ЕГЭ по математике. Расчет ошибок ранжирования

М. М. Луценко
Петербургский государственный университет путей сообщения Императора Александра I
ml4116@mail.ru

***Аннотация*. Приведены результаты моделирования ЕГЭ по математике по данным за 2011, 2012 годы. Найдено распределение входного потока тестируемых по уровням знаний. Указанные значения латентных параметров и предлагаемая методика их расчета использована для расчета ошибок ранжирования, произведенного по результатам тестирования. Оказалось, что только у 67% тестируемых положение в группе изменилось менее чем на 10%. Примерно 17% тестируемых существенно улучшили свое положение в группе, а другие 16% существенно ухудшили положение.**

Ключевые слова: тестирование; модель Раша; ЕГЭ; Item Response Theory

I. Введение

Контроль знаний был и остается важнейшей частью учебного процесса. Без объективного контроля сложно оценивать как качество обучения, так и проводить сравнение различных методик. В последние годы процедура контроля знаний упростилась и автоматизировалась. Она стала более объективной и, в конечном итоге, свелась к оценке уровня знаний учащегося. Современные теории оценивания предполагают, что каждый учащийся имеет некоторый уровень знаний (число θ), а экзаменатор пытается оценить его. Для определения уровня знаний экзаменатор проводит тестирование, по результатам которого строится его оценка или формируется решение *d*. Таким образом, тест оказывается инструментом, позволяющим оценить уровень знаний и принять решение о достаточности этих знаний при выполнении тех или иных работ.

Тестирование, как и всякий измерительный инструмент, имеет свои точность, надежность, разрешающую способность, которые необходимо учитывать при принятии тех или иных решений. В настоящее время по результатам тестирования абитуриенты ранжируются, и после этого принимаются управленческие решения: зачисление в то или иное высшее учебное заведение и т. д. В работах [4, 5, 6] показано, что точность и надежность оценки уровня знаний тестируемого могут быть довольно низкие, а, следовательно, принимаемые решения будут содержать ошибки. В настоящей работе мы вместо изучения ошибок оценивания сравним априорное и апостериорное (по результатам тестирования) ранжирование. Авторам не известны аналитические формулы, позволяющие измерить степень искажения априорного ранжирования, вызванного процедурой тестирования, а поэтому было принято решение о создании имитационной модели [7, 8].

В настоящей работе мы используем теорию параметризации педагогических тестов (Item response theory, IRT) [1, 2, 3, 4]. В рамках этой теории заданиям тестам и тестируемым приписываются числа: трудности заданий и уровни знаний тестируемых. Эти латентные характеристики позволяют анализировать, как тест, так и входные потоки тестируемых независимо выполненных тестов. В работе мы рассчитаем уровни трудностей заданий ЕГЭ по математике за 2011-2013 годы и найдем распределение входных потоков тестируемых по уровням их знаний.

Данные о результатах ЕГЭ в Санкт-Петербурге взяты из аналитического отчета предметной комиссии, подготовленного В.Б. Некрасовым, Г.И. Вольфсоном, и опубликованного на сайте [9]. В качестве базовой математической модели использовалась модель Раша. Оценки уровней трудностей заданий и уровней знаний находились по методу моментов [3]. Некоторые результаты частично опубликованы авторами [7, 8]. Теоретико-игровой подход к оценке уровня знаний тестируемых был рассмотрен автором в работах [4, 5, 6].

II. Модель Раша

Рассмотрим базовую модель массового тестирования. Обозначим через *n* количество участников тестирования, а их уровни знаний через – θ_i; $i = 1, 2, \ldots, n$. Мы считаем, что каждому участнику тестирования предлагается один и тот же вариант теста, состоящий из *N* заданий различной трудности τ_j; $i = 1, 2, \ldots, N$. Предполагается, что выполнение каждого задания оценивается по дихотомному принципу. Иными словами, за правильно выполненное задание тестируемый получает один балл, в противном случае он получает ноль баллов. Множество всех таких нулей и единиц образуют прямоугольную таблицу размера $n \times N$, называемую матрицей первичных баллов.

Современный подход к расчету уровня знаний тестируемого основан на теории параметризации педагогических тестов. При этом под тестом мы понимаем набор заданий, каждое из которых может быть выполнено некоторыми учащимися за отведенное время.

Перечислим основные положения IRT.

- Каждый тестируемый имеет некоторый уровень подготовки (знаний) θ из множества возможных (допустимых) уровней подготовки $\Theta \subseteq R$, а каж-

дое задание теста имеет свою трудность $\tau \in R$, выразимую вещественным числом.

- Заданию трудности τ приписана неубывающая на множестве Θ функция выполнимости этого задания $p_\tau(\theta)$, значение которой – вероятность выполнения этого задания тестируемым с уровнем подготовки θ. Функция $p_\tau(\theta)$ называется характеристической функцией задания или Item characteristic curve (ICC).
- Оценка уровня знаний тестируемого происходит по результату выполнения им теста, содержащего N заданий с характеристическими функциями $p_{\tau_1}(\theta), p_{\tau_2}(\theta),\ldots, p_{\tau_N}(\theta)$, каждое из которых выполняется независимо от других заданий.
- Как сложность (трудность) задания τ, так и уровень знаний тестируемого θ можно измерять в одинаковых единицах, а характеристическая функция $p_\tau(\theta)$ зависит лишь от разности $\theta-\tau$ - величины превышения уровня знаний тестируемого над трудностью выполняемого задания. Таким образом, существует функция выполнимости заданий теста $p(t)$, такая что $p_\tau(\theta) = p(\theta-\tau)$.

Предположение о выразимости степени подготовленности (уровня знаний) тестируемого вещественным числом - существенная часть теории параметризации педагогических тестов. Оно значительно упрощает наше представление о величине знаний тестируемого. Предположение о монотонности функций $p_\tau(\theta)$ говорит, о согласованности уровня знаний θ с порядком на множестве вещественных чисел. Существование одной характеристической функции для всех заданий теста существенно упрощает все вычислительные процедуры. Кроме того, предположение 4 позволяет считать, что задания можно сравнивать по трудности их выполнения и это сравнение можно проводить независимо от процедуры тестирования.

В модели Раша латентные параметры θ и τ измеряются в одинаковых единицах, логитах. Вероятность выполнения задания тестируемым зависит лишь от величины превышения $t=\theta-\tau$ уровня знаний θ тестируемого над трудностью τ задания. Характеристическая функция заданий находится по формуле:

$$p(t) = [1+\exp(-t)]^{-1}. \quad (1)$$

Используя характеристическую функцию выполнимости заданий (1), мы обозначим через $p_{i,j}$ вероятность того, что j-е задание теста выполнено i-м тестируемым. Если уровень знания i-го тестируемого равен θ_i, а уровень трудности задания есть τ_j, то

$$p_{i,j} = (1+\exp(\tau_j-\theta_i))^{-1}. \quad (2)$$

Обозначим через c_j – наблюдаемое число тестируемых, решивших задание с номером j из N заданий теста (число первичных баллов j-го задания); через b_i – количество тестируемых верно выполнивших ровно i заданий теста. Оценки $\hat\theta_0, \hat\theta_1,\ldots,\hat\theta_n$ и $\hat\tau_1, \hat\tau_2,\ldots,\hat\tau_N$ соответствующих параметров модели могут быть получены по методу моментов из решения следующей системы уравнений:

$$\begin{cases} \sum_{j=1}^{N} p_{i,j} = b_i; \quad i=\overline{1,n}; \\ \sum_{i=1}^{n} p_{i,j} = c_j; \quad j=\overline{1,N}; \end{cases} \quad (3)$$

Так как тест содержит только заданий N, то мы можем считать, что вся группа тестируемых разбита на N+1 подгруппу, в каждой из которых один и тот же уровень знаний. Обозначим через n_k число тестируемых набравших ровно k баллов, тогда система уравнений (3) примет вид:

$$\begin{cases} \sum_{j=1}^{N} p_{k,j} = k; \quad k=\overline{0,N}; \\ \sum_{k=0}^{N} n_k p_{k,j} = c_j; \quad j=\overline{1,N}. \end{cases} \quad (4)$$

Заметим, что $\sum_{k=0}^{N} n_k = n$.

Нелинейная система (4) содержит $2N+1$ уравнение и $2N+1$ переменную. Ее решения $\hat\theta_0, \hat\theta_1,\ldots,\hat\theta_n$ будут оценками уровней знаний для каждой из $N+1$ подгрупп тестируемых, а величины $\hat\tau_1, \hat\tau_2,\ldots,\hat\tau_N$ оценками уровней трудностей заданий теста.

III. Расчет латентных параметров модели

Рассчитаем латентные параметры θ и τ для заданий теста и участников ЕГЭ по математике в Санкт-Петербурге в 2011 году [7, 8]. В экзамене участвовало (обрабатывалось тестов) n = 20908 чел. (шт.) и каждый участник мог получить от нуля до N = 30 баллов. Все задания разделялись на две группы B и C. При выполнении заданий группы B тестируемому начислялся один балл. При выполнении заданий группы C тестируемому начислялся от 2 до 4 баллов (в зависимости от номера задания). Однако при частичном выполнении задания из группы C им мог быть начислен неполный балл, но кратный единице. Например, за выполнение задания C6 максимальный балл составлял 4 балла, однако тестируемый мог получить любое число баллов от 0 до 4 в зависимости от полноты выполнения задания C6. Таким образом, с заданием C6 мы можем связать 5 зависимых между собой подзаданий, каждое из которых имеет свою трудность, и за выполнение каждого из них начисляется ровно один балл.

Если теперь каждое подзадание части C считать отдельным заданием, то мы можем говорить о 30 заданиях, трудности которых мы обозначим через $\tau_j \quad j=\overline{1,30}$. По результатам тестирования всех тестируемых можно раз-

бить на 31 подгруппу. Обозначим через θ_k (средний) уровень знаний тестируемых в k-й подгруппе $k=\overline{0,30}$, а через n_k – число тестируемых выполнивших k-ое заданий теста (получивших k первичных балла). Из системы (6) мы получим следующую систему уравнений:

$$\begin{cases} \sum_{j=1}^{30} p_{k,j} = k; \quad k=\overline{0,30}; \\ \sum_{k=0}^{30} n_k p_{k,j} = c_j; \quad j=\overline{1,30}; \end{cases} \tag{5}$$

Решение системы нелинейных уравнений (5) можно свести к следующей экстремальной задаче:

$$f(\theta,\tau) = \sum_{k=0}^{30}\left(\sum_{j=1}^{30} p_{k,j} - k\right)^2 + \\ + \sum_{j=1}^{30}\left(\sum_{k=0}^{30} n_k p_{k,j} - c_j\right)^2 \to \min, \tag{6}$$

где $\tau=(\tau_1,\tau_2,\ldots,\tau_{30})$, $\theta=(\theta_0,\theta_1,\ldots,\theta_{30})$.

Для нахождения минимума целевой функции мы воспользовались надстройкой ПОИСК РЕШЕНИЙ Excel MS Office. При решении задачи (6) приходилось учитывать, что значения характеристической функции (2) при значениях близких к (–6) практически равны нулю, а при значениях близких к 6 практически равны 1. Кроме того из свойств нелинейной системы следует, что она имеет бесконечное число решений, получаемых друг из друга сдвигом.

Из результатов расчета видно, что уровни знаний тестируемых распределены примерно равномерно, кроме как у наиболее (15 логит) и наименее (–14 логит) подготовленных тестируемых.

Тест ЕГЭ 2011г. был составлен так, что сильные учащиеся могли выполнить все задания теста. Хотя модель Раша плохо оценивает уровни знаний за предельными значениями трудностей заданий, однако, можно с уверенностью сказать, что уровни знаний сильных учащихся существенно превосходят знания основной группы учащихся. А именно, учащие набравшие 100 баллов имеют уровни знаний свыше 15 логит, и уровень превышения над наиболее сложными заданиями составляет примерно 5,6 логит, то есть вероятность, с которой самые сильные учащиеся выполняют самое сложное задание, составляет 0,9963. Для учащихся набравших 98 баллов (29 первичных баллов) аналогичная вероятность равна 0,6019. Оцениваемый уровень знаний учащихся, не выполнивших ни одного задания, не превышает –14 логит, или вероятность, с которой они могли бы выполнить самое простое задание, менее 0,0004. Уровни знаний тестируемых, набравших менее 24 баллов (менее 4 первичных баллов) оказались менее чем –5,30 логит, а учащиеся набравшие 24 и более баллов имели уровень знаний превышающий –3,94 логита. Иными словами учащийся, выполнивший 4 и более заданий, выполнял самое простое задание B2 с вероятностью более чем 0,8964, и следующее по сложности задание B5 с вероятностью 0,6834.

IV. Моделирование экзамена

Попробуем оценить надежность принимаемых решений об уровне знаний тестируемых по результатам ЕГЭ. Для этого смоделируем процесс тестирования, используя найденные нами значения параметров.

Предположим, что тестируемый с уровнем знаний θ выполняет тест, содержащий N заданий, с вектором трудностей заданий $\tau=(\hat{\tau}_1,\hat{\tau}_2,\ldots,\hat{\tau}_N)$. Мы считаем, что каждое задание частей B и C он выполняет независимо друг от друга.

Для моделирования процесса тестирования мы запускаем датчик случайных чисел, который порождает равномерно распределенную на отрезке [0, 1] случайную величину ξ. Для каждого задания группы B мы порождаем свою случайную величину и считаем, что задание выполнено, если значение этой случайной величины меньше $p(\theta-\hat{\tau}_j)$, где через $\hat{\tau}_j$ обозначена трудность j-го задания. Так как подзадания заданий группы C зависимы, то число баллов, полученных при выполнении соответствующего задания, определяется по результату реализации одной случайной величины.

Например, заданию C6 будет соответствовать один из пяти наборов $(0,0,0,0)$, $(1,0,0,0)$, $(1,1,0,0)$, $(1,1,1,0)$ $(1,1,1,1)$ в зависимости от того между какими вероятностями окажется значение случайной величин.

Итак, по результату выполнения теста тестируемым с уровнем знаний θ мы получаем вектор (протокол) $z(\theta)=(z_1,z_2,\ldots,z_N)$ выполнения заданий теста, в котором на k-м месте стоит единица, если это задание (подзадание) выполнено и ноль, если оно не выполнено

Составим группу тестируемых из 1000 чел. Мы считаем, что их уровни подготовленности тестируемых θ распределены по непрерывному закону, построенному нами ранее, как кусочно-линейная аппроксимация выборочного закона распределения. В силу непрерывности априорного распределения все тестируемые имеют различные уровни знаний и, следовательно, могут быть ранжированы. Путь $(\theta_1,\theta_2,\ldots,\theta_n)$ – упорядоченная последовательность уровней подготовленности тестирования. По результатам тестирования они снова могут быть упорядочены $(r_1,r_2,\ldots,r_n)$, но уже по полученным баллам за тест. В этом векторе число r_i – положение i-го тестируемого в группе.

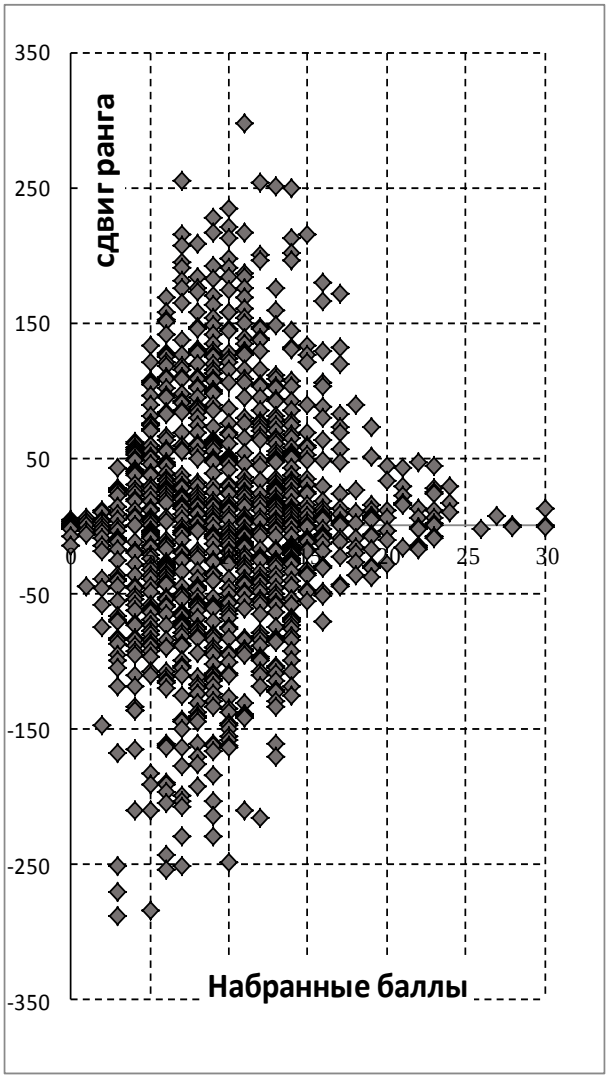

Рис. 1. Отклонение априорного ранга от апостериорного для группы из 1000 чел.

Пусть – апостериорные положения тестируемых, тогда $\Delta_i = i - r_i$ – величина сдвига ранга (положения) в группе. На рис показаны величины сдвигов для 1000 человек в зависимости от числа набранных баллов.

В результате сравнения априорного и апостериорного рангов (положений тестируемого в группе из 1000 чел) получилось, что:

- у 332 человек (у 66,4% участников) положение в группе отклонилось от первоначального менее чем на 50 позиций;
- у 128 человек (у 25,6% участников) положение в группе изменилось более чем на 50 позиций, но менее чем на 100;
- у 40 человек (у 8% участников) положение в группе изменилось более чем на 100 позиций.

Полученные результаты позволяют по-новому определить качество тестирования и сделать следующие выводы об уровне надежности полученных оценок.

Вывод. В результате тестирования выпускников школ по математике в 2011 г. в СПб у 1766 человек ранг изменился более чем на 4400 позиций, что, возможно, привело к ошибкам при зачислении в ВУЗ. А для России, где участвовало 738746 человек, у 59100 ранг изменился более чем на 14775 позиций.

Аналогичные расчеты были сделаны и для ЕГЭ по математике за 2012, 2013 гг. Найденные величины латентных параметров использовались для моделирования процедуры тестирования. Были найдены распределения ошибок ранжирования, которые во многом совпали результатами, найденными нами для СПб.

Список литературы

[1] Крокер Л., Алгина Дж. Введение в классическую и современную теорию тестов: Учебник. М.: Логос, 2010.

[2] Анастази А., Урбина С. Психологическое тестирование. 7-е изд. СПб.: Питер, 2007. 688 с.

[3] Нейман Ю.М., Хлебников В.А. Введение в теорию моделирования и параметризации педагогических тестов, 2000, М., 168 с.

[4] Луценко М.М. Теория статистических решений. Ч. 1: учеб.пособие / М.М. Луценко. СПб.: Петербургский гос. ун-т путей сообщения, 2011.

[5] Мягкие оценки уровня знаний тестируемого / М.М. Луценко, Тез. докл. XX Международная конференция по мягким вычислениям и измерениям. СПб, 2017 / СПбГЭТУ «ЛЭТИ». Т. 2. С. 167-170.

[6] Луценко М.М., Шадринцева Н.В. О точности педагогического тестирования, Известия Петербургского университета путей сообщения, СПб.: Петербургский гос. Ун-т путей сообщения, 2011. Вып 4(29), с. 250-258.

[7] Гарец С.Б., Елисеева Д.В., Соснина А.С. Имитационная модель изменения ранга тестируемого, неделя науки СПбГПУ : материалы научно-практической конференции с международным участием. Институт информационных технологий и управления СПбГПУ. СПб. : Изд-во Политехн. ун-та, 2014, с.83-85.

[8] Луценко М.М., Кенесбай М.Н. Ошибки при ранжировании по результатам тестирования, VIII Московская международная конференция по исследованию операций (ORM2017): Москва, Труды. Том 2/ Отв. ред. Е.З. Мохонько. М.: Изд-во ФИЦ ИУ РАН, 2016, с. 46-47

[9] www.ege.spb.ru/ege/statistika-i-analitika/ege-2013 (Официальный информационный портал ЕГЭ, Санкт-Петербург).