Методика анализа качества проверки заданий с развёрнутым ответом ЕГЭ по физике

Гиголо Антон Иосифович

кандидат технических наук, ФГБНУ ФИПИ, член федеральной комиссии по разработке КИМ для ГИА по физике, kim@fipi.ru

Ключевые слова: ЕГЭ по физике, задания с развёрнутым ответом, экспертная проверка, обобщённая система оценивания, анализ третьей проверки, рейтинг качества проверки.

В действующей модели КИМ ЕГЭ по физике предлагается пять заданий с развёрнутым ответом, которые оцениваются двумя экспертами с учётом правильности и полноты ответа. В экзаменационных материалах по физике используется две обобщённых системы оценивания заданий с развёрнутым ответом: для оценивания качественных задач и для оценивания расчётных задач. В экзаменационном варианте перед каждым типом задания предлагается инструкция, в которой приведены общие требования к оформлению ответов. В критериях оценивания экзаменационного варианта к каждому заданию приводится подробная инструкция для экспертов, в которой указывается, за что выставляется каждый балл — от нуля до максимального балла.

Максимальный первичный балл за любое из заданий с развёрнутым ответом составляет 3 балла. При проверке экзаменационных работ несущественным расхождением в баллах, выставленных двумя экспертами, считается расхождение в 1 балл. Расхождение в 2 и более балла оценки за выполнение любого задания считается существенным и приводит к необходимости третьей независимой проверки.

Работа московской предметной комиссии по физике организуется в соответствии с рекомендациями ФГБНУ ФИПИ. Для более качественной проверки заданий с развёрнутым ответом перед началом работы группа ведущих и старших экспертов во главе с председателем организует детальный анализ экзаменационных заданий и критериев оценки. В процессе обсуждения выявляются возможные варианты альтернативных решений, отличающихся от авторского подхода, типичные и характерные возможные ошибки в решениях учащихся и т.п. По результатам обсуждения составляются дополнительные рекомендации по оцениванию.

В процессе проверки экзаменационных работ в каждой аудитории работает консультант из числа старших или ведущих экспертов, который помогает решить сложные и неоднозначные случаи, возникающие в процессе работы эксперта. Председатель предметной комиссии, если возникает такая необходимость, дополнительно проводит оперативное согласование между аудиториями, для того, чтобы обеспечить согласованность работы всей предметной комиссии. Однако, несмотря на все вышеперечисленные меры, при проверке возникают случаи существенных расхождений, и экзаменационная работа попадает на третью проверку.

Региональный центр обработки информации (РЦОИ) выдаёт председателю предметной комиссии только общую статистику по проверке экзаменационных работ, в которой указаны общее количество проверенных работ, количество про-

веренных заданий как по каждому эксперту, так и по работе всей комиссии в целом. Особенно важной информацией является количество и процент работ, попавших на третью проверку.

Пары экспертов, проверяющих ту или иную работу, создаются системой РЦОИ случайно, а значит, без проведения детального анализа невозможно выявить причины появления существенных расхождений и определить эксперта, оценка которого привела к этому расхождению. В общей статистической информации РЦОИ приводится для каждого эксперта только процент работ, попавших на третью проверку, что, к сожалению, не даёт полной объективной картины: какой из двух экспертов, проверяющих работу, совершил ошибку.

Если же при оценивании работ наблюдаются расхождения в 1 балл по отдельному заданию, то в подобных случаях оценка округляется в пользу учащегося, и при этом считается, что расхождения несущественные. Однако следует иметь в виду, что в самом худшем случае при проверке пяти заданий с развёрнутым ответом могут возникать расхождения до 5 баллов, как в «плюс», так и в «минус», и подобные расхождения максимальных баллов двух экспертов в процессе проверки не определяются и не корректируются. Конечно, подобные случаи встречаются достаточно редко: по статистике в 2016 г. в среднем процент подобных ситуаций составил 0,03% расхождения на одну работу в 5 баллов и 0,29% — в 4 балла.

Из приведённого выше следует, что для более качественной и согласованной проверки, а значит и для уменьшения случаев существенных расхождений, возникающих при проверке, требуется более детальный анализ статистических данных, имеющихся в распоряжении председателя предметной комиссии. Именно эту проблему мы и решали в процессе анализа статистических данных.

Выделим *две основные задачи*, решение которых необходимо получить на основе анализа статистических данных:

1. Детальное исследование *только случаев третьей проверки*, выявлений причин, повлекших за собой третью проверку, и определение эксперта, оценки которого стали причиной этого.

2. Исследование *всех результатов проверки* на предмет выявления несущественных расхождений, определение согласованности проверки.

Решение этих двух задач позволит составить рейтинг качества проверки всех экспертов предметной комиссии.

Исходный статистический массив данных формируется РЦОИ и содержит идентификационный номер работы, информацию о эксперте, проверившем эту работу, номер варианта, первичные баллы за задания с развёрнутым ответом и всю работу в целом, тестовый балл за всю работу, статус проверки (первая или третья) и номер протокола, в котором были выставлены оценки эксперта, а главное, оценки по каждой экзаменационной работе, проверенной двумя экспертами. Если же в этих оценках присутствуют существенные расхождения, то в данных присутствует строка, соответствующая оценкам третьего эксперта.

Подобный статистический массив для работ, проверенных московской предметной комиссией, содержит более 20 000 строк, поэтому анализ этой информации в ручном режиме не представляется возможным. Для анализа результатов работы предметной комиссии была разработана и написана программа статистической обработки результатов, по результатам работы которой были выявлены и сгруппированы все причины существенных расхождений, проанализированы несущественные расхождения и составлен рейтинг качества проверки экспертов.

При таком количестве исходной информации анализировать отдельные случаи существенных расхождений, запрашивая и перепроверяя каждую «проблемную» работу отдельно, крайне затруднительно. Поэтому при разработке программы авторы руководствовались принципом: «*Финальная оценка третьего эксперта — эталон, с которым производится сравнение оценок первого и второго эксперта в спорных случаях*».

Также в работе программы решаются дополнительные задачи:

– определение причин появления третьей проверки;

– выявление характера расхождения;

– учёт и суммирование всех ошибок каждого эксперта;

– расчёт процента относительных ошибок по каждому эксперту;

– расчёт весового «коэффициента ошибок».

развёрнутым

Таблица 1

	28	29	30	31	32
Эксперт 1	3	2	3	X	0
Эксперт 2	3	1	2	0	X
Эксперт 3				X	0

Таблица 2

	28	29	30	31	32
Эксперт 1	1	X	3	X	0
Эксперт 2	2	2	2	X	1
Эксперт 3		2			

Таблица 3

	28	29	30	31	32
Эксперт 1	1	2	3	0	X
Эксперт 2	3	1	2	0	X
Эксперт 3	3				

Все выявленные ошибки в работе эксперта структурируются, ранжируются и суммируются, в результате чего формируется рейтинг качества проверки.

Для иллюстрации принципов, заложенных в программу статистической обработки, подробно остановимся на *типичных случаях возникновения существенных расхождений*, приводящих к третьей проверке.

1. Техническая ошибка

При работе экспертов возникают случаи технических ошибок, в результате которых один эксперт ошибочно заносит в протокол баллы не соответствующие номеру оценённого задания. В таблице 1 приведён пример фрагмента статистического массива в случае возникновения третьей проверки.

Видно, что Эксперт 2 совершил техническую ошибку и переставил местами оценки за задания № 31 и № 32. Таким образом, Эксперт 1 в данной спорной ситуации ошибок не совершил, а *Эксперт 2 совершил две технические ошибки*.

В таблице 2 приведён ещё один случай технической ошибки, Эксперт 1 не нашёл в работе учащегося задание № 29.

В этом случае *Эксперт 1 совершает одну техническую ошибку*. Второй эксперт оценил работу верно.

2. Существенные расхождения (критическая и грубая ошибка)

К существенным относятся расхождениям на 2 или 3 балла. Однако в каждом конкретном случае третьей проверки ситуация может быть совершенно различной. В таблице 3 приведён случай расхождения на 2 балла в задании № 28.

Видно, что третий эксперт согласился с оценкой Эксперта 2, поэтому *Эксперт 1* занижает оценку на 2 балла и *совершает грубую ошибку «–2БАЛЛА»*.

В табл. 4 также приведён случай расхождения на 2 балла в задании № 28, однако ситуация совершенно иная.

Третий эксперт в этом случае ставит промежуточную оценку 1 балл. Таким образом, *Эксперт 1* занижает на 1 балл и *совершает ошибку «–1 БАЛЛ»*, а *Эксперт 2* завышает, *совершая ошибку «+1 БАЛЛ»*.

Таким образом, все работы, попавшие на третью проверку, анализируются, выявляются причины существенных расхождений, и фор-

Таблица 4

	28	29	30	31	32
Эксперт 1	0	2	3	0	X
Эксперт 2	2	1	2	0	X
Эксперт 3	1				

мируется личная статистика каждого эксперта по всем возможным допущенным ошибкам:

- *критическим ошибкам «±3 БАЛЛА»,*
- *грубым ошибкам «±2 БАЛЛА»,*
- *ошибкам «±1 БАЛЛ»,*
- *«техническим ошибкам»*.

Для того, чтобы составить рейтинг качества проверки всех экспертов, необходимо учесть не только количество проверенных каждым экспертом работ, но *число пустых заданий* (оценка X в протоколе) во всех проверенных работах.

Несмотря на то, что количество проверенных работ, а также страниц экзаменационных работ, у каждого эксперта примерно одинаково, распределение по пустым заданиям существенно отличается, что, несомненно, сказывается и на качестве проверки. На рисунке 1 приведён график распределения проверенных и пустых заданий по каждому эксперту Московской предметной комиссии.

Из представленного графика видно, что количество проверенных и пустых заданий сильно разнится у разных экспертов. В самом сложном положении оказался эксперт, в работах которого 23% пустых заданий, и 77% заданий, требующих оценивания. При этом в более простой ситуации оказался тот эксперт, у которого проверять требовалось всего 45 % заданий, а пустых оказалось 55%. В среднем число проверенных заданий составило 66%, а пустых 34%. Значит, при составлении рейтинга качества проверки экспертов следует учитывать количество реально проверенных заданий, а не проверенных работ.

После суммирования всех ошибок различного типа и усреднения по количеству проверенных работ или проверенных «непустых» заданий каждый эксперт получает оценку качества проверки — «***Процент ошибок***». Среднее значение допущенных ошибок проверки, усреднённое по количеству работ по предметной комиссии, в 2016 г. составило 7,26%. Если учесть, что в каждой работе 5 заданий, при этом количество «непустых» заданий составило 66%, то можно усреднить по каждому реально проверенному заданию.

Всего в 2016 г. во время первой и второй проверок было проверено 18 418 работ, из которых на третью проверку попало 1 740

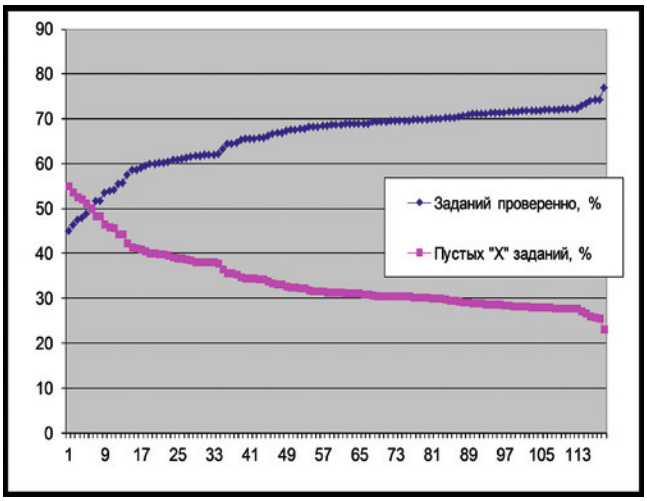

Рис. 1. Распределение количества проверенных и пустых заданий между членами предметной комиссии в 2016 году

Таблица 5

	–3	–2	–1	Тех	+1	+2	+3
Средний % ошибок	2,4%	18,8%	23,3%	7,3%	24,1%	20,4%	3,6%
Средний % ошибок на 1 работу	0,18%	1,43%	1,78%	0,56%	1,84%	1,56%	0,28%
Средний % ошибок на 1 проверенное задание	0,06%	0,43%	0,54%	0,17%	0,56%	0,47%	0,08%
Среднее количество ошибок на одного эксперта	0,3	2,2	2,8	0,9	2,8	2,4	0,4

(9,45%) работ, в которых было допущено 1 405 ошибок различного типа. Таким образом, если детально проанализировать все возникшие при проверке ошибки, то в среднем на одну проверенную работу пришлось 7,63% ошибок, а на одно проверенное «непустое» задание 2,31%.

В таблице 5 проанализированы и приведены средние значения по различным типам допущенных ошибок.

Весовой «коэффициент ошибок»

Анализируя различные типы ошибок, приводящих к появлению третьей проверки, приходишь к выводу, что они имеют разную степень значимости. Критические или грубые ошибки в случае оценки абсолютно правильного решения или решения, содержащего незначительные недостатки, в 0 баллов или высокие баллы за неверное или содержащее серьёзные физические ошибки решение должны рассматриваться и учитываться более серьёзно, чем расхождение с третьим экспертом в 1 балл. Ошибки технического характера также являются существенными, поскольку путаница в протоколе с оценками за то или иное задание или ненайденное, а потому и не оценённое задание, всегда приводит к появлению третьей проверки. Поэтому была предложена и разработана процедура весового суммирования допущенных ошибок.

В таблице 6 приведены весовые коэффициенты ошибок различного типа и их «допустимое» количество на 100 проверенных экспертом работ. Так, например, вес критической ошибки «±3 балла» максимален и равен 2, техническая ошибка имеет вес 1,5, а ошибка «±1 балл» суммируется с весом 0,7.

Пороговое значение *весового «Коэффициента ошибок»* в случае, если допущены 1 критическая и 2 грубые ошибки, 3 технические и 10 ошибок, составило 16,9. Таким образом, чем меньше значение коэффициента ошибок, тем лучше и качественнее выполнил свою работу эксперт, а значит, его оценки были согласованы лучше.

Анализируя полученные результаты, можно сделать вывод о том, что большинство экспертов предметной комиссии успешно справляются со своей работой. Примерно 20% имеют значения коэффициента ошибок меньше 5, а значит, совершают за время своей работы всего 1–2 грубые ошибки или порядка 5 несущественных расхождений с третьим экспертом. Работа этих экспертов является некоторым эталоном для всех. Почти половина предметной комиссии (45,4%) получили коэффициент ошибок от 5 до 10. Значение коэффициента ошибок ещё примерно у 30% экспертов лежит в допустимом диапазоне и не превышает его порогового значения. Этим экспертам следует более внимательно отнестись к своей работе, чтобы улучшить свой результат в будущем. Пороговое значение коэффициента ошибок было превышено только у 5% экспертов, работа кото-

Таблица 6

Ошибки	±3 балла	±2 балла	Технические	1 балл
Весовой коэффициент	2	1,7	1,5	0,7
«Допустимое» количество ошибок на 100 работ	1	2	3	10

Таблица 7

	28	29	30	31	32
Эксперт 1	2	2	3	2	3
Эксперт 2	3	1	2	2	3

рых в предметной комиссии может быть признана неудовлетворительной.

«Коэффициент завышения/занижения»

Отдельной задачей стоял анализ выставленных оценок в работах, которые не попали на третью проверку, т.е. в них не было выявлено технических ошибок или существенных расхождений. В таблице 7 приведён пример фрагмента статистического массива в случае незначительных расхождений.

Видно, что Эксперт 1 в заданиях № 28 и № 29 ставит на 1 балл больше, чем Эксперт 2, а в задании № 28 наоборот. В заданиях № 31 и № 32 эксперты солидарны и их оценки совпадают.

Проанализировав весь массив оценок для каждого эксперта, можно установить общее количество «завышений» или «занижений» оценок. Конечно, без «эталонной» проверки достоверно определить, кто из пары экспертов поставил в данном случае более правильную оценку, невозможно. Однако на достаточно большом количестве проверенных заданий (в среднем 550 заданий на каждого эксперта) можно увидеть некоторую корреляцию и сделать вывод, пусть и оценочный, о систематическом занижении или завышении. Таким образом, для каждого эксперта рассчитывается «Коэффициент завышения/занижения»:

$$K_{+/-} = \frac{\sum N_{+} - \sum N_{-}}{N_{\text{Общ Заданий}}} \cdot 100\%.$$

На рисунке 3 приведён график распределения значения «Коэффициента завышения/занижения» для экспертов Московской предметной комиссии в 2016 г.

Из графика видно, что количество «завышающих» и «занижающих» экспертов практически одинаково (60 на 59), максимальный процент завышения 28%, а процент занижения — 21%, при этом в среднем эксперты как завышают, так и занижают примерно в 8% проверенных заданий.

***

Подводя итоги проделанной работы, можно говорить и том, что были разработаны подход и методика детального анализа результатов работы Московской предметной комиссии по физике. На основе разработанной методики была написана программа автоматической обработки статистического массива данных, сформированного РЦОИ, которая позволила:

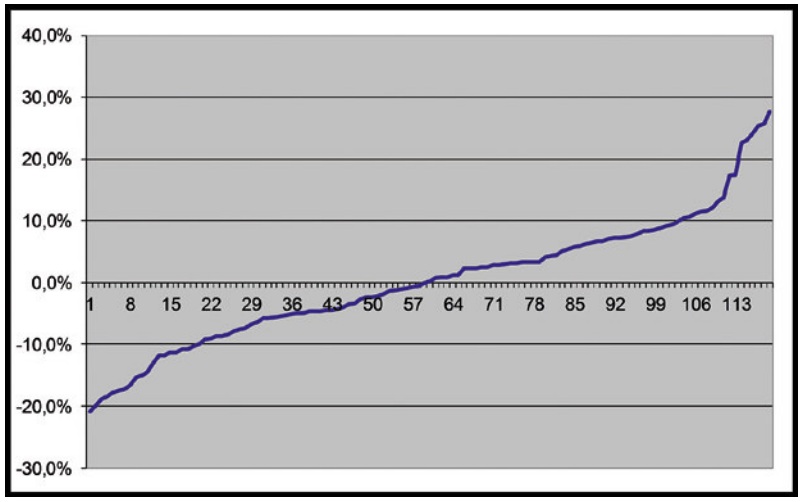

Рис. 2. Распределение «Коэффициента ошибок» среди экспертов

– проанализировать и выявить конкретные ошибки в работе каждого эксперта;

– рассчитать количественные оценки качества и согласованной проверки каждого эксперта;

– построить для каждого эксперта предметной комиссии карту личных показателей проверки (количество конкретных допущенных ошибок, средний процент ошибок, весовой коэффициент ошибок, коэффициент завышения/занижения);

– получить количественные показатели качества и согласованной проверки предметной комиссии в целом.

Полученные статистические данные позволят улучшить показатели работы как каждого эксперта в отдельности, так и предметной комиссии в целом. Разработанная методика анализа результатов работы предметной комиссии и рассчитанные индивидуальные показатели качества проверки дают возможность председателю рекомендовать экспертов для дальнейшей работы, более эффективно формировать предметную комиссию, а значит существенно уменьшить количество третьей проверки и повысить показатели согласованности проверки.

Разработанные подход и методика анализа могут быть использованы в работе как в предметных комиссиях ЕГЭ по другим предметам, так и коллегами в предметных комиссиях ОГЭ.