TECHNOLOGICAL AND ECONOMIC DEVELOPMENT OF ECONOMY
Baltic Journal on Sustainability
2008
14(3): 388–401

MATHEMATICAL MODELLING OF FORECASTING THE RESULTS OF KNOWLEDGE TESTING

Aleksandras Krylovas[1], Natalja Kosareva[2]

Vilnius Gediminas Technical University, Saulėtekio al. 11, LT-10223 Vilnius, Lithuania
E-mails: [1]aleksandras.krylovas@fm.vgtu.lt, [2]natalja.kosareva@ fm.vgtu.lt

Received 5 May 2008; accepted 5 September 2008

Abstract. In this paper a mathematical model for obtaining probability distribution of the knowledge testing results is proposed. Differences and similarities of this model and Item Response Theory (IRT) logistic model are discussed. Probability distributions of 10 items test results for low, middle and high ability populations selecting characteristic functions of the various difficulty items combinations are obtained. Entropy function values for these items combinations are counted. These results enable to formulate recomendations for test items selection for various testing groups according to their attainment level. Method of selection of a suitable item characteristic function based on the Kolmogorov compatibility test, is proposed. This method is illustrated by applying it to a discreet mathematics test item.

Keywords: testing, logistic model, item characteristic function, generating function, entropy, probability distribution.

Reference to this paper should be made as follows: Krylovas, A.; Kosareva, N. 2008. Mathematical modelling of forecasting the results of knowledge testing, *Technological and Economic Development of Economy* 14(3): 388–401.

1. Introduction

The theoretical and practical aspects of measuring the knowledge and other person's features are not only objects of common and discipline didactics (Bitinas 2002), but also objects of mathematical modelling. Mathematical background of the Classical Test Theory (Anastasi, Urbina 1997) is based on normal probability distribution. The assumption is made so, that any measured feature – tested person's knowledge level, IQ coefficient or acquired skills application efficiency, etc – is normally distributed. This enables to analyse deviations of this measured feature from some standard values (Hopkins 1998). However, when the number of tested people is not large, when not suitable formalized knowledge is measured, when not standardised, but teacher's created tests are applied, the suggestion about normal distribution is invalid.

ISSN 1392-8619 print/ISSN 1822-3613 online doi: 10.3846/1392-8619.2008.14.388-401
http://www.tede.vgtu.lt

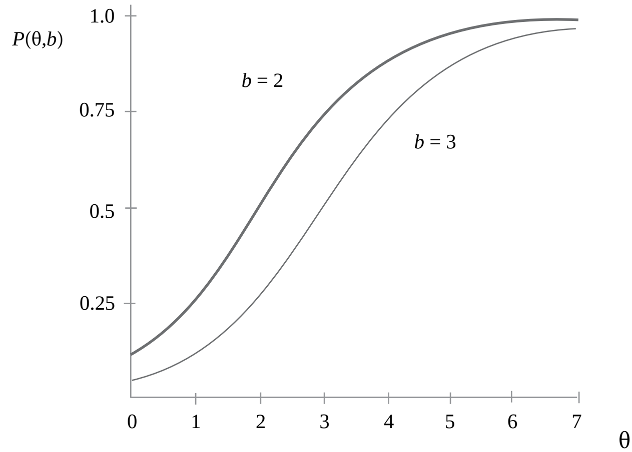

Fig. 1. Graph of function $P(\theta,b)$ for b =2; 3

In such cases the probabilistic test theory, also known as Item Response Theory (IRT), is invoked (Kiseliova, Kiseliovas 2004). The main task of this theory (Hambleton, Swaminathan, Rogers 1991) is, according to the testing outcome, to evaluate some persons' measured ability – latent trait of the model. This latent parameter could be an attainment level in a specific branch of science or internal person's feature, for example, anxiety, self-sufficiency, the ability to concentrate, etc. IRT enables to apply tests, which are corresponding to the abilities of tested people of different groups and different individuals and to gauge estimates of tested features on the same scale. We can choose individual assignments for the tested people: for stronger persons – more difficult, for weaker persons – easier.

One of essential features of IRT is that Item Characteristic Curve (ICC) parameters are invariable and they are independent of the probability distribution of investigative population parameters (Lord 1980; Reeve 2002). This feature is called *Invariance of Parameters* of item characteristic function. Estimations of item parameters can be derived from any tested population; in all cases the obtained item parameters estimations will be the same. Another any less important feature of IRT, opposite to Classical Test Theory, is that the estimation of measured latent parameter is independent of the difficuly of the selected assignment (Wright 1968; Slinde, Linn 1979). Thus, the basis of IRT is the assumption, that probability of correct response to the test item depends on the difficulty of test item and person's measured feature – *construct*. The author of this theory is a Danish mathematician and statistician George Rasch (Rasch 1960).

Let us suppose, that θ is tested feature and b is item difficulty. In Rasch model probability of a correct response to the test item $P(\theta,b)$ is expressed by formula

$$P(\theta,b)=\frac{1}{1+e^{b-\theta}}. \tag{1}$$

This model is called one-parameter logistic (1PL) model and function $P(\theta,b)$ – item characteristic function. In Fig. 1 function $P(\theta,b)$ graphs are presented with different

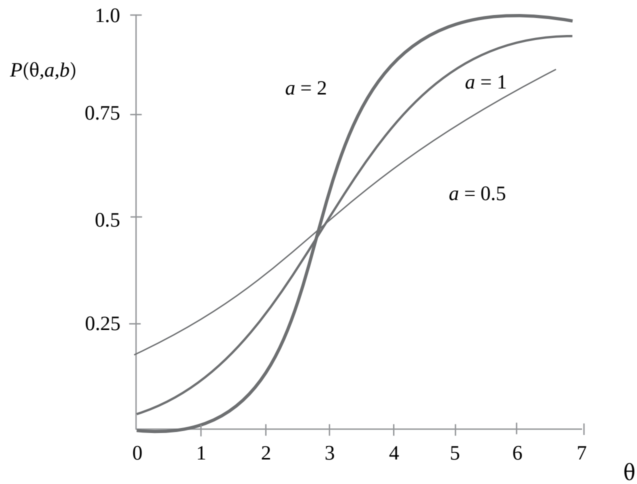

Fig. 2. Graph of function $P(\theta$, a, $b)$ for b = 3, a = 0.5, 1, 2

parameter b values (b = 2 and b = 3). When parameter b increases, function $P(\theta,b)$ graph is scrolling to the right preserving its shape. The tested person with a high ability is more likely to give the correct response to the test item than a person with a low ability, so item characteristic function (1) is not decreasing by the parameter θ. Notice, that latent model parameter θ and test difficulty b are reflected on abscissa axis. Test difficulty b is the point on abscissa axis, where item characteristic function gains the value of 0.5.

A. Birnbaum's two-parameter logistic (2PL) model and three-parameter logistic (3PL) model (Birnbaum 1968) are Rasch model generalization when supplementary parameters are introduced. These parameters are item discrimination (in 2PL model) and probability of guessing a correct response (in 3PL model). In 2PL model probability of a correct response to the test item $P(\theta,a,b)$ is calculated by formula

$$P(\theta,a,b)=\frac{1}{1+e^{a(b-\theta)}}. \tag{2}$$

Here θ is measured feature, b – item difficulty, a – item discrimination parameter. The less the value of the parameter a, the more sloping item characteristic function (2) is and the less item discrimination. In the Fig. 2 function $P(\theta,a,b)$ graphs are presented with the same value of item difficulty parameter (b = 3) and different parameter a values (a = 0.5; a = 1; a = 2). Item characteristic function (2) for any values of parameters a and b is not decreasing according to parameter θ, because with a higher attainment level of tested person, the probability of correct response to the item does not decrease.

The alternative models were also analysed in Item Response Theory. One of them is *normal ogive model,* when, instead of the logistic function (2), the distribution function of normal random value is used (Uebersax 1999):

$$P(\theta,a,b)=\Phi\left(\frac{\theta-b}{\sigma}\right). \tag{3}$$

Here mean of normal distribution b is corresponding to item‘s difficulty, and standard deviation θ is a reciprocal value of item‘s discrimination: $\theta = \frac{1}{a}$.

However, it is not convenient to use item characteristic function (3), therefore usually logistic functions (1) or (2) are chosen. Rapid development of Item Response Theory is associated with coming of powerful personal computers, which permit to automatise routine calculating procedures (Baker 2001, 1992). Programs BICAL (Wright, Mead 1976), LOGIST (Wingersky *et al.* 1982), BILOG (Mislevy, Bock 1986), performing selection of test items of various difficulty, evaluation of item parameters, evaluation of tested person‘s latent parameter, were created.

The newest articles on IRT are concerned with computerized adaptive tests, i.e. individualized tests that are optimal to 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).

Distinctly from Classical Test Theory (CTT) and from probabilistic test theory (IRT), in this paper we are not aiming to evaluate the investigated feature of tested persons. We will analyse the possibility to maximize the amount of information, obtained during knowlegde testing. This aim is achieved when distribution of obtained test points is not normal, but, contrarily, is not even close to it. Maximum amount of information is obtained, when possible test points 0, 1, ..., n are achieved with the same probability $p = \frac{1}{n+1}$. This test is oriented towards the effective *norms-referenced* knowledge assessment, when comparative attainment of tested people is measured (Girdzijauskas 1999), i.e. when weaker testees have to be distinguished from stronger ones. Traditional assessment, called *criterion-referenced*, when the student‘s assimilated part of program is measured, this value in the article will be denoted $p \in [0;1]$.

The first problem to be solved in this paper is to construct the effective knowledge test. The other our set task was to increase the test reliability by choosing separate test items. The third problem – having particular test item to select suitable item characteristic function, i.e. to determine, to which group of items – easy, middle or hard (there may be more groups of this kind) – this item is referred.

We propose a mathematical model, which does not have requirements for function shape, to forecast the results of knowledge testing. But we restrict ourselves to the *segments of linear functions*, which are near to the logistic functions. The main difference between these functions is when function values are near 0 or 1. Segments of linear functions are more easily applied in practice because of the simplicity of their analytical expression.

2. Mathematical model

Let us suppose, that $p \in [0;1]$ is the attainment level of testee, i.e. some criterion-referenced knowledge estimation. Norms-referenced knowledge assessment of testees will be performed with *n* items test, denoting that $k_j(p)$ is the probability that testee, whose knowledge level is *p*, will give the correct response to *j*-th test item. Let us select any of the set of non-decreasing functions $k_j(p):[0;1] \to [0;1]$, which could be test items characteristic functions and construct *n* items collection (test) *T*:

$$T = \{k_1(p), k_2(p), \ldots, k_n(p)\}. \tag{4}$$

Let us denote that *K*(*p*) is the number of correct responses to the test items by the testee, whose attainment level is *p*. *K*(*p*) is a discreet random variable obtaining values $m = 0,1,\ldots,n$ with probabilities

$$P(K(p) = m) = t_m(p), m = 0,1,\ldots,n. \tag{5}$$

Supposing that function *f*(*p*), described as

$$f(p) = \begin{cases} f_0(p) \geq 0, p \in [0;1] \\ 0, p \notin [0;1] \end{cases} \qquad \int_0^1 f_0(x)dx = 1,$$

is probability density function of testees population knowledge level, distribution of number of correct responses to the test items in the whole population *K* could be found by the formula

$$P(K = m) = p_m = \int_0^1 t_m(p) f_0(p)dp, \tag{6}$$

here $m = 0,1, \ldots, n$.

Willing to find probabilities $t_m(p)$ (4), we will use the essential suggestion of Item Response Theory – *local independence assumption*, i.e. we will consider that when knowledge level value *p* is fixed, responces to the different test items are **independent** random values. Suppose, that a student, whose knowledge level is *p*, is answering *n* test items. Then the number of correct responses *K*(*p*) is the sum of independent random variables with Bernoulli distribution. Therefore, random variable *K*(*p*) generating function (Kruopis 1993) is equal to product:

$$\psi_n(p,x) = \prod_{j=1}^{n}\left(1 - k_j(p) + k_j(p)x\right) = t_0(p) + t_1(p)x + \ldots + t_n(p)x^n. \tag{7}$$

Random variable generating function has an important feature: the coefficients, near various *x* degrees in the generating function polynomial (7), are equal to probabilities of obtaining correspondent values by this random variable. From (6) and (7) we can find the distribution of test results *K* in the whole population, i.e. the probabilities p_m, $m = 0,1, \ldots, n$.

The amount of information, given by the test *T*, is expressed by entropy (Stakenas 1996):

$$E\left(k_1, k_2, \ldots, k_n, f_0\right) = \sum_{m=0}^{n} p_m 1n \frac{1}{p_m}. \quad (8)$$

Entropy function achieves its maximum value, when $p_0 = p_1 = \ldots = p_n = \frac{1}{n+1}$, i.e. when the number of testees giving the correct responses to the 0, 1, ..., n test items will be approximately equal. For example, when test is made of one item, entropy functions value will be maximal, when approximately the same number of testees will give correct and incorrect responses to the item.

Differently from IRT, where logistic functions (1), (2) are used, we will describe item characteristic functions as the segments of linear functions:

$$k_j\left(p, \alpha, \beta\right) = \begin{cases} 0, p < \alpha \\ \frac{p-\alpha}{\beta-\alpha}, \alpha \le p \le \beta \\ 1, p > \beta \end{cases}. \quad (9)$$

In the Fig. 3 graphs of functions k_E (p, 0, 0.5), k_M (p, 0, 1), k_H (p, 0.5, 1) are presented. These functions are called respectively characteristic functions of easy, middle and hard item and enable us to describe *fuzzy* sets of items (Zadeh 1965; Ustinovichius *et al.* 2006). These characteristic functions are the analogues of 2PL model functions $P\left(\theta, a, b\right)$ (2), because parameters α and β describe not only the difficulty of test item $\frac{\alpha+\beta}{2}$, but also discrimination $\frac{1}{\beta-\alpha}$. The particular values of parameters α and β are found performing experiments. The procedure of finding model parameters is called *item calibration*.

Let us describe density function of testees population knowledge level $f_0\left(p\right)$ as the segments of linear functions:

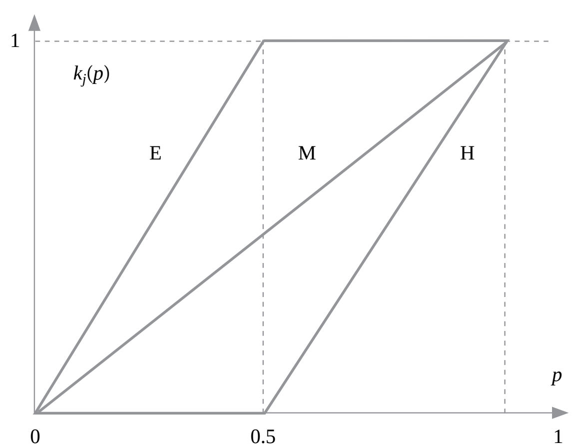

Fig. 3. Different difficulty items characteristic functions k_E (p, 0, 0.5), k_M (p, 0, 1), k_H (p, 0.5, 1)

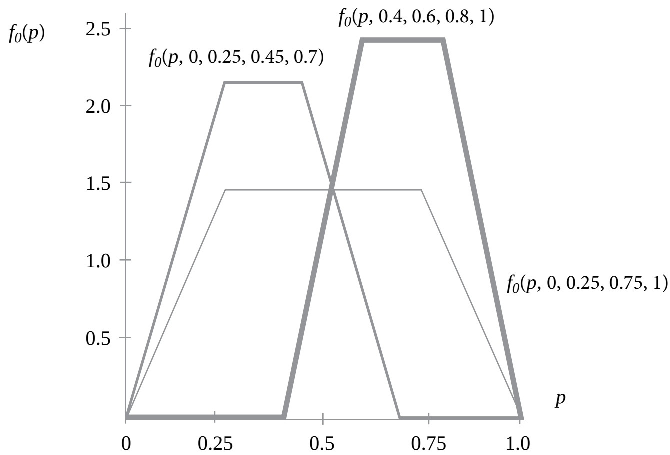

Fig. 4. Knowledge level density function of low ability (f_0 (p, 0, 0.25, 0.45, 0.7)), middle ability (f_0 (p, 0, 0.25, 0.75, 1)) and high ability (f_0 (p, 0.4, 0.6, 0.8, 1)), population

$$f_0\left(p,a,b,c,d\right)=\begin{cases}0,\ p\le a\\ M\cdot\dfrac{p-a}{b-a},a<p<b\\ M,b\le p\le c\\ M\cdot\dfrac{d-p}{d-c},c<p<d\\ 1,\ p\ge d\\ M=\dfrac{2}{d-a+c-b}\end{cases} \tag{10}$$

In general case this function has the form of trapezium; in some cases it may have the form of triangle. Parameters a, b, c, d for each specific population are found by performing experiments. For example, in the Fig. 4 knowledge level density functions for 3 populations – having low, middle and high abilities, are represented.

3. Selection of the best test

Let us describe testees populations of 3 types – having low f_0 (p, 0, 0.33, 0.33, 0.67), middle f_0 (p, 0.25, 0.5, 0.5, 0.75) and high f_0 (p, 0.33, 0.67, 0.67, 1) abilities. To investigate this model the program which counts probability distribution of test results and entropy function values for low, middle and high abilities populations with various item characteristic functions of easy k_E (p, 0, 0.5), middle k_M (p, 0, 1), and hard k_H (p, 0.5, 1) items combinations was created. The calculations were performed with 10 items test, according to formulas (7) and (8). In Fig. 5 the normalized entropy function's graphs for various 10 items combinations with 3 testees populations are presented. For example, the mark 235 on the axis of abscissas expresses that test is constructed from 2 easy, 3 middle and 5 hard items. Items are ranged on

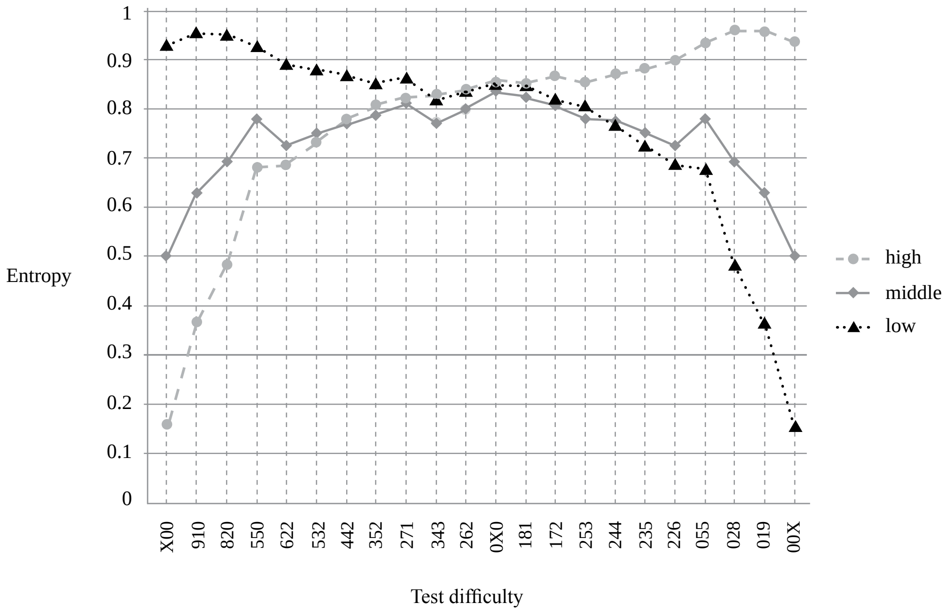

Fig. 5. Entropy functions for 10 items tests for the populations of low, middle and high attainment level

the axis of abscissas according to increasing difficulty of the test: from 10 easy items (X00) to 10 hard items (00X).

Entropy function value for the population of low abilities was the biggest (96 % of maximum value), when test was constructed of 9 easy and 1 middle items (910) and of 8 easy and 2 middle items (820). For the population of high abilities, the same entropy value was reached when test was constructed of 9 hard and 1 middle items (019) and 8 hard and 2 middle items (028). For the population of middle abilities, the highest entropy value (from 73 % to 80 % of maximum value) was obtained with many item characteristic functions combinations – 442, 334, 262, 253, 235, ...

This program allow us to forecast the distribution of the number of correct responses to the test items for populations of low, middle and high abilities and differently selected test items collections. The examples of such forecasts for 100 testees of middle abilities are presented in Table 1.

Table 1. Distribution of correct responses for the population of 100 people of middle attainment for test of 10 different difficulty items

Test difficulty	**Distribution of correct responses to 10 items test**										
	0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
820	0	0	0	1	2	3	6	10	22	35	21
442	0	0	2	5	11	21	27	21	10	3	0
037	12	29	28	15	8	5	2	1	0	0	0

4. Maximising the amount of information obtained from the test

Let us suppose, that test T is constructed of 1 item, the item characteriscic function is $k(p)$, testees population knowledge level density function is $f_0(p)$. The question is – for which values of item characteristic function entropy function value (and also information amount obtained from test) will be the biggest? We shall compare the results when function $k(p)$ is segments of linear functions (9) and when it is a logistic function, described by equations (1) and (2).

Let us suppose, $k(p)$ is described by Eq. (9). Then maximum entropy value will be reached, when α and β satisfy the equation

$$\int_0^1 k(p)\, f_0(p)\, dp = \frac{1}{2}. \tag{11}$$

If knowledge level density function of testees population $f_0(p)$ has the form of trapezium (10) and is symmetric in point of the mean Ep of the population parameter p, whose values are in the interval [0; 1], then the solutions of the equation (11) will be all such functions $k(p)$ for which the test item difficulty coincides with the mean of population parameter p:

$$Ep = \frac{\alpha+\beta}{2}. \tag{12}$$

The equation (11) has an infinite number of solutions satisfying the equation (12). When we fix test discrimination $\frac{1}{\beta-\alpha}$, we will obtain the equation with one unknown parameter – test difficulty $\frac{\alpha+\beta}{2}$.

If the test consists of 2 items, whose characteristic functions are respectively $k_1(p)$ and $k_2(p)$, the biggest entropy will be reached for solving the system of equations:

$$\begin{cases} \int_0^1 k_1(p) k_2(p)\, f_0(p)\, dp = \frac{1}{3} \\ \int_0^1 \left(k_1(p) + k_2(p)\right) f_0(p)\, dp = 1 \end{cases}. \tag{13}$$

Likewise, in the 3 items test case, the biggest entropy will be reached when solving the system of equations:

$$\begin{cases} \int_0^1 k_1(p) k_2(p) k_3(p)\, f_0(p)\, dp = \frac{1}{4} \\ \int_0^1 \left(k_1(p)k_2(p) + k_1(p)k_3(p) + k_2(p)k_3(p)\right) f_0(p)\, dp = 1 \\ \int_0^1 \left(k_1(p) + k_2(p) + k_3(p)\right) f_0(p)\, dp = \frac{3}{2} \end{cases}.$$

Now, suppose that $k(p)$ is a logistic function, described by equation (1). Then in case of one item test maximum entropy value will be obtained, when parameter b satisfies the equation

$$\int_{-\infty}^{+\infty} \frac{1}{1+e^{b-r}} f_0(r)\,dr = \frac{1}{2}\cdot \tag{14}$$

If the distribution of population knowledge level density function $f_0(r)$ is symmetric in point of the mean *Er* of the population parameter *r*, for example, in case of normal probability distribution, entropy function achieves its maximum value when the item difficulty coincides with the mean of population parameter *r*:

$$Er = b. \tag{15}$$

We got the same result for two-parameter logistic model. We can see that the characteristic function, described by Eq. (9) and logistic function (1) and (2) features, are similar in this case.

5. Evaluating the difficulty of the item

For practical applying the model, we must establish to which set of items – easy, middle or hard – the test item should be referred. Let us deal with an example of evaluating one item. Assume, that sets of items are described by the formula (9), where parameters α and β are selected like in the paragraph 3: k_E (*p*, 0, 0.5), k_M (*p*, 0, 1), k_H (*p*, 0.5, 1). We will deal with one discrete mathematics test item, proposed in 2006–2007 spring semester interim examination taken by 405 VGTU students (Krylovas, Suboc, Kosareva 2007: 254–258). The students must determine, which of two proposed Boolean functions‘ systems is full (Krylovas 2005). The following Boolean functions systems were proposed:

Table 2. 10 variants of discrete mathematics test item. Question: which of two Boolean functions systems is full?

$\{\oplus,1,\vee\};\{\Rightarrow,1\}$	$\{\oplus,\vee\};\{\oplus,1,\&\}$
$\{\oplus,\Leftrightarrow,\vee\};\{\vee,\neg\}$	$\{\Rightarrow,\Leftrightarrow\};\{\oplus,\Leftrightarrow\}$
$\{\Rightarrow,\neg\};\{\oplus,\Leftrightarrow\}$	$\{\vee,\&\};\{\Rightarrow,0\}$
$\{\vee,\neg\};\{\oplus,1,\vee\}$	$\{\Rightarrow,\&\};\{\Rightarrow,1\}$
$\{\downarrow\};\{\Rightarrow,\vee\}$	$\{\oplus,\vee\};\{\mid\}$

It was established (Krylovas *et al.* 2007a, b: 249–253), that these 10 test items variants are equivalent by the difficulty. Therefore, we can join these variants and deal like with one test item. We call such test items *parallel variants* (Krylovas, Raulynaitis 2003). Let us divide the interval [0; 1] to *n* equal length intervals and assign all tested students to these intervals according to their discreet mathematics knowledge level. In each interval we will count the values of empirical item characteristic funcion $\hat{F}_n(p)$, as a proportion of correct responses to the item and all observations, which belong to this interval. Function $\hat{F}_n(p)$ has the feature of distribution function $0 \le \hat{F}_n(p) \le 1$, when $0 \le p \le 1$. However, it is not always non-decreasing.

Kolmogorov test lets us verify the hypotheses about coincidence of function $\hat{F}_n(p)$ and respectively item characteristic functions k_E, k_M and k_H. Suppose, that n_i tested students belong to the i-th interval, and m_i of them gave the correct responses to the test item. Kolmogorov test statistic (Kruopis 1993), which is used to test the hypothesis $H_0 : \hat{F}_n(p) \equiv k(p)$ is:

$$D_n = \max_{0 \le p \le 1} |\hat{F}_n(p) - k(p)|.$$

We can count statistic D_n in the following way. Primarily we count statistics

$$D_n^+ = \max_{i=1,2,\dots,n} \left[\frac{m_i}{n_i} - k\left(\frac{i}{n}\right) \right],$$

$$D_n^- = \max_{i=1,2,\dots,n} \left[k\left(\frac{i}{n}\right) - \frac{m_{i-1}}{n_{i-1}} \right], \text{ here } \frac{m_0}{n_0} = 0 .$$

Then $D_n = \max\left(D_n^+, D_n^-\right)$. Hypothesis H_0 is rejected when $D_n > D_n(0.05)$. Here $D_n(0.05)$ is statistic D_n probability distribution critical value with the significance level of 0.05. If n = 10, then $D_n(0.05) = 0.40925$.

We obtained these values of Kolmogorov statistics: $D_{E,10} = 0.46$; $D_{M,10} = 0.28$; $D_{H,10} = 0.68$, hence we cannot reject hypothesis $H_0 : \hat{F}_n(p) \equiv k_M(p)$, while hypotheses $H_0 : \hat{F}_n(p) \equiv k_E(p)$ and $H_0 : \hat{F}_n(p) \equiv k_H(p)$ are rejected with significance level 0.05. Conclusion – the considered item may be assigned to the middle difficulty items having characteristic function $k_M(p, 0, 1)$.

In Table 3 all observations grouped into 10 intervals are presented. Data are grouped according to the attainment level of tested students. m_i is the number of correct responses to the test item by students in the i-th interval; n_i – number of all testees in the i-th interval:

Table 3. Data of correct responses to the test item, grouped in 10 intervals according to the attainment level of tested students

Interval	m_i	n_i	$\frac{m_i}{n_i}$
[0; 0.1]	9	43	0.1731
]0.1; 0.2]	13	23	0.3611
]0.2; 0.3]	10	27	0.2703
]0.3; 0.4]	14	23	0.3784
]0.4; 0.5]	22	19	0.5366
]0.5; 0.6]	37	12	0.7551
]0.6; 0.7]	43	6	0.8776
]0.7; 0.8]	32	2	0.9412
]0.8; 0.9]	30	0	1.0000
]0.9; 1]	40	0	1.0000

6. Conclusions

The proposed mathematical model to forecast the results of knowledge testing allow us to construct norms-referenced estimation tests, optimally applied to the knowledge level of tested population. The novelty of this investigation is the applying segments of linear functions as item characteristic functions and also as population knowledge level density function. However, this shape of functions is not the restriction of the model and it is selected because of being convenient to apply in practice.

Computational experiments were performed when the knowledge level of tested population is low, middle or high, and test items – easy, middle or hard. Distribution of knowledge level of tested population and item characteristic functions were chosen as segments of linear functions. This program enables to increase the test reliability by choosing separate test items. It is important that we can select individual test, which is the best for particular testees population and forecast the results of knowledge testing. This allows to improve the process of knowledge testing.

Practical evaluation of model parameters (item calibration) is accomplished by mathematical statistics methods. In this paper it was shown, how we can assign the specific discrete mathematics test item to one of three item sets (easy, middle or hard items).

This model may be used for other grouping the population and item difficulty levels or another number of fuzzy sets. However, the determination of such sets requires gathering and analysing the empirical data and this will be the object of further author‘s investigation.

It is intended to analyse the stability of model in regard to function parameters and also by replacing segments of linear functions with other functions in the future.

References

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.

Anastasi, A.; Urbina S. 1997. *Psychological testing* (7th edition). Prentice Hal. 721 p.

Baker, F. B. 1992. *Item Response Theory: Parameter Estimation Techniques*. Marcell Dekker.

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*, 397–479. Reading, MA: Addison-Wesley.

Bitinas, B. 2002. *Basics of Educational Diagnostics*. Vilnius Pedagogical University. 200 p.

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.

Girdzijauskas, S. 1999. *Student‘s knowledge control and evaluation*. Vilnius University. 38 p.

Hambleton, R. K.; Swaminathan, H. and Rogers, H. J. 1991. *Fundamentals of Item Response Theory*. Sage Publications, Newburg Park.

Hopkins, K. D. 1998. *Educational and Psychological Measurement and Evaluation*. – 8th ed. Boston. Walsh & Associates, Inc. 486 p.

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

Kiseliova, D.; Kiseliovas, A. 2004. *Diagnostics of Mathematical Skills*. Research Monograph. Shiauliai University Publishing. First Book. 412 p. Second Book 484 p.

Kruopis, J. 1993. *Mathematical Statistics*. Vilnius: Mokslo ir enciklopedijų leidykla. 416 p.

Krylovas, A. 2005. *Discrete Mathematics*. Vilnius.

Krylovas, A.; Raulynaitis, J. 2003. The experience in parallelism of a probability theory task, *Lithuanian Mathematical Journal*, spec. issue 43: 357–360.

Krylovas, A.; Suboc, O.; Kosareva, N. 2007a. Statistical analysis of students marks, obtained by different methods, *Lithuanian Mathematical Journal*, spec. issue 47: 254–258.

Krylovas, A.; Suboc, O.; Kosareva, N. 2007b. Analysis of equivalence of parallel variants in discrete mathematics, *Lithuanian Mathematical Journal*, spec. issue 47: 249–253.

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. M. *Applications of Item Response Theory to Practical Testing Problems*. Lawrence Erlbaum Assoc.

Mislevy, R. J. and Bock, R. D. 1986. PC-BILOG 3: *Item Analysis and Test Scoring with Binary Logistic Models*. Mooresville, IN: Scientific Software, Inc.

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.

Reeve, B. B. 2002. *An Introduction to Modern Measurement Theory*. National Cancer Inst.

Slinde, J. A.; Linn, R. L. 1979. A note on vertical equating via the Rasch model for groups of quite different ability and tests of quite different difficulty, *Journal of Educational Measurement* 16: 159–165.

Stakenas V. 1996. *Information Coding*. Vilnius University. 127 p.

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.

Ustinovichius, L.; Zavadskas, E. K.; Migilinskas, D.; Malewska, A.; Nowak, P.; Minasowicz, A. 2006. Verbal analysis of risk elements in construction contracts, in *Cooperative Design, Visualization, and Engineering: third International Conference, CDVE 2006*, Mallorca, Spain, September 17–20, 2006: Proceedings. Lecture Notes in Computer Science 4101: 295–302.

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.

Wingersky, M. S.; Barton, M. A.; Lord, F. M. 1982. *LOGIST: A computer program for estimating examinee ability and item characteristic curve parameters (LOGIST 5, version 1)*. Princeton NJ: Educational Testing Service.

Wright, B. D. 1968. Sample free test calibration and person measurement, in *Proceedings of the 1967 Invitational Conference on Testing Problems*, 85–101.

Wright, B. D.; Mead, R. J. 1976. *BICAL: Calibrating Items with the Rasch Model*. Research Memorandum No. 23. Statistical Laboratory, Department of Education, University of Chicago.

Zadeh, L. A. 1965. Fuzzy sets, *Information and Control* 8(3): 338–353.

ŽINIŲ TIKRINIMO REZULTATŲ PROGNOZĖS MATEMATINIS MODELIAVIMAS

A. Krylovas, N. Kosareva

Santrauka

Straipsnyje pasiūlytas matematinis modelis žinių tikrinimo rezultatų tikimybiniam skirstiniui gauti. Aptarti šio modelio ir užduočių sprendimo teorijos (IRT) logistinio modelio skirtumai ir panašumai. Išnagrinėti 10 klausimų testo rezultatų tikimybiniai skirstiniai silpnai, vidutinei ir stipriai testuojamųjų populiacijoms parenkant įvairias testo klausimų sunkumo funkcijų kombinacijas. Apskaičiuotos entropijos funkcijos reikšmės. Gauti rezultatai leidžia formuluoti rekomendacijas testo klausimams parinkti skirtingoms testuojamųjų grupėms pagal jų žinių lygį. Pasiūlytas tinkamiausios klausimo charakteristinės funkcijos parinkimo būdas, grindžiamas Kolmogorovo kriterijumi. Ši procedūra iliustruojama taikant ją konkrečiam diskrečiosios matematikos testo klausimui.

Reikšminiai žodžiai: testavimas, logistinis modelis, klausimo charakteristinė funkcija, generuojančioji funkcija, entropija, tikimybinis skirstinys.

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

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