Методология

МЕТОД НЕЗАВИСИМОГО ШКАЛИРОВАНИЯ РЕЗУЛЬТАТОВ ЕГЭ

Юрий Каргин,
Ноябрьский колледж профессиональных
и информационных технологий
kargin04@yandex.ru

В предлагаемой статье рассматривается проблема качества оценок ЕГЭ. Актуальность этой проблемы состоит в том, что разработчики и организаторы проведения педагогического контроля знаний в федеральном масштабе кардинально изменили систему шкалирования результатов ЕГЭ в 2011 году. Вследствие чего существующая в России система оценивания знаний из метрической области оказалась сдвинутой в область описательных характеристик и неметрических оценок. Предложена новая методика шкалирования результатов ЕГЭ. Методика проверена на данных, взятых из официально опубликованных результатов ЕГЭ-2012[1].

Ключевые слова: педагогическое тестирование, педагогические измерения, шкалирование.

«Сделай так просто, как возможно,
но не проще этого».
А. Эйнштейн

О предисловии к официальной методике

Предисловие к описанию реально применяемой сейчас технологии шкалирования результатов ЕГЭ начинается с постановки це-

1 http://www.ege.edu.ru

ли: «Для объективной оценки уровня выполненной работы каждого участника ЕГЭ, по сравнению с другими участниками экзамена...». Далее дипломатично представлены методологические основания. Методика шкалирования «разработана ведущими отечественными специалистами на основе признанных международных тестологических моделей». Непонятно, кого авторы предисловия имеют в виду «под специалистами» и о каких «тестологических» моделях идёт речь. Для начала напомним, что таких моделей просто нет, но есть модели педагогического измерения. Путаница и подмена в таких делах нежелательны.

Естественно, что библиографические списки или другие указания на этот счёт не приводятся, как не указываются и публикации отечественных или зарубежных авторов по обоснованию используемой технологии.

Действующая методика шкалирования результатов ЕГЭ

Приведём кратко содержание методики шкалирования и некоторые результаты её применения. С формальной точки зрения процедура преобразования наблюдаемых первичных баллов ПБ в тестовые ТБ сводится к следующим математическим правилам (с округлением до ближайшего большего целого значения):

$$ТБ=\begin{cases}\dfrac{ТБ1}{ПБ1}\cdot ПБ, \text{ при } ПБ\le ПБ1,\\ ТБ1+\dfrac{ТБ2-ТБ1}{ПБ2-ПБ1}\cdot\\ \cdot(ПБ-ПБ1),\\ \text{при } \quad ПБ1<ПБ\le ПБ2,\\ ТБ2+\dfrac{100-ТБ2}{ПБмакс-ПБ2}\cdot\\ \cdot(ПБ-ПБ2),\\ \text{при } \quad ПБ2<ПБ\le ПБмакс.\end{cases}$$

Здесь ПБ1 — наименьший первичный балл, «получение которого свидетельствует об усвоении участником экзамена основных понятий и методов по соответствующему общеобразовательному предмету», ПБ2 — наименьший первичный балл, «получение которого свидетельствует о высоком уровне подготовки участника экзамена».

Эти значения определяются «на основе экспертизы демонстрационного материала по данному общеобразовательному предмету специалистами общего образования, ссузов и вузов различного профиля из разных субъектов РФ».

Первичным баллам ПБ1 и ПБ2 ставятся в соответствие установленные распоряжениями Рособрнадзора тестовые баллы ТБ1 и ТБ2 по каждому общеобразовательному предмету. Эти значения совпадают с прошлогодними или устанавливаются после проведения основного экзамена (только значения ТБ1 для предметов по выбору).

Так, для обязательных русского языка и математики таблица граничных первичных и тестовых баллов, а значит, и правила преобразования определяются ещё до проведения экзаменов, а для предметов по выбору один параметр (ТБ1) из четырёх определяется после статистической обработки результатов основного экзамена. Максимально возможные значения первичного балла ПБmax для каждого предмета свои, а максимальное значение тестового балла для любого предмета устанавливается равным ТБmax = 100.

Эти правила преобразования первичных баллов ПБ в тестовые баллы ТБ можно представить и графиком (рис. 1). Причём в отличие от аналогичного рисунка из представленного на официальном портале ЕГЭ, мы представляем его в том виде, в котором это принято делать в теоретических работах по педагогическим измерениям — по вертикальной оси наблюдаемые значения и по горизонтальной оси соответствующие оценки.

В завершение в табл. 1 приведём утверждённые Рособранадзором граничные значения первичных и тестовых баллов, а также средние значения первичных баллов ПБср и тестовых баллов ТБср, полученные

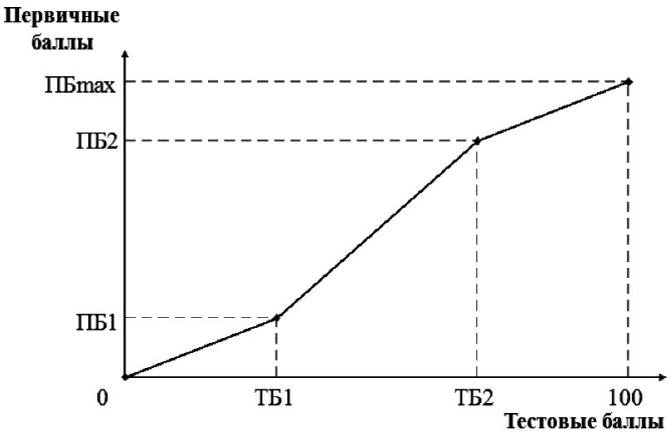

Рис. 1. Правила преобразования результатов ЕГЭ

Таблица 1

Значения граничных и средних первичных и тестовых баллов в 2012 г.

№ п/п	Предмет	ПБмакс	ПБ1	ТБ1	ПБ2	ТБ2	ПБср	ТБср
1	Русский язык	64	17	36	54	73	42,1	61,1
2	Математика	32	5	24	15	63	10,3	44,6
3	Обществознание	59	15	39	48	72	31,2	55,2
4	История	58	13	32	46	72	28,8	51,1
5	Физика	51	12	36	33	62	20,6	46,7
6	Химия	66	14	36	58	80	35,3	57,3
7	Биология	69	17	36	60	79	35,0	54,0
8	География	54	14	37	44	69	31,6	55,8
9	Информатика	40	8	40	35	80	21,7	60,3
10	Иностранные языки	80	16	20	65	82	48,2	60,8
11	Литература	42	8	32	36	73	24,6	56,3

усреднением по всем участникам экзамена по предмету (в строке «иностранные языки» приведены средние значения для английского языка).

Анализ методики шкалирования результатов ЕГЭ

Первый вопрос к авторам методики шкалирования возникает уже из предисловия к ней. Как следует понимать фразу «для объективной оценки уровня выполненной работы»? В науке «под объективными знаниями» понимают такие, которые не зависят от человека, от его суждений. Объективность противостоит субъективности. Данная методика основывается на экспертных заключениях, т.е. на субъективных суждениях, называть её или результаты её применения объективными ошибочно. Может, правильнее было бы сформулировать «для квалифицированной оценки...».

Поскольку теоретические основания принятой технологии шкалирования результатов ЕГЭ авторами не представлены, остаётся обратиться к анализу практики её применения. Для этого представим данные табл. 1 в более наглядном виде. На рис. 2 приведены графики правил преобразования для трёх предметов — обязательных экзаменов по русскому языку и математике и претендующему в будущем на обязательный экзамен по иностранному языку.

Для корректного сопоставления графиков по разным предметам с разными диапазонами изменения первичных

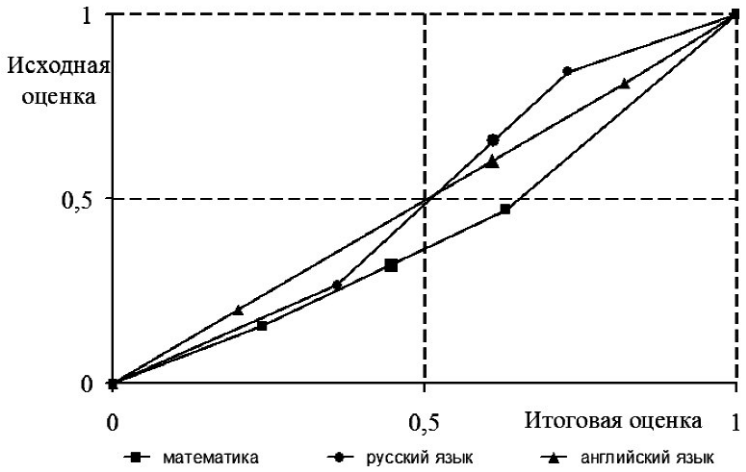

Рис. 2. Правила преобразования результатов ЕГЭ по предметам

баллов мы привели исходные данные к нормированным на единицу шкалам ПБ/ПБмакс и ТБ/100. Тогда нормированный на единицу первичный балл равен доле набранных испытуемым первичных баллов и представляет собой первичную оценку уровня выполнения экзаменационной работы, а нормированный на единицу тестовый балл равен доле набранных испытуемым тестовых баллов и представляет собой итоговую оценку уровня подготовленности по предмету. Граничные значения отмечены маркером, средние значения — маркером большего размера.

Следующий рисунок отражает экспертные заключения о граничных значениях. В основе официальной методики шкалирования результатов ЕГЭ лежит выделение экспертами трёх областей изменения первичных и тестовых баллов по уровням подготовки экзаменующихся — область недостаточного усвоения основных понятий и методов по соответствующему общеобразовательному предмету, центральная область и область высокого уровня подготовки участников экзамена.

Для каждой области связь между ПБ и ТБ линейная, т.е. в каждой области свой коэффициент перевода приращения первичного балла в приращение тестового балла, а значит, и своя степень разрешения испытуемых. На рис. 3 приведены эти области по одиннадцати предметам, сами области выделены разными оттенками, линией внутри центральной области обозначено усреднённое по всем испытуемым значение итоговой.

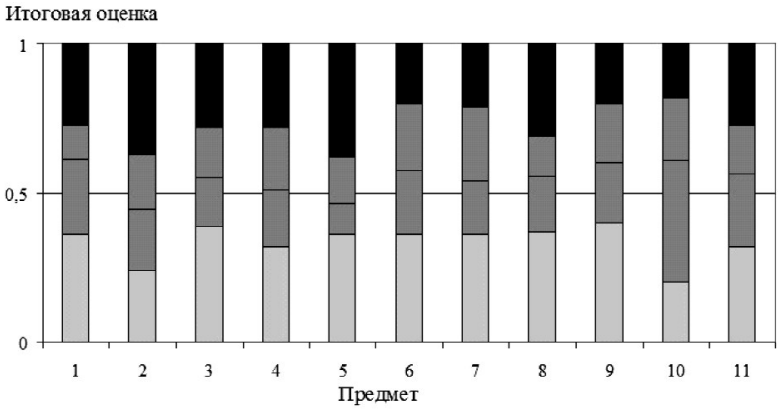

Рис. 3. Оценки уровней подготовленности по предметам

Приведём некоторые пояснения к этому рисунку. Первый столбец соответствует русскому языку. Самый светлый оттенок соответствует оценкам недостаточного уровня подготовки и его граничное значение для русского языка равно $^{36}/_{100}$ = 0,36. Нижняя граница оценок высокого уровня подготовки — 0,73, который выделен самым тёмным оттенком. И наконец, линией в центральной области выделена усреднённая оценка 0,611 по всем участникам экзамена по русскому языку. Второй столбец по математике с соответствующими значениями оценок — 0,24; 0,63; 0,446 и т.д. по остальным предметам.

Если обратиться к рис. 1, то представленное на нём кусочно-линейное преобразование шкалы первичных баллов в тестовую шкалу напоминает логистическую функцию. Тогда можно предположить, что данная технология есть некоторая упрощающая модификация математической теории измерений (IRT). Подобная логика рассуждений, конечно, допустима. Только при этом следует понимать, что линейные или другие упрощающие аппроксимации неизбежно ведут к потере точности[2].

Действительно, и в логистической кривой можно выделить относительно «пологие» области низкого и высокого уровней подготовленности испытуемых, и центральную область среднего уровня подготовки с более выраженным угловым коэффициентом.

Однако практика применения данной методики шкалирования не позволяет принять эту версию. Для этого достаточно обратиться к рис. 2. Если для русского языка ломаная линия преобразования баллов

2 Познакомиться с этими направлениями исследований можно по публикациям психометрического исследовательского центра национального института образовательных измерений CITO, Нидерланды (www.cito.com).

действительно напоминает логистическую функцию, то для математики она имеет более выраженный угловой коэффициент в области высокого уровня подготовленности, а для английского языка ломаная линия вообще вырождается в прямую с постоянным угловым коэффициентом.

Для иностранных языков правила преобразования первичных баллов в тестовые имеют достаточно простой вид:

$$\text{ТБ} = \frac{100}{\text{ПБмакс}} \cdot \text{ПБ}.$$

В этом случае получение тестового балла можно трактовать как пересчёт первичного балла на шкалу процентов. Часто именно так интерпретируется тестовый балл в среде учителей, выпускников школ и их родителей. Тогда значения ТБ1 = 20 можно трактовать примерно так: если экзаменующийся набрал менее 20% баллов из максимально возможного числа, то его уровень подготовленности по предмету неудовлетворительный.

Именно этот простейший способ приведения к единой стобалльной шкале результатов различных экзаменов является самым широко используемым в практике работы учителя. В нём заложены и здравый смысл, и достаточно внятная интерпретация результата. *Принципиальным недостатком такой измерительной шкалы является её неметрический характер*. Она лишь упорядочивает испытуемых по уровню подготовленности, но не указывает меру этой подготовленности.

Однако и официальная технология шкалирования не является метрической. Этот вывод следует из её опоры на экспертные заключения, из выделения в единой шкале областей различных свойств, из отсутствия единицы измерения, а также из отсутствия оценок точности таких измерений. А если учесть, что отсутствует и внятная интерпретация результатов такого оценивания по учебным предметам, то процентная система вообще является предпочтитель-ней.

Неметрический характер официальной системы шкалирования

Неметрический характер официальной системы шкалирования проявляется в виде парадоксальных последствий при сопоставлении результатов экзамена по различным предметам. Одно из основных назначений применения результатов ЕГЭ состоит в формировании итогового проходного балла в вуз.

Официальная система формирует проходной балл через прямое сложение набранных тестовых баллов по выделен-

ным предметам. Математические действия с неметрическими величинами (т.е. величинами, не имеющими меры), пусть даже обозначенными в виде чисел, неправомерны даже для результатов одной измерительной системы — экзамена по одному предмету. Тем более, прямое сложение тестовых баллов по различным предметам оцененных в различных порядковых шкалах уже напоминает шутку по скрещиванию ежа и ужа для получения метра колючей проволоки.

Действительно, сравнивая результаты оценивания учебных достижений выпускников по различным предметам, начинаешь понимать, что прежняя конкурсная система приёма в вуз была по некоторым позициям даже более адекватна. В прежней системе шкала результатов была единая для всех вузов и предметов. Там за недостаточную подготовленность ставили «двойку» по любому предмету, а за очень хорошую подготовленность по любому предмету любая комиссия ставила «пятёрку». Иногда эти требования корректировали под профиль специальности обучения.

Теперь за недостаточную подготовленность по одному предмету устанавливаются одни баллы, а по другому предмету — другие. Например, недостаточная подготовленность по иностранным языкам оценивается в диапазоне от 0 до 20 тестовых баллов, а по информатике от 0 до 40 тестовых баллов.

Такая же ситуация со средним и высоким уровнем подготовки. Этот разнобой в итоговых оценках в зависимости от уровня подготовленности экзаменующихся по различным предметам достаточно наглядно выражен на рис. 3. А затем эти баллы ещё и складывают. О профориентации в данной системе речи вообще не идёт.

Некорректность и противоречивость принятой системы оценивания

Несложно привести пример, демонстрирующий указанную некорректность в принятой системе шкалирования. Один выпускник набирает по математике 63 балла, что соответствует пусть наименьшему, но баллу, «получение которого свидетельствует о высоком уровне подготовки участника экзамена».

Другой выпускник с тем же баллом 63 по русскому языку показывает самый настоящий средний уровень подготовки (табл. 1). Т.е. выпускникам за явно разные, по мнению экспертов, достижения приписываются одинаковые баллы. Или иначе, за средние дости-

жения по русскому языку и математике в данной системе приписываются заметно различающиеся тестовые баллы: 61 и 45 соответственно.

Что делать?

Эти противоречия можно частично устранить даже в рамках данной системы. Для этого достаточно упростить методику шкалирования. Пусть предметные экспертные комиссии устанавливают по содержанию демонстрационных вариантов граничные значения первичных тестовых баллов ПБ1 и ПБ2 с тем же содержательным смыслом. Но соответствующие им тестовые баллы ТБ1 и ТБ2 должны принимать целочисленные значения и быть одинаковыми для любого предмета. Если остановиться на стобалльной системе, то можно предположить следующие крайние значения: ТБ1 = 30, ТБ2 = 70. Но более адекватной будет целочисленная шкала с пониженными значениями, например ТБ1 = 10, ТБ2 = 20 и ТБмакс = 30. Аргументация преимуществ подобной единой шкалы тестовых баллов состоит в следующих рассуждениях.

Единая шкала не только снимет непонимание значения оценок или ошибочную процентную трактовку стобалльной шкалы тестовых баллов, но и даст внятную трактовку самого численного результата. Итоговый результат менее 10 баллов будет указывать на неудовлетворительную, по мнению экспертов, подготовленность к экзамену, а, например ,результат 15 баллов будет трактоваться как примерно средний уровень подготовленности по учебному предмету и т.д.

Неточность оценок

Второй аргумент относится к точности оценок. Любой здравомыслящий учитель-предметник скажет, что разница 1 первичный балл (что соответствует примерно 2 тестовым баллам) совершенно незначительна. По оценкам О.Г. Деменчёнка[3], погрешность вообще составляет не менее 5 (а реально 9) тестовых баллов. Справедливости ради следует отметить, что его оценки проводились для другой системы шкалирования. Но этот вывод следует из количества заданий и справедлив для количества заданий менее 100. Из этого вывода следует, что 100-балльную шкалу можно без потери точности сжать вплоть до 15 или 20-балльной.

И наконец, очень важный с практической точки зрения вопрос формирования конкурсного балла. Действительно, достаточно трудоёмко подготовить экзаменационные задания

3

Деменчёнок О.Г. Погрешность баллов Единого государственного экзамена //Педагогические измерения. №4. 2011. С. 3–17.

по различным предметам с близкими распределениями заданий по уровню трудности. Но подготовить наборы экзаменационных заданий по учебным предметам с близкими уровнями трудности самих наборов заданий вполне посильная задача. По итогам экзаменов с 2009 по 2013 год она не решена.

Достаточно проследить результаты по двум основным экзаменам — русскому языку и математике. В среднем экзаменующийся по русскому языку набирает первичный балл со значением около $^2/_3$ от ПБмакс, а экзаменующийся по математике лишь $^1/_3$ от ПБмакс. Процедура шкалирования несколько нивелирует эти различия и доводит их в 2012 году до значений ТБср = 61,1 по русскому языку и ТБср = 44,6 по математике, но эти различия всё равно остаются очень значительными.

Причина такого различия одна. Набор экзаменационных заданий по русскому языку для экзаменующихся существенно легче набора заданий по математике. Корректировать эти различия и должна процедура шкалирования. По-хорошему, процедура шкалирования должна различающие средние первичные баллы привести к единым тестовым баллам. Почему это очевидное требование уравнивания результатов экзаменов по различным предметам не выполняется? Непонятно.

Расчёт тестовых баллов по предложенной выше схеме с единой стобалльной шкалой даёт значения ТБср = 57,1 по русскому языку и ТБср = 51,1 по математике, т.е. различия существенней уменьшаются. При переходе к 30-балльной единой шкале тестовых баллов эти значения будут соответственно равны 16,8 и 15,3. Т.е. предложенная система эффективней решает задачу выравнивания результатов экзаменов по различным предметам.

Преимущества 30-балльной шкалы

Преимущества 30-балльной шкалы проявляются при сравнении результатов экзаменующихся. Только в этом случае разницу 1 тестовый балл можно начинать воспринимать как разницу в уровне подготовленности. Аргументацию преимуществ 100-балльной шкалы, связанную с попыткой увеличить разрешающую способность экзамена, в лучшем случае можно назвать неубедительной.

Дело в том, что разница между значениями 50 и 51 тестовых баллов никаких различий в подготовленности испытуемых не отражает, эти результаты для одномоментного экзамена неразличимы. Кто это не пони-

мает, тот заблуждается в правилах интерпретации результатов. Но разница между значениями 15 и 16 тестовых балов в 30-балльной шкале уже может дать некоторые основания для выделения различий в уровне подготовленности экзаменующихся.

Недостатки принятой сейчас системы шкалирования оценок ЕГЭ

В завершение этой первой части работы выделим ещё раз основные выводы анализа принятой системы шкалирования. Кусочно-линейное преобразование шкалы первичных баллов в тестовую шкалу сводится к формированию трёх порядковых шкал последовательно оценивающих экзаменующихся с недостаточным уровнем подготовленности, с достаточным уровнем подготовленности и с высоким уровнем подготовленности.

Разрешающие свойства этих трёх шкал фактически задаются мнением экспертной группы. Интерпретация тестовых баллов сводится к ранжированию испытуемых и позволяет отнести испытуемого к одной из трёх групп. Точность таких оценок ничем не подтверждается[4].

Единственным общим признаком тестовых шкал по различным предметам является их диапазон от 0 до 100 баллов. Правила назначения промежуточных значений тестовых баллов для каждого предмета свои и никак с другими предметами не связаны. Сопоставлять промежуточные значения тестовых баллов по различным предметам некорректно. Тем более ошибочно составлять простым суммированием тестовых баллов конкурсный балл абитуриента.

Таким образом, мы вынуждены констатировать, что предложенная методика шкалирования результатов ЕГЭ некачественно решает взятую на себя задачу «объективной оценки уровня выполненной работы каждого участника ЕГЭ, по сравнению с другими участниками экзамена...».

Трудность тестового задания

На наш взгляд, основной источник проблем действующей сейчас методики шкалирования заложен в отсутствии какой-либо научной основы. Официальная методика носит скорее умозрительный, искусственный характер и, по-видимому, не имеет под собой ни теоретического обоснования, ни достаточного эмпирического подтверждения.

Конечно, ограничиваться только критикой действующей

4 В официальной технологии точность получаемых оценок не обсуждается. Поэтому важно понимать, что оценка надёжности экзаменационных вариантов по значению коэффициента альфа Кронбаха отражает лишь меру взаимной согласованности заданий одного теста.

методики шкалирования без формулировки предложений, устраняющих выделенные недостатки, было бы неправильным. Эта задача решается в заключительных частях данной работы. В данной части кратко остановимся на понятии меры трудности тестового задания. Надеемся, что эти рассуждения помогут читателю лучше понять логику измерения педагогических отношений, заложенную в альтернативном варианте системы шкалирования результатов ЕГЭ.

Единого и определённого мнения в понимании содержания понятия «трудность тестового задания» в педагогической литературе пока не существует. Если исключить интуитивные объяснения и остановиться только на тех подходах, которые обсуждают и используют количественные способы оценки меры трудности учебного задания, то можно выделить два основных направления в решении этой проблемы.

Одни исследователи предполагают, что трудность задания определяется включённым в него учебным содержанием, меру трудности задания можно определить ещё до проведения тестовых испытаний, трудность задания никак не связана с результатами его выполнения испытуемыми. Но это неверно.

Другие основываются на стохастическом характере выполнения испытуемым учебного задания и полагают, что трудность задания проявляется лишь в ходе его выполнения и меру трудности можно оценить по результатам выполнения задания большой группой испытуемых. Несложно привести примеры таких подходов.

Разрабатывая теорию дидактических систем, В.П. Беспалько[5] в качестве «кирпичика знаний» использует понятие учебного элемента и его количественной характеристики — объём учебной информации. Эта величина рассчитывается по содержанию учебного материала, имеет единицу измерения бит, обычно изменяется в диапазоне от нескольких десятков до тысяч бит и в среднем учебный элемент содержит 350 бит информации. По мнению автора, такой подход даёт не только возможность планировать учебный процесс исходя из дидактических целей, но и оперативно управлять им на основе строго диагностируемого контроля усвоением знаний.

Для количественного описания свойств теста (В.П. Беспалько одно тестовое задание называет тестом) автор вводит два понятия — трудность теста и сложность задания. Первый показатель может принимать целые значения от 1 до 4, каждое из которых соответствует одному из уровней усвоения — узнавание, воспроизведение,

5 *Беспалько В.П.* Образование и обучение с участием компьютеров. М.: Издательство Московского психолого-социального института, 2002.

умение и творческое применение. Второй показатель определяется подсчётом числа существенных операций в задании.

Иначе дело обстоит в статистической теории тестов. Здесь используется следующая логика: чем меньше испытуемых правильно выполняют тестовое задание, тем оно труднее. Тогда количественным показателем меры трудности задания может служить доля неправильных ответов всеми испытуемыми на тестовое задание. По доле неправильных ответов несложно ранжировать тестовые задания в порядковой шкале или группировать задания в классы трудности, качественно описывать эти классы. В этом подходе предполагается, что трудность тестового задания уже обусловлена не только его содержанием, но и затруднениями испытуемых при его выполнении.

Конечно, вопрос выявления причин трудности тестового задания может быть и самостоятельным. Но в контексте проведения единого госэкзамена он относителен, имеет промежуточное значение. Он актуален для решения более важного вопроса — вопроса оценивания подготовленности испытуемых. В конечном итоге тестовые задания для этого разрабатываются и используются. И в этом смысле более важны функции, свойства, которые проявляет тестовое задание при проведении контроля знаний.

В этой связи можно провести некоторую аналогию с проблемой определения фундаментального экономического понятия «цена». Одни учёные-экономисты природу цены объясняют объективными затратами на производство товара, другие — массовой субъективной оценкой полезности. Каждая из этих позиций имеет свои аргументы, свои методы прогноза и обоснования наблюдаемых фактов. Однако все экономисты одинаково выделяют основную функцию цены товара — роль коэффициента обмена между разнородными товарами.

Пример

Изложим эту мысль на примере. Продавец обменивает на рынке два гуся на овцу, т.е. для него цена овцы в два раза выше цены одного гуся. Для одного акта обмена именно это соотношение ценностей товаров случайно. Однако если это отношение в среднем удовлетворяет всех участников товарного обмена, то оно перестаёт быть случайным и становится в данных условиях реальным. Это соотношение уже не зависит от конкретных участников обмена, не зависит от валюты, в которой выражены цена овцы или цена гуся и вообще от на-

личия денег, и не зависит от того, как пытаются люди рассчитывать цены товаров — затратами на их производство или полезностью.

Подобная ситуация встречается достаточно часто. Физики считают, что они только сейчас близки к пониманию смысла понятия «массы тела», хотя история измерения масс тел, разработки методов их измерения и применения этих знаний на практике насчитывает тысячелетия. Эти же рассуждения справедливы и при измерениях длины, промежутков времени и др. В классическом естествознании под измерением как раз и понимают определение отношения величины к принятому эталону.

Единица измерения, основанная на отношениях

Если подобную логику применить к проблеме измерения трудности тестового задания, то задача состоит в определении отношения меры трудности одного задания к мере трудности другого. В таких измерениях всегда присутствует произвол, связанный с выбором эталона. Предложить эталонное тестовое задание, трудность которого удобно принять за единицу измерения, мы не можем. Отсюда и термин «опорное», который мы используем для тестового задания с функциями эталона в данном тесте.

И наконец, необходимо отметить ещё одну особенность альтернативного метода. С формальной точки зрения в альтернативном методе присутствуют две равноправные группы показателей — уровня трудности тестовых заданий и уровня подготовленности испытуемых. Результатом обработки тестовых данных альтернативным методом является оценка отношений не только уровней трудности тестовых заданий, но и уровней подготовленности испытуемых. Отсюда и применяемый ниже альтернативный метод можно характеризовать как метод измерения педагогических отношений.

Альтернативный метод шкалирования результатов ЕГЭ

Теоретической основой альтернативного метода обработки результатов теста служит математическая модель, задающая результат тестового испытания в зависимости от показателей модели[6]:

- вероятность правильного ответа испытуемым на тестовое задание $P = 0{,}5^{\frac{\beta}{\theta}}$ или для вероятностных показателей

6

Каргин Ю.Н. Педагогические измерения в шкале отношений// Педагогические измерения. №2. 2012. С. 3–26.

$$P = \log_{0,5} u \times \log_{0,5} v;$$

- относительные показатели уровня подготовленности испытуемых θ и уровня трудности тестовых заданий β, значения которых измеряются в опорных единицах;
- вероятностные показатели уровня подготовленности испытуемых u и уровня трудности тестовых заданий v, значения которых измеряются в вероятностных единицах и однозначно связаны с относительными показателями соотношений $u = 0,5^{\frac{1}{\theta}}$ и $v = 0,5^{\beta}$.

Экспериментальной основой для проведения педагогических измерений служат тестовые испытания и полные данные ответов испытуемых на тестовые задания. Если говорить о ЕГЭ, то доступ к полным данным имеется у организаторов проведения экзамена и по каким-то причинам для всех желающих закрыт. В опубликованной статистике приведены только распределения значений тестового балла ТБ, т.е. после применения к наблюдаемым первичным баллам ПБ процедуры шкалирования.

Результаты выполнения отдельных заданий всеми испытуемыми вообще не приведены. Отсюда и предлагаемый ниже метод обработки результатов ЕГЭ основывается на неполных данных. Он содержит лишь обобщающие оценки уровня трудности всего набора тестовых заданий по предмету без детализации по заданиям.

Таким образом, основой для проведения последующих здесь оценок служит опубликованная статистика распределения значений тестовых баллов испытуемых по отдельным предметам. По известным правилам перевода первичных баллов в тестовые и необходимым граничным значениям несложно вернуться к исходным данным. Для этого достаточно применить обратные преобразования расчёта первичных баллов испытуемых по значениям тестовых баллов (округление не указано):

$$\text{ПБ} = \begin{cases} \dfrac{\text{ПБ1}}{\text{ТБ1}} \cdot \text{ТБ}, \text{ при } \text{ТБ} \le \text{ТБ1}, \\ \text{ПБ1} + \dfrac{\text{ПБ2} - \text{ПБ1}}{\text{ТБ2} - \text{ТБ1}} \cdot \\ \cdot (\text{ТБ} - \text{ТБ1}), \\ \text{при } \text{ТБ1} < \text{ТБ} \le \text{ТБ2}, \\ \text{ПБ2} + \dfrac{\text{ПБмакс} - \text{ПБ2}}{100 - \text{ТБ2}} \cdot \\ \cdot (\text{ТБ} - \text{ТБ2}), \\ \text{при } \text{ТБ2} < \text{ТБ} \le 100. \end{cases}$$

Набранный i-м испытуемым первичный балл, отнесённый к максимально возможному значению ПБmax, можно трактовать как наблюдаемую долю X_i правильных ответов испытуемым на все задания те-

ста. Это значение именуются здесь как исходная оценка уровня выполнения испытуемым экзаменационной работы. Усреднённое по всем испытуемым значение исходной оценки обозначим через X. Это значение фактически опубликовано (табл. 1), оно легко рассчитывается по правилу $X=ПБ_{ср}/ПБ_{макс}$.

Основной показатель теста

Первый и основной показатель теста, значение которого отражает отношение уровня подготовленности группы испытуемых к уровню трудности набора тестовых заданий, вычисляется по формуле: $\gamma_0 = {}^{1}\!/_{\log_{0,5} X}$. Для хорошего теста значение γ_0 должно быть близким к единице. В этом случае уровень трудности тестовых заданий соответствует (близок по значению) уровню подготовленности испытуемых. Если следовать рекомендациям работы[7], для удовлетворительного теста значение этого показателя должно лежать в диапазоне от 0,71 до 1,41 (точнее от $\sqrt{2}/2$ до $\sqrt{2}$). Оценим это значение по материалам ЕГЭ 2012. Как и прежде для иллюстрации будем преимущественно использовать результаты испытаний по обязательным предметам — русскому языку и математике.

Для русского языка и для математики средние значения исходных оценок X соответственно равны ${}^{42,1}\!/_{64} = 0{,}66$ и ${}^{10,3}\!/_{32} = 0{,}32$, отсюда значения основного показателя теста γ_0 равны 1,65 и 0,61. Они явно выпадают из удовлетворительного диапазона. Отсюда первый вывод — уровень трудности набора заданий ЕГЭ по русскому языку заметно занижен, а для математики заметно завышен по отношению к уровню подготовленности испытуемых.

Гипотетически, предыдущие оценки можно трактовать и иначе, а именно высоким уровнем подготовленности экзаменующихся по русскому языку и низким по математике. Но такое предположение вряд ли удастся обосновать. Более того, отечественная система подготовки школьников именно по математике находит широкое признание за рубежом и сомневаться в низком качестве этой подготовки нет убедительных оснований. А вот предположить, что в среднем уровень подготовленности испытуемых по различным предметам примерно одинаковый, вполне допустимо. Статистическим основанием для такого предположения может служить и большое количество экзаменующихся и достаточно большое количество учителей, занимающихся их подготовкой.

7 Там же. С. 16.

Частные показатели теста

На следующем этапе необходимо оценить уровень выполненной экзаменационной работы по предмету каждым участником ЕГЭ. Для решения этой задачи в альтернативном подходе существуют различные приближённые методы[8]. Одним из самых простых, предлагающих достаточно надёжные и хорошо интерпретируемые результаты, являются решения первого приближения (элементарный метод[9]).

В основе этого метода лежат исходные предположения об уровнях трудности тестовых заданий и/или об уровнях подготовленности испытуемых (т.н. система отсчёта) и определение опорного участника теста. В системе отсчёта «нейтральная группа испытуемых» усреднённое по всем экзаменующимся значение уровня подготовленности $u_0 = 0{,}5$. Испытуемого с таким усреднённым уровнем подготовленности будем называть опорным, относительный показатель его подготовленности θ_0 примем за единицу измерений, и все последующие измерения относительных показателей теста будем проводить в этих единицах измерения.

Тогда уровень подготовленности испытуемого с наблюдаемой долей X_i правильных ответов рассчитывается по формулам: $\theta_i \approx {}^{1}\!/_{\gamma_0 \log_{0,5} X_i}$ — относительное значение $u_i \approx X_i^{\gamma_0}$ — вероятностное значение. Т.к. официальная статистика не содержит данных ответов испытуемых по отдельным тестовым заданиям, то мы можем проводить только усреднённые оценки по самому набору тестовых заданий: $\beta = \log_{0,5} X$ — относительный уровень трудности набора тестовых заданий; $v = X$ — вероятностное значение этого уровня.

Альтернативный метод шкалирования

Если перейти к обозначениям официальной методики шкалирования результатов ЕГЭ, то правила преобразования первичных баллов ПБ в 100-балльные тестовые баллы ТБ испытуемых альтернативным методом примут вид:

$$\text{ТБ} = 100 \cdot (\text{ПБ}/\text{ПБ}_{\max})^{\gamma_0}.$$

где показатель подготовленности группы испытуемых по отношению к набору экзаменационных заданий можно записать в более традиционном виде через натуральный логарифм

$$\gamma_0 = \ln 2 / \ln (\text{ПБ}_{\max} / \text{ПБ}_{ср})$$

На рис. 4а и 4б мы привели графики, иллюстрирующие официальные и альтернативные правила преобразования первичных баллов ПБ в тесто-

8 Там же. С. 19.

9 *Каргин Ю.Н.* Элементарное решение основной задачи педагогических измерений // Педагогические измерения. 2011. № 4. С. 50–67.

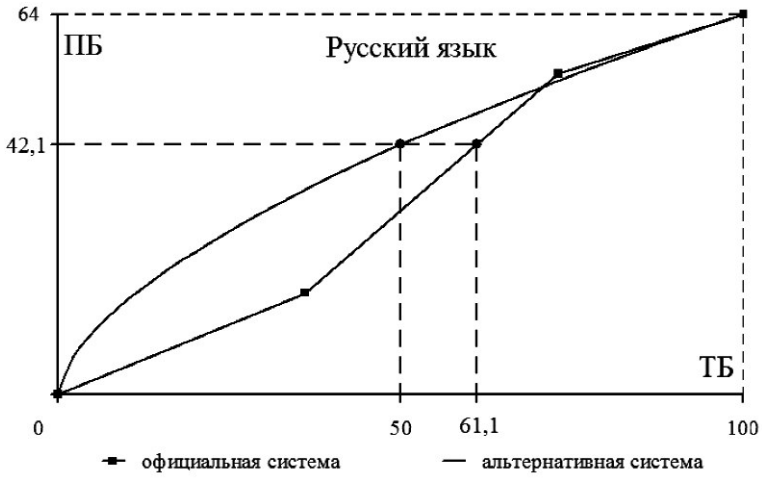

Рис. 4а. Правила преобразования результатов ЕГЭ по русскому языку

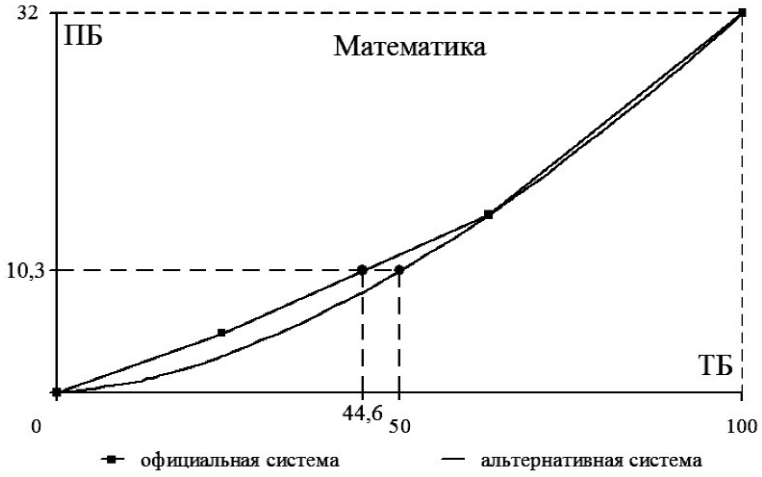

Рис. 4б. Правила преобразования результатов ЕГЭ по математике

вые баллы ТБ для двух предметов. По графикам можно даже визуально выделить области хорошего согласия правил шкалирования и области их существенного различия.

По виду графиков правил преобразования для русского языка (рис. 4а) можно говорить о достаточно хорошей согласованности систем шкалирования в области высокого

уровня подготовленности участников экзамена — в области высоких значений экзаменационных баллов. Для испытуемых с низким и средним уровнями подготовленности графики существенно различаются и о согласованности систем шкалирования говорить не приходится. На рис. 4б для математики удовлетворительная согласованность проявляется во всём диапазоне изменения экзаменационных баллов.

Напомним, что официальная система шкалирования основана на субъективных экспертных мнениях и не застрахована от свойственных человеку заблуждений. Именно в этом мы видим причину имеющихся разногласий. Альтернативная система шкалирования напротив, носит формализованный характер и никак не связана с субъективными оценками. Фактически характер кривой шкалирования задаётся только значением основного показателя тестовых испытаний — средним наблюдаемым значением первичного балла ПБср.

В частности, если набор экзаменационных заданий оптимально соответствует группе экзаменующихся, что подтверждается равенством $ПБср = 0{,}5 \cdot ПБ_{макс}$, то правила преобразования сводятся к простому переводу в шкалу процентов $ТБ = 100 \cdot ПБ/ПБ_{max}$. Именно это правило находит широкое применение в учительской среде. И оно оправданно, но только для хорошо подобранных под уровень подготовленности учащихся контрольных заданий. В этом случае график преобразований имеет вид прямой пропорциональности.

Любые измерения предполагают оценку их точности. Выше указывалось о расчётах О. Деменчёнка и оценке погрешности результатов ЕГЭ значением в 9 тестовых баллов. Мы можем только подтвердить его выводы, подчёркивая, что это минимальные оценки. В статистике для повышения надёжности выводов принято использовать разные методы решения статистической задачи. Следуя этой логике, мы провели альтернативные оценки точности педагогических измерений с применением углового преобразования Фишера (критерий φ^*). Этот метод позволяет оценить достоверность различий между процентными долями двух выборок. Рассматривая результат экзамена как процентную долю набранных баллов от максимально возможного значения, мы получили, что минимальные значения погрешности составляют от 11 до 15 баллов (уровень значимости 0,1). Таким образом, если строго подходить к вопросу разделения испытуемых по уровню подго-

товленности, то следует применять не 100-балльную, а 10-балльную шкалу. Только в этом случае разница оценок в один тестовый балл будет статистически значимой.

Высокая погрешность при проведении педагогических измерений — известная проблема, но и игнорировать её — значит уходить от ответственности за результаты измерений. Напомним, в представлении официальной технологии расчёта тестовых баллов вопросы о точности оценок никак не обсуждаются. Способы снижения погрешности педагогических измерений видятся сегодня в развитии и разработке рекомендаций по двум направлениям. Первое связано с конструированием научно-обоснованных тестов через разработку тестовых заданий с заданными метрическими свойствами, например с основой на систему измерений Раша. Второе направление связано с переходом от единого экзамена к сериям экзаменационных испытаний.

Указание правил преобразования наблюдаемых данных в метрические оценки ещё не дают их понимания. Правила преобразования представляют собой своего рода ключ к интерпретации и пониманию результатов экзамена. Мы уже указывали об отсутствии возможности внятно прокомментировать результаты ЕГЭ официальной методики. Официальные результаты фактически позволяют только ранжировать экзаменующихся по предмету и отнести каждого из них к одной из трёх групп уровня подготовленности, выделенных экспертами. Индивидуальные значения тестовых баллов другой информации не несут. Другое дело — измерения уровней подготовленности в метрической шкале с единицей измерения, пусть даже без «привязки» к описательным требованиям КИМ.

Интерпретация результатов альтернативного метода шкалирования результатов ЕГЭ

Начнём с описания результатов измерений трудности заданий. Значения вероятностного показателя уровни трудности наборов тестовых заданий по русскому языку и математики в данной системе измерений равны соответствующим долям X:$v = 0{,}66$ для русского языка и $v = 0{,}32$ для математики. Эти значения содержат следующий содержательный смысл — опорный испытуемый правильно выполняет усрёднённое задание по русскому языку с вероятностью 0,66, а по математике с вероятностью 0,32.

Достаточная простая зависимость относительного показателя трудности набора тестовых заданий β через основной параметр теста $\beta = 1/\gamma_0$ есть следствие выбранной системы отсчёта измерения. Значения этого показателя для русского языка и для математики соответственно равны 0,60 и 1,64. Что показывают эти значения? Т.к. для оптимального набора заданий $\gamma_0 = 1$ и $\beta = 1$, то набор заданий по русскому языку в $^1/_{0,6} = 1,65$ раз легче оптимального набора. Аналогично, набор заданий по математике в 1,64 раза трудней оптимального для данной группы набора заданий. Из этих оценок следует и другое более выразительное высказывание — уровень трудности набора заданий по математике в $^{1,64}/_{0,60} = 2,7$ раза выше уровня трудности набора заданий по русскому языку.

На наш взгляд, такие различия в экзаменационных заданиях единого аттестационного мероприятия желательно не допускать. Даже если они выявляются по результатам выполнения заданий, то их обязательно следует корректировать процедурой шкалирования. Официальная система шкалирования эти различия явно не фиксирует, и тем более не корректирует.

Обратимся к оценкам испытуемых. В выбранной системе отсчёта альтернативного метода, независимо от экзамена, значение подготовленности усреднённого испытуемого всегда равно единице, или в обозначениях 100-балльной шкалы, значение тестового балла всегда равно 50. Любопытно оценить в этой системе уровень подготовленности такого испытуемого, который набрал первичный балл равный половине максимально возможного значения. Для такого испытуемого имеем следующее индивидуальное значение исходной оценки $X_i = ПБ_i/ПБ_{max} = 0,5$. Тогда рассчитывая относительное значение этого испытуемого по формуле $\theta_i \approx 1/\gamma_0 \log_{0,5} X_i$, имеем результаты: $\theta = 0,60$ — по русскому языку и $\theta = 1,64$ — по математике.

В первом случае испытуемый показал достаточно низкий уровень подготовленности, он в 1,65 раза ниже усреднённого, а во втором случае достаточно хороший результат — в 1,64 раза выше усреднённого испытуемого. Т.е. этот испытуемый очень значимо, опять примерно в 2,7 раза лучше подготовлен к математике, чем к русскому языку.

По-видимому, здесь целесообразно ещё раз прокомментировать смысл высказывания «уровень подготовленности испытуемого А в 2 раза выше уровня подготовленности испытуемого В». Эта фраза означает, что если испытуемый В

правильно выполнит задание с вероятностью P, то испытуемый А с той же совместной вероятностью P правильно выполнит два таких задания.

Вероятностные значения подготовленности рассчитываются по формуле $u_i \approx X_i^{\gamma_0}$ и для рассмотренного выше испытуемого принимают значения 0,32 для русского языка и 0,65 для математики. Стобалльная альтернативная система шкалирования такие результаты оценит значениями тестовых баллов 32 и 65. Различия в 33 тестовых балла очень существенны, они равны третьей части всего диапазона измерений. Действующая система шкалирования такие результаты экзаменов оценивает тестовыми баллами 51 и 65 с разницей 13 баллов. Если учесть погрешности измерений от 11 до 15 баллов, то эти различия минимальные.

Другой пример. Испытуемый по обоим предметам набирает по официальной системе 50 итоговых тестовых баллов. По мнению экспертов, такой испытуемый показывает средний уровень подготовленности к предмету. Во всяком случае, в массовом сознании именно так интерпретируется этот результат. Пересчёт этих значений к результатам альтернативной системы даёт 30 тестовых баллов по русскому языку и 54 балла по математике. Различия существенны.

Интерпретация значений вероятностных показателей, на наш взгляд, ещё более наглядна. Она может иметь процентный вид. Напомним, тестовые баллы ЕГЭ процентной интерпретации не имеют. Правда, она иногда ошибочно применяется.

Опять вернёмся к нашему испытуемому, набравшему первичный балл в половину от максимального. В альтернативной системе шкалирования он оценивается значениями ТБ = 32 балла для русского языка и ТБ = 65 баллов для математики. Приведём смысл этих значений.

Трактовка тестовых баллов альтернативной системы шкалирования проводится относительно оптимального набора экзаменационных заданий, для которых ПБср = 0,5·ПБмакс. Оптимальные наборы экзаменационных заданий имеют одинаковый уровень трудности по любому предмету, равный $\beta = 1$. Тогда значение ТБ = 32 балла по русскому языку означает 32-процентную подготовленность к оптимальному экзамену. Аналогично для математики. Наш испытуемый правильно выполнит около 65% заданий оптимального тестового набора. Если хотите, альтернативная система шкалирования корректирует действующий набор экзаменационных заданий под оптимальный набор и предъявляет оцен-

ку относительно этого оптимального набора заданий.

Заключение

В заключение можно опять привести аналогию с деньгами. Английский фунт стерлингов примерно в 2,7 раза дороже новой турецкой лиры, как и экзамен по математике в 2,7 раза труднее экзамена по русскому языку. Логика существующей системы шкалирования эти различия или не замечает и просто считает количество банкнот у каждого покупателя, в независимости от их национальной принадлежности, или устанавливает курс валют исходя из мнений экспертов местного рынка.

Логика альтернативной системы шкалирования другая. Она сначала устанавливает действующий рыночный курс валют (уровень трудности экзаменационных материалов) и уже исходя из этого — покупательскую способность (уровень подготовленности).