Fuzzy assessment of test results

O. Sokolov[1] **O. Molchanova**[2]

National Aerospace University
Dept. of Informatics
Chkalov Street 17
61070 Kharkov, Ukraine
[1]Email: oleksandr_sokolov@yahoo.com
[2]Email: molchanova@mail.ru

Abstract

The objective of this research is to study various assessment methods used in modern test systems. Methods based on classical test theory and IRT (item response theory) have been considered in this paper. The approaches to improve the quality of assessment of the ability level as a function of the diversity of evaluation points have been proposed. Statistical analysis packages SPSS, the tests analysis WinSteps and MatLab packages have been applied as software tools.

Keywords: item response theory, ability, difficulty, assessment scales, logit

1. Introduction

Development of information technologies has contributed significantly to the teaching methods and students assessment [1]. Tutorial programs, interactive tests, different monitoring studies and state programs for automated evaluation of knowledge have been introduced recently.

Testing has been applied widely in distance education and during implementation of Bologna system for student's self-education. Automated testing application has been expanded to the manufacturing, where personnel management is transformed into a continuous process of training (of course, with the subsequent testing and assessment of trainees). The distinct feature of such systems is that the role of teacher in the process of learning and assessment is much narrower, and the results have been evaluated automatically. The latter has been caused by the requirement for simultaneous estimation of a large number of trainees, and by the ideology of automated learning itself – self-consistent learning and independent evaluation. One of the major tasks is the comparability of the results of different tests, ranging of students level of knowledge, formation of the final scoring for the test sets. Use of so-called "raw" scores, i.e., totals for the successful implementation of items resulting from the test might be applied to the very limited extent (if testing is limited to the identifying of the level of knowledge on particular topic and can not be integrated with other results). Effectiveness of the test score depends not only on the quality of the test, but also on the methods of comparison and interpretation of primary (raw) score of test group [2].

Therefore one can assume as important to analyze the existing methods of comparison and integration of scores of various tests, study the quality of the students group assessment, understanding the diversity of evaluation points as a quality criterion for estimating methods. All these issues have been addressed in this article.

1.1. Assessment in the classical test theory

The classical test theory is based on the concept of the raw scores converting into a unified scale on the basis of the baseline information type analysis.

In accordance with the typology of pedagogical measurements, the types of scales can be conveniently presented as a hierarchy, proposed by S.Stevens (Fig. 1).

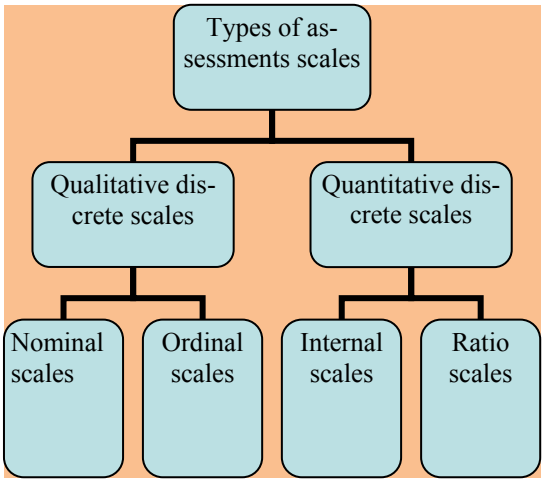


Fig. 1. Typology of scales by S. Stevens

Selection of estimation method depends on the purpose of testing analysis and source data characteristics. Basically, two types of evaluation can be applied - ranking and criterial ones. The task of ranking is focused on students positioning, without making a conclusion how much one student is better than other. For this purpose, qualitative characteristics are sufficient. Criterial evalua-

tion is focalized on the comparability of results and can be resolved using quantitative scales only. Despite of long (nearly a century) period of use, the classical theory of testing, and recommended linear transformations of raw scores allow to improve students comparability, however do not change the nature of an ordinal scale. The scales which have been considered below are related to such transformations.

Primarily, this is a Z-score, which is based on conversion of "raw" score r

$$z = \frac{r - M}{\sigma},$$

where M and σ - the expectation and root-mean-square deviation of entire assembly respectively.

This scale has several advantages, including the capacity to compare variations in the "strong" and "weak" groups.

A large number of other linear transformations is based on the Z-scale - IQ – (100+15z), CEEB – (500+100z), Veksler – (10+3z), T-score – (50+10z). Each of these scales has been applied in practical pedagogical measurements and features of different interpretations.

A group of methods based on the percentile transformation of raw scores provide the best comparability of results within the classical testing theory. Thus, to determine the students relative position in the group one can set his rank in percentiles - the fraction of students in the group, whose results are not above the primary score of the student. Despite of apparent advantage - the opportunity to rank the relative position of student at the scale set, a comparison of different samples (i.e. different tests, set of sessions on one subject, etc.) has been complicated by the fact that the percentile distribution is closely related to the frequency distribution of the same sample [3].

Unfortunately, even percentile assessments are difficult to compare with each other if they are received on different samples. The obvious solution of this problem is sample standardization, which is contraction all scores to a common scale. To do this, all empirical density frequency distributions of raw scores are converted to the same "reference" distribution - the normal distribution with given mean and variance. Usually centered and normalized normal distribution is applied.

This method has been named as equal-percentile normalization (ESP) and is applied in many countries for knowledge assessment. For example, in Ukraine, the method anticipates transition of raw scores to the scale [100,200], for which a reference distribution N(150,15) has been utilized.

Thus, we will apply ESP as the main method in the comparative analysis.

1.2 Assessment in the item response theory

Item response theory (IRT) is based on the study of interrelationship of item difficulty, student's abilities and probability of correct answer. The basic model, reflecting this interrelationship is named after G. Rush.

Success in the item resolving has probabilistic nature. Let us assume that the probability that the student resolves the item correctly (success rate) is determined as a function of the student level of knowledge a and level of item difficulty d :

$$p = p(a,d) = \frac{a}{a+d} = \frac{a/d}{1+a/d} = \frac{\lambda}{1+\lambda}. \qquad (1)$$

A symbol λ means the ratio of training and challenges latent parameters.

Equation (1) is actually a Rasch model, according to which the probability of success does not depend on each argument, but on their relationship. Let us study some properties of this function.

The unit measurement for readiness and difficulties is the same. If to assign unit difficulty d0=1 (the same can be applied to unit ability a0=1), then the difficulty of all items can be compared to the unit difficulty and ability can be compared in the same way.

(If the difficulty d of the item is less than 1, then it is d times easier than the unit difficulty, if its value is more than 1, then, respectively – d times more difficult).

It is obvious that $d, a, \lambda \in [0, \infty), p \in [0,1]$

If, $\lambda \to 0, \frac{a}{d} \to 0, p \to 0$ then the student is totally unprepared and is will never be able to complete the item.

If, $\lambda \to \infty, \frac{a}{d} \to \infty, p \to 1$ then the student, whose level of ability is many times higher than the difficulty of the item, will pass the test successfully for sure.

Function arguments (1) can not be measured directly, but the value of the function, i.e. the probability is available for measurement on test results. This is the basic idea of the IRT theory - a known probability to estimate the difficulty of items and level of student's ability. Based on the type of function (1) it is obvious that this problem has no correct solution. Inverse function allows to determine the parameter λ only on measured value *p*, i.e.,

$$\lambda = \frac{p}{1-p} = \frac{p}{q} \qquad (2)$$

and found only the ratio of latent parameters of ability and difficulty. If we have reference item with the unit difficulty then one can identify corresponding value of ability and position it on the measurement scale. This is a second fruitful IRT idea, allowing to solve the problem of various populations standardization.

2. Logistic Rasch function

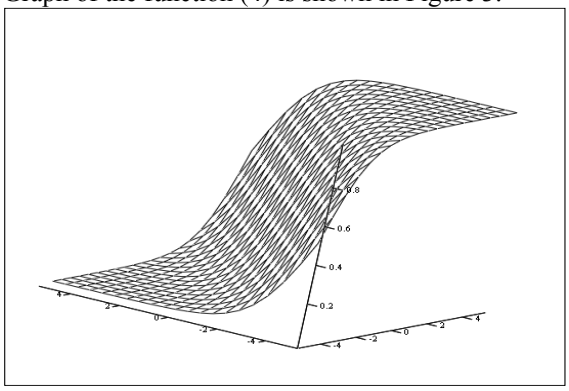
In practice it is convenient to express the level of ability arguments and difficulties not in a linear but a logarithmic scale

$$\ln a = \theta, \ln d = \delta \Leftrightarrow a = e^{\theta}, d = e^{\delta}. \quad (3)$$

The function of success takes the form

$$p = \frac{e^{\theta}}{e^{\theta} + e^{\delta}} = \frac{1}{1 + e^{-(\theta - \delta)}}. \quad (4)$$

Formula (4) is called the basic logistic model Rush. Graph of the function (4) is shown in Figure 3.



p

Fig. 3. Rasch logistic function

Arguments of abilities and difficulty $\theta, \delta \in (-\infty, \infty)$ are measured on a scale with a unit of 1 logit.
It is obvious that

$$\frac{1}{1 + e^{-1}} = 0.731$$

that means that 1 logit distance increases the probability of success in 0.731 times.

2.1. Assessment of latent parameters on the basis of raw scores

Let us consider the testing scheme including *k* items. Let us assume that *n* people are involved in testing. The result of each item accomplishment is assessed by dichotomous principle.

Let us consider $R = (r_{ij})$ as a response matrix ($i = 1, \ldots, n; j = 1, \ldots, k$).

Elements r_{ij} are random variables: they take the value 1 with probability

$p_{ij} = p(\theta_i, \delta_j)$.

Calculation of the raw scores of participants and items gives:

$$b_i = \sum_{j=1}^{k} r_{ij}, i = 1, \ldots, n, \quad (5)$$

$$c_j = \sum_{i=1}^{n} r_{ij}, j = 1, \ldots, k. \quad (6)$$

It is obvious that the set of values $b_i \in \{0, 1, \ldots, k\}$, i.e. total number of participants, can be divided strictly in $k+1$ groups by the "raw" score value.
Let us rewrite formulas (2) and (4) as follows,.

$$\lambda = \frac{p}{1-p} = \frac{p}{q}, \; p = \frac{e^{\theta}}{e^{\theta} + e^{\delta}}, q = \frac{e^{\delta}}{e^{\theta} + e^{\delta}},$$

then

$$\lambda = e^{\theta - \delta}. \quad (7)$$

Note that the symbol λ means the ratio of latent parameters of ability and difficulty in a linear scale. Logarithm of (7) gives the discrepancy of latent parameters of ability and difficulty θ, δ in the logit scale.
Let us assign this parameter as follows

$$l = \ln \lambda = \theta - \delta. \quad (8)$$

It is obvious that for each participant and each item one can determine the value of

$$l_{ij} = \theta_i - \delta_j, i = 1, \ldots, n; j = 1, \ldots, k \quad (9)$$

if corresponding probabilities are not equal to 0 or 1.
Moreover, the value l_{ij} can be measured on the basis of raw scores. Unknown quantities in the system of equations (9) are the parameters of the level of ability and difficulty θ_i, δ_j
Since all the participants who get the same "raw" score, have the same ability level of training (according to Rasch model), the number of equations is reduced from $n \times k$ to $(k+1) \times k$ that if $n \gg k$ gives a substantial reduction of the system of equations.

Let us denote as n_b the number of participants which got the same "raw" score $b = 0, 1, \ldots, k$.

Let us denote $\Theta(b)$ as a level of group abilities.

Then the system of equations (9) can be rewritten as:

$$l_j(b) = \Theta(b) - \delta_j, \quad (10)$$

$j = 1, ..., k; b = 1, \ldots, k-1$.

Cases $b = 0, b = k$ mean that either the participant did not fulfill any one item, or fulfilled all items k.

For these groups, the calculation of $l_j(b)$ values must be performed on a special procedure.

For the rest values $l_j(b)$ they can be defined as follows:

$$l_j(b) = \ln \frac{p_j(b)}{q_j(b)} \quad (11)$$

where - $p_j(b)$ relative frequency of correct solutions of the *j* item by the participants who obtained the same "raw" score.

. Taking into account, for example, $l_j(0) = -5, l_j(k) = 5,$ and possible similar cases in other groups b, with the possible values $p_j(b) = 0, p_j(b) = 1,$ we have an inconsistent sys-

tem $k\times(k+1)$ of equations (10), containing $2k+1$ that unknown values.

In this case, the coefficient matrix is singular and its rank is $2k$, i.e. the number of independent equations is less than the number of unknown parameters. Therefore, one of the values must be assumed free, expressing through it the remaining unknowns. This value specifies the origin point of the scale. It is convenient to combine the origin point with a mean value Θ of the θ parameter in logits.

System of equations (10) can be resolved in several ways - through a system of normal equations, using method of moments or maximum likelihood method. These methods are described in detail in [3] and applied in the WinSteps software to estimate the latent parameters.

As it was previously noted, despite of individual correlation of the item difficulty with the level of ability of the participant which is declared in the name of the theory, in practice, we have the same number of student's subgroups, getting the same score in the logit scale as in "raw" scores. That is, the canonical IRT theory does not improve the quality of estimation as a function of the diversity of initial scores, while the rest of its advantages are obvious. Thus, it looks actual to improve the theory in the direction of increasing the diversity of participants with different test scores. Solution of this problem can be obtained only when the participants who got the same score (even in the logit scale), are then redistributed within subgroups according to the difficulty of items, for example, in lexicographical order, either through conversion of ability levels for fixed values of the items difficulty. These modifications of the canonical theory of IRT are offered in this paper.

2.2 Improving the quality of assessment in modern test theory

Suppose that in solving the system of equations (10) were obtained values for levels of ability in groups $b=0,1,\ldots,k$, and levels of tests difficulty δ_j, $j=1,\ldots,k$.

Consider a group with the same level of abilities, $\Theta(b^*)=\theta_{n_1}=\ldots=\theta_{n_{\kappa^{\delta^*}}}$,

where $n_1,\ldots,n_{\kappa^{\delta^*}}$ are the number of participants in this group. For these participants the number of items performed correctly, is the same (their "raw" scores are equal, i.e., for dichotomous tests the number of correct answers and items is equal as well). But the items themselves may be different.

Let us sort the group $\left\{n_1,\ldots,n_{\kappa^{\delta^*}}\right\}$ in lexicographical order, i.e. first place in the group will take a participant, who accomplished the most difficult item with the highest level of difficulty in this group. If there are several such participants, the same procedure has been applied but for the second on the difficulty item, and so on.

Thus, all participants within the group will be ranked based on the difficulty of items.

Such solution is acceptable to rank in the framework one test. If one needs to change the quantitative values of the student's abilities level, an iteration procedure based on the obtained earlier Rasch model parameters for this test and the change in indicators of the level of ability for fixed values of the difficulty of the item has been proposed.

Along with the one-parameter Rasch model (4), a two-parameter model has been more widely used in the testing practice in the following form:

$$p(\theta)=\frac{1}{1+e^{-a(\theta-\delta)}} \qquad (12)$$

in which the probability of the student success with the level of ability θ is defined as the difficulty of the item δ and the discriminatory properties of the item a - the coefficient of discrimination.

We propose for all members of the group with the same level of ability $\Theta(b^*)$ to conduct an iterative refinement procedure of ability by the formula, which is commonly used in the method of moments together with the formula for clarification of items difficulty:

$$\hat{\theta}_{j,s+1}=\hat{\theta}_{j,s}+\frac{\sum_{i=1}^{k}a_i\left(r_{j,i}-p\left(\hat{\theta}_{j,s}\right)\right)}{\sum_{i=1}^{k}a_i^2\left(p\left(\hat{\theta}_{j,s}\right)\left(1-p\left(\hat{\theta}_{j,s}\right)\right)\right)}, \qquad (13)$$

$$j=n_1,\ldots,n_{\kappa^{\delta^*}},\ s=0,1,2,\ldots$$

This iterative procedure is performed for all elements of the set $\Theta(b^*)=\left\{\theta_{n_1},\ldots,\theta_{n_{\kappa^{\delta^*}}}\right\}$, whose elements are the same at the beginning of the iterative procedure, ie $\theta_{n_1,0}=\ldots=\theta_{n_{\kappa^{\delta^*},0}}=\Theta(b^*)$.

The disadvantage of lexicographic method is obvious - it cannot be applied for the formation of an extended table of scores within one group.

The disadvantage of the iterative formula (13) is its bulkiness and the possibility to receive divergent processes. Furthermore, the perception of this iterative procedure on the phenomenological level is rather complicated. At the same time the basic principle in testing is the clarity and accessibility in the interpretation of scores by both the students and other parties related to testing.

Therefore, it looks practical to apply an approach based on the accrual of additional points for complex items or removal of points for the simple ones. But this is not quite understandable and obvious. Obviously, one indicator of item difficulty for supporting the decision to increase or decrease of scores within the group is not enough. Two-parameter Rasch model implies that after constructing a model based on a table of raw scores the

success probability depends on item discrimination and its difficulty.
Therefore, the technique of using fuzzy model to form transformation,

$$F : a, \delta \to \Delta\theta \qquad (14)$$

defined on the set of values $a > 0, \delta \in [\delta_{\min}, \delta_{\max}]$ that determines encouragement or punishment of the student in relation to his score.
Transformation (14) is convenient to present as a set of rules, getting into the linguistic variables defined on the respective domains. For example,
if *discriminatory item* a *is High* ***and*** *item* δ *is difficult,* ***then*** *encouraging is considerable.*
The following example demonstrates the application of the set of rules.

3. Example

Let us consider the example of test results analysis for the group of 13 persons ($n = 13$) with a test consisting of three items ($k = 3$).The response matrix $R = (r_{ij})$ is shown in Table 1.This test divided groups into two categories – with "raw" scores 1 and 2.
Consider the application of classical EPN method.

Table 1

№ member	Item 1	Item 2	Item 3	total score
1	0	1	1	**2**
2	1	0	1	**2**
3	1	1	0	**2**
4	1	0	0	**1**
5	1	0	0	**1**
6	1	0	0	**1**
7	1	0	0	**1**
8	1	0	0	**1**
9	0	1	0	**1**
10	0	1	0	**1**
11	0	1	0	**1**
12	0	0	1	**1**
13	0	0	1	**1**

Frequency analysis of the results has been presented in Table 2. Also in this table the translation into a scale of 100-200 on the basis of the standard features package MatLab - norminv has been presented.

Table 2

Raw score	frequency	Cum. frequency	Percentile	EPN
0	0	0	0%	100
1	10	76,9%	38%	145
2	3	100%	88%	167
3	0	100%	100%	200

Consider the application of Winsteps software to construct a two-parameter Rasch model (12) and analyzing the test using IRT (Table 3).

Table 3

№ member	Item 1	Item 2	Item 3	Raw score	level of ability (logit)
1	0	1	1	2	**0,72**
2	1	0	1	2	**0,72**
3	1	1	0	2	**0,72**
4	1	0	0	1	**-0,72**
5	1	0	0	1	**-0,72**
6	1	0	0	1	**-0,72**
7	1	0	0	1	**-0,72**
8	1	0	0	1	**-0,72**
9	0	1	0	1	**-0,72**
10	0	1	0	1	**-0,72**
11	0	1	0	1	**-0,72**
12	0	0	1	1	**-0,72**
13	0	0	1	1	**-0,72**
	-0,57	**0,1**	**0,47**	The difficulty of the item (logit)	

If to assign three logit to 200 points , a linear transformation of the logit scale to the scale of 100-200 results in the logit can be accomplished using the formula

$$150+16{,}667 \cdot \text{logit} \qquad (14)$$

and the resulting values - 162 and 138 points respectively.

However, as in the case of EPN application, the IRT does not change the number of participants tested with the same score. This example demonstrates that the level of items difficulty is different, and it is obvious that this factor should be taken into consideration when ranking participants with the same points. Let us apply the proposed methods to improve the quality of estimation.
With the lexicographic ordering of the participants, obtained 2 raw scores, it is easy to see that the following order will be correct: $1 \succ 2 \succ 3$ since the participant № 1 resolved the most difficult item with 0.47 logit and following item with the difficulty in the 0.1 logit, while the participant № 2 – more simple combination.
With regard to participants who got the raw score of 1, then the order would be:

$$13 \equiv 12 \succ 11 \equiv 10 \equiv 9 \succ 8 \equiv 7 \equiv 6 \equiv 5 \equiv 4$$

This approach ranks the participants within the groups, however does not add quantitative information to the resulting scores. Let us consider the application of the iterative procedure (13).
As a result of Rasch model constructing (12) in the WInSteps package the following values of the coefficients for items discrimination have been obtained

$$a_1 = 0{,}43; a_2 = 1{,}04; a_3 = 1{,}18.$$

The initial values of the levels of abilities are the last column of Table 3.

The results of the iterative procedure (13), starting from step 2, have been shown in Table 4.

Table 4

№ Account	Iteration 2	Iteration 3	Iteration 4	Iteration 5
1	1,6575	1,8891	1,9096	1,9097
2	0,6856	0,6857	0,6857	0,6857
3	0,4625	0,4683	0,4683	0,4683
4	-1,1946	-1,5062	-1,5442	-1,5448
5	-1,1946	-1,5062	-1,5442	-1,5448
6	-1,1946	-1,5062	-1,5442	-1,5448
7	-1,1946	-1,5062	-1,5442	-1,5448
8	-1,1946	-1,5062	-1,5442	-1,5448
9	-0,2226	-0,1996	-0,1997	-0,1997
10	-0,2226	-0,1996	-0,1997	-0,1997
11	-0,2226	-0,1996	-0,1997	-0,1997
12	0,0004	0,0240	0,0240	0,0240
13	0,0004	0,0240	0,0240	0,0240

The listing of calculation in the MatLab package is presented in Figure 4,

```
Persons=1;Items=3;
R=[1 1 0];% Change this raw for each
D1new=[0,72];                %case
D1=D1new;
D=[];
for k=1:10 %Number of iterations
   D1=D1new;
D=[D,D1];

a=[0,43 1,04 1,18];
b=[-0,57 0,1 0,47];
for j=1:Persons
nn=0;
for i=1:Items
   p=1/(1+exp(-a(i)*(D1(j)-b(i))));
   nn=nn+a(i)*(R(j,i)-p);
end;
dn=0;
for i=1:Items
   p=1/(1+exp(-a(i)*(D1(j)-b(i))));
   dn=dn+a(i)*a(i)*p*(1-p);
end;
dDj=nn/dn;
D1new(j)=D1(j)+dDj;
end;
end;
```

Fig. 4. The program listing for iterative procedure (13)

Using formula (14), we can present the results obtained in a single table

Table 5

№ member	total score	EPN	IRT	Variant IRT
1	**2**	200	162	182
2	**2**	200	162	161
3	**2**	200	162	158
4	**1**	161	138	124
5	**1**	161	138	124
6	**1**	161	138	124
7	**1**	161	138	124
8	**1**	161	138	124
9	**1**	161	138	146
10	**1**	161	138	146
11	**1**	161	138	146
12	**1**	161	138	150
13	**1**	161	138	150

It should be noted that, after the iterative modification of the points, the participants with a lower initial score did not reach the level of points of the higher group, i.e. the segregation within one group does not intersect with other groups, which does not contradict the notion of estimation validity.

Fig. 5 shows the graphs constructed for two-parameter Rasch model for the test.

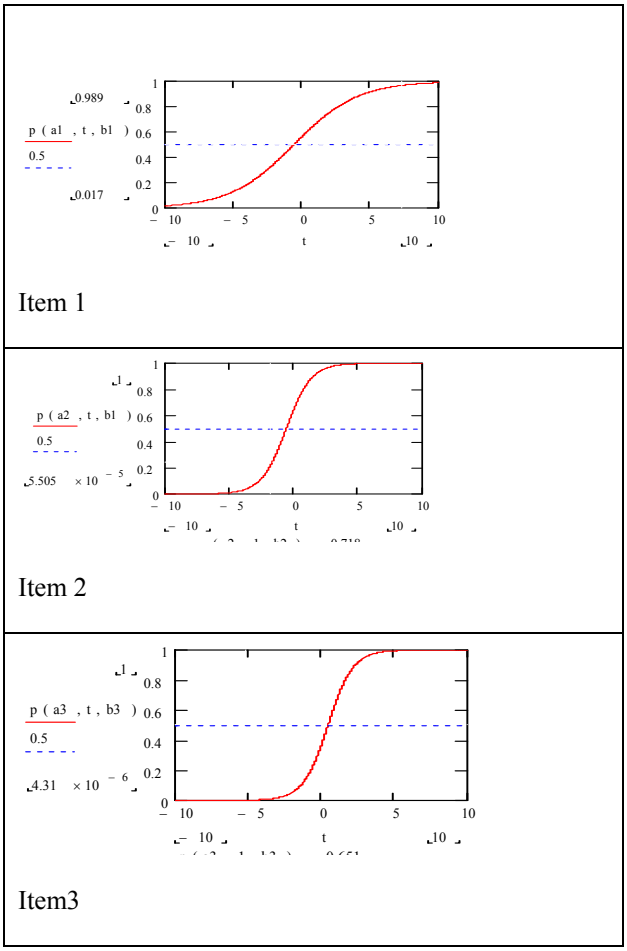


Fig. 5. Rasch Model

Obviously, for high discrimination difference a one logit difference causes higher difference in the likelihood in the mid of the range than in the case of low discrimination difference.

$$p_1(1) = 0.663;\ p_1(0) = 0.561;$$
$$p_2(1) = 0.718;\ p_2(0) = 0.474;$$
$$p_3(1) = 0.651;\ p_3(0) = 0.365.$$

The value of difficulty determines the bias of curves along the x-axis. Consider the transformation (14) for the formation of new values in the groups. Then the rules can be defined as follows:

If *the item discrimination* a *is "high",* ***and*** *the item difficulty* δ *is " complex",* ***then*** *the encouragement – is "high" ;*

If *the item discrimination* a *- is "average",* ***and*** *the item difficulty* δ *is "medium difficulty"* ***then*** *encouragement is "absent";*

If *the item discrimination* a *- is "low"* ***and*** *the item difficulty* δ *is "simple",* ***then*** *penalty is "big".*

All the linguistic variables are defined on appropriate domains. Fig. 6 shows the attribute function for the variable difficulty of the item.

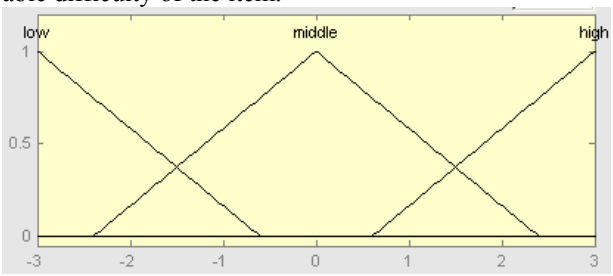


Fig. 6. Linguistic variable "difficulty of the item"

Figure 7 shows Mamdani Model for the set of rules for encouragement formation.

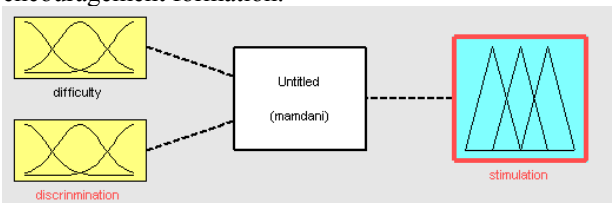


Fig. 7. A Rule Set Model

The inference mapping is shown in Fig.8.

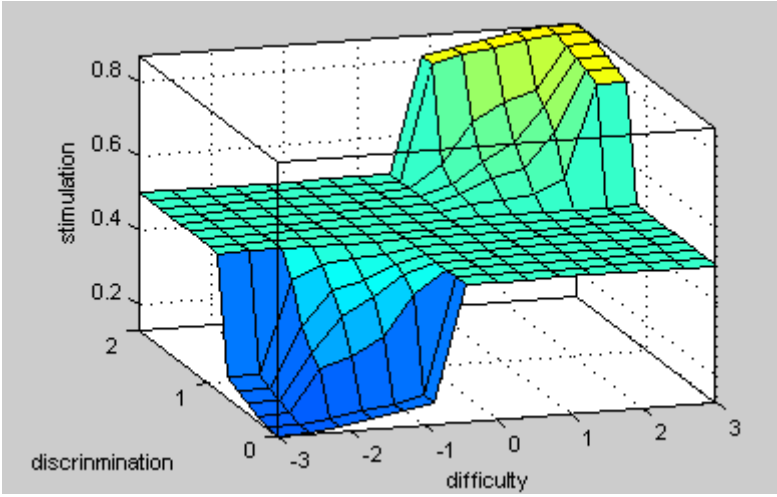


Fig.8. Mapping (14)

Conclusion

The application of classical test theory and modern IRT theory allows to improve the performance of comparability of test results. However, the best results can be achieved by modifying a modern test theory by ranking within the groups with the same score based on the difficulty of items. A fuzzy score modification model, that combines the properties of the estimates ranking lexicographic method and iterative procedure for calculating of additions to the estimates in the group have been proposed.

References

[1] Modern means of assessment of learning outcomes / H, H, Samylkina - M: Bean, Knowledge Lab, 2007-172p. (in Rus.)

[2] Chelyshkova M, B, Theory and practice of designing pedagogical tests: Textbook - M: Logos, 2002 – 432p. (in Rus.)

[3] Introduction to Modeling and parameterization of pedagogical tests / S, M, Neumann, B, A, Khlebnikov,-M: Moscow, 2000 – 168p. (in Rus.)

[4] Frank B, Baker, The basics of item response theory, – Univ, of Wisconsin,2001, – 185p,

[5] Chiu-Keung Law. Using fuzzy numbers on educational grading system/Fuzzy Sets and Systems 83 (1996) p.311-323.