[1] **Angelina G. Kirkova-Bogdanova***

E-Test – from Writing Test Items to Statistical Validation

Abstract: - Education, assessment, and learning are interconnected and influential components of the learning process. As technology continues to advance, electronic assessment methods have evolved to reliably evaluate knowledge and skills across the cognitive scale. While online test examinations offer clear benefits, such as convenience and objectivity, there are challenges in creating and programming the tests in electronic systems. Developing a valid and reliable electronic test requires specific knowledge and skills.

This paper aims to provide a comprehensive overview of the process involved in creating a valid and reliable electronic didactic test. It summarizes the requirements and recommendations for crafting effective test questions, and explores the development of a test specification, which involves distributing questions based on learning units and cognitive levels. Additionally, the paper discusses the contrast groups method for establishing the cut-off point.

It is important to view the test as a comprehensive assessment of learning objectives at various cognitive levels, rather than a mechanical collection of exam questions. The qualities of the educational test, particularly its objectivity, are determined by statistical indicators. The paper also delves into various methods for determining the test's validity and reliability.

Keywords: assessment, education, electronic test, validation.

I. INTRODUCTION

Three activities, interdependent and interrelating, make up the complex learning process. These are teaching, learning and assessment. Assessment is the process of systematically documenting knowledge, skills, and attitudes [1], [2]. It is a dynamically changing element, the development of which depends on changes in education goals, determined by the social environment, and on the rapid growth and mass penetration of educational technologies.

The main tasks of assessment in higher education today are:

- Assessment of competencies - a wide range not only of knowledge and skills but also attitudes, skills for applying what has been learned in different contexts of real practice.
- Providing feedback to support learning.
- Cultivating critical thinking and self-assessment skills in students, which can only be developed through their involvement in the assessment process.

These tasks can be successfully performed in e-learning environments that have a rich collection of activities implementing different types of assessment, including modern methods such as self-assessment and peer assessment. The online learning environment requires assessment models that are not only suitable for the new way of acquiring knowledge and skills, but also that provide a balance of assessment methods for both lower and higher intellectual skills, both quantitative and qualitative forms of knowledge [3]. Electronic evaluation methods are developing in parallel with the technologies for their implementation. They have evolved to focus on the process rather than the result and enable the building and following of a development plan in higher education. They can reliably assess knowledge and skills along the entire cognitive scale – from information reproduction to critical evaluation and self-assessment skills.

The progress of the students towards the fulfilment of the set educational goals can be tracked in an electronic environment through various activities, one of which is the electronic test. Online testing has its undoubted advantages – machine-timed and more strictly managed; assessment is automatic and objective, which increases students' confidence; paper is saved - a problem to which we are becoming more and more sensitive; summary presentation of the results of the individual student and the group as a whole, including in graphic form; do not require a competent evaluator, very often the evaluator is the software itself; once programmed, the electronic test can be used repeatedly. A disadvantage is the complexity of the preparation of the test questions and the test itself, further aggravated by the subsequent programming in the electronic system. Preparing a valid and reliable online test requires additional knowledge and skills from the test maker. Test modules in eLearning content management systems typically support different modes of operation and have many options for effective assessment, whether

[1] Medical University Plovdiv, Faculty of Public Health, Department of Medical Informatics, Biostatistics and E-learning
* Corresponding Author Email: angelina.kirkova@mu-plovdiv.bg

summative or formative. It also has tools to limit cheating. Adaptive tests can be programmed where subsequent questions depend on student responses. An important functionality is the statistical analysis of the test and test questions.

II. WRITING TEST ITEMS FOR AN E-TEST

The first and most important task is to prepare the test questions. In terms of rules and advice for preparing test questions, the literature is abundant. The instructional design experts Smith and Ragan suggest that language should be kept as simple as possible, the difficulty should not come from decoding complex and convoluted text. Also be careful not to suggest the answer to the question, for example by gender and number agreement [4]. Of course, the statement must be clear, but the simplicity of the language should not compromise the terminological apparatus of the discipline.

For preparing questions and tests, Tornyova [5] advises:

- to ensure unambiguity of statements.
- statements to be clear, concise and logical.
- to avoid specific words.
- the questions are formulated briefly, clearly, accurately and correctly.
- the wrong answers look as credible as possible.
- uniformity of all answers in terms of grammar and style.
- that there is no system in the arrangement of the possible answers.

The composition of the test questions (items) must be tailored to the environment in which the test will be conducted. If the environment is electronic, it is a good idea to first familiarize yourself with what types of questions it supports, how their evaluation is programmed, and the features of automatic evaluation of open-ended questions.

To save time and effort in checking dozens of tests, and to demonstrate objectivity in grading, we recommend automatic grading. Closed questions are appropriate in this case. Other advantages of multiple-choice questions are the immediate measurement of knowledge and skills, efficient administration - students do not waste time writing, and teachers assess quickly with a template. Multiple-choice questions with more than one correct answer are not good choices for final grading because of the subjectivity of awarding or deducting points for a correct or incorrect answer. Although it is technologically possible to program to deduct points for a wrong answer, didactically this is not a good practice, and students can mark all answers as correct, thereby fooling the system. However, these questions are good when we use them for or as learning, because in learning content management systems they allow feedback at different levels - on a correct answer, on a wrong choice or on the whole question. We can create a valid and reliable test only from multiple-choice questions [6].

Before we start compiling the items, we can make a table in which we can distribute them by learning units and by cognitive levels. This table, part of the test specification, can help us distribute the questions evenly. Lorin Anderson [7] revised Bloom's Taxonomy considering the increasing penetration of technology in the assessment process in higher education.

The more questions we measure the cognitive level to a study unit, i.e. the larger the values in the cells of the table, the more reliable the statistical results in assessing the reliability of the test. Of course, these values cannot be infinitely large, the recommendation is that they should not be less than three. At least three questions are required to check the achievement of a lesson objective at a specific cognitive level. It is recommended to include more questions because some of them may be dropped due to poor statistics. The internal consistency of the measurement scale also depends on the number of items in it – it is greater with more questions.

Learning units can be sections, topics, lessons or topic points - any self-contained part to which lesson objectives are attached. Test questions are closely tied to lesson objectives, so most instructional design models recommend that item writing be at the goal-setting stage [4].

One way to limit cheating is to give more questions that measure comprehension, transfer, and analysis of what has been learned than questions that measure declarative knowledge. Of course, there is a certain number of concepts and facts that must be known for students to reason about them, but these are the questions where it is easiest to cheat on a distance test.

A feature of the test questions that we determine at the stage of their creation is the weight. It represents the relative share of the task in forming the total score. Expresses how important is the achievement of the learning objective checked with the corresponding test item. It may depend on the learning content, for example, basic concepts, procedures, etc. It also depends on the cognitive level to which the task is assigned in the specification table. Higher cognitive levels should have more weight.

Weight is a subjective parameter. To improve objectivity, Cartwright and Mussio [8] suggest that it be calculated as the median of the severity ratings of a group of experts. Weight is optional, it is quite possible to develop a valid and reliable test in which all tasks have equal weight. Weight and difficulty are two different characteristics of test questions. Weight is determined by the test author at the development stage, while difficulty is a statistical quantity calculated after the exam test is run.

III. VALIDITY AND RELIABILITY

The didactic test is a modern form of assessment, with established advantages, recognized by all researchers, regardless of differences in assessment classification [9]. Objectivity, guaranteeing unbiased evaluation, is the main advantage of the test. Objectivity indirectly increases the motivation for learning, the reduction of which is one of the main problems of educational theorists today [10]. The objective test requires serious preliminary preparation and statistical analyses. There are two main requirements for assessment - validity and reliability.

Assessment is valid when it verifies the achievement of precisely those knowledge and skills to be assessed. Evidence in valid assessment covers a range of skills and knowledge in a variety of contexts and situations [11]. Validity can be construct, content, face, and criterion.

A construct valid test measures exactly that characteristic (knowledge, skill, attitude) that it is designed to measure through observable and measurable indicators of that characteristic. It is achieved by carefully designing the test questions so that they describe in measurable categories the construct being assessed. The statistical procedure for determining construct validity is factor analysis. Exploratory Factor Analysis (EFA - Exploratory Factor Analysis) is a statistical method that reveals whether there are hidden variables (factors) and how many there are, as well as which item refers to which factor. Exploratory factor analysis is a useful tool when constructing a new measurement scale when the researcher wishes it to be homogeneous and measure the same characteristic [12]. The difficulty here comes from the large number of participants in the sample required. The ratio of the studied variables to the number of respondents should be at least 1:10-15 [12], that is, to conduct a factor analysis of a test of 10 questions, it should be completed by 100-150 students. This is necessary to fulfil the conditions for conducting factor analysis:

- minimum permissible value of the sample adequacy measure KMO (Kaiser-Meyer-Olkin) is 0.6; the closer it is to 1, the more applicable factor analysis is;
- Bartlett's Test for sphericity is statistically significant at $p<0.05$.

Content validity – determines whether the test measures all aspects of the characteristic being assessed. It is achieved by asking questions, at least three on each lesson objective/point of the taught topic. We avoid questions that are not taught or do not directly relate to lesson objectives.

Face validity – external, overall, whether the test is fit for purpose. The easiest way to determine face validity is to ask a colleague to comment on the assessment tool.

Criterion validity – determines how closely the created test resembles results from another standardized and validated measurement. It can be determined statistically by looking for a correlation between the results of the two measurements – a high correlation is an indicator of good criterion validity. It can also be determined by T-test if we have the population mean of a criterion measurement with the same scale, for example from literature.

When an assessment procedure is reliable, it produces the same results when repeated or similar when assessed on another cohort with the same didactic characteristics. Conducting it with other reliable methods should not yield statistically significantly different results. According to Knight [11], measures of real, rater-independent phenomena are reliable if they fulfil the following conditions:

- objective – they are not compromised by the prejudices of the observer.
- accurate – measurement methods are stable and sensitive.
- repeatable – the measurement procedure does not change in all cases of its application.
- analytically sound – the assessment is set correctly according to the requested scale, and the results are accurately entered into a suitable statistical software package.

There are statistical methods for determining the reliability of an examination test and for determining the statistical parameters of test tasks, which are discussed in the following points. The e-learning system Moodle, for example, has a built-in statistical module that provides evidence of the reliability of e-testing. This is extremely useful because it provides us with information based on which we can adjust questions, answers, and distractors to improve statistical performance.

An important stage in the preparation of the didactic test is its testing with a small, randomly selected group of students, about 30-40 students, under strict transcription control and subsequent statistical processing of the results, which will objectively prove the validity and reliability of the test.

Validity and reliability are not the only criteria for qualitative assessment. Other characteristics include impartiality, accessibility, usability, transparency, relevance, and inclusion. Unbiased evaluation implies an equal start for all evaluated. Prior agreement between the teacher and the students is necessary for the conditions under which the procedure will be conducted, the criteria/indicators for achieving each grade from the selected scale, for the assessed competencies, for the time and place of the test. Electronic assessment should be available to all students, they should be informed in advance about the necessary technical resources, a trial test should be conducted if possible, and the design of the method should not take an unreasonable amount of time and money. To have transparency - students clearly and unequivocally understand the assessment criteria to prepare adequately. Exam tasks should be relevant and ensure demonstration of knowledge and skills authentically. The time of the assessment is planned so that all students can participate. If they do not have enough time to prepare, assessment can be an obstacle for learners to show what they have learned, and this directly affects the motivation for learning.

IV. STATISTICAL EVALUATION OF RELIABILITY

The test should not be seen as a mechanical collection of exam questions but as a complex of units assessing the achievement of learning objectives at different cognitive levels.

The qualities of the didactic test are determined by its statistical indicators. The performance of statistical test validation procedures is usually underestimated, so we will pay special attention to the didactic test validation procedure as a tool for measuring cognitive processes.

There are several methods of determining test reliability. We will discuss some of them.

4.1 Test-retest method

By definition, a reliable test produces the same results every time it is administered to the same students. Therefore, the most logical way to check whether we have created a reliable test is to offer it twice to the same group of students - a test and a retest. This is of course not done in one day, but with an interval of several days to several weeks. It is also important that the number of participants in the sample is not less than 30, even some authors recommend N≥50. This is necessary to avoid checking for the normality of metric variables and to have a reason to apply parametric methods for statistical analysis [12].

Statistical processing includes:

1. Paired Samples T-test. The test questions are analyzed. The pairs of variables that are defined in the t-test are each student's first measurement and second measurement scores for each test question. We remove questions where there is a statistically significant difference in means between the first and second tests.
2. Determination of correlation coefficients according to Pearson Ri of the variables expressing the results on questions, between the first and second administration of the test. Reliable, accurately measuring questions have a high correlation between each measurement. We remove questions that have a statistically significant low correlation, i.e. $Ri \geq 0.500$, $p \leq 0.05$.
3. Determination of the reliability coefficient of the Spearman-Brown test by the formula $Rsb = 2R/(1+R)$, where R is the Pearson correlation coefficient between the scores (test results) of the first and second testing. According to [13], a sufficiently reliable test has $Rsb > 0.06$, but we should not work with tests whose reliability is lower than 0.8.

Reliability measurement by this method can be done with specialized statistical software or with MS Excel, in the Data Analysis module.

The "test-retest" method is often used, but more so in psychological research. In didactic tests, it is not suitable because the results on the second run may be compromised by the fact that the participants are already familiar with the questions and may have searched for the correct answers, which would have led to higher results. Fatigue and the fact that the test is already familiar can also affect the result.

4.2 Split-half method

This method determines the internal consistency of the measurement scale, in our case – of the didactic test by dividing it into two parts [14], [15]. This is a very convenient method because it requires a one-time test. The test is divided into two halves, with even questions forming one half and odd-numbered questions forming the other half. The basis of this method is the fact that routine and fatigue are the same for two consecutive questions, for example between the 1st and the 2nd or between the 30th and the 31st. A Pearson correlation coefficient was

calculated between the scores from the two halves. If the test is good, there should be a high correlation between the two halves. This means that strong students perform well on both even and odd questions of the test, while weak students perform poorly regardless of whether the question is even or odd. The compromise here is that the correlation is underestimated by halving the elements of the strings being compared.

To determine the reliability, we apply the Spearman-Brown method, discussed in the previous point, as a parameter in the formula is the obtained Pearson correlation coefficient.

4.3 Determination of internal consistency using the Cronbach's Alpha method

According to [16], the coefficient α from the "Cronbach's Alpha" method does not determine the reliability of the test, but its lower limit. As the tasks are more reliable and measure the same score, α approaches its maximum value of 1. This coefficient depends on the number of items that form the measurement scale. The larger the number of exam questions, the larger the value for α we will get. It is not suitable to calculate for tests with less than 20 questions. For dichotomous variables, i.e. questions are scored as true (1) and false (0), the particular variant of Cronbach's α is used - the Kuder-Richardson coefficient (KR) [17], [18]. Cronbach's α value is comparable to the Spearman-Brown coefficient. We consider the scale reliable if $\alpha \geq 0.8$.

The choice of method for determining reliability depends on the purpose of the test, the context in which it will be conducted, the number of exam questions, and the number of students tested. It is good practice to subject each test to a reliability check. If there are statistical deviations in groups with similar didactic characteristics, the test may be compromised, which would lead to unfair scoring. As university teachers, we need to be sensitive to the problem of objectivity and equity of assessment to be fair in assigning a numerical measure of learners' competencies.

V. DETERMINATION OF STATISTICAL PARAMETERS OF TEST QUESTIONS

Analysis of the test items is necessary to detect possible defects in the measurement qualities of the test due to poorly functioning items that do not adequately measure the achievement that the entire test measures and/or are of inappropriate difficulty [16]. The multiple-choice questions that we recommend when building an electronic test are subject to statistical processing to obtain valuable diagnostic information.

5.1 Difficulty

The difficulty shows how well the students are doing in giving the correct answer. It is expressed in the ratio between the number of students who solved the task correctly and all students. It takes values from 0 to 1, where 0 means that no one gave a correct answer, and 1 - everyone solved the investigated task correctly. Very easy and very difficult tasks are not a good measure, and tasks whose difficulty is 0 or 1 should be removed from the test because they have no diagnostic value.

Taib and Yusoff [19] recommend that the difficulty is in the range of 0.20-0.80. According to Alashka [16], the optimal task has a difficulty between 0.41 and 0.70.

5.2 Correlation between item and score (Corrected Item-Total Correlation)

This method of evaluating test units shows the correlation between the results of the studied question and the total score from which the value of the corresponding question is subtracted (this is what the correction consists of). Good questions correlate with the total score, and those with a negative correlation disrupt the consistency of the scale.

Analysis by this method can be done in statistical processing software and is part of the reliability analysis of the scale using Cronbach's Alpha method.

There are other methods of analyzing a test task according to the correlation between achievement on the specific task and the total score. Zijlmans et al. [20] consider three other methods – MS, λ6 and CA and compare them with Corrected Item-Total Correlation. The authors point out the advantages of the methods but provide no evidence that they are better than the known and commonly used method discussed in this section.

5.3 Discrimination index DI

This parameter indicates the ability of the test task to distinguish strong from weak students. The concept here is that strong learners are most likely to solve a given task while weak learners are not. This is assumed if solving the task is the result only of the personal efforts of the test participants, of their knowledge and skills. Therefore, this characteristic indicates how random, contrived, or copied the response is, or whether it truly measures student achievement.

To determine the DI, we first determine the strong and weak groups of students. When listed in descending order of total score, students with a score up to the 27th percentile form the strong group, and those with a score after the 73rd percentile form the weak group. DI is the difference between the average score of students from the strong group and the average score of students from the weak group, with a correctly solved task having a value of 1 and an incorrect one - 0. This parameter takes values from -1 to 1, with a negative value indicating a lack of discrimination efficiency and tasks with DI≤0 should be removed from the test or edited. Taib and Yusoff [19] indicated 0.40 as the cutoff for DI. According to Alashka [16], the task is good at 0.31<DI<0.40 and very good at 0.41<DI<1.00.

5.4 Distractor analysis

Distractors (incorrect answers in test items) should sound believable and be approximately as long as the correct answer. Writing distractors is not an easy task, a good strategy is to use the most common mistakes students make. There is a method for analyzing distractors and it is based on the relative contribution of a given distractor to the total number of responses. According to Alashka [16], distractors that collected less than 5% of responses are not preferable.

In other methods of assessment of achievements, for example, presentation of written work, demonstration of skills, project, etc. the assessed activity is a single component. The test question does not exist by itself, but in the context of the purpose of the test and depends on the other items. To measure abilities and cognitive functioning at different levels we use a collection of questions, and whether a particular item contributes to this process depends on its relationship with the rest of the exam questions. Test questions are not created arbitrarily, but always in agreement with the test specification. This approach contributes to achieving good statistical indicators of the items and good reliability of the whole test.

VI. Determining a cut-off point

An important stage in the preparation of the test is the determination of the MRLC (Minimum Required Level of Competence). MRLC can be determined using the contrast group method [21]. In this method, the cut-off value for a successfully passed test is determined by the intersection of the frequency distributions of the scores of two groups of students, determined by experts, according to the results of a previous assessment or according to their self-assessment, as a strong and a weak group.

VII. Conclusion

E-testing is growing in popularity. During the Covid-19 pandemic, the electronic test was one of the most widely used forms of testing. Objective assessment not only increases trust in the examiner but also ensures a reliable assessment of the assessee's competencies. There are available software tools for the statistical analysis of didactic tests, including statistical modules in e-learning systems such as Moodle. The habit of keeping track of test statistics and the ability to correctly interpret the results are important starting points for creating a fair electronic test assessment of student achievement.

References

[1] J. Heil and D. Ifenthaler, "Online Assessment in Higher Education: A Systematic Review," Online Learning, vol. 27, no. 1, 2023.

[2] N. J. Rao and S. Banerjee, "Classroom Assessment in Higher Education," Higher Education for the Future, vol. 10, no. 1, 2023.

[3] M. Northcote, "Online assessment in higher education: The influence of pedagogy on the construction of students' epistemologies," Issues In Educational Research, vol. 13, no. 1, pp. 66-84, 2003.

[4] P. Smith and T. Ragan, Instructional Design, 2nd ed., New York: John Wiley & Sons, Inc, 1999.

[5] B. Tornyova, " Docimology in pandemic conditions," MU-Plovdiv, 2020. [in Bulgarian]

[6] A. Kirkova-Bogdanova, "Computer literacy of healthcare students from Medical University - Plovdiv," in CBU International Conference Proceedings, Prague, 2017.

[7] D. Clark, 12 Jan 2015. [Online]. Available: http://www.nwlink.com/~donclark/hrd/bloom.html. [Accessed 19 януари 2015].

[8] F. Cartwright and J. Mussio, "Development of test: A manual," Strategies for Policy in Science and Education, vol. 21, no. 1, pp. 112-127, 2013.

[9] S. Ruskov and J. Ruskova, " The problem of assessment in higher education," in Scientific Proceedings XXI International Conference "Trans&Motauto '13", 2013. [in Bulgarian]

[10] R. Stancheva, "A culture of test assessment in education - homegrown practices and global standards," in Culture of Education and Education in Culture, vol. 2, Union of Scientists in Bulgaria, 2010, pp. 38-44. [in Bulgarian]

[11] P. Knight, "A Briefing on Key Concepts Formative and Summative, Criterion Norm Referenced Assessment," 2001. [Online]. Available: http://www.heacademy.ac.uk. [Accessed 10 December 2006].

[12] Z. Ganeva, Let's reinvent statistics with IBM SPSS Statistics., Elestra Ltd, 2016. [in Bulgarian]

[13] A. Lazarova, "Types of errors when working with statistical scales in online-based surveys," Science and business, no. 7, pp. 4-10, 2017. [in Bulgarian]

[14] T. Pronk, D. Molenaar, R. W. Wiers and J. Muure, "Methods to split cognitive task data for estimating split-half reliability: A comprehensive review and systematic assessment," Psychonomic Bulletin & Review, vol. 29, pp. 44-54, 2021.

[15] E. R. Van Norman and D. C. Parker, "A Comparison of Split-Half and Multilevel Methods to Assess the Reliability of Progress Monitoring Outcomes," Journal of Psychoeducational Assessment, vol. 36, no. 6, 2017.

[16] R. Alashka, "Application of probabilistic models for the analysis of exam and test results. Dissertation thesis for the award of the educational and scientific degree "Doctor", SU "Kliment Ohridski", Sofia, 2017. [in Bulgarian]

[17] V. Vijayagopal and K. Prabu, "A trust and energy-based efficient routing scheme using kuder-richardson reliability coefficient for manets," International Journal of Scientific & Technology Research, vol. 9, pp. 890-894, 2020.

[18] D. Uyanah and U. I. Nsikhe, "The Theoretical and Empirical Equivalence of Cronbach," International Research Journal of Innovations in Engineering and Technology, vol. 7, no. 5, pp. 17-23, 2023.

[19] F. Taib and M. S. B. Yusoff, "Difficulty index, discrimination index, sensitivity and specificity of long case and multiple choice questions to predict medical students' examination performance," Journal of Taibah University Medical Sciences, vol. 9, no. 2, pp. 110-114, 2014.

[20] E. Zijlmans, J. Tijmstra, L. A. van der Ark and K. Sijtsma, "Item-Score Reliability as a Selection Tool in Test Construction," Frontiers in Psychology, vol. 9, p. doi.org/10.3389/fpsyg.2018.02298, 2019.

[21] S. A. Shrock and W. C. Coscarelli, Criterion-Referenced Test Development: Technical and Legal Guidelines for Corporate Training, San Francisco: Pfeiffer, 2007.