ISSN 2029-7564 (online)
SOCIALINĖS TECHNOLOGIJOS
SOCIAL TECHNOLOGIES
2012, 2(1), p. 40–52.

INTEGRAL CRITERIA FOR MEASURING THE QUALITY OF TEACHER EVALUATION

Olga Navickienė

Mykolas Romeris University, Lithuania, navickiene@mruni.eu

Aleksandras Krylovas

Mykolas Romeris University, Lithuania, krylovas@mruni.eu

Abstract

Purpose—*To construct the knowledge evaluation quality integral criteria, which allows to ascertain whether the different teachers properly assess the students' knowledge. The criteria has been tested setting up the educational experiment and examining the six mathematics lecturers' assessments.*

Design/methodology/approach—*This research involved the Mykolas Romeris University students of Public Administration degree second year and Management of Organizations degree first year. The test questions for students were designed using the mathematical knowledge assessment information system, which allows for closed-ended mathematical test, to obtain statistical data about test takers, to perform quality analysis of the test; in the middle and the end of the semester.*

Findings—*The construction technique for the evaluation quality criteria of the students' working results assessment, which were performed by six different lecturers, during practical trainings, seminars, laboratory and other sessions is proposed in this article.*

Research limitations/implications—*The constructed evaluation criteria is universal: it does not depend on the particular subject; it can be applied to several groups, courses or lecturers. It depends on three calculated indicators I, S, K, which show in two ways*

ISSN 2029-7564 (online)
https://www.mruni.eu/lt/mokslo_darbai/st/apie_leidini/
https://www.mruni.eu/en/mokslo_darbai/st/apie_leidini/index.php

obtained estimates of the measured information compatibility of degrees, marks matching and correlation terms.

Practical implications*—The integral criteria has been tested examining the six mathematics lecturers' assessments.*

Originality/ Value*—Constructing the criteria have been used the educational measurement models of authors of this article and other researchers, however its' connection to general (integral) criteria, best of authors' knowledge, is original and have not be researched before.*

Keywords*: knowledge evaluation, quality of teacher evaluation, evaluation of teachers, mathematical modeling.*

Research type*: research paper.*

1. Introduction

This paper analyses how the different lecturers assess achievements of students during Applied Mathematics and Quantitative Methods practical trainings and the construction of the evaluation quality criteria, which have to show whether lecturers properly assess the students' knowledge. Constructing the criteria have been used different researchers' educational measurement models. These models constitute the new integral criteria, which have been tested setting up real educational experiments. This study has not only theoretical but also practical significance. The *aim* of this research is to construct the knowledge evaluation quality integral criteria and to test it examining the six mathematics lecturer assessments. This research involved practical training lecturer assessments, which are compared with the same student knowledge assessments for 20 questions of closed-end tests, of 10 academic groups (262 students). Tests have equal variants of problems (Krylovas, Raulynaitis, 2003; Krylovas *et al*, 2002), i.e. each student gets a different but equal in difficulty test variant (Krylovas *et al*, 2007). The authors of this article used these tests for a number of years in the didactical research. The tests were organized for all groups in two months and four month after the beginning of the semester. Prior to this, teachers had to evaluate students' working results during practical trainings on a scale from 1 to 10; assessment criteria had been stated during the first practical training.

The object of lecturer's assessment named knowledge, as it is understood in the broad sense. Basically it is a certain *construct* (Kardelis, 2007), which, in our opinion, is closely related to the student's knowledge. Lecturer evaluation accounted for 30% of the Applied Mathematics course cumulative grade, and the students were motivated enough to get a high evaluation for the practical trainings. Thus, the article examines the whole complex of elements of teachers' work: encouragement of students' activeness, the evaluation of their efforts and results, etc. We propose a universal, i.e. unrelated to the subject and number of objects evaluated; technique of constructing criteria and demonstrating how it has been applied to the practical assessment. The p*ractical aim*

of this study is to ascertain whether the different lecturers assess properly the students' knowledge.

The idea to construct the knowledge evaluation quality integral criteria has been suggested in the previous article (Navickienė, Krylovas, 2012). The analysis of the indicator of informativeness of assessments have been done extra and new research data examined, however previous research data are presented for comparison in this article.

2. Theoretical background

Knowledge evaluation is one of the most important elements of the study process and the objectivity of the evaluation is absolutely essential to ensure the quality of education (Dranevičienė, 2005). The lack of objectivity can be a source of conflict, what is especially relevant for students' tight competition for e.g. state-funded places and so on. However, an objective assessment of knowledge depends not only on teacher's experience, integrity, or other individual characteristics (Gage, Berliner, 1994; Peterson, 1987). This is a complex phenomenon, widely considered in the relevant literature (Anastasi, Urbina, 1997). A based assessment is considered to be a satisfactory solution to the problem (Bulajeva, 2007), when the knowledge of the test takers is being compared with each other's, for example, the state examinations. However, this method do not applied to a small number of students and on the other hand, the regulatory documents often require criteria-based assessment, evaluating the level of certain skills acquisition by the specialist (Pukelis, Savickienė, 2003). The methodological and practical aspects of such assessment have been addressed in the literature (Hopkins, 1998; Andziulienė, 2004; Butėnas, 2009). The authors of the articles (Kriauzienė *et al*, 2010; Krylovas *et al*, 2002) applied a various statistical methods to examine the problem of knowledge evaluation. These studies show a number of differences in obtained estimates clearly depending on assessors (Krylovas *et al*, 2006; Blanton *et al*, 2006). In recent years, interest has grown in using classroom observation as a means to several ends, including teacher development, teacher evaluation, and impact evaluation of classroom-based interventions (Hill *et al*, 2011; Rani, 2004). Measures of teacher effects are of interest as a means of answering at least two broad questions: 1. Do teachers have differential effects on student outcomes? 2. How effective is an individual teacher at producing growth in student achievement, and which teachers are most or least effective? (McCaffrey *et al*, 2003). Existing studies employ a variety of empirical models (Lefgren, Sims, 2012; Harris *et al*, 2010; Kane, Straiger, 2008; Koretz, 2002; Medley *et al*, 1984).

3. Research methodology

Examined is the assessment of student results in Applied Mathematics and Quantitative Methods practical trainings, which were performed by six different lecturers. Lecturers performed practical trainings in ten academic groups of students.

Student's progress assessed in accordance with the cumulative grade system. The final assessment consisted of: the first and the second tests – 35% each, 30%—the work during practical trainings, i.e. 3 points out of 10 in ten-point rating system. This study was accomplished after two months from the beginning of the semester, i.e. after 12 practical trainings, and after four months from the beginning of the semester, i.e. after 10 more practical trainings. The students took the first and the second tests, preceded by lecturers' evaluation of each student's work during the practical trainings in ten-point system. The students work during the practical trainings have been organised at the discretion of each lecturer of the practical trainings. The lecturer could organize independent tasks, the defence of obligatory written tasks or any other type of assessment of student progress. These tasks account for only 3 points of the final grade for Applied Mathematics and Quantitative Methods and were stated during the first practical training. Also, all students were required to submit two written tasks that lecturer could have evaluated in point system or limited the evaluation with pass or fail. In this article lecturers bear the initials AK, ON, LG, RK, JK, TL.

The AK lecturer's assessment is calculated by taking into account the attendance estimate with the weight of 1/3 and the activeness estimate with the weight of 2/3. Attendance L calculated as the part of the practical trainings attended to the total number of practical trainings (22 practical trainings per the first two months of the semester). Student activeness A is calculated in the following way: the number of correctly solved problems in the semester (at the blackboard or shown to the lecturer; and so during independent tasks) divided by 7—the points that were collected by students who study well, although a few of them had much better results.

The JK and LG lecturers scored of equal value the attendance calculated as the part of the practical trainings attended to the total number of practical trainings (22 practical trainings), the activeness and the one independent task.

The ON lecturer's assessments includes the part of the practical trainings attended to the total number with the weight 1/2, the activeness: the independent correctly problems solving at the blackboard; with the weight of 1/12, two independent tasks of 5 problems each and eight homework assignments (30 problems) with weights 1/6, the defence for two written tasks: all the problems of written tasks solved correctly and any of their solution explained orally; with the weight of 1/12.

The RK lecturer's assessment calculated by taking into account the attendance estimate with the weight of 1/3, two independent tasks estimate with the weight of 1/2 and the defence for two written tasks estimate with the weight of 1/6.

The TL lecturer's assessment calculated by taking into account the attendance and the activeness estimates with weights of 1/3, two independent tasks and the one homework estimates with weights 1/6.

Students working results assessments by all lecturers during practical trainings are given in *Table 1*.

Table 1. Lecturer assessment techniques (scored)

Lecturer	Attendance	Activeness	Independent tasks	Homework	Defence of written tasks
AK	1	2	–	–	–
JK	1	1	1	–	–
LG	1	1	1	–	–
ON	1,5	0,25	0,5	0,5	0,25
RK	1	–	1,5	–	0,5
TL	1	1	0,5	0,5	–

Lecturer assessments were compared with the results of 20 questions of close-end tests. Tests have equal variants of problems, i.e. each student gets a different but equal in difficulty test variant. For this purpose, three indicators *I, S, K*, depending on the amount of information available from the estimates, the difference between the lecturer's and the test's estimates; and the correlation coefficients between these values, were constructed. Each of these indicators has the higher value the better lecturer's and test's estimates are matched. This corresponds to a requirement for the validity (Anastasi, Urbina, 1997) that indicators reflect the particular characteristics of the construct which they are intended to measure. Indicators are dimensionless values varying from 0 to 100. Therefore, any weighted average also will have the quality of validity. Determination of the weights of indicators requires empirical data analysis, and is the object of our further research.

In this article the calculated indicators are treated as certain rank values, arithmetic operations with which are not performed. They are calculated the same way for an individual academic group, for the unions of groups, corresponding to the lecturer and the whole flow of tested students. This allows comparing the indicator values with the average value calculated for the whole flow and construct the evaluation quality criteria.

4. Indicators

Supposing the number of points $t_0 = 0$, $t_1 = 1$, ..., $t_n = n$ (in our case $n = 20$) gets respectively $k_0, k_1, ..., k_n$ students in a test. With that the amount of information acquired from a test is calculated in the following way (Stakėnas, 1996):

$$entr(k_0, ..., k_n) = -\sum_{j=0}^{n} \frac{k_j}{k} \ln \frac{k_j}{k},$$

here. It is noticing that only the dimension of the amount of information depends on the base of the logarithm (for example, when the base of the logarithm equals 2, the information is measured in well-known *bits*). The largest volume of information

corresponds to the same possible test score distribution: $k_0 = k_1 = \cdots = k_n = \frac{k}{n+1}$. In this case we get $entr = \ln(n+1)$. It is convenient to express the amount of information acquired from the test with the dimensionless value $\frac{entr(k_0, \ldots, k_n)}{\ln(n+1)}$, which shows the relative amount compared with potential maximum (Krylovas, Kosareva, 2008a).

Let $l_1, l_2, \ldots, l_{10}$ be the number of students, who have got mark 1, 2, ..., 10 by lecturer. Then the entropy function is calculated in the following way:

$$entr(l_1, \ldots, l_{10}) = -\sum_{j=0}^{n} \frac{l_j}{l} \ln \frac{l_j}{l},$$

here $l = \sum_{j=1}^{10} l_j$. The function takes the maximum value $\ln 10$, when.

The indicator of informativeness of assessments is defined as follows:

$$I = \frac{entr(l_1, \ldots, l_{10})}{entr(k_0, \ldots, k_n)} 100,$$

showing the amount of information acquired from the estimates of lecturers, compared with the estimates of the test. All the indicators are expressed with dimensionless values varying from 0 to 100. They can be interpreted as percentages. It is noticing that the value of *I* theoretically could be greater than 100, but the authors' experience suggests that this does not happen in practice. It is worth mentioning that the articles (Krylovas, Kosareva, 2008a; Krylovas, Kosareva, 2008b) examine the construction of the tests, maximising the amount of information. It guarantees that $I \leq 100$.

Let r_i be the student's i test mark (the points of the test are converted into a ten-point scale), p_i —the same student's estimate by the lecturer, n—the number of students taking the first test. Now let us consider the indicator of the coincidence of the estimates:

$$S = \frac{100}{1 + \frac{1}{n} \sum_{i=1}^{n} |r_i - p_i|}.$$

The maximum value of *S* represents the case when all test's and lecturer's estimates coincide. In this case, the number of inversions $\sum_{i=1}^{n} |r_i - p_i|$ (Liutikas, 1983; Jaurienė *et al*, 1983; Jaurienė, 1997) is equal to zero in the denominator of the expression. The experience (Raulynaitis, Krylovas, 2002; Krylovas, Raulynaitis, 2004) suggests that this quantity takes rather high values and does not characterised as being stable. The constructed indicator *S* takes the larger value the better students' estimates, which obtained in two ways, coincide.

The third indicator of correlation of assessments is defined as follows:

$$K = \begin{cases} 100r, & kai \quad r \geq r_0, \\ 0, & kai \quad r \geq r_0, \end{cases}$$

here $r = \frac{n\sum_{i=1}^{n} t_i r_i - \left(\sum_{i=1}^{n} t_i\right)\left(\sum_{i=1}^{n} r_i\right)}{\sqrt{\left(n\sum_{i=1}^{n} t_i^2 - \left(\sum_{i=1}^{n} t_i\right)^2\right)\left(n\sum_{i=1}^{n} r_i^2 - \left(\sum_{i=1}^{n} r_i\right)^2\right)}}$ is well-known Pearson's correlation coefficient (between the lecturer's and the test's estimates) (Čekanavičius, Murauskas, 2006), r_0—its critical value, indicating when statistical hypothesis H^0: $r = 0$ is rejected. Under the assumption of normal distribution $r_0 = \frac{t_{\alpha,n-2}}{\sqrt{n-1}}$ of random value r, $t_{\alpha,n-2}$—Student distribution with n—2 degrees of freedom α-level critical value (Čekanavičius, Murauskas, 2006, p. 166). For the purposes of this paper, $r_0 = 0.2$.

The correlation coefficient is a popular measure of the compatibility of the different evaluations (Krylovas *et al*, 2002). It usually takes values $0.5 \le r \le 0.7$, indicating moderate or strong correlations (Raulynaitis, Krylovas, 2002).

We propose to compare the values of indicators of lecturers not to each other, but to values of *I, S, K*, calculated for all students groups. Let denote:

$$i^d = \begin{cases} +, \textbf{kai } I^d \ge I, \\ -, \textbf{kai } I^d \le I, \end{cases}$$

$$s^d = \begin{cases} +, \textbf{kai } S^d \ge S, \\ -, \textbf{kai } S^d \le S, \end{cases}$$

$$k^d = \begin{cases} +, \textbf{kai } K^d \ge K, \\ -, \textbf{kai } K^d \le K, \end{cases}$$

here the index *d* denotes the value of the lecturer's indicator. For example, the value of the AK lecturer's indicator of informativeness of assessments after the first test I^d I^d is compared to the value of the same indicator *I*, of the whole students flow. Thus, each lecturer is assigned to one of the four sets of pluses and minuses: (+ + +), (+ + –), (+ – –), (– – –). In the first case, attention should be paid to the lecturer's assessment system as to a good example of one, the last (the fourth) – should be critically reviewed. In the second and third cases there is no sufficient reason to conclude that the assessment differs significantly from the average values.

5. Analysis of the indicator *I*

The indicator *I* is original and, to the best of our knowledge, has not yet been examined in the relevant literature. Therefore, we will analyse its modification:

$$I_{mod} = \frac{1}{2}\left(\frac{entr(l_1, \ldots, l_{10})}{entr(k_0, \ldots, k_n)} + \frac{entr(k_0, \ldots, k_n)}{entr(l_1, \ldots, l_{10})}\right)100.$$

The values of the modified indicator are in the *Table 2* and *3*.

Table 2. The values of the modified indicator I_{mod}

Group	1	2	3	4	5	6	7	8	9	10	I
I_{mod_1}	104	104	105	102	107	103	105	104	102	102	105
I_{mod_2}	100	108	101	101	103	117	100	102	128	0	105

Table 3. The values of the modified indicator *I* by lecturers

Lecturer	AK	JK	LG	ON	RK	TL
I_{mod_1}	104	104	102	107	103	105
i_1^d	–	–	–	+	–	+
I_{mod_2}	*[1]	112	101	103	117	104
i_2^d	*	+	–	–	+	–

1 Practical trainings of the second part of the semester to the lecturer's AK group were performed by the TL lecturer.

We can see that all the values of modified indicator are greater than 100 and do not characterised by the big variation. Indices *1* and *2* of the indicators are respectively the first and second tests. Thus it is left previously constructed indicator *I*, although its value can be artificially increased. However if a lecturer increase the value of the indicator *I*, he will simultaneously decrease the value of the indicator *S*.

It is interesting to notice that lecturer estimates of the informativeness and the values of the integral indicator according to the requirements of the modified indicator I_{mod} are rather poor. Due to the empirical data is insufficient in quantity it is early to reach a conclusion on this modified criteria. We leave these issues for further research.

6. Results and findings

The results of the educational experiment are given in this paragraph.

Table 4. The number of students

Group	1	2	3	4	5	6	7	8	9	10	Total
Number of students in each group	29	32	27	26	31	32	29	23	17	16	262
Number of students, who have taken the first test	27	30	26	26	31	29	28	23	13	15	248
Number of students, who have taken the second test	14	12	15	14	22	21	18	15	6	4	141

The values of the three indicators are calculated to the each academic group of students. The *Microsoft Excel* tables have been used for these calculations.

Table 5. The values of the three indicators

Group	1	2	3	4	5	6	7	8	9	10	
I_1	78	77	75	83	68	81	74	76	90	83	75
I_2	92	62	89	86	78	56	97	79	48	0	73
S_1	35	25	25	41	42	44	43	37	36	28	34
S_2	39	32	34	37	49	45	43	38	38	44	40
K_1	63	55	21	63	64	68	63	61	55	57	51
K_2	74	0	25	27	53	39	27	37	92	0	31

Indices 1 and 2 of indicators *I, S, K* are respectively the first and second tests. Calculated values of indicators of all students are given in the last column. Given values joined to groups of same lecturers are given in *Table 6*. The JK lecturer had practical trainings in following groups: 2, 3, 9 and 10; the lecturer TL – 7 and 8 groups. It is noticeable that indicators values are determined almost the same not only to one academic group, but also their unions.

Table 6. The values of the indicators of each lecturer

Dėstytojas	AK	JK	LG	ON	RK	TL
I_1	78	77	83	68	81	74
I_2	*	62	86	78	56	74
S_1	35	27	41	42	44	40
S_2	*	35	37	49	45	38
K_1	63	50	63	64	68	60
K_2	*	30	27	53	39	31

It is obvious that none of the six lecturers has the highest or lowest values of all of three indicators (see *Table 6*). This means that there is no sufficient reason to believe that any of lecturer‘s assessment system is the best or worst of all, comparing systems to each other.

7. Conclusions

The values of the integral criteria of all lecturers after two tests are given in *Table 7*.

Table 7. The values of the integral criteria

	AK	JK	LG	ON	RK	TL
i_1^d	+	+	+	–	+	–
i_2^d	*	–	+	+	–	+
s_1^d	+	–	+	+	+	+
s_2^d	*	–	–	+	+	–
k_1^d	+	–	+	+	+	+
k_2^d	*	–	–	+	+	+

The values of the integral criteria of lecturers by groups are given in *Table 8*.

Table 8. The values of the integral criteria by groups

	1	2	3	4	5	6	7	8	9	10
i_1^d	+	+	+	+	–	+	–	+	+	+
i_2^d	+	–	+	+	+	–	+	+	–	–
s_1^d	+	–	–	+	+	+	+	+	+	–
s_2^d	–	–	–	–	+	+	+	–	–	+
k_1^d	+	+	–	+	+	+	+	+	+	+
k_2^d	+	–	–	–	+	+	–	+	+	–

This article proposes the construction technique for the evaluation quality criteria of the students' working results assessment during practical trainings, seminars, laboratory and other sessions. The criterion does not depend on the particular subject and the number of students, teachers or academic groups. The constructed criteria depends on three calculated indicators *I*, *S*, *K*, which show in two ways obtained estimates of the measured information compatibility of degrees, marks matching and correlation terms. These three indicators are intrinsically interesting characteristics and could be an object of a new empirical research.

The constructed integral criteria tested assessing the evaluation quality of teachers of practical trainings of Applied Mathematics and Quantitative Methods. It analysed 262 Mykolas Romeris University students' assessments. The analysis has shown that the constructed criteria is characterised by the big stability, comparing the results of different teachers and students. It is noticing that assessments of first and second parts of semester are similar. The assessment system of one of the lecturers should be critically reviewed; however assessments of the second part may not be statistically significant due to the relatively small number of test takers (see *Table 4*). It should be

noted that comparing the first and the second part of the results of only one of ten groups of estimates significantly changed. It stands to reason that to test the characteristics of constructed criteria is needed more empirical data. It would be our further research.

Literature

Anastasi, A.; Urbina, S. 1997. *Psychological Testing*. Upper Saddle River (NJ): Prentice Hall.

Andziulienė, B. 2004. *Žinių ir gebėjimų testavimas*. Klaipėda: Klaipėdos universiteto leidykla.

Blanton, L. P., *et al*. 2006. Models and Measures of Beginning Teacher Quality. *The Journal of Special Education*. 40(2): 115-127.

Bulajeva, T. 2007. *Žinių ir kompetencijų vertinimas: kaip sukurti studentų pasiekimų vertinimo metodiką: metodinė priemonė*. Vilnius: UAB „Petro osfsetas".

Butėnas, R. 2009. Studentų pasiekimų ir kompetencijų vertinimas. *Tarptautinės mokslinės–praktinės konferencijos „Šiuolaikinio specialisto kompetencijos: teorijos ir praktikos dermė" straipsnių rinkinys*. 13–16.

Čekanavičius, V.; Murauskas, G. 2006. *Statistika ir jos taikymai I*. Vilnius: TEV.

Dranevičienė, N. 2005. Žinių tikrinimas ir vertinimas—studijų proceso tobulinimo veiksnys. *Respublikines mokslines praktinės konferencijos medžiaga. Straipsnių rinkinys*. 143–146.

Gage, N. L.; Berliner, D. C. 1994. *Pedagoginė psichologija*. Vilnius: Alma Littera.

Harris, D., *et al*. 2010. Value-Added Models and the Measurement of Teacher Productivity. CALDER Working Paper No. 54.

Hill, H. C., *et al*. 2011. When Rater Reliability Is Not Enough: Teacher Observation Systems and a Case for the Generalizability Study. *Educational Researcher*. 41: 56-64.

Hopkins, Kenneth D. 1998. *Educational and Psychological Measurement and Evaluation*. Boston: Allyn and Bacon.

Jaurienė, J. 1997. Inversijos intensyvumo mato reikšmių skalės nustatymas remiantis tiesinės koreliacijos koeficiento reikšmių skale. *Pedagogika*. 1(22): 111.

Jaurienė, J., *et al*. 1983. Inversijos intensyvumo mato taikymas lyginant studentų mokymosi pažangumo rezultatus. *Mokymo ir auklėjimo klausimai*. PMTI, Vilnius. p. 136–138.

Kane, T. J.; Straiger, D. O. 2008. *Estimating teacher impacts on student achievement: An experimental evaluation*. Cambridge, MA: National Bureau of Economic Research.

Kardelis, K. 2007. *Mokslinių tyrimų metodologija ir metodai: vadovėlis*. Šiauliai: Lucilijus.

Koretz, D. 2002. Limitations in the use of achievement tests as measures of educators' productivity. *Journal of Human Resources*. 37(4):752-777.

Kriauziene, R., *et al*. 2010. Studentų matematikos žinių vertinimo subjektyvumo problema: teoriniai ir praktiniai aspektai. *Socialinės technologijos*. 1(1): 121–138.

Krylovas, A.; Kosareva, N. 2008a. Žinių tikrinimo matematinis modelis. *Lietuvos matematikos rinkinys*. 48/49: 217–221.

Krylovas, A.; Kosareva, N. 2008b. Mathematical modeling of forecasting the results of knowledge testing. *Technological and economic development of economy: Baltic journal on sustainability*. 14 (3): 388–401.

Krylovas, A., *et al*. 2007. Diskrečiosios matematikos žinių tikrinimo testų lygiagrečiųjų variantų lygiavertiškumo tyrimas. *Lietuvos matematikos rinkinys*. 47 (spec. nr.): 249–253.

Krylovas, A., *et al*. 2006. Dėstytojų nuomonės apie studentų klaidas. *Lietuvos matematikos rinkinys*. 46 (spec. nr.): 158–162.

Krylovas, A.; 2004. Raulynaitis, J. Studentų matematikos žinių kompleksinis vertinimas semestro metu. *Lietuvos matematikos rinkinys*. 44 (spec. nr.): 477–481.

Krylovas, A.; Raulynaitis, J. 2003. Vieno tikimybių teorijos uždavinio išlygiagretinimo patirtis. *Lietuvos matematikos rinkinys*. 43 (spec. nr.): 357–360.

Krylovas, A., *et al*. 2002. Apie matematikos žinių įvairių vertinimų suderinamumą. *Lietuvos matematikos rinkinys*. 42 (spec. nr.): 397–401.

Lefgren, L.; Sims, D. 2012. Using Subject Test Scores Efficiently to Predict Teachers Value-Added. *Educational Evaluation and Policy Analysis*. 34(1): 109-121.

Liutikas, V. 1983. Inversijos intensyvumo matas ir jo taikymas mokinių mokėjimui įvertinti. *Mokymo ir auklėjimo klausimai*. PMTI, Vilnius. p. 133–136.

McCaffrey, D. F., *et al*. 2003. *Evaluating value-added models for teacher accountability*. Santa Monica, CA: Rand.

Medley, D. M., *et al*. 1984. *Measurement-based evaluation of teacher performance: An empirical approach*. New York: Longman.

Navickienė, O.; Krylovas, A. 2012. Studentų žinių vertinimo kokybės kriterijų modeliavimas. *Studijos šiuolaikinėje visuomenėje // Studies in modern society*. 3(1) (In press).

Peterson, D. K. 1987. Teacher Evaluation with Multiple and Variable Lines of Evidence. *American Educational Research Journal*. 24(2): 311-317.

Pukelis, K; Savickienė, I. 2003. Studijų kokybės vertinimo sistemų lyginamoji analizė: pasaulinė patirtis. *Studijų kokybės užtikrinimo sistemos modeliavimas pasaulinės patirties kontekste: konferencijos pranešimų medžiaga*. Kaunas: VDU leidykla. P. 15–27.

Raulynaitis, J.; Krylovas, A. 2002. Matematikos pažymių koreliacinė analizė ir sesijos rezultatų prognozė. *Lietuvos matematikos rinkinys*. 42 (spec. nr.): 438–443.

Stakėnas, V. 1996. *Informacijos kodavimas*. Vilnius: VU leidykla.

Rani, J. S. 2004. *Educational Measurement and Evaluation*. New Delhi: Discovery Publishing House.

INTEGRALINIS KRITERIJUS DĖSTYTOJŲ VERTINIMO KOKYBEI MATUOTI

Olga Navickienė

Mykolo Romerio universitetas, Lietuva, navickiene@mruni.eu

Aleksandras Krylovas

Mykolo Romerio universitetas, Lietuva, krylovas@mruni.eu

Santrauka. *Šiame darbe taikant edukometrinius metodus analizuojami šešių skirtingų dėstytojų, vertinančių studentų darbą taikomosios matematikos ir kiekybinių metodų dalyko praktinių užsiėmimų metu, subjektyvumo įtaka galutiniam studentų žinių vertinimo rezultatui. Studentų žinios vertinamos pagal kaupiamojo balo sistemą. Galutinį įvertį sudaro: pirmas ir antras testai po 35 %; darbas praktinių užsiėmimų metu – 30 %, t. y. 3 balai dešimties balų vertinimo sistema. Šis tyrimas buvo atliktas po dviejų semestro mėnesių, t. y. po*

12 praktinių užsiėmimų, ir po keturių semestro mėnesių – po kitų 10 praktinių užsiėmimų. Studentai laikė pirmą ir antrą testus, prieš kuriuos pratybų dėstytojai buvo įvertinę pažymiu dešimties balų sistema kiekvieno studento darbą pratybų metu. Kiekvienas pratybų dėstytojas savo nuožiūra organizavo studentų darbą praktinių užsiėmimų metu. Dėstytojų vertinimai lyginami su 20 klausimų uždarojo testo laikymo rezultatais. Testai turi išlygiagretintus užduočių variantus, t. y. kiekvienas studentas gauna skirtingą, bet lygiavertį variantą. Šio tyrimo teorinis ir praktinis tikslas yra sukonstruoti studentų praktinių užsiėmimų vertinimo kokybės integralinį kriterijų, kuris rodo, ar dėstytojai tinkamai parinko vertinimo metodikas. Kriterijui konstruoti naudojami įvairių autorių tyrimais pagrįsti didaktinių matavimų matematiniai modeliai, kurie sudaro integralinį kriterijų. Kriterijus tikrinamas autorių atliktu edukologiniu eksperimentu. Dėstytojų vertinimo metodikos lyginamos remiantis trimis indikatoriais, priklausančiais nuo įverčiais teikiamos informacijos kiekio, nuo skirtumo tarp dėstytojo ir testo įverčių, ir nuo koreliacijos koeficientų tarp pastarųjų dydžių. Kiekvienas iš šių indikatorių įgyja tuo didesnę reikšmę, kuo geriau suderinti dėstytojo ir testo vertinimai. Indikatoriai yra bedimensiniai dydžiai, įgyjantys reikšmes nuo 0 iki 100, ir traktuojami kaip tam tikri ranginiai dydžiai, su kuriais aritmetiniai veiksmai neatliekami. Apskaičiuotos indikatorių reikšmės lyginamos ne tarpusavyje, o su reikšmėmis, apskaičiuotomis visoms studentų grupėms. Taigi kiekvienam dėstytojui priskiriamas vienas iš keturių pliusų ir minusų rinkinys: (+ + +), (+ + –), (+ – –), (– – –). Pirmuoju atveju dėstytojo vertinimo sistema laikoma geruoju pavyzdžiu, o ketvirtuoju – dėstytojo vertinimo sistemą reikia kritiškai peržiūrėti. Antruoju ir trečiuoju atvejais nėra pakankamo pagrindo teigti, kad vertinimas smarkiai skiriasi nuo vidutinių reikšmių. Sukonstruotas kriterijus išbandytas vertinant taikomosios matematikos dalyko pratybų dėstytojų vertinimo kokybę. Parodyta, kad dėstytojai tinkamai parinko vertinimo metodikas. Taip pat buvo bandoma modifikuoti šį kriterijų, bet geresnių rezultatų negauta. Sukonstruotas vertinimo kriterijus yra universalus: jis nepriklauso nuo dėstomojo dalyko, taikomas atskiroms grupėms, srautams arba dėstytojams. Jis priklauso nuo trijų apskaičiuojamų indikatorių, kurie yra savaime įdomios statistikos ir galėtų būti naujų empirinių tyrimų objektu.

***Raktažodžiai*:** *žinių vertinimas, dėstytojų vertinimo kokybė, dėstytojų vertinimas, matematinis modeliavimas.*