УДК 378.146

Оригинальная статья

# Моделирование адаптивного теста по результатам классического тестирования

**А. И. Безруков, С. А. Акимова***

ФГБОУ ВО «Саратовский государственный технический университет имени Гагарина Ю.А., г. Саратов, Российская Федерация

Адрес: 410054, Российская Федерация, г. Саратов, ул. Политехническая, д. 77

* akimovasa@yandex.ru

## Аннотация

Тестирование широко используется учителями и преподавателями вузов, как для текущего, так и для аттестационного контроля знаний. По сравнению с традиционными методами оценки тестирование — более быстрый, менее трудоемкий и более объективный метод. Однако тестирование часто вызывает нарекания. Его называют «лотереей», вспоминая случаи, когда удачливый троечник выполнял тест на «отлично», а отличник не мог справиться с тем же тестом. Тестирование — случайный процесс, в котором мы можем гарантировать только вероятности, а не его результат. Чтобы снизить воздействие случайных факторов, тест должен соответствовать рекомендациям теории педагогических измерений (IRT). Такие параметры теста, как: метод проведения тестирования, содержание и количество заданий в тесте, их распределение по трудности выполнения, рейтинговая шкала и т. д., должны задаваться исходя из целей тестирования, с учетом подготовленности аудитории. Обычно в педагогической практике используются классические алгоритмы тестирования, однако их цель — с максимальной точностью оценить уровень подготовленности испытуемых — не соответствует цели проведения тестирования — расставить оценки, т. е. отнести каждого испытуемого к одной из заданных категорий. Более подходящим для этого является адаптивное тестирование, базирующееся на алгоритме Байеса, цель которого — наиболее точно оценить вероятности принадлежности тестируемого к каждой категории. Настоящая публикация посвящена сравнительному анализу алгоритмов классического и адаптивного тестирования. Для их сопоставления на результатах классического тестирования был имитирован адаптивный тест. Для сравнения достоверности результатов тестирования были рассчитаны матрицы рассеяния. Эксперимент показал, что достоверность адаптивного тестирования выше, чем классического, при том что число выполненных заданий составляет 60 % от заданий при классическом тестировании. Результаты работы показывают важность внедрения адаптивных методов в системы компьютерного тестирования, используемые при контроле знаний.

**Ключевые слова:** компьютерное тестирование знаний, теория педагогических измерений; классическое и адаптивное тестирование, достоверность результатов тестирования, алгоритм Байеса, имитационное моделирование тестирования

**Финансирование:** исследование выполнено при финансовой поддержке РФФИ в рамках научного проекта № 20-013-00783 «Развитие методов анализа данных для оценки компетенций, формируемых в процессе обучения».

**Конфликт интересов:** авторы заявляют об отсутствии конфликта интересов.

**Для цитирования:** Безруков А. И., Акимова С. А. Моделирование адаптивного теста по результатам классического тестирования // Современные информационные технологии и ИТ-образование. 2023. Т. 19, № 2. С. 498-507.

Original article

# Modeling an Adaptive Test Based on the Results of Classical Testing

**A. I. Bezrukov, S. A. Akimova***
Saratov State Gagarin Technical University, Saratov, Russian Federation
Address: 77 Polytechnicheskaya St., 410054 Saratov, Russian Federation
* akimovasa@yandex.ru

## Abstract

Testing is widely used by teachers and lecturers of universities, both for current and for attestation control of knowledge. Compared to traditional assessment methods, testing is the fastest, least time-consuming and more objective method. However, testing is often controversial. Some call it the "lottery", recalling cases when a successful "C" student completed a test with "excellent", but an "A" student could not cope with the same test. Testing is a random process in which we can only guarantee the probabilities, not its outcome. To reduce the impact of random factors, the test should comply with the recommendations of the item response theory (IRT). Test parameters such as: the method of testing, the content and number of tasks in the test, their distribution according to the difficulty of performing, the rating scale, etc., should be set based on the goals of testing, taking into account the preparedness of the audience. Usually, in pedagogical practice, classical testing algorithms are used, but their goal: to assess the level of preparedness of the subjects with maximum accuracy, does not correspond to the purpose of testing: assign each subject to one of the given categories. More suitable for this purpose is adaptive testing based on the Bayesian algorithm, the purpose of which is to most accurately estimate the probabilities of belonging to each category. This publication is devoted to a comparative analysis of classical and adaptive testing methods. To compare them, an adaptive test was simulated on the results of classical testing. To compare the reliability of the test results, scatter matrices were calculated. The experiment showed that the reliability of adaptive testing is higher than that of classical testing, despite the fact that the number of completed tasks is 60% of the tasks in classical testing. The results of the work show the importance of introducing adaptive methods into computer testing systems used in knowledge control.

**Keywords:** computer knowledge testing, item response theory, classical and adaptive testing; reliability of test results, Bayesian algorithm, testing simulation

**Funding:** the research was carried out with financial support from the Russian Foundation for Basic Research within the framework of scientific project No. 20-013-00783 "Development of Data Analysis Methods for Assessing Competencies Formed in the Learning Process".

**Conflict of intersts:** The authors declare no conflict of interests.

**For citation:** Bezrukov A.I., Akimova S.A. Modeling an Adaptive Test Based on the Results of Classical Testing. *Modern Information Technologies and IT-Education*. 2023;19(2):498-507.

## Введение

Наиболее популярные в России системы компьютерного тестирования АСТ и Moodle реализуют классические алгоритмы тестирования с подсчетом результатов как суммы баллов выполненных тестовых заданий. Балл за каждое задание указывает разработчик теста исходя из трудности выполнения этого задания и его полезности для оценки уровня подготовленности. Одной из проблем, свойственных такому подходу, является субъективность процедуры назначения баллов, ведь автор теста назначает их исходя из своих представлений о трудности, а у тестируемых представления о трудности заданий могут быть совершенно другие.

С математической точки зрения, при классическом тестировании решается следующая задача. Считается, что каждый испытуемый обладает некой скрытой (латентной) характеристикой, называемой «уровень подготовленности». Задача тестирования — оценить эту характеристику с максимальной точностью. Такая постановка адекватна конкурсному тестированию, по результатам которого формируется элитная группа (например, команда для участия в олимпиаде).

При текущем контроле знаний учащихся решается совершенно другая задача: нужно отнести каждого учащегося к одному из заранее заданных классов, например «отличник», «хорошист», «троечник» и «неуспевающий». Адекватным способом такого контроля является адаптивное тестирование. При адаптивном тестировании происходит последовательное уточнение вероятностей принадлежности тестируемого к каждому из указанных классов. Каждое следующее задание выбирается так, чтобы его успешное выполнение или невыполнение давало максимальную информацию о том, какому же классу принадлежит тестируемый.

Многочисленные теоретические публикации [1-5], а также наши эксперименты на имитационной модели компьютерного тестирования[1] говорят о том, что применение адаптивного тестирования для целей текущего контроля значительно эффективнее классического тестирования (для достижения сопоставимого уровня достоверности при адаптивном тестировании требуется выполнить почти в два раза меньше заданий).

## Обработка результатов классического тестирования

В качестве исходных данных использованы результаты тестирования по дисциплине «Дискретная математика», проведенного с использованием Moodle.

Для имитации адаптивного тестирования нам потребуется оценить некоторые дополнительные характеристики каждого задания, которые в классических системах компьютерного тестирования (например, в Moodle) не используются или назначаются произвольно.

Уровень трудности задания в популярных системах тестирования (например, в АСТ и Moodle) задается составителем теста и отражает его представления о трудности заданий. Соотношение трудностей заданий для тестируемых может существенно отличаться от представлений разработчика. В теории педагогических измерений (Item Response Theory, IRT) уровни подготовленности испытуемых и трудности заданий принято измерять в логитах (логарифмах отношения шансов)[2] [6-12]. Обозначим:

$N_t(i)$ — количество попыток выполнения задания $i$;

$N_s(i)$ — количество успешного выполнения задания $i$.

Тогда трудность выполнения задания, выраженная в логитах, равна логарифму отношения шансов не выполнить или выполнить задание:

$$\delta(i) = \ln\left(\frac{N_t(i)-N_s(i)}{N_s(i)}\right) \quad (1)$$

Еще одна важная характеристика задания — вероятность угадывания правильного ответа без выполнения задания. Вероятность угадывания для каждого задания $C(i)$ оценим из следующих предположений: если ответ в задании выбирается из $n$ альтернатив, $C(i) = \frac{1}{n}$. Например, если нужно выбрать один правильный ответ из пяти предложенных, $C(i) = 0{,}2$. При множественном выборе $k$ правильных ответов из $n$ предложенных вероятность угадывания резко снижается $C(i) = \frac{1}{C_n^k}$, а если $k$ неизвестно, $C(i) = \frac{1}{\sum_k C_n^k}$. Если в задании требуется сопоставить объекты из двух списков, каждый длиною $n$, мы получим $n!$ различных комбинаций $C(i) = \frac{1}{n!}$. Вероятность угадывания в задании с открытым ответом существенно ниже рассмотренных, поэтому в нашем исследовании мы будем считать ее равной нулю.

Для подведения итогов тестирования нам нужно как-то сформулировать свое представление о том, кого можно считать отличником, а кого двоечником. Обозначим: $P_c^+(i)$ — вероятность того, что испытуемый, принадлежащий классу $c$, успешно справится с заданием $i$. Мы можем задать эти вероятности экспертно (например, вероятность того, что отличник справится с самым сложным заданием, равна 90 %; хорошист — 60%; троечник — 20%; неуспевающий — 2%). Однако это весьма трудоемко, а назначить правдоподобные вероятности, соответствующие здравому смыслу, весьма проблематично. Проще для каждого класса назначить средний (типовой) уровень подготовленности $\theta_c$, выраженный в логитах, а затем вычислить нужные вероятности с помощью трехпараметрической модели Раша[3] (формулы Бирнбаума[4]) [13-17]:

$$P_c^+(i) = P\left(\theta_c, \delta(i)\right) = C(i) + (1 - C(i)) \cdot \frac{\exp\left(\alpha\cdot\left(\theta_c-\delta(i)\right)\right)}{1+\exp\left(\alpha\cdot\left(\theta_c-\delta(i)\right)\right)} \quad (2)$$

[1] Безруков А. И. Имитационная модель для оценки достоверности и точности результатов тестирования / А. И. Безруков, С. А. Акимова, Г. Г. Погожильская // Математическое моделирование и информационные технологии в исследованиях по физике и педагогике. Саратов : Центр Просвещение, 2017. С. 8-15. EDN: ZBDRTT

[2] DeMars C. Item response theory: Understanding statistics measurement. Oxford University Press, Inc., 2010. 144 p.

[3] Partchev I. A visual guide to item response theory. Friedrich Schiller University Jena, 2004. 61 p. URL: https://www.stat.cmu.edu/~brian/PIER-methods/For%202013-02-28/Readings/Interactive%20pdf%20on%20Item%20Response%20Theory.pdf (дата обращения: 11.02.2023).

[4] Birnbaum A. Some Latent Trait Models and Their Use in Inferring and Examinee's Ability // Statistical Theories of Mental Test Scores ; ed. by F. M. Lord, M. Novick. Addison-Wesley Publ., 1968. P. 397-479.

где $C(i)$ — вероятность угадывания задания $i$; $\alpha \approx 1{,}71$ — средняя чувствительность заданий.

Для оценок $\theta_c$ разобьем весь диапазон трудностей, выраженных в логитах, на четыре интервала и выберем в каждом интервале типичное значение $\theta_c$. Используя формулу (2), вычислим вероятности $P_c^+(i)$ ($c$=2, 3, 4, 5; i=1…N) и проведем их валидацию (оценим их соответствие нашим представлениям). Вероятность того, что представитель данного класса не справится с заданием, естественно равна дополнительной вероятности:

$$P_c^-(i) = 1 - P_c^+(i) \quad (3)$$

Если нас что-то не устроит, выберем другие $\theta_c$ и повторим процедуру. Эта процедура аналогична заданию рейтинговой шкалы, отличие заключается в том, что в результате мы получаем согласованные и осмысленные с точки зрения педагогики оценки вероятностей выполнения для каждого задания.

Байесовский алгоритм адаптивного тестирования позволяет уточнять вероятности принадлежности испытуемого каждому из классов, используя формулу Байеса для условных вероятностей [18-20]. Если у нас есть событие A и несколько взаимно исключающих гипотез, объясняющих появление этого события: $B_1$, $B_2$, … $B_n$, то вероятность каждой гипотезы можно оценить по формуле:

$$P(B_c|A) = \frac{P(A|B_c)\cdot P(B_c)}{P(A)} \quad (4)$$

В нашем случае $A$ — успешное выполнение задания данным студентом;

$B_c$, — событие «студент принадлежит классу $c$».

Так как принадлежность разным классам события взаимоисключающие, вероятность события A можно оценить по формуле полной вероятности:

$$P(A) = \sum_{c=1}^{4} P(A|B_c)\cdot P(B_c) \quad (5)$$

Вставляя (5) в (4) получаем:

$$P(B_c|A) = \frac{P(A|B_c)\cdot P(B_c)}{\sum_{c=1}^{4} P(A|B_c)\cdot P(B_c)} \quad (6)$$

Формула (6) позволяет последовательно уточнять вероятности принадлежности испытуемого к классам после выполнения каждого задания. Обозначим $P_c^k(j)$ — вероятность того, что на шаге $k$ студент $j$ принадлежит классу $c$. Будем считать $P(B_c) = P_c^{k-1}(j)$ априорными (найденными при выполнении предыдущих шагов) вероятностями принадлежности к классам. $P(A|B_c) = P_c^+(i)$ — вероятность, того, что студент из класса $c$ выполнит задание $i$. Тогда апостериорная вероятность, учитывающая, что испытуемый $j$ справился с заданием, будет равна:

$$P_c^k(j) = P(B_c|A) = \frac{P(A|B_c)\cdot P(B_c)}{\sum_{c=1}^{4} P(A|B_c)\cdot P(B_c)} = \frac{P_c^+(i)\cdot P_c^{k-1}(j)}{\sum_{c=1}^{4} P_c^+(i)\cdot P_c^{k-1}(j)} \quad (7)$$

Если испытуемый с заданием не справился:

$$P_c^k(j) = P(B_c|\neg A) = \frac{P(\neg A|B_c)\cdot P(B_c)}{\sum_{c=1}^{4}(1-P(A|B_c))\cdot P(B_c)} = \frac{P_c^-(i)\cdot P_c^{k-1}(j)}{\sum_{c=1}^{4} P_c^-(i)\cdot P_c^{k-1}(j)} \quad (8)$$

Если у нас нет сведений о том, к какому классу, скорее всего, принадлежит испытуемый, будем считать, что на первом шаге априорные вероятности принадлежности каждому классу одинаковы: $P_c^0(j) = \frac{1}{4}$,

c = 2, 3, 4, 5.

В качестве первого задания выбирается среднее по трудности из реально выполненных данным студентом заданий. По результату его выполнения вероятности принадлежности классам уточняются по формулам (7) или (8) в зависимости от того, справился или не справился испытуемый с этим заданием. Каждое следующее задание выбирается так, чтобы факт его выполнения или невыполнения давал максимальную информацию о принадлежности классам. Для выбора такого задания используется информационная функция, максимум которой достигается на требуемом задании[5]:

$$Inf(i,\theta) = \alpha \cdot \left(\frac{1-P}{P}\right) \cdot \left(\frac{P-c}{1-c}\right)^2, \quad (9)$$

где $P$ — вероятность выполнения задания $i$ испытуемым с уровнем $\theta$.

Для каждого ранее неиспользованного задания вычисляем взвешенную по вероятности информационную функцию.

$$InfPW(i) = \sum_{c=2}^{5} Inf(i,\theta_c)\cdot P_c^0 \quad (10)$$

Выбираем задание $i^*$, для которого эта функция максимальна, и предложим его испытуемому. Вычисляем новые апостериорные вероятности по формулам (7) или (8). Алгоритм завершает работу, если число шагов (выполненных заданий) достигает указанного значения или максимальное изменение вероятностей принадлежности классом меньше заданного значения $Err$.

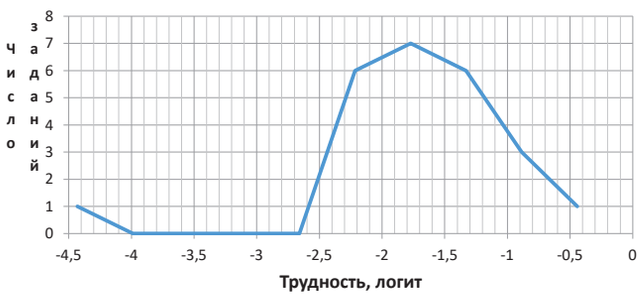
$$\max_i |P_c^k(i) - P_c^{k-1}(i)| < Err \quad (11)$$

Для сопоставления эффективности классического и адаптивного алгоритмов проведем численный эксперимент.

В качестве исходных данных используем реальные результаты тестирования, проведенного в среде Moodle. Для этого выгрузим отчет о тесте в Excel-файлы. Если при тестировании использовался случайный выбор заданий, нам потребуется матрица предъявления заданий каждому студенту. В данной работе использован тест по дисциплине «Дискретная математика», содержащий 25 заданий, разработанный в среде Moodle. Тестирование проводилось с целью аттестации 85 студентов. Для дальнейшего анализа нами на языке VBA была написана программа, сокращающая трудоемкость вычислений.

Оценим трудности заданий в логитах. Трудности каждого задания в логитах оценивались по результатам его выполнения по формуле (1). Результаты расчета проиллюстрированы на графике (Рис. 1), диапазон трудности — от -4,4 до 0,44 логит.

[5] Partchev I. A visual guide to item response theory. Friedrich Schiller University Jena, 2004. 61 p. URL: https://www.stat.cmu.edu/~brian/PIER-methods/For%202013-02-28/Readings/Interactive%20pdf%20on%20Item%20Response%20Theory.pdf (дата обращения: 11.02.2023).




Р и с. 1. Распределение заданий по трудности

F i g. 1. The distribution of tanks by difficulty levels

*И с т о ч н и к:* здесь и далее в статье все таблицы и рисунки составлены авторами.

*S o u r c e:* Hereinafter in this article all tables and figures were made by the authors.

Чтобы с достаточной достоверностью оценить как слабо, так и хорошо подготовленных испытуемых, распределение трудностей заданий в тесте должно быть близко к равномерному. Следует отметить, что в исследуемом тесте явно не хватает как простых, так и трудных заданий [17]. Такой тест более пригоден для оценки по двухбалльной шкале с уровнем отсечения порядка -1,75 логит.

Оценка фактической трудности заданий после пробного тестирования с последующим анализом распределения заданий по трудности позволяет оценить пригодность создаваемого теста для решения поставленных задач тестирования и улучшить тест путем добавления заданий недостающих трудностей.

Для каждого задания был определен его тип, что представлено в таблице 1.

Т а б л и ц а 1. **Типы заданий**

T a b l e 1. **Types of tasks**

| Тип | Описание |
|---|---|
| Список | Выбор из списка предлагаемых ответов одного или нескольких правильных решений |
| Соответствие | Установить соответствие между объектами двух списков. Например, между терминами и их определениями |
| Значение | Требуется ввести конкретное значение результата |

В соответствии с указанным типом задания программа запрашивала требуемые характеристики задания и вычисляла вероятность угадывания ответа без выполнения задания, что указано в таблице 2.

Т а б л и ц а 2. **Распределение заданий по типам**

T a b l e 2. **Distribution of tasks by type**

| Тип | Число заданий | Доля |
|---|---|---|
| Список | 21 | 84% |
| Сопоставление | 0 | 0% |
| Значение | 4 | 16% |

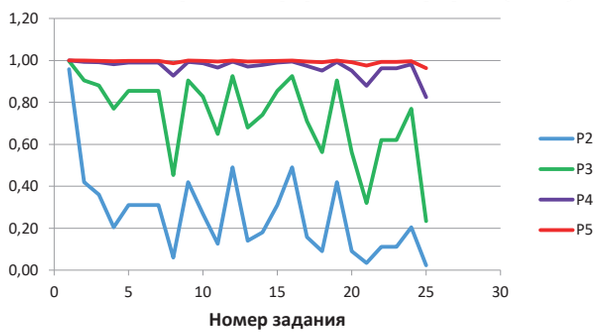
Большее число заданий имеет тип «Список», поэтому средняя вероятность угадывания равна 20 %. Это существенно снижает достоверность результатов тестирования. Зададим рейтинговую шкалу. В таблице приведен один из возможных вариантов разделения диапазона трудностей заданий на интервалы.

Т а б л и ц а 3. **Разделение диапазона трудностей заданий на интервалы**

T a b l e 3. **Dividing the range of task difficulties into intervals**

| Уровень, логит | | | Оценка |
|---|---|---|---|
| От | До | Средний | |
| -3 | -2,2 | -2,6 | Неудовлетворительно |
| -2,2 | 0 | -1,1 | Удовлетворительно |
| 0 | 1 | 0,5 | Хорошо |
| 1 | 2 | 1,5 | Отлично |

Для данной рейтинговой шкалы, используя формулу Бирнбаума (2), вычислим вероятности того, что испытуемый, имеющий уровень подготовки $\Theta_c$, выполнит задание трудности $\delta_i$. Результаты вычислений проиллюстрированы на графике (Рис. 2).




Р и с. 2. Вероятности выполнения заданий различной трудности

F i g. 2. Probabilities of completing tasks of various difficulty levels

Вычисленные вероятности позволят оценить рейтинговую шкалу с точки зрения ее эффективности. В таблице приведены математические ожидания суммы баллов, которые наберут типовые представители каждого класса, если за каждое решенное задание им будут начислять по 0,4 балла, и они будут выполнять классический тест из 25 заданий.

Т а б л и ц а 4. **Математические ожидания суммы баллов тестируемых**

T a b l e 4. **Expected value of test takers' total scores**

| Оценка | Неуд. | Уд. | Хор. | Отл. |
|---|---|---|---|---|
| Сумма баллов | 2,64 | 7,35 | 9,68 | 9,94 |

Изменим наш тест: назначим баллы пропорционально вычисленной трудности задания, так, чтобы отличник в среднем получал сумму баллов, равную 10, а за задание с минимальной трудностью $\delta_{min}$ баллы не начислялись:

$$Ball(i) = k \cdot (\delta(i) - \delta_{min}), \quad (12)$$

где $Ball(i)$ — бал, начисляемый при успешном выполнении задания i;

k — коэффициент, вычисляемый из выражения:

$\sum_i Ball(i) \cdot P^{+}_{c=5}(i) = 10$.

В таблице 5 приведены средние значения баллов, получаемых испытуемыми из различных классов.

В таблице 5 приведены средние значения баллов, получаемых испытуемыми из различных классов.

Таблица 5. **Средние значения баллов тестируемых студентов**
Table 5. **Average scores of tested students**

| Класс | 2 | 3 | 4 | 5 |
|---|---|---|---|---|
| Одинаковые баллы | 2,64 | 7,35 | 9,68 | 9,94 |
| Баллы пропорциональны трудностям | 2,05 | 6,84 | 9,60 | 9,92 |

Видно, что нарастание среднего балла при учете трудности задания происходит несколько быстрее, чем при использовании одинаковых баллов, но существенной роли это не играет. Из этого следует, что простой пересчет баллов в соответствии с фактическими трудностями заданий нецелесообразен.

Для оценки достоверности результатов тестирования построим матрицу рассеяния: таблицу, столбцы которой соответствуют классам испытуемых, а строки классам, к которым отнесены испытуемые в результате тестирования. В ячейки таблицы запишем соответствующие вероятности. Например, на пересечении столбца «отличник» и строки «троечник» запишем вероятность того, что испытуемый с уровнем подготовленности, равным типовому уровню отличника, получит при тестировании оценку «Удовлетворительно». Очевидно, чем меньше сумма недиагональных элементов такой матрицы, тем достовернее результаты тестирования.

Для оценки таких вероятностей проведем численный эксперимент. Пусть у нас есть 10 заданий, для каждого из которых указаны: трудность *δ(i);* балл *Ball(*i), начисляемый при его успешном выполнении; и вероятности *{Pc} c=2, 3, 4, 5* того, что студент из класса *c* справится с этим заданием, которые мы вычислим по формуле Бирнбаума.

Построим матрицу всех возможных ситуаций *M(s,i)*: если задание выполнено, соответствующий элемент равен 1, если нет, то 0. Для 10 заданий матрица будет иметь $2^{10}$=1024 стокситуаций и 10 столбцов, соответствующих заданиям. Для каждой ситуации вычислим суммарный балл *SumBal(s)l:*

$$SumBall(s) = \sum Ball(i), \quad (13)$$

где суммирование ведется по всем заданиям, выполненным в данной ситуации. Будем считать события выполнения или невыполнения разных заданий независимыми. Тогда вероятность появления этой ситуации для испытуемого из каждого класса будет равна произведению:

$$P_c(s) = \prod_i P_c^*(i,s), \quad (14)$$

где:

$$P_c^*(i,s) = \begin{cases} P_c(i) \text{ если задание выполнено} \\ 1 - P_c(i) \text{если не выполнено} \end{cases}$$

Для вычисления элементов матрицы рассеяния введем обозначения $P_{f,t}$ — вероятность того, что испытуемый, фактически принадлежащий классу f, будет отнесен к классу *t*. Тогда:

$$P_{f,t} = \sum P_f, \quad (15)$$

где суммирование ведется по всем ситуациям, приводящим к суммарному баллу, соответствующему классу *t*.

В таблице 6 приведена матрица рассеяния для выбранной рейтинговой шкалы. Слева указана шкала («уровень в баллах»), справа приведены вероятности отнесения типового студента к каждому из классов.

Таблица 6. **Матрица рассеяния для выбранной рейтинговой шкалы**
Table 6. **Scattering Matrix for the selected rating scale**

| | Уровень в баллах | | Отнесено к классу | | | |
|---|---|---|---|---|---|---|
| | От | До | 2 | 3 | 4 | 5 |
| Неудовлетворительно | 0 | 2 | 0,674 | 0,254 | 0,065 | 0,007 |
| Удовлетворительно | 2 | 3 | 0,450 | 0,357 | 0,161 | 0,030 |
| Хорошо | 3 | 4 | 0,274 | 0,373 | 0,267 | 0,078 |
| Отлично | 4 | 5 | 0,141 | 0,316 | 0,357 | 0,163 |

Если рейтинговая шкала перекрывает весь диапазон трудностей, сумма вероятностей, записанных в столбцах «2», «3», «4» и «5» в каждой строке равна единице, т. к. любого испытуемого можно отнести к одному из заданных классов. А вот суммы по столбцам могут отличаться от единицы: если сумма меньше 1, к данному классу было отнесено неоправданно мало ситуаций, если больше — слишком много.

Качество теста можно оценить суммой недиагональных элементов, отражающих ошибочное отнесение испытуемого. Для данной матрицы эта сумма равна 2,507. Отметим, что элементы матрицы рассеяния зависят от распределения баллов по заданиям и от рейтинговой шкалы. Изменяя эти характеристики, мы можем добиться минимальной суммы недиагональных элементов.

С увеличением числа заданий количество возможных ситуаций экспоненциально возрастает. Так для 30 заданий оно равно $2^{30} \approx 10^9$. Подобные расчеты уже нельзя выполнить на персональном компьютере. В этих случаях элементы матрицы рассеяния приходится оценивать с помощью имитационной модели [21].

Для сопоставления качества классического и адаптивного тестирования проведем численный эксперимент с использованием результатов классического тестирования. Для каждого испытуемого выпишем задания, которые он выполнял при тестировании, и результат выполнения этих заданий. Используя формулу (9), вычислим информационные функции $Inf(i,\theta_c)$ для каждого задания *i* и класса *c*.

Имитацию адаптивного тестирования проведем для каждого испытуемого следующим образом.

1. Будем считать, что априорные вероятности принадлежности испытуемого каждому из четырех классов одинаковы и равны 0,25. Выберем из списка заданий, выполняемых данным испытуемым, задание среднего уровня трудности.
2. В зависимости от того, выполнил или не выполнил испытуемый это задание, вычислим апостериорные вероятности по формулам (7) или (8). Пометим это задание как использованное.
3. Сравним, насколько новые оценки вероятностей отличаются от предыдущих. Если максимальное отличие ниже заданного уровня (например, 0,001) или уже все задания использованы, закончим имитацию тестирования.
4. Найдем еще неиспользованное задание, для которого взвешенная информационная функция максимальная. Повторим шаги, начиная со второго.

При оценке элементов матрицы рассеяния воспользуемся следующими предположениями. Для каждого испытуемого алгоритм вычисляет вероятности принадлежности каждому классу. Будем считать, что тест правильно отнес всех испытуемых к классам. Тогда вероятность того, что двоечник будет отнесен к классу двоечников, равна отношению суммы вероятностей отнесения к двоечникам каждого из испытуемых к числу испытуемых, отнесенных к двоечникам. Аналогично вероятность отнесения к троечникам равна отношению суммы вероятностей отнесения двоечников к троечникам к числу двоечников. В таблице представлена матрица рассеяния, вычисленная подобным образом.

Таблица 7. **Матрица рассеяния адаптивного теста**

Table 7. **Adaptive Test Scattering Matrix**

| Класс | 2 | 3 | 4 | 5 | ∑ по строкам |
|---|---|---|---|---|---|
| 2 | 0,7865 | 0,2097 | 0,0037 | 0,0001 | 1 |
| 3 | 0,0379 | 0,9167 | 0,0448 | 0,0006 | 1 |
| 4 | 0,0000 | 0,1860 | 0,6972 | 0,1168 | 1 |
| 5 | 0,0000 | 0,0008 | 0,3667 | 0,6325 | 1 |
| ∑ по столбцам | 0,8245 | 1,3131 | 1,1124 | 0,7500 | |

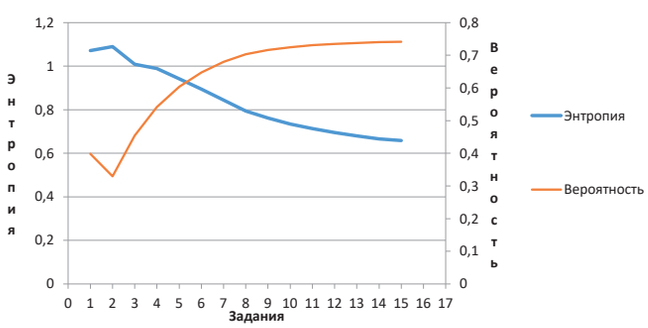



Рис. 3. Изменение вероятности и неопределенности принадлежности классу в зависимости от числа выполненных заданий

Fig. 3. Changes in the probability and uncertainty of class membership depending on the number of completed tasks

Сумма недиагональных элементов равна 0,97, т. е. меньше, чем при классическом тестировании. При этом общее число выполняемых заданий составило 65 % от исходного числа заданий. Изменяя правило окончания алгоритма (например, если максимальное расхождение вероятностей будет не 0,001, а 0,002, сумма недиагональных элементов увеличится до 1,037, т. е. останется меньше, чем при классическом тестировании, но число заданий сократится до 60 % от первоначального).

На рисунке 3 показано, как меняется энтропия (неопределенность) принадлежности классам и вероятность принадлежности «правильному» классу одного из испытуемых с увеличением числа заданий.

Предъявляя первые задания, алгоритм как бы грубо оценивает испытуемого. После второго задания энтропия начинает устойчиво снижаться, а одна из вероятностей повышаться. Остальные задания только уточняют (повышают вероятность) найденной гипотезы о принадлежности к испытуемого к одному из классов. Анализ подобных графиков позволяет правильно выбрать баланс между достоверностью результатов тестирования и трудоемкостью теста [22-25].

## Заключение

Проведенная имитация адаптивного тестирования демонстрирует преимущество этого метода перед классическим тестированием, если цель тестирования — промежуточная аттестация. Рассмотренные в статье модели выбора рейтинговой шкалы для конкретного теста и построения матрицы рассеяния позволят разработчикам тестов более осознано назначать параметры тестирования и оценивать качество тестов. Периодическое отслеживание и статистический анализ результатов тестирования позволяют постоянно уточнять распределение трудностей заданий в тесте, разумно назначать рейтинговую шкалу для данного теста и целей его использования, сокращать время и трудоемкость проведения тестирования без снижения (а иногда с существенным повышением) достоверности результатов тестирования.

Внедрение предлагаемых методик в существующие системы компьютерного тестирования, организация ведения баз тестовых заданий позволят одновременно повысить достоверность результатов тестирования и снизить трудоемкость проведения тестирования.

**Благодарности:** авторы выражают благодарность анонимным рецензентам этой публикации за изучение и содержательные замечания.

**Acknowledgments:** The authors would like to thank the anonymous reviewers of this publication for their review and insightful comments.

## Список использованных источников

[1] Деменченок О. Анализ моделей для адаптивного тестирования // Педагогические измерения. 2011. № 1. С. 3-18. EDN: OJOMQT

[2] Белоус Н. В., Куцевич И. В. Модель адаптивного контроля знаний // Радиоэлектроника, информатика, управление. 2010. № 1(22). С. 39-44. EDN: SAZQLJ

[3] Корнилов Д. Е. Сравнительный анализ адаптивной и классической модели тестирования // Перспективы развития информационных технологий. 2014. № 19. С. 119-124. EDN: SEFOEJ

[4] A flexible online platform for computerized adaptive testing / S. Oppl [et al.] // International Journal of Educational Technology in Higher Education. 2017. Vol. 14. Article number: 2. https://doi.org/10.1186/s41239-017-0039-0

[5] Nurakhmetov D. Reinforcement Learning Applied to Adaptive Classification Testing // Theoretical and Practical Advances in Computer-based Educational Measurement. Methodology of Educational Measurement and Assessment ; ed. by B. Veldkamp, C. Sluijter. Cham : Springer, 2019. P. 325-366. https://doi.org/10.1007/978-3-030-18480-3_17

[6] Optimising Moodle quizzes for online assessments / S.H.P.W. Gamage [et al.] // International Journal of STEM Education. 2019. Vol. 6. Article number: 27. https://doi.org/10.1186/s40594-019-0181-4
[7] Romanenko D. Computer Testing as a Form of Students' Knowledge Control // CEUR Workshop Proceedings. 2020. Vol. 2770. P. 196-203. URL: https://ceur-ws.org/Vol-2770/paper23.pdf (дата обращения: 11.02.2023).
[8] Maslova O., Kuklina A. Adaptive testing model based on IRT techniques // The Scientific Heritage. 2020. Vol. 43. P. 41-44. EDN: AGLIOQ
[9] Riza L. S., Nurjanah N. S., Wihardi Y. The development of computer adaptive test and item response theory with 4 parameters based the logistics model // Journal of Physics: Conference Series. 2019. Vol. 1280. Article number: 032033. https://doi.org/10.1088/1742-6596/1280/3/032033
[10] Wulandari F., Hadi S., Haryanto. Computer-based Adaptive Test Development Using Fuzzy Item Response Theory to Estimate Student Ability // Computer Science and Information Technology. 2020. Vol. 8, issue 3. P. 66-73. https://doi.org/10.13189/csit.2020.080302
[11] The Integration of Classical Testing Theory and Item Response Theory / Z. F. Hu [et al.] // Psychology. 2021. Vol. 12. P. 1397-1409. https://doi.org/10.4236/psych.2021.129088
[12] Kimura T. The impacts of computer adaptive testing from a variety of perspectives // Journal of Educational Evaluation for Health Professions. 2017. Vol. 14. Article number: 12. https://doi.org/10.3352/jeehp.2017.14.12
[13] Sijtsma K. Developments in Measurement of Persons and Items by Means of Item Response Models // Behaviormetrika. 2001. Vol. 28. P. 65-94. https://doi.org/10.2333/bhmk.28.65
[14] Relevance and advantages of using the item response theory / S. L. V. Bortolotti [et al.] // Quality & Quantity. 2013. Vol. 47. P. 2341-2360. https://doi.org/10.1007/s11135-012-9684-5
[15] Koliada M., Bugayova T., Miklashevich N. Intelligent Testing Systems Based on Adaptive Algorithms // CEUR Workshop Proceedings. 2020. Vol. 2770. P. 172-186. URL: https://ceur-ws.org/Vol-2770/paper21.pdf (дата обращения: 11.02.2023).
[16] Almansour F. M., Alroobaea R., Ghiduk A. S. An Empirical Comparison of the Efficiency and Effectiveness of Genetic Algorithms and Adaptive Random Techniques in Data-Flow Testing // IEEE Access. 2020. Vol. 8. P. 12884-12896. https://doi.org/10.1109/ACCESS.2020.2966433
[17] Стариченко Б. Е., Гизатуллин М. Г., Истомина Е. А. Оценка уровня подготовленности и качества измерительных тестовых материалов с помощью интерактивной формы Test_Results // Педагогическое образование в России. 2016. № 7. С. 104-112. https://doi.org/10.26170/po16-07-15
[18] Буравлев А. И., Переверзев В. Ю., Голяева Н. В. Оценка качества подготовки на основе байесовской процедуры адаптивного критериально-ориентированного тестирования // Среднее профессиональное образование. 2006. № 9. С. 18-21. EDN: KTOGIR
[19] Адаптивная модель тестирования нескольких компетенций на основе алгоритма Байеса / В. Н. Гусятников, Т. Н. Соколова, А. И. Безруков, И. В. Каюкова // Современные наукоемкие технологии. 2022. № 1. С. 40-46. https://doi.org/10.17513/snt.39007
[20] Applying Bayesian Network to Assess the Levels of Skills Mastering in Adaptive Dynamic OER-Systems / I. Nekhaev [et al.] // Software Engineering Perspectives in Intelligent Systems. CoMeSySo 2020. Advances in Intelligent Systems and Computing ; ed. by R. Silhavy, P. Silhavy, Z. Prokopova. Vol. 1294. Cham : Springer, 2020. P. 1090-1116. https://doi.org/10.1007/978-3-030-63322-6_94
[21] Безруков А. И., Грахольская Л. В. Имитационная модель для выбора стратегии адаптивного тестирования // Математическое и компьютерное моделирование в экономике, страховании и управлении рисками. 2020. № 5. С. 145-151. EDN: KYXXJM
[22] Synthesizing adaptive test strategies from temporal logic specifications / R. Bloem [et al.] // Formal Methods in System Design. 2019. Vol. 55. P. 103-135. https://doi.org/10.1007/s10703-019-00338-9
[23] Rodríguez I., Rubio D., Rubio F. Complexity of adaptive testing in scenarios defined extensionally // Frontiers of Computer Science. 2023. Vol. 17. Article number: 173206. https://doi.org/10.1007/s11704-022-1673-9
[24] Assessment of Test Items Quality and Adaptive Testing on the Rasch Model / A. Kostikov [et al.] // Information and Communication Technologies in Education, Research, and Industrial Applications. ICTERI 2021. Communications in Computer and Information Science ; ed. by V. Ermolayev [et al.]. Vol. 1698. Cham : Springer, 2022. P. 252-271. https://doi.org/10.1007/978-3-031-20834-8_12
[25] Тагирова Л. Ф., Зубкова Т. М. Интеллектуальная система адаптивного тестирования // Научно-технический вестник информационных технологий, механики и оптики. 2023. Т. 23, № 4. С. 757-766. https://doi.org/10.17586/2226-1494-2023-23-4-757-766

*Поступила 11.02.2023; одобрена после рецензирования 18.04.2023; принята к публикации 21.05.2023.*

Об авторах:

**Безруков Алексей Иосифович**, доцент кафедры информационно-коммуникационных систем и программной инженерии Института прикладных информационных технологий и коммуникаций, ФГБОУ ВО «Саратовский государственный технический университет имени Гагарина Ю.А. (410054, Российская Федерация, г. Саратов, ул. Политехническая, д. 77), кандидат экономических наук, доцент, **ORCID: https://orcid.org/0000-0002-8726-6404,** bezr_alex@mail.ru

**Акимова Светлана Александровна**, доцент кафедры информационно-коммуникационных систем и программной инженерии Института прикладных информационных технологий и коммуникаций, ФГБОУ ВО «Саратовский государственный технический университет имени Гагарина Ю.А. (410054, Российская Федерация, г. Саратов, ул. Политехническая, д. 77), кандидат физико-математических наук, доцент, **ORCID: https://orcid.org/0009-0009-7289-7105**, akimovasa@yandex.ru

*Все авторы прочитали и одобрили окончательный вариант рукописи.*

## References

[1] Demenchenok O. *Analiz modelej dlya adaptivnogo testirovaniya* [Analysis of models for adaptive testing]. *Pedagogicheskie izmereniya* = Pedagogical measurements. 2011;(1):3-18. (In Russ., abstract in Eng.) EDN: OJOMQT

[2] Belous N.V., Kutsevich I.V. *Model' adaptivnogo kontrolya znanij* [Model of adaptive control of knowledge]. *Radioelektronika, informatika, upravlenie* = Radio Electronics. Computer Science. Control. 2010;(1):39-44. (In Russ., abstract in Eng.) EDN: SAZQLJ

[3] Kornilov D.E. *Sravnitel'nyj analiz adaptivnoj i klassicheskoj modeli testirovaniya* [Comparative analysis of adaptive and classical testing models]. *Perspektivy razvitiya informacionnyh tekhnologij* = Prospects for the development of information technologies. 2014;(19):119-124. (In Russ., abstract in Eng.) EDN: SEFOEJ

[4] Oppl S., Reisinger F., Eckmaier A. et al. A flexible online platform for computerized adaptive testing. *International Journal of Educational Technology in Higher Education*. 2017;14:2. https://doi.org/10.1186/s41239-017-0039-0

[5] Nurakhmetov D. Reinforcement Learning Applied to Adaptive Classification Testing. In: Veldkamp B., Sluijter C. (eds.) Theoretical and Practical Advances in Computer-based Educational Measurement. Methodology of Educational Measurement and Assessment. Cham: Springer; 2019. p. 325-366. https://doi.org/10.1007/978-3-030-18480-3_17

[6] Gamage S.H.P.W., Ayres J.R., Behrend M.B. et al. Optimising Moodle quizzes for online assessments. *International Journal of STEM Education*. 2019;6:27. https://doi.org/10.1186/s40594-019-0181-4

[7] Romanenko D. Computer Testing as a Form of Students' Knowledge Control. *CEUR Workshop Proceedings*. 2020;2770:196-203. Available at: https://ceur-ws.org/Vol-2770/paper23.pdf (accessed 11.02.2023).

[8] Maslova O., Kuklina A. Adaptive testing model based on IRT techniques. *The Scientific Heritage*. 2020;(43):41-44. EDN: AGLIOQ

[9] Riza L.S., Nurjanah N.S., Wihardi Y. The development of computer adaptive test and item response theory with 4 parameters based the logistics model. *Journal of Physics: Conference Series*. 2019;1280:032033. https://doi.org/10.1088/1742-6596/1280/3/032033

[10] Wulandari F., Hadi S., Haryanto. Computer-based Adaptive Test Development Using Fuzzy Item Response Theory to Estimate Student Ability. *Computer Science and Information Technology*. 2020;8(3):66-73. https://doi.org/10.13189/csit.2020.080302

[11] Hu Z.F., Lin L., Wang Y.H., Li J.W. The Integration of Classical Testing Theory and Item Response Theory. *Psychology*. 2021;12:1397-1409. https://doi.org/10.4236/psych.2021.129088

[12] Kimura T. The impacts of computer adaptive testing from a variety of perspectives. *Journal of Educational Evaluation for Health Professions*. 2017;14:12. https://doi.org/10.3352/jeehp.2017.14.12

[13] Sijtsma K. Developments in Measurement of Persons and Items by Means of Item Response Models. *Behaviormetrika*. 2001;28:65-94. https://doi.org/10.2333/bhmk.28.65

[14] Bortolotti S.L.V., Tezza R., de Andrade D.F. et al. Relevance and advantages of using the item response theory. *Quality & Quantity*. 2013;47:2341-2360. https://doi.org/10.1007/s11135-012-9684-5

[15] Koliada M., Bugayova T., Miklashevich N. Intelligent Testing Systems Based on Adaptive Algorithms. *CEUR Workshop Proceedings*. 2020;2770:172-186. Available at: https://ceur-ws.org/Vol-2770/paper21.pdf (accessed 11.02.2023).

[16] Almansour F.M., Alroobaea R., Ghiduk A.S. An Empirical Comparison of the Efficiency and Effectiveness of Genetic Algorithms and Adaptive Random Techniques in Data-Flow Testing. *IEEE Access*. 2020;8:12884-12896. https://doi.org/10.1109/ACCESS.2020.2966433

[17] Starichenko B.E., Gizatullin M.G., Istomina E.A. Assessment of the level of readiness and quality of test materials using the online form test_results. *Pedagogical Education in Russia*. 2016;(7):104-112. (In Russ., abstract in Eng.) https://doi.org/10.26170/po16-07-15

[18] Buravlev A.I., Pereverzev V.Yu., Golyaeva N.V. *Ocenka kachestva podgotovki na osnove bajesovskoj procedury adaptivnogo kriterial'no-orientirovannogo testirovaniya* [Evaluation of the quality of training based on the Bayesian procedure of adaptive criterion-oriented testing]. *Srednee professional'noe obrazovanie* = Secondary vocational education. 2006;(9):18-21. (In Russ., abstract in Eng.) EDN: KTOGIR

[19] Gusyatnikov V.N., Sokolova T.N., Bezrukov A.I., Kayukova I.V., Adaptive model for testing several competencies based on the Bayes algorithm. *Modern high technologies*. 2022;(1):40-46. (In Russ., abstract in Eng.) https://doi.org/10.17513/snt.39007

[20] Nekhaev I., Zhuykov I., Manukyants S., Maslennikov A. Applying Bayesian Network to Assess the Levels of Skills Mastering in Adap-

tive Dynamic OER-Systems. In: Silhavy R., Silhavy P., Prokopova Z. (eds.) Software Engineering Perspectives in Intelligent Systems. CoMeSySo 2020. *Advances in Intelligent Systems and Computing.* Vol. 1294. Cham: Springer; 2020. p. 1090-1116. https://doi.org/10.1007/978-3-030-63322-6_94

[21] Bezrukov A.I., Grakholskaya L.V. *Imitacionnaya model' dlya vybora strategii adaptivnogo testirovaniya* [Simulation model for choosing an adaptive testing strategy]. *Matematicheskoe i komp'yuternoe modelirovanie v ekonomike, strahovanii i upravlenii riskami* = Mathematical and computer modeling in economics, insurance and risk management. 2020;(5):145-151. (In Russ., abstract in Eng.) EDN: KYXXJM

[22] Bloem R., Fey G., Greif F. et al. Synthesizing adaptive test strategies from temporal logic specifications. *Formal Methods in System Design*. 2019;55:103-135. https://doi.org/10.1007/s10703-019-00338-9

[23] Rodríguez I., Rubio D., Rubio F. Complexity of adaptive testing in scenarios defined extensionally. *Frontiers of Computer Science.* 2023;17:173206. https://doi.org/10.1007/s11704-022-1673-9

[24] Kostikov A., Vlasenko K., Lovianova I., Volkov S., Kovalova D., Zhuravlov M. Assessment of Test Items Quality and Adaptive Testing on the Rasch Model. In: Ermolayev V., et al. (eds.) Information and Communication Technologies in Education, Research, and Industrial Applications. ICTERI 2021. *Communications in Computer and Information Science*. Vol. 1698. Cham: Springer; 2022. p. 252-271. https://doi.org/10.1007/978-3-031-20834-8_12

[25] Tagirova L.F., Zubkova T.M. Intelligent adaptive testing system. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics.* 2023;23(4):757-766. (In Russ., abstract in Eng.) https://doi.org/10.17586/2226-1494-2023-23-4-757-766

*Submitted 11.02.2023; approved after reviewing 18.04.2023; accepted for publication 21.05.2023.*

About the authors:

**Alexey I. Bezrukov**, Associate Professor of the Department of Information and Communication Systems and Software Engineering, Institute of Applied Information Technologies & Communication, Saratov State Gagarin Technical University (77 Polytechnicheskaya St., 410054 Saratov, Russian Federation), Cand. Sci. (Econ.), Associate Professor, **ORCID: https://orcid.org/0000-0002-8726-6404**, bezr_alex@mail.ru

**Svetlana A. Akimova**, Associate Professor of the Department of Information and Communication Systems and Software Engineering, Institute of Applied Information Technologies & Communication, Saratov State Gagarin Technical University (77 Polytechnicheskaya St., 410054 Saratov, Russian Federation), Cand. Sci. (Phys.-Math.), Associate Professor, **ORCID: https://orcid.org/0009-0009-7289-7105**, akimovasa@yandex.ru

*All authors have read and approved the final manuscript.*