

Апробация заданий перспективных моделей КИМ ЕГЭ

**Полежаева
Мария Викторовна**

кандидат педагогических наук, начальник
отдела научно-методической экспертизы
и психометрических исследований ФГБНУ «ФИПИ»,
polejaeva@fipi.ru

**Орехова
Светлана Васильевна**

руководитель Центра экспертизы и аналитики
ФГБНУ «ФИПИ», svorekhova@fipi.ru

**Рыжко
Екатерина Борисовна**

начальник организационно-аналитического
отдела ФГБНУ «ФИПИ»,
ryzhko@fipi.ru

**Нурминский
Алексей Игоревич**

ведущий аналитик отдела научно-методической
экспертизы и психометрических исследований
ФГБНУ «ФИПИ», org@fipi.ru

Ключевые слова: единый государственный экзамен, контрольные измерительные материалы, перспективная модель, апробация заданий, выборка, статистические характеристики заданий.

Апробация заданий КИМ ЕГЭ (далее — апробация, апробационное исследование) проводится в целях проверки возможности применения этих заданий при проведении государственной итоговой аттестации (далее — ГИА). В течение двух последних лет апробировались новые модели заданий для перспективных моделей КИМ ЕГЭ по образовательным программам среднего общего образования в соответствии с требованиями федерального государственного образовательного стандарта среднего общего образования.

Обозначенная цель достигается посредством анализа статистических характеристик как апробируемых вариантов в целом (надёжность, распределение трудности заданий, средний первичный балл, корреляционные характеристики, информационные кривые и др.), так и отдельных заданий (трудность, дифференцирующая способность и др.), результатов анализа экспертных характеристик апробируемых заданий (корректность формулировок, значимость для ГИА, соответствие действующим стандартам и наличие проверяемого материала в учебниках федерального перечня Минпросвещения России и др.). Также анализируется время выполнения отдельных заданий и работы в целом, качество инструкций, отношение обучающихся к работе и др.

На основе результатов анализа проводится доработка заданий, отдельные задания могут отбраковываться. Далее формулируются выводы о возможности применения в КИМ ЕГЭ новых моделей заданий в ближайшие 2–3 года.

Апробационное исследование традиционно состоит из нескольких этапов, комплексно обеспечивающих проведение процедуры¹:

1) подготовка и организация исследования (определение сроков проведения, формирование выборки, определение и разработка технологии

¹ СТО ФИПИ 59576103.1.1–2020. Требования к качеству образовательных проектов и систем, применяемых для оценки качества образования: стандарт организации ФГБНУ «ФИПИ». — URL.: <https://doc.fipi.ru/o-nas/dokumenty-fipi/Standart-STO-FIPI-59576103.1.2-2020.pdf>.

проведения процедуры и обработки результатов апробации и др.);

2) подготовка инструментария для проведения исследования (основного — непосредственно предмет апробации, и сопроводительного — дополнительные материалы, средства, обучающие и подготовительные мероприятия и др.);

3) непосредственно проведение исследования;

4) сбор и обработка данных о результатах апробации;

5) анализ результатов, в том числе анализ полученных статистических данных, а также полученных экспертных заключений;

6) подготовка отчётов и принятие решений на основе проведённого исследования.

При подготовке и организации исследования формирование выборки для апробационного исследования проводилось с учётом требований, которые традиционно предъявляются к формированию выборки для массовых исследований. Так, при проведении апробации перспективных моделей заданий ЕГЭ в 2020–2021 гг. выборка для исследований формировалась таким образом, чтобы в нём приняли участие представители субъектов Российской

Федерации всех федеральных округов от Дальнего Востока до республик Северного Кавказа, что обеспечило участие в исследовании обучающихся образовательных организаций в том числе и тех территорий, специфика которых часто играет свою роль в результатах массовых испытаний оценки качества образования. В апробации приняли участие обучающиеся 10–11-х классов из 20 регионов Российской Федерации. Количество участников варьировалось по регионам, но общий объём выборки позволил сделать выводы о трудности и качестве разработанных материалов по учебным предметам. Общее количество участников апробации по всем предметам, кроме немецкого, французского, испанского и китайского языков, варьируется от 1442 (английский язык) до 9040 (русский язык) (диаграмма 1).

Релевантность данного объёма выборки обоснована и с позиций международного опыта в апробации тестовых заданий. Так, американский College Board — организация, которая разрабатывает и администрирует стандартизированные тесты и программы обучения, используемые средними и средне-профессиональными учебными заведениями, колледжами и университетами, — апробирует вопросы на выборке

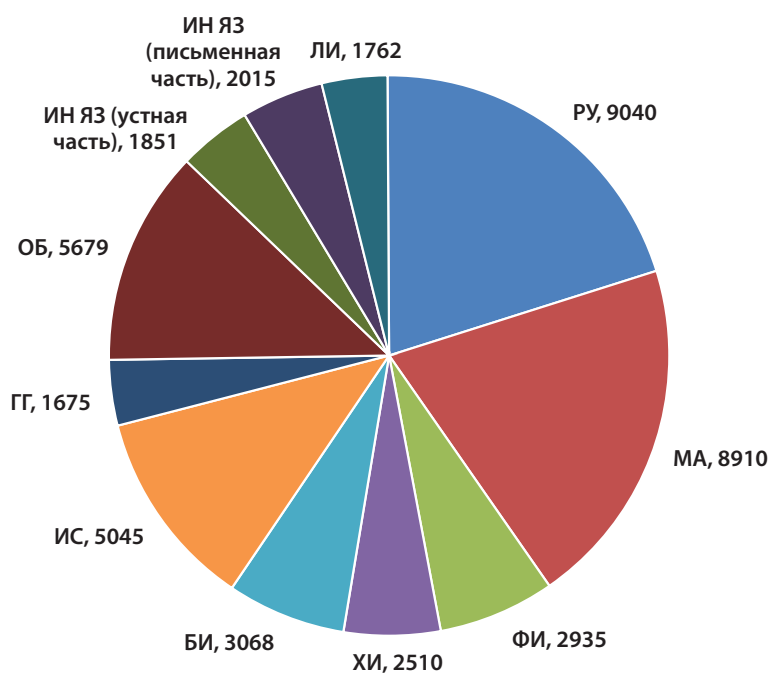


Диаграмма 1. Распределение количества участников апробации по предметам

от 1000 до 3000 учащихся², в Канаде апробации включают около 2000 учащихся³, а принципы датской национальной системы тестирования включают обязательную пилотную проверку новых вопросов на выборке в 500–700 учащихся⁴.

Внутри региона отбор образовательных организаций (далее — ОО) для участия в апробации, как правило, ведётся вероятностно-пропорциональным методом с учётом наполняемости школ и их территориально-административной принадлежности. Выбор ОО осуществляется на основе полного списка всех организаций планируемой совокупности.

Выборка ОО должна включать не менее 150 ОО в целом, при этом предполагается не менее 400 участников апробации на каждый вариант КИМ.

В качестве переменных внутренней стратификации выборки используются расположение, тип, наполняемость ОО.

Выборка для массовых обследований на федеральном уровне должна обладать следующими основными характеристиками:

- 1) репрезентативность по количественной и качественной составляющим;
- 2) мотивированность тестируемых, эквивалентная мотивации единого государственного экзамена;
- 3) подготовленность обучающихся, эквивалентная подготовке выпускников, участвующих в едином государственном экзамене.

Репрезентативность выборки обеспечивается работой группы специалистов, которая формирует эту выборку на основе имеющихся данных об образовательных организациях как на федеральном, так и на региональных уровнях.

Мотивированность тестируемых может обеспечиваться рядом специальных мер, при которых учащимся не будут безразличны результаты их работы. Эти меры могут определяться специальными требованиями к процедуре проведения тестирований.

В некоторых странах существует практика включения в состав материалов реального экзамена апробируемых заданий (какие именно из заданий являются апробируемыми, для участника экзамена остаётся неизвестным). Результаты выполнения апробируемых заданий не учитываются при подсчёте баллов итоговой оценки за экзамен, но анализируются с позиций перспективы использования в последующие годы (как, например, в экзаменах SAT в США). Однако данная практика не применяется в России, и вопросы мотивации учащихся при выполнении апробируемых заданий требуют дополнительных решений. Как отмечают и зарубежные специалисты, подобные апробации, когда обучающиеся знают о том, что выполнение заданий не влияет на оценки и поэтому выполняют задания без высокой мотивации, нередко приводят к некорректным выводам о сложности заданий ввиду более низкого уровня выполнения теста, а значит, к использованию впоследствии более простых заданий для настоящего экзамена. Тем не менее нельзя не учитывать и обратное мнение — что для некоторых обучающихся позитивную роль в выполнении апробируемых заданий играет отсутствие волнения, неизбежного на настоящем экзамене, и на апробации их успеваемость может быть лучше именно из-за отсутствия напряжённости (как отмечают, например, и датские исследователи)⁵. В целом наиболее очевидные доводы для участия в апробации включают такие преимущества для учащихся, как возможность ознакомиться с процедурой оценивания и самим форматом испытания, что позволит снизить тревожность перед настоящим экзаменом, а для преподавателей несомненно полезна оперативная информация об уровне учащихся. Это убедительная мотивация для участников, нередко входящая сразу в официальные анонсы проводимых апробаций (как, например, в программу апробационных испытаний канадской провинции Альберта)⁶.

² См.: <https://accuplacer.collegeboard.org>.

³ См.: <https://www.cmec.ca/Publications/Lists/Publications/Attachments/124/PCAP2007-Report.en.pdf>.

⁴ Wandall J. National Tests in Denmark-CAT as a Pedagogic Tool // Journal of Applied Testing Technology. — 2011. — Vol. 12. — Spec. iss. May. — URL.: https://www.testpublishers.org/assets/documents/JATT_Vol_12_special_3-CAT_as_a_pedagogic_tool.pdf.

⁵ Beuchert L.V., Nandrup A.B. The Danish national tests at a glance // Nationaløkonomisk tidsskrift. — København: Gyldendal, 2018. — Vol. 156 (1). — P. 30–66. — ISSN 0028-0453, ZDB-ID 280945-X.

⁶ См.: <https://www.alberta.ca/assets/documents/edc-field-test-rules-and-request-guide.pdf>.

Подготовленность обучающихся определяется используемыми при освоении учебного содержания образовательными технологиями и учебными программами. Эти характеристики можно определить посредством анкетирования участников апробационных исследований.

В 2020–2021 годах апробация заданий перспективных моделей КИМ ЕГЭ проводилась по следующим учебным предметам: математике, русскому языку, физике, химии, биологии, географии, обществознанию, истории, литературе, иностранным языкам (английскому, немецкому, французскому, испанскому и китайскому). По каждому предмету были подготовлены два или четыре апробационных варианта, а также комплекты сопроводительной документации для проведения апробационного исследования. Варианты включали в себя задания новых моделей и задания из вариантов ЕГЭ прошлых лет (так называемые якорные задания), статистические характеристики которых имеются в распоряжении разработчиков КИМ. Приоритетом апробации являлось испытание максимального количества различных по содержанию, формам и проверяемым умениям заданий, соответственно, задача выравнивания вариантов по статистическим характеристикам была второстепенной. Тем не менее по структуре (количество и форма заданий, максимальное количество баллов за них) варианты формировались единообразно. Все варианты прошли содержательную и тестологическую экспертизу в соответствии с разработанным и действующим в ФГБНУ «ФИПИ» регламентом проведения экспертизы.

Апробация проводилась в два этапа: осенью 2020 г. и весной 2021 г. В варианты первого этапа не включались задания, для выполнения которых требовался материал, который изучается в 11-м классе позднее.

Для обеспечения технологических процессов, связанных с проведением исследования, использовались решения системы «АВВУУ Мониторинг» компании «РТК-Иксора», обеспечивающие максимальное приближение процедуры исследования к условиям проведения ЕГЭ. В частности, задания апробации предъявлялись только в определённое время в едином

формате (в соответствии с разработанными оригинал-макетами апробационных вариантов для каждого отдельного учебного предмета). Варианты с заданиями были распечатаны для каждого участника апробации непосредственно в ОО вместе с машиночитаемыми бланками, в которые необходимо было внести информацию, аналогичную той, которую требуется внести в бланки ЕГЭ (минимальная идентификационная информация и ответы: краткие и/или развёрнутые (письменные) ответы на задания апробационного варианта). Поскольку апробировались в том числе и задания устной части ЕГЭ по иностранным языкам, необходимо было реализовать организационно-технологические элементы проведения апробации устной части экзамена, включая запись устных ответов участников апробации и оценивание экспертами устных ответов по аудиозаписи. Как и при проведении ЕГЭ, проверка кратких ответов осуществлялась автоматизированно, оценивание развёрнутых ответов (устных и письменных) проводилось экспертами в соответствии с критериями оценивания выполнения заданий. Выбор технологического решения был обусловлен и возможностями системы «АВВУУ Мониторинг», которая позволяет представить, в том числе, статистические данные обо всех этапах проведения и обработки материалов апробации, некоторые статистические характеристики заданий, а также изображения бланков с ответами участников апробаций и звуковые файлы с записью устных ответов для более детального содержательного анализа.

При проведении анализа результатов апробации рассматривались следующие статистические характеристики заданий: трудность, дифференцирующая способность и точечно-бисериальный коэффициент корреляции⁷. К статистическим характеристикам заданий, используемых в КИМ для ГИА, предъявляются определённые требования.

⁷ Полежаева М.В., Нурминский А.И., Орехова С.В., Полежаев В.Д. Подходы к определению данных статистических характеристик более подробно рассмотрены в статье: Влияние формы задания на его статистические характеристики (на примере контрольных измерительных материалов ЕГЭ) // Педагогические измерения. — 2018. — № 2. — С. 97–103.

При анализе результатов выполнения задания учитывается доля участников, которые по каким-либо причинам не выполнили задание: пропустили его или не успели выполнить. Если данное число превышает 5%, то важно понять причину невыполнения (незнание правильного ответа; нежелание выполнять задание по причине его трудоёмкости; отсутствие времени на выполнение, поскольку задание расположено в конце варианта) и принять соответствующие меры для исправления ситуации. Столь низкий процент выполнения допустим только для заданий высокого уровня сложности, которые предназначены для дифференциации абитуриентов вузов с высоким конкурсом.

Значение параметра «дифференцирующая способность» может варьироваться от -1 до 1 и должно быть не менее $0,2$. Индикатором, достаточным для включения задания в экзаменационную работу, являются характеристики ТБКК для каждой категории ответа (на $0, 1, 2$ балла и т.д.). В экзаменационную работу не включаются задания, для которых:

- $\text{ТБКК} < 0,2$ для максимального балла;
- $\text{ТБКК} > 0$ для неверного ответа (нулевого балла). Такое значение параметра говорит о низком качестве задания, т.е. о том, что неверный ответ дали участники со средней (или выше среднего) подготовкой;
- $\text{ТБКК} \leq 0$ для ненулевого балла означает проблему, показывая, что балл получен участниками с низким уровнем подготовки;
- процент выполнения по каждой категории (на каждое значение балла) $< 5\%$. Это означает, что неправильно сформирована шкала оценивания и критерии;
- ТБКК не возрастает при переходе от более низкого балла к более высокому. Данный факт является главным основанием исключения задания из экзаменационного варианта, поскольку указывает на то, что задание не обладает дифференцирующей способностью, т.к. более низкий балл могут получить участники с более высокой подготовкой по предмету.

Сравнение результатов выполнения «якорных» заданий в рамках апробационного исследования и полученных по ито-

гам ЕГЭ позволяет предположить, что процент выполнения заданий новых моделей в рамках ЕГЭ может быть несколько выше процента выполнения, полученного по результатам апробации.

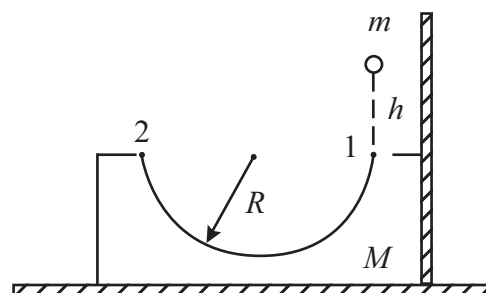
Анализ результатов показал, что большинство заданий отвечает предъявляемым требованиям, лишь некоторые задания оказались слишком сложны для выборки.

Рассмотрим, например, задание 16 в апробационных вариантах КИМ по физике. В данной линии предлагалась расчётная задача с дополнительным (по сравнению с используемой в КИМ ЕГЭ моделью) условием — описанием физической модели. От обучающихся требовалось обосновать применимость выбранных законов и формул к процессам, описанным в задаче.

Пример 1

Шарик небольшого размера, имеющий массу m , падает с высоты $h = 0,1$ м в точку 1, являющуюся верхней точкой полусферической выемки радиуса R , сделанной в платформе массой $M = 4m$. Платформа находится вплотную у вертикальной стенки на горизонтальной гладкой плоскости. Шарик, двигаясь по гладкой сферической поверхности и не вращаясь, останавливается в точке 2, находящейся на противоположном конце диаметра сферической поверхности. Чему равен радиус R ? Сопротивлением воздуха пренебречь.

Какие закономерности Вы использовали для описания движения шарика и платформы? Обоснуйте их применимость к данному случаю.



Задание имеет крайне низкие результаты выполнения. 0 баллов получили более 96% участников, выполнявших данный вариант. Максимальный балл, равный 4, получили только около 0,5% участников. Дифференцирующая способность задания также довольно низка — около 0,07.

На основе результатов анализа работ участников апробации можно рекомендовать корректировку системы оценивания задания с выделением обоснования физической модели в отдельный критерий на 1 балл. Разработчики также приняли решение о необходимости дополнительной апробации заданий с изменёнными критериями и более широким спектром тематики для формирования окончательного вывода о возможности использования в КИМ ЕГЭ.

При анализе результатов апробации разработчики рассматривали не только статистические характеристики, но и отзывы педагогической общественности и результаты оценивания конкретных работ. В качестве примера задания, имеющего хорошие статистические характеристики, но требующего доработки, можно привести задание 13 из апробационного варианта по истории (пример 2).

Пример 2

Сравните ход, результаты Ливонской и Северной войн.

Сформулируйте две общие характеристики.

Заполните таблицу, отражающую различия: сформулируйте две линии (критерии) сравнения в форме вопросов, предполагающих однозначные ответы «да» или «нет», и заполните соответствующие ячейки таблицы.

Различия

Линии (критерии) сравнения	Ливонская война	Северная война

Выполнение задания оценивалось по трём критериям: общим характеристикам (2 балла), линиям (критериям) сравнения (2 балла) и корректности указания различий по линиям (критериям) (1 балл). Средние проценты выполнения по этим линиям составили соответственно 8,4, 16,8 и 15,4%, дифференцирующие способности: 0,38, 0,62 и 0,63. Наблюдался рост ТБKK с увеличением баллов. Тем не менее при этом были выявлены недостатки в формулировке критериев оценивания и предложена доработка задания⁸.

По итогам апробации можно сделать вывод о том, что большинство заданий может быть включено в КИМ ЕГЭ 2022 г. Отдельные модели заданий рекомендованы к включению в КИМ ЕГЭ 2023 г. Были выявлены задания, которые требуют доработки, определены пути их совершенствования. Эта работа будет проводиться с учётом того факта, что мотивация участников исследования и уровень выполнения ими заданий значительно ниже мотивации участников ЕГЭ и уровня выполнения заданий ими. Это подтверждает и анализ статистики выполнения «якорных» заданий.

Кроме того, апробация предоставила широкий спектр образцов выполнения заданий с развёрнутым ответом, которые будут использоваться для совершенствования критериев оценивания новых моделей заданий и для подготовки методических рекомендаций для экспертов по проверке ответов на задания с развёрнутым ответом.

⁸ Доработка модели задания описана в статье: Артасов И.А., Мельникова О.Н. Особенности перспективной модели КИМ ЕГЭ по истории // Педагогические измерения. — 2021. — № 2.