Eighth Grades in in the Public Schools of Des Moines. lowa, with the Author's Measures of Musical Talent].
9. Seashore, C. E. (1938). *Muzichna psihologiy.* [Psychology of music]. London.
10. Seashore, C. E. (1937). *Ob"yektivnyy analiz muzykal'nogo ispolneniya.* [Objective Analysis of Musical Performance].
11. Seashore, C. E. (1956). *Rukovodstvo po izmereniyu muzykal'nykh talantov.* [Measures of Musical Talents Manual].

ВІДОМОСТІ ПРО АВТОРА

ЛЮ Хунюй – аспірант факультету мистецтв імені Анатолія Авдієвського Українського державного університету імені Михайла Драгоманова.

Наукові інтереси: мистецька освіта, музичні здібності, музичний слух, інтегральний підхід.

INFORMATION ABOUT THE AUTHOR

LIU Hunyuy – is a graduate student at the Faculty of Arts named after Anatoly Avdiyevskii of Mykhailo Drahomanov Ukrainian State University.

Circle of scientific interests: art education, musical abilities, musical ear, integral approach.

Стаття надійшла до редакції 16.10.2024 р.

UDC 371.263:51
DOI: https://doi.org/10.59694/ped_sciences.2024.10.136

PAPP Gabriella Gaborivna –
teacher in Ferenc Rakoczi II Transcarpathian Hungarian College of Higher Education, Department of Mathematics and Informatics,
ORCID iD: https://orcid.org/0000-0001-9470-7119
e-mail: papp.gabriella@kmf.org.ua

EXAMINING E-TEST AND ITS GOODNESS INDICATORS

ПАП Габріелла Габорівна. **ДОСЛІДЖЕННЯ ЕЛЕКТРОННИХ ТЕСТІВ ТА ІНДИКАТОРІВ ЇХ ДОБРОЯКІСНОСТІ**

Вимірювання рівня знань за допомогою електронних тестів є непростим завданням, якщо ви створюєте тести самостійно. Ми стикаємося з такими питаннями, як: «Чи вдалося нам створити об'єктивний тест?», «Чи вимірює е-тест відповідно, чи досягнуто мети тесту?», «Яка надійність е-тесту?».

Ключові слова*: електронний тест, індикатори доброякісності, вимірювання рівня знань, надійність електронного тесту.*

PAPP Gabriella Gaborivna. **EXAMINING E-TEST AND ITS GOODNESS INDICATORS**

Measuring knowledge levels using e-tests is not easy if you create the tests ourselves. We are faced with questions such as: «Have we succeeded in creating an objective test?», «Does the e-test measure accordingly, has the aim of the test been achieved?», «What is the reliability of the e-test?».

The study aims to search and examine relevant literature to get a clearer picture of tests and their correct creation, taking into account objectivity, validity and reliability. Also, we would like to present the results of the testing conducted during the 2022-2023 academic year. We have examined the goodness indicators of our e-tests within the framework of the subject «Probability Calculation, Mathematical Statistics and Econometrics» course, modified by previous results and used in the measurement of knowledge levels, to find answers to the above questions.

Keywords: *e-test, goodness indicators, measuring knowledge levels, reliability of the e-test.*

Statement and justification of the relevance of the problem. As technology advances, we need to integrate different tools and platforms into education increasingly. A tool used correctly can also improve student demotivation. One of the obvious options is electronic testing of students, which requires the teacher to know the characteristics that a test should have to develop it properly. These properties are called the goodness indicators of the e-test.

Analysis of recent research and publications. Studying the scientific literature provides a detailed description of the goodness indicators and their properties.

The concept of tests, the problems of their development, and their goodness indicators (objectivity, validity and reliability) are discussed

by Csapó B. [2], Demkanin P. [4], Molnár Gy. [9] Goforth C. [6]. However, despite numerous scientific publications, the majority of teachers do not know test theory well enough.

The purpose of the article is to search for and review the relevant literature to edit tests professionally. Also, it presents the results of a study we have carried out, taking into account the goodness indicators of the self-made e-test.

Presentation of the main research material. In contrast to classical test theory, modern (probabilistic) test theory (Item Response Theory), which is a new generation of test theory, uses probabilistic instruments to characterise item properties [9]. The aim of using any measurement instrument is to provide an accurate and reliable measure and objective evaluation of the property under test [10].

In the literature on tests and testing, we can find different characteristics or properties of tests. They can be divided into two groups, within which a certain hierarchy prevails:

- Basic or main characteristics of tests (validity, reliability),
- Other characteristics of tests (difficulty, sensitivity, relevance, objectivity) [4].

The term validity, when applied to test scores, refers to the consistency (accuracy) with which scores measure a given cognitive ability [5]. A fundamental prerequisite for validity is the development of a detailed test specification, which should include clear test objectives [4]. These have two aspects: what they measure and how they measure validity [5].

According to the literature, a good test is valid if you know what you want to measure, how you want to measure it, and if the test measures what you want to measure [4]. The fundamental task in the test design process is to define the measurement objective, the requirements and the test content [10]. Then, we examine the property that the test measures what it is designed to measure [2]. This requires first of all to demonstrate to the users of the test scores (1) the meaning of the scores and (2) the appropriateness of using the scores [5].

One of the most important properties of the test is its reliability [2]. Test reliability is an indicator of the accuracy of measurement [4], which estimates measurement errors at the group level [10].

The reliability coefficient is an indicator of the degree of error associated with a score and is thus important information for evaluating the meaningfulness and usefulness of scores [5]. Each score that a student achieves on a test can be divided into two parts in terms of accuracy:

- the correct score – it should reflect what the learner actually knows,
- the incorrect score – part of the result but not reflecting the learner's actual knowledge. The error values can be used to find out what the learner does not know or what the learner's learning problem is [4].

Reliability can be estimated using various methods [10], including determining the correlation between half-tests [2] and using a universal indicator, Cronbach's alpha [10], where comparisons of variances are also used [2]. Cronbach created the formula in 1951:

$$\alpha = \frac{n}{n-1}\left(1 - \frac{\sum_i v_i}{v_t}\right)$$ [1].

Here n is the number of items, V_t is the variance of total scores, and V_i is the weighted variance of item scores [1].

Ideally, the reliability of a test is 1, which means that if such a test were administered repeatedly, all students would score the same on each item as on the first measurement. In practice, however, reliability 1 does not occur [4].

Reliability can be quantified by a margin between 0 and 1 [3], as the resulting reliability coefficient α ranges between 0 and 1. Although the standards for producing a «good» α coefficient are entirely self-imposed and depend on the theoretical knowledge of the scale in question, many method experts recommend a minimum α coefficient between 0.65 and 0.85 (or higher in many cases); α coefficients lower than 0.5 are generally unacceptable [4], [5], [6].

The reliability value depends on the number and quality of items. In general, the higher the number of items, the more reliable the measurement [10]. If the number of items is small or the average correlation is low, the Cronbach's alpha will be low [7]. Therefore, formal tests usually consist of a relatively large number of items. From a practical point of view, a test should consist of at least 20-30 items, and increasing the number of items can increase reliability [5], [10].

Csapó refers to the relationship between task reliability and validity [2]. According to Demkanin et al. (2015), reliability is a necessary but not sufficient condition for validity [4]. For a test to be valid, it must also be reliable [3], i.e. it must have a good reliability index. The converse is not true, if the validity of a test is poor, its reliability can be very high [2]. Reliability and accuracy are therefore a prerequisite for validity, but not the only prerequisite [4].

Test objectivity means that the test is relevant, unbiased, non-subjective [2], always measures everyone equally [8]. Some authors consider it as an essential characteristic (factor) that influences reliability. It also means excluding (or strongly reducing) random or subjective factors in testing [4]. More specifically, the result of a test measurement is independent of who performs the test measurement [2].

The objectivity of the knowledge assessment tests used in schools can be raised to a satisfactory

level by following a few simple rules [3]. The objectivity of testing is ensured by:

(a) test objectivity,

(b) the objectivity of the scoring,

(c) objectivity of the testing process [4].

Test objectivity is achieved through the correct selection of items, clear wording of questions and response options, adequate coverage of items, and eliminating any doubt about the correct answers to items [4], ensuring that the test result is independent of the respondent [2]. Objectivity of the assessment is ensured by precise and equal scoring rules and by scoring (marking) identical answers in the same way [4], so that its results are independent of who does the correcting, and coding, i.e. scoring of the results [2]. Objectivity of the testing process is achieved by providing equal testing conditions for all students tested [4].

Research and results. Having familiarised ourselves with the relevant literature on test theory and test's goodness indicators, we aimed to develop an e-test and evaluate its success.

The research subject was probability, one of the most difficult chapters (according to students) in mathematics education. It was applied in two 1-th grade groups of their Bachelor's programme of the Ferenc Rakoczi II Transcarpathian Hungarian College of Higher Education, in the spring semester of 2022-2023 academic year for the first chapter of the practical lessons of the subject «Probability Calculation, Mathematical Statistics, and Econometrics». Students wrote one input, three intermediate, and one output e-test during the chapter. The study involved 21 participants, however, we could use the results of 18 students for the lack of output test.

The tests that are part of the research were created in the OnlineTestPad interface. The input and output tests consist of 10 items, include both closed and open tasks, and for a more accurate measurement, they include one-choice and multiple-choice items with one or more correct answers, pairing, and short-answer items. The intermediate tests are 4 items, they include one-choice tasks.

Since the reliability of the test depends on the number of correct solutions too, we examined and analysed the output test. In this test, each task was worth a maximum of 1 point, a total of 10 points. The range of the test is 4.25, which is not much exceeded by its average - 4.56.

Students we divided into 5 groups based on their performance, this was done to test the difficulty of the tasks. These group results are shown in Figure 1, where G1 is the group of students with the highest scores and G5 is the group with the lowest scores.

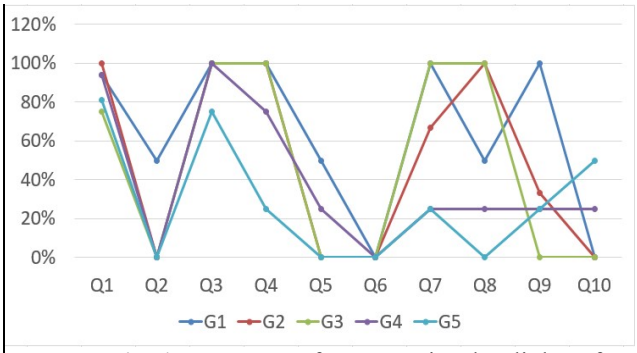

Fig. 1. Group performance in the light of items.

The test we consider quasi-valid based on its difficulty, as the questions are designed to be differentiating and to measure knowledge. The success of this is demonstrated in Figure 1. We can see that Q6 was the most unsuccessful, with no student getting the correct answer, while Q3 had the highest number of correct answers.

The reliability of the output test was assessed using Cronbach's alpha, which was calculated using the appropriate functions of the Microsoft Excel spreadsheet program. Our calculations give a reliability of 0.14 which is below the «good» reliability. The low reliability may be due to the small number of items, as we deviated from the amount specified in the literature because of the difficulty of the subject.

In addition to the differentiating difficulty of the items, the objectivity of the test we sought by providing clear and test-theoretically defined correct answers and distractors. Objectivity in scoring was facilitated by setting up the e-test on the web platform. The test was written in the practical class under the supervision of the instructor to ensure objectivity in the testing process.

Conclusions and prospects of further scientific investigations. In conclusion, it should be noted that the test we have presented, which we have carried out, has not been entirely successful. Nevertheless, we can provide answers to the above questions.

The test can be considered objective based on the choice of platform, settings, and question creation. It had a differentiating effect and measured knowledge levels, thus achieving the objective of the test. Reliability is rather low, which may be due to the small number of items, as the recommended number of items based on the literature we not respected for this test due to the difficulty of the subject.

In the future, it is worth further developing the e-test, adding more items, and if possible increasing the number of respondents to more accurately demonstrate reliability.

СПИСОК ДЖЕРЕЛ

1. Cronbach L. J. Coefficient alpha and the internal structure of tests. *Psychometrika*, *1951. 16*, 297–334. [Електронний ресурс].: Режим доступу: https://link.springer.com/article/10.1007/BF02310555
2. Csapó B. Tudásszintmérő tesztek. In Falus I. (Ed.), *Bevezetés a pedagógiai kutatás módszereibe* (4. kiad., pp. 277–316). 2004. Budapest: Műszaki K. [Електронний ресурс].: Режим доступу: https://core.ac.uk/download/pdf/84775002.pdf
3. Csíkos C., & B. Német M. A tesztekkel mérhető tudás. *Csapó Benő (szerk.) Az iskolai tudás* 2002. (pp. 91–123). Budapest: Osiris Kiadó. [Електронний ресурс].: Режим доступу: https://publicatio.bibl.u-szeged.hu/11931/1/CsBeno_Iskolai_tudas_2002.pdf
4. Demkanin P., Hajdúk M., Hanuljakova H., Kubiš T., Lapitka M., & Malčík M. Metodika tvorby testových úloh a testov (Mgr. Timotej Kubiš.). 2015. Bratislava. [Електронний ресурс].: Режим доступу: https://www.researchgate.net/publication/349279914_Metodika_tvorby_testovych_uloh_a_testov
5. Ebel R., & Frisbie D. Essentials educational measurement. Prentice Hall. 1991.
6. Goforth C. Using and Interpreting Cronbach's Alpha. *University of Virginia Library*. 2015. [Електронний ресурс].: Режим доступу: Hidegkuti I., & Balázs K. Tesztelmélet. 2015. [Електронний ресурс].: Режим доступу: https://www.researchgate.net/publication/305678594_Tesztelmelet
7. Magyar A., & Molnár G. Számítógép alapú adaptív és rögzített formátumú tesztelés összehasonlító hatékonyságvizsgálata. *Magyar Pedagógia*, 2013. *113*(3), 181–193.
8. Molnár G. Az ismeretek alkalmazásának vizsgálata modern tesztelméleti (IRT) eszközökkel. *Magyar Pedagógia*, *2003. 103*(4), 423–446.
9. Molnár E. K., & Vígh T. *A tantervelmélet és a pedagógiai értékelés alapjai.* «Mentor(h)áló 2.0 Program» TÁMOP-4.1.2.B.2-13/1-2013-0008 projekt. 2013. [Електронний ресурс].: Режим доступу: http://www.jgypk.hu/mentorhalo/tananyag/Tantervelmlet_s_a_pedaggiai_rtkels_alapjai/index.html

REFERENCES

1. Cronbach, L. J. (1951). *Koefitsiyent al'fa i vnutrishnya struktura testiv.* [Coefficient alpha and the internal structure of tests].
2. Csapó B. (2004). *Tudásszintmérő tesztek. In Falus I.* [Tudásszintmérő tesztek. In Falus I]. Budapest.
3. Csíkos, C., & B. Német, M. (2002). *A tesztekkel mérhető tudás.* [A tesztekkel mérhető tudás]. Budapest.
4. Demkanin, P., Hajdúk, M., Hanuljakova, H., Kubiš, T., Lapitka, M., & Malčík, M. (2015). *Metodyka stvorennya testovykh ulokh i testiv.* [Metodika tvorby testových úloh a testov (Mgr. Timotej Kubiš)]. Bratislava.
5. Ebel, R., & Frisbie, D. (1991). *Osnovy osvitn'oho vymiryuvannya. Prentis Kholl.* [Essentials educational measurement Prentice Hall].
6. Goforth, C. (2015). *Vykorystannya ta interpretatsiya Al'fa Kronbakha.* [Using and Interpreting Cronbach's Alpha]. Virginia.
7. Hidegkuti, I., & Balázs, K. (2015). *Számítógép alapú adaptív és rögzített formátumú tesztelés összehasonlító hatékonyságvizsgálata.* [Számítógép alapú adaptív és rögzített formátumú tesztelés összehasonlító hatékonyságvizsgálata].
8. Magyar A., & Molnár G. (2013). *Az ismeretek alkalmazásának vizsgálata modern tesztelméleti (IRT) eszközökkel.* [Számítógép alapú adaptív és rögzített formátumú tesztelés összehasonlító hatékonyságvizsgálata].
9. Molnár G. (2003). *A tantervelmélet és a pedagógiai értékelés alapjai.* [Az ismeretek alkalmazásának vizsgálata modern tesztelméleti (IRT) eszközökkel].
10. Molnár, E. K., & Vígh, T. (2013). *A tantervelmélet és a pedagógiai értékelés alapjai. Proekt «Prohrama Mentor(h)áló 2.0».* [A tantervelmélet és a pedagógiai értékelés alapjai. «Mentor(h)áló 2.0 Program»].

ВІДОМОСТІ ПРО АВТОРА

ПАП Габрієлла Габорівна – старший викладач кафедри математики та інформатики Закарпатського угорського інституту імені Ференца Ракоці ІІ, студентка науково-технічного факуль-тету Докторської школи математичних та комп'ютерних наук Інституту математики Дебреценського університету.

Наукові інтереси: створення та використання електронних тестів на уроках математики.

INFORMATION ABOUT THE AUTHOR

PAPP Gabriella Gaborivna – senior teacher of the Department of Mathematics and Informatics of the Transcarpathian Hungarian Institute named after Ferenc Rakoczy II, student of the scientific and technical faculty of the Doctoral School of Mathematical and Computer Sciences of the Institute of Mathematics of the University of Debrecen.

Circle of scientific interests: creating and using electronic tests in mathematics lessons.

Стаття надійшла до редакції 10.11.2024 р.