DOI:10.20535/2410-8286.225018

COMPUTER ADAPTIVE LANGUAGE TESTING ACCORDING TO NATO STANAG 6001 REQUIREMENTS

Piotr Gawliczek[1], Viktoriia Krykun[2], Nataliya Tarasenko[3], Maksym Tyshchenko[4], Oleksandr Shapran[5]

[1]*University of Warmia and Mazury in Olsztyn, Poland*

[2,3,4,5]*National Defence University of Ukraine named after Ivan Cherniakhovskyi, Kyiv, Ukraine*

pgawliczek@gmail.com

The article deals with the innovative, cutting age solution within the language testing realm, namely computer adaptive language testing (CALT) in accordance with the NATO Standardization Agreement 6001 (NATO STANAG 6001) requirements for further implementation in foreign language training of personnel of the Armed Forces of Ukraine (AF of Ukraine) in order to increase the quality of foreign language testing. The research provides the CALT method developed according to NATO STANAG 6001 requirements and the CALT algorithm that contains three blocks: "Starting point", "Item selection algorithm", "Scoring algorithm" and "Termination criterion". The CALT algorithm has an adaptive ability, changing a complexity level, sequence and the number of items according to the answers of a test taker. The comparative analysis of the results of the CALT method piloting and the paper-and-pencil testing (PPT) in reading and listening according to the NATO STANAG 6001 requirements justifies the effectiveness of the three-level CALT method. It allows us to determine the following important benefits of CALT: test length reduction, control of measurement accuracy, objective assessment, improved test security, generation of a unique set of items, adaptive ability of the CALT algorithm, high motivation of the test takers, immediate score reporting and test results management. CALT is a qualitative and effective tool to determine test takers' foreign language proficiency level in accordance with NATO STANAG 6001 requirements within the NATO Defence Educational Enhancement Programme. CALT acquires a special value and relevance in the context of the global COVID 19 pandemic.

Keywords: CALT method; CALT algorithm; foreign language training; military personnel; NATO STANAG 6001.

Introduction

Nowadays the language training system of military personnel within the Department of Defence of Ukraine is supposed to provide foreign language training compatible with the armed forces of NATO countries. Such tendency caused the increase in the number of language test sessions, which in turn requires increasing the efficiency of the testing procedure. Computerized adaptive language testing (CALT) is considered to be one of the effective means to resolve this task.

It is important to notice that NATO developed a special language standard (NATO STANAG 6001, 2016) for testing a language proficiency level of personnel of the armed forces of NATO member-countries and the Armed Forces of Ukraine (AF of Ukraine) should follow their requirements in the testing to define their levels of foreign language proficiency. All higher military educational institutions have sufficient material and technical resources for a full-fledged implementation of CALT into the language training process, but AF of Ukraine don't have an effective CALT method that will meet the NATO STANAG 6001 requirements (NATO STANAG 6001, 2016).

Topicality of the study is defined by the necessity of a search of the effective ways to improve the language training of personnel of the AF of Ukraine and by the need to create the CALT method. The experience of NATO partner countries is an important aspect in the development of the CALT method. The cooperation of Ukraine and NATO partner countries within the Defence Education Enhancement Programme (NATO DEEP) deserves a special attention. Considering the NATO DEEP Ukraine programme as the demand driven activity, it is important to meet respective suggestions related to the changing situation. The Advanced Distributed Learning tools are playing the essential role, as far as the education and training process of the AF of Ukraine is concerned.

One of the demands formulated during the NATO DEEP meeting in 2019 was devoted to the intention of Ukraine to implement the CALT method. The CALT development and implementation process was perceived as the solution focusing on: practical recommendations concerning the development of the CALT method; administrative issues related to CALT. All the above mentioned and the experience of international partners of Ukraine (the representatives of Bulgaria and Poland) within the framework of NATO DEEP

indicates and reinforces the need to develop an effective CALT method in accordance with NATO STANAG 6001 requirements.

Literature review

Computerized adaptive testing (CAT) comes from decades of research and academic publications (Weiss & Kingsbury, 1984; Vispoel, Rocklin, & Wang, 1994; Sands, Waters, & McBride, 1997; Weiss, 2004; Thompson & Weiss, 2011; Thompson, 2016; Wang et al., 2019; Yigit et al., 2019; Albano et al., 2019; Chen et al., 2020). CAT is "a sophisticated method of delivering examinations, and has nearly 40 years of technical research" (Thomson & Weiss, 2011, p. 1). CAT is based on "a complex of algorithms which adapt the test to each test-taker while controlling the content distribution, item exposure, and test length. CATs have been shown to reduce test length by up to 90% without a loss of precision. At the same time, to achieve such a reduction, it is crucial that the CAT developer performs research studies to simulate the performance of CATs" (Thompson & Weiss, 2011).

In the 1960s, the Office of Naval Research sponsored the work on CAT. "Some of the nation's most eminent psychometricians such as Drs. Frederick Lord, Darrell Bock, David Weiss and others were involved in this effort" (Sands, Waters, & McBride, 1997, p. 9). In October 1996, the US Department of Defense became the first organisation to use CAT-derived scores for personnel selection. US Department of Defense has become the first employer to adapt CAT for its employment system (Sands, Waters, & McBride, 1997, p. 9). This revolutionized innovation made testing fairer and faster.

The International Association for Computerized Adaptive Testing (IACAT) invests significant efforts in the development of theory, techniques, technologies and instrumentation available for adaptive measurement in all relevant human, institutional, and social characteristics. IACAT is an organisation that is established exclusively for scientific, educational, literary, and charitable purposes. Adaptive testing is being developed in various scientific areas of testing (Gibbons & deGruy, 2019; Lin et al., 2019; Mâsse et al., 2020). Thus, CAT has become a particularly wide application for solving educational problems (Weiss & Kingsbury, 1984; Maia, Lilley & Barker, 2003; Babcock & Weiss, 2012; Oppl et al., 2017). The scientists (Blake, 2011; Hambleton,& Zaal; Newhouse & Cooper, 2013) study three main types of the computer adaptive testing: *in pyramidical* testing the test taker first gets the items of medium complexity, and then depending on his/her answers, easier or more complex items are provided; (Larkin & Weiss, 1975); *flexilevel* testing starts with the complexity level chosen by a test taker, and then depending on his/her answers, the next item is easier or more complex than the previous one; it goes that way until the test taker's knowledge level is defined (Lord, 1971); *stradaptive* testing is held with the help of the bank of items divided according to complexity levels. If a test taker gives the correct answer to the item, his/her next item is suggested at a higher level of complexity and vice versa (Weiss, 1973, p. 121).

Nowadays the testing as a method of a foreign language skills control is thoroughly studied by a wide range of scientists (Bachmann, 1990, 2000; Bachmann, Davidson & Milanovic, 1996; Canale & Swain, 1980; Larson, 1999; Fulcher, 1999, 2017; Holzknecht et al., 2021; Monfils & Manna, 2021). Among the modern development of CALT we can emphasise British tests, such as: Key English Test, Preliminary English Test, First Certificate in English, Certificate in Advanced English, Certificate of Proficiency in English, International English Language Testing System meant to define the foreign language proficiency of candidates for studying or employment in English-speaking countries. These tests are developed and based on the researches in language testing (Larson, 1999; Fulcher, 1999, 2017; Bachman, 2000; Chapelle & Voss, 2017; Wigglesworth & Frost, 2017; Mizumoto, Sasao & Webb, 2019; Monfils & Manna, 2021) and are supported by the programmes for scientific researches concerning language tests' validity.

Taking into consideration all the above mentioned, we may conclude that CALT is the means to define the level of foreign language competence of a test taker, where current items given to a test taker depend on the results of his/her previous answers. "However, CALT cannot test the multidimensional nature of language, and as such it fails to assess its communicative and functional nature" (Meunier, 1994, p. 23). CALT shows a diagnostic value in the definition of the level of skills maturity in receptive language skills (reading and listening), but the latest foreign researches consider the possibility of creating the Automated Writing Evaluation system (Hockly, 2019) where its advantages and disadvantages are elaborated.

The aim of this article is to justify the effectiveness of the CALT method according to NATO STANAG 6001 requirements and to present the CALT method piloting results. We hypothesise that the quality of a foreign language testing will be increased in case of implementation of the CALT method, which is developed according to NATO STANAG 6001 requirements.

Methods

Research Design

The CALT method piloting was carried out with the aim to confirm the validity of the developed theoretical provisions. The experiment consisted of two stages: development of the CALT method and its piloting. The results of PPT were then compared against those of CALT to address the research hypothesis. In order to achieve the aim of the research, a set of scientific research methods was used: theoretical − analysis, generalisation, comparison and systematisation of psychological and pedagogical, educational and methodological materials, international experience and NATO standards to clarify the state of the research issues; empirical − surveys, study and generalisation of the scientific experiment to obtain the results of the CALT method piloting; mathematical and statistical methods for processing the data obtained during the experiment.

Participants

The military students of foreign language courses, "Air Radio Communication of Aircrews" and "UN Military Observer" held at the National Defence University of Ukraine named after Ivan Cherniakhovskyi were involved in the piloting. 114 test-takers participated in the experiment. They took PPT according to NATO STANAG 6001 in two language skills (reading and listening) during the first phase of the CALT method piloting. These respondents were involved in the CALT session during the next phase no more than 10 days after the first phase to obtain valid and reliable results.

Instruments and Procedures

From an architectural perspective, CAT is composed of five components (Weiss & Kingsbury, 1984; Thompson, 2007). The first component is *a calibrated item bank*, and is therefore developed as a test content and the remaining four components (*starting point, item selection algorithm, scoring algorithm, termination criterion*) refer to algorithms in the CAT system (Weiss & Kingsbury, 1984; Thompson, 2007). The CALT method is based on the Item Response Theory (Hambleton, Swaminathan & Rogers, 1991) which is a psychometric paradigm designed specifically to address the well-known shortcomings of the Classical Test Theory (Frick, 1992) and taking into account the experience of leading Ukrainian scientists (Fedoruk, 2008; Hrabar, 2010; Kravchenko & Plakasova, 2010; Serhiienko, Malezhyk & Sitkar, 2012) and above mentioned foreign scientists in the field of CAT. Therefore, we offered the CALT method which includes calibrated items bank according to the defined complexity parameters and the algorithm that were developed with a purpose to improve the efficiency of the language testing according to the NATO STANAG 6001 requirements. The CALT method provides the assessment of a language proficiency in receptive language skills (reading and listening). Thus, we believe that the CALT method will increase the quality of foreign language testing of military officers of AF of Ukraine in accordance with the NATO STANAG 6001 requirements.

The CALT algorithm (Fig. 1) consists of three consecutive blocks. The *"Starting point"* block is responsible for generation of a random set of items of three complexity levels (Level 1, Level 2, Level 3) according to NATO STANAG 6001 requirements for fixing the start time and processing of other necessary settings to begin the testing. The CALT algorithm programs 45 items for each test taker, when it defines at least 15 items for each level. Language testing starts with presenting the Level 2 items.

The *"Item selection algorithm"* block presents the main procedures concerning logic of transition between different levels of complexity. The CALT algorithm is developed according to the principle of feedback, it means that the next task of a higher level of complexity is offered after the test taker's correct answer is given, and the wrong answer shows the necessity to offer the easier item. In such a way each test taker has an individual and unique set of items. The CALT adapts uniquely to each test taker in difficulty and the number of items.

The *"Scoring algorithm"* and *"Termination criterion"* are presented by the third block of the CALT algorithm the purpose of which is to determine the foreign language proficiency level based on the results of a matrix analysis. The block provides the analytical information on the progress of the testing process, the complexity of the items, and aggregates information on groups of test-takers, the actual time spent on a test in general and other necessary data for reporting. The number of correct answers to achieve the target level is 11, and the number of mistakes cannot be more than four at each level, which is determined by the requirements of a specification of the NATO STANAG 6001 language test. The CALT method allows to define the level of a test taker's foreign language competence formation by cutting down the number of items. The CALT algorithm also defines automatically the language proficiency level of a test taker and shows the test result immediately after finishing the testing session.

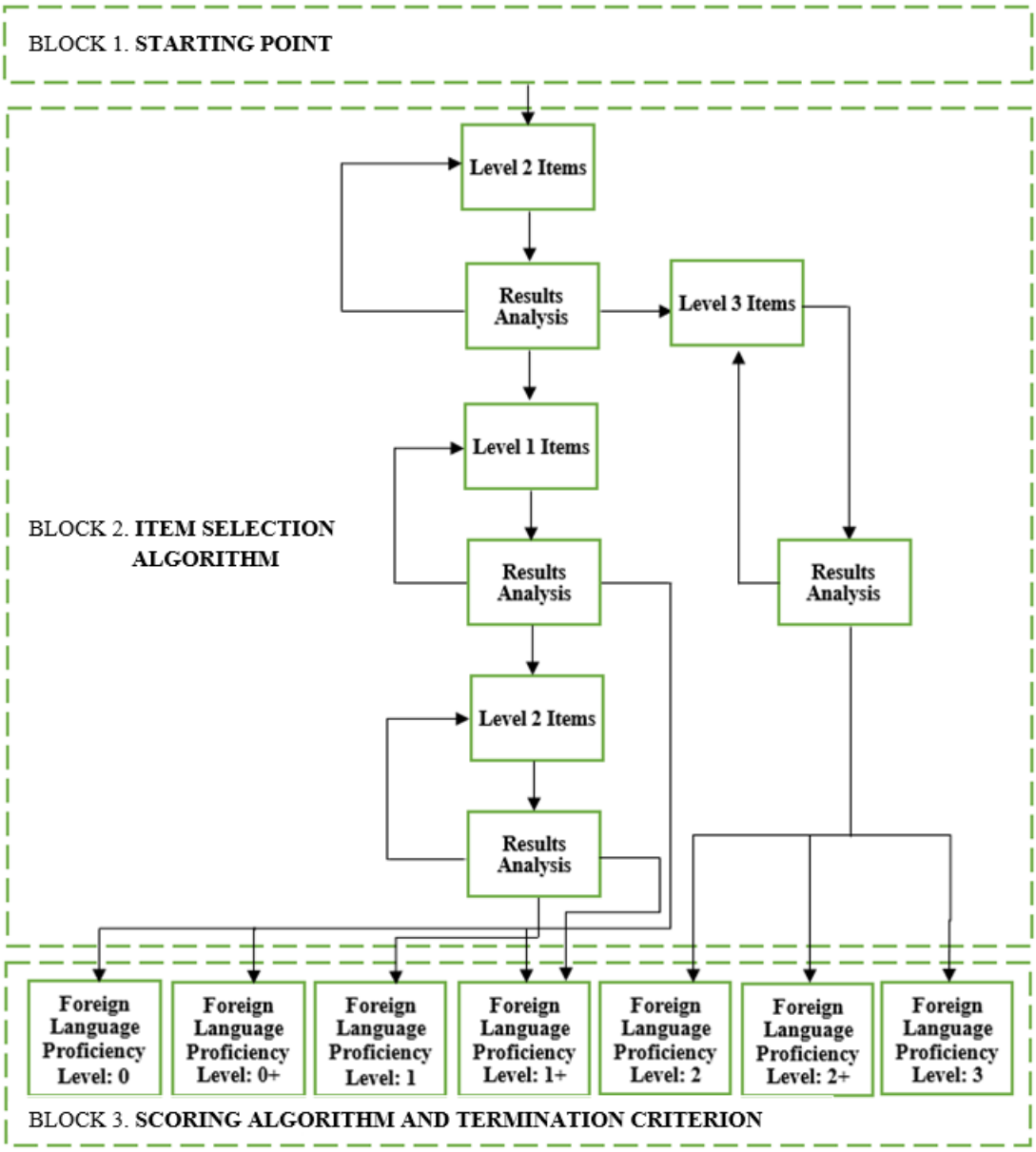

Figure 1. Structure of the CALT algorithm

Source: authors' development based on the research results conducted in November 2019 at the National Defence University of Ukraine named after Ivan Cherniakhovskyi within the International Scientific Project "Computer Adaptive Language Testing – CALT" within the framework of the NATO Defence Educational Enhancement Programme

Data Analysis

The data analysis of the experimental research is based on statistical and mathematical methods − for quantitative and qualitative analysis of the experimental results; graphical method − for their visualisation. After CALT and PPT test-takers were suggested a survey on a test-takers' readiness for CALT. The structure of the CALT algorithm, results of CALT and PPT and a survey were interpreted and represented graphically.

Ethical issues

The collection of data has been performed according to general standards of research ethics and previously approved by the Research Board of Language Testing and Research Centre and the Research Advanced Distributed Learning Centre of the National Defence University of Ukraine named after Ivan Cherniakhovskyi. The Research Board consisted of researchers from both centres who did not conduct classes either participated in the experimental research to keep the objectivity of the process and results. Test takers gave their consent for participation in the experiment. Before the CALT method piloting started, the test takers were informed about the circumstances of experimental research, the confidentiality of the observation data, their right to familiarise with the CALT results and were asked for their permission to publish the results of the experiment. It was noticed that the results do not deal with individual abilities assessment and the CALT data will not affect the academic performance.

Results

In order to justify the effectiveness of the CALT method we conducted piloting of CALT and provided the comparative analysis of results of CALT and PPT in receptive language skills (reading and listening) according to the NATO STANAG 6001 requirements (Figure 2).

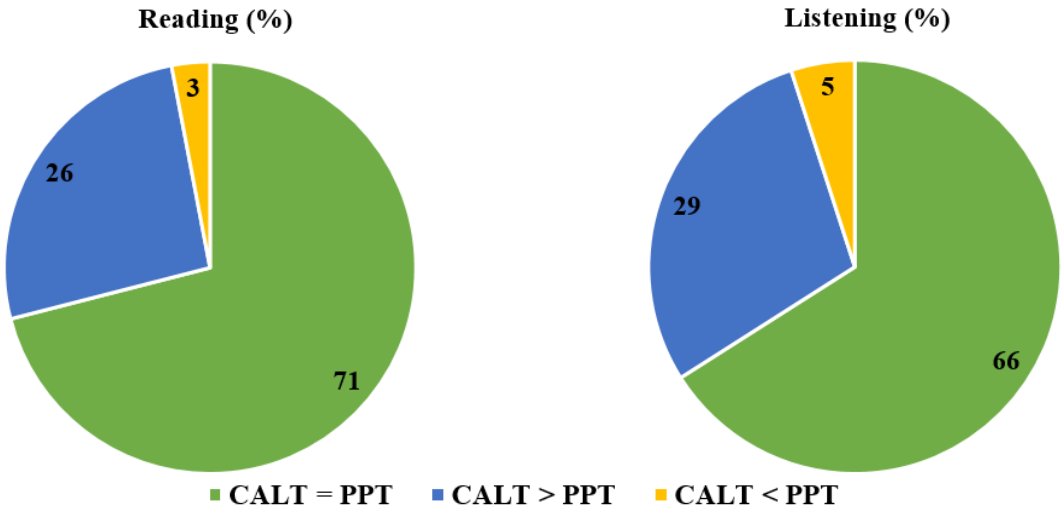

Figure 2. Comparative analysis of results of computer adaptive language testing (CALT) and paper-and-pencil testing (PPT) methods

Source: authors' development based on the experimental results conducted in December 2020 at the National Defence University of Ukraine named after Ivan Cherniakhovskyi within the International Scientific Project "Computer Adaptive Language Testing – CALT" within the framework of the NATO Defence Educational Enhancement Programme

The results of CALT show that a high percentage of the test takers have confirmed their language proficiency level in listening (66%) and reading (71%) skills. Thus, it indicates the validity and reliability of the CALT method. The results in listening comprehension test (29% of test-takers) and in reading comprehension test (26% of test-takers) demonstrate a discrepancy between these testing forms in favour of the CALT method. This discrepancy is caused by different levels of test takers' readiness for CALT (Padmavathi, 2016; Jamieson 2005; Beckers & Schmidt, 2003) which is confirmed by the results of the survey (Fig.3). The analysis of the survey results and its comparison with the CALT and PPT results shows that the number of respondents (31% of test-takers) had an experience of a computer testing which influenced the CALT results. And only 6% of participants were slightly confused during the CALT session because of their low level of the computer skills.

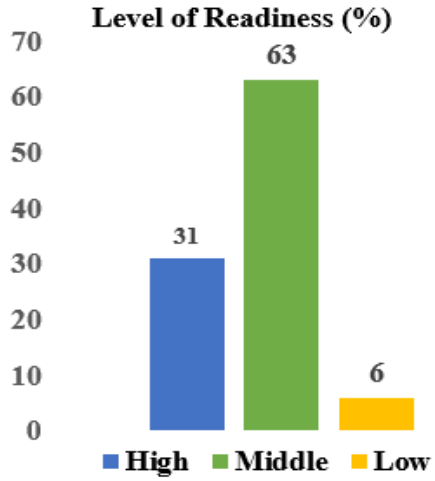

Fig.3 The survey results on test-takers' readiness for CALT

Source: authors' development based on a survey conducted in February 2020 at the National Defence University of Ukraine named after Ivan Cherniakhovskyi within the International Scientific Project "Computer Adaptive Language Testing – CALT" within the framework of the NATO Defence Educational Enhancement Programme

Discussion

The results of the present research highlight the potential of utilizing CAT in foreign language testing in accordance with the NATO STANAG 6001 requirements. We agree with Mizumoto, Sasao & Webb that "CAT can indeed measure test-takers' ability more efficiently than the paper-and-pencil counterpart" (Mizumoto, Sasao & Webb, 2019, p.119). Comparing the results of the experiment we can claim that the use of the CALT method provides more effective determination of the test takers' language proficiency level in reading and listening skills according to NATO STANAG 6001 than a PPT method. This can be explained by the benefits of CALT which are mentioned below. Besides the reliability, credibility and objectivity of obtained results, CALT can *reduce a test length* by 50 % or more (Weiss & Kingsbury, 1984). Due to the adaptive function of the CALT algorithm, the average time of testing is decreased by 22 % while PPT had time limits and fixed set of items. The process of a test results evaluation has been greatly accelerated due to the CALT ability to define automatically the language proficiency level according to the NATO STANAG 6001 requirements.

One of the benefits of CALT is *a control of measurement accuracy* (Mizumoto, Sasao & Webb, 2019). CALT can measure all test takers with the same degree of precision that means *an objective assessment* of the testing results. All test takers had the same level of accuracy in their scores, but they have an individual set of items. CALT allows to improve *a test security* due to the unique sets of items for each test taker. Thus, the *generation of a unique test* is another benefit of CALT.

We agree with Eggen (2018), who considers that "the using an item response theory-calibrated item bank, a CAT algorithm ensures that each test taker receives an optimal test. The algorithm selects items from the item bank tailored to the ability of the test taker, as determined from the test taker's responses during the testing" (Eggen, 2018). Thus, the main benefit of CALT is *an adaptive ability of the CALT algorithm* that is very flexible and adapts to test takers' competence during a testing session. We found out *a higher motivation of top test takers* because they did not waste their time on easy items, and at the same time we can state that the test takers of a low foreign language proficiency level were not discouraged by too difficult items. Other very important benefits of CALT are *an immediate score reporting* and *a test results management* (registration, collection, creation of item bank and database, reporting etc.).

Limitations

The test takers took PPT only in two language skills (reading and listening) in the first phase of piloting. This research does not provide a complete picture of the testing process when all four language skills are included (reading, listening, writing and speaking). The CALT test takers had an opportunity to take a three-level test (Level 1, Level 2, Level 3) while PPT provides the test items of only Level 1 and Level 2 according to the NATO STANAG 6001 requirements.

Conclusions

In order to justify the CALT method for its further implementation into the language training system of AF of Ukraine, a special CALT method was developed. This method required the development of a special CALT algorithm acting according to NATO STANAG 6001, taking into account the experience of the international partners of Ukraine and the analysis of the scientific studies on this issue. The CALT algorithm is characterised by a change of the complexity level, sequence and the number of items according to the answers of a test taker. This approach significantly differentiates CALT from a traditional PPT format when the number of correct answers is the main indicator of a test result, and CALT focuses on the test taker's foreign language proficiency level.

The use of the CALT method allows the following: to reduce a test length, time of taking CALT compared to PPT; to generate a unique test set for each test taker from the items bank according to the adaptive principle; to obtain reliable and objective results and to process them promptly; to provide an immediate reporting of test results; to improve a test security; to determine automatically a foreign language proficiency level according to the NATO STANAG 6001 requirements; to monitor the test taker's success in order to make adjustments to the process of foreign language learning. It is worth mentioning that the effectiveness of the CALT method is confirmed by the comparative analysis of the results of CALT and PPT. That fact proves our hypothesis that the quality of foreign language testing of the personnel of AF of Ukraine will be increased after the implementation of the CALT method developed according to NATO STANAG 6001 requirements.

CALT can be used during the military personnel language testing sessions and in the selection process of military personnel, employees of AF of Ukraine and other specialists of the defence and security sector of Ukraine for foreign language courses both in our country and abroad as well as for fulfilling military tasks in

the foreign environment. It should be noted that the theoretical issues concerning the development of CALT according to the NATO STANAG 6001 requirements are poorly covered by scientific studies. Moreover, the issue of the computer-aided design of adaptive language tests according to NATO STANAG 6001 requires extensive research, as there are no clear criteria for the automation of this process.

Acknowledgements. This research is carried out within the International Scientific Project "Computer Adaptive Language Testing – CALT" within the framework of the NATO Defence Educational Enhancement Programme – NATO DEEP (Brussels, Belgium).

References:

Albano, A. D., Cai, L., Lease, E. M., & McConnell, S. R. (2019). Computerized adaptive testing in early education: Exploring the impact of item position effects on ability estimation. *Journal of Educational Measurement, 56*(2), 437-451. https://doi.org/10.1111/jedm.12215

ATrainP-5. NATO STANAG 6001: Language Proficiency Levels. Edition A Version 2. (2016). *North Atlantic Treaty Organization, Standardization Office. Retrieved 12 June 2017.* Retrieved from www.natobilc.org

Babcock, B., & Weiss, D. J. (2012). Termination criteria in computerized adaptive tests: do variable-length CATs provide efficient and effective measurement? *J. Comput. Adap. Test, 1*, 1–18. https://doi.org/10.7333/1212-0101001

Bachman, L. (2000). Modern language testing at the turn of the century: assuring that what we count counts. *Language Testing, 17*, 1–42. https://doi.org/10.1191/026553200675041464

Bachman, L. F. (1990). *Fundamental Considerations in Language Testing*. Oxford, Oxford University Press. Retrieved from https://www.academia.edu/28794667/Fundamental_Considerations_in_Language_Testing

Bachman, L. F., Davidson, F. & Milanovich, M. (1996). The use of test methods in the content analysis and design of EFL proficiency tests. *Language Testing, 13*, 125–1

Beckers, J.J., Schmidt, H.G. (2003) Computer experience and computer anxiety. *Computers in Human Behavior, 19*, 6, 785-797, https://doi.org/10.1016/S0747-5632(03)00005-0

Blake, R. J. (2011). Current trends in online language learning. *Annual Review of Applied Linguistics, 31*, 19-35. https://doi.org/10.1017/S026719051100002X

Canale, M., & Swain, M. (1980) Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics, 1(1),* 1–47. https://doi.org/10.1093/applin/I.1.1

Chapelle, C., & Voss, E. (2017). Utilizing Technology in Language Assessment. *Language Testing and Assessment. Encyclopedia of Language and Education, 3rd ed.* (pp.149-161). https://doi.org/10.1007/978-3-319-02261-1_10

Chen, C. -., Wang, W. -., Chiu, M. M., & Ro, S. (2020). Item selection and exposure control methods for computerized adaptive testing with multidimensional ranking items. *Journal of Educational Measurement, 57*(2), 343-369. https://doi.org/10.1111/jedm.12252

Eggen, TJHM. (2018). Multi-Segment Computerized Adaptive Testing for Educational Testing Purposes. *Front. Educ, 3*, 111. https://doi.org/10.3389/feduc.2018.00111

Fedoruk, P. I. (2008). Adaptive tests: general provisions. *Mathematical machines and systems. 1*, 115−127. Retrieved from http://dspace.nbuv.gov.ua/handle/123456789/748

Frick, T. W. (1992). Computerized adaptive mastery tests as expert systems. *Journal of Educational Computing Research, 8(2),* 187−213. https://doi.org/10.2190/J87V-6VWP-52G7-L4XX

Fulcher, G. (1999). Computerizing an English language placement test. *ELT Journal, 53*, 289–299. https://doi.org/10.1093/elt/53.4.289

Fulcher, G. (2017). Criteria for Evaluating Language Quality. *Language Testing and Assessment. Encyclopedia of Language and Education, 3rd ed.* (pp.179–192). https://doi.org/10.1007/978-3-319-02261-1_13

Gibbons, R. D., & deGruy, F. V. (2019). Without wasting a word: Extreme improvements in efficiency and accuracy using computerized adaptive testing for mental health disorders (CAT-MH). *Current Psychiatry Reports, 21*(8) https://doi.org/10.1007/s11920-019-1053-9

Hambleton, R. K., & Zaal, J. N. (1991). Advances in educational and psychological testing: theory and applications. *Springer Sciences & Business Media, LLC*. https://doi.org/10.1007/978-94-009-2195-5

Hambleton, R.K., Swaminathan, H. & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park, CA, Sage Publications.

Hockly, N. (2019). Automated writing evaluation. *ELT Journal, 73(1)*, 82–88. https://doi.org/10.1093/elt/ccy044

Holzknecht, F., McCray, G., Eberharter, K., Kremmel, B., Zehentner, M., Spiby, R., & Dunlea, J. (2021). The effect of response order on candidate viewing behaviour and item difficulty in a multiple-choice listening test. *Language Testing, 38*(1), 41-61. https://doi.org/10.1177/0265532220917316

Hrabar, E. V. (2010). Basic types of tests in foreign languages in US pedagogy. *Scientific notes of Ternopil National Pedagogical University named after Volodymyr Hnatyuk. Series: Pedagogy, 1*, 194–201. Retrieved from http://nbuv.gov.ua/UJRN/NZTNPU_ped_2010_1_37

ICAT (International Association for Computerized Adaptive Testing). Retrieved from http://www.iacat.org/

Jamieson, J. (2005). Trends in computer-based second language assessment. *Annual Review of Applied Linguistics*. 25. 228-242. https://doi.org/10.1017/S0267190505000127.

Kravchenko, O. M. & Plakasova, Zh.M. (2010). Model of intellectual controlling subsystem with multilevel adaptive testing. *East European Journal of Advanced Technologies, 4/2 (46),* 21–25.

Larkin, K. C. & Weiss, D. J. (1975). *An empirical comparison of two-stage and pyramidal adaptive ability testing (Research Report, 75-1)*. Minneapolis: Psychometrics Methods Program, Department of Psychology, University of Minnesota. Retrieved from https://eric.ed.gov/?id=ED106317

Larson, J. (1999). Considerations for testing reading proficiency via computer-adaptive testing. In M. Chalhoub-Deville (Ed.), *Studies in language testing, Vol. 10. Issues in computer-adaptive testing of reading proficiency* (pp.71-90). Cambridge: University of Cambridge Press.

Lin, G.-H., Huang, Y.-J., Chou, Y.-T., Chiang, H.-Y., & Hsieh, C.-L. (2019). Computerized adaptive testing system of functional assessment of stroke. *Journal of Visualized Experiments, 2019*(143). https://doi.org/10.3791/58137

Lord, F. M. (1971). The self-scoring flexilevel test. *Journal of Educational Measurement, 8*, 147–151. https://doi.org/10.1111/j.1745-3984.1971.tb00918.x

Maia, M., Lilley, M. & Barker, T. (2003). Computer-Adaptive Testing in Higher Education: the way forward? *XXXVIII Cladea - Latin American Council of Schools of Administration: Lima.* https://doi.org/10.13140/2.1.2074.7520

Mâsse, L. C., O'Connor, T. M., Lin, Y., Hughes, S. O., Tugault-Lafleur, C. N., Baranowski, T., & Beauchamp, M. R. (2020). Calibration of the food parenting practice (FPP) item bank: Tools for improving the measurement of food parenting practices of parents of 5–12-year-old children. *International Journal of Behavioral Nutrition and Physical Activity, 17*(1). https://doi.org/10.1186/s12966-020-01049-9

Meunier, L. (1994). Computer Adaptive Language Tests (CALT) Offer a Great Potential for Functional Testing. Yet, Why Don't They? *CALICO Journal, 11*(4), 23-39. Retrieved February 25, 2021, from http://www.jstor.org/stable/24152755

Mizumoto, A., Sasao, Y., & Webb, S. A. (2019). Developing and evaluating a computerized adaptive testing version of the word part levels test. *Language Testing, 36*(1), 101-123. https://doi.org/10.1177/0265532217725776

Monfils, L. F., & Manna, V. F. (2021). Time to achieving a designated criterion score level: A survival analysis study of test taker performance on the TOEFL iBT® test. *Language Testing, 38*(1), 154-176. doi:10.1177/0265532220940709

Newhouse, C. P., & Cooper, M. (2013). Computer-based oral exams in Italian language studies. *ReCALL, 25(3)*, 321-339. https://doi.org/10.1017/S0958344013000141

Oppl, S., Reisinger, F., Eckmaier, A Helm, C. (2017). A flexible online platform for computerized adaptive testing. *Int J Educ Technol High Educ, 14*(2). https://doi.org/10.1186/s41239-017-0039-0

Padmavathi, M. (2016) A study of student-teachers' readiness to use computers in teaching: an empirical study. I-manager's Journal on School Educational Technology, 11(3). Retrieved from https://pdfs.semanticscholar.org/ce38/ddfc90c28af40dfddd9a741fe703a632565f.pdf?_ga=2.127548237.1317622877.1612358315-1008902291.1612358315

Sands, W.A., Waters, B.K. & McBride, J.R. (1997). *Computerized adaptive testing. From inquiry to operation*. Washington, American Psychological Association. Retrieved from https://pdfs.semanticscholar.org/4e9c/c706ea17628f970389a25b2d268b52320e13.pdf

Serhiienko, V.P., Malezhyk, M.P., & Sitkar T.V. (2012). *Computer technologies in testing: a textbook.* Lutsk: Printing house "Volyn Polygraph". Retrieved from https://www.coursehero.com/file/64399499/KTTpdf/

Thompson, N. A. & Weiss, D. A. (2011). A Framework for the Development of Computerized Adaptive Tests. *Practical Assessment, Research, and Evaluation, 16*, Article 1. https://doi.org/10.7275/wqzt-9427

Thompson, N. A. (2007). A practitioner's guide for variable-length computerized classification testing. *Practical Assessment Research & Evaluation, 12 (1)*. https://doi.org/10.7275/fq3r-zz60

Thompson, N.A. (2016). *User's manual for SIFT: Software for investigating test fraud.* Minneapolis, Assessment Systems Corporation.

Vispoel, W.P., Rocklin, T.R., & Wang, T. (1994). Individual differences and test administration procedures. A comparison of fixed-item, computerized adaptive, and self-adapted testing. *Applied Measurement in Education, 7*, 53–59.

Wang, C., Weiss, D. J., & Shang, Z. (2019). Variable-length stopping rules for multidimensional computerized adaptive testing. *Psychometrika, 84*(3), 749-771. https://doi.org/10.1007/s11336-018-9644-7

Weiss, D. J. & Kingsbury, G. G. (1984). Application of computerized adaptive testing to educational problems. *Journal of Educational Measurement, 21*, 361–375. https://doi.org/10.1111/j.1745-3984.1984.tb01040.x

Weiss, D. J. (1973). The stratified adaptive computerized ability test (Research Report 73-3). *Minneapolis: University of Minnesota, Department of Psychology.* Retrieved from http://iacat.org/sites/default/files/biblio/we73-3.pdf

Weiss, D. J. (2004). Computerized adaptive testing for effective and efficient measurement in counseling and education. *Measurement and Evaluation in Counseling and Development, 37* (2), 70–84. https://doi.org/10.1080/07481756.2004.11909751

Wigglesworth, G. & Frost, K. (2017). Task and Performance-Based Assessment. *Language Testing and Assessment. Encyclopedia of Language and Education, 3rd ed.*, 121-133. https://doi.org/10.1007/978-3-319-02261-1_8

Yigit, H. D., Sorrel, M. A., & de la Torre, J. (2019). Computerized adaptive testing for cognitively based multiple-choice data. *Applied Psychological Measurement, 43*(5), 388-401. https://doi.org/10.1177/0146621618798665

Received: February 18, 2021
Accepted: March 31, 2021