TECHNOLOGICAL AND ECONOMIC DEVELOPMENT OF ECONOMY
ISSN 2029-4913 print/ISSN 2029-4921 online
2011 Volume 17(1): 42–61
doi: 10.3846/13928619.2011.553994

A NUMERICAL EXPERIMENT ON MATHEMATICAL MODEL OF FORECASTING THE RESULTS OF KNOWLEDGE TESTING

Natalja Kosareva[1], Aleksandras Krylovas[2]

[1]Vilnius Gediminas Technical University, Saulėtekio al. 11, LT-10223 Vilnius, Lithuania
[2]Mykolas Romeris University, Ateities g. 20, LT-08303 Vilnius, Lithuania
E-mails: [1]Natalja.Kosareva@vgtu.lt (corresponding author); [2]Krylovas@mruni.eu

Received 22 March 2010; accepted 20 October 2010

Abstract. In this paper the new approach to the forecasting the results of knowledge testing, proposed earlier by authors, is extended with four classes of parametric functions, the best fitting one from which is selected to approximate item characteristic function. Mathematical model is visualized by two numerical experiments. The first experiment was performed with the purpose to show the procedure of selecting the most appropriate item characteristic function and adjusting the parameters of the model. Goodness-of-fit statistic for detecting misfit of the selected model is calculated. In the second experiment a test of 10 items is constructed for the population with latent ability having normal distribution. Probability distribution of total test result and test information function are calculated when item characteristic functions are selected from four classes of parametric functions. In the next step it is shown how test information function value could be increased by adjusting parameters of item characteristic functions to the observed population. This model could be used not only for knowledge testing but also when solving diagnostic tasks in various fields of human activities. Other advantage of this method is the reduction of resources of testing process by more precise adjustment of the model parameters and decreasing the standard error of measurement of the estimated examinee ability. In the presented example the methodology is applied for solving the problem of microclimate evaluation in office rooms.

Keywords: Item Response Theory (IRT), mathematical modelling, item characteristic function, generating function, Monte Carlo Method.

Reference to this paper should be made as follows: Kosareva, N.; Krylovas, A. 2011. A numerical experiment on mathematical model of forecasting the results of knowledge testing, *Technological and Economic Development of Economy* 17(1): 42–61.

JEL Classification: I31, C15, C38, C63.

1. Introduction

Measuring the knowledge and other mental features is the problem which has it‘s particularity because of difficulty to determine the object of investigation and deficiency of measuring instruments. One will agree that it is much more difficult to measure person‘s attainment in some knowledge field than his physical properties. A very important stage is to create the appropriate instruments – questionnaires that allow getting maximum information about a measured feature. For the construction of a "good" questionnaire we must be able to choose the most informative subset of items from the whole item bank. This subset of items must be suited to the population under investigation so that the information supplied by test reaches it‘s maximum value. In this article we will deal with dichotomous test items, when there are only two possibilities to answer – test item may be responded correctly or incorrectly. In practice there can be more than two answer categories in questionnaires. For such cases polytomous latent variable models are developed. Another approach is when all incorrect answer categories are joined to one category and we derive dichotomous model as well. Knowledge testing problem is the object of investigation of Item Response Theory (Rasch 1960).

The last articles on IRT are concerned with computerized adaptive tests, i.e. individualized tests that are optimal for each individual (Eggen, Verschoor 2006); latent class analysis (LCA) – a statistical method used to identify a set of discrete, mutually exclusive latent classes of individuals based on their responses to a set of observed categorical variables (Lanza *et al.* 2007); new technologies such as heuristic search and machine learning approaches, including neural networks to automatically identify the most informative subset of test items when the item bank is very large (El-Alfy, Abdel-Aal 2008); tests of model misfit to validate the use of a particular model in IRT (Wells, Bolt 2008); evaluation of the standard error of the estimated latent variable score (Hoshino, Shigemasu 2008); new IRT software development (Rizopoulos 2006).

El-Alfy, Abdel-Aal (2008) proposed a new approach of abductive network modelling to automatically select most informative subset of test items without serious loss of accuracy. This method was compared to three parameter logistic IRT model (3PL). The accuracy of IRT-based model was slightly better, nevertheless the new abductive network approach enable to reduce number of test items from 45 to 12 which classified an evaluation population with 91% accuracy.

Van Barneveld (2007) analyzed the effect of aberrant response patterns on test construction. Data was generated using two item response models – the three parameter logistic IRT model (3PL) alone and combined with Wise‘s examinee persistence model. Item parameters were estimated using the maximum marginal likelihood estimate approach with Bayesian priors on the item parameters using the program BILOG-MG (Zimowski, Muraki, Mislevy & Bock 1996). Tests were constructed using an optimal item selection method. Items with the largest item information estimates at each of the targeted cut-off ability points were selected for the optimal test. Biased item parameter estimates, item and test information estimates were obtained from responses from poorly motivated examinees.

Wells, Bolt (2008) investigated a nonparametric method for detecting misfit when using the two-parameter logistic model (2PL). Two nonparametric statistics for detecting misfit based on the (Douglas, Cohen 2001) approach were examined. The results were compared to other well known goodness-of-fit statistics $S-X^2$ (Orlando, Thissen 2000) and BILOG‘s G^2 (Mislevy, Bock 1982). For all studied conditions the methods based on the nonparametric approach exhibited more power to detect the misfit while also controlling Type I error rate. It is hypothesized that nonparametric statistics provide a more informative description of the nature of misfit, which can help in diagnosing the cause of misfit (e.g., guessing).

Savalei (2006) proposed the approximation of standard normal distribution with a logistic distribution with scaling constant 1.749 based on minimizing the Kullback-Leibler (KL) information function. This approximation is compared with Item Response Theory logistic function, in which another constant 1.702 is used. The new approximation gives better fit on the tails of the distribution.

Hoshino, Shigemasu (2008) proposed a formula to evaluate the variance of the estimated latent variable score when the true values of the structural parameters are not known and must be estimated. It is shown that the appropriate accuracy is reached when the number of subjects and items are both large. For all conditions considered the standard errors of ability parameters using the proposed method were less than those using familiar standard errors as the inverse of the test information.

Eggen, Verschoor (2006) investigated computerized adaptive tests (CAT), which select an optimal test for each individual. Such test is realized by selecting, on the basis of the results of previously selected items, the most informative item from the item bank. The optimal selection of item often means that item will be chosen for the individual student, which has a 50% of probability of answering correctly. But such tests are often too difficult for students and this fact has its negative side effects. To eliminate these effects two item selection procedures giving easier or more difficult tests were analyzed for both one (1PL) and two (2PL) parameter logistic models. The first procedure based on the success probability points of selected items shows good results in ability estimates measurement precision only for 1PL model. Another item selection procedure based on maximum information at shifted ability level gives good results for both 1PL and 2PL models.

Rizopoulos (2006) developed the package **ltm** for the well known open source statistical software **R** for the analysis of multivariate dichotomous and polytomous data using Item Response Theory logistic models. Parameter estimates are obtained under marginal maximum likelihood using the Gauss-Hermite quadrature rule. This package is suitable not only for unidimentional latent variable models but also when there is a small set of latent variables which explain the observed data.

St-Onge *et al.* (2009) compared parametric and nonparametric Item Characteristic Curve estimation methods on the effectiveness of Person-Fit Statistics (PFS). For both large and small sample sizes, the accuracy of the PFS was greater when used with the parametric models.

The aim of this paper is to propose the model for forecasting the results of knowledge testing when the best fitting item characteristic function is selected from 4 classes of parametric functions. Prior distribution of knowledge level of the population could be chosen from 4 classes of probability density functions. However, this model allows using parametric functions of another form, the main restriction of the model is that item characteristic function has to be nondecreasing and items have to be mutually independent.

Earlier it was shown (Krylovas, Kosareva 2008a, b) how segments of linear functions could be used as an item characteristic function and also as a probability density function of the population knowledge level. It was shown how this approach could be used to construct norms-referenced latent trait estimations to select test items which are optimally fitted to the examined population. In (Krylovas, Kosareva 2009a) the generalization of this model with wider set of item characteristic functions and probability density functions was presented. This model could be used not only for knowledge testing, but also for solving diagnostic tasks in various fields of human activities. The problems of decision making in the information deficiency conditions were analyzed in (Zavadskas *et al.* 2009, 2010).

2. Rasch and normal ogive models

The well known models in latent trait testing theory are the One Parameter Logistic model (1PL) and normal ogive model. These models describe the conditional probability of correct response to the item i given ability level $p : P\left(x_i = 1 \middle| p\right)$. This functional relation is denoted $k_i(p) = P\left(x_i = 1 \middle| p\right)$ and its graph is called item characteristic curve (ICC). In 1PL the logistic function is applied to describe this relation (Rasch 1960):

$$k_i(p) = \frac{1}{1 + e^{-\left(p - b_i\right)}},$$

here b_i is the difficulty parameter of item i. The function above, which is called *item characteristic function,* belongs to the class of logistic functions. Rasch model is taking place if some constraints on the model are satisfied (Molenaar 2007):

i. Unidimensionality of latent trait p. This means that p is one-dimensional quantity which reflects person's ability to answer test items correctly. At a time one mental property is measured, the influence of other latent traits is treated as negligible.

ii. Conditional independence of items given person and conditional independence of persons given item. Given the person's ability p the elements of the response vector are independent. On the other hand, person's response to the item is independent of other respondent's responses to this item. Respondents do not influence the responses of each other.

iii. Monotonicity of item response functions. The item response function $P\left(x_i = 1 \middle| p\right)$ is nondecreasing function of p.

iv. Sufficiency of total test score. The total score $\sum x_i$ is a sufficient statistic for p.

Formula (1) represents generalization of Rasch módel (Birnbaum 1968) were supplementary parameters are discrimination parameter a_i (2PL model) and additionally the probability of random guessing c_i of the item i (3PL model):

$$P\left(x_i = 1 \middle| p\right) = c_i + (1 - c_i)\frac{1}{1 + e^{-a_i\left(\theta - b_i\right)}}. \tag{1}$$

In 1PL, 2PL and 3PL models probabilities of correct response to the items are "S" – shaped functions. 2PL and 3PL logistic models satisfy only i. – iii. conditions, while condition iv. is generally not required.

In the normal ogive model the link function between given ability level p and the probability is standard normal probability distribution function (Uebersax 1999):

$$k_i(p) = \Phi\left(\frac{p - b_i}{a_i}\right),$$

with the same interpretation of parameters b_i and a_i. We suggest using wider class of link functions that enables more precise approximation of ICC and as a consequence more efficient tests.

3. Problems and proposals for their solution

Before the examinee testing process begins each item must be calibrated according to the selected model. Due to the restrictions on one class of parametric functions (either on the logistic functions or on the standard normal probability distribution function) the calibration process results in biases of the item parameter estimates. These biases cause the biases in the test information function value's estimates $I(\hat{p})$ and consequently in the precision of examinee ability estimates $\hat{p}$. The other sequel of biases in item parameter estimates is that the standard error of measurement of the estimated examinee ability, when overestimated at a given ability level, results in excess number of items proposed to examinees with intention to reach the nominal precision (Van Barneveld 2007). This enlarges the resources of testing process.

According to the formula (2) (Lord 1980) the standard error of measurement is inversely proportional to the square root of test information function $I(\hat{p})$:

$$SE(\hat{p}) = \frac{1}{\sqrt{I(\hat{p})}} = \frac{1}{\sqrt{\sum_i a_i^2 k_i(\hat{p})\left(1 - k_i(\hat{p})\right)}}. \qquad (2)$$

Our proposal is that the reduction of the bias in item parameter estimates is possible **not only by increasing the number of examinees** in calibration group **or/and number of test items** proposed to the examinees **but also by expanding the set of item characteristic functions** $k(p)$ which we select the best fitting one from. In the proposed model the best fitting ICC is selected from one of the 4 classes of parametric functions depending on one or two parameters. These functions are – 2 parameter logistic function restricted in the interval [0; 1] described by (3); arccotangent function (4); segments of linear functions (5); segments of 2 parabolas (6).

$$k_1(p;a;b) = \frac{1}{1 + e^{-a(p-b)}}, \ a \ge 0, b \in [0;1], \qquad (3)$$

$$k_2(p;a) = \frac{2}{\pi} \operatorname{arccot}\left(\frac{a \ln(p)}{\ln(1-p)}\right), \ a \ge 0, \qquad (4)$$

$$k_3(p;a;b) = \begin{cases} 0, \ p < a \\ \dfrac{p-a}{b-a}, \ a \le p \le b, \\ 1, \ p > b \end{cases} \qquad (5)$$

$$k_4(p;a;b) = \begin{cases} 0, \ p \le a \\ 2\left(\dfrac{p-a}{b-a}\right)^2, \ a < p \le \dfrac{a+b}{2}, \\ 1 - 2\left(\dfrac{b-p}{b-a}\right)^2, \ \dfrac{a+b}{2} < p \le b. \\ 1, \ p > b \end{cases} \qquad (6)$$

In the Figs 1–2 graphs of functions $k_1(p)$ – $k_4(p)$ with various parameter values are imaged.

It is notable, that all these functions have their definition range in the interval [0; 1], so functions $k_1(p)$ – $k_4(p)$ are defined for ability levels from this interval. The interpretation of examinee's ability level $p \in [0; 1]$ is the proportion of maximum value of ability score $\left(p_{\max} = 1\right)$, which the examinee possesses. In IRT the ability value (usually noted by θ) in theory belongs to the interval $(-\infty;+\infty)$ and in practice it is in the interval $(-3;+3)$. The ability level of the examinee in IRT can not be estimated when number of correct responses to the N items test is equal to its minimum or maximum value (0 or N respectively). Our model enables estimation of the ability level in such marginal cases with the number in the interval [0; 1]. The particular estimated ability value depends on the selected ICC of the model.

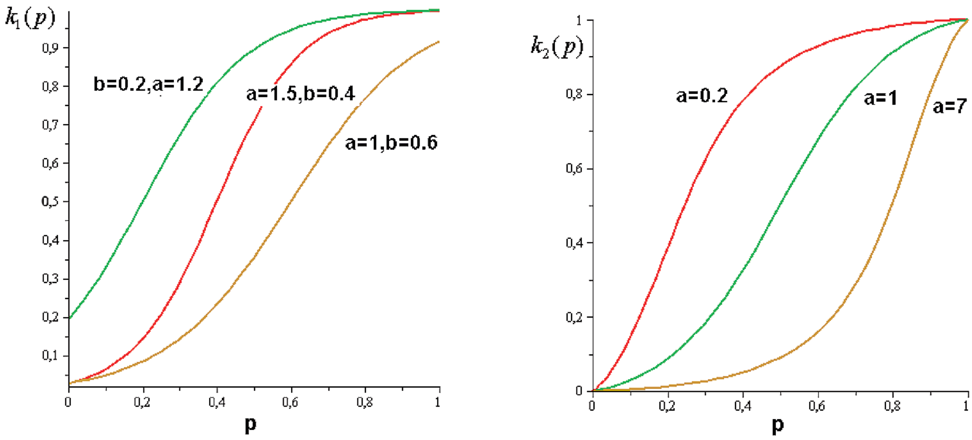

Fig. 1. Graphs of functions $k_1(p)$ and $k_2(p)$

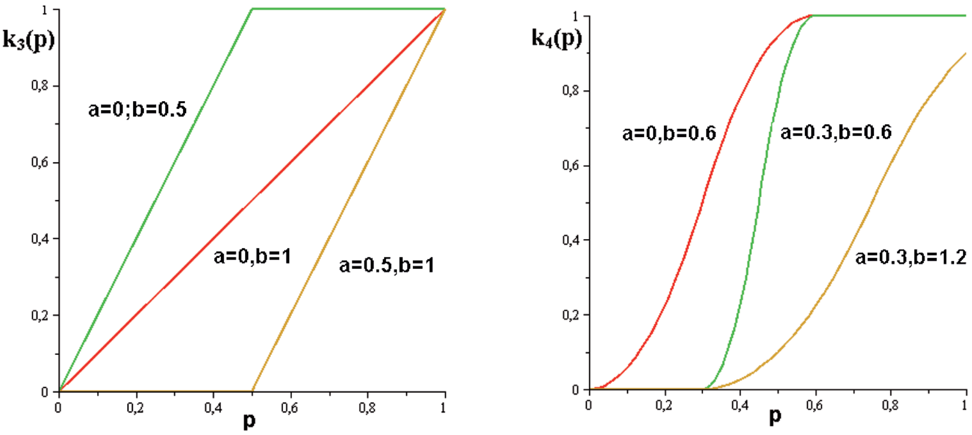

Fig. 2. Graphs of functions $k_2(p)$ and $k_3(p)$

Function (3) is the two parameters logistic function (2PL). When restricted in the interval [0; 1] it has the attractive property to be similar with the three parameter logistic function (1) for low item discrimination parameter a and difficulty parameter b values. This function obtains values greater than zero for low p values. So we get the effect of guessing without guessing parameter of 3PL. Likewise for low parameter a and high parameter b values function's (3) value is less than 1 for high ability levels p. This can also improve the estimate of ICC in some situations.

The selection of the most appropriate model from these function classes and estimation of item parameters that best fit the observed proportions of correct responses could be done as described in (Baker 2001). Examinees are grouped into ability intervals based on their ability scores. The interval [0; 1] is divided to J intervals of equal length with m_j examinees in the j -th interval. The total number of examinees is $M=\sum_{j=1}^{J} m_j$. The examinees within the same interval have the same ability score p_j (we have taken this point at the middle of the j -th interval). Let r_j examinees of ability score p_j answered the item correctly. Then the observed proportion of correct responses to the item at ability score p_j is $P(p_j)=\frac{r_j}{m_j}$. Our purpose is to select the function from 4 function classes, which will provide the best accuracy of the approximation to the observed proportions of correct responses to the item.

At the first step the best approximation of ICC from the 2 parameters logistic functions class $k_1(p;a;b)$ described by (3) is found. The initial estimates of item parameters a_1 and b_1 are established a priori. Then values of $k_1(p_j;a_1;b_1)$ are computed for all ability scores $p_j, j=1,2,...J$ and the distance

$$d_1(k_1)=\sqrt{\frac{1}{M}\sum_{j=1}^{M}\left(k_1\left(p_j,a_1,b_1\right)-P\left(p_j\right)\right)^2}\ , \tag{7}$$

is calculated. In the next iteration adjustments to the estimated parameters a_2 and b_2 , which improve the agreement between $k_1(p;\ a;\ b)$ and observed proportions of correct responses are found. So, $d_2(k_1)<d_1(k_1)$ and this process is continued until the improvement of the agreement becomes very small. Then current values of parameters a_{n_1} and b_{n_1} are fixed and they are considered item parameter estimates for $k_1(p;a;b)$. This procedure is repeated for functions $k_2(p)$ – $k_4(p)$ determined by (4)–(6) at the next steps, and the minimum value is chosen from the four distances $d_{n_1}(k_1)-d_{n_4}(k_4)$. The corresponding ICC is the best fitting model to the observed data which is chosen from 4 classes of functions (3)–(6).

Goodness-of-fit X^2 statistic's value for detecting misfit of the selected model is defined as follows:

$$\mathrm{X}^2=\sum_{j=1}^{J} m_j\frac{\left(k\left(p_j\right)-P\left(p_j\right)\right)^2}{k\left(p_j\right)\left(1-k\left(p_j\right)\right)}\ , \tag{8}$$

here $k\left(p\right)$ is the best fitting model ICC, found in previous step.

X^2 statistic has χ^2 distribution with $J-s-1$ degrees of freedom when $k\left(p\right)$ is suitable for the observed data. Here J is the number of grouping intervals, s is the number of parameters of the model under investigation. For example, $s = 1$ for the arccotangent function (4) and $s = 2$ for the functions (3), (5), (6). The observed value of statistic (8) is compared with criterion value which is equal to χ^2 distribution with $J - s - 1$ degrees of freedom critical value $\chi^2_{0.05}(J - s - 1)$. If calculated X^2 statistic's value is greater than criterion value $\chi^2_{0.05}(J - s - 1)$ then corresponding ICC does not fit the data and vice versa. There is the requirement to have more than 5 observations in each interval for X^2 statistic (Bagdonavicius, Kruopis 2007). When number of the observed data in some interval is less than 5, the adjacent intervals may be joined together and *J* equals to the number of intervals after concatenation.

4. Experiment 1

The primary aim of this experiment is to demonstrate how the best fitting ICC could be selected from 4 classes of parametric functions (4)–(6). Suppose that the ability level p has Beta distribution $B(3;3)$. Beta distribution is convenient to use for approximation of the ability level distribution because it's definition range is the interval [0; 1] and it could be either symmetric or not – depending on the parameter values. 3000 observations were randomly generated from the Beta distribution $B(3;3)$. Item responses drawn from Bernoulli distribution with probabilities $k_2(p;1)$ (arccotangent function (4) with parameter $a = 1$) were generated for $j = 1,2,...,3000$. The data were grouped into 31 equal length intervals according to the ability scores. The observed proportions of correct responses were calculated for each group. Then the iterative process of unknown parameter a value adjustment was made by minimizing distances $d(k_2)$. The same procedure was performed for functions from other function classes $k_1(p;a;b)$, $k_3(p;a;b)$, $k_4(p;a;b)$. The obtained values of parameter estimates, distances $d_{n_i}(k_i)$, $i = 1,2,3,4$, values of X^{2}statistic (8) for detecting misfit of the selected model and criterion values are presented in Table 1.

The best adjustment to the observed proportions of correct responses was achieved with the function $k_2(p;a)$ and parameter value estimate $\hat{a} = 0.96$. Very similar result was obtained with the modified logistic function $k_1(p;a;b)$ and parameter estimates $\hat{a} = 1.23$, $\hat{b} = 0.49$ and arccotangent function $k_4(p;a;b)$ with $\hat{a} = -0.04$, $\hat{b} = 1.02$. The adjustment of the best fitting function $k_3(p;a;b)$ to the observed data was worse, values of X^2 statistic exceeded criterion

Table 1. The best fitting function's $k_1(p)$ – $k_4(p)$ parameter values estimates, distances $d_{n_i}(k_i), i = 1,2,3,4,$ X^2 statistic values and criterion values, when p has Beta distribution $B(3;3)$ and actual ICC is generated from $k_2(p;1)$

	$\hat{a}$	$\hat{b}$	$d_{n_i}(k_i)$	X^2	**Criterion value**
$k_2(p;a)$	0.96	–	0.032 174	23.879 05	37.65*
$k_1(p;a;b)$	1.23	0.49	0.033 303	22.661 93	36.415*
$k_3(p;a;b)$	0.13	0.85	0.047 247	40.201 87	36.415*
$k_4(p;a;b)$	–0.04	1.02	0.034 40	20.612 24	36.415*

* there were joint intervals.

value, so the conclusion about the misfit of the model $k_3(p;a;b)$ is done. It is notable that due to the randomness of the experiment the best accuracy of the distance with the function from the class $k_2(p;a)$ wasn't reached with the true parameter value a = 1 though the number of observations is large (3000).

5. Mathematical model of experiment 2

Let us suppose that the probability distribution of knowledge level p is known:

$$P\left(p \le x\right) = \int_0^x f(p)dp, \text{ here } f(p) = 0 \text{ for } p \notin [0;1].$$

The model was applied to 4 classes of probability density functions $f(p)$: segments of linear functions $f_1(p)$, Beta distribution $f_2(p)$, Normal distribution when normalized in the interval [0; 1] $f_3(p)$ and histogram function $f_4(p)$. These functions are defined for $p \in [0;1]$ and their parameters are chosen in such way that functions satisfy two features of probability density functions:

$$f(p) \ge 0, \text{ for } \forall p \in [0;1], \tag{9}$$

$$\int_0^1 f(p)dp = 1. \tag{10}$$

Segments of linear functions $f_1(p)$ are represented by trapezium or triangle depending on the parameter values (Krylovas, Kosareva 2008a). Beta distribution probability density function initially satisfies features (9) and (10). The probability density function of normal distribution is restricted with $p \in [0;1]$ and multiplied by suitable normalize constant so that the feature (10) holds. Histogram is also defined for $p \in [0;1]$.

Graphs of functions $f_1(p)$ – $f_4(p)$ with various parameter values are represented in the Fig. 3.

If N test items are given to the examinee, the total test result S would be the number of correctly responded items. Random variable $S = \sum_{i=1}^{N} X_i$ is gaining values from 0 to N. The responses to the test items for the fixed latent ability p due to the condition ii) stated above are **independent** random variables X_i having Bernoulli distributions with probabilities $k_i(p)$, $i = 1,2,...,N$. So S has **generalized** Binomial distribution (Bagdonavicius, Kruopis 2007) with the probabilities $p_i(p) = P(S = i \mid p)$, $i = 0,1,...,N$, which are equal to the coefficients near the corresponding x degrees in the generating function polynomial of the random variable S:

$$\Psi(p;x) = \prod_{i=1}^{N}(1 - k_i(p) + k_i(p)x) = p_0(p) + p_1(p)x + p_2(p)x^2 + ... + p_N(p)x^N,$$

$$\text{here } p_0(p) = P(S = 0 \mid p) = \prod_{i=1}^{N}(1 - k_i(p)),$$

$$p_1(p) = P(S = 1 \mid p) = \sum_{i=1}^{N} k_i(p) \prod_{\substack{j=1 \\ j \ne i}}^{N}(1 - k_j(p)),$$

$$p_N(p) = P(S = N \mid p) = \prod_{i=1}^{N} k_i(p).$$

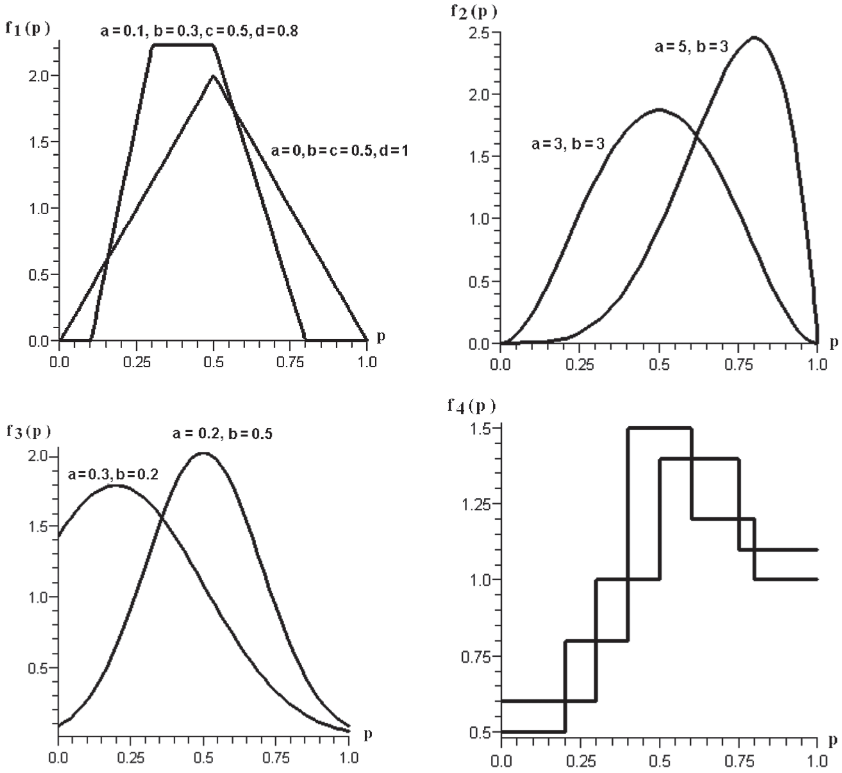

Fig. 3. Graphs of functions $f_1(p) - f_4(p)$

The probability distribution of total test result S in the whole population is received by integrating probabilities $p_i(p)$ multiplied by the probability density function $f(p)$:

$$p_i = P(S = i) = \int_0^1 p_i(p) f(p) dp, \quad i = 0,1,\ldots,N.$$

The test information function I is described as follows:

$$I(k_1, k_2, \ldots, k_n; f) = -\sum_{i=0}^{N} p_i \ln p_i \ . \tag{11}$$

The normalized value of the function I is the percentage of the test information function (11) from the maximum value, which is reached when all probabilities are equal to $P(S = i) = \frac{1}{N+1}, i = 0,1,2,\ldots,N$. Our purpose is to choose test items that maximize the value of the test information function.

6. Experiment 2

The probability distribution of total test result S and values of the test information function for distribution classes $f_1(p)–f_4(p)$. and various combinations of item characteristic functions (3)–(6) could be calculated according to the model. On the other hand the distribution of total test result S was obtained by Monte Carlo method. Data were simulated for M (700, 1200, 1800, 2500 and 3000) examinees with p values drawn from Normal distribution $N(0.2;0.2)$ normalized in the interval [0; 1]. Examinees were responding to 10 test items. Items were selected by choosing 2 or 3 diagnostic operators from each function class $k_1(p)$ – $k_4(p)$.

Let $d\left(P;\hat{P}_M\right)=\sum_{i=0}^{N}\left|p_i-\hat{p}_{iM}\right|$ be a measure of distance between the distributions obtained by the model (P) and by generating the responses of M examinees $\left(\hat{P}_M\right)$. 50 random samples were generated to calculate maximum, average values and standard deviations (STD) of the distances $d\left(P;\hat{P}_M\right)$ ($n=50$). The results are presented in Table 2. The same trend of the results was observed for other probability density functions $f_1(p), f_2(p), f_4(p)$ of latent ability and various combinations of item characteristic functions (3)–(6), so the conclusion about the stability of these results could be drawn.

Maximum values, averages and standard deviations of $d\left(P;\hat{P}_M\right)$ are decreasing as the sample size increases. It was shown in this data simulation example that the mathematical model describes real processes correctly. This model when applied for ICC functions chosen from 4 parametric classes of functions guarantees better approximation precision of the observed ICC function and as a consequence better accuracy of the probability distribution function of total test result S.

Probability distribution of total test result S and the value of the normalized test information function I (11) for the described data calculated by the model are presented in Table 3. The histogram of probabilities is shown in Fig. 4.

Table 2. Maximum values, averages and standard deviations of $d\left(P;\hat{P}_M\right)$ calculated for 10 item test with p drawn from Normal distribution $N(0.2;0.2)$, normalized in the interval $[0;1]$

Sample size M	**Average** $d\left(P;\hat{P}_M\right)$	**STD** $d\left(P;\hat{P}_M\right)$	$\max d\left(P;\hat{P}_M\right)$
700	0.088 9748	0.022 660 155	0.1527
1200	0.072 717 4	0.017 215 457	0.1165
1800	0.058 508 62	0.016 722 36	0.093 77
2500	0.046 498 24	0.013 388 334	0.079 48
3000	0.043 409 16	0.010 700 308	0.073 63

Table 3. Probability distribution of total test result S and the value of the normalized test information function I for 10 item test and p values drawn from normal distribution $N(0.2;0.2)$ normalized in the interval $[0;1]$

I	**Probability distribution of total test result *S***										
	0	1	2	3	4	5	6	7	8	9	10
0.92	0.1598	0.1978	0.1600	0.1193	0.0893	0.0713	0.0618	0.0532	0.0419	0.0298	0.0161

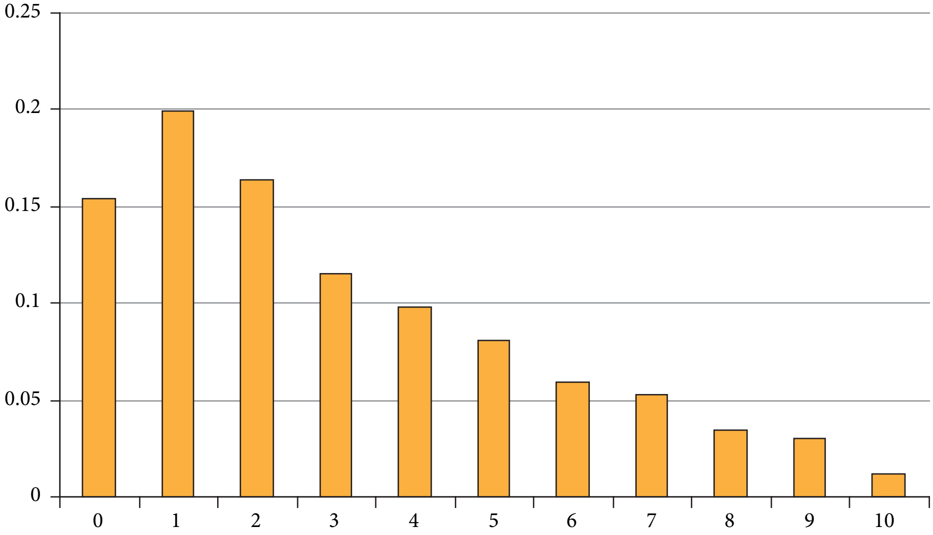

Fig. 4. The histogram of probabilities of total test result *S*

The test is sufficiently good for this population, the value of the test information function reaches 92% of its maximum value. Nevertheless, we can see from the histogram that the test is too difficult for this group of examinees as the probabilities of lower grades exceed the probabilities of higher grades. We can increase the value of the test information function by substituting difficult items with easier ones (for example, by reducing parameter's a value in $k_2(p;a)$) or substituting items with low discrimination parameter values with items that have higher values of this parameter. In this experiment parameters of 4 items were changed:

$$k_2(p;a): a = 0.5 \Rightarrow a = 0.3\,,$$

$$k_1(p;a;b): a = 0.3, b = 0.8 \Rightarrow a = 0.1, b = 0.8\,,$$

$$k_1(p;a;b): a = 0.2, b = 0.5 \Rightarrow a = 0.1, b = 0.5\,,$$

$$k_4(p;a;b): a = 0.3, b = 0.8 \Rightarrow a = 0.1, b = 0.6\,.$$

The results of the new test are presented in Table 4 and Fig. 5. The value of the improved normalized test information function is 0.96.

Table 4. Probability distribution of total test result S and the value of the normalized test information function I of the improved test

I	Probability distribution of total test result *S*										
	0	1	2	3	4	5	6	7	8	9	10
0.96	0.1408	0.1638	0.1339	0.1113	0.0940	0.0787	0.0684	0.0639	0.0612	0.0532	0.0307

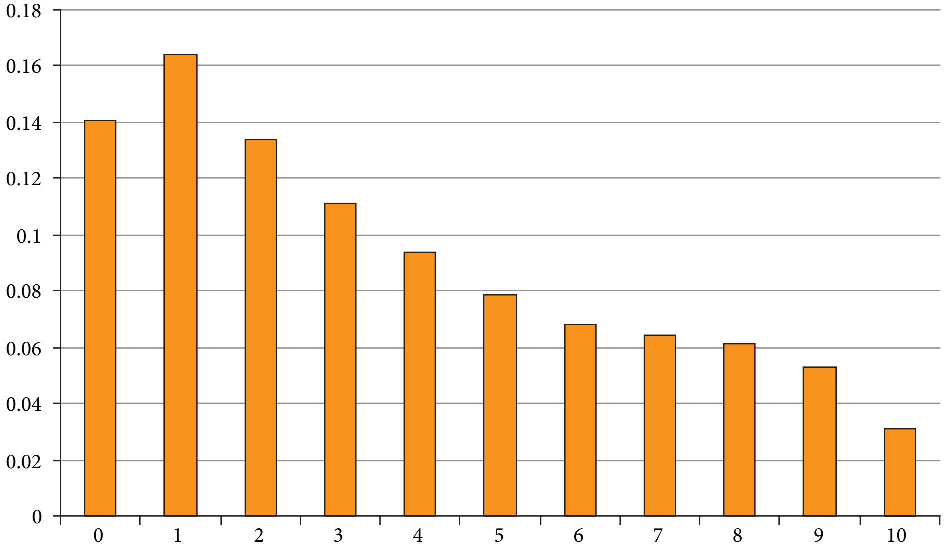

Fig. 5. The histogram of probabilities of total test result S for improved test

7. Case study: some examples of diagnostic operators for evaluation of microclimate in office rooms

We will now show how the proposed model could be applied for solving a practical decision-making problem. It is necessary to emphasize that the example is fitted only for demonstrative purposes and we do not try to reach very precise results. This is because of insufficient number of observations (14) and shortage of test length (only 3 items). In practice if one wants to obtain good precision in parameter calibration procedure the recommended number of observations is 500. Nevertheless our method gives suitably accurate results in this example.

In (Zavadskas, Turskis 2010) the problem of evaluation of microclimate in office rooms was solved by applying ARAS method for multicriteria decision-making. 6 microclimate evaluation parameters were analyzed in the paper: 1) air turnover inside the premises; 2) air humidity; 3) air temperature; 4) illumination intensity during work hours; 5) air flow rate; 6) dew point. According to these parameters and estimates of 38 experts comparison criterion of 14 office rooms (denoted by p) was calculated. In the example below 2 problems will be solved – office rooms will be grouped into clusters according to the test results and on the next step the comparison criterion p for 14 office rooms will be evaluated.

Let us denote NR – number of room, RH – relative air humidity, T – temperature, I – illumination intensity during work hours (parameters RH, T and I will be used to construct diagnostic operators), R – rank of the office room. All parameter values are presented in (Zavadskas, Turskis 2010). Notice that only three of six parameters which are given in Table 5 will be used.

Table 5. Measurement results in 14 rooms from (Zavadskas, Turskis 2010)

NR	RH (%)	T (°C)	I (lx)	p	R
1	46	18	390	0.671	4
2	32	21	360	0.656	6
3	32	21	290	0.627	10
4	37	19	270	0.632	9
5	38	19	240	0.546	14
6	38	19	260	0.558	13
7	42	16	270	0.566	12
8	44	20	400	0.772	2
9	44	20	380	0.773	1
10	46	18	320	0.6	11
11	48	20	320	0.677	3
12	48	20	310	0.663	5
13	49	19	280	0.633	8
14	50	16	250	0.651	7

Comparison criterion p was calculated using 6 parameters. As we see from Table 5, according to the 2 parameters RH and T, estimations of rooms 2 and 3, 5 and 6, 8 and 9, 11 and 12 will coincide. Let us construct the test from 2 dichotomous items K_{RH} (RH > = 44), K_T (T > = 20). In Table 6 the results of two items test are presented.

Therefore 2 items test lets us group office rooms into three clusters correctly enough. According to the criterion p, NR4 in the group TS2 = 0 must be substituted with NR10. NR12 in the group TS2 = 2 must be substituted with NR1. However there are not enough data in the 2 item test to distinguish NR11 from NR12.

With the intention to improve the test a third item K_I (I > = 320) will be added to it. The new 3 item test denoted by TS3 will give 4 clusters, represented in Table 7.

Table 6. The results of 2 items test K_{RH} (RH > = 44), K_T (T > = 20)

The number of positively responded items TS2	Rooms, which correspond to this value of TS2
TS2 = 0	NR4, NR5, NR6, NR7
TS2 = 1	NR1, NR2, NR3, NR10, NR13, NR14
TS2 = 2	NR8, NR9, NR11, NR12

Table 7. The results of 3 items test K_{RH} (RH > = 44), K_T (T > = 20), K_I (I > = 320)

The number of positively responded items TS3	Rooms, which correspond to this value of TS3
TS3 = 0	NR4, NR5, NR6, NR7
TS3 = 1	NR3, NR13, NR14
TS3 = 2	NR1, NR2, NR10, NR12
TS3 = 3	NR8, NR9, NR11

Group TS3 = 0 coincides with TS2 = 0. TS3 = 3 correctly includes 3 rooms NR8, NR9, NR11 possessing highest ranks. TS2 = 1 is spited into 2 groups TS3 = 1 and TS3 = 2. The next improvement is that NR12 goes to TS3 = 2. The only improvement required is to move NR10 from TS3 = 2 to TS3 = 0.

So, third item lets us distribute office rooms into 4 groups more precisely. However there is no reason to expect better results of ranking comparing with ARAS method, because only 3 of 6 parameters were used. Better results could be achieved by including additional items to the test.

The comparison of classification results obtained by 3 items test (4 clusters) and by ARAS method is represented in Fig. 6.

It is notable, that theoretical characteristics of diagnostic operators depend on the chosen mathematical model (Krylovas, Kosareva 2009b). It is important that this methodology has only natural restrictions on the shape of diagnostic operator that assures **the principle of diagnostic operator's validity** – subject with higher comparison criterion value has bigger probability to respond to the test item positively. Thresholds of diagnostic operators (44 for RH, 20 for T and 320 for I) are selected so that approximately one half of testees would positively respond to the item. In this case test information function (11) achieves it‘s maximum value.

The best approximation to the probability of positive response to the first item K_{RH} (RH> = 44) was selected from the class of two parabolas segments functions (6), where $a = 0.53$, $b = 0.75$. In Fig. 7 best approximations to the first item empirical distribution function P(RH> = 44|p) selected from 4 function classes (3)–(6) are represented.

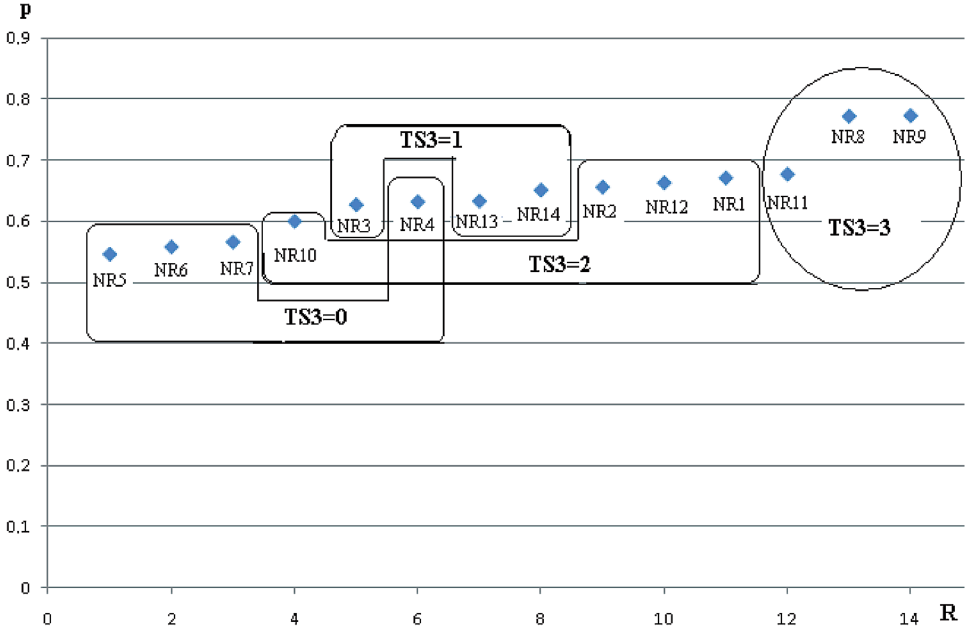

Fig. 6. The comparison of classification results obtained by 3 items test and ARAS method

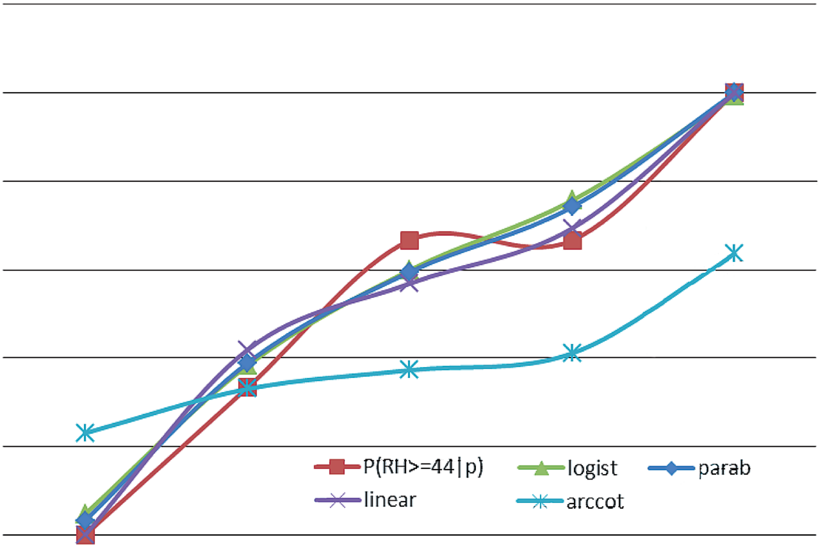

Fig. 7. Approximations to the function P(RH> = 44|*p*) from the logistic, two parabolas, segments of linear and arccotangent function classes

Distances $d_1(k_1) - d_4(k_1)$ were calculated for the data grouped in 5 intervals by the formula (7): $d_1(k_1) = 0.0590$, $d_2(k_1) = 0.4339$, $d_3(k_1) = 0.0594$, $d_4(k_1) = 0.0547$. The closest to the empirical distribution function is 2 parabolas function ($d_4(k_1) = 0.0547$) with parameter values $a = 0.53, b = 0.75$. It's graph is represented by dark blue line in Fig. 7.

Theoretical characteristics of the second K_T (T> = 20) and third K_I (I> = 320) diagnostic operators coincide, since coincide their empirical distribution functions. The best approximation was selected from the logistic function class (3), where $a = 120, b = 0.67$. So, we have calibrated three item characteristic functions $k_1(p) - k_3(p)$ and are ready to go to the next step.

Quantitative evaluation of comparison criterion value p could be made by applying maximum likelihood method (Harris, Stocker 1998). For each office room the observed response vector is (w_1, w_2, w_3). Here $w_i = 0$ in case of negative response to the item i and $w_i = 1$ in case the item i was answered positively. Maximum likelihood function is equal to the probability of the observed response vector:

$$l(p) = \prod_{i=1}^{3} k_i(p)^{w_i} (1 - k_i(p))^{1-w_i} .$$

Criterion value estimation $\hat{p}$ for each office room is the value that maximizes loglikelihood function (12), i.e. natural logarithm of likelihood function:

$$L(p) = \ln l(p) = \sum_{i=1}^{3} (w_i \ln k_i(p) + (1 - w_i) \ln(1 - k_i(p))). \qquad (12)$$

Estimated criteria values $\hat{p}$ and rank values $\hat{R}$ are presented in Table 8.

Table 8. Criteria values p, corresponding estimated values $\hat{p}$ and response vectors for 14 office rooms

NR	p	$\hat{p}$	w_1	w_2	w_3
1	0.671	0.67	1	0	1
2	0.656	0.68	0	1	1
3	0.627	0.67	0	1	0
4	0.632	0.54	0	0	0
5	0.546	0.54	0	0	0
6	0.558	0.54	0	0	0
7	0.566	0.54	0	0	0
8	0.772	0.72	1	1	1
9	0.773	0.72	1	1	1
10	0.6	0.67	1	0	1
11	0.677	0.72	1	1	1
12	0.663	0.67	1	1	0
13	0.633	0.66	1	0	0
14	0.651	0.66	1	0	0

Since responses to all 3 items are identical for rooms NR4, NR5, NR6, NR7 and for NR8, NR9, NR11, also for NR13, NR14, etc., estimated criteria values $\hat{p}$ coincide for rooms in these groups. Value of the Spearman's rank correlation coefficient between p and $\hat{p}$ equals 0.841. We can see that even 3 item test gives estimated criteria values accurate enough. Better results could be expected for tests with more items.

8. Conclusions

In this paper the investigation of mathematical model of forecasting the results of knowledge testing proposed by the authors in (Krylovas, Kosareva 2008a, b; 2009a) is continued. This model does not require apriori information about probability distribution of ability level in the population of examinees and allows selecting item characteristic functions from a variety of forms. This enables to apply the model for the different probability distribution functions which occur in practice.

In the paper the Monte Carlo experiments were performed to show the technique of evaluating the best fitting parameters of the model. The unknown parameters of the item characteristic function were selected by the method proposed by (Baker 2001). Then the best fitting function was chosen from four function classes. Values of X^2 goodness-of-fit statistic for detecting misfit of each model were calculated. The experiment demonstrated that parameters of the model could be steadily reconstructed using standardized statistical procedures when the number of numerical experiments is rather big. These results show that in cases when the number of real experiments was not very big, the proposed model would still enable one to construct efficient tests for attainment measuring. This is the object of the authors' further investigations.

In the second numerical experiment the responses to 10 items test with ICC from four function classes and increasing number of examinees were generated for normal-ability population. The probability distribution of total test result and test information function value were calculated. It was shown that the results of the test could be efficiently improved by selecting relevant parameters of ICC. Obviously, the set of ICC functions, from which we choose the best fitting one, could be expanded with other classes of parametric functions. It must be mentioned that the precision of the results depends not only on how good item characteristic curves are approximated but also on the precision of probability density function of p value approximation.

The proposed mathematical model could be used not only for knowledge testing but also for solving diagnostic tasks in various fields of human activities – medicine, sports, geology, technical diagnostics and others. As an example, evaluation of microclimate in office rooms was performed by applying this methodology.

References

Bagdonavicius, V.; Kruopis, J. 2007. *Mathematical Statistics*. Vilnius. 359 p.

Baker, F. 2001. The Basics of Item Response Theory. *ERIC Clearinghouse on Assessment and Evaluation*. University of Maryland, College Park, MD.

Birnbaum, A. 1968. Some latent trait models and their use in inferring an examinee's ability, in F. Lord & M. Novick (Eds.). *Statistical Theories of Mental Test Scores.* Reading, MA: Addison-Wesley, 397–479.

Douglas, J. & Cohen, A. S. 2001. Nonparametric item response function for assessing model fit, *Applied Psychological Measurement* 25(3): 234–243. doi:10.1177/01466210122032046

Eggen, T. J.; Verschoor, A. J. 2006. Optimal Testing with Easy or Difficult Items in Computerized Adaptive Testing, *Applied Psychological Measurement* 30(5): 379–393. doi:10.1177/0146621606288890

El-Alfy, E. M.; Abdel-Aal, R. E. 2008. Construction and Analysis of Educational Tests Using Abductive Machine Learning, *Computers & Education* 51(1): 1–16. doi:10.1016/j.compedu.2007.03.003

Harris, J. W.; Stocker, H. 1998. *Handbook of Mathematics and Computational Science.* New York: Springer-Verlag.

Hoshino, T.; Shigemasu, K. 2008. Standard Errors of Estimated Latent Variable Scores with Estimated Structural Parameters, *Applied Psychological Measurement* 32(2): 181–189. doi:10.1177/0146621607301652

Lanza, S. T.; Collins, L. M.; Lemmon, D. R.; Schafer, J. L. 2007. PROC LCA: A SAS Procedure for Latent Class Analysis, *Structural Equation Modeling: A Multidisciplinary Journal* 14(4): 671–694.

Lord, F. 1980. *Applications of Item Response Theory to Practical Testing Problems*. Lawrence Erlbaum Assoc.

Krylovas, A.; Kosareva, N. 2008a. Mathematical modelling of forecasting the results of knowledge testing, *Technological and Economic Development of Economy* 14(3): 388–401. doi:10.3846/1392-8619.2008.14.388-401

Krylovas, A.; Kosareva, N. 2008b. Mathematical model for knowledge testing, *Lithuanian Mathematical Journal* (48/49): 217–221.

Krylovas, A.; Kosareva, N. 2009a. The investigation of mathematical modeling of diagnostic test, *Lithuanian Mathematical Journal* (50): 202–207.

Krylovas, A.; Kosareva, N. 2009b. Mathematical modeling of diagnostic tests, in *Knowledge-Based Technologies and OR Methodologies for Strategic Decisions of Sustainable Development (KORSD-2009)*, 5th international conference, 120–125.

Mislevy, R.; Bock, R. 1982. *BILOG: Item analysis and test scoring with binary logistic models* [Computer Program]. Mooresville IN: Scientific Software.

Molenaar, I. 2007. Parametric and nonparametric item response theory models in health related quality of life measurement, in Mounir Mesbah, Bernard F. Cole ir Mei-Ling Ting Lee (Eds.). *Statistical Methods for Quality of Life Studies: Design, Measurements and Analysis*, 143–154.

Orlando, M.; Thissen, D. 2000. Likelihood-based item-fit indices for dichotomous item response theory models, *Applied Psychological Measurement* (24): 50–64. doi:10.1177/01466216000241003

Rasch, G. 1960. *Probabilistic Models for some Intelligence and Attainment Tests*. Copenhagen: Danish Institute for Educational Research. Expanded edition: 1980, Chicago: The University of Chicago Press. 199 p.

Rizopoulos, D. 2006. ltm: An R package for latent variable modeling and item response theory analyses, *Journal of Statistical Software* 17(5): 1–25.

Savalei, V. 2006. Logistic approximation to the normal: The KL rationale, *Psychometrica* (71): 763–767.

St-Onge, Ch.; Valois, P.; Abdous, B.; Germain, S. 2009. A Monte Carlo Study of Item Characterictic Curve Estimation on the Accuracy of Three Person-Fit Statistics, *Applied Psychological Measurement* 33(4): 307–324. doi:10.1177/0146621608329503

Uebersax, J. S. 1999. Probit latent class analysis with dichotomous or ordered category measures: conditional independence/dependence models, *Applied Psychological Measurement* 23(4): 283–297. doi:10.1177/01466219922031400

Van Barneveld, C. 2007. The Effects of Examinee Motivation on Multiple-Choice Item Parameter Estimates, *Applied Psychological Measurement* 31(1): 31–46. doi:10.1177/0146621606286206

Wells, C. S.; Bolt, D. M. 2008. Investigation of a Nonparametric Procedure for Assessing Goodness-of-Fit in Item Response Theory, *Applied Measurement in Education* 21(1): 22–40. doi:10.1080/08957340701796464

Zavadskas, E. K.; Turskis, Z. 2010. A new additive ratio assessment (ARAS) method in multicriteria decision-making, *Technological and Economic Development of Economy* 16(2): 159–172. doi:10.3846/tede.2010.10

Zavadskas, E. K.; Kaklauskas, A.; Turskis, Z.; Tamošaitienė, J. 2009. Multi-attribute decision-making model by applying grey numbers, *Informatica* 20(2): 305–320.

Zavadskas, E. K.; Turskis, Z.; Tamošaitienė, J. 2010. Risk Assessment of Construction Projects, *Journal of Civil Engineering and Management* 16(1): 33–46. doi:10.3846/jcem.2010.03

Zimowski, M. F.; Muraki, E.; Mislevy, R. J.; Bock, R. D. 1996. *BILOG-MG: multiple-group item analysis and test scoring.* Chicago: Scientific Software International.

ŽINIŲ TESTAVIMO PROGNOZĖS MATEMATINIO MODELIO TYRIMAS SKAITINIU EKSPERIMENTU

N. Kosareva, A. Krylovas

Santrauka. Šiame straipsnyje žinių tikrinimo rezultatų prognozės matematinis modelis, pasiūlytas ankstesniuose autorių darbuose, praplėstas keturiomis parametrinių funkcijų klasėmis, iš kurių parenkama tinkamiausia funkcija klausimo charakteristinei funkcijai aproksimuoti. Matematinis modelis vizualizuojamas atliekant du skaitinius eksperimentus. Pirmojo eksperimento tikslas buvo parodyti tinkamiausios klausimo charakteristinės funkcijos ir šio modelio parametrų parinkimo procedūrą. Siekiant nustatyti modelio tinkamumą, buvo skaičiuojama suderinamumo kriterijaus statistikos reikšmė. Antrajame eksperimente buvo sukonstruotas 10 klausimų testas populiacijai, turinčiai normalųjį žinių lygio skirstinį. Testo rezultatų tikimybinis skirstinys ir testo informacijos funkcijos reikšmė buvo apskaičiuojamos, kai klausimo charakteristinės funkcijos parenkamos iš keturių parametrinių funkcijų klasių su skirtingomis parametrų reikšmėmis. Kitame žingsnyje parodyta, kaip galima padidinti testo informacijos funkcijos reikšmę parenkant klausimo charakteristinių funkcijų parametrus, atitinkančius stebimą populiaciją. Šis modelis galėtų būti pritaikytas ne tik testuojant žinias, bet ir sprendžiant diagnostinius uždavinius įvairiose žmogaus veiklos srityse. Kitas šio metodo privalumas yra testavimo proceso sąnaudų sumažinimas mažinant vertinamo žinių lygio standartinę matavimo paklaidą. Pateiktas šios metodikos taikymo pavyzdys sprendžiant mikroklimato biuro patalpose vertinimo uždavinį.

Reikšminiai žodžiai: užduoties sprendimo teorija, matematinis modeliavimas, klausimo charakteristinė funkcija, generuojančioji funkcija, Monte Karlo metodas.

Natalja KOSAREVA. Dr, Associate Professor, Dept of Mathematical Modelling, Vilnius Gediminas Technical University. Doctor (mathematics, 1986). Research interests: mathematical modelling of attainment tests, mathematical statistics in education, information technologies.

Aleksandras KRYLOVAS. Dr (HP), Professor, Head of Dept of Mathematical Modelling, Vilnius Mykolas Romeris University. Doctor (mathematics, 1987), (HP – 2006). Research interests: mathematical modelling, asymptotic analysis, didactics of mathematics.