УДК 51-37:51-77

РАЗРАБОТКА ПРОГРАММНОГО ПРОДУКТА ДЛЯ ОЦЕНКИ КАЧЕСТВА ТЕСТИРОВАНИЯ В РАМКАХ МОДЕЛИ РАША

А. Р. Лобанева

DEVELOPMENT OF SOFTWARE FOR TESTING EVALUATION ACCORDING TO RASCH MEASUREMENT

A. R. Lobanyova

В данной статье рассмотрено использование теории Раша для анализа качества теста, описана разработка программного обеспечения для оценки качества тестирования согласно модели Раша, проанализированы данные, полученные с помощью разработанного программного обеспечения.

The paper focuses on using Rasch measurement to analyze the quality of a test. The development of software to evaluate the quality of testing according to Rasch model is described, the data obtained with the developed software are analyzed.

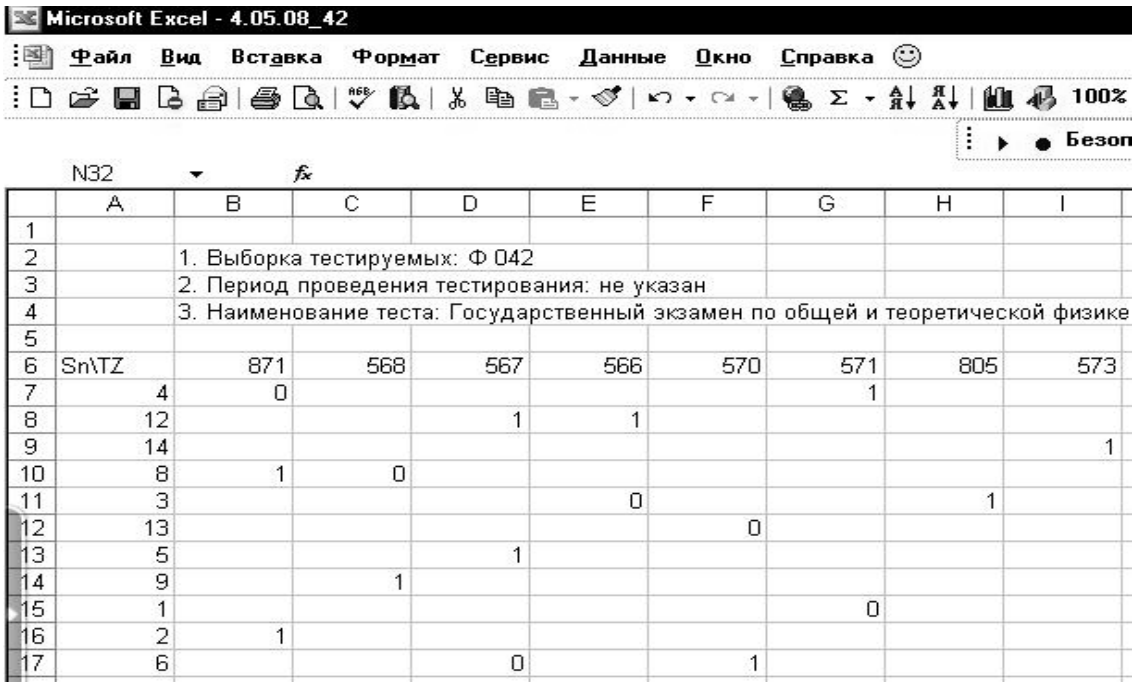
Ключевые слова: тестирование, компьютерное тестирование, теория Раша, сложность тестовых заданий, программное обеспечение.

Keywords: testing, computer testing, Rasch measurement, test items complexity, software.

Оценка качества тестирования

В образовательном процессе для контроля знаний учащихся все чаще стали использовать компьютерное тестирование на основе различных программ и порталов, таких как, например, INDIGO, I-exam, АСТ-тест и др. Тестирование удобно, обладает низкой трудоемкостью, высокой объективностью и позволяет исключить влияние человеческого фактора при проведении и проверке. Однако для того чтобы составленный тест достоверно оценивал знания испытуемых, нужно следить за качеством, как тестовых заданий, так и всего тестирования в целом.

Контроль качества теста удобно проводить, оценивая результаты тестирования, полученные при его выполнении различными группами учащихся. Комплекс АСТ-тест, используемый Кемеровским государственным университетом для организации сеансов тестирования, сохраняет результаты испытуемых в виде матриц тестирования (рис. 1).



	A	B	C	D	E	F	G	H	I
1									
2		1. Выборка тестируемых: Ф 042							
3		2. Период проведения тестирования: не указан							
4		3. Наименование теста: Государственный экзамен по общей и теоретической физике							
5									
6	Sn\TZ	871	568	567	566	570	571	805	573
7	4	0					1		
8	12			1	1				
9	14								1
10	8	1	0						
11	3				0			1	
12	13					0			
13	5			1					
14	9		1						
15	1						0		
16	2	1							
17	6			0		1			

Рис. 1. ***Матрица тестирования комплекса АСТ-тест***

Система оценки в тестировании дихотомическая, т. е. {0;1}, где 0 – задание выполнено неправильно или не выполнено, 1 – задание выполнено правильно. Строка 6 матрицы в примере (рис. 1) показывает номер вопроса (TZ), первый столбец – это номер тестируемого (Sn).

Для подробного анализа результатов тестирования было разработано несколько математических мо-

делей, основными из которых являются Классическая теория тестирования (CTT) и Современная теория тестирования или Item Response Theory (IRT).

Статистическая обработка матриц тестирования позволяет Классической теории тестирования с одной стороны, объективно определить результаты испытуемых, с другой – оценить качество самого теста, тестовых заданий, надежность.

CTT обладает такими достоинствами как: хорошо разработанный математический аппарат; простота и наглядность получаемых выводов.

Однако классическая теория тестирования имеет такие существенные недостатки:

1) зависимость оценки знаний испытуемого от трудности теста. Например, для слишком трудного тестирования баллы будут одинаково плохи для всех тестируемых вне зависимости от их реальных знаний;

2) зависимость трудности тестирования от выборки испытуемых. Если тестирование проводилось на группе испытуемых с высокой успеваемостью, то тестирование ошибочно может быть оценено как легкое.

Несмотря на перечисленные недостатки, оказывающие значительное влияние на точность оценки тестов, Классическая теория широко используется российскими учеными [1]. Существуют так же различные отечественные программные решения для оценки качества тестирования согласно CTT. Например, комплекс АСТ-тест предлагает в дополнение программу АСТ-тест Statistica, предназначенную для получения характеристик качества тестовых заданий на основе данных результатов тестирования.

Item Response Theory (IRT) или Современная теория тестирования – это одна из теорий, в которой смогли отказаться от недостатков CTT. Данная теория является частью общей теории латентно-структурного анализа – Latent Trait Theory (LTT). LTT следует понимать как теорию измерения латентных качеств, т. е. качеств неподдающихся непосредственному измерению (например, уровень знания, понимание дисциплины и т. д.). Следует отметить, что в современной литературе все чаще IRT указывается просто как второе название LTT.

IRT построена на предположении о существовании функциональной связи между наблюдаемыми результатами тестирования и латентными параметрами испытуемых. Несмотря на то, что именно латентные качества испытуемых приводят к наблюдаемым результатам тестирования, на практике ставится обратная задача. Таким образом, основной задачей IRT является переход от некоторых индикаторных (сумма баллов испытуемого) переменных к латентным параметрам [2].

Одна из наиболее удачных моделей для решения этой задачи была предложена датским математиком Георгом Рашем (Georg Rasch). Так появилась Rasch measurement – теория тестирования Раша. Другое ее название – однопараметрическая модель IRT [3].

Математический аппарат однопараметрической модели Раша

Обозначим уровни подготовки испытуемых как θi (i=1…n), где n – количество испытуемых, а трудности заданий теста – βj (j=1…m), где m – количество вопросов. Теория Раша устанавливает связь между множествами параметров θi и βj. Причем θi и βj оцениваются в одной и той же шкале, а функция успеха испытуемого имеет в качестве аргумента разность (θi - βj). Если (θi - βj) > 0 и велика, вероятность успеха i-го испытуемого в j-м задании велика. При условии, что (θi - βj) < 0, но |θi - βj| велик, вероятность успеха i-го испытуемого в j-м задании мала. В качестве единой шкалы Раш предложил ввести интервальную шкалу логитов.

Интервальной является шкала, в которой значимыми считаются расстояния между измеряемыми величинами. Логит – это единица измерения уровней подготовленности участников тестирования и трудностей тестовых заданий в рамках логистических моделей тестирования (от logo – слова, речь) [4]. В общем виде преобразование шкалы логитов имеет вид:

$$\theta 1 = \alpha + \gamma\theta \qquad (1)$$

где θ1 – оценки уровней подготовки тестируемых, θ – оценки параметров в шкале логитов, α – константа, служащая для определения начала шкалы (константа переноса), γ – коэффициент шкалирования для определения размерности шкалы.

$$\beta 2 = \alpha + \gamma\beta, \qquad (2)$$

где β2 – оценки уровней трудности заданий, β – оценки параметров в шкале логитов, α – константа переноса, γ – коэффициент шкалирования [1].

Преобразование исходного балла в логиты направлено на то, чтобы убрать из оценки тестируемого зависимость от трудности заданий теста и из трудности заданий зависимость от конкретной выборки испытуемых [5].

Условная вероятность правильного выполнения i-м испытуемым с уровнем подготовки θi различных по трудности β заданий вводится как:

$$P_i\{x_{ij} = 1 \mid \vartheta_i\} = f(\vartheta_i - \beta), \quad i = 1 \ldots n, \qquad (3)$$

где xij – результат тестирования, θi – логит подготовки i-го испытуемого, β – независимая переменная (xij = 0, если ответ i-го испытуемого на j-й вопрос не верен, xij = 1, если верен).

Аналогично, вероятность правильного выполнения j-го задания трудности βj разными испытуемыми с уровнем подготовки θ:

$$P_j\{x_{ij} = 1 \mid \beta_j\} = \varphi(\vartheta - \beta_j), \quad j = 1 \ldots m, \qquad (4)$$

где xij – результат тестирования, βj – логит трудности j-го задания, θ – независимая переменная [6].

В теории Раша данные функции имеют вид:

$$P_i(\beta) = \{1 + \exp[-1.7(\vartheta_i - \beta)]\}^{-1}, \qquad (5)$$

где β – независимая переменная, θi – логит подготовки i-го испытуемого.

$$P_j(\vartheta) = \{1 + \exp[-1.7(\vartheta - \beta_j)]\}^{-1}, \qquad (6)$$

где θ – независимая переменная, βj – логит трудности j-го задания (1,702 – фактор шкалирования).

Согласно теории Раша можно оценить качество составленного теста и уровень знаний учащихся по матрице тестирования, составленной АСТ-сервером.

Изначально, как и в Классической теории, проводится подсчет правильных p_j и неправильных q_j ответов каждого испытуемого (i = 1…n).

Затем вычисляются их доли. Доля правильных ответов:

$$p_i = \frac{\sum_{j=1}^{m} x_{ij}}{m}, \quad (7)$$

где xij – ответ i-го испытуемого на j-й вопрос, m – количество вопросов.

Соответственно, доля неверных ответов i-го испытуемого:

$$q_i = 1 - p_i, \quad (8)$$

где p_j – доля правильных ответов испытуемого.

Так же выполняется подсчет долей правильных p_j и неправильных q_j ответов на каждое задание (j = 1…m) по аналогичным формулам:

$$p_j = \frac{\sum_{i=1}^{n} x_{ij}}{n}, \quad (9)$$

где xij – ответ i-го испытуемого на j-й вопрос, n – количество испытуемых.

$$q_j = 1 - p_j, \quad (10)$$

где p_j – доля правильных ответов на задание.

Определяются предварительные оценки логитов уровней подготовки i-го тестируемого:

$$\vartheta_i^0 = \ln \frac{p_i}{q_i}, \quad (11)$$

где p_j – доля правильных ответов испытуемых, q_j – доля неправильных ответов испытуемых.

Предварительные логиты трудности заданий:

$$\beta_j^0 = \ln \frac{q_j}{p_j}, \quad (12)$$

где p_j – доля правильных ответов на задания, q_j – долянеправильных ответов на задания.

Вычисляется средние значения логитов уровня подготовки:

$$\overline{\vartheta} = \frac{\sum_{i=1}^{n} \vartheta_i^0}{n}, \quad (13)$$

где ϑ_i^0 – предварительные значения логитов подготовки испытуемых, n – количество испытуемых.

Средние значения логитов трудности заданий:

$$\overline{\beta} = \frac{\sum_{j=1}^{m} \beta_j^0}{m}, \quad (14)$$

где β_j^0 – предварительные значения логитов трудности заданий, m – количество заданий.

Последующие вычисления производятся для того, чтобы свести оценки логитов уровня подготовки тестируемых ϑ_i^0 и логитов трудности заданий β_j^0 в единую шкалу.

Дисперсия V по множеству ϑ_i^0:

$$V = \frac{\sum_{i=1}^{n} (\vartheta_i^0)^2 - n \cdot \overline{\vartheta}^2}{n-1}, \quad (15)$$

где ϑ_i^0 – предварительные значения логитов подготовки испытуемых, $\overline{\theta}$ – среднее значение логитов подготовки испытуемых, n – количество испытуемых.

Дисперсия U по множеству β_j^0:

$$U = \frac{\sum_{j=1}^{m} (\beta_j^0)^2 - m \cdot \overline{\beta}^2}{m-1}, \quad (16)$$

где β_j^0 – предварительные значения логитов трудности заданий, $\overline{\beta}$ – среднее значение логитов трудности заданий, m – количество заданий.

Поправочный коэффициент для логитов уровня подготовки X:

$$X = \sqrt{\frac{1 + U/2.89}{1 - U \cdot V/8.35}}, \quad (17)$$

где U – дисперсия по множеству β_j^0, V – дисперсия по множеству ϑ_i^0.

Поправочный коэффициент для логитов трудности заданий Y:

$$Y = \sqrt{\frac{1 + V/2.89}{1 - U \cdot V/8.35}}, \quad (18)$$

где V – дисперсия по множеству ϑ_i^0, U – дисперсия по множеству β_j^0.

Таким образом, окончательно имеем для значений логитов уровня подготовки тестируемых ϑ_i и логитов трудности заданий β_j выражения:

$$\vartheta_i = \overline{\beta} + X \cdot \vartheta_i^0, \quad (19)$$

где $\overline{\beta}$ – среднее значение логитов трудности заданий, X – поправочный коэффициент, ϑ_i^0 – предварительные значения логитов подготовки испытуемых.

$$\beta_j = \overline{\vartheta} + Y \cdot \beta_j^0, \quad (20)$$

где $\overline{\theta}$ – среднее значение логитов подготовки испытуемых, Y – поправочный коэффициент, β_j^0 – предварительные значения логитов трудности заданий.

Получившиеся формулы (19 – 20) имеют большое практическое значение. С их помощью можно получить объективные оценки параметров испытуемых и заданий, не зависящие друг от друга и выраженные в единой шкале.

На основе полученных вычислений можно построить характеристические кривые испытуемых и заданий. Например, для кривой заданий βj будет параметром, а θ переменной изменяющейся в пределах [-5,5] с интервалом в 0,5 логита. Теоретически θ и β

могут изменяться в интервале (-∞, +∞), но на практике чаще всего при значениях меньше -5 и больше +5, вероятность равна 0 или 1 [7].

По расположению этих кривых относительно друг друга можно сделать вывод о качестве тестирования, а так же выявить и исключить неэффективные задания (рис. 2). В местах, где кривые значительно отстают друг от друга, вероятно, следует добавить задание с промежуточной трудностью. Если же графики накладываются один на другой, то, следовательно, их логиты трудности одинаковы. Это означает, задания соответствующие данным кривым, возможно, имеет смысл удалить из тестирования, как не вносящие вклада в картину трудностей.

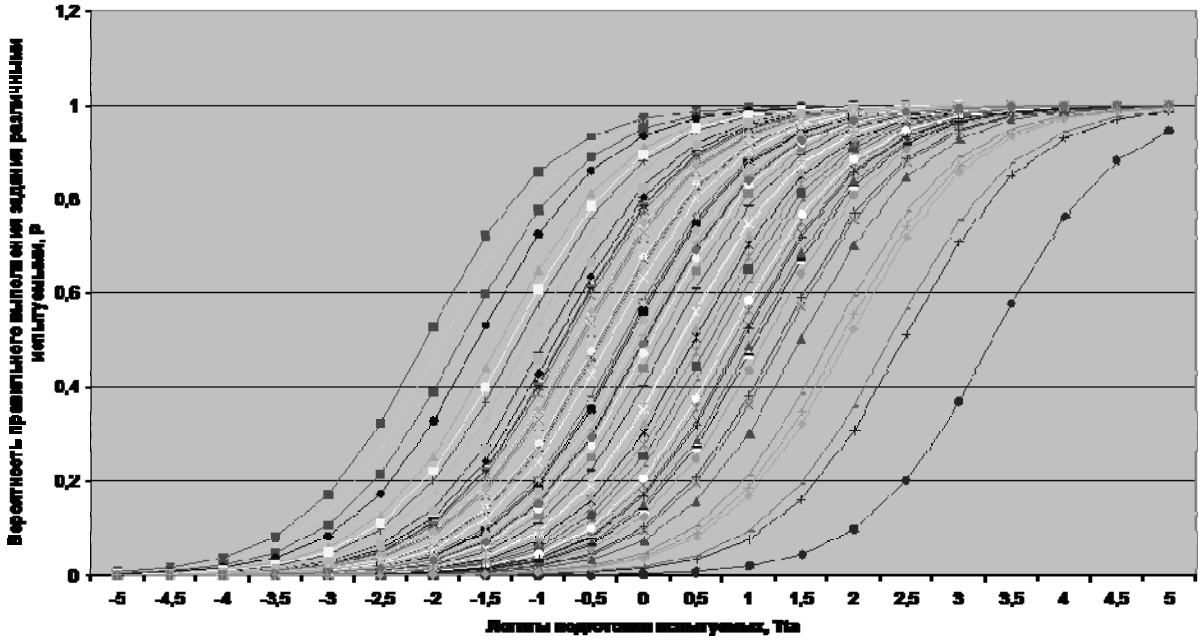


Рис. 2. ***Характеристические кривые заданий теста***

Кроме того, оценить сбалансированность теста по трудности, также можно, рассчитав сумму логитов трудности заданий:

$$\sum_{\beta} = \frac{\sum_{j=1}^{m} \beta_j}{m}, \qquad (21)$$

где β_j – предварительные значения логитов трудности заданий, m – количество заданий.

Если $\sum_{\beta} > 0$, то значит заданий с положительными βj в тесте больше, чем с отрицательными. Следовательно, у теста повышенная трудность. Очевидно, что считается оптимальным, если $\sum_{\beta} \approx 0$ [8].

Таким образом, видно, что в модели Раша, как и в IRT в целом, не ставятся и не решаются фундаментальные проблемы валидности и надежности теста. Тест заранее считается валидным. Вся процедура сводится к получению оценок параметров трудности задания и к измерению латентных качеств испытуемых.

Достоинства и недостатки однопараметрической модели Раша

У математической модели Г. Раша можно выделить ряд достоинств:

1. Единая интервальная шкала для измерения уровня подготовленности и трудности заданий.
2. Возможность дифференциации учеников по уровню подготовленности, а заданий по трудности.
3. Отсутствует зависимость оценки уровня подготовленности от теста.
4. Отсутствует зависимость оценки трудности заданий от уровня подготовленности испытуемых (это означает, что можно использовать данную систему, как для оценки знаний, так и для определения качества теста).
5. Существует возможность соотнести любого тестируемого с любым заданием, что позволит предсказать вероятность правильного выполнения заданий данным испытуемым.
6. Относительная простота математического аппарата. Это обусловлено введением только одного параметра уровня знаний для испытуемого и одного параметра трудности для задания [9].

Основным недостатком данной теории является высокая сложность математико-статистической обработки по сравнению с CTT.

Кроме того, в модели Раша не рассматривается «крутизна» характеристических кривых заданий, по умолчанию этот параметр считается одинаковым для всех кривых.

Однако существуют другие математические модели IRT, которые способны исключить последний недостаток. Например, параметр, определяющий «крутизну» называется дифференцирующей силой заданий и вводится в рассмотрение в моделях А. Бирнбаума (A. Birnbaum), двухпараметрической и трехпараметрической (учитывается возможность случайного выбора правильного ответа в заданиях закрытого типа) [10].

Программное обеспечение для оценки тестирования в рамках модели Раша

Статистическая обработка результатов тестирования методами модели Раша представляет собой длительный и трудоемкий процесс, поэтому имеет смысл его автоматизация.

Система АСТ-Тест сохраняет матрицы тестирования в программе Microsoft Office Excel. Как известно, для работы с приложениями Office существует общая языковая платформа Visual Basic for Applications (VBA). В связи с этим для программной обработки матриц тестирования было закономерным выбрать именно средства VBA.

В разработанном ПО можно выделить две функциональные части:

1. Сбор первичных данных. Точность статистической обработки повышается с увеличением количества тестируемых, поэтому в программе реализована возможность объединения матриц тестирования одного и того же тестирования разными группами испытуемых.

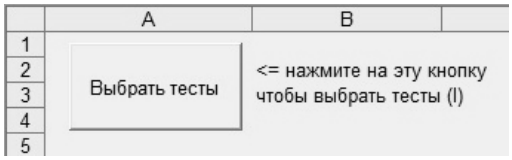


Рис. 3. ***Скриншот кнопки, позволяющей выбрать матрицы для обработки***

2. Статистическая обработка результатов тестирования и их интерпретация. После нажатия на кнопку программа производит расчеты согласно алгоритму, разработанному в рамках Однопараметрической модели Георга Раша.

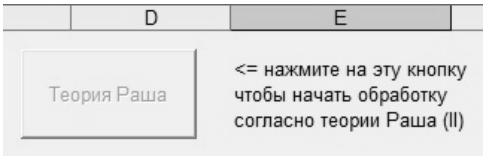


Рис. 4. ***Скриншот кнопки, позволяющей обработать матрицы тестирования согласно модели Раша***

Результаты работы программы при оценке тестирования согласно модели Раша (5 – 21):

1. Расчет логитов трудности заданий и уровня подготовки тестируемых (19 – 20).

2. Вывод о сбалансированности тестирования по трудности на основе среднего значения суммы логитов трудности заданий (21) (рис. 5).

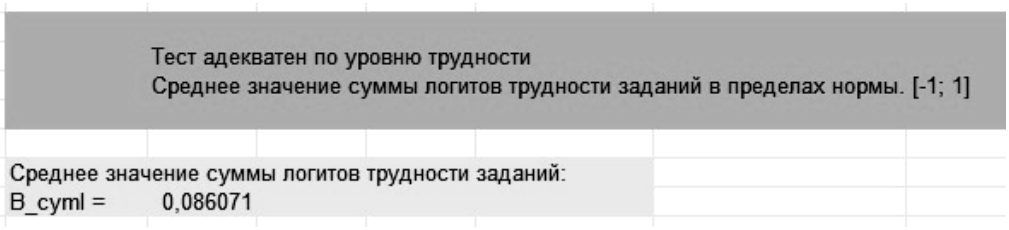


Рис. 5. ***Вывод о сбалансированности тестирования по трудности***

3. Построение характеристических кривых тестовых заданий и испытуемых (рис. 6, 2). Это поможет сделать выводы относительно того, какие задания по уровню трудности не делают вклада в тестирование или же вопросы с какой трудностью следует добавить в тест. Чтобы оценить знания испытуемых при помощи теории Раша следует рассматривать Характеристические кривые испытуемых (рис. 6). Данные кривые характеризуют вероятность решения испытуемым того или иного задания.

При оценивании качества тестирования согласно модели Раша строятся характеристические кривые заданий (рис. 2). На графике (рис. 2) построены кривые соответствующие 80 вопросам, заданным испытуемым. Количество тестируемых составляет 120 человек. Из рис. 2 видно, что характеристические кривые распределены достаточно неравномерно, особая неоднородность отмечается для кривых соответствующих трудным заданиям – кривые значительно отстают друг от друга. Т. о. рекомендуется для улучшения качества тестирования добавить в базу сложные задания с промежуточными уровнями трудностей. Так же из графика очевидно, что имеет место наложение кривых в области средних трудностей заданий. Поэтому одной из рекомендаций по оптимизации тестирования может служить предложение сократить базу заданий, исключив вопросы с одинаковыми логитами трудностей.

Заключение

Реализованный программный продукт имеет большое практическое значение. Он позволяет составителям тестов не только быстро и качественно оценить свой инструментарий по трудности, выявить и переработать тестовые задания, не соответствующие принятым требованиям, а так же получить объективные оценки испытуемых, не зависящие от трудности тестирования. Это в целом положительно скажется на образовательном процессе.

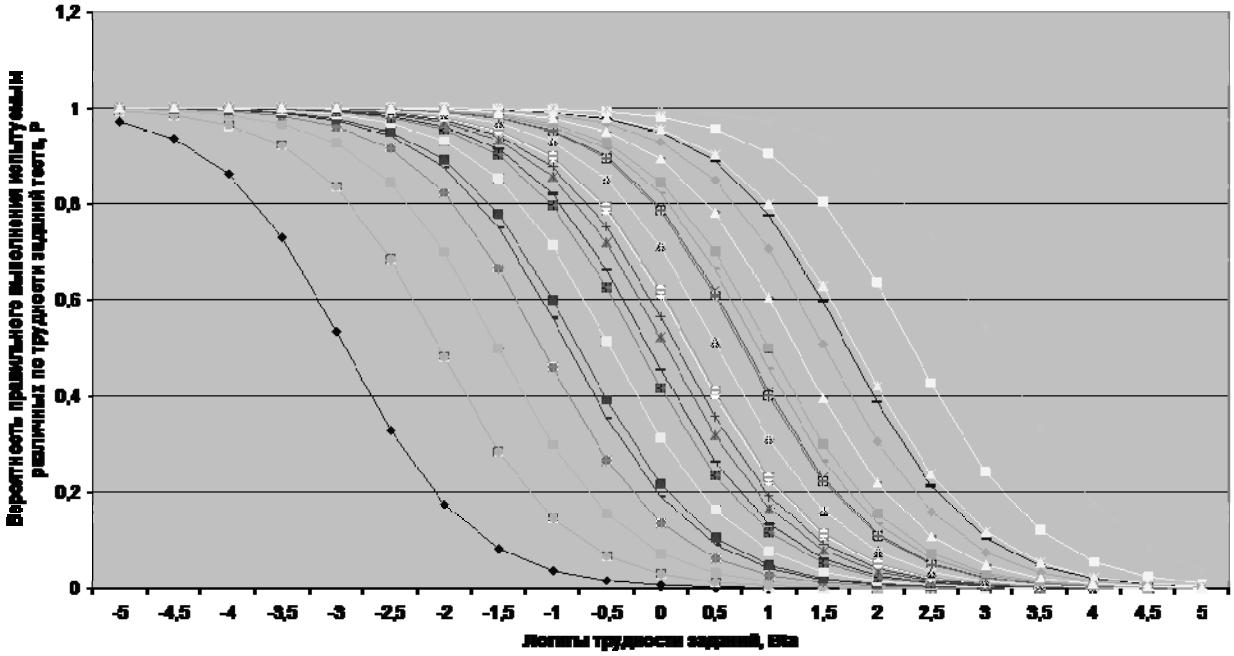


Рис. 6. ***Характеристические кривые испытуемых***

Литература

1. Челышкова, М. Б. Теория и практика конструирования педагогических тестов: учебное пособие. – М.: Логос, 2002. – 432 с.

2. Аванесов, В. С. Педагогическое измерение латентных качеств. Педагогическая диагностика / В. С. Аванесов. – 2003. – № 4. – С. 69 – 78.

3. Аванесов, В. С. Item Response Theory: основные понятия и положения. Педагогические измерения / В. С. Аванесов. – 2007. – № 2. – С. 3 – 28.

4. Национальная психологическая энциклопедия. – 2001. – Режим доступа: http://vocabulary.ru/dictionary/4/-word/logit (дата обращения: 18.04.2013).

5. Колпаков, А. В. Уровни измерений в педагогике / А. В. Колпаков, А. А. Захаров // сб. материалов Всерос. науч. метод. конференции «Дистанционное образование, состояние и перспективы развития». – 1998. – С. 42 – 43.

6. Компьютерное тестирование обучающихся [электронный ресурс]: учебное пособие (мультимедийные учебные материалы) / О. Г. Альтшулер [и др.]. – Кемерово: [б. и.], 2011.

7. Звонников, В. И. Современные средства оценивания результатов обучения / В. И. Звонников, М. Б. Челышкова. – М.: Академия, 2007. – 224 с.

8. Ким, В. С. Тестирование учебных достижений: монография / В. С. Ким. – Уссурийск: УГПИ, 2007. – 214 с.

9. Магранова, Ю. В. Теория тестирования как основа оценивания уровня знаний в современной системе образования / Ю. В. Магранова // Методы социологических исследований: сб. статей. – М.: Теис, 2006. – С. 199 – 208.

10. Проскурин, А. А. Математические модели оценки знаний / А. А. Проскурин // Интеллектуальные технологии и системы: сборник учебно-методических работ и статей. – М.: Эликс+, 2005. – С. 197 – 210.

Информация об авторах:

Лобанева Анастасия Рамильевна – магистрант физического факультета КемГУ, 8-950-265-58-77, lobanyova-n@mail.ru.

Anastasia R. Lobanyova – Master's Degree student at the Faculty of Physics, Kemerovo State University.

Научный руководитель:

Павлова Татьяна Юрьевна – кандидат физико-математических наук, доцент кафедры информационных технологий в образовании КемГУ, 8-913-290-29-15.

TatianaYu. Pavlova – research advisor, Candidate of Physicsand Mathematics, Associate Professorat the Department of TinI Education, Kemerovo State University.