АВТОМАТИЗИРОВАННЫЕ ИНФОРМАЦИОННЫЕ СИСТЕМЫ В ОБРАЗОВАНИИ И НАУКЕ

УДК: 51-7
Doi: 10.17223/16095944/61/10

В.М. Карнаухов
Московский государственный университет природообустройства, г. Москва, Россия

ТОЧНОСТЬ ОЦЕНОК ЕГЭ В ЗАВИСИМОСТИ ОТ КОЛИЧЕСТВА И ТРУДНОСТИ ЗАДАНИЙ ГРУППЫ «С»

Представлены три метода получения оценок уровней подготовленности абитуриентов на ЕГЭ. Первый из них используется в настоящее время на ЕГЭ, второй был предложен автором в своих предыдущих работах, третий метод – метод первичных баллов – достаточно известен в теории тестирования и был обоснован автором в своих работах. В статье автор исследует зависимость точности вышеупомянутых методов от количества и трудности заданий теста ЕГЭ группы «С». На основании полученных графиков и таблиц автор дает конкретные рекомендации с целью повышения точности оценок, выставляемых на ЕГЭ.

Ключевые слова: модель Раша, метод Монте-Карло, функция шкалирования, метод шкалирования, метод первичных баллов, латентные параметры, уровень подготовленности, уровень трудности.

В статье используются следующие термины и понятия [1–2].

Первичный балл – число баллов, набранных участником тестирования при выполнении заданий теста.

Уровень подготовленности участника тестирования и уровень трудности задания теста – латентные параметры тестирования, недоступные для непосредственного измерения, отражающие соответственно уровень обученности испытуемого и уровень сложности задания.

Логит – единица измерения латентных параметров тестирования, диапазон изменения которой совпадает с числовой прямой.

Процентный логит – единица измерения латентных параметров, диапазон изменения которой совпадает с интервалом [0%, 100%].

Математическая модель Раша – модель тестирования, венцом которой является формула для вероятности решения участником тестирования с заданным уровнем подготовленности задания теста с заданным уровнем трудности.

Имитационное моделирование – компьютерное моделирование тестирования на основе метода Монте-Карло.

В статье описаны три основных метода получения оценок уровней подготовленности абитуриентов на Едином государственном экзамене. Среди них два прямых метода: метод шкалирования, используемый в настоящее время на ЕГЭ, и модифицированный метод шкалирования, предложенный автором в работе [4]. Эти методы позволяют переводить первичные баллы, набранные абитуриентами, на шкалу процентных логитов, характеризующих уровни подготовленности абитуриентов. Третий метод, исследуемый в этой работе, обоснован автором в работе [3] – метод первичных баллов. Этот метод является косвенным методом, позволяющим за два шага получать оценки уровней подготовленности абитуриентов. Первый шаг состоит в получении оценок латентных параметров уровней подготовленности, измеряемых в логитах. На втором шаге логиты переводятся в процентные логиты.

После описания методов получения оценок уровней подготовленности абитуриентов в статье следует изложение основных этапов имитационного моделирования процесса тестирования. Моделирование использует известную в теории математическую модель тестирования известного датского математика Г. Раша [1–2]. Обсуждаются некоторые элементы компьютерной программы, осуществляющей имитационное моделирование тестирования.

Результатом работы вышеупомянутой программы являются графики и таблицы, которые выявляют характер зависимости точности трех методов от числа и трудности заданий теста ЕГЭ группы «С». На основании полученного материала автор делает конкретные выводы и рекомен-

дации, направленные на повышение точности оценок ЕГЭ.

Метод № 0 – метод шкалирования. В методике шкалирования результатов ЕГЭ, используемой в 2011–2014 гг., реализуется поэтапное установление соответствия тестовых и первичных баллов для каждого общеобразовательного предмета, по которому проводится ЕГЭ.

I этап.

Сначала в диапазоне первичных баллов от нуля до максимального первичного балла ПБmax для каждого общеобразовательного предмета ЕГЭ выбираются два значения первичных баллов: ПБ1 и ПБ2, разделяющие группы участников с различным уровнем подготовки по данному предмету.

Величина ПБ1 выбирается как наименьший первичный балл, получение которого свидетельствует об усвоении участником экзамена основных понятий и методов по соответствующему общеобразовательному предмету. Он определяется на основе экспертизы демонстрационного варианта по данному общеобразовательному предмету специалистами общего образования, ссузов и вузов различного профиля из разных субъектов РФ. Экспертиза осуществляется с учетом уровня сложности каждого задания и значимости проверяемого им содержания, умения, навыка, способа деятельности в контексте общеобразовательного предмета. При этом требования к значению ПБ1 соответствуют требованиям, которые использовались при определении ПБ1 прошлого года (для обеспечения эквивалентности шкал двух лет).

Величина ПБ2 определяется профессиональным сообществом как наименьший первичный балл, получение которого свидетельствует о высоком уровне подготовки участника экзамена, а именно, о наличии системных знаний, овладении комплексными умениями, способности выполнять творческие задания по соответствующему общеобразовательному предмету.

Если спецификация экзаменационного варианта не изменилась по сравнению с прошлым годом, то ПБ1 и ПБ2 также остаются неизменными. Если же структура экзаменационной работы или сложность заданий контрольных измерительных материалов поменялись, то устанавливаются новые значения ПБ1 и ПБ2 с учетом имеющихся изменений.

II этап.

Первичным баллам ПБ1 и ПБ2 ставятся в соответствие тестовые баллы ТБ1 и ТБ2 по каждому общеобразовательному предмету.

Для всех предметов в качестве величин ТБ1 выбираются минимальные тестовые баллы ЕГЭ 2013 г., установленные распоряжениями Рособрнадзора. Данные значения совпадают с минимальными баллами ЕГЭ 2012 г.

Тестовые баллы ТБ2 по всем предметам, кроме географии и истории, устанавливаются равными аналогичным баллам 2012 г. По сравнению с 2012 г. на 1 балл уменьшился ПБ2 по географии и на 1 балл увеличился ПБ2 по истории. Это связано с изменением структуры экзаменационных работ по этим предметам. В табл. 1 представлены значения ПБ1 и ПБ2, ТБ1 и ТБ2 на 2013 г.

Таблица 1

Значения граничных первичных и тестовых баллов в 2013 г.

Предмет	ПБ1	ТБ1	ПБ2	ТБ2
Русский язык	17	36	54	73
Математика	5	24	15	63
Обществознание	15	39	48	72
История	13	32	47	72
Физика	12	39	33	62
Химия	14	36	58	80
Биология	17	36	60	79
География	14	37	43	69
Информатика	8	40	35	84
Иностранные языки	16	20	65	82
Литература	8	32	36	73

III этап.

По каждому общеобразовательному предмету определяется соответствие между первичным и тестовым баллами на основе следующей процедуры. Первичному баллу 0 ставится в соответствие тестовый балл 0, а максимальному первичному баллу ПБmax ставится в соответствие тестовый балл 100. Все промежуточные первичные баллы между 0, ПБ1, ПБ2 и ПБmax переводятся в тестовые, пропорционально распределенные между соответствующими значениями тестовых баллов: 0, ТБ1, ТБ2 и 100. На рис. 1 представлена получаемая зависимость.

Если промежуточные первичные баллы соответствуют дробным значениям тестовых, то производится округление тестового балла до ближайшего большего целого числа. Указанная процедура позволяет согласовывать тестовые

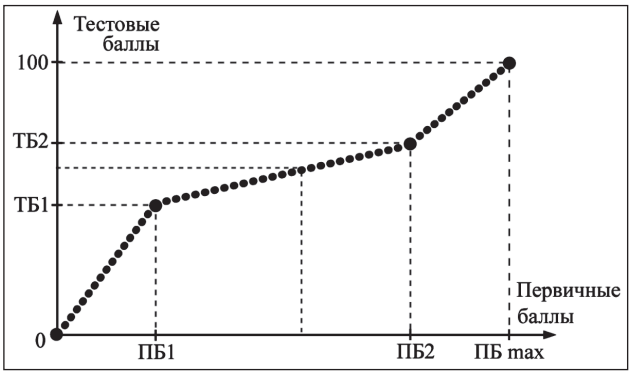

Рис. 1. Соответствие между тестовыми и первичными баллами

баллы одинаково подготовленных участников 2011–2013 гг. и обеспечивает сравнительную сопоставимость результатов экзамена по годам.

Метод № 1 – метод модифицированного шкалирования. Описанный выше метод шкалирования можно усовершенствовать [4]. Например, можно рассмотреть семейство функций перевода первичных баллов в тестовые, которые отличаются между собой только значениями в точках ПБ1 = 5 и ПБ2 = 15. Исследование этого семейства приводит нас к наиболее эффективной функции зависимости уровня подготовленности от первичного балла. Ломаная линия зависимости изображена на рис. 2.

Для полученной линии ТБ1 = 36, ТБ2 = 58, тогда как для прежней, используемой на ЕГЭ, ТБ1 = 24 , ТБ2 = 63. Используя на практике полученные значения, можно добиться выигрыша в точности примерно в 2,3 %.

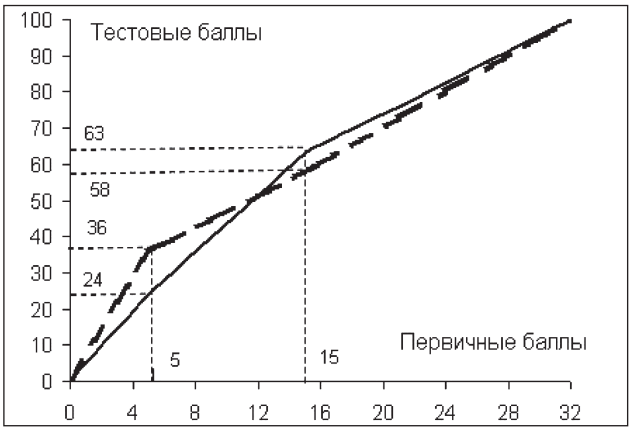

Рис. 2. Изменение соответствия между первичными и тестовыми баллами

Метод № 2 – метод первичных баллов. Третьим методом получения оценок уровней подготовленности абитуриентов является метод первичных баллов. Этот метод описан и обоснован в авторской работе [3]. Согласно этому методу оценки θ_i, i=0,...,K уровней подготовленности участников тестирования (где i – число набранных тестовых баллов на экзамене; K – максимально возможное число набранных баллов) вычисляются по формуле:

$$\theta_i = \bar{\theta}_i + \bar{\theta}_{cp}, \ , i = 1, ..., K-1.$$

В этой формуле используются следующие величины:

$$1)\bar{\theta}_i = ln\left(\frac{i}{K-i}K_1\right), \text{ причем } K_1 = \frac{\sum_{r=1}^{K-1}(K-r)\cdot N_r}{\sum_{r=1}^{K-1} r\cdot N_r},$$

где N_r – число участников тестирования, набравших r тестовых баллов.

$$2)\ \bar{\theta}_{cp} = \frac{\sum_{i=1}^{K-1}\bar{\theta}_i \cdot N_i}{N - N_0 - N_K}.$$

Для крайних значений i: $i = 0$ и $i = K$ используются следующие оценки:

$\theta_0 = -\theta_{max}$, $\theta_K = \theta_{max}$, где $\theta_{max} = 5$.

Точность вышеперечисленных трех методов выражается табл. 2, полученной в работе [4].

Таблица 2

Точность различных методов оценки латентных параметров тестирования

№	Метод	Средняя погрешность, %	Максимальная погрешность, %
1	Прямой метод шкалирования	6,9	27
2	Прямой метод модифицированного шкалирования	4,6	22
3	Косвенный метод первичных баллов	4,5	19

Как видно из табл. 2, косвенный метод первичных баллов является наиболее точным, что будет также подтверждено ниже.

Имитационное моделирование. Для исследования зависимости точности вышеперечисленных методов от количества и трудности заданий группы «С» была разработана авторская программа, моделирующая при помощи метода Монте-Карло процесс проведения ЕГЭ для абитуриентов в коли-

честве $N = 500$ и теста, состоящего из n_B заданий группы В и n_C заданий группы «С». При этом число n_B фиксировано и равно 15, что наблюдалось на последнем ЕГЭ, уровень сложности этих задач $\delta_B = -1$, а число n_C меняется в пределах от 1 до 8. Уровень сложности заданий группы «С» изменяется от 1 до 3. Благодаря вариативности количества заданий группы «С» и их сложности можно было провести запланированное исследование.

Процесс ЕГЭ моделировался достаточно большое количество раз (число итераций равно 20). Для каждого моделирования вычислялись две характеристики:

1) среднее отклонение σ_{cp} оценки уровня подготовленности абитуриента от истинного значения этого латентного параметра;

2) наибольшее отклонение σ_{max} оценки уровня подготовленности от истинного значения этого латентного параметра.

Далее вычисленные характеристики усреднялись по всем итерациям.

Для моделирования процесса тестирования использовался метод Монте-Карло. Опишем процесс компьютерной имитации процесса тестирования.

1) Вначале моделируются истинные уровни подготовленности участников θ_i, i=1,...,N и истинные уровни трудностей заданий δ_j, j=1,...,M. Уровни подготовленности участников смоделированы как реализации нормальной случайной величины $N(0,1)$ по формуле $\theta_i = F_N^{-1}(r_i)$, где $F_N(x)$ – функция распределения нормированной нормальной случайной величины, т.е. $N(0,1)$, которая определяется по формуле

$$F_N(x) = 0.5 + \frac{1}{\sqrt{2\pi}} \int_0^x e^{-x^2/2} dx,$$

$F_N^{-1}(r_i)$ – обозначение функции, обратной к функции $F_N(x)$. Значение обратной функции вычисляется в точке r_i, представляющей собой очередную реализацию датчика случайных чисел на отрезке (0,1).

В силу правила 3 сигм все реализации выше определенной случайной величины будут находиться в интервале $\theta_i \in (-3;3)$.

Уровни трудностей заданий смоделированы как реализации нормальных случайных величин

$$\left(\Delta = \frac{0{,}1}{3}\right): \delta_j \in N(\delta_j^{cp}; \Delta),\ j=1,...,M,$$

$$\delta_j^{cp} = -1,\ j=1,...,n_B,$$

$$\delta_j^{cp} = 1,2,3,\ j=n_B+1,...,n_B+n_C.$$

В силу правила 3 сигм и малости Δ задания с одним номером в различных вариантах будут мало отличаться друг от друга.

2) Для каждого абитуриента и для каждого задания вычисляются первичные баллы. Для этого по формуле

$$p_{ij} = \frac{1}{1+e^{-(\theta_i-\delta_j)}},\quad i=0.,...,N,\quad j=1,...,M$$

вычисляются вероятности p решения i-м абитуриентом j-го задания. Затем абитуриенту начисляется первичный балл B за решение задания по формуле

$$B = \begin{cases} 0, & r \ge p \\ \left[\frac{r \cdot m}{p}\right] + 1, & r < p, \end{cases}$$

где r – очередная реализация датчика случайных чисел на (0;1);

m – максимальное число баллов за решение задачи, причем $m = 1$ для $j = 1, ..., n_B$,

$m = 2$ для заданий группы «С», сложность которых равна $\delta_C = 1$,

$m = 3$ для заданий группы «С», сложность которых равна $\delta_C = 2$,

$m = 4$ для заданий группы «С», сложность которых равна $\delta_C = 3$,

квадратные скобки обозначают целую часть их содержимого.

В результате использования вышеописанной программы были получены результаты, сведенные в таблицу, аналогичную табл. 3.

Результаты исследования зависимости погрешности методов от количества заданий группы «С». Если найти для каждого значения n_C среднее арифметическое (математическое ожидание) погрешности для различных наборов заданий группы «С», то можно построить график зависимости погрешности от числа заданий группы «С» (рис. 3). Для заданий группы «В» в программе была установлена трудность $\delta_B = -1$. В результате описанной выше имитации процесса ЕГЭ также был получен график, представленный на рис. 3.

Таблица 3

Зависимость погрешности от количества и трудности сложных заданий группы «С»

n_C	Трудность заданий δ_C	Метод 2	Метод 0	Метод 1	n_C	Трудность заданий δ_C	Метод 2	Метод 0	Метод 1
1	1 2 3	4.6281 4.4543 4.4100	6.7966 6.9570 6.8008	4.9084 5.0878 5.0187	2	1 1 1 2 1 3 2 2 2 3 3 3	4.3697 4.3861 4.3578 4.3602 4.2813 4.4406	7.4526 7.5015 7.4624 7.1132 7.0604 6.8974	5.3735 5.3837 5.3836 5.1190 4.9999 4.8620
	...	...	...	...		...	...	...	...
7	1111111 1111112 1111113 ...	3.8854 3.8495 3.8606 ...	9.9057 9.8466 9.9334 ...	7.2770 7.2649 7.3101 ...	8	11111111 11111112 11111113 ...	3.8468 3.8089 3.8370 ...	10.1372 10.3168 10.2994 ...	7.5312 7.7378 7.7174 ...

Комментарий к рис. 3:

1) Погрешность для метода № 3 с ростом n_C уменьшается. Этот метод ведет себя естественным образом: с увеличением количества информации точность увеличивается.

2) Методы № 1 и 2 ведут себя неестественно в смысле, описанном выше. Погрешность растет с увеличением n_C. Напомним, что на текущий момент $n_C = 6$.

3) Для метода № 1, который используется на ЕГЭ, увеличение n_C на единицу приводит к увеличению погрешности на 0,5 %. Поэтому дальнейшее увеличение n_C не рекомендуется.

4) Погрешность методов № 1 при $n_C = 6$ практически в 2,25 раза больше погрешности № 3. Поэтому рекомендуется заменить метод № 1 на метод № 3 с целью повышения точности выставления оценок.

5) Погрешность метода № 1 при $n_C = 6$ практически в 1,4 раза больше погрешности № 2. Поэтому рекомендуется провести модификацию метода № 1 в соответствии с методом № 2 с целью повышения точности выставления оценок.

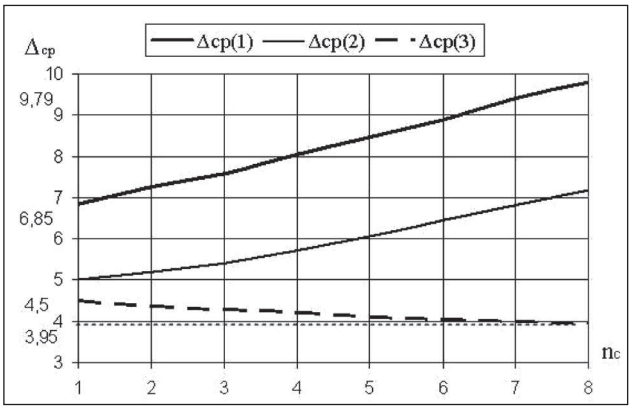

Рис. 3. Зависимость погрешности 3 методов от количества заданий группы «С» при $\delta_в = -1$

Результаты исследования зависимости погрешности методов от трудности заданий группы «С». Для исследования заданной зависимости было произведено три среза для $n_C = 4$, $n_C = 6$, $n_C = 8$. При этом сложность определяется специальном набором уровней сложности, равных 1, 2, 3, причем в каждом из них каждый последующий уровень не меньше предыдущего. Таким образом, каждый такой набор можно представить числом (в дальнейшем будем называть это число весом набора) в 4-й системе счисления, например, 1123 соответствует десятичному числу $1 \cdot 4^3 + 1 \cdot 4^2 + 2 \cdot 4 + 3 = 64 + 16 + 8 + 3 = 91$. Поэтому наборы можно расположить на одномерной шкале по возрастанию их весов, соответствующих этим наборам. В силу этой возможности удалось

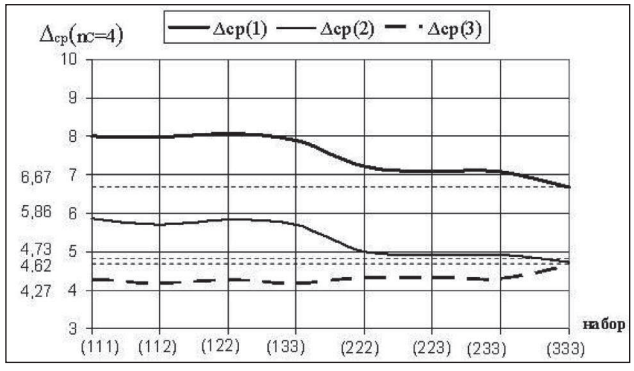

Рис. 4. Зависимость погрешности от трудности заданий группы «С» при $n_C = 4$

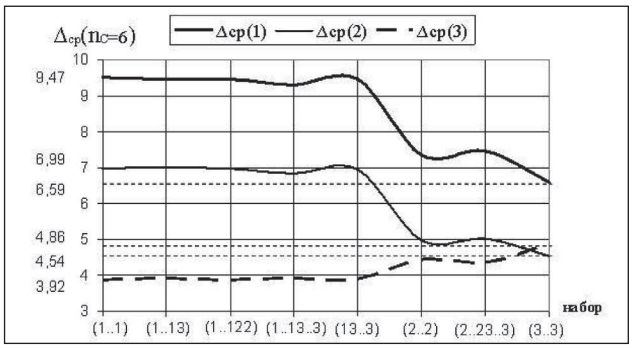

Рис. 5. Зависимость погрешности от трудности заданий группы «С» при $n_C = 6$

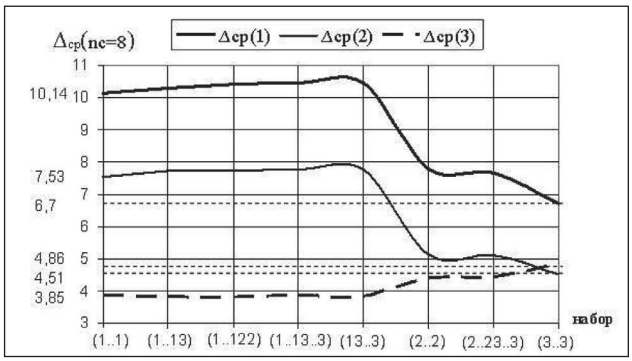

Рис. 6. Зависимость погрешности от трудности заданий группы «С» при $n_C = 8$

получить одномерные зависимости погрешности от наборов сложности заданий группы «С» при фиксированных значениях n_C (рис. 4–6).

Комментарий к рис. 4–6:

1) Для всех срезов и для всех методов погрешности постоянны, начиная с единичного набора (1...1) и заканчивая набором (13...3), причем погрешность метода № 3 меньше погрешности метола № 1 в 2,5 раза и меньше погрешности метода № 2 в 1,7 раза.

2) С набора (2...2) происходит резкое изменение погрешности в сторону уменьшения (на 26–35 %) для методов № 1 и 2, и в сторону увеличения (на 24 %) для метода № 3, причем к концу шкалы, заканчивающейся набором (3…3), модифицированный метод шкалирования дает погрешность, меньшую, чем метод первичных баллов.

3) В конце весовой шкалы наборов преимущество метода первичных баллов нивелируется по сравнению с методами шкалирования: метод № 3 «лучше» метода № 1 в 1,4 раза (погрешность метода № 1 в 1,4 раза больше погрешности метода № 3), а методы № 3 и 2 по точности примерно одинаковы.

Выводы-рекомендации:

1) Для выставления оценок на ЕГЭ рекомендуется использовать метод первичных баллов, точность которого при фиксированном наборе заданий группы «В» в 2,25–2,5 раза выше, чем у метода шкалирования, который используется в настоящее время на экзамене.

2) При использовании метода шкалирования рекомендуется использовать наименьшее возможное число заданий группы «С» в силу того, что с увеличением числа задач на единицу погрешность растет на 0,5 %. При использовании метода первичных баллов рекомендуется использовать 6 заданий группы «С», так как, начиная с $n_C = 6$, происходит стабилизация погрешности и дальнейшее увеличение n_C не приводит к существенному выигрышу в точности.

3) При использовании метода шкалирования наиболее выгодно использовать наборы заданий группы «С», начиная с набора (2…2) и заканчивая набором (3…3). При использовании метода первичных баллов выгоднее использовать наборы заданий группы «С», начиная с набора (1…1) и заканчивая набором (1…3).

ЛИТЕРАТУРА

1. *Rasch G.* Probabilistic Models for Some Intelligence and Attainment Tests. – Copengagen Denmark: Danish Institute for Educational Research, 1968.

2. *Нейман Ю.М., Хлебников В.А.* Введение в теорию моделирования и параметризации педагогических тестов. – М., 2000. – 169 с.

3. *Карнаухов В.М.* Модель Раша как игровая модель // Открытое и дистанционное образование. – 2014. – № 4 (56). – С. 69–76.

4. *Карнаухов В.М.* Точность оценок ЕГЭ для различных методик // Открытое и дистанционное образование. – 2015. – № 2. – С. 20–28.

Karnaukhov V.M.

Moscow State University of Environmental Engineering, Moscow, Russia

ACCURACY OF EXAM ESTIMATES IN DEPENDENCE ON THE NUMBER AND COMPLEXITY OF TASKS OF C GROUP

Keywords: Rasch's model, Monte-Carlo method, function scaling, method scaling, the method of primary points, latent parameters, the level of entrants' proficiency, the level of complexity.

The article presents three methods for estimates of levels of entrants' proficiency at the United State Examination. The first of them is currently used at the exam, the second was proposed by the author in his previous works, the third method is a method of primary points, which is known enough in testing theory; it was proved by the author in his works. In the article the author investigates the dependence of the accuracy the methods mentioned above on the number and complexity of test tasks of C group. On the basis of the obtained graphs and tables, the author gives specific recommendations to improve the accuracy of ratings at the exam.

The article describes the three main methods for obtaining estimates of levels of entrants' proficiency at the Unified State Exam. Among them there are two direct methods: a method of scaling used at present at the exam, and a modified scaling method proposed by the author in his previous works. These methods make it possible to translate the primary points obtained by entrants to the scale of percentage logits, characterizing the levels of entrants' proficiency. The third method, being investigated in this work, was proved by the author in his previous works, i.e. the method of initial points. This method is an indirect method that makes it possible to assess levels of entrants' proficiency by means of two steps. The first step consists in obtaining estimates of the latent parameters of proficiency levels, which are measured in logits. In the second step the logits are translated into percentage logits.

Then, the article describes the main stages of the simulation of testing process followed by description of methods of estimates of levels of entrants' proficiency. The simulation uses a well-known mathematical model of testing suggested by famous Danish mathematician G. Rasch. Some elements of a computer program performing simulation of testing are discussed.

The results of the program are diagrams and tables that show the character of dependence of accuracy of the three methods on the number and complexity of tasks of C group. On the basis of the obtained material the author makes specific findings and gives recommendations aimed at improving estimates accuracy at the exam.

1) For estimating at the exam it is recommended to use the method of initial points; its accuracy in a fixed set of tasks of B group is 2.25 - 2.5 times higher than the scaling method, which is currently used at the exam.

2) When using the scaling method, it is recommended to use the smallest possible number of the tasks of C group, because the error grows to 0.5 % when the number of tasks increases to a unit. When using the method of the primary points, it is recommended to use 6 tasks of C group, because stabilization of errors takes place since $n_C = 6$ and further increase of n_C does not lead to a significant accuracy.

3) When using the scaling method, it is more profitable to use the sets of tasks of C group starting from (2…2) and ending with (3…3). When using method of the primary points, it is more advantageous to use the sets of tasks of C group starting from (1…1) and ending with (1…3).

REFERENCES

1. *Rasch G.* Probabilistic Models for Some Intelligence and Attainment Tests. – Copengagen Denmark: Danish Institute for Educational Research, 1968.

2. *Nejman Ju.M., Hlebnikov V.A.* Vvedenie v teoriju modelirovanija i parametrizacii pedagogicheskih testov. – M., 2000. – 169 s.

3. *Karnauhov V.M.* Model' Rasha kak igrovaja model' // Otkrytoe i distancionnoe obrazovanie. – 2014. – № 4 (56). – S. 69–76.

4. *Karnauhov V.M.* Tochnost' ocenok EGJe dlja razlichnyh metodik // Otkrytoe i distancionnoe obrazovanie. – 2015. – № 2. – S. 20–28.