TASK QUARTERLY 15 No 2, 191–202

METHODS FOR REFINING ASSESSMENT OF TEST-TAKERS BASED ON ITEM RESPONSE THEORY MODEL

OLEKSANDR SOKOLOV[1], OLGA MOLCHANOVA[1] AND WIESŁAW URBANIAK[2,3]

[1]*Department of Informatics, National Aerospace University, Chkalova 17, 61070 Kharkiv, Ukraine*
oleksandr_sokolov@yahoo.com, molchanova@mail.ru

[2]*Faculty of Mathematics, Physics and Technical Sciences, Kazimierz Wielki University, Weyssenhoffa 11, 85-072 Bydgoszcz, Poland*
wurban@ukw.edu.pl

[3]*Faculty of Informatics and Mechatronics, University of Economy, Garbary 2, 85-229 Bydgoszcz, Poland*

(Received 2 February 2011; revised manuscript received 5 March 2011)

Abstract: This research is devoted to assessment methods used in different test systems, including e-learning systems. The methods considered here are based on classical test theory and item response theory (IRT). We propose a new approach for improving the quality of assessment by diversifying evaluation points.

Keywords: test-taker, item response theory, ability, difficulty, assessment scale, logit

1. Introduction

Testing is widely applied in distance education and in students' self-education [1]. Automated testing applications 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). In these systems, the role of a teacher in the learning and assessment process becomes less significant, and the results are evaluated automatically. The latter is dictated by the need for simultaneous assessment of a large number of trainees, and by the possibility of automated learning, which offers self-consistent learning and independent evaluation. The major tasks in these are the comparability of the results of different tests, ranging of students' level of knowledge, and preparation

of a final scoring system for test sets. The so-called raw scores can be considered as simplest ones in the educational assessment and applied in the limited extent. (*i.e.* when testing is limited to identifying the level of knowledge of a particular topic and thus cannot be integrated with other results). The effectiveness of a test score depends not only on the quality of the test, but also on the methods for comparing and interpreting primary (raw) scores of test groups [2].

Therefore, it is important to analyze the existing methods of comparison and integration of scores of various tests, as well as to study the quality of assessment of student groups, while taking into account the variety (spectrum) of possible scores (or evaluation points) as a quality criterion for assessing methods. We address all these issues in this article.

1.1. Classical test theory

Classical test theory is based on converting raw scores into a unified scale using baseline information analysis.

In accordance with the typology of pedagogical measurements, scales can be conveniently presented as a hierarchy, *e.g.* as proposed by Stevens (Figure 1).

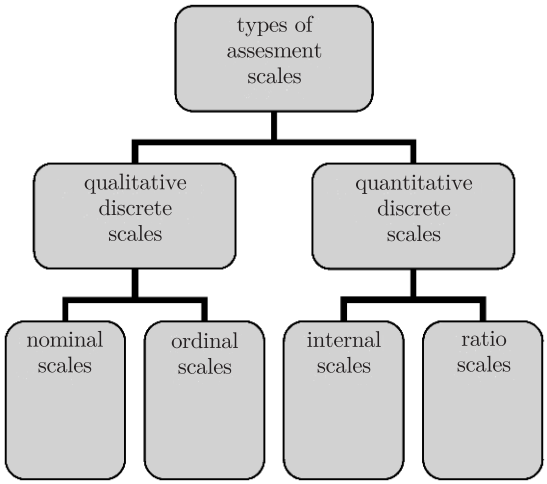

Figure 1. Typology of scales by Stevens

The selection of an assessment method depends on the purpose of the test and on the characteristics of source data. Basically, two types of evaluation can be applied – ranking and criterial evaluation. The purpose of ranking is to order the scores of students, without concluding to what degree one student is better than another. For this purpose, qualitative characteristics are sufficient. Criterial evaluation is focused on the comparability of results and can be carried out using solely quantitative scales. Despite the fact that they have been in use for almost a century, classical test theory, and the recommended linear transformations of raw scores still allow to improve students' comparability, but they do not change

the nature of an ordinal scale. The scales considered below are related to such transformations.

A prime example is the Z-scorebased on the conversion of a raw score r:

$$z = \frac{r - M}{\sigma} \tag{1}$$

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

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

A large number of other linear transformations is based on the Z-scale, *e.g.* IQ – $(100+15z)$, CEEB – $(500+100z)$, Veksler – $(10+3z)$, T-score – $(50+10z)$ [2]. Each of these scales was applied in practical pedagogical measurements and uses for different interpretations of tests [3]. features of different interpretations.

A group of methods based on the percentile transformation of raw scores provides the best comparability of results in the framework of classical testing theory. Thus in order to determine the relative position of a student in a group, it suffices to express their rank in percentiles, *i.e.* a fraction of students in the group, whose results are below or equal to the primary score of the student. Despite an apparent advantage, *i.e.* an opportunity to rank the relative position of a student on a scale, the comparison of different samples (*i.e.* different tests, sets of sessions on one subject, *etc.*) is complicated by the fact that the percentile distribution is closely related to the frequency distribution of the sample, for which it was obtained [3].

Unfortunately, even percentile assessments are difficult to compare with each other, if they are obtained for different samples. The best solution of this problem is the standardization of the samples, which expresses all scores using a common scale. In order to achieve this, all empirical density frequency distributions of raw scores are converted to the same reference distribution, *i.e.* a normal distribution with a given mean and variance. Usually in such cases a centered and normalized normal distribution is applied.

This method is termed equal-percentile normalization (EPN) and is applied in many countries for the assessment of knowledge. For example, in Ukraine, the method converts raw scores to the scale $[100,200]$, which uses a reference distribution $N(150,15)$ (normal distribution with mean value of 150 and standard deviation of 15).

Thus, we apply the EPN as the main method in our comparative analysis.

1.2. Item response theory

Item response theory (IRT) is based on the study of the relationship between the difficulty of an item, student's abilities and the probability of a correct answer. The basic model which reflects this relationship was named after Rush [4].

A success of test-taker in solving a task from the item has a probabilistic nature. Let us assume that the probability that the student solves an item

correctly (success rate) is determined as a function of the level of student's knowledge a and level of item's difficulty d:

$$p = p(a,d) = \frac{a}{a+d} = \frac{a/d}{1+a/d} = \frac{\lambda}{1+\lambda} \tag{2}$$

where λ stands for the ratio of the latent parameters of training level and difficulty.

Equation (2) corresponds to the the Rasch model [4], %[[Literature reference needed.]] according to which the probability of success does not depend on arguments *per se*, but on their relationship. Let us study some properties of this function.

The unit measurement for readiness and difficulties is the same. If we assign a unit difficulty $d_0 = 1$ (and similarly for unit ability $a_0 = 1$), then the difficulty of all items can be compared to the unit difficulty and the ability can be compared in the same way. (If the difficulty d of an item is lower than 1, then this item is $1/d$ times easier than the unit difficulty. If its value is higher than 1, then, the item is d times more difficult).

Hence $d, a, \lambda \in [0,\infty)$, $p \in [0,1]$:

- If $\lambda \to 0$, $\frac{a}{d} \to 0$, $p \to 0$, then the student is completely unprepared and is unable 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, is bound to successfully pass the test.

Function arguments in Equation (2) cannot be measured directly, but the value of the function, *i.e.* the probability, is available for the measurement based on test results. Basically, in the IRT, we must know the probability in order to estimate the difficulty of the items and the level of student's ability. Based on the type of function (2), it is obvious that this problem does not have a correct solution. An inverse function allows to determine the parameter λ only on the measured value p, *i.e.*:

$$\lambda = \frac{p}{1-p} = \frac{p}{q} \tag{3}$$

and to find only the ratio of the latent parameters of ability and difficulty. If we have a reference item with unit difficulty, it is possible to identify the corresponding value of ability as well as to position it on a scale. This is another advantage of the IRT, since it allows to solve the problem of the standardization of various populations.

2. Rasch's logistic function

In practice, it is convenient to express the level of ability and difficulty arguments not on a linear but rather on a logarithmic scale:

$$\ln a = \theta,\ \ln d = \delta \Leftrightarrow a = e^{\theta},\ d = e^{\delta} \tag{4}$$

The function of success takes the form:

$$p = \frac{e^{\theta}}{e^{\theta} + e^{\delta}} = \frac{1}{1+e^{-(\theta-\delta)}} \tag{5}$$

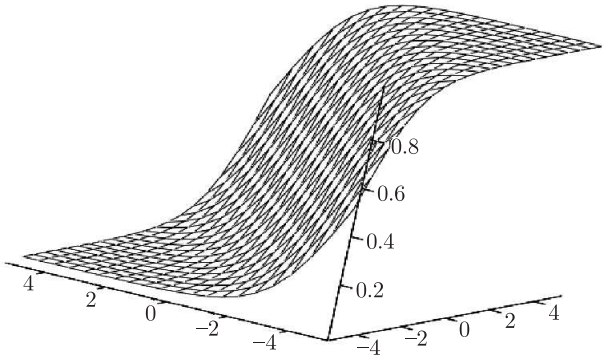

Figure 2. Rasch's logistic function

Formula (5) is termed Rasch's basic logistic model. A graph of the function (5) is shown in Figure 2.

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

Clearly:

$$\frac{1}{1+e^{-1}} = 0.731 \tag{6}$$

This means that the distance of 1 logit increases the probability of success 0.731-fold.

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

Let us consider a testing scheme with k items. We assume that n people are involved in the testing. The results for each item are assessed by the dichotomous principle.

Let us consider $R = (r_{ij})$ as a response matrix:

$$(i = 1,\ldots,n;\ j = 1,\ldots,k) \tag{7}$$

where r_{ij} are random variables, which take the value of 1 with probability

$$p_{ij} = p(\theta_i,\delta_j) \tag{8}$$

The calculation of the raw scores of participants and items yields:

$$b_i = \sum_{j=1}^{k} r_{ij},\ i = 1,\ldots,n \tag{9}$$

$$c_j = \sum_{i=1}^{n} r_{ij},\ j = 1,\ldots,k \tag{10}$$

A set of values $b_i \in \{0,1,\ldots,k\}$, *i.e.* the total number of participants, can be divided strictly into $k+1$ groups according to the value of the raw score.

Let us rewrite formulas (3) and (5) as follows:

$$\lambda = \frac{p}{1-p} = \frac{p}{q},\ p = \frac{e^{\theta}}{e^{\theta}+e^{\delta}},\ q = \frac{e^{\delta}}{e^{\theta}+e^{\delta}}, \tag{11}$$

then

$$\lambda = e^{\theta-\delta} \tag{12}$$

Note that λ is the ratio of the latent parameters of ability and difficulty on a linear scale. The logarithm of (12) gives the discrepancy between the latent parameters of ability and difficulty θ, δ on the logit scale.

Let us assign this parameter as follows:

$$l = \ln\lambda = \theta - \delta \tag{13}$$

For each participant and each item, we can determine the value of

$$l_{ij} = \theta_i - \delta_j;\ i = 1,\ldots,n;\ j = 1,\ldots,k \tag{14}$$

if the corresponding probabilities are not equal to 0 or 1.

Moreover, l_{ij} can be measured on the basis of raw scores. The unknown quantities in the system of equations (14) are the level of ability and difficulty θ_i, δ_j, respectively.

Since all participants who receive the same raw score, have the same level of ability (according to the Rasch model), the number of equations is reduced from $n \times k$ to $(k+1) \times k$ so that for $n \gg k$ the system of equations is substantially reduced.

n_b is the number of participants which received the same raw score $b = 0,1,\ldots,k$. $\Theta(b)$ is the level of group ability. Thus the system of equations (14) can be rewritten as:

$$l_j(b) = \Theta(b) - \delta_j;\ j = 1,\ldots,k;\ b = 1,\ldots,k-1 \tag{15}$$

If $b = 0$, $b = k$, then either the participant did not complete any item, or they completed all items k.

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

For the remaining values of $l_j(b)$, they can be defined as follows:

$$l_j(b) = \ln\frac{p_j(b)}{q_j(b)} \tag{16}$$

where $p_j(b)$ is the relative frequency of the correct solution of item j completed by the participants who obtained the same raw score.

Taking into account *e.g.* $l_j(0) = -5$, $l_j(k) = 5$, and similar cases in different groups b, for which *e.g.* $p_j(b) = 0$, $p_j(b) = 1$, we obtain an inconsistent system of $k \times (k+1)$ equations (15) containing $2k+1$ unknown values.

In this case, the coefficient matrix is singular and its rank is $2k$, *i.e.* the number of independent equations is lower than the number of unknown parameters. Therefore, one of the values must be assumed *ad hoc*, and the

remaining unknowns must be expressed with respect to it. This value specifies the origin point of the scale. It is convenient to combine the origin point with the mean value Θ of parameter θ expressed in logits.

The system of equations (15) can be solved in several ways, *e.g.* by menas of a system of normal equations, using the method of moments or the maximum likelihood method. These methods are described in detail in [3] and applied in the WinSteps software in order to estimate the latent parameters.

Despite the fact that the theory (accordingly named) assumes correlation between the difficulty of an item and the level of ability of a participant, in practice, the same number of subgroups of students that have obtained the same score on the logit scale as on the raw-score scale. This means that the canonical IRT does not improve the quality of the estimation understood as a function of the diversity of the initial scores. Thus, it is possible to improve this theory so as to increase the diversity of participants with different test scores. This problem can be solved only when participants who obtained the same score (even on the logit scale), are subsequently redistributed within the subgroups according to the difficulty of the items, for example, either in lexicographical order, or by converting the ability levels for fixed values of the difficulty of the items. These modifications of the canonical IRT are given in this paper.

2.2. New approach to diverse scores

Suppose that in solving the system of equations (15), we obtained the values of the ability levels in groups $b=0,1,\ldots,k$, and the levels of test difficulty δ_j, $j=1,\ldots,k$.

Let us consider a group with the same level of ability:

$$\Theta(b^*)=\theta_{n_1}=\ldots=\theta_{n_{\kappa^{\delta^*}}} \tag{17}$$

where $n_1,\ldots,n_{\kappa^{\delta^*}}$ are the numbers of participants belonging to this group. For these participants, the number of items completed correctly is identical (their raw scores are equal, *i.e.*, for dichotomous tests, the number of correct answers and items is equal as well). However, we assume that the items can be different.

Let us sort the group $\{n_1,\ldots,n_{\kappa^{\delta^*}}\}$ in lexicographical order, *i.e.* the first position in the group is taken by a participant who completed the most difficult item. If there are several such participants, the second most difficult item is taken into account, and so on. Thus, all participants in the group are ranked based on the difficulty of items. Such a solution is acceptable when ranking in the framework of one test. If, however, a change in the quantitative values of ability levels is required, we propose applying an iterative procedure based on the parameters of the Rasch model obtained earlier for this test as well as changing the indicators of ability levels for fixed values of item difficulty.

Along with the one-parameter Rasch model (5), in practice, a two-parameter model is widely used . It has the following form:

$$p(\theta) = \frac{1}{1+e^{-a(\theta-\delta)}} \tag{18}$$

where the probability of success of a student with an ability level of θ is defined in terms of the difficulty of the item δ and the discriminatory properties of the item a, *i.e.* the coefficient of discrimination.

For all members of the group with the same ability level $\Theta(b^*)$, we conduct an iterative refinement procedure for the ability level using the formula commonly applied in the method of moments:

$$\hat{\theta}_{j,s+1} = \hat{\theta}_{j,s} + \frac{\sum\limits_{i=1}^{k} a_i \left(r_{j,i} - p(\hat{\theta}_{j,s})\right)}{\sum\limits_{i=1}^{k} a_i^2 \left(p(\hat{\theta}_{j,s})(1-p(\hat{\theta}_{j,s}))\right)} \tag{19}$$

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

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 identical at the beginning of the iterative procedure, *i.e.* $\theta_{n_1,0} = \ldots = \theta_{n_{\kappa^{\delta^*}},0} = \Theta(b^*)$.

3. Example

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

Table 1. The response matrix

No. 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**

Let us consider the application of a classical EPN method.

The frequency analysis of the results is presented in Table 2. Table 2 also shows the conversion of the scores to a scale of 100–200 (column EPN).

Table 2. The frequency analysis of results

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

The results for the two-parameter Rasch model (18) and the analysis of the test using IRT are shown in Table 3.

Table 3. The results of two-parameter Rash model

No. member	item 1	item 2	item 3	raw score	ability level (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**
item difficulty (logit)	**−0.57**	**0.1**	**0.47**		

If we assign 3 logit to 200 points, a linear transformation of the logit scale into a scale of 100–200 points on the logit scale can be performed using the formula:

$$150 + 16.667 \cdot \text{logit} \tag{20}$$

and thus we obtain the values of 162 and 138 points, respectively.

However, as in the case of the EPN, the IRT does not change the number of participants with the same score. This example demonstrates that here the level of item difficulty is different, and thus this factor must be taken into

consideration when ranking participants with identical results. Let us apply the proposed methods in order to improve the quality of our estimation.

With the lexicographic ordering of the participants, who got 2 total score, it is easy to notice that the following order is correct: $1 \succ 2 \succ 3$, since participant No. 1 solved the most difficult item (worth 0.47 logit) and the subsequent item (worth 0.1 logit), while participant No. 2 solved a simpler combination.

For the participants who obtained the raw score of 1, the order would be:

$$13 \equiv 12 \succ 11 \equiv 10 \equiv 9 \succ 8 \equiv 7 \equiv 6 \equiv 5 \equiv 4 \tag{21}$$

This approach ranks the participants within their respective groups, however, it does not add any quantitative information to the resulting scores. Let us consider the application of the iterative procedure (19).

As a result of applying the Rasch model (18) in the WinSteps package, the following values of the discrimination coefficients for the items, were obtained:

$$a_1 = 0.43;\ a_2 = 1.04;\ a_3 = 1.18 \tag{22}$$

The initial values of the ability levels are presented the final column of Table 3.

The results of the iterative procedure (19), starting from step 2, are shown in Table 4.

Table 4. The results of the iterative procedure

No. 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

Figure 3 presents the listing of the MatLab code implementing the calculation.

Using Equation (20), we can present the obtained results in a single table, see Table 5.

```
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;
```

Figure 3. Program listing for the iterative procedure (19)

Table 5. The comparision of results

No. 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 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. This does not contradict the validity of our estimation.

4. Conclusion

The application of a classical test theory and the IRT allows to improve the comparability of test results. However, the best results can be obtained by modifying the IRT by ranking within groups with the same score based on the difficulty of items. Our future work will be devoted to a fuzzy score modification model, which combines the properties of the estimates of the ranking lexicographic method and the iterative procedure for calculating of makeweight to the estimates in group.

References

[1] Samylkina H H 2007 *Modern Means of Assessment of Learning Outcomes*, *Bean, Knowledge Lab, p. 172* (in Russian)
[2] Chelyshkova M B 2002 *Theory and Practice of Designing Pedagogical Tests*, Textbook, Logos, p. 432 (in Russian)
[3] Neumann S M and Khlebnikov B A 2000 *Introduction to Modeling and Parameterization of Pedagogical Tests*, Moscow, p. 168 (in Russian)
[4] Baker F B 2001 *The Basics of Item Response Theory*, University of Wisconsin, p. 185