate the feasibility of constructing polymorphic tests, which corresponds to global trends in the assessment of academic achievement.

Thus, a thorough analysis of the psychological and pedagogical literature Lissitz Hou and Slater (2012) made it possible to prove that, although psychometric indicators of assessment results on tests with closed and open questions may be equivalent, questions with constructed or freely constructed answer is not can be replaced by closed questions with the choice of the correct answer from several proposed. We can explain it solving problems of open and closed forms activate different types of thinking, i.e., aimed at identifying different cognitive skills. Alharbi (2017). Second, the authors, citing experimental data, argue that a test based on questions of one form may be the cause of deterioration in the results of a particular tested category, because some students cope better with open-ended questions, and worse – with closed, and others, on the contrary, are more successful in writing tests with MC, no the same questions of the CR (Lissitz, Hou & Slater, 2012; Onaiba & Jannat, 2019). In addition to individual characteristics of students' cognitive activity, the test is to some extent determined by gender: the results of a study conducted by researchers from Stanford University (Reardon et al., 2018) showed that according to the prevailing form of questions in the test can explain about 25% of gender differences in test results. That is, the form of tasks affects the results of their performance by students of different genders (Akhavan Masoumi & Sadeghi, 2020; Reardon et al., 2018).

Another argument in favor of polymorphic tests is the results of an experimental study that showed: students can answer MCQs successfully without knowing the answer, so the authors conclude that it is inappropriate to use such tests for final assessment, because the assessment of the number of correct answers is too high (McKenna, 2019). Therefore, in the tests of international comparative studies (PIRLS, PISA, TIMMS), most of the tasks are presented in an open form, as they are characterized by a high level of the diagnostic potential of test results (Martin, von Davier & Mullis, 2020; Mullis & Prendergast, 2017; Organisation for Economic Co-operation and Development, 2020).

We see an urgent need for empirical research to substantiate the optimal structure of the test in the perspective of updating the provisions on the polymorphism of the test for the final assessment of student achievement. This question requires an in-depth empirical study of the effects of sequence, correlation and combination of different question forms on quantitative indicators of student responses for the reliability of pedagogical measurement, as well as to ensure the reliability of test tools. The review of scientific works enabled to generalize that the researchers consider the issue (of the order of the questions in the test as the factor of impact on numerical indicators of students'

performance) in the following aspects: (1) question order by the level of complexity or in any other order (Monk & Stallings, 1970; Ollennu & Etsey, 2015; Şad, 2020; Zhuk & Vashchenko, 2020); (2) placement of questions taking into account their form (closed, open test questions) (Mollenkopf, 1950); (3) the effect of the previous or the next question on the results of solving a certain task (effect of knowledge transfer) (Gray, 2004); (4) the connection between the order of the questions with the confidence of students in the correctness of their solution; (5) the relationship of the order of the questions with such effects: the "effect of practice", "effect of fatigue", "effect of the impact on the course of cognitive activity" of students in the testing process (Gray, 2004; Pools & Monseur, 2021) etc. At the same time, the issue of the effect of question order in the test on the outputs today remains a debatable issue due to the ambiguity of the conclusions of numerous studies conducted taking into account the specifics of a particular subject area of knowledge.

3. Research Method

The presented article proposes a new approach to the complex research and analysis of linguistic tests, the results of which highlight general trends and determine the directions of further research.

3.1 Study Objectives

The aim of the study, the results of which are presented in the article, is to test the methodology developed by the authors to assess the degree of influence of a set of functional and variable factors on the results of testing students in Ukrainian to further determine the optimal test structure. The test results of the proposed study are considered as generalized numerical values of students' answers to test tasks, which is an effective indicator of testing (OI), which reflects the set of influences of several individual factors (OI = f {F1; F2; F3; Fn}). The study examined the influence of such factors: (1) “test questions form” factor (F1); (2) combination of different questions forms in one test (F2); (3) cognitive factor (F3); (4) “profile specialization of class” factor (F4); (5) affective factor (F5); (6) the ratio of questions of different forms in the test (F6); (7) the order of the questions of different forms in the test (F7).

3.2 Participants

The research was conducted in several stages in the situation of real schooling in Ukrainian language lessons. At the first stage, testing of 8th-grade students (N = 77) and 9th-grade students (N = 87) was conducted. Participants in the study (N = 164, the average age M = 14.04 (SD = 0.65)) were students of specialized classes (economic – Econ, chemical-biological – CB, physical-mathematical – PM); in the second phase, testing 2 and testing 3 were conducted. Participants in testing 2 were students of 9th grades (N = 81), 95% of them participated in the first phase of the study, studying in 8th grades; the average age of participants M = 14.66 (SD=1.33). Participants in test 3 were 87 students of 9th grades (M = 14.28 (SD = 0.44)), 47.12% were female.

3.3 Instrument

To achieve the goal of the study, the authors have developed a number of tools such as tests and a questionnaire.

1. Two variants of parallel in content and complexity criterion-oriented linguistic tests aimed at verifying the mastery of lyceum students of the basic elements of the current at the time of testing the program in the Ukrainian language. Each of the 4 blocks of the test contained tasks of one form: block I - MC questions, which provided the choice of one correct answer with using the text (MCtext); block II - MC with the choice of one and several correct answers - (MCsingle + multiple); block III - matching questions (MQ); block IV - open-ended questions that provided a student-designed response (CR questions). The content of the tasks of each block was identical (aimed at mastering the same elements of the curriculum), the operational material was different, but was selected taking into account the equivalent complexity. In the proposed design of tests, the authors provided the opportunity to compare the results of test tasks of different forms with the same sample of study participants in the same test conditions (to eliminate random unaccounted for factors and differences in different samples). Analysis of the reliability of the tests of each option based on the calculation of Cronbach's alpha coefficient made it possible to conclude the internal consistency of the tests (Variant I = 0.641, Variant II = 0.673). The same calculations were performed for each possible combination of the described task blocks - the value for them ranges from 0.6 to 0.7. Thus, Cronbach's alpha coefficient calculations have shown the suitability of the tests for use in the study.
2. Questionnaire of students' attitude to different forms of test questions (by the method of semantic differential, which involves assessing individual blocks of tasks on the proposed bipolar traits (scaling procedure), the intensity of each of the 12 pairs of traits ranged from -3 to +3). Such a questionnaire was used in the study to study the affective factor influencing the test results.

3. For a refinement study, stem-equivalent tests were developed with a variable ratio of response forms - closed MC questions with the choice of one correct answer (MC_{single}) and open-ended (CR questions) (Figure 1).

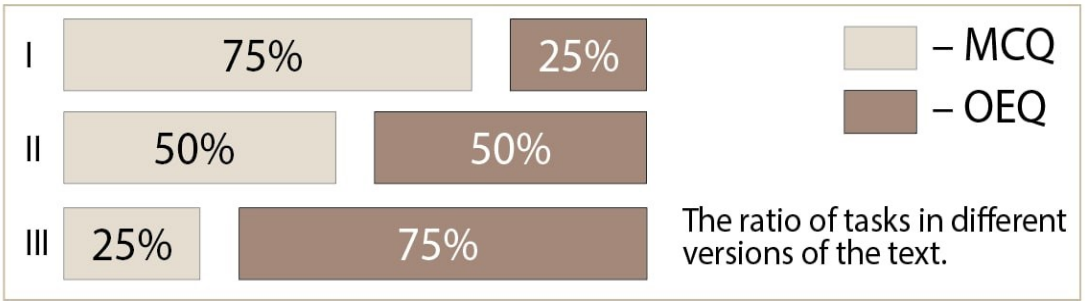

Figure 1. The Structure of Tests with Different Ratios of Questions

Source: Authors

It was assumed that such a structure of tests will allow drawing conclusions about the impact of factor 6 - the ratio of different forms of tasks in one test - on the test results. Since these tests were developed in the refinement phase, the authors constructed them from the most commonly used forms of tasks in school practice and taking into account the results of the first phase of the study. The combination of these question forms revealed a low level of influence on test results, which allowed to eliminate factors 1 ("test questions form") and 2 ("combination of different question forms in one test"), and provided the ability to identify the impact of different ratios of MC and CR questions on students' performance. The reliability of each version of the authors' tests, with which students worked in the second phase of the study, was determined by calculating α Cronbach using two-way analysis of variance, which confirmed the suitability of the tool for use in the proposed study (Table 1).

Table 1. Descriptive Statistics of Collected Data of Testing 2

% of MCQ	N	M	SD	SE	Min	Max	α Cronbach
75%	26	9,5	2,5	0,49	4	16	0,634
50%	26	11,38	3,62	0,72	1	19	0,819
25%	29	9,59	3,64	0,68	2	18	0,827
Total	81						

1. The test constructed from the MCQs of various complexity. These question forms (MC_{single} and MC_{text}) were selected taking into account the results of the first phase of the study. It is assumed that the MC text questions are characterized by increased operational complexity because to find the answer to the question the student must "go beyond" the test task and analyze additional language-and-operational material (text, sentences). Such activity increases time for answering and complicates the cognitive activity of the student. It may be a "hidden" factor that affects the test results in each case. In conventional MCQ, the activity takes place "inside" the test task (Figure 2).

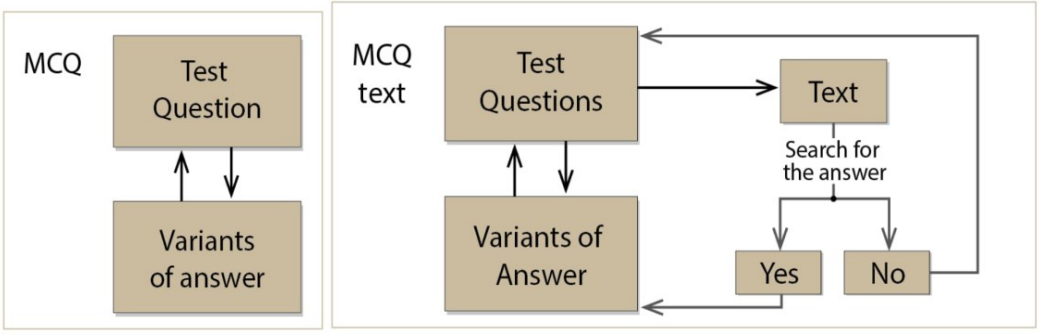

Figure 2. Operational Complexity of the Questions (MC_{single} and MC_{text})

Source: Authors

In the first version of the test (variant I), the questions are ordered from more complex to simpler (in descending order of difficulty), in the second version (variant II) - from simpler to more complex (in ascending order of difficulty). The choice of question forms for testing 3 is due to the results obtained in the first phase of the study on the feasibility of combining certain question forms and the impact of cognitive complexity of tasks on test results (these results are described below in the relevant section). This allowed us to investigate the effect of the order of questions of different operational complexity on the students' effectiveness and determine its impact power, eliminating other factors studied in the first phase (the effect of the question forms and their combination in one test on test results). The reliability of the authors' tests, with which students worked, was determined by calculating α Cronbach using two-way analysis of variance (Table 2).

Table 2. Descriptive Statistics of Collected Data of Testing 3

Source of variation	N	M	SD	SE	Min	Max	a Cronbach
Variant I	44	9,89	3,04	0,46	4	17	0,572
Variant II	43	8,95	2,91	0,44	4	15	0,515
Total	87						

3.4 Process

The first phase of the study (testing 1). The content of the developed tests was based on the material of the a 8-th grade. Therefore, testing was conducted in two sessions - for a 8-th grade at the end of the school year and for a 9-th grade at the beginning of the school year, each testing session was held on the same day for all participants. The duration of the tests and filling in the questionnaire forms was 45 minutes.

An important condition for comparing the performance of different blocks is their interchangeability, which we tried to achieve not only with the content construct, but also compliance with the Number Correct scoring rule (Kastner & Stangl, 2011) – this rule made it possible to take into account the partial knowledge in the assessment and not to reduce the scores for incorrect answers.

The second phase of the study (testing 2 and testing 3) was conducted to refine and expand the results of the first phase. These tests were conducted in the second and the first semesters of participants' studying in the 9th grade, respectively, in the real educational process. Execution of tests by each group of participants lasted about 35 minutes. The results were calculated in each case according to the dichotomous system of evaluation of test tasks (0 and 1 point).

3.5 Data Analysis

Processing of results, in particular analysis of variance, factor analysis, etc., was carried out using MS Excel and STATISTICA software packages.

Testing 1. It is determined that the total test results, as well as the results of each sample (class) and the results grouped by each block of tasks, comply with the law of normal distribution. The correctness of the hypothesis about the equality of secondary populations (by educational level - 8th grade students (N1 = 77; M1 = 7.32; SD1 = 16.25) and 9th grade students (N2 = 87; M2 = 7.80; SD2 = 12.76)) was tested by various statistical methods. According to the Wilcoxon-Manny-Whitney criterion, the psychometric characteristics of the test results of these samples coincide at the level of 0.05 - the empirical value of 0.8896 with a critical value of 1.96. Comparison of the test results of these samples showed a weak statistical power (0.056), the measure of the effect of d was also weak (= 0.12665) (Cohen, 1988). Statistical equality of the samples was checked by Student's t-test: t (1.96) = 0.827. Thus, according to certain statistical indicators, the average population is equal, which allowed them to be combined into a single sample (N = 164; M = 14.04; SD = 0.65).

The influence of the studied factors on the results of testing a single sample was determined by multidimensional analysis of variance (in each case, the accuracy of calculations exceeded 99% ($p < 0.01$)). Snedekor's method was used to determine the strength of the influence of factors.

The effective (numerical) indicators of testing in the qualitative dimension are compared with the verbal-numerical results of the survey by the method of semantic differential. To study the generalized perceptual-emotional attitude of students to questions of different forms, which they worked with during linguistic testing, factor analysis of survey results and dominant features and depth factors that characterize the affective factor influencing the results of tasks (Hryvko & Zhuk, 2019).

Testing 2. As expected, testing 2 was aimed at comparing the results of tests with different ratios of MCQs and CRQs. Accordingly, to verify the significance of the difference in test results between the three test structures, the authors used a one-way analysis of variance. As shown in Table 3, the difference between the three test structures is statistically significant. Further analysis showed that the results of responses to different forms of test questions also have large differences (Tables 4 and 5).

Table 3. Impact of Test Structure on Test Results

Source of variation	SS	Df	MS	F	Sig.*
Between groups	66,67885	59	1,13015	5,244003	1,29E-31
Within groups	323,2692	1500	0,215513		
Total	389,9481	1559			

*Note: N=81; *p<.0001*

Table 4. Results of Answers to MCQs

Source of variation	SS	Df	MS	F	Sig.
Between groups	0,688661	2	0,34433	6,506805	0,002478
Within groups	3,968889	75	0,052919		
Total	4,65755	77			

Table 5. Results of Answers to CRQs

Source of variation	SS	Df	MS	F	Sig.
Between groups	1,361481	2	0,680741	10,53936	9,25E-05
Within groups	4,844274	75	0,06459		
Total	6,205755	77			

The effect of the question form in both cases is statistically significant (Table 4: 17.48%, $p < .01$; Table 5: 26.84%, $p < .0001$). The impact power degree of the question forms on the test results was determined with the application of the method of Snedecor.

Comparison of assessment objects average was carried out according to Cohen (1988).

In each variant, the effect of the question format turned out to be significant: variant I: $F(1.61) = 8.61$, $p < .0001$; variant II: $F(1.59) = 3.61$, $p < .0001$; variant III: $F(1.61) = 7.76$, $p < .0001$. The effect of gender is also significant: variant I: $F(1.53) = 2.73$, $p < .0001$; variant II: $F(1.53) = 5.25$, $p < 0.0001$; variant III: $F(1.49) = 5.78$, $p < .0001$.

Testing 3. As with test results 2, an one-way analysis of variance was used to determine the significance of the difference in test results between the two test structures (this time differing in the question order). This difference is statistically significant, as is clear from Table 6 ($p < .0001$).

Table 6. The Effect of Test Structure on Test Results

Source of variation	SS	Df	MS	F	Sig
Between groups	71,2186	39	1,826118	8,582874	1,25E-43
Within groups	357,4419	1680	0,212763		
Total	428,6605	1719			

The use of two-way analysis of variance made it possible to establish that on each variant the impact of two factors on the test results is statistically significant: the factor of the order of different format questions in the test (variant 1: $F(1.60) = 10.39$, $p < .0001$, = 0.177; option 2: $F(1.59) = 8.12$, $p < .0001$; = 0.148), and gender (option 1: $F(1.39) = 2$, 33, $p < .0001$, = 0.090; Option 2: $F(1.39) = 2.06$, $p < .0001$; = 0.083). However, the effect of the question order in the test on test results is almost twice the effect of the gender factor.

Table 7. The Difference of the Students' Performance on the Tasks of Different-Formats Test Segments

Variants	Objects of assessment	% of correct answers		Impact of the test item format (by Snedekor)	Cohen's d
		MC_{text}	MC		
	Test	53,86%	45,0%	16,52%	0,579
I	Males	52,40%	43,20%	20,94%	0,568
	Females	51,58%	47,37%	11,14%	0,418
	Test	49,53%	40,23%	13,49%	0,551
II	Males	44,76%	35,71%	14,46%	0,568
	Females	54,09%	44,54%	13,78%	0,574

*Note: *p<0001*

As can be seen from Table 7, the differences in the mean values of the results of different test segments in each variant are significant, but comparable. Comparison of the test results between variants I and II as a whole shows an average difference (Cohen's d = 0.306). In the MC_{text} and MC segments, the difference between variant I and II is 0.253 and 0.271, respectively. The difference in the results of performing these segments 1 and 2 of test variants by the males (0.497 and 0.452) and females (0.091 and 0.107) shows that the greatest differences are observed in males; there are practically no differences among females. Comparison of the mean values of the assessment objects was carried out according to Cohen's method.

Table 8. The Difference in Test Performance in the Gender Aspect

Variants	Objects of assessment	% of correct answers		Impact of the test item format (by Snedekor)*	Cohen's d
		Males	Females		
	Test	47,80	51,58	15,82%	0,243
I	MC_{text}	52,40	55,79	8,38%	0,185
	MC	43.20	47,37	**23,16%**	0,227
	Test	40,24	49,32	14.16%	0,652
II	MC_{text}	44,76	54,09	7,15%	0,585
	MC	35,71	44,54	**20,67%**	0,357

*Note: *p<0001*

4. Main Results of the Research

One of the main factors influencing the performance of testing in our study is directly considered the form of the test task. The author's hypothesis about the presence of the influence of this factor was confirmed by the results of the analysis of variance, which showed that the strength of such a factor can be detected in up to 35%.

A comparative analysis of test results for the Variants I and II clearly showed that depending on the texture of the test fluctuates the strength of the factor under consideration on test results. On average, this figure is 45.5% for Variant I and 27.8% for Variant II, however, It should be noted that the tendency of weaker manifestation of the strength of the impact of the question form on OI in Variant II is equally observed in the results of students of different educational levels (when dividing the sample into age groups).

The tendency of similarity of indicators of the force of influence on RP of the factor "form of test task" is revealed at a comparison of results of testing of groups of pupils of 8 and 9 classes (on educational level) and classes of various profiles (PM, Econ, CB). The similarity of indicators of the strength of this factor on the test results of students studying in classes of different profiles and at different educational levels can be explained by the peculiarities of the subject - the state language, which is regularly used in teaching and life of tested students. That is, the content of testing in the proposed study, in contrast to the content of other educational areas, has a constant activity. Thus, we can conclude that the impact of training profile on test results depends on the subject field of assessment. However, the probability of these conclusions should be tested experimentally in further research (in the subject field of other fields of education).

The study found that the variation of different question forms combinations in one test affects the strength of this

factor impact. It is theoretically proved that high diagnostic reliability, which minimizes the impact of external factors (taking into account compliance with psychometric quality criteria - reliability, validity, etc.) on test results (up to 10%), is characterized by a test using the following combinations of question forms: (1) closed with a possible choice of one or more correct answers, and closed tasks on the text, involving the choice of one correct answer; (2) open-ended tasks with a student-designed answer and closed-ended text-based tasks with only one correct answer; (3) open-ended tasks involving students' constructed answers and closed-ended tasks with one or more correct answers. Any combination of three or more question forms causes a certain "distortion" of the results because in each case the impact power of the considered factor on the test results is more than 20% (Table 9).

Table 9. Calculation Results of the Effect of Test Structure on Tests' Scores

Combinations of questions forms (variants of the test structure)	**The impact power** (of the test structure on tests' score)	**Significance level of data***	**Test reliability** (α Cronbach)
I- II	*8,45%*	*F(2,01)=16,14*	*.566*
III-IV	17,30%	F(2,01)=35,32	.671
II-III	35,13%	F(2,01)=89,82	.6545
II-IV	*9,02%*	*F(2,01)=17,25*	*.6431*
I-IV	*8,79%*	*F(2,01)=16,8*	*.6849*
I-II-III	30,79%	F(1,80)=73,98	.6427
I-III-IV	20,35%	F(1,79)=42,89	.6526
II-III-IV	21,75%	F(1,79)=46,58	.6434
I-II-III-IV	21,27%	F(1,67)=45,31	.6373
I-III	32,89%	F(2,01)=81,37	.661

*Note: N=164; *p<.0001*

These tables clearly show that the most reliable and internally consistent combination of tasks (α Cronbach = .6849), which has a significantly lower level of impact on test results, is SingleMC for text + OEQ (Impact power = 8.79%).

Table 10. Differences in the Performing Results

% of MCQs in the test	**Objects of assessment**	**% of correct answers**		**Degree of question formats effect (by Snedekor)**	**Cohen's d**
		MCQ	**CRQ**		
75%	Test	53,30%	30,00%	35,6%	**1,093**
	Males	49,80%	23,53%	44,16%	**1,306**
	Females	60,00%	42,22%	19,93%	**0,849**
50%	Test	51,54%	61,54%	3,92%	0,396
	Males	54,67%	62,67%	2,77%	0,324
	Females	47,27%	60,00%	5,74%	0,474
25%	Test	73,10%	39,54%	46,28%	**1,351**
	Males	73,75%	40,00%	31,74%	**1,323**
	Females	72,31%	38,97%	44,98%	**1,338**

At the same time, the formation of evidentiary conclusions requires clarification of the results. The test, which was covered analysis, contained an equal number of different form questions due to the need to minimize the possibility of question numbers impact on test results. Such test constructions made it possible to find out which question forms should be combined into one test to avoid "distortion" of the test results. However, different question forms are often combined into one test in different proportions (in school practice). Therefore, we performed testing 2 within our study to find out what the ratio of different question forms should be (in particular those that according to the results of testing 1 should be combined in one test). It may enable to find out the real level of academic achievement of

students in a random sample (determine some ideal external structure of the test, which will allow to assess the knowledge (skills) of each student, regardless of his cognitive style of thinking) (Bridgman B., Morgan R., etc.). We were able to find that the factor "question forms" has the least effect on the results of testing in the Ukrainian language if the ratio of MCQ to CRQ is 1: 1 (50%). In the case of a ratio of 3: 1 (75% / 25%), as well as 1: 3 (25% / 75%), the strength of the impact of this factor increases by an average of 2. 5-3, and sometimes 4 times (Table 10).

According to the analysis of the performing the tests with different proportions of MC and CR questions in terms of gender characteristics of the studied audience, it was found that a significant gender effect is found in the testing with using variant I of the test (which consists of 75% of MCQs). In this case, the results of performing both MCQs and CRQs by females are higher (from 10 to 20%) than the results of males. In the case of another question ratio (1: 1; 1: 3), the results of males and females are close in numerical value, the factor "gender" in this case is insignificant (Table 11).

Table 11. The Difference in the Test Performing in Gender Aspect

% of MCQs	Objects of assessment	% of correct answers		Degree of the question formats effect (by Snedekor)	Cohen's d
		Males	Females		
	Test	43,24%	55,66%	24,93%	**0,862**
75%	Closed test	49,80%	60,00%	20,15%	**0,656**
	Open test	23,53%	42,22%	21,66%	**0,760**
50%	Test	58,67%	53,64%	1,24%	0,218
	Closed test	54,67%	47,27%	2,12%	0,292
	Open test	62,67%	60,00%	2,71%	0,102
25%	Test	48,44%	47,31%	0,06%	0,050
	Closed test	73,75%	72,31%	0,08%	0,058
	Open test	40,00%	38,97%	0,04%	0,040

The number and nature of operations that must be performed in the process of the task determines the presence of the operational-activity component of complexity, which was found in the process of analyzing the cognitive complexity of the students offered parallel tasks of various forms. In the course of research in the analysis of tasks that involved students to match two sets and which are identical parallel tasks with the ability to choose several or only one correct answer for cognitive complexity. Such tasks are characterized by cyclical actions (operations) during their solution: with each repetition of a certain action varies the choice of "extra" descriptors (compared to tasks where you need to choose one correct answer), which in turn significantly reduces the diagnostic accuracy of test results and makes it possible the probability of "guessing" when finding the answer by the tested student. This hypothesis was confirmed by the authors in the process of conducting a general analysis of the comparison of test results (testing 1) (compliance tasks showed a higher level of performance compared to parallel test tasks of other units) with analysis of the impact on test structure results: reduction of diagnostic accuracy (Table 9) identified in the presence of the structure of the test tasks that provide for the establishment of compliance and which are combined with other closed test tasks.

The complexity of mental operations is determined by the corresponding "behavioral dimension of problem-solving" or "skill measurement", which correlates with the taxonomy of levels (or categories) of cognitive actions and is regulated by the predicted number of cognitive operations in the form of a task, the formulation of the task, the accompanying material to the task and in the form of a response. This means that not only different question forms can be different in complexity, but also questions presented in the same form according to the presence/absence and nature of the accompanying material (additional operational material – sentences, text, diagrams, etc.), as well as answering forms in accordance with the question form. For greater clarity, it is necessary to compare test tasks with the choice of only one correct answer to test knowledge of a particular theoretical construct with test tasks that involve reading the text or performing certain operations to select one correct answer and test the application and the theoretical construct knowledge. Another example - the open tasks to explain and open tasks to fill in the gaps will be characterized by different cognitive complexity.

At the same time, the impact of tasks of different complexity on the test results is associated with the order of their

placement in the test. Thus, research of Ollennu and Etsey (2015) (conducted at the level of the Basic Education Certificate Examination (BECE)) showed a statistically significant difference in the results of tests with different order of multiple choice tasks (random, from simple to complex, from complex to simple). According to the results of statistical analysis of the study, the authors concluded that changing the order of tasks affects student performance in testing in English, Maths, and Science, and therefore rearranging tasks in the test is unacceptable. In contrast, there are studies that show a slight effect of the order of tasks on the complexity of the test (Hohensinn & Baghaei, 2017) and studies that refute the impact of randomization of tasks on the level of complexity of the test and test results (Satti et al., 2019).

The placement of the questions, taking into account their form, is associated with different complexity of different types of questions of closed and open forms, which at the same time depends on the subject on which the educational material is tested. Thus, according to a study of Mollenkopf (1950), the question order does not affect the results of testing students in Math, however, under the same organizational conditions, such an impact is observed in the case of testing students' verbal abilities. Accordingly, the results of testing 3 confirmed the influence of the factor "order of the test questions" on the results of testing in the Ukrainian language, which may indicate the presence of "fatigue effect" (Gray, 2004) or the “effect on student productivity” during testing (Ollennu & Etsey, 2015; Pools & Monseur, 2021), but this thesis requires additional empirical evidence. At the same time, the analysis of the gender factor of differences in the results of testing showed that changing the question order affects the effectiveness of testing males (decreases in the case of the increasing complexity of tasks). In testing females, such an effect is almost absent.

5. Discussions

The results of the proposed research conducted in the subject field of linguistic testing correlate with the results of research conducted on the subject of other educational fields, in particular on the impact on test results in the form of tasks (e.g. Jonick et al., 2017). Closed forms (Melovitz Vasan et al., 2018), substantiation of expediency of constructing multidimensional polymorphic tests to assess students' learning achievements with different abilities, thinking styles and cognitive learning models (e.g. Kan, Bulut & Cormier, 2019).According to the results of our study, the difference between the tasks presented in a different order (in descending and increasing operational complexity) varies within 5%: the average result of the first variant of the test is 49.05% (tasks are placed in descending order of complexity), the second variant (increasing complexity) - 44.9% (these are results of three samples of 9th-grade students). These results coincide with the results of the study of Monk and Stallings (1970). The authors state that although the results showed a change in test complexity within 5%, they have important prognostic value: despite the fact that for a small sample such a difference is insignificant and acceptable, for large-scale testing it is will significantly affect the results of a large number of students, so in such a situation, the assessment of the possibility of random question organization in the test should be avoided. This thesis is reinforced by the identification of gender "sensitivity" to the factor "order of test questions". Such conclusions are a strong argument to confirm the position on the inexpediency of randomization of tasks in the process of normative-oriented testing, the purpose of which is the ranking or selection of students by academic achievement. In the case of criterion-based assessment, randomization is acceptable, especially in the context of the remote assessment of students using computer-based testing, in particular, to avoid cheating. At the same time, in order to ensure the diagnostic potential of such an assessment, it is necessary to limit the possibilities of returning to previous tasks, as well as the testing time. Randomization of test tasks is also permissible in terms of formative assessment, which is essentially educational, i.e. its purpose is not so much a test score as the teacher's determination and awareness of the student of his real at the time of testing academic achievements and progress.

However, according to the analysis of seven factors affecting the results of testing, we can offer an ideal externally-organized design of a criterion-oriented test in the Ukrainian language (for in-school control and classroom assessment), which combines MCQs and OEQs in equal proportions in descending complexity of questions order. Such a test design takes into account the gender characteristics of their perception by testing participants.

We consider it necessary to note that the presented results of the complex study are preliminary and need clarification and control verification on a larger sample and in slightly changed conditions, in particular with other rules of assessment of test tasks (Kastner & Stangl, 2011).

6. Conclusions

The paper experimentally proved that the form of the test question can affect the test results with variable strength up to 35% and depends on the external organization of the test. At the same time, the results of tasks parallel in content and form (which was differed only in the means of language and operational material) in Variant I and Variant II are tendentiously similar, but the strength of the factor "test question form" on test results varies. This fact indicates the presence of the factor "language design of the task", which necessitates further research aimed at determining the predictors of the impact of language tools and ways of formulating the task on the results of its implementation.

Based on the results of a comprehensive study, we can assume that the diagnostic reliability of the results of criterion-oriented testing in the Ukrainian language can be increased by including open tasks in the test equal proportion of closed tasks. This assumption is confirmed by the analysis of the affective factor. We found that the formation of students' interest and personal attitudes to perform tasks depends on the form of the task. The same content of parallel tasks is perceived differently by students in different blocks of the test, which allowed us to conclude that the test question form is a sensory-afferent stimulus of the operational image of activity, on which its further implementation depends (Hryvko & Zhuk, 2019). This conclusion proves once again that in order to form a motivational optimum and positive attitudes of students to perform the test in its structure should combine different forms of test questions and tasks of different cognitive complexity, which will allow to clarify and verify the test results.Given the presented results, we consider the study of the issue of language and operational material of the test questions as an impact factor of their understanding and implementation in various subject areas as a promising direction of further research.

References

Akhavan Masoumi, G., & Sadeghi, K. (2020). Impact of test format on vocabulary test performance of EFL learners: the role of gender. *Lang Test Asia, 10*, 2. https://doi.org/10.1186/s40468-020-00099-x

Alharbi, A. (2017). Review on the effects of test formats on achievement and retention. *International Interdisciplinary Journal of Education, 6*(4), 273 - 278.

AlKhatib, H. S., Brazeau, G., Akour, A., & Almuhaissen, S. A. (2020). Evaluation of the effect of items' format and type on psychometric properties of sixth year pharmacy students clinical clerkship assessment items. *BMC Medical Eduation, 20,* 190. https://doi.org/10.1186/s12909-020-02107-3

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice: designing and developing useful language tests.* Series: Oxford applied linguistics (1st ed.). Oxford, UK: Oxford University Press.

Gray, K. E. (2004). *The effect of question order on student responses to multiple choice physics questions.* (Master's Thesis, Kansas State University, Manhattan, KS). Retrieved from https://www.phys.k-state.edu/ksuper/dissertations/gray.pdf

Hancock, G. R. (1994). Cognitive complexity and the comparability of multiple-choice and constructed-response test formats. *The Journal of Experimental Education*, *62*(2), 143-157.

Hohensinn, C., & Baghaei, P. (2017). Does the position of response options in multiple-choice tests matter? *Psicológica*, *38*, 93-109.

Hryvko, A. V., & Zhuk, J. O. (2019). Using the means of computing technologies in the process of experimental research of the students' emotive-estimate relation to different forms of test tasks in Ukrainian language. *Information Technologies and Learning Tools. Theory, Methods and Practice of Using ICT in Education*, *70*(2), 285-297.

Johnson, J. (2010). Designing with the mind in mind. In: B. Begole, & J. Kim (Eds.), *Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems* (pp. 2501-2502). New York, NY: Association for Computing Machinery. https://doi.org/10.1145/2702613.2706667.

Jonick, C., Schneider, J., & Boylan, D. (2017). The effect of accounting question response formats on student performance. *Accounting Education, 26*(4), 291-315.

Kan, A., Bulut, O., & Cormier, D. C. (2019). The impact of item stem format on the dimensional structure of mathematics assessments. *Educational Assessment*, *24*(1), 13-32.

Kastner, M., & Stangl, B. (2011). Multiple-choice and constructed response tests: do test format and scoring matter*? Procedia - Social and Behavioral Sciences*, *12*, 263-273.

Lissitz, R. W., Hou, X., & Slater, S. (2012). The contribution of constructed response items to large scale assessment: measuring and understanding their impact. *Journal of Applied Testing Technology*, *13*(3). Retrieved from http://www.jattjournal.com/index.php/atp/article/view/48366.

Marengo, D., Miceli, R., & Settanni, M. (2016). Test unidimensionality and item format: Do mixed item formats threaten test unidimensionality? Results from a standardized math achievement test. *Testing, Psychometrics, Methodology in Applied Psychology*, 23(1), 25-36.

Martin, M. O., von Davier, M., & Mullis, I. V. S. (2020). *Methods and procedures: TIMSS 2019 technical report*. Amsterdam, The Netherlands: International Association for the Evaluation of Educational Achievement.

McKenna, P. (2019). Multiple choice questions: answering correctly and knowing the answer. *Interactive Technology and Smart Education*, *16*(1), 59-73. https://doi.org/10.1108/ITSE-09-2018-0071

Melovitz Vasan, C. A., DeFouw, D. O., Holland, B. K., & Vasan, N. S. (2018). Analysis of testing with multiple choice versus open-ended questions: Outcome-based observations in an anatomy course. *Anatomic Science Education*, *11*(3), 254-261. https://doi.org/10.1002/ase.1739

Mollenkopf, W. G. (1950). An experimental study of the effects of item analysis data of changing item placement and test time limit. *Psychometrika, 15*(3), 291-315.

Monk, J. J., & Stallings, W. M. (1970). Effects of item order on test scores. *The Journal of Educational Research, 63*(10), 463-465.

Mozaffari, F., Mohammad Alavi, S., & Rezaee, A. (2017). Investigating the impact of response format on the performance of grammar tests: selected and constructed. *Journal of Teaching Language Skills, 36*(2), 103-128. https://doi.org/10.22099/jtls.2017.23918.2154

Mullis, I. V. S., & Prendergast, C. O. (2017). Developing the PIRLS 2016 achievement items. In M. O. Martin, I. V. S. Mullis, & M. Hooper (Eds.), *Methods and Procedures in PIRLS 2016* (pp. 1.1-1.29). Amsterdam, The Netherlands: International Association for the Evaluation of Educational Achievement.

Organisation for Economic Co-operation and Development, (OECD). (2020). *PISA 2018 technical report*. Retrieved from https://www.oecd.org/pisa/data/pisa2018technicalreport/

Ollennu, S., N., N., & Etsey, Y. K. A. (2015). The impact of item position in multiple-choice test on student performance at the basic education certificate examination (BECE). *Universal Journal of Educational Research, 3*(10), 718-723. https://doi:10.13189/ujer.2015.031009

Onaiba, A., & Jannat, F. (2019). Test method effect and test-takers' scores: a critical review of the pertinent literature. *Scientific Journal of Faculty of Education, Misurata University-Libya, 1*(14), 3-22.

Orlov, A., Ponomareva, T., Chukajev, O., & Pazuhina S. (2017). *Technologies for assessing the results of the educational process in a university in the context of a competency-based approach* (2nd ed.). Moscow, Russian Federation, Berlin, Germany: Direkt-Media.

Pools, E., & Monseur, C. (2021). Student test-taking effort in low-stakes assessments: evidence from the English version of the PISA 2015 science test. *Large-Scale Assessments in Education 9,* 10. https://doi.org/10.1186/s40536-021-00104-6

Reardon, S. F., Kalogrides, D., Fahle, E. M., Podolsky, A., & Zárate, R. C. (2018). The relationship between test item format and gender achievement gaps on math and ELA tests in fourth and Eighth grades. *Educational Researcher*, *47*(5), 284-294. https://doi.org/10.3102/0013189X18762105

Şad, S. N. (2020). Does difficulty-based item order matter in multiple-choice exams? (Empirical evidence from university students). *Studies in Educational Evaluation, 64,* 100812. https://doi.org/10.1016/j.stueduc.2019.100812

Satti, I., Hassan, B., Alamri, A., Khan, M., & Patel, A. (2019). The effect of scrambling test item on students' performance and difficulty level of MCQS test in a college of medicine, KKU. *Creative Education*, *10,* 1813-1818. https://doi.org/10.4236/ce.2019.108130

Schladitz, S., Groß Ophoff, J., & Wirtz, M. (2017). Effects of different response formats in measuring educational research literacy. *Journal for Educational Research (Online), 9(*2), 137-155.

Srivastava, A., Dhar, A., & Aggarwal C. S. (2004). Why MCQ. *Indian Journal of Surgery, 66*, 246-248.

Zhuk, Y., & Vashchenko, L. (2020). Features of the influence of the composition of the test in biology on high

school test scores. *Ukrainian Educational Journal*, *1*, 20-31.