UDC 37.091.26:004.9

N. Kruglova, O. Dykhovychnyi, D. Lysenko

APPLICATION OF IRT AND MIRT MODELS
TO ANALYSIS OF ANALYTICAL GEOMETRY TESTS

Abstract: The article conducts research on the construction of analytical methods of computer tests in higher mathematics. These computer tests contain test items of various types, including problems with "embedded answers" which include several related subitems. Also, on the basis this method quality analysis of the control work was conducted. The technic is based on the methods of Classical Test Theory (CTT) and Item Response Theory (IRT), which have proven their effectiveness in statistical analysis of tests. The focus of the work is on the use of Multidimensional Item Response Theory (MIRT) models, which allows to analyze simultaneously the whole vector of students' competencies and to scrutinize them more carefully. The study also uses one-dimensional IRT models, the results of which are compared with the use of MIRT. Among the one-dimensional models, the well-known Muraki and Birnbaum models were selected, and among the multidimensional models, the choices were the two-dimensional 2-PL and GPCM. The multidimensional models involved in the study are compensatory. The application of non-compensatory models was not considered. The comparison of compliance with the data of different models was carried out on the basis of special information criteria. One-dimensional models were discovered to be somewhat more effective.

The R programming environment was chosen as the main toolkit, which provides a powerful set of software for the test statistical analysis. The **mirt** package is selected as the main software package.

Tests results from a modular test in analytical geometry have been chosen as data for research. The test was written by 105 students from IT stream, specialty 121, at FIOT NTUU "Igor Sikorsky Kiev Polytechnic Institute". The test was placed on the MOODLE platform and was conducted remotely. The analysis of test results on the basis of selected models demonstrated the consistency between the results of both one-dimensional and multidimensional models. However, multidimensional models also allow to expound the analysis of various additional competences, in this case - knowledge of vector algebra and knowledge of lines, planes, and surfaces in space.

The analysis showed that the test was, in general, composed correctly; also the analysis allowed to systematize the items by complexity; and for questions such as "embedded answers " to detail the complexity of the subitems.

Evaluating the overall results of the application of one-dimensional and multidimensional models of IRT, it should be noted their effectiveness in the analysis of both tests in higher mathematics and in the control of knowledge in other disciplines.

Keywords: computer tests; "embedded answers (cloze)"; CTT; IRT; MIRT; MOODLE.

Problem statement

The article presents mathematical methods of test analysis based on one-dimensional and multidimensional IRT models. These methods allow for a more detailed analysis of the quality of the test, namely, to identify problematic issues that need to be improved: have too high or too low complexity; excessive probability of guessing; problems that are solved with the help of additional software. MIRT models generalize one-dimensional models by introducing additional latent parameters that correspond to certain competencies of the examinee. Due to this, the set of assessed competencies of examinees is expanding; also, each examinee is characterized not only by the level of training in a particular field but also by skills in related fields of knowledge.

Let us illustrate the method of the research of tests on modular control work on analytical geometry. 105 students of Faculty of Informatics and Computer Science (FICT) National Technical University of Ukraine (NTUU) "Igor Sikorsky Kyiv Polytechnic Institute" 121-th specialty of the IT stream wrote a test. The control work was placed on the MOODLE platform and it was conducted remotely.

The test consisted of 10 test items: 2,3,4,7,10 items are "numerical" type, and 1,5,6,8,9 items are "embedded answers". For the "numerical" type we used dichotomous models, and for "embedded answers" we used polytomous models.

A feature of the test is the presence of items such as "embedded answers". Test questions such as "embedded answers" (EA) differ in that they are performed step by step and can reveal several competencies of examinees at the same time. These test items can be used in combination with questions such as "numerical", "correspondence", "multivariate question", "calculation" and require the selection of an appropriate mathematical model for qualitative analysis of test results.

Knowledge of vector algebra (competence 1), knowledge of lines and planes (competence 2) were tested, which in turn require the ability to convert algebraic expressions to the form required by the condition of the test problem in this test. This skill was especially used in problems to determine the equations of lines and planes because they have many equivalent notations. Therefore, one of the competencies was fixed in these items, and the other was already determined by using this condition.

Obviously, the question arises: what the best way is to analyze the relevant tests, in which there are test questions of different types; in particular, the type of "EA".

Analysis of recent research and publications

Statistical analysis of tests, including pedagogical ones, is based on two main theories: Classical Test Theory (CTT) [1, 2] and Item Response Theory (IRT) [3 - 7]. Studies with the combined application of both theories are described in [8]. This approach is quite acceptable and can be applied in pedagogy or sociology. For example, to analyze the quality of psychological tests in [9 - 11] instead of traditional CTT used IRT, which significantly clarified the test results.

The experience of testing in higher mathematics at technical faculties "KPI. I. Sikorsky "are given in [12, 13]. Using IRT models for designing tests in physics described in detail in [14]. The relationship between different types of test items and statistical characteristics of the test was studied in [15]. In [16] the possibility of using the Rash polytomous model in compliance problems in the Unified State Exam (USE) was investigated.

Using the platform MOODLE new models forms of tests, for example, such as "embedded answers" for testing control in higher mathematics National Technical University of Ukraine (NTUU) "Igor Sikorsky Kyiv Polytechnic Institute" requires the use of new IRT models for the analysis of such items. One possible way to analyze such models is to use Multidimensional IRT (MIRT) [17]. MIRT algorithms in the R [18] environment are implemented in the **mirt** package.

Item definition

Two items were set in the work:

1. Choose the optimal model for the analysis of control work, which consists of various items, including items such as "EA";

2. Based on the chosen model to analyze the results of the control work and draw conclusions.

We have selected the following models of one-dimensional IRT. Let:

- n be the number of examinees,
- m be the number of items in the test.

1) *Birnbaum Model (2-PL)* [5]

The probability of the correct answer of the i -th examinee to the j -th item of the test is determined as follows:

$$p_{ij} = \frac{1}{1 + e^{-a_j(\theta_i - b_j)}}, i = \overline{1, n}, j = \overline{1, m},$$

where $a_j, j = \overline{1, m},$ are discrimination parameters of the item; $\theta_i, i = \overline{1, n},$ is an ability parameter of the i-th examinee; $b_j, j = \overline{1, m},$ are difficulty parameters of the item.

2) *Muraki Model* [4].

A problem is called polytomous if the problem j has sublevels m_j . One of the following models, which describes the following item, specifies the probability of reaching the i-th examinee of the k-th level of the j-th item:

$$p_{ijk} = \frac{e^{\sum_{s=0}^{k} a_j(\theta_i + d_{js})}}{\sum_{k=0}^{m_i} e^{\sum_{s=0}^{k} a_j(\theta_i + d_{js})}}, i = \overline{1, n}, j = \overline{1, m}, k = \overline{1, m_j},$$

where: a_j are discrimination parameters of the item; $d_{jk}, j = \overline{1, m}, k = \overline{0, m_j},$ are difficulty

parameters of the k-th level of the j-th item; $\theta_i, i = \overline{1,n},$ is the ability parameter of the i-th examine $\theta_i, \ i = \underline{1, N}$; m_j is the number of item levels.

Note that the Birnbaum model is a partial case of the Muraki model for the case when $m_j = 1$.

Consider some multidimensional **MIRT** models. **MIRT** can be considered as a generalization of IRT in the case of multidimensional data. The main feature of these models is that the level of readiness of the i-th examinee is an s-measurable vector $(\theta_{i1}, \theta_{i2}, ..., \theta_{is})$. MIRT models are divided into two classes: compensatory and non-compensatory.

Compensatory models are additive, i.e., are supposed to compensate for poor preparedness examinees on one basis due to the high level of knowledge and skills to other conditions.

3) *Compensatory 2-PL model* [17].

The probability of the correct answer of the *i*-th examinee to the *j*-th dichotomous item of the test is determined by the formula:

$$P_{ij} = \frac{e^{\sum_{l=1}^{s}(a_{jl}\theta_{il}+d_j)}}{1+e^{\sum_{l=1}^{s}(a_{jl}\theta_{il}+d_j)}}, i = \overline{1,n}, j = \overline{1,m}.$$

where: $d_j, j = \overline{1,m},$ - difficulty parameter of the item; $a_{jl}, j = \overline{1,m}, l = \overline{0,s},$ - the discrimination parameter of the item *j*; $\theta_{il}, i = \overline{1,n}, l = \overline{0,s},$ - ability parameter of the i-th examinee $\theta_i, \ i = \underline{1, N}$; s - measurability of the ability vector.

$$B_j = \frac{-d_j}{\sqrt{\sum_{q=1}^{s} a_{jq}^2}}, j = \overline{1,m}, \text{ - multidimensional difficulty parameter of the item.}$$

$$A_j = \sqrt{\sum_{q=1}^{s} a_{jq}^2}, j = \overline{1,m}, \text{ - multidimensional discrimination parameter of the item.}$$

4) *Multidimensional Generalized Partial Credit Model* [19].

In this model, the best quality of achieving the i-m examiner of the k-th level of the j-th question is determined by the formula:

$$P_{ijk} = \frac{e^{k\sum_{l=1}^{s}a_{jl}\theta_{il}+\sum_{q=0}^{k}d_{jq}}}{\sum_{v=0}^{m_i} e^{v\sum_{l=1}^{s}a_{jl}\theta_{il}+\sum_{q=0}^{v}d_{jq}}}, \ i = \overline{1,n}, \ j = \overline{1,m}, k = \overline{1,m_j}.$$

where: $d_{jq}, j = \overline{1,m}, q = \overline{1,m_j},$ - the complexity of the *q*-th level of the *j*-th items; $a_{jl}, j = \overline{1,m}, l = \overline{0,s},$ - the discrimination parameter of the item *j*; $\theta_{il}, i = \overline{1,n}, l = \overline{0,s},$ - ability

parameter of the i-th examinee θ_i, $i = \overline{1, N}$; m_j - number of level of the item; s - measurability of the ability vector.

Similar to the previous one:

$$B_{jk} = \frac{-d_{jk}}{\sqrt{\sum_{q=1}^{s} a_{jq}^2}}, \ j = \overline{1, m}, \ k = \overline{1, m_j},$$

the complexity of the *k*-th step of the *j*-th items.

Non-compensatory models are multiplicative. They are used when insufficient knowledge in a certain area cannot compensate for good skills or the skills in another area of expertise.

The main research material

The analysis of the control work on analytical geometry was carried out using a mixed model: some questions were dichotomous type "numerical", and some polytomous type "embedded answers".

Three approaches were used:

A. One-dimensional mixed model (2-PL + GPCM);

B. Multidimensional model with two features;

C. Multidimensional model with two features except for one feature;

Questions 1-3, 7 tested the ability to use the properties of vectors, and questions 4-6, 8-10 tested the knowledge of curves and lines in the plane, lines and planes in space.

Problems 2, 3, 4, 7, 10 were "numerical" questions, so they can be analyzed by a dichotomous model. Items 1, 5, 6, 8, 9 were of the "embedded answers" type. Problems 1 and 6 contained 4 questions, 5, 8 and 9 problems contained 3 questions. We will analyze them using a polytomy model.

Initially, Cronbach's coefficient was calculated for the control work

$$(\alpha = 0.7054).$$

This is a fairly high value, which indicates the high reliability of the test.

Initially, the adequacy of the models was checked using special information criteria (Table 1).

Table 1. **The value of the criteria for the models**

Criterion	**A**. One-dimensional mixed model (2-PL + GPCM)	**B**. Multidimensional model with two features	**C**. Multidimensional model with two features except for one feature
Number of iterations	18	78	1142
AIC	1441.768	1449.803	1477.354

Ended tabl 1.

Criterion	**A**. One-dimensional mixed model (2-PL + GPCM)	**B**. Multidimensional model with two features	**C**. Multidimensional model with two features except for one feature
AICc	1461.405	1504.47	1506.688
BIC	1513.425	1558.616	1562.281
SABIC	1428.127	1429.089	1461.187
G2	534.6	514.64	560.18

As can be seen from table 1, the least important criteria are for model A. Slightly better is model A.

A. One-dimensional mixed model (2-PL+GPCM)

First, consider the one-dimensional case A. Since the Birnbaum model is a partial case of the Muraki model, the **mirt** function of the **mirt** package was used.

Table 2. **Coefficients of the one-dimensional Muraki model**

Item	b_1	b_2	b_3	b_4	a
1	0.372	-1.316	-2.876	-2.056	0.448
2	-0.621				0.561
3	-3.096				0.916
4	-0.703				1.539
5	4.931	-0.188	-7.779		0.403
6	3.192	-4.286	1.498	-5.796	0.457
7	-1.913				1.504
8	-1.631	0.381	-2.397		0.655
9	-2.431	-1.706	-3.271		0.787
10	-1.346				1.253

Table 2 shows:

− in general, dichotomous items is composed correctly. They have moderate complexity and positive differentiating ability. The simplest are items 3.7;

− 1st item has a low complexity, subitems are relatively uniform;

− the 3rd subitem of the 5th item and the 4th subitem of the 6th item have the lowest complexity. The first questions of the 5th and 6th problems are the most difficult, which is also reflected in the graphs of the characteristic curves;

− applying the χ^2 criterion, we see that the model is adequate (p-value = 0.138> 0.05) Here are the texts of the relevant items (Figure 1):

Item 1

Задано пряму $l : 2x - y - 5 = 0$ та точку $M_1(-1; -2)$. Знайдіть:

1) проекцію точки M_1 на пряму l - точку O.

$O($ ⬚ ; ⬚ $)$.

2) Точку M_2, яка симетрична точці M_1 відносно прямої l.

$M_2($ ⬚ ; ⬚ $)$.

Item 3

Знайдіть гострий кут між площиною $\alpha : x + y + 7 = 0$ та площиною, яка містить точки $M_1(1; -2; 5)$, $M_2(-1; 5; 2)$, $M_3(1; 3; 0)$. У відповіді вкажіть градусну міру кута (тільки число).

Відповідь: ⬚

Item 5

Задано точки A(-1;-3) та B(2;1).

а) Запишіть загальне рівняння прямої l, що перпендикулярна до прямої АВ та проходить через точку В.

l: 3x+ ⬚ y+(⬚)=0.

б) Знайдіть відстань d від прямої l до початку координат.

d= ⬚

Item 6

Задано вектори: $\bar{a} = \bar{i} + 2\bar{j} - \bar{k}$, $\bar{b} = 2\bar{i} + 3\bar{k}$, $\bar{c} = 3\bar{i} - 2\bar{j}$, $\bar{d} = 2\bar{a} - \bar{b}$.

1) Обчисліть $\overline{a}\overline{c}\overline{d}=$ ⬚

2) Обчисліть $\bar{a} \times \bar{b}$ і введіть у відповідні комірки координати вектора $\bar{a} \times \bar{b}=$

=(⬚ ; ⬚ ; ⬚).
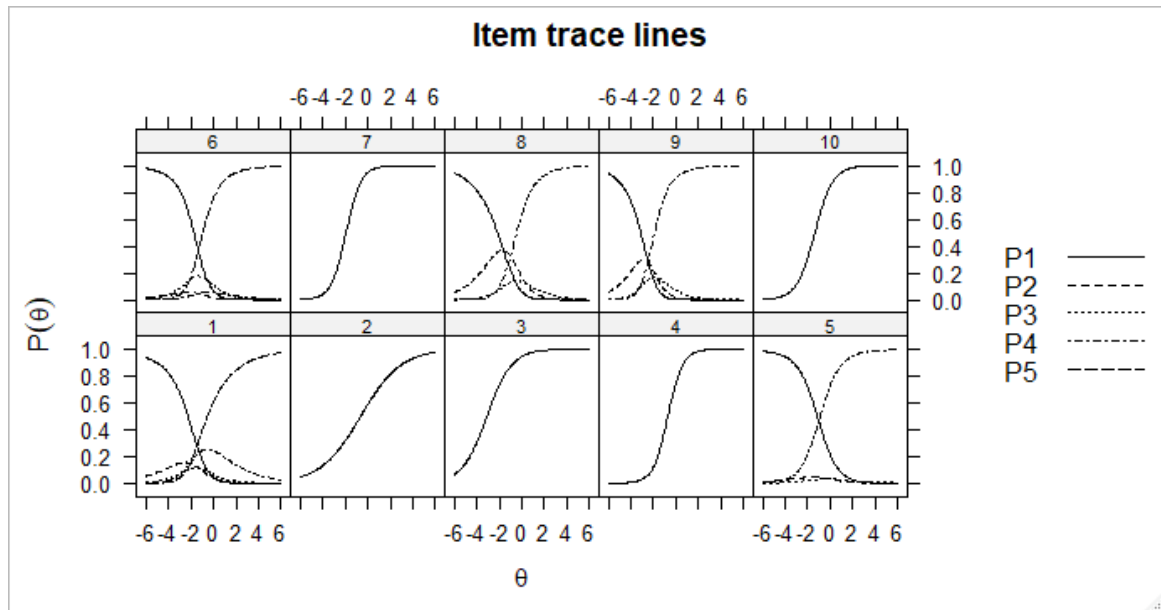
Figure 1. The texts of the relevant items

Figure 2. Characteristic item curves

Characteristic (Fig. 2) item curves correspond to the estimated parameters. They show that some subitems, for example, in the 5th and 6th "sag" due to high complexity.
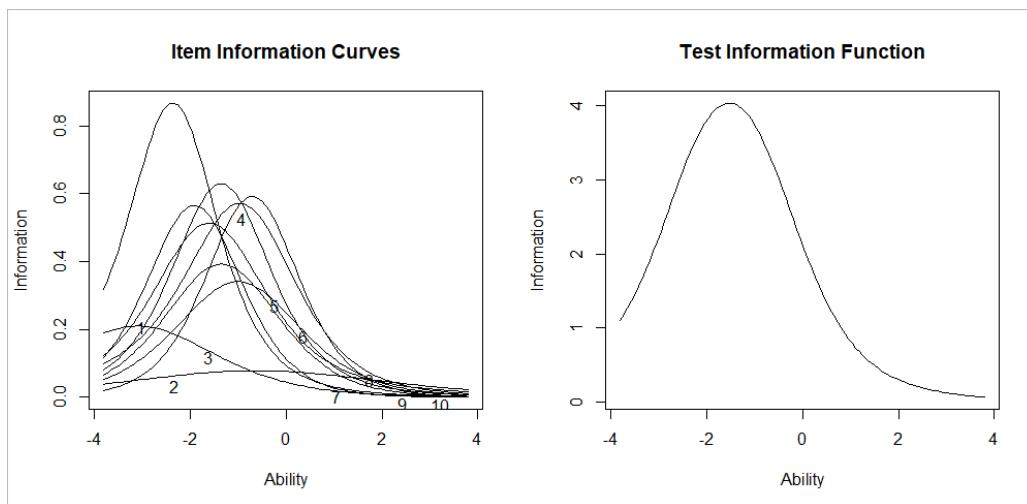

Figure 3. The information curve of the test

The information curve (Figure 3) of the test has a bell-shaped shape. Therefore, the test is passed normally. But the maximum of the curve is shifted to the left. This means that the test is most informative for poorly prepared students.

B. Multidimensional model with two features

We use a multidimensional model 2-PL model to analyze the same items. Let's assume $(\theta_{i1}, \theta_{i2})$ levels of readiness of the test subject in vector algebra (competence 1) - θ_{i1}, knowledge of lines, curves, planes and the ability to make appropriate algebraic transformations (competence 2) - θ_{i2}.

Estimates of the parameters of dichotomous problems for the compensatory 2-PL model are collected in Table 3, and the coefficients of polytomous problems of the type "BB" according to the multidimensional GPCM model are collected in Table 4.

Table 3. **Coefficients of dichotomous problems of the two-dimensional model**

Item	a_{j1}	a_{j2}	d_j	A_j	B_j
2	-0,499	0,33	0,35	0,60	-0,59
3	-1,559	-0,71	3,528	1,71	-5,90
4	-2,515	-0,064	1,426	2,52	-2,38
7	-1,043	1,249	2,978	1,63	-4,98
10	-1,777	0	1,969	1,78	-3,29

Table 4. **Coefficients of polytomous problems of a two-dimensional model**

Item	a_{j1}	a_{j2}	d_{j1}	d_{j2}	d_{j3}	d_{j4}	A_j	B_{j1}	B_{j2}	B_{j3}	B_{j4}
1	-0,31	0,322	-0,164	0,43	1,722	2,643	0,45	0,36	-0,95	-3,81	-5,84
5	-0,27	0,277	-2,00	-1,927	1,206	-	0,39	5,19	5,00	-3,13	-
6	-1,09	2,48	1,6	5,821	6,718	9,599	2,71	-0,59	-2,15	-2,48	-3,54
8	-0,61	0,126	1,035	0,769	2,343	-	0,62	-1,66	-1,23	-3,76	-
9	-0,68	0,319	1,83	3,123	5,684	-	0,75	-2,45	-4,17	-7,59	-

Analyzing tables 3 and 4, we see:

− that items in different competencies have the resolution of different characters, but the overall resolution of the items A_j are positive;

− analyzing B_j, we see that all items are moderately complex;

− the simplest is again item 3.7, and the highest resolution has item 4;

− items 1 and 9 are generally too simple;

− the 1st and 2nd subitems of the 5th item are complex;

− in the 6th complex is the 1st subitem, and 2-4 is again a simple 4;

− in general, the resolution of the items is integral and the complexity of the test is low.

Figure 4 shows the characteristic surfaces (right) and level lines for items (left). Analyzing the level lines, we see that most of them, except for problem 3, have a positive slope. This is quite understandable because it is illogical to assume that ignorance of vectors can be compensated by knowledge of lines, planes or surfaces. That is, most likely, it is inappropriate to use a compensatory model for test analysis. The high density of contour lines in Problem 6 is explained by their high resolution.

C. Multidimensional model with two features except for one feature

The analysis of control work on the multidimensional model in which only one competence is chosen was carried out. That is, part of the questions was focused on the test of competence 1, and part - on the test of the second competence 2. The results are shown in tables 5,6. Resolution 0 excludes the corresponding competence.

Table 5. **Coefficients of dichotomous items**

Item	a_1	a_2	d	B
2	0,235	0	0,288	-1,22553
3	0,589	0	2,518	-4,27504
4	0	1,778	1,172	-0,65917
7	1,119	0	2,34	-2,09115
10	0	1,771	1,967	-1,11067

Table 6. **Coefficients of polytomous problems of the two-dimensional model**

	a_1	a_2	d_1	d_2	d_3	d_4	B_1	B_2	B_3	B_4
1	0	9.035	9,583	16,974	22,49	22,678	-1,06	-1,88	-2,49	-2,51
5	0	0,336	-2,044	-1,986	1,156	-	0,23	0,22	-0,13	-
6	0,37	0	-1,584	0,303	-0,41	2,249	0,18	-0,03	0,05	-0,25
8	0	0,595	0,987	0,713	2,294	-	-0,11	-0,08	-0,25	-
9	0	0,902	2,04	3,525	6,149	-	-0,23	-0,39	-0,68	-

Analysis of tables 5,6 shows:

− again the simplest are 3 and 7 items. Resolution is inalienable. Although the items are relatively easy.

− also in the 5th and 6th items the first subitems are difficult.
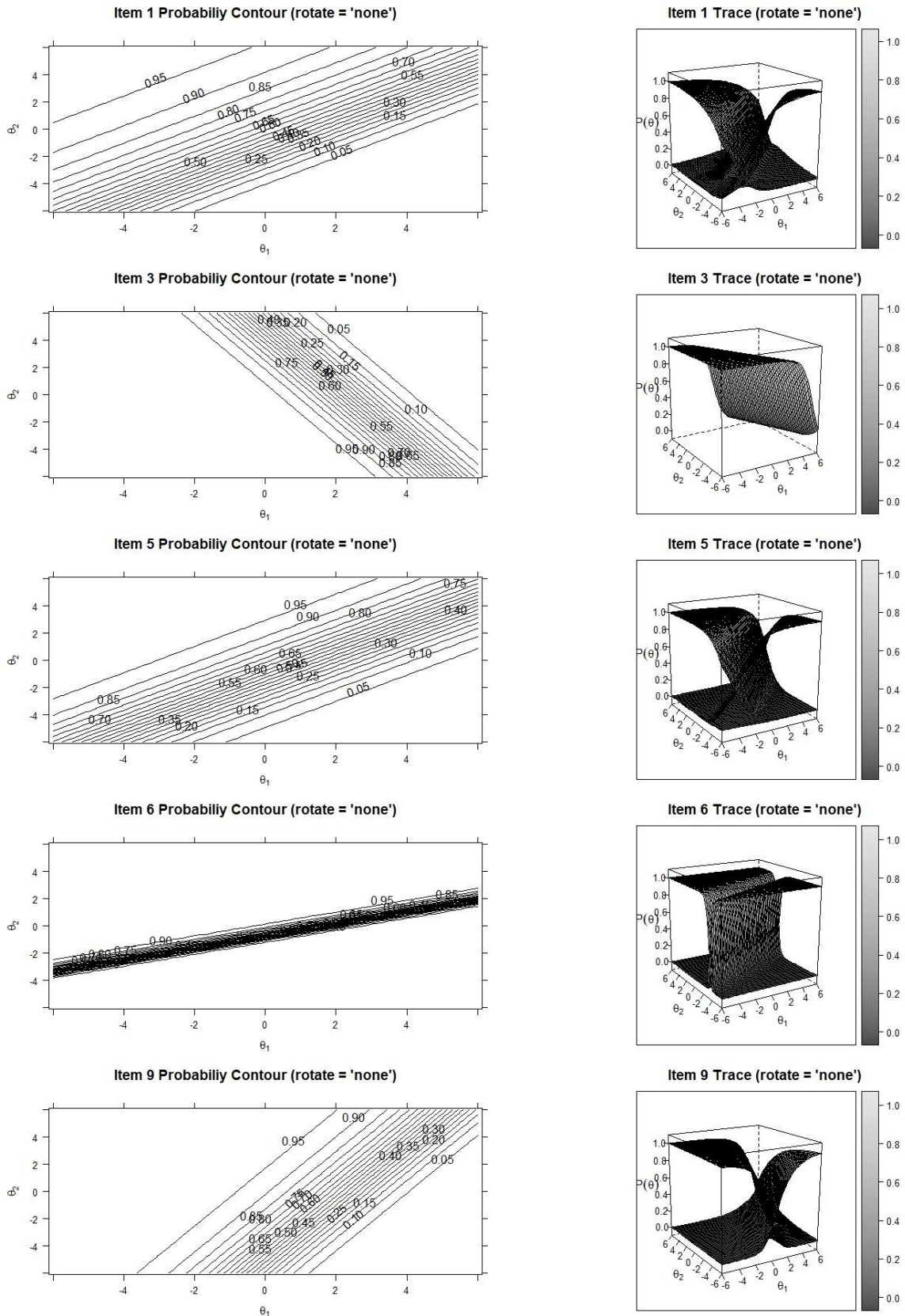
− resolutions are positive.

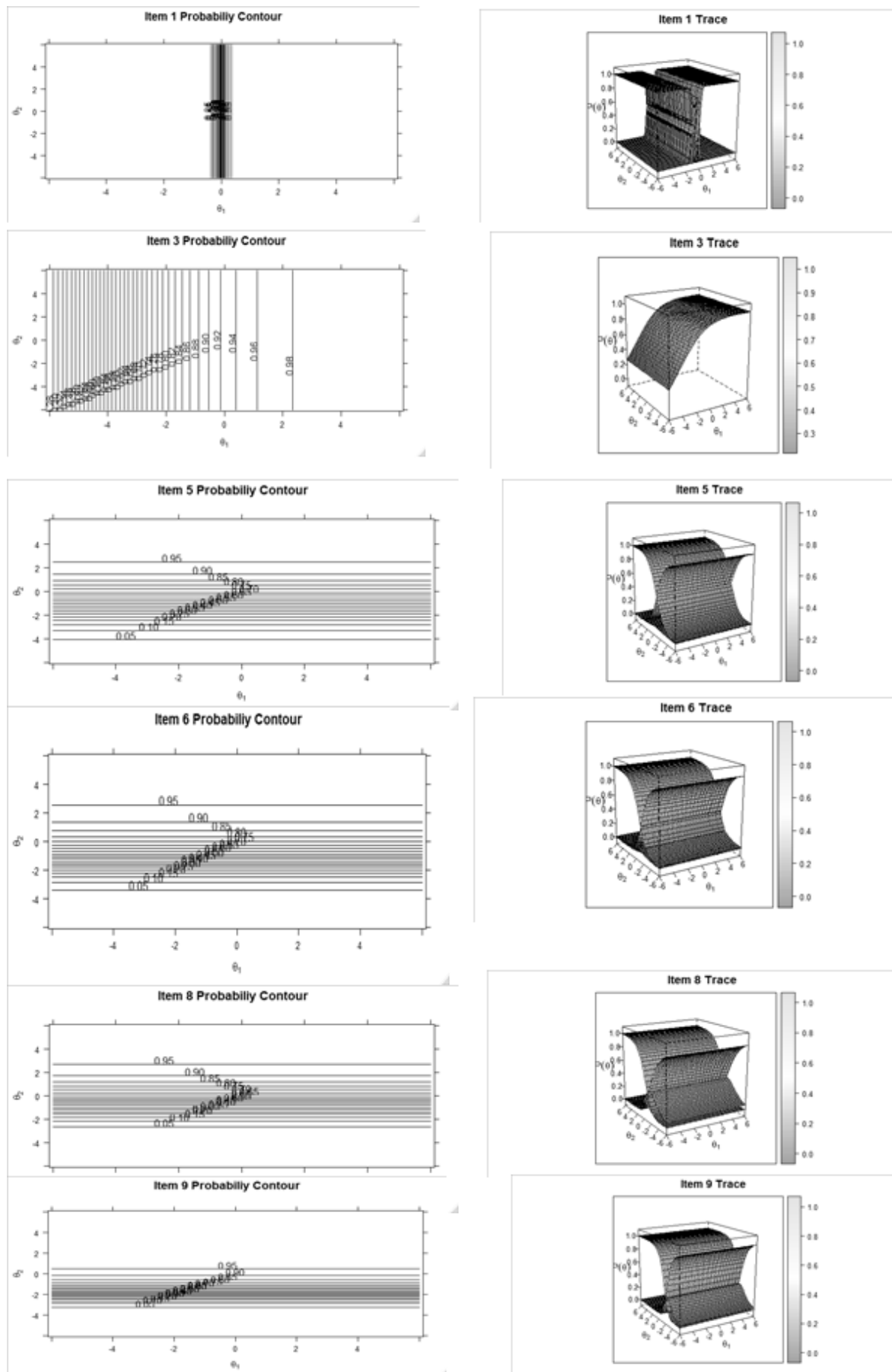*Figure 4.* The characteristic surfaces (right) and level lines for items (left)

Figure 5. The characteristic surfaces (right) and level lines for items (left)

The ensembles of probability contours are fairly uniform, except for the first and ninth problems. The high density of these graphs is explained by the high resolution of the items, which is reflected in the corresponding characteristic surfaces.

Conclusions

1. All three types of models (A, B, C) were adequate to the data, and therefore allow the analysis of tests.

2. The control work is generally "accurate" and can be used in the future to monitor knowledge, although it was somewhat simplistic.

3. The usage of "embedded answers" items is quite appropriate and helps to prevent the cheating and usage of online resources.

4. All three types of models can be used to analyze EA-type problems.

5. The use of multidimensionality (MIRT) makes it possible to separate the impact of individual competencies on the test result and more accurately identify gaps in knowledge.

6. The question of non-compensatory models remains open, the use of which, according to the authors, is appropriate in the further analysis of tests.

REFERENCES

1. Crocker L., Algina J. Introduction to Classical and Modern Test Theory. Belmont: CA:Wadsworth, 2006. 527 p.

2. Челышкова М.Б. Теория и практика конструирования педагогических тестов. Москва: Логос, 2002. 432 с.

3. Rasch G. Probabilistic Models for Some Intelligence and Attainment Tests. Copenhagen: Danish Institute for Educational Research, 1960. 184 p.

4. Muraki E., Carlson E. Full-information factor analysis for polytomous item responses // Applied Psychological Measurement. 1995. № 19 (1). P. 73-90.

5. Linden W., Hambleton R. Handbook of modern item response theory. CRC Press, 2018. 576 p.

6. Baker F. The Basics of Item Response Theory. College Park, MD: ERIC Clearinghouse on Assessment and Evaluation// 2001. URL: http://echo.edres.org:8080/irt/baker/ (дата звернення: 24.04.2020).

7. Лісова Т.В. Моделі та методи сучасної теорії тестів. Ніжин: Видавець ПП Лисенко М.М., 2012. 112 с.

8. Диховичний О.О., Дудко А.Ф. Комплексна методика аналізу якості тестів з вищої математики// Науковий часопис НПУ імені М. П. Драгоманова, серія 2 : Комп'ютерно-орієнтовані системи навчання. 2015. №. 15. С. 139–144.

9. Dykovychnyi O., Zlyvkov V., Kruglova N., Lukomska S., Kotukh O. Using the multidimensional models to the teacher authenticity scale adaptation// Actual Problems of Psychology. 2018. № 14 (1). P. 137–146.

10. Dykhovychnyi O.O., Kruglova N.V., Moskalov I. O. Using of mathematical models for analyzing of the results in psychological Guilford test// Mathematics in Modern Technical University. 2018. № 1. P. –75–89.

11. Dykovychnyi O., Zlyvkov V., Kruglova N., Lukomska S., Kotukh O. Authenticity of the english language teacher's: the validation of authenticity questionnaire using item response theory// Science progress in European countries: new concepts and modern solutions. 2019. P. 335–347.

12. Алєксєєва І.В., Гайдей В.О., Диховичний О.О., Коновалова Н.Р., Федорова Л.Б. Застосування сучасних математичних моделей педагогічного тестування у формуванні та аналізі тестових завдань комплекту «Вища математика»// Дидактика математики: проблеми і дослідження: міжнародний збірник наукових робіт. Донецьк: Вид-во ДонНТУ. 2010. № 33. P. 50–56.

13. Алєксєєва І.В., Гайдей В.О., Диховичний О.О., Коновалова Н.Р., Федорова Л.Б. Застосування математичних моделей тестів у комплекті дистанційної освіти «Вища математика»// Математичні машини і системи. 2010. № 4. P. 89–98.

14. Ким В.С. Разработка тестов по физике. Учебное пособие. Владивосток: Дальневосточный федеральный университет. 2015. 228 с.

15. Дятлова К.Д., Михалева Т.Г. Исследование влияния разнообразия форм тестовых заданий на статистические характеристики теста// Вопросы тестирования в образовании. 2006. №4. С. 65–75.

16. Карданова Е.Ю. О применимости политомической модели Г. Раша к тестовым заданиям различных форм, оцениваемым политомически// Вопросы тестирования в образовании. 2006. №4. С. 44–56.

17. Reckase M. D. Multidimensional item response theory. New York: Springer. 2009. 354 p.

18. The Comprehensive R Archive Network // URL:.http://cran.us.r-project.org (дата звернення: 21.01.2021).

19. Yao L., Schwarz R. A multidimensional partial credit model with associated item and test statistics. An application to mixed format tests// Applied Psychological Measurement. 2006. № 30. P. 469–492.