The research of the effect of different question forms and their combination variability on the results of testing

Antonina Hryvko*, and *Yurii* Zhuk

Monitoring and Assessment of the Education Quality Department, Institute of Pedagogy of the National Academy of Pedagogical Sciences of Ukraine, 52-D Sichovykh Striltsiv Str., Kyiv, 04053, Ukraine

Abstract. The purpose of the research was to identify and study the features of the effect of four questions forms and their combination in polymorphic test on the output indicators of secondary school students' testing. In the study, conducted in the subject field of studying native (Ukrainian) language, took part students of 8th and 9th-grades (N=164). The tools (criterion-oriented author's tests, questionnaire) and methods of results' analysis (factor analysis, dispersion analysis, etc.) used in the research process, made it possible to prove the existence of the effect of factor "question form" on the test scores (impact power is up to 35%), and also to determine combination of question forms that is characterized by the highest diagnostic accuracy. Results of a study indicate that it is advisable to construct a polymorphic linguistic test. The combination of the multiple-choice questions with one correct answer with using the text and constructed-response questions can ensure the clarity of output indicators of testing (with the least effect of question forms on them), in case evenly balance of these questions in the test. The disadvantages of the study are pointed out and prospects for their continuation and deepening, connected with further study of the optimal empirically valid test structure, are substantiated.

1 Introduction

1.1 Problem statement

The presence of discussed in the pedagogical community shortcomings of education quality evaluation through the use of test technologies on the one hand, and the use of test technologies to evaluate educational achievements of education seekers on the other, including for the further decision-making on the test subjects (high-stakes tests), updates researches related to the study of peculiarities of test technologies usage to objectively control students' educational achievements. Such studies should be aimed to determine the optimal questions forms and test structure in order to improve the method of the construction assessment tool in terms of functional and variation factors that affect the results of its application.

The purpose of the proposed article is to elucidate the study results of factors "question form" and "variability of combining different forms of questions in one test" in the context of effect study of a set of functionally-variational factors on the test results.

1.2 The rationale for the relevance of the study

Despite the bunch of studies, the problem of the effect of question forms on the test results remains relevant in the theory of testing. A number of scientific papers argue that the test with questions of close type and selected responses is easier than the open-ended test [1; 2; 3], which has led scientists to conclude that it is appropriate to combine questions of different forms in the test. In contrast to the situation about the different complexity of closed and open-ended questions and their corresponding impact on overall assessment results, some publications [4; 5; 6] highlight the results of studies that prove the equivalence of such questions in terms of statistical significance and reliability. In addition, some papers (for example, [7]) provide empirical information (correlation index between total scores from different subjects tests with construct-response (CR) questions and without them), which, according to the authors, is the evidence that conclusions for the test results will not change significantly, if you remove open-ended questions from the test or replace them with multiple-choice (MC) questions of the corresponding difficulty. M. Kastner and B. Stangl [8] explain the ambiguity of research results, comparing multiple-choice questions and open-ended questions, and the lack of consensus among scientists on the equivalence of diagnostic potential for such questions: the authors experimentally demonstrated the dependence of such studies results on the type of compared questions forms and scoring methods for evaluating tests and thereby questioned the reliability of some of them. The presence of different scientists' conflicting views on the problem of comparative assessment of the CR and MC questions importance makes it necessary to consider the few

* Corresponding author: av.hryvko@gmail.com

investigated aspects of this problem. In particular, the problem of measuring the effect of different questions forms of different types and different variants of combining them in the test on the output indicators of testing is not solved. It determines the relevance of studies, aimed at determining the optimal structure of the test, as well as finding out whether different question forms affect the test results of students performance who study in certain profile classes (physical, mathematical, humanities, etc.). The study presented in the proposed article is an attempt to solve the identified questions.

2 Theory

2.1 The theoretical basis of the research hypothesis

The formulation of the proposed study hypothesis (about the effect of functional and variational factors on the test results) is based on Bachman-Palmer theory of facets, developed in the subject field of linguistic testing [9], which proves the need to study different aspects of linguistic achievements evaluation of the tested students, including the coherence of results evaluation and factors, distinguished by researchers from the testing organization procedures and a set of characteristics toolkit. The authors of the theory divided these factors into 5 categories: (1) testing environment; (2) test structure (headings); (3) questions' characteristics; (4) characteristics of predicted responses; (5) relationship between questions and answers [9]. In accordance with the context of our research and on the basis of the assumptions based on the results of theoretical analysis of the scientific literature, we have distinguished functional and variation factors that affect on the results of testing: (1) uncontrollable – respondents' characteristics: personal (cognitive – knowledge and ability, affective – perceptive-emotional evaluation of the activity object – test questions) and group (in particular, class profile); (2) controlled – content and externally organized tests' constructs. The meaningful constructs of the test are, in particular, test questions content (displaying a fragment of the subject content in a test form) and the content of the test in general (optimally reduced reflection education content in the questions system [10]). Test content characteristics are always determined by the purpose of assessment (for example, thematic assessment, general determination of educational attainment levels of students at a particular stage of study; students selection by the level or orientation of their knowledge). As the subject content will always be wider than the test content [10], the issue of selecting the most relevant control material is one of the most problematic and requires the participation of experts. Selection of the test content for the final control is subordinated to the questions of determining levels of mastering basic constructs of the subject by the tested participants, which obviously determines the need for its compliance with the curriculum content and requirements of the standard. In this aspect, predictability of questions content impact on the results of their solution depends on externally organized constructs of the test.

Externally organized constructs of the test determine conditions, means and methods of ordering and expression of its content: setting characteristics (physical testing conditions, time of test questions and the whole test completion); general test characteristics (its specification) - purpose, structure, sequence and number of questions, instructions for the procedure of conducting and results evaluation; test questions format and form of predicted answers. The format means a method of initiating or carrying out certain actions that are accomplished by questions form, language design of the questions (formulation methods, volume and type of material), programming the method of processing material (determines the number of actions required to complete the questions). Forms of questions are ways of ordering its elements: closed (selective – MC question with one correct answer, with several correct answers, with the choice of the most correct answer, with the choice of true or false statement; the questions of establishing the sequence; the matching questions) and open-ended (constructed with a regulated response – add-on questions, freely-constructed – extended response questions). The response form is a method of answering defined by the form of a question. The choice of the appropriate form of the test questions is connected with the necessity to take into account the criterion of pedagogical measurement validity.

2.2 Analysis of scientific works on the topic of the research

An unreasonable choice of the test questions forms leads to a distortion of its content expression and complication of its understanding by tested students [11]. Therefore, scientists [12; 13] mostly associate the choice of test questions forms with the test complexity parameters, which are determined by psychometric test results. G. R. Hancock in his study compared MC and CR tests, designed according to Bloom's taxonomy levels. The author had experimentally demonstrated that the two formats measure similar cognitive constructs at the same taxonomic level but at different levels of complexity [14]. At the same time, researchers [15; 16; 17; 18; 19] believe that the structure of the test should contain different questions forms (polymorphic tests). Such a statement is in line with global trends in assessing academic achievement in an international context. Thus, in R. W. Lissitz et al. study [7], based on the analysis of the psycho-pedagogical literature, it is proved that, despite the possible equivalence of psychometric indicators of the test results, closed questions with the choice of the correct answer from a number of the offered ones cannot replace open questions with a constructed or freely constructed response, because of: first, in the process of solving MC and CR questions, different kinds of thinking activity are activated, and therefore its results are manifestations of different cognitive skills; secondly, the research presented experimental data, which show that some students perform better MC questions and worse – CR ones and others, on the contrary, more successfully write tests with CR questions than MC questions, so a test constructed from single format questions may be a reason of results deterioration of a particular tested category [7]. In this regard, as well as with the high level of questions diagnostic potential that involve a constructed or freely constructed form of response, in international benchmarking tests (PIRLS, PISA, TIMMS), most of the questions are presented in open forms. For the same reason, in Russia in the Unified State Examination in all subjects, since 2016, the share of expanded answer questions has gradually increased [20].

Consequently, the analysis of scientific works on the subject under consideration showed the existence of contradictions in the views of scientists regarding the possibility of using only one form of questions in the test and vice versa – the expediency of combining different forms of questions in accordance with the requirements of pedagogical measurements validity, as well as varying complexity / equivalence of MC and CR questions and the ability to adjust test results by constructing a test of a specific format. The relevance of provision concerning polymorphic tests for the final assessment of students' academic achievement confirms the need for empirical studies to substantiate the optimal structure of the test, which, accordingly, requires the study of the effect of the combination of different test questions forms on the numerical indicators of students' responses (test scores) to ensure the reliability of toolkit and pedagogical measurement.

3 Research method

The presented research is exploratory (in the study proposed a new approach and research methodology, developed by the authors of the article, also the results of their approbation are revealed). The article covers the results of the analysis of the linguistic tests, which at the same time determine the general tendencies and directions of further researches.

3.1 Study objectives

The proposed study is a component of scientific researches, *the purpose* of which is to determine the degree of functional and variational impact factors on the output test indicators (OI) – numerical values of students' answers to test questions (test scores) from the subject: Ukrainian language. *Factors under study:* since the proposed study considers OI as a manifestation of the impacts set of isolated factors (OI = $f\{F_1; F_2; F_3; F_n\}$), the following statistical hypotheses about the effect on the characteristics of the OI were tested during the analysis of the results: (1) "test questions form" factor (F_1); (2) combination of different questions forms in one test (F_2); (3) cognitive factor (F_3); (4) "profile specialization of class" factor (F_4); (5) affective factor (F_5).

3.2 Participants

The study was conducted in the subject field of Ukrainian language studying in the situation of the real educational process between students of 8th and 9th grades (N=164) of the following profiles: chemical-biological (CB), physical-mathematical (PM), economic (Econ). From the total amount of students: the number of 8th-grade students was 77, 9th-grade students – 87. The average age of the participants was M=14.04 (SD=0.65).

3.3 Instrument

In order to study the impact of the above factors, criterion-oriented tests from the subject "Ukrainian language" were developed in two variants (as expected, parallel in content and difficulty). In terms of content, the test questions offered to the lyceum students were consistent with the current Ukrainian language curriculum and provided an assessment of the most important knowledge and skills. Each variant of tests contained 4 blocks of different questions forms, aimed at testing students' mastering of the same studying constructs – knowledge and skills: MC questions, which provided the choice of one correct answer with using the text, hereinafter referred to as MC_{text} (block I), MC with the choice of one and several correct answers – $MC_{single+multiple}$ (block II), matching questions – MQ (block III), CR questions (open-ended questions that provided a student-designed response) (block IV). The tests included instruction on how to complete the questions each of the blocks. Such tests construction made it possible to compare the results of different questions forms with the same students (and not with different samples of students). To determine the consistency of the internal tests, their reliability was analyzed and based on Cronbach's alpha coefficient calculation [21] for each variant (0.641 for 1st variant and 0.673 for 2nd variant), as well as for each of the possible combinations of the described questions' blocks. This value ranges from 0.6 to 0.7, which makes it possible to conclude that the tests are suitable for usage in the research process.

To study the affective factor, a questionnaire was developed for students' attitudes toward different forms of the questions they solve, using the method of semantic differential (SD questionnaire) [22], by which, using a bipolar scaling procedure, students evaluated isolated blocks of test questions (by their forms).

As tests contained mostly the 8th-grade material, testing in the 8th-grades was carried out at the end of the school year (in one day), in the 9th-grades – at the beginning of the school year (in one day). Examinees were instructed to answer the questions of each form. The students took 45 minutes to complete the tests and the questionnaire in paper-and-pencil format.

According to M. Kastner and B. Stangl's study, MC and CR questions can be considered as interchangeable (which is an important prerequisite for comparing their results) if the Number Correct scoring rule is applied [8]. Thereafter, the proposed study obtained this rule, which allowed to take into account partial knowledge and not to reduce scores for incorrect answers.

3.4 Data Analysis

The test results were processed using statistical methods, such as variance analysis, factor analysis, etc. in MS Excel and STATISTICA software package.

According to the results of testing, compliance with the law of the normal distribution of both total and partial results is determined (for each questions block and in each sample). The characteristics of the comparative samples of the result indicators – numerical values of the test results of 8th and 9th-grades students – coincide with the level of 0.05 (according to the Wilcoxon-Mann-Whitney criterion (empirical value 0.8896, critical 1.96). Comparison of test results by 8th-grade students (N_1=77; M_1=7.32; SD_1=16.25) and 9th-grade students (N_2=87; M_2=7.80; SD_2=12.76) showed weak statistical power (0.056); a measure of effect d=0.12665 (that is weak) [23]. To check the statistical hypothesis of the equality of two samples, the Student's T-test criterion was applied, which showed that t (1,96) =.827. Accordingly, obtained statistics indicators confirmed the correctness of the hypothesis on the equality of the average general totality, which allowed to combine test scores, done by students of 8th and 9th-grades into a single sample (N=164; M=14.04; SD=0.65).

According to the results of the single sample, a multivariate dispersion analysis of the studied factors influence on them was carried out. The accuracy of calculations in each case exceeded 99% (p<0.01). The power of impact factors was determined by the method of J. Snedecor [24].

According to the results of processing verbal-numerical assessments on the SD questionnaire (analysis of averages, factor analysis), were identified dominant features (markers), expressing the generalized perceptual-emotional students' attitude to the proposed questions forms, and the profound factors which characterize the perception of different questions by the respondents. The results of the survey were compared with the OI of testing (test scores) in qualitative terms, given the different measurement systems and tools [25].

4 Main results of the research

The form of the test questions in the study was considered as an impact factor the output test indicators. The results of the variance analysis confirmed the hypothesis about the presence of this factor impact, whose power is up to 35%.

Comparative analysis of the test results of 8th and 9th-grades students on 1st and 2nd variants showed that the impact power of considered factor on OI varies depending on the texture of the test – for the 1st variant it is on average 45.5%, for 2nd – 27.8%. Besides, the tendency of a lesser expression of the impact power of the questions forms on OI in the second variant is observed in both 8th and 9th-grade students. Comparison of test results of students' groups by the profile of classes (CM, PM, Econ.) and educational level (8th and 9th-grades) revealed tendentious proximity of impact power indicators on OI factor "test questions form". This leads to the conclusion about the dependence of impact of profile studying on the test results from the subject field of assessment – in the subject field of Ukrainian language, indicators of the impact power of this factor on the results of different samples (classes of different profiles and educational levels) are tendency similar, which is explained by the presence of constant activity support in mastering and using the state language in the process of studying and everyday life.

At the same time, the analysis revealed that the impact power of the factor under consideration changes according to the variations of combining different questions forms in one test. It is determined that the highest diagnostic reliability, which (in addition to compliance with psychometric quality criteria – reliability, validity, etc.) defined by the smallest expression of the impact power on the test results (up to 10%), characterizes test, which combines the following forms of questions: (1) MC_{text} and $MC_{single+multiple}$; (2) MC_{text} and CR; (3) $MC_{single+multiple}$ and CR questions (Table 1).

Table 1. 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			

As can be seen from the table, the most reliable and internally consistent combination of questions (*α* Cronbach's = .6849), which at the same time has a low level of impact on the test results (Impact power = 8.79%) is MC_{text} + CR questions.

An analysis of the cognitive complexity of the parallel questions offered to students in different forms [14] revealed the presence of an operational-activity component of complexity, which is determined by the number and nature of operations that need to be done in the process of completing a question. For example, compare a questions: to fill in gaps and an explanation questions, questions with the choice of one correct answer to the question that test *knowledge* of a certain theoretical construct, and questions that involve text reading or performing certain actions to select one correct answer and test *knowledge and application* of theoretical construct). Thus, in the proposed tests, the matching questions, which by cognitive complexity are identical to the parallel questions (with the choice of one and more than one correct answers), are characterized by cyclic repetition of operations and less variability of "extra" descriptors choice (than in the questions with the choice of right answer) that increases the probability of "guess" and decreases the diagnostic accuracy of the results. This thesis is supported by the analysis of the overall test results comparison (the level of questions' completion in each case is higher than the level of other blocks parallel questions implementation) with the results of the impact analysis of combining different questions forms in the test: in each of the test variants, which reduce the diagnostic accuracy of test results (Table 1), there are matching questions that are combined with other closed questions.

5 Conclusions, discussions and prospects of further researches

Generalization of the study results allows us to conclude about the presence of impact factor "questions form" on the test results (up to 35%). The variability of its power indicators shows the probability of a change in test result indicators according to the features of the external test organization. Variation of impact power indicators on OI factor "questions form" in the first and second tests' variants while maintaining the results of different samples of the general tendency of performing the comparable questions forms proves the existence of impact factor "linguistic design of questions": parallel in content and form, questions in two variants of the used in the study test were aimed at checking the assimilation of the same element of the educational material, but by means of different language-operating material (usage of different tokens, phrases and sentences to apply theoretical knowledge).

Presented results also lead to the assumption that CR questions should be included into the Ukrainian language test, which can increase the diagnostic accuracy of the test results, provided a certain ratio of CR and MC questions number. This assumption is confirmed by the consideration of an affective factor, which was analyzed by students' verbal-numerical evaluation of the proposed questions forms (the analysis is presented in a separate article [25]). According to the students' survey results, it is found out that the same questions are perceived by students in different ways depending on its form. Comparison of the affective factor analysis results with the testing results makes it possible to conclude that the form of the question is a sensory-afferent stimulus of the operative way of activity, which influences students' interest and motivation formation to perform the proposed questions. Considering the peculiarities of criterion-oriented assessment of students' level of mastery identified in the test components of educational material (irrespective of the group indicators level with subsequent results' ranking, as in the case of regulatory-oriented testing), this conclusion is significant in terms of test structuring and proves the feasibility of combining different forms and different cognitive complexity different questions in one test, and both for the formation of positive attitudes and motivational optima in the process of their implementation by students, and for testing and refining processes of test results.

Therefore, conclusions on the results of presented study in the subject field of linguistic testing confirm scientists' opinion who have studied external features of tests in other subject areas regarding the existence of effect of questions forms on the testing results (e.g., [26]), as well as different diagnosis power of MC and CR questions ([13], [3]) and expediency of constructing polymorphic tests (e.g., [18]) as a multidimensional toolkit for assessing knowledge of students with different abilities and cognitive models of cognition.

However, the dependence of OI on the test structure which was determined in the process of the proposed scientific study makes it necessary to take it into account in further scientific researches, aimed at deepening and explaining the presented results, which are preliminary and need clarification. Thus, the test used in the experiment, as noted earlier, consisted of 4 blocks that contained the same number of different questions forms, aimed at testing students' learning of the same educational constructs, but these questions were characterized by different degrees of cognitive and operating load, which made it impossible to evaluate dichotomous test performance and, consequently, made it difficult to compare different questions forms. Thus, we consider it desirable to carry out refinement studies on a larger sample: (1) using stem-equivalent individual tests, constructed from the questions of each of four investigated forms, which will give the opportunity to increase the number of comparable questions and increase results reliability (in the present study, questions of various forms were contained in one test that significantly limited the number of compared questions in each block of the test), as well as to compare the test results obtained by following different scoring rules ([27], [8]); (2) using tests, constructed from questions forms, the combination of which in one test, according to our study, has the least effect on test results (MC_{text} + CR). Conclusions about the impact of the test structure on the test results determine the following areas of further research: determination of the most appropriate ratio of different questions forms in the structure of the polymorphic test, the study of the problem of questions language-operational material as an impact factor of their understanding and evaluating in various subject areas.

The proposed study was performed for state funds in the framework of the investigation "Methodology of state final certification in the form of external independent evaluation of high school graduates" of the department of monitoring and quality assessment of general secondary education of the Institute of Pedagogy of the National Academy of Pedagogical Sciences of Ukraine.

References

1. I.P. Karpova, *Sravnenie otkrytyh i vyborochnyh testov* (Comparison of open and sample tests). Otkryitoe obrazovanie 3, 32–38 (2010)
2. A. Orlov, T. Ponomareva, O. Chukajev, S. Pazuhina, Tekhnologii otsenivaniya rezultatov obrazovatelnogo protsessa v vuze v kontekste kompetentnostnogo podkhoda (*Technologies for*

assessing the results of the educational process in a university in the context of a competency-based approach), 2nd edn. (Direkt-Media, Moscow, Berlin, 2017).

3. Ch.A. Melovitz Vasan, D.O. DeFouw, B.K. Holland, N.S. Vasan, Analysis of testing with multiple choice versus open - ended questions: Outcome - based observations in an anatomy course. Anat Sci Educ, **11**(3), 254–261 (2018)
4. D. Marengo, R. Miceli, M. Settanni, 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 (2016)
5. F. Mozaffari, S. Mohammad Alavi, A. Rezaee, Investigating the Impact of Response Format on the Performance of Grammar Tests: Selected and Constructed. Journal of Teaching Language Skills **36**(2)*,* 103–128 (2017)
6. S. Schladitz, J. Groß Ophoff, M. Wirtz, Effects of different response formats in measuring educational research literacy. Journal for educational research online **9**(2), 137–155 (2017)
7. R. W. Lissitz, Xiaodong Hou, Sh. Slater, The Contribution of Constructed Response Items to Large Scale Assessment: Measuring and Understanding their Impact. Journal of Applied Testing Technology **13** (3) (2012), http://www.jattjournal.com/index.php/atp/article/view/48366. Accessed 21 Jul 2019.
8. M. Kastner, B. Stangl, Multiple-Choice and Constructed Response Tests: Do Test Format and Scoring Matter? Procedia – Social and Behavioral Sciences **12**, 263–273 (2011)
9. L.F. Bachman, A.S. Palmer, *Language Testing in Practice: Designing and Developing Useful Language Tests*, 1st edn (Oxford University Press, New York, 1996)
10. V. Avanesov, *Soderzhaniye testa i testovykh zadaniy* (The content of the test and test items). (2007), http://testolog.narod.ru/Theory61.html. Accessed 21 June 2019
11. V. Avanesov, *Problemy kachestva pedagogicheskikh izmereniy* (The problem of the quality of pedagogical measurements) (2008), http://testolog.narod.ru/EdMeasmt2.html#_ftn23. Accessed 6 July 2019
12. K. Djatlova, T. Mikhaleva, Issledovaniye vliyaniya raznoobraziya form testovykh zadaniy na statisticheskiye kharakteristiki testov (Investigation of the influence of the forms variety of test tasks on the statistical characteristics of the test). Voprosyi testirovaniya v obrazovanii 4, 65– 75 (2006)
13. M.E. Martinez, Cognition and the question of test item format. Educational Psychologist **34** (4), 207–218 (1999)
14. G.R. Hancock, Cognitive Complexity and the Comparability of Multiple-Choice and Constructed-Response Test Formats. The Journal of Experimental Education **62/2,** 143–157 (1994)
15. J.C. Alderson, *Assessing Reading* (Cambridge University Press, UK, Cambridge, 2000)
16. B. Clay, *Is This a Trick Question? A Short Guide to Writing Effective Test Question* (Kansas Curriculum Center, 2001).
17. T. Hudson, *Teaching Second Language Reading* (Oxford University Press, Oxford, 2007)
18. A. Kan, O. Bulut, D.C. Cormier, The Impact of Item Stem Format on the Dimensional Structure of Mathematics Assessments. Educational Assessment **24**(1), 13–32 (2019)
19. H. Wainer, D. Thissen, Combining Multiple-Choice and Constructed-Response Test Scores: Toward a Marxist Theory of Test Construction. Applied Measurement in Education **6**(2), 103–118 (1993)
20. O.A. Reshetnikova, Chto okazyvayet vliyaniye na izmeneniye kontrolnykh izmeritelnykh materialov dlya gosudarstvennoy itogovoy attestatsii? (What influences the change of control measuring materials for state final certification?). Pedagogicheskie izmerenija 2, 5–9 (2016)
21. C. Hoyt, Test reliability estimated by analysis of variance. Psychometrika **6**, 153–160 (1941)
22. Ch.E. Osgood, G. J. Suci, P. Tannenbaum, *The Measurement of Meaning* (University of Illinois Press, Champaign, 1957)
23. J. Cohen, *Statistical power analysis for the behavioral sciences,* 2nd edn. (Lawrence Erlbaum Associates, Hillsdale, 1988)
24. G.W. Snedecor, *Statistical Methods Applied to Experiments in Agriculture and Biology* (Collegiate Press, Ames, 1937)
25. A.V. Hryvko, Yu.O. Zhuk, Vykorystannia zasobiv IKT u protsesi eksperymentalnoho doslidzhennia emotyvno-otsinnoho stavlennia uchniv do riznykh form testovykh zavdan z ukrainskoi movy (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 **70**(2), 285–297 (2019). doi:10.33407/itlt.v70i2.2621
26. C. Jonick, J. Schneider, D. Boylan, The effect of accounting question response formats on student performance. Accounting Education **26**(4), 291–315 (2017)
27. E. Lesage, M. Valcke, E. Sabbe, Scoring methods for multiple choice assessment in higher education – Is it still a matter of number right scoring or negative marking? Studies in Educational Evaluation **39***,* 188–193 (2013)