

Прогнозирование индивидуальных результатов испытуемых по разным вариантам гомогенных тестов на основе современной теории тестирования IRT

**Баранова
Татьяна Альбертовна**

начальник отдела ГАОУ ДПО «Московский центр качества образования»

**Белобородов
Владимир
Николаевич**

кандидат физико-математических наук,
ГАОУ ДПО «Московский центр качества образования»,
доцент НИЯУ МИФИ

**Татур
Александр Олегович**

кандидат физико-математических наук,
научный консультант ФГБНУ «ФИПИ», эксперт ГАОУ ДПО
«Московский центр качества образования»,
mcko@edu.mos.ru

Ключевые слова: вероятность выполнения задания, способность испытуемого, трудность задания, прогнозируемый балл, прогнозирование индивидуальных результатов.

Как оценить возможности предсказания результатов обучающегося в тестировании, если относительно недавно он уже выполнял тест по тому же учебному предмету? Очевидно, что на результат второго тестирования в сравнении с первым могут оказать существенное влияние: степень новизны проверяемого материала, личные обстоятельства учащегося, качество работы учителя, уровень сложности используемых заданий и другие факторы. Для того чтобы учесть различия первого и второго теста в части сложности заданий, целесообразно обратиться к современной теории тестирования IRT (Item Response Theory).

Вероятность правильного выполнения дихотомического (однобалльного) задания в рамках теории IRT в модели Раша¹ определяется выражением

$$P = \frac{e^{(\theta-\beta)}}{e^{(\theta-\beta)} + 1} = \frac{1}{e^{(\beta-\theta)} + 1}, \quad (1)$$

Здесь θ — способность испытуемого, β — трудность задания. Эта вероятность при любых действительных значениях способностей и трудностей лежит в интервале от 0 до 1, что обеспечивает уровень выполнения варианта из таких заданий от 0% до 100%. При использовании политомических (многобалльных) заданий выражение (1) может в рамках теории IRT определять условную вероятность совершения очередного шага, то есть увеличения набранного балла на единицу².

¹ Rasch G. Probabilistic Models for Some Intelligence and Attainment Tests. — Chicago.: Mesa Press, 1980, 199 p.

² Линда Крокер, Джеймс Алгина. Введение в классическую и современную теорию тестов. — М.: Логос, 2010. — 668 с.

Для определения средней трудности заданий теста можно предположить, что среднее значение способности равно нулю $\langle \theta \rangle = 0$, а также то, что как распределение способностей, так и их индивидуальные значения не изменяются по отношению к проверочным материалам. Эти предположения основаны на хорошо известных фактах: отличники и двоечники — группы достаточно стабильные во времени. Распределение по оценкам на интервалах порядка года тоже, как правило, не изменяются. Серьёзные изменения в средних результатах ожидаемы при изменении контингента, но в данном исследовании анализируются результаты одних и тех же учащихся по первому и второму тестированиям. Средняя вероятность (процент выполнения для выборки учащихся и по всем заданиям теста) с учётом указанных предположений может быть оценена так:

$$\langle P \rangle = \frac{1}{e^{\langle \beta \rangle} + 1}, \quad (2)$$

Из выражения (2) получаем

$$\langle \beta \rangle = \ln \frac{1 - \langle P \rangle}{\langle P \rangle}$$

и оценку индивидуальной вероятности выполнения первого варианта:

$$P_1 = \frac{1}{\frac{1 - \langle P_1 \rangle}{\langle P_1 \rangle} e^{-\theta} + 1}, \quad (3)$$

откуда определяется способность :

$$e^{-\theta} = \frac{1 - P_1}{P_1} \cdot \frac{\langle P_1 \rangle}{1 - \langle P_1 \rangle} \quad (4)$$

Аналогично для второго варианта можно написать

$$P_2 = \frac{1}{\frac{1 - \langle P_2 \rangle}{\langle P_2 \rangle} e^{-\theta} + 1} \quad (5),$$

или

$$P_2 = \frac{1}{\left(\frac{1 - \langle P_2 \rangle}{\langle P_2 \rangle} \right) \cdot \left(\frac{\langle P_1 \rangle}{1 - \langle P_1 \rangle} \right) \cdot \frac{1 - P_1}{P_1} + 1} =$$

$$= \frac{P_1}{P_1 + (1 - P_1) \cdot \langle C_1 \rangle / \langle C_2 \rangle}, \quad (6)$$

$$\text{где } \langle C_i \rangle = \frac{\langle P_i \rangle}{1 - \langle P_i \rangle} \text{ — усреднённые}$$

шансы по первому и второму тестированиям при $i = 1, 2$ соответственно. Вероятность P_i вычисляется как отношение набранного испытуемым в первом тестировании балла к максимальному баллу при первом тестировании. Прогнозируемый балл вычисляется как произведение вероятности P_2 , рассчитанной по формуле (6) и максимального балла при втором тестировании. Следует отметить, что соотношение максимальных баллов при первом и втором тестированиях может быть произвольным.

Таким образом, для прогноза индивидуального результата по второму тестированию нужно знать:

- индивидуальный процент выполнения первого теста;
- средний процент выполнения первого теста;
- оценку среднего процента выполнения второго теста или её эмпирическое значение, если второй тест был уже использован в аналогичной выборке.

Апробация рассматриваемого подхода была проведена на основе известных результатов двух последовательных тестирований *одних и тех же* школьников по некоторым учебным предметам.

Математика

Рассмотрим результаты исследований по математике. В качестве результатов первого тестирования были использованы данные диагностики обучающихся восьмых классов по математике (Тест 1), а второе тестирование этих же учащихся проводилось в девятом классе через год (Тест 2). В качестве примера расчёта индивидуальных значений прогнозируемых баллов для Теста 2 на основе данных Теста 1 рассмотрим результаты обучающихся одной из школ, представленные в таблице 1.

В последнем столбце таблицы представлены разности баллов расчётных значений прогноза и реальных значений баллов, полученных при втором тестировании. Максимальные отклонения

Таблица 1

Участник №	Фактические баллы		Прогноз для Теста 2, балл	Отклонение баллов прогноза и Теста 2
	Тест 1, балл (макс. б. 12)	Тест 2, балл (макс. б. 19)		
1	12	18	19	1
2	7	13	11	-2
3	7	9	11	2
4	2	6	3	-3
5	7	13	11	-2
6	10	18	16	-2
7	10	16	16	0
8	6	11	10	-1
9	9	13	14	1
10	5	9	8	-1
11	9	16	14	-2
12	12	15	19	4
13	6	10	10	0
14	5	14	8	-6
15	11	15	18	3
16	11	15	18	3
17	5	9	8	-1
18	10	7	16	9
19	9	12	14	2
20	3	6	5	-1
21	11	16	18	2
22	6	9	10	1
23	9	15	14	-1
24	7	15	11	-4
25	8	11	13	2
26	12	16	19	3
27	9	13	14	1
28	8	12	13	1
29	7	15	11	-4
30	7	11	11	0

(разности прогнозируемого и реального баллов) составляют, соответственно, 9 и -6 баллов. Среднее значение модулей отклонений прогноза от реального балла для данной школы не более 2,5 баллов. То есть с точностью около 2,5 баллов (без округления прогнозируемый балл может быть дробным) прогноз с большой вероятностью совпадёт с реальным значением для второго тестирования. Отклонения не более чем на 3 балла наблюдаются у 26 участников из 30.

Был проведён аналогичный расчёт и для всех обучающихся (4 287 человек), участвовавших в обоих тестированиях. На рисунке 1 представлено распределение участников тестирований в зависимости от отклонения полученного балла от его расчётного (прогнозируемого) значения. Распределение имеет вид, близкий к нормальному распределению (Рис. 1).

Среднее значение отклонения прогноза от реального балла равно -0,27, а среднеквадратичная ширина гистограммы откло-



Рис. 1. Гистограмма отклонений прогноза от реального балла. Математика

нений равна 3,2 балла. Средняя ширина распределения по баллам второго тестирования составляет 3,5, а ошибка определения второго балла равна 1,6.

Имеет смысл сравнить средние проценты выполнения тестов. Для первого теста $\langle P_1 \rangle = 52\%$, для второго $\langle P_2 \rangle = 60\%$ для рассматриваемой выборки. Средний прогноз процента выполнения для второго теста по предложенной методике составил $\langle P_{2II} \rangle = 58\%$. То есть средний прогноз близок к реальному среднему результату, но находится в интервале между реальными средними процентами по первому и второму тестированиям.

Важно отметить, что большинство заданий по математике для 9-го класса существенно отличалось от заданий для 8-го класса, за исключением нескольких. В частности, очень близки были задания на определение значений арифметических выражений (задания В1). Так для первого варианта теста 8-го класса нужно было найти значение выражения: $-8,5 - (-2,7)$, а для первого вари-

анта теста 9-го класса нужно было определить значение выражения: $6,3 + 3(-1,1)$. Проценты выполнения этих заданий для обучающихся, получивших указанные в таблице 2 отметки, приведены ниже.

Из таблицы видно, что проценты выполнения заданий для каждой категории обучающихся в 8-х и 9-х классах практически совпадают. На рисунке 2 представлено распределение участников в зависимости от разности баллов, полученных за второе и первое тестирования при выполнении задания В1. Всего было 4 287 участников.

Из диаграммы видно, что обучающихся, ухудшивших свой результат, 8%, а улучшивших результат — 17%. То есть улучшивших свой результат на 9% больше, чем ухудшивших. Большинство (75%) не изменили результат при выполнении заданий на вычисление значения арифметического выражения. Вследствие высокого процента выполнения этого задания в 8-х и 9-х классах (79% и 88%) этот результат является естественным.

Таблица 2

Класс	Процент выполнения, %	Процент выполнения для получивших за работу отметку, %			
		«2»	«3»	«4»	«5»
8	79	45	87	95	100
9	88	48	86	97	100



Рис. 2. Распределение участников по приращению балла за выполнение однотипного задания по математике при двух тестированиях

Русский язык

Аналогичный анализ проведён для тестирований по русскому языку. Первое тестирование обучающихся проводилось в 7-м классе, а второе — для этих же обучающихся, но уже через год в 8-м классе. Выборка школьников, участвовавших в двух тестированиях, составила 1 748 человек.

Средние проценты выполнения тестов по фактическим результатам и по прогнозу практически совпали (54%). Результаты на гистограмме отклонений имеют вид, близкий к нормальному распределению (Рис. 3).

Среднее значение отклонения прогноза от реального балла равно $-0,03$, а среднеквадратичная ширина гистограммы отклонений равна 3,6 балла. Стандартное отклонение и ошибка второго тестирования составляли 4,0 и 2,0 балла соответственно.

Физика

Также выполнены сравнения результатов тестирований по физике в восьмых и девярых классах, проводившихся с интервалом год и один месяц. Общая выборка составила 348 человек.

Средний процент выполнения заданий по первому тестированию составил 52%. По второму тестированию реальный средний процент выполнения равен 57%, а средний прогнозируемый процент составляет 56%.

Результаты на гистограмме отклонений имеют вид, близкий к нормальному

распределению, но с видимыми флуктуациями, которые могут быть связаны с небольшим объёмом выборки (Рис. 4).

Основная масса отклонений находится в интервале $(-2, +2)$. Среднее значение отклонения прогноза от реального балла равно $-0,15$, а среднеквадратичная ширина гистограммы отклонений равна 3,4 балла. Стандартное отклонение и ошибка для второго тестирования составляли 4,2 и 2,1 балла соответственно.

Обществознание

По обществознанию анализировались результаты тестирования в 8-х классах и тестирования в 9-х классах, проводившегося через девять месяцев после первого. Объём выборки учащихся, участвовавших в обоих тестированиях, составил 1 105 человек. Результаты на гистограмме отклонений прогноза от реального балла имеют вид, близкий к нормальному распределению (Рис. 5).

В случае обществознания практически совпали все три средних процента выполнения: первого тестирования, второго тестирования и прогноза результатов второго тестирования: 70%. Этот результат очевиден, если учесть, что индивидуальные проценты выполнения заданий по первому и второму тестированиям будут совпадать, если равны средние шансы выполнения двух вариантов. А средние шансы в данном контексте совпадают, если совпадают средние проценты выполнения вариантов двух тестирований.



Рис. 3. Гистограмма отклонений прогноза от реального балла. Русский язык.

Среднее значение отклонения прогноза от реального балла равно $-0,003$, а среднеквадратичная ширина гистограммы отклонений равна 4,2 балла. Стандартное отклонение и ошибка для второго тестирования составляли 4,2 и 2,4 балла соответственно.



Рис. 4. Гистограмма отклонений прогноза от реального балла. Физика.



Рис. 5. Гистограмма отклонений прогноза от реального балла. Обществознание.

Выводы

Отклонение прогноза не более чем на одну ошибку теста от истинного балла испытуемых следует ожидать приблизительно в двух третях случаев. Прогноз отличается от реального балла более чем на одну ошибку приблизительно в половине случаев. То есть можно считать, что одна шестая часть результатов значительно отличается от прогнозируемых. Это могут быть те случаи, для которых не выполняется предположение о неизменности способности испытуемого. Строгое выполнение этого предположения означает неизменность рейтингов испытуемых в рамках двух тестирований. Это не всегда так по причинам как субъективного, так и объективного характера. Перескоки между соседними текущими отметками обучающихся: 2–3, 3–4, 4–5 происходят в процессе обучения гораздо чаще, чем между экстремальными отметками: 2–5, например.

Предложенная методика апробировалась также на парных тестированиях по истории в 10-х и 11-х классах, по химии в 8-х и 9-х классах, по английскому языку в 7-х и 8-х классах, а также на диагностиках метапредметных умений в 8-х и 9-х классах. Результаты апробации по этим предметам аналогичны представленным выше результатам.

Описанная методика даёт разумные результаты для прогноза выполнения гомогенных тестов. Нами, в частности, было установлено, что для выборки школьников, принимавших участие в первом тестировании по математике, во втором тестировании получены более высокие результаты (60%), чем результаты тех школьников, которые в первом тестировании участия не принимали (56%). При этом в первом тестировании результаты выборки и всех участников практически совпали: средний процент выполнения был 52%. Этот факт может свидетельствовать о том, что мероприятия мониторингового контроля обладают свойством положительного воздействия на результаты дальнейшего обучения.

Использование нескольких предварительных результатов испытуемых для прогноза позволяет учесть тренд изменения способностей по формулам типа (1), (3)–(5), а затем вычислять прогнозируемые проценты выполнения нового теста.

Методику прогнозирования результатов можно рекомендовать при проведении регулярных диагностических мероприятий в образовательных организациях для планирования деятельности по управлению качеством образования.