Simulation Model for Computerized Testing of Learning Success in Quality Management Systems

Alexander N. Alexeyev [1,*], Nataliya A. Konovalova [1], Kateryna A. Lozova [1], Elena N. Korol [2]

[1] Sumy State University, Faculty of Technical Systems and Energy Efficient Technologies, Sumy, Ukraine
[2] Sumy State Pedagogical University, the Department of Preschool and Primary Education, Sumy, Ukraine

alekseev.aleksandr.nik@gmail.com, konovalova.nataliya11@yandex.ua, katarina_lozovaya@ex.ua, korol.9@mail.ru

Abstract. A computerized test successfully complements and enhances traditional methods for the assessment of knowledge. This article introduces a simulation model of computerized testing of learning success that is complementary to the existing methods used for knowledge evaluation. The simulation model combines possibilities of computerized testing with mathematical rationale in examiner's decision-making during oral knowledge assessment. Application of the simulation model enables one to make mathematically precise decisions in the majority of standard procedures of test development, during computerized testing, and in the analysis of its results. The new features of computerized testing, introduced here within the framework of the simulation model, help diminish the limitations of computerized testing that arise from the impossibility of utilizing diagnostic potential of a human examiner in traditional testing procedures.

Keywords. simulation model, educational measurement, computerized testing of learning success, key stages of computerized tests, test tasks.

Key Terms. ICTTool, QualityAssuranceProcess, Teaching Methodology, Teaching Process, Technology.

1 Introduction

Fast and accurate evaluation of knowledge formation remains to be a relevant task for long-existing forms of learning. Moreover, it has become increasingly important for the comparatively recently emerged distance learning or blended learning (partial implementation of distance learning technologies into classes that are conducted traditionally). The most important characteristics of different forms of learning remains

[*] Professor of Department of Manufacturing Engineering, Machines and Tools, Doctor of Pedagogy., Associate Professor

ICTERI 2016, Kyiv, Ukraine, June 21-24, 2016

the objective monitoring of students' academic achievements and the construction of effective teaching methods based on that.

The development of theory and practice of computerized tests makes it possible to increase the precision of pedagogical measurements as tools for objective knowledge control. Computerized testing carries out a number of pedagogical functions assigned to tests, hence becoming an effective means for evaluating the results of learning at all stages of education, from an entrance test to a comprehensive final exam.

This article describes a simulation model of computerized tests that on the one hand draws upon modern information and communication technologies and on the other is maximally reliant on the traditions of active participation of an instructor in students' knowledge assessment. Combination of the advantages of computerized testing with mathematical grounding of examiner's decision-making expands the range of effective applications of test-based knowledge assessment. The authors hope that the use of the simulation model developed by them will contribute to the further development of quality management systems at the institutions of higher learning.

2 Antecedents of simulation model for computerized testing of learning success

The classical period of the development of computerized testing theory to a great extent expanded the field of rational usage of computerized tests. In studies by A. Birnbaum, H. Gulliksen, G. F. Kuder, F. M. Lord, M. Novick, G. Rasch and others attempts were made to create an objective tool for observations in the fields of psychology, sociology, pedagogy, and other behavioral sciences. C. Spearman, one of the founders of the classical approach in the testing theory, proposed using methods of physical measurements in psychology. In pedagogy, this approach is called educational measurement. Increase in reliability of results of educational measurement in this period is due to the introduction into the testing theory of certain provisions of mathematical statistics, as well as of the elements of correlation theory aimed to justify the reliability and validity of the tests.

The 1970s witnessed emergence of a new direction in the theory of educational measurement – one connected with the Item Response Theory (IRT). Scholars J. B. Bjorner, B. Gandek, R. K. Hambleton, H. J. Rogers, S. J. Sinclair, M. H. Stone, H. Swaminathan, J. E. Ware, B. D. Wright et al. significantly contributed to the development of this new direction. The mathematical logistic models proposed by G. Rasch and A. Birnbaum were used to construct tests, or educational measurements. The goal of such measurements was an obtainment of numerical equivalents that were identified with the estimates of the measured variable. The measured variable was associated with the level of academic achievement, which was allowed in a certain way to reflect the latent parameter of the test-takers – their level of preparation.

The modern development of theory and technologies of educational measurement happens as a continuation of approaches founded in the previous period. The progress in development of the new testing methods has been driven by the applied and theoretical research of such scholars as F. B. Baker, R. Draney, G. R. Engelhard,

G. G. Kingsbury, D. J. Weiss, and M. Wilson. One of the most dynamically developing directions today is the one related to the design of adaptive tests, where new test questions are chosen based on one's performance on the previous questions. As the information and telecommunication technologies improve, the computerized testing of learning success becomes more and more prevalent in the theory and practice of educational measurement.

At the same time, many researchers (F. M. Bernt, A. C. Bugbee, D. C. Buhr, M. F. Johnson, S. M. Legg, K. C. Moe, R. Sutton and others) note the salient disadvantages of computerized tests that have not been resolved to date. Their findings, the results of our studies, suggest that testing designed based on most of the modern techniques still remains biased. Therefore, if no action is taken, the substitution of oral control with computerized control of learning success would not increase the reliability of educational measurement. Moreover, the exclusion of teachers from the monitoring process does not allow using the invaluable diagnostic capabilities of an instructor.

Nobody but an instructor, through conversation and additional probing questions, can determine whether a student's seemingly expressionless answer means the absence of knowledge on the subject or his or her mere nervousness. The instructor also has more opportunities to formulate questions not only by taking into account the student's responses to previous questions, but also depending on the content of the tested study material. For courses that require unconventional thinking and experiential approach, it is often difficult to create adequate and easily conveyed test questions. Hence such test questions frequently present difficulties for students. On the other hand, the fact that test design is still largely a subjective process also remains to be a problem. At the time of test creation, it is up to each of the test makers to decide upon the requirements for the number and complexity of tests to be included in a given assessment. Obviously, students with the same level of preparation are likely to score differently in such case, with students that had more simple test questions receiving higher grades than those whose test questions were more complex.

The objectivity of the results of computerized tests is also vulnerable to the inconsistency in the definition of evaluation criteria. It is certainly possible to introduce uniform requirements to testing. However, these might still be the same only for a given group of students, whereas in another group of students, or when tested by another instructor, a simple change in the grading criteria might change test results dramatically.

Therefore, with a steady ever-increasing usage of testing in knowledge assessment, there is a pressing need to create a model of computerized control of learning success that would utilize all the advantages of the testing method and would also maximally draw on the experience of active participation of instructor in diagnosing students' learning success, gained in the course of traditional knowledge assessment process.

3 Key stages of control

To solve this problem, the author's team has developed a simulation model of computerized control of learning success, which combines technological capabilities of computer-based testing with mathematical justification used in an instructor's decision -making procedures. In this diagnosis, the identity of the examiner is replaced, as much as it is possible, with his or her mathematical model.

The figure below shows a diagram of a multi-level computerized test, which has advanced measurement capabilities. Similarly to other approaches to the organization of testing procedure, the control is comprised of three phases: test design, test administration, and analysis of test results. The test design and analysis of test results phases rely on well-known theoretical positions, grounded in wide usage of statistical methods to increase of accuracy and objectivity of testing. In the test administration phase, mathematical methods that model diagnostic functions of an instructor are used to increase the reliability of results of educational measurement.

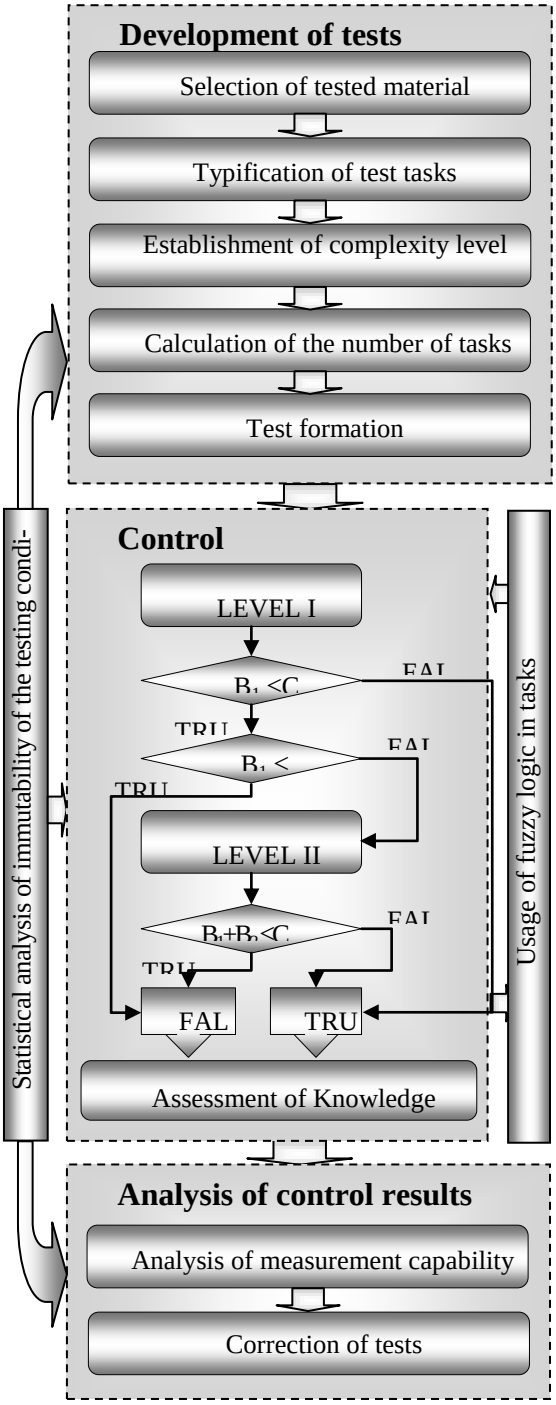

Fig. 1. Simulation model technological scheme

4 Test design

According to the scheme provided here, test design starts with the selection of test material. In this simulation model, it is supposed that this part of test design - similarly to many other testing methods - is done by experts that comprise a group of test makers. At the time of material selection, the experts are first-most guided by the ultimate goals of testing. In consideration of these goals, the experts decide upon the types of knowledge and skills that are most important for the goals set, as well as on the sufficient level of their demonstration by students.

After selecting the test content, test makers proceed to the design of test questions. The tested material is divided into separate parts, on which students can then be tested using sample test tasks. Provisions of the IMS Global Learning Consortium are placed at the basis of classification. These provisions are processed in such a way as to empower an instructor with more possibilities for formulation of test questions that would be maximally close to the content of the tested material. A total of 13 types of standardized test tasks are included into the proposed simulation model.

In addition to the recommendations of IMS, the simulation model contains special types of tasks that enable an instructor to check the extent to which the student's knowledge and skills have been formed. These include tasks on the control and sequence of actions. The design of a test task on control is a set of graphical images that reflect separate states of a certain object, and test takers are evaluated on their ability to manage it. The image shows targets, and the visible or invisible boundaries of these targets correspond to the contours of the object's organs of control. As one uses the pointer of a mouse to click on the required target, a graphical image of the object is substituted with a simulated control action. In a test task on the sequence of actions, object management happens with the mouse-click on control keys. In both types of test tasks, there is an option of setting an allowed interval of time between the mouse clicks.

The adjustment of the complexity of test tasks is possible through the procedures of design and corrective calculations, which are included into the simulation model. Expert assessment, which is accomplished using the method of paired comparisons, lies at the basis of design calculations (which are performed when prototypes of test tasks are created). Execution of such an expertise is most justified when a given test contains many tasks, and hence when it is difficult to preserve a single strategy and to have a comparable level of complexity for each of the tasks in the test. The corrective calculations procedure uses classical approach, which is based on the statistical processing of test results: expert grade estimates are refined taking into account students' performance on the test. It is assumed that the more students have answered a given test question incorrectly, the higher was its level of difficulty.

Once the level of test difficulty is determined, the test maker can move on to the next stage of test design: defining the necessary number of test questions in a given test. In the simulation model, the method of choosing a reasonable number of test questions is grounded in an assumption that it is important to account for both the quantity and the complexity of each task. Here, the total number of test questions is determined in such a way, that the cumulative complexity of one test would be com-

parable to that of another test. (For instance, in order to compare test results in physics and in chemistry, it is critical that the total complexity of tasks for the test in each of these subjects would be comparable.)

As a rule, test tasks have different levels of complexity. This is reflected in the assignment of unequal quantitative characteristics of test complexity measures. Since the tasks selected for a given test are chosen at random, while the method for calculating the number of tasks to be included into a test requires the tasks' cumulative complexity to remain constant, the authors recommend using genetic algorithms to design tests. In accordance with them, the process of test creation is seen as a successive change in the populations of species, whose genomes are random collections of test tasks of varying complexity. To generate different test versions (species of new populations) we apply operators of selection, crossover, mutation and survival. Such cyclical execution of operators is repeated until the total complexity of all tasks in a test does not reach optimal, i.e. as close as possible to the specified one.

5 Conducting iterative control measures

In the simulation model of computerized testing of learning success, the step during which the test is actually carried out is built on the basis of mathematical modeling of diagnostic capabilities of an examiner. Similarly to an oral testing procedure, in which an examiner can deem necessary to continue and ask a student additional questions which would help her determine the student's true level of knowledge, the simulation model provides for both basic and additional examination sessions. The procedure enables such a multilevel control via the employment of an apparatus of statistical analysis that resembles one used in engineering for the development of plans for the selective acceptance control.

Analogous to how the conclusion about the satisfactory quality of products that are manufactured in hundreds of thousands of pieces is made by means of an inspection of just a sample of them, the conclusion about the extent to which students' learning has been successful is evaluated by means of the statistical processing of the results of tests which have a limited number of questions. Comparison of the cumulative number of points received for the test with the values specified for the acceptance and rejection criteria makes it possible to make a final conclusion about the need to have an additional session of control.

If, upon completion of all tasks in the test, a student scores above the acceptance threshold, then his knowledge is evaluated as sufficient for a corresponding grade. Analogously, if a student knows the tested material worse than the rejection threshold, a conclusion is made that the student knows the tested material worse than the level of knowledge required for a given grade. However, when the number of points that a student receives for the test lies within the range of the pre-set acceptance and rejection values, the conclusion is made that it is impossible to determine the student's true level of knowledge and additional sessions of control are then carried out.

To expand the adaptive capabilities of a simulation model, the authors modernized the genetic algorithm for the selection of test tasks for additional test sessions. To

accommodate for such changes, a survival operator is altered and includes a criterion, which takes into account results of the preliminary test sessions. Here, the more poorly the student performs in questions on a certain topic in the previous sessions, the more likely is a question on this topic to show up in the additional testing session.

The testing methodology that is based on a simulation model stands out among most other existing methodologies in that, similarly to an oral exam, it enables a student to express their level of confidence in the correctness of a given answer in case the knowledge they possess does not allow them to give a definitive answer to the test question. Mathematical apparatus of fuzzy logic is used to make this functionality in the simulation model possible. A student that is being tested in the traditional way has to give a definitive answer to the test question by choosing one of several answer choices or by formulating their own answer choice using a limited set of words, letters, numbers, or graphical symbols. When giving an answer, a student has to formulate a response which would contain conclusions about the truthfulness of an expressed judgment using terminology of strict logic and hence has no way to express doubt or specify how far, in their opinion, the answer deviates from truth. Application of the fuzzy logic apparatus, on the other hand, allows a student to operate not only with the classical values of logical variables such as "false" and "truth", but also to use the interim values that smoothly transition from the one extreme value ("false") to the other extreme value ("truth"). This capability hence liberates a student from the necessity to make conjectures about an answer and go beyond their own knowledge on the topic. Such solution thus helps avoid introduction of an additional error into the results of computerized control of learning success.

In the computerized control of learning success nowadays, the prevalent methodology is one in which the resulting grade is assigned through a comparison of the total number of gained points with some linear and, less frequently, nonlinear scale of assessment. Grading scale in such an approach is typically set based on the probability of guessing the right answer or based on the expert assessments. However, both options are not the best ones for the creation of such a grading scale. In the first case, usage of such a scale would be justified if the probability of the randomly picked answer choice being correct does materialize: the student does not know the answer but happens to guess it correctly. Such a grading scale quickly becomes inaccurate if the probability of randomly selected correct answer does not materialize: the student actually knows the answer and hence responds correctly. In the latter case, the student’s knowledge of the subject is underestimated in such a grading scale. On the other hand, empirical grading scales are not universal. Here, expertise assessments should be carried out maximally often since the continuously changing conditions, in which the knowledge is being gained, to a large degree predetermine the students’ efforts at achieving a given level of knowledge. Therefore, the grading scale used in the simulation model is constructed based on the comparison of test results among students in the class. Similarly to oral testing, when an examiner that has to decide on a grade takes into account not only his assessment of the correctness and fullness of an answer but also other students' answers, the grading scale in the simulation model is based on the distribution of grades in the tested groups of students. To realize such

an approach this study adapted a method for building a five-point criteria scale for grading introduced by T.D. TenBrink.

Considering the fact that the change in the content of material covered in class or organizational and methodological supplements for it have a roughly the same effect on all the students, such method makes additional test sessions unnecessary for the conclusion of the assessment process in new conditions and the assignment of final grades. This is accomplished on the basis of the selective characteristics of the grade distribution parameters.

6 Analysis of test results

The mathematical rationale of the examiner's decision-making process mitigates significantly the disadvantages of computerized testing as of a tool for educational measurement. Additionally, the simulation model includes the stage for the analysis of test results, which rests on the traditional approaches. This stage includes procedures for evaluating measurement capabilities of individual test tasks and of the entire test using adapted for the use in simulation model indicators of distinctive capabilities and reliability. Furthermore, it is suggested to use the probability characteristics of impossibility of the extreme marks, as well as to use the specific for the simulation model criterion of abnormal amount of time spent on test completion. To identify the test items with an unsatisfactory measurement capability in the simulation model, the authors suggest using characteristics of impossible (more than 95%) probability of scoring only at the highest or only unsatisfactorily, and of impossible probability of abnormally spent time on completing the test. Distinctive capability of a test task is measured using the biserial correlation coefficient (discrimination index). The extent to which a test is reliable is characterized by the correlation of marks obtained for different parts of the test. (In the simulation model a change in the approach of dividing the test into parts was made: the selection of tasks is done at random, however in such a way that the total complexity of both parts of the test would be the same).

The level of knowledge and learning effectiveness are integrated indicators of many factors that influence the learning process. Students' performance on tests is dependent on the students themselves, on their instructors, on the methodological and organizational support of the learning process, as well as on other factors. Any changes made to the learning process, including changes to the procedures of knowledge control, can cause distortion to the statistical picture of test results. In the simulation model, most of the decisions rely on the statistical analysis of test results, and hence it is necessary to measure statistical significance of the changes that occurred in the course of the semester with a coefficient of reliability of statistical differences.

7 Conclusion

The simulation model for computerized testing of learning success makes it possible to make mathematically precise design solutions for the majority of standard procedures of test development, implementation and results analysis. The authors do not

deny the fact that any testing procedure, including one on the basis of the simulation model, cannot fully replace an expert examination board, in which subjective evaluation and pedagogical expertise of its individual members make it possible to give overall a fuller and more objective evaluation of each student's knowledge. However, such a method is not always possible in the conditions of today's computer-based learning. Creation of expert committees is further limited by economic considerations and is implemented in the rare cases when different supervisory committees are created to ascertain a student's inability to master a discipline, or in controversial cases, etc. Most universities are forced to find their own ways to make educational process in the environment of market relationships economically feasible and, based on the need to reduce expenses related to the educational process, increasingly switch to various forms of test-based knowledge control. The mathematical justification for the examiner's decision-making procedure, which lies within the framework of the model proposed here, will significantly mitigate the weaknesses of computerized testing as a tool for educational measurement.

8 References

1. Alekseev, A. N. Multi–level computerized control of learning success of quality of knowledge by quantitative parameters. Computer modeling and new technologies, 11(3), 43 – 45 (2007)
2. Alekseev, A., Aleksieieva, M., Lozova, K., Nahorna, T. Using Fuzzy Logic in Knowledge Tests ICTERI 2015: 11th International Conference on ICT in Education, Research, and Industrial Applications. – Lviv,. 51 – 61 (2015)
3. Baker, F. B. The Basics of Item Response Theory, ERIC Clearinghouse on Assessment and Evaluation, University of Maryland, College Park, MD (2001)
4. Birnbaum A. Some Latent Trait Models and Their Use in Inferring and Examinee's Ability. In Lord F. M., Novick M. Statistical Theories of Mental Test Scores. Addison-Wesley Publ. Co. Reading, Mass, 397 – 479 (1968)
5. Bugbee, A. C., Bernt, F. M. Testing By Computer: Findings in Six Years of Use 1982 – 1988, Journal of Research on Computing in Education, Vol. 23, 1, 87 – 100 (1990).
6. Gulliksen H. Theory of mental tests. – N-Y. Willey (1950)
7. Hambleton, R. K., Swaminathan, H, Rogers H. J. Fundamentals of Item Response Theory. N-Y.: SAGE Publications (1991)
8. IMS Global Learning Consortium. Accessible at http://www.imsglobal.org/question Ten-Brink, T. D. An educator's guide to classroom assessment - Boston: Houghton Mifflin (2003)
9. Kuder, G. F., Richardson, M. W. The theory of the estimation of test reliability // Psychometrika, 1937, v.2, N3, 151 – 160 (1937)
10. Legg, S. M., Buhr, D. C. Computerized Adaptive Testing with Different Groups, Educational Measurement: Issues and Practice, 23 – 27 (1992)
11. Lord F. M., Novick M. Statistical Theories of Mental Test Scoires. Addison-Westley Publ. Co. Reading, Mass (1968)
12. Moe, K. C., Johnson, M. F. Participants Reactions to Computerized Testing, Journal of Educational Computing Research, Vol. 4, 1, 79 – 86 (1988)
13. Rasch, G, Probabilistic Models for Some Intelligence and Attainment Tests //With a Foreword and Afterworld by B. D. Wright.- Chicago: The Univ. Press (1968)

14. Spearman C. Correlation calculated from faulty data //British Journal of Psychology, Vol.3, 2, 271 – 295 (1910)
15. Sutton, R. Equity and Computers in the Schools: A Decade of Research, Review of Educational Research. Vol. 61, 4, 475 – 503 (1991)
16. Volkov, N. I., Alexeyev, A. N., Kochevsky, A. N. Composing of test using the software tool SSUQuestionnaire. Education Technologies on Electronic Platforms in Engineering Higher Education: symposium, Bucharest, 297 – 304 (2005)
17. Volkov, N. I., Alexeyev, A. N., Kochevsky, A. N. Software tool for computer–aided testing of students: types of questions Bulletin of the Petroleum and Gas University of Ploiest, 3 (I,LVII), 41 – 51 (2005)
18. Ware, J. E., Gandek, B., Sinclair, S. J., Bjorner, J. B. Item response theory and computerized adaptive testing: Implications for outcomes measurement in rehabilitation. Rehabilitation Psychology, 50, 71-78 (2005)
19. Weiss D. J., Kingsbury G. G. Application of computerized adaptive testing to educational problems // Journal of Educational Measurement, № 21, 361—375 (1984)
20. Wilson, M., Engelhard, G. R., Draney, R, Objective measurement. Theory into practice. England: Ab. Publ. Corporation, Vol. 4, 37 – 50 (1997)
21. Wright, B. D., Stone, M. H. Best Test Design. Chicago: Messa Press (1979)