М.Б. Челышкова, А.Г. Шмелев

ШКАЛИРОВАНИЕ РЕЗУЛЬТАТОВ ЕДИНОГО ГОСЭКЗАМЕНА: ПРОБЛЕМЫ И ПЕРСПЕКТИВЫ

Как известно, начиная с 2001 г. в Российской Федерации вводится единый госэкзамен для всех выпускников общеобразовательных учебных заведений (принятая аббревиатура — ЕГЭ). По проблематике ЕГЭ уже выпущено немало научно-методических и научно-организационных материалов. Среди них сборники статей (Болотов, 2002), материалы конференций (Хлебников, 2001–2003). В настоящей статье рассматриваются научно-методические и социально-психологические проблемы обработки и представления результатов ЕГЭ. Проблема шкалирования является одной из центральных в системе методического обеспечения ЕГЭ. Вместе с тем до настоящего времени она решена далеко не так, чтобы удовлетворить все стороны, вовлеченные в процесс проведения и использования результатов ЕГЭ. Прежде всего в данном случае имеются в виду интересы самих учащихся и их родителей, а также самого массового отряда педагогов — школьного учительства.

Достоинства и недостатки нынешнего подхода

В 2001–2003 гг. результаты ЕГЭ обрабатывались в соответствии с моделью шкалирования, разработанной в Центре тестирования Минобразования России (ЦТМО) — организации, ответственной за технологию проведения ЕГЭ (Нейман, 2002). С точки зрения пользователей результатов основные черты этой модели состоят в следующем:

1) В едином методическом центре в Москве из регионов собираются все протоколы экзамена по каждому предмету и для каждого протокола (ответов одного учащегося) подсчитываются так называемые «первичные» (или «сырые») баллы ЕГЭ. Эти баллы отражают число правильных ответов на все задания экзамена с весовыми коэффициентами, разными для заданий разных типов. Известно, что экзамен ЕГЭ состоит из частей A, B и C, где A — задания с выбором из предложенных вариантов, B — задания с кратким свободным от-

ветом, С — задания, с развернутым свободным письменным ответом. Практически для всех предметов принята упрощенная схема весовых коэффициентов: задания А и В дают коэффициент 1, задания С — от 2 до 4. Ответы А и В проверяются автоматизированно — на компьютере. Первичные баллы по заданиям типа «С» выставляют эксперты. При этом эксперты исходят в своих оценках из предлагаемого авторами заданий диапазона оценок — от 0 до 2 (для более легких заданий), от 0 до 4 (для более сложных) и т.п.

2) После подсчета первичных баллов производится пересчет в стандартизированные, так называемые «тестовые баллы», которые измеряются на 100-балльной шкале. При этом применяется однопараметрическая модель Раша в модификации Ю.М. Неймана, которая позволяет сохранить монотонность преобразования первичных баллов в стандартизированные, несмотря на взвешивание эмпирической трудности заданий. Калибровка заданий по трудности при этом происходит весьма приближенно — с точностью не до отдельного конкретного задания в отдельном варианте, а до «типового задания», занимающего определенное место (по теме и типу задания) во всех вариантах. Но уже такая калибровка дает эффект нормализации распределения баллов (в логике известной теоремы Муавра-Лапласа о приближении биномиального распределения к нормальному), хотя нынешний алгоритм шкалирования в ЕГЭ и не реализует сам по себе принцип форсированной нормализации (см. ниже в параграфе, посвященном зарубежному опыту).

3) Для стобалльной шкалы тестовых баллов предметная комиссия Минобразования разрабатывает рекомендации по переводу в пятибалльные отметки. Так как для тестовых шкал, принятых в ЕГЭ 2002–2003 гг., фактически действует модель нормального распределения с параметрами 50+/-15 (то есть, матожидание принималось равным 50, а среднее квадратическое, или стандартное, отклонение — 15), Минобразование получило возможность управлять балансом традиционных оценок (в масштабе страны в целом), рекомендуемых школам для учета в аттестатах. При этом верхняя граница «двойки» в районе 30 баллов отсекает от распределения нижнюю группу численностью примерно в 10 процентов по всем предметам. Симметрично нижняя граница «пятерки» в районе 70 тестовых баллов отсекает верхнюю группу численностью также в 10 процентов. А точка 50 на этой шкале примерно соответствовала медиане и часто утверждалась (предметными комиссиями) как граница между «четверкой» и «тройкой».

Описанный здесь подход явился несомненным шагом вперед в плане создания более удобной, более стандартизированной

и легкой в использовании шкалы, чем шкалы, применявшиеся ранее в Централизованном тестировании (см. Нейман, Хлебников, 2000). В целом данный подход вполне находится в русле мировых научных тенденций в области педагогических измерений. Наиболее ценным следствием этого подхода для практиков явился тот факт, что определенным отрезкам шкалы тестовых баллов фактически поставлены в соответствие определенные вероятности эмпирической встречаемости учащихся с определенным уровнем подготовки. Это облегчает приемным комиссиям вузов планирование приема по результатам ЕГЭ. Вот как примерно выглядит соответствие между определенными точками на шкале тестовых баллов ЕГЭ и процентильными баллами (процентами от выборки испытуемых, выполнявших тест, — см. словарь Балыхина, 2000):

Ниже 30	Ниже 40	Выше 50	Выше 60	Выше 70
Менее 10%	Менее 25%	50%	Менее 25%	Менее 10%

Табл. 1

Однако не все выглядит так просто, как это сформулировано в табл. 1. На самом деле особенности алгоритма шкалирования, применяемого Центром тестирования, таковы, что этот алгоритм дает в случае различных предметов определенные (хотя иногда и малозначительные) отклонения распределения тестовых баллов от того, что мы видим в табл. 1. В силу этого вы не увидите подобной простой таблицы буквально нигде, кроме данной статьи,— ни в печатных материалах Центра тестирования, ни на сайте ЦТМО *www.rustest.ru*. Более того, Ю.М. Нейман нигде не говорит о том, что приводит результаты шкалирования к нормальному распределению с параметрами 50+/-15, ибо в строго математическом смысле это не так.

В этой ситуации участники ЕГЭ (сами учащиеся, их родители, рядовые учителя, неискушенные в математике) жалуются, что применяемая в настоящее время в едином экзамене шкала тестовых баллов им непонятна, то есть не удовлетворяет критерию «прозрачности». Жалобы на это обстоятельство составили едва ли не четвертую часть от общего числа всевозможных жалоб, поступивших в редакцию портала информационной поддержки ЕГЭ в Интернете *http://www.ege.edu.ru* (анализ остальных жалоб выходит за пределы тематики данной статьи). Например, на экзамене по русскому языку в ходе ЕГЭ-2003 десятки учащихся возмущались, почему им «срезали» баллы — заменили после шкалирования их более высокие первичные баллы на более низкие тестовые. Они не понимали ни смысла, ни процедурного механизма подобного пересчета. Оказалось крайне трудным делом объяснять выпускникам средней шко-

лы, которые не проходят ни основы статистики, ни теорию измерений, почему сравнительно легкий тест (каким является русский язык для большинства носителей русского языка как родного) при наличии положительной асимметрии (то есть при сдвиге медианы к высокому полюсу на шкале первичных баллов) дает автоматически более низкие стандартизированные (тестовые) баллы по сравнению с первичными при данной процедуре шкалирования именно для сильных учащихся (для высокой группы).

Другой (и более серьезный!) недостаток применяемой в ЕГЭ 100-балльной шкалы заключается в том, что она вызывает сплошь и рядом ассоциацию со шкалой процентов, в то время как никоим образом не является таковой. Если сертификаты Централизованного тестирования (см. образцы в изданиях Центра тестирования Минобразования России) еще снабжаются так называемыми «рейтинг-баллами», проинтерпретированными на обороте сертификатов в терминах процентов (что дает учащимся и другим пользователям шанс не путать тестовые баллы и проценты), то свидетельства ЕГЭ такой дополнительной информацией не снабжаются (см. образец в сборнике «Единый государственный экзамен. Сборник нормативных документов», 2002 г.). В свидетельствах ЕГЭ можно найти по каждому предмету только один тестовый балл, и нет никаких сведений о том, какое место занял учащийся, располагающий этим баллом, среди всех сдававших ЕГЭ в этом году.

Опросы, проведенные на портале *www.ege.ru*, показали, что не только учащиеся и учителя школ, но и преподаватели вузов слишком часто интерпретируют тестовые баллы ЕГЭ как отражающие процент решенных заданий или процент учащихся, набравших более низкий балл. Многие вузовские приемные комиссии, устанавливая свой «проходной» балл, например, равным 90, ошибочно полагают, что отсекают тем самым либо 90% возможных абитуриентов, либо людей, допустивших более 10% ошибок в ходе ЕГЭ.

Даже нынешний министр образования В.М. Филиппов в пылу полемики по телевидению или радио не раз допускал оговорки, что нижняя граница пятерки — балл 70 — указывает на то, что учащийся решил не менее 70% заданий.

Статистические исследования данных ЕГЭ 2001–2002 гг.

Есть определенные проблемы научного характера, связанные с подходом, выбранным для шкалирования результатов ЕГЭ. В частности, не имеет пока позитивного ответа вопрос о правомерности применения модели Г. Раша к эмпирическим данным при преобразовании сырых баллов выпускников в шкалированные баллы. Сомнения в правомерности порождает анализ характеристик распределения

сырых баллов выпускников, проведенный в рамках НИР 2000–2002 гг. по трем случайно выбранным вариантам четырех предметов (математика, история, русский язык, химия). Полученные по результатам НИР данные говорят о том, что многие характеристики распределения сырых баллов не в полной мере отвечают совокупности необходимых условий, без которых использование модели Раша для шкалирования в теории педагогических измерений считается недопустимым, поскольку может привести к неправильной интерпретации результатов шкалирования.

Среди наиболее важных необходимых условий в рамках НИР анализировались условия адекватности эмпирических данных ЕГЭ требованиям модели Раша, параллельности вариантов, одномерности пространства измерений, обуславливающей возможность представления результатов выпускников на одной шкале по различным частям КИМ, и ряд других не менее важных условий. Выполнение перечисленных условий необходимо для того, чтобы оценка уровня подготовленности испытуемых была независима от трудности заданий теста. В свою очередь, на инвариантности оценок параметра подготовленности строится переход к интервальной шкале баллов, то есть реализация того преимущества, которое и заставляет в основном обратиться к сложной в использовании, но эффективной модели современной теории измерений в образовании. Сложность здесь видится, конечно, не в технической реализации алгоритмов теории, а в выполнении всей совокупности необходимых условий, без которых использование модели Раша при шкалировании теряет всякий смысл, поскольку создается лишь видимость реализации всех преимуществ этой теории. На деле же объективность оценок испытуемых, их сопоставимость и представимость в интервальной шкале, позволяющей в отличие от шкалы сырых баллов интерпретировать разность тестовых баллов по одному или по различным вариантам теста, не достигаются.

Проверка адекватности эмпирических данных требованиям модели измерения осуществляется специальной процедурой, получившей в тестологической литературе название «подгонка данных» («Within population item-fit»). По результатам проверки выбраковываются эмпирические данные, не удовлетворяющие требованиям модели измерения, что приводит к удалению части заданий до шкалирования результатов выпускников по модели Раша. Оставшиеся задания дают основания для построения одномерной шкалы баллов, поскольку являются внутренне согласованными, однородными по содержанию и работают на оценивание одной и той же переменной. Положение осложняет необходимость проведения процедуры под-

гонки не только по заданиям, но и по испытуемым, которых тоже следует удалять из обработки данных в силу несогласованности результатов испытуемых с требованиями модели измерения. Поэтому в мировой практике принято проводить длительную апробацию и коррекцию тестов на репрезентативных выборках учащихся для того, чтобы добиться в ситуации экзамена адекватности эмпирических данных тестирования требованиям модели измерения, сохранив тем самым всю совокупность испытуемых и заданий теста.

Анализ данных по трем случайно выбранным вариантам КИМ по математике, истории, русскому языку и химии показал наличие проблем. В соответствии с критерием отбраковки заданий, предлагаемому теорией и используемому в практике деятельности западных служб тестирования, удалению подлежат от 50% до 70% заданий и не менее 10% результатов выпускников по анализируемым вариантам. Ситуация вполне прогнозируемая, поскольку использование модели Раша при шкалировании предполагает длительную (не менее 2–3 лет) тщательную отработку теста, отсутствующую в условиях ЕГЭ. В этой связи возникают сомнения в том, что проблема шкалирования результатов ЕГЭ на сегодняшний день решена удовлетворительно, по крайней мере анализ адекватности эмпирических данных ЕГЭ требованиям модели Раша говорит о необходимости изменения подхода к шкалированию.

Дополнительным подтверждением последнего утверждения служат результаты углубленного содержательного анализа характеристик КИМ, основанного на обработке эмпирических данных ЕГЭ и последующей интерпретации. Дело в том, что при использовании математических моделей современной теории следует учитывать, что процедура построения шкалы латентных переменных порождает вероятностную версию шкал Guttman, последние попадают в класс моделей, известных как жестко детерминированные. В них предполагается, что задания теста отбираются в порядке нарастания их трудности по определенным, тщательно структурированным элементам содержания дисциплины. При этом считается, что любой испытуемый с правильной структурой знаний, справившийся с данным заданием теста, может наверняка успешно выполнить все предыдущие, более легкие задания.

Шкалирование по Рашу в определенной степени преодолевает трудности построения шкалы Guttman, поскольку является вероятностной версией и отражает вероятностную сущность тестовых процессов. Согласно модели Раша, о правильном выполнении любого задания испытуемым можно говорить лишь с некоторой вероятностью, и прогнозировать успешность можно лишь в том случае,

если эта вероятность близка к единице. Это означает, что каждое задание теста, данные которого обрабатываются с помощью модели Раша, должно иметь высокую биссериальную корреляцию с критерием — общим показателем по тесту — и являться весьма дискриминативным в некоторой точке на континуальной оси измеряемой переменной.

Таким образом, шкалирование по Рашу означает специальный отбор заданий для теста в порядке нарастания трудности из банка данных. В критерии отбора помимо прочих соображений должно входить требование того, что правильное выполнение испытуемым какого-либо задания означает высокую вероятность правильного выполнения предыдущих более легких заданий теста, наоборот, неправильное выполнение задания позволяет прогнозировать с высокой вероятностью неправильное выполнение последующих более трудных заданий теста. Это требование легко применимо к хорошо структурированным дисциплинам, однако по многим предметам, в частности, как показал анализ данных ЕГЭ-2002, по истории, его выполнить достаточно сложно.

В целом можно сказать, что используемая модель измерения должна соответствовать объекту измерения, что для многих предметов в силу слабой структурированности содержания и отсутствия специальной методики отбора заданий в КИМ ЕГЭ не выполняется на практике. Как показал анализ, по многим предметам задания КИМ далеко не всегда коррелируют на соответствующем уровне значимости с общими показателями по измеряемой переменной, что дает основания для заключения о непригодности модели Раша для шкалирования данных ЕГЭ.

Еще одна проблема с правомерностью использования модели Раша для шкалирования результатов ЕГЭ связана с требованием одномерности. Построение шкалы типа Guttman, нарастание трудности заданий и структурированность содержания на основе экспертных оценок не гарантирует одномерность. Возможно, что легкое, среднее и весьма трудное задание образуют шкалу типа Guttman, но каждое из них измеряет что-то свое, поэтому без специальных исследований и проведения факторного анализа непонятно, по какой переменной измерения строится шкала результатов испытуемых. Исследования результатов ЕГЭ, проведенные по данным 2001–2002 гг. (по трем предметам), достаточно однозначно ответили на вопрос о правомерности предположения об одномерности. Анализ характера распределения частот ответов на части КИМ А, В и С и соответствующая обработка данных распределений позволили сделать выводы о том, что они представляют собой самостоятельные совокупности, которые не должны формальным

образом без учета природы данных объединяться в одно множество в процессе шкалирования результатов выпускников.

Таким образом, для ситуации шкалирования результатов выпускников, по крайней мере по трем предметам, характерна многомерность, при которой задания КИМ оценивают выпускника по целому набору переменных, к числу которых могут быть отнесены его знания и умения, его способности к обучению и т.п. Поэтому при оценивании выпускников на концептуальном уровне прежде всего необходимо решить, что именно будет измерять этот набор переменных, как интегрировать оценки по отдельным переменным, какую способность, какой уровень подготовки выпускника можно получить в результате такой интеграции, а затем путем статистического анализа эмпирических данных сделать вывод о том, состоялось ли измерение и позволяет ли такой набор измеряемых переменных сделать вывод о возможностях обучения выпускников в вузах. Поскольку пространство измерений в ЕГЭ многомерно, что не вызывает сомнений, то следует выделить подшкалы и определить, что меряет каждая одномерная подшкала, а затем разработать корректные методы интеграции и интерпретации данных. В целом можно сделать вывод о необходимости дальнейшей работы по выравниванию вариантов, исследованию размерности пространства измерений, созданию репрезентативных выборок, без которых немыслимо проведение подобных работ.

Итак, возникает очевидное противоречие между выбранным подходом к шкалированию результатов ЕГЭ и требованиями теории на фоне явного непонимания подхода со стороны пользователей в силу его расхождения с массовыми представлениями о тестовых шкалах. Какую же стратегию следует выбирать в этой ситуации: следует ли подтягивать массовые представления до уровня научных, добиваясь популяризации азов теории педагогических измерений, или следует упростить процедуру шкалирования, добиваясь ее большей прозрачности для населения?

Зарубежный опыт

Теперь посмотрим, как обстоит в этом отношении дело в западных странах с развитыми традициями использования количественных шкал для оценки образовательных достижений.

Самые популярные шкалы оценки образовательных достижений в западных странах отличаются от нашей традиционной пятибалльной шкалы (на самом деле четырехбалльной) наличием двух совмещенных систем — очковой (scores) и отметочной (grades). Очковая система, как правило, выглядит как 100-балльная шкала, а отметочная задается в простейшем случае путем равномерного деления 100-балльной (Gronlund, Linn, 1990):

0–20	21–40	41–60	61–80	81–100
E	D	C	B	A

Табл. 2

Такой подход позволяет более тонко дифференцировать оценки внутри каждой отметки (категории достижений) — 20 ступенек внутри категории «A», столько же внутри «B» и т.п. Хотя следует отметить, что в чистом виде такой подход скорее применяется лишь для текущего, но не для итогового контроля.

В плане итогового контроля особого внимания заслуживают две разные традиции, представленные американской и британской школами. В США педагогические измерения развивались в XX столетии в большей мере под влиянием психометрики. В США уделялось больше внимания тестам с выбором ответа, так как последние позволяют применять более строгие математико-статистические модели анализа результатов (Standards for educational and psychological tests, 1974). В Великобритании развивалась традиция, придающая большее значение экзаменованию продуктивных умений — способности к выводу теорем, порождению текста, обоснованию ответа и т.п. В этом смысле британская школа ближе к нашей отечественной. Но, как мы увидим ниже, подход к шкалированию результатов ЕГЭ оказался у нас пока ближе к американскому.

Как известно, в США не существует централизованной государственной системы образовательных экзаменов. Но при этом огромной популярностью пользуются тесты, разработанные ETS (Education Testing Service) — фирмой-лидером в данной области. Миллионы американских выпускников школ, желающих поступить в университеты, выполняют тест SAT (Scholastic Aptitude Test, 1998), разработанный ETS и проходящий ежегодное обновление (каждый год появляются новые варианты SAT — подобно тому, как ежегодно обновляются задания ЕГЭ). Результаты теста SAT выражаются на шкале тестовых баллов с параметрами 500+/-100 (аналогичная шкала применяется в более широко известном в России тесте TOEFL, также разработанном фирмой ETS). Применяемая при шкалировании процедура форсированной нормализации (с помощью функции обратного нормального интеграла) дает однозначное соответствие между определенными точками на шкале SAT стандартизированных баллов и процентильными баллами:

Ниже 300	Ниже 400	Выше 500	Выше 600	Выше 700
Менее 3%	Менее 16%	50%	Менее 16%	Менее 3%

Табл. 3

Следует отметить, что тысячебалльная шкала с параметрами 500+/-100 использовалась нами в России в 1997–2001 гг. для фиксации результатов компьютерной олимпиады «Телетестинг» (см. Шмелев, 2000). Выбор подобной шкалы был продиктован использованием в «Телетестинге» (так же, как и в тесте SAT) исключительно заданий с выбором ответа.

Сравнивая табл. 1 и табл. 3, мы можем, казалось бы, констатировать высокую степень сходства двух подходов — российского и американского. Более того, таблица 1 кажется даже более удобной для практического использования. Но тут же стоит зафиксировать 2 существенных различия:

1. Американская шкала SAT оперирует 1000-балльной, а не 100-балльной системой оценок, что исключает риск неправильных ассоциаций с процентами и подталкивает к явному использованию таблицы.

2. Нынешняя шкала российского ЕГЭ в силу специфики алгоритма шкалирования не всегда подчиняется закономерности, описанной в табл. 1 (об этом мы уже писали выше).

Теперь рассмотрим, как обстоит дело со шкалированием в Великобритании — в стране, где первые специализированные организации, занимающиеся разработкой экзаменационных технологий, созданы уже полтора века назад (например, Экзаменационный синдикат в Кембридже). Выпускники основной школы сдают экзамен GCSE — на «общий сертификат о среднем образовании». Эта система экспортируется в десятки стран мира, причем не только те, которые входят в Британское содружество наций (Cambridge International Examination, 2000). По каждому предмету экзамен состоит из частей, в которых собраны задания определенного типа: на выбор ответа, с кратким ответом, структурированные вопросы, с развернутым ответом, эссе, практические работы. На каждую часть экзамена отводится определенное время. За каждую часть присваивается определенное количество очков в процентах к общему баллу. В некоторых случаях балл GCSE набирается по принципу «портфолио» («портфель достижений»), так как включает накопление очков за выполнение практических работ, за получение определенных оценок в школе и т.п. Практически по каждому предмету существует 2 версии экзаменов по уровням: «ядерный» (core) и «расширенный» (extended), что у нас чаще обозначается в терминах курсов или учебных программ — «базовый» и «углубленный». Расширенный вариант, как правило, включает ядерный как подмножество. Достижения по выполнению «ядерного» варианта фиксируются на 100-балльной шкале очков, а по выполнению «расширенного»

варианта — на 200-балльной шкале (хотя и не всегда). Причем считается, что учащийся, который сдает экзамен на расширенном уровне, набирает больше 100 процентов очков (!). Затем 200-балльная шкала GCSE по каждому предмету разбивается на 8 градаций (grades) по уровням достижений: A+, A, B, C, D, E, F, G. Учащиеся не обязаны выполнять экзамен в «расширенном» варианте, но это ограничивает их достижения. Выполнение экзамена на «ядерном» уровне дает возможность получить градации (отметки) не выше C. Для получения более высоких оценок следует выполнять «расширенный» вариант экзамена.

Вот как выглядит, например, таблица начисления баллов за экзамен по естествознанию (physical science):

Номер компонента	Название Компонента	Длительность (в мин.)	Вес в общем балле (в %)
1	Выбор ответа	45	40
2	С кратким ответом	60	40
3	Структурированные вопросы и со свободным ответом	75	80
4	Школьная оценка практических навыков	—	20
5	Тест практических навыков	90	20
6	Альтернативное задание практическому тесту	60	20

Комбинация компонент в зависимости от курса (curriculum):

Курс	Компоненты		
Ядерный (базовый)	1 2 4	1 2 5	1 2 6
Расширенный (профильный)	1 2 3 4	1 2 3 5	1 2 3 6

Таким образом, мы видим, что компоненты 4, 5 и 6 оказываются альтернативными, то есть 20 очков можно набрать либо путем засчитывания школьного балла, либо путем выполнения практической работы в момент сдачи экзамена. За расширенный курс ставится до 180% очков. При этом вес части со свободным ответом достигает почти 45% очков от возможного максимума за расширенный курс естественных наук.

Конечно, в таком виде британская система шкалирования (точнее баллирования) выглядит более запутанной и менее логически цельной, менее формализованной, чем американская. Но она име-

ет ряд достоинств, которые следует учесть именно в контексте нашей уже складывающейся отечественной модели ЕГЭ:

1) Учащимся легко ориентироваться в том, как складывается их итоговая оценка по частям экзамена, ибо вес каждой части заранее объявлен.

2) Внутри отдельных частей экзамена сохраняется возможность применения более сложных математических процедур шкалирования (предполагающих калибровку заданий по статистике ответов и т.п.).

3) Введение весовых (долевых) отношений между частями экзамена гарантирует высокий вес неформализованных заданий со свободным ответом (подобно тому, как в экономическом планировании защищается определенная статья бюджета).

4) Учащийся свободен в выборе варианта экзамена (базового или расширенного) и заранее знает, во сколько очков обернется для него тот или иной выбор.

5) Сохраняется универсальность ранговых шкал учебных достижений (отметок), так как шкалы отдельных вариантов экзамена оказываются вложенными (базовая шкала вложена в расширенную).

6) Использование расширенной шкалы, на которой можно набрать больше 100% очков, создает благоприятную психологическую атмосферу, защищающую самооценку учащихся, которые в состоянии освоить только лишь базовую программу, но набирают при этом все-таки почти 100% очков.

Как видим, по ряду признаков наш ЕГЭ ближе к британской системе экзаменов. В частности, по наличию в каждом экзамене части «С» — заданий со свободным развернутым ответом. Но применяется в нашем ЕГЭ не британская, а американская система шкалирования. В результате учащиеся у нас понимают, как подсчитываются первичные баллы, сколько очков весит каждое задание типа «С» и сколько весит эта часть в общей сумме первичного балла, но… никто не понимает, сколько эта часть весит в итоговом тестовом балле. И эта не единственный недостаток действующей системы шкалирования результатов нашего единого экзамена.

Наличие в британской системе расширенной шкалы до 200% наводит на мысль о том, что и в России — с ее резкими контрастами в уровне образовательной подготовки учащихся из городов и сел, из разных регионов — было бы гораздо гуманнее применять не шкалу со средним значением в районе 50 (что многими до сих пор интрерпретируется как исключительно низкий процент усвоения школьного материала — «больше половины»), а шкалу, которая обеспечивала бы социально-психологическую защищенность

выпускников школ, не получивших в своих школах (не по своей вине) высокого качества образовательных услуг. Возможно, что применение такой более гуманной системы шкалирования снизило бы в определенной степени массовое недовольство населения внедрением ЕГЭ.

Субъективные шкалы

Чтобы создать шкалу, понятную для населения, следует изучить, какие представления об оценивании образовательных достижений стихийно сложились в головах у массы педагогов и самих учащихся в Российской Федерациии. Для этого авторы данной статьи провели в течение ряда последних лет несколько различных эмпирических исследований, опрашивая школьников, студентов, педагогов, методистов, участвовавших в различных конференциях. Не утомляя здесь читателя подробностями данных эмпирических исследований, позволим себе лишь суммировать основные результаты.

Существующая пятибалльная система оценивания явно испытывает в России определенную девальвацию. Она выразилась в частности в том, что изначальный смысл, который приписывается градациям пятибалльной шкалы в самих известных названиях («отлично, «хорошо», «удовлетворительно»), уже давно фактически трансформировался. Оценка «отлично» вовсе не воспринимается подавляющим большинством как полное освоение всего материала плюс освоение дополнительного материала. Оценка «отлично» рассматривается лишь как относительная категория, указывающая на более высокую степень превосходства над средним уровнем, чем это достигается в случае оценки «хорошо». Вот какие смысловые градация имеют наши привычные оценки фактически:

Двойка	Тройка	Четверка	Пятерка
Явно ниже среднего уровня	Несколько ниже среднего уровня	Несколько выше среднего уровня	Явно выше среднего уровня

Табл. 4

Как видим, такая трактовка смысла оценок находится в полном соответствии с тестологическим подходом, описанном в табл. 1 и 3.

А как все-таки соотносятся наши традиционные оценки с охватом учебного материала? Для выяснения этого мы опрашивали сотни респондентов о том, какой процент материала усваивают, с их точки зрения, «отличники», какой усваивают «хорошисты», какой «троечники». Выяснилось, что в сознании педагогов-методистов чаще присутствует следующая шкала (см. табл. 5).

Можно уверенно сказать, что статистика ЕГЭ опровергла данные идеализированные представления педагогов-методистов.

Двойка	Тройка	Четверка	Пятерка
Менее 50% материала	От 50 до 70%	От 70 до 90%	Не менее, чем 90%

Табл. 5

По многим предметам «отличники» явно недотягивали до первичных баллов в 90 процентов от максимума.

А в сознании самих учащихся эта же шкала чаще всего выглядит более реалистичной:

Двойка	Тройка	Четверка	Пятерка
Менее 40% материала	От 40 до 60%	От 60 до 80%	Не менее, чем 80%

Табл. 6

Легко видеть, что наши учащиеся уже приблизились к международному стандарту, представленному в табл. 2.

Совершенно очевидно, что оптимальная шкала ЕГЭ должна больше соответствовать распространенным субъективным представлениям. Без этого единому экзамену очень трудно завоевать реальную популярность у населения.

Принципы

Было бы неправильно ограничиться в данной статье лишь критикой действующего подхода и аналитическими аргументами. Авторы считают своим долгом сформулировать конструктивные предложения по модификации шкалы единого госэкзамена. При этом еще раз изложим принципы, на которых должна быть построена такая шкала:

1) Принцип прозрачности. Шкалирование результатов ЕГЭ должно быть понятным для самых широких масс участников ЕГЭ и его организаторов. Люди должны видеть и понимать, как в самой процедуре шкалирования реализуется принцип объективности в оценивании образовательных достижений и справедливости конкурсного отбора в вузы.

2) Принцип гуманизма. Шкала ЕГЭ должна обеспечивать социально-психологическую защищенность выпускников школ, которые не ставят своей целью дальнейшее изучение данного предмета в вузе.

3) Принцип единства. Шкала ЕГЭ должна быть универсальной и оценивать в различных количественных показателях достижения учащихся, прошедших как программу базовой школы, так и программу профильной (специализированной) школы. Было бы неправильно оценивать одним и тем же числом достижения по математике выпускников гуманитарных гимназий и физико-математических лицеев.

4) Принцип научности. Шкалирование результатов ЕГЭ должно базироваться на достижениях математизированной теории педагогических измерений. Не следует отказываться от возможности

калибровать отдельные задания на основе реальных статистических данных массового экзамена, тем более что в ходе самого ЕГЭ мы получаем весьма репрезентативные результаты.

5) Принцип программирующего воздействия. Шкалирование результатов ЕГЭ не должно развиваться в отрыве от конструирования контрольно-измерительных материалов (КИМ). Напротив, от веса, придаваемого определенным частям экзамена, должна зависеть значимость каждой части в контексте КИМов в целом. Более того, следует, по-видимому, открыто признать, что взвешивание частей экзамена — это задача, имеющая самое прямое отношение к государственной образовательной политике. То или иное решение этой задачи имеет весьма серьезные последствия, а именно — сказывается на приоритетах в самом учебном процессе, на том, какое внимание будут уделять педагоги формированию тех или иных учебных умений.

6) Принцип соответствия интересам конкурсного отбора в вузы. Результаты ЕГЭ должны информировать приемные комиссии вузов о том, какое место учащийся занял в общероссийском рейтинг-листе образовательных достижений.

7) Принцип плавного отказа от привычных представлений. Привычная для российской школы «пятибалльная оценочная система» должна отмирать постепенно — путем сосуществования в рамках ЕГЭ с более дробными шкалами, а также с расширенными шкалами, соответствующими интересам профильных школ и вузов.

8) Принцип поддержки ГИФО. Шкала ЕГЭ должна логичным образом увязываться с возможным выделением государственных именных финансовых обязательств. (Хотя авторы статьи и не считают нынешнюю технологию ЕГЭ достаточно защищенной для того, чтобы выдержать риск начисления крупных сумм денег по результатам ЕГЭ).

Перспективная модель

На основании вышеизложенных принципов, конечно, возможно выдвижение разных моделей, но мы позволили бы предложить следующий подход:

1) 100-балльная шкала ЕГЭ сохраняется только для базовой версии ЕГЭ, которая включает задания с выбором ответа (от 30 до 60 очков по разным предметам), с кратким ответом (от 10 до 40 очков по разным предметам), а также школьный балл за итоговую контрольную работу (от 10 до 30 очков по разным предметам).

2) Стобалльная шкала ЕГЭ использует только первичные баллы и интерпретируется очень просто — в терминах процента правильно решенных заданий.

3) Для 100-балльной шкалы ЕГЭ заранее объявляются границы отметок с использованием традиционной «пятибалльной шка-

лы»: «пятерка» — от 81 до 100, «четверка» — от 61 до 80, «тройка» — от 41 до 60.

Все три изложенные выше правила призваны обеспечить следующие принципы: прозрачности (принцип 1), гуманизма (2) и «плавного отказа» (7)

4) К стобалльной шкале ЕГЭ добавляется «расширенная шкала ЕГЭ», которая включает от 100 до 200 очков (по разным предметам). Дополнительные 50 или 100 первичных очков учащиеся получают за решение заданий типа «С» (с развернутым ответом) или (и) за выполнение практических работ (например, за прохождение устной беседы по иностранному языку). Эти дополнительные задания учитывают программу «профильной школы».

Для соблюдения принципа единства баллы по расширенной шкале должны отображаться не в виде обычных «пятерок» и «четверок», а в виде баллов, превышающих «пятерку». Простейший вариант отображения таков:

Шестерка	Семерка	Восьмерка	Девятка	Десятка
От 101 до 120	От 121 до 140	От 141 до 160	От 161 до 180	Не менее, чем 181

Табл. 7

При этом важно подчеркнуть, что в итоговый балл за расширенный экзамен входит оценка за базовую часть. Например, учащийся Петров набрал за дополнительное задание 82 очка, но за базовую часть только 70 очков. Тогда его итоговый результат будет равен не 182 очкам, а только 152 очкам («восьмерка», а не «десятка»).

Правила 4 и 5 обеспечивают выполнение принципов «единства» (принцип 3), «программирующего воздействия» (5), «конкурсного отбора» (6).

Ну а как же обеспечить в таком случае выполнение принципа научности? Для этого мы считаем необходимым указывать в свидетельстве по итогам ЕГЭ как минимум 2 разных результата: первичный балл (в виде процента от максимального балла) и тестовый балл. Второй следует получать с помощью применения современной процедуры шкалирования, позволяющий учесть различный вклад заданий по данным объективной статистики. Но было бы гораздо лучше выражать тестовый балл не на 100-балльной шкале, а на 1000-балльной. Чтобы никто не путал этот балл с процентами.

На переходный период (пока внедрение ЕГЭ не завершилось) не следует избегать наличия и сосуществования по результатам ЕГЭ сразу нескольких оценок, в которых используются различные принципы шкалирования. Ведь новое познается с опорой на старое.

Хватило же ума нашему правительству после 1000-кратной деноминации рубля в Росии в 1998 г. не сразу изымать из обращения купюры с огромным количеством нулей. Точно так же в англоязычных странах переходили к десятичной системе мер длины — указывали рядом с сантиметрами старые футы и дюймы. Этот подход, учитывающий множественность шкал, является психологически щадящим. Если люди имеют рядом с новой и непонятной мерой старую и понятную, то они легче привыкают к новой — с опорой на старую.

Было бы еще более разумным шагом не ограничиваться в свидетельствах о результатах ЕГЭ двумя числами (первичным и тестовым баллом), но указать и третье число — «рейтинг-балл», определяющий место, который занял данный экзаменуемый среди всех учащихся в России в текущем году. Ведь подобная традиция фиксации результатов фактически принята во всех видах спорта: рядом с результатам в секундах (метрах, килограммах и т.п.) указывается место, которое занял спортсмен с этим результатам в состязании.

Множественность оценок в свидетельствах ЕГЭ не повысит, а наоборот снизит неопределенность в головах пользователей результатов. Это доказал опыт и Централизованного тестирования (как уже говорилось выше, в сертификатах ЦТ указываются два показателя), и «Телетестинга» (в сертификатах «Телетестинга» давались 4 показателя — кроме тестового и рейтингового балла, давались первичные баллы и рекомендуемые традиционные отметки).

Использование трех (или даже четырех!) различных показателей в свидетельствах ЕГЭ позволит учесть принцип 8 — «поддержка ГИФО». Дело в том, что ГИФО никак нельзя привязывать к первичным баллам и отметкам (столь понятным и простым для населения), но можно привязывать лишь к тестовым и рейтинг-баллам. Это обеспечит защищенность бюджета государства от незапланированного перерасхода средств на образовательные нужды. А главное — это повысит защищенность ЕГЭ от соблазна местных организаторов «повысить всем своим сразу».

Предложенный нами подход несомненно улучшил бы понимание результатов ЕГЭ самыми широкими категориями пользователей — от учащихся, родителей, школьных учителей до работников приемных комиссий вузов. Хотя вполне возможно, что данный подход еще встретит весьма горячие контраргументы со стороны специалистов, обожающих споры в форме высказываний типа «А версты все же гораздо удобнее, чем километры».

Помимо выбора оптимального подхода к шкалированию, для повышения объективности и сопоставимости шкалированных баллов выпускников необходимо также проведение ряда неотлож-

ных работ, направленных на повышение качества КИМ, процедур применения, обработки данных и интепретации результатов выполнения тестов, поскольку возможность оптимизации процедур шкалирования и выравнивания находится в прямой зависимости от качества выборки, используемой для построения шкалы и от качества КИМ. Здесь логика очень проста: если не обеспечено должное качество самих КИМ, выполнение условий параллельности вариантов, требований к характеристикам заданий и адекватность характеристик распределения эмпирических данных ЕГЭ требованиям моделей измерения, то нет и не может быть корректных процедур шкалирования и выравнивания.

В этой связи, поскольку значительное число вариантов КИМ по результатам анализа является статистически значимо различающимися, то необходимо наибольшие усилия сосредоточить на работе по улучшению параллельности вариантов. Следует повысить качество проведения апробации заданий и вариантов, причем выполнять подгонку эмпирических данных под требования используемых моделей измерения при помощи коррекции статистических свойств КИМ еще на стадии апробации.

Основываясь на международном опыте (для корректного шкалирования 180 000 результатов испытуемых необходимо не менее трех недель и значительная предварительная работа), для шкалирования результатов по обязательным экзаменам нужно пойти по принятому в зарубежных тестовых службах пути, обеспечивающему максимально возможную корректность результатов испытуемых при массовом тестировании и максимальную эффективность процесса шкалирования. Предлагаемый подход основан на создании априорной (до начала тестирования) репрезентативной выборки, на которой затем строится шкала в процессе массового тестирования. Построение шкалы на небольшой, но репрезентативной выборке при больших объемах тестирования позволяет значительно сократить время обработки всего массива данных, поскольку результаты остальных учащихся просто отображаются на готовую шкалу.

Литература

1. Балыхина Т.М. Словарь терминов и понятий тестологии. М: МГУП, 2000.
2. Единый государственный экзамен. Сборник статей. / Под ред. В.А. Болотова. М.: Логос, 2002.
3. Единый государственный экзамен. Сборник нормативных документов. М.: Минобразования РФ, 2002.
4. Нейман Ю.М. Шкалирование результатов единого госэкзамена. М: ЦТМО, 2002.
5. Нейман Ю.М. Хлебников В.А. Введение в теорию моделирования и параметризации педагогических тестов. М.: Прометей, 2000.
6. Рэш Дж. Индивидуальный подход к анализу вопросов // Математические методы в социальных науках. М.: Прогресс, 1973. С. 91–116.
7. Развитие системы тестирования в России. / Под ред. В.А. Хлебникова. Материалы ежегодной Всероссийской конференции. М: ЦТМО, 2001–2003.
8. Челышкова М.Б. Разработка педагогических тестов на основе современных математических моделей. М: МИСИС, 1995.
9. Тесты для старшеклассников и абитуриентов. Телетестинг. / Под ред. А.Г. Шмелева. М.: Первое сентября, 2000.
10. Cambridge International Examination. Cambridge (UK): Local Examinations Syndicate, 2000.
11. Gronlund N.E., Linn R.L. Measurement and Evaluation in Teaching. 6th edition. N.Y.-L.: Macmillan, 1990.
12. Manual for Scholastic Aptitude Test. Prinston (N.J.): ETS, 1998.
13. Standards for educational and psychological testing. Washington: American Psychological Association, 1974.