Технический отчет по проекту SAM
(Student Achievement's Monitoring)

Оглавление

1. Проект SAM: цель разработки, теоретическая основа, результат
2. Психометрические характеристики тестов SAM
2.1. Результаты статистического анализа данных апробации теста по математике
2.2. Результаты статистического анализа данных апробации теста по русскому языку
3. Обоснование валидности SAM
3.1. Валидность по содержанию
3.2. Конструктная валидность
3.3. Критериальная валидность
4. Анализ функционирования заданий по отношению к различным группам испытуемых (DIF анализ)
4.1. DIF анализ заданий теста по математике
4.2. DIF анализ заданий теста по русскому языку
5. Оценивание участников тестирования
6. Первичный анализ результатов тестирования
Заключение

1. Проект SAM: цель разработки, теоретическая основа, результат

Проект посвящен разработке инструмента мониторинга учебных достижений SAM, применяемого на уровне отдельной школы или муниципальной школьной системы. Цель проекта состояла в создании набора педагогических тестов, позволяющих сочетать измерение предметных образовательных результатов с их качественной характеристикой на основе соотнесения с некоторой таксономией педагогических целей.

Модель SAM разработана в рамках теории культурного развития, предлагаемой научной школой Л.С.Выготского [1; 2; 3; 4; 6; 17]. Ключевым элементом модели выступает трехуровневая таксономия присвоения деятельностного культурного содержания, вписанная в возрастной контекст [5; 12; 24]. Согласно этой таксономии существует три качественных уровня овладения культурным способом действия, которые для краткости можно обозначить как формальный, рефлексивный и функциональный (рис.1.1). Каждый из этих уровней связан с определенным типом ориентировки в предметном содержании и проявляется в способности решать соответствующий класс задач.

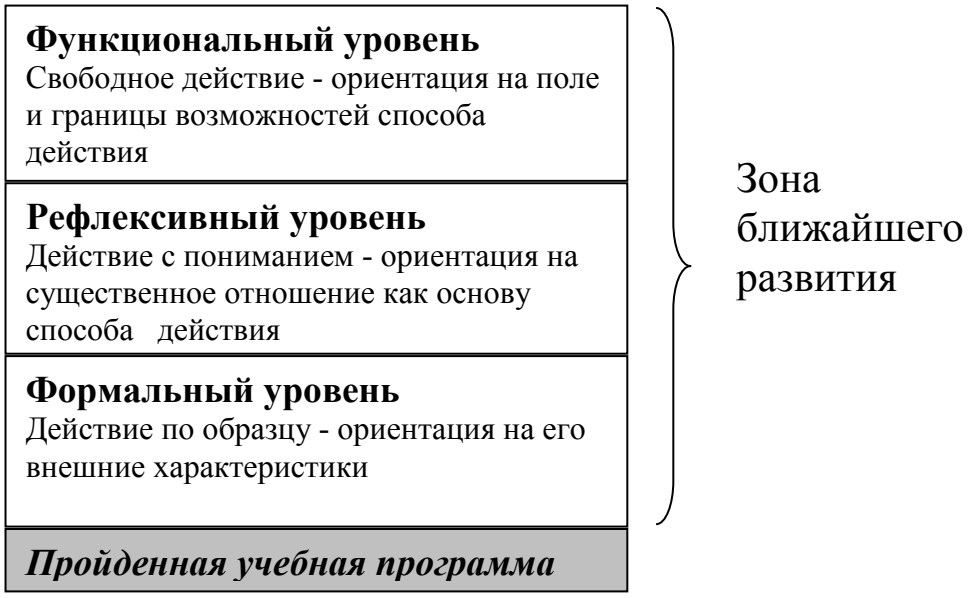

Рис. 1.1. Уровневая модель присвоения деятельностного предметного содержания

Описанная уровневая схема положена в основу построения тестов SAM.

Уровням освоения способа действия были поставлены в соответствие индикаторы, т.е. типы задач. При создании предметного теста для каждого относительно целостного раздела учебного предмета разрабатываются блоки задач. Каждый блок включает три задачи – 1-го, 2-го и 3-го уровня, образующие естественную иерархию по трудности. Блок задач служит детектором, выявляющим уровень присвоения соответствующего предметного материала (определяется по самой трудной задаче блока, которую решил тестируемый). Совокупность блоков, покрывающая основные разделы учебного предмета, позволяет выявить внутреннюю картину его присвоения учащимися.

В качестве иллюстрации ниже приведены блоки задач по математике и русскому языку:

Блок задач по математике

1 Какое получится число, если 10472 разделить на 34? Ответ: _____
2 Рассеянный Петя, переписал из учебника пример на умножение двух чисел. Первый множитель он записал правильно: 7. А во втором множителе по ошибке переставил местами цифры. Из-за этого в ответе у него получилось число 147. Какой ответ должен был получить Петя, если бы он правильно переписал пример? Ответ: _____
3 Какой самый большой результат может получиться, если в выражении AB5 + BC2 заменить буквы цифрами (разные буквы заменяются разными цифрами)? Ответ: _____

Примечание. Все три задачи относятся к разделу «Числа и вычисления». Первое задание предполагает прямое применение алгоритма вычисления. Второе требует анализа ошибочного арифметического действия (с учетом позиционного принципа) и построения программы по его коррекции. Наконец, третье задание предполагает «обыгрывание» позиционного принципа для определения одного из конкретных значений, удовлетворяющих требованию получить максимальное значение выражения.

Блок задач по русскому языку

1 Отметь предложение, в котором поставлены НЕ ВСЕ нужные запятые. 1. Осенний листопад. Листья летят, скачут, плывут. 2. Мальчишки просидели в засаде до вечера, но ушли почти ни с чем. 3. Мама сидела за компьютером и писала какой-то доклад. 4. На массивном столе с зелёным сукном лежал ноутбук папки с бумагами, калькулятор.
2 Поставь знаки препинания в соответствии с описанными ситуациями. 1. Один ученик неверно расставил знаки препинания, и у него получилось, что море выбросило на берег предметы и людей. **Прибоем выбросило на берег корабль__испанцев__лодку__рыбака__катер.**

2. Другой ученик верно расставил знаки препинания, и у него получилось, что море выбросило на берег только предметы. **Прибоем выбросило на берег корабль__испанцев __лодку__ рыбака__катер.**
3 Переформулируй предложение так, чтобы смысл его сохранился, но оно стало простым с однородными членами. **Вода непрестанно трудится, и время от времени края горных уступов обрушиваются.** _____ _____

Примечание. Задачи этого блока связаны с расстановкой знаков препинания в простом предложении с однородными членами. В первой задаче дан набор простых предложений, предполагающих прямое применение соответствующего правила. Вторая задача требует выделения однородных членов в условиях смысловой неоднозначности предложений, т.е. предполагает выявление и учет существенных в данном случае смысловых отношений. Третья задача предполагает активное владение понятием простого предложения с однородными членами, позволяющим преобразовывать сложное высказывание в простое при сохранении его содержания, т.е. рассчитана на анализ вариантов и выбор адекватного.

Из подобных блоков, построенных на содержании основных разделов учебной дисциплины, составляются варианты теста. В соответствии с этим, тест может рассматриваться как система из трех субтестов, каждый из которых представляет собой набор задач одного уровня из различных разделов предметного содержания (рис. 1.2.).

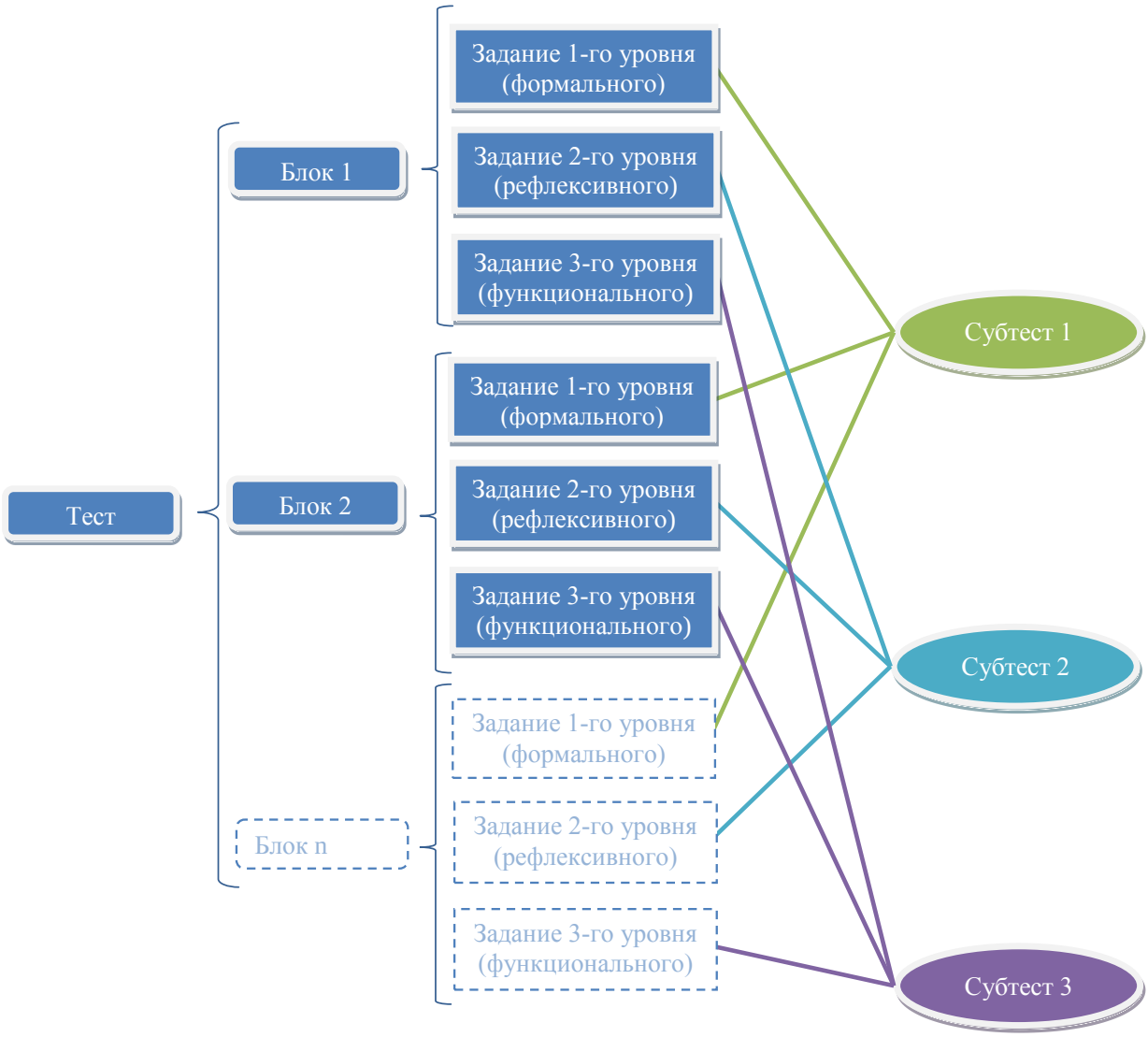

Рис. 11.2. Структура теста

В результате тест может выполнять две функции: а) измерения предметных образовательных результатов (45 задач, покрывающих основные разделы предмета) и б) диагностики их качественных уровней (15 задачных блоков).

В тестах SAM использовались задания открытой формы с кратким ответом, закрытой с выбором одного ответа из 4-5 предложенных, задания на установление соответствия, требующие построений и другие. Задачи в тесте предъявлялись блоками. Последовательность предъявления блоков значения не имела.

Для оценивания заданий использовался дихотомический подход: за правильный ответ ученик получает 1 балл, за неправильный ответ (или отсутствие ответа) – 0 баллов. Таким образом, максимальное количество первичных баллов, которые может набрать участник за выполнение теста, равно 45. Максимальное количество первичных баллов по каждому субтесту (т.е. по каждому уровню) равно 15.

Каждый предметный тест выполнен в нескольких вариантах с близкими статистическими характеристиками. При этом во все варианты включаются общие задания, позволяющие провести процедуры выравнивания и получить оценки участников на единой шкале. Число общих заданий в различных вариантах составляет не менее 6 (не менее двух общих блоков). Тесты могут предлагаться в бланковой и компьютерной формах.

В рамках модели SAM разработаны тесты по двум предметам – математике и русскому языку – имеющие общую структуру.

Полное описание модели и инструментария SAM представлено в книге: SAM (School Achievement Monitoring): Инструмент мониторинга учебных достижений школьников // *под ред. Нежнова П.Г., Кардановой Е.Ю.*, 2011.

Литература

1. Выготский Л.С. История развития высших психических функций //Собр. соч.: В 6 т., т. 3. М.: Педагогика, 1983.
2. Выготский Л.С. Мышление и речь //Собр. соч.: В 6 т., т. 2. М.: Педагогика, 1982.
3. Гальперин П.Я. Психология как объективная наука. М., 1998.
4. Давыдов В.В. Теория развивающего обучения. М.:ИНТОР, 1996.
5. Диагностика учебной успешности в начальной школе. М.: ОИРО, 2009.
6. Запорожец А.В. Психология действия. Москва-Воронеж: НПО «МОДЭК», 2000.
7. Исаев Е.И. Психологическая характеристика способов планирований у младших школьников // Вопр. психологии. 1983. № 2.
8. Лернер И.Я. Процесс обучения и его закономерности. М.: Знание, 1980Магкаев В.Х. Экспериментальное изучение планирующей функции мышления в младшем школьном возрасте // Вопросы психологии, 1974, № 5.
9. Магкаев В.Х. Экспериментальное изучение планирующей функции мышления в младшем школьном возрасте // Вопросы психологии, 1974, № 5.
10. Микулина Г.Г., Савельева О.В. К психологической оценке качества знаний у младших школьников // Психологическая наука и образование. 1997. № 2.
11. Нежнов П.Г. Опосредствование и спонтанность в модели «культурного развития» // Вестн.Моск.Ун-та. Серия 14. Психология. 2007, № 1.
12. Нежнов П.Г., Карданова Е.Ю., Эльконин Б.Д. Оценка результатов школьного образования: структурный подход // Вопросы образования, 2011, № 1, стр. 26-43.
13. Нежнов П.Г., Медведев А.М. Исследование теоретического анализа у школьников // Вопросы психологии, 1990, № 5.
14. Савельева О.В. Психологические критерии качества знаний младших школьников // Автореф… дисс. канд. пед. наук. М., 1989.

15. Симонов В.П. Диагностика степени обученности учащихся: Учебно-справочное пособие. М.: МПА, 1999
16. Эльконин Б.Д. Введение в психологию развития. М., 1994.
17. Эльконин Д.Б. Избранные психологические труды. М.: Педагогика, 1989.
18. Карданова Е.Ю. Моделирование и параметризация тестов: основы теории и приложения. – М.: Федеральный центр тестирования, 2008, 304 с.
19. Kardanova, E. (2010): The development of the toolkit for assessment of subject competences of primary school students. The paper presented at the 36-th Annual Conference IAEA 2010 “Assessment for the Future Generations”. Bangkok
20. Карданова Е.Ю., Нейман Ю.М. Проблема выравнивания в современной теории тестирования. Вопросы тестирования в образовании, 2003, № 8, с. 21-40.
21. Kardanova, E., Nezhnov, P. (2011): School Achievements Monitoring Toolkit: Assessment Framework. The paper presented at the 37-th Annual Conference IAEA 2011. Manila.
22. Bloom, B., Englehart, M. Furst, E., Hill, W., & Krathwohl, D. (1956), *Taxonomy of educational objectives: The classification of educational goals. Handbook I: Cognitive domain*, New York, Toronto, Longmans, Green.
23. Mullis, I. V.S., Kennedy, A. M., Martin, M.O., Sainsbury, M. (2006), PIRLS 2006 *Assessment Framework and Specifications, 2nd Edition*, TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College.
24. Nezhnov, P. (2011). SAM – toolkit to assess primary school students’ academic achievements. CADMO. Innovations in assessment to meet changing needs. ANNO XIX, 1, pp.85-98.
25. SAM (School Achievement Monitoring): Инструмент мониторинга учебных достижений школьников // *под ред. Нежнова П.Г., Кардановой Е.Ю.*, 2011, 104 с.

2. Психометрические характеристики тестов SAM

В отчете приведены результаты статистического анализа данных апробации тестов SAM по математике и русскому языку, которая проводилась в бланковой форме в регионах Российской Федерации весной 2012 г. Общее число участников апробации составляет более 5000 учащихся 4-х классов общеобразовательных школ.

Анализ результатов апробации проводился как в рамках классической теории тестирования, так и в соответствии с современной теорией тестирования IRT.

2.1. Результаты статистического анализа данных апробации теста по математике

В апробации были использованы два варианта теста по математике. Тест состоял из 45 заданий, объединенных в 15 блоков. В табл. 2.1 представлены суммарные статистические показатели по двум вариантам теста **в рамках классической теории тестирования**, на рис.2.1 - гистограммы распределения первичных баллов участников тестирования.

Табл. 2.1 Суммарные результаты теста по математике

	1 вариант	2 вариант
Число испытуемых	3018	2941
Средний первичный балл	26	27
Стандартное отклонение	8.37	8.55
Коэффициент асимметрии	-0.21	-0.37
Коэффициент эксцесса	-0.56	-0.36

Средний коэффициент решаемости	0.59	0.61
Средний индекс дискриминативности	0.44	0.46
Средний скор. коэф. точ.-бис. корреляции	0.39	0.39
Показатель надежности (KR20)	0.90	0.91
Стандартная ошибка измерения	2.61	2.61

1 вариант 2 вариант

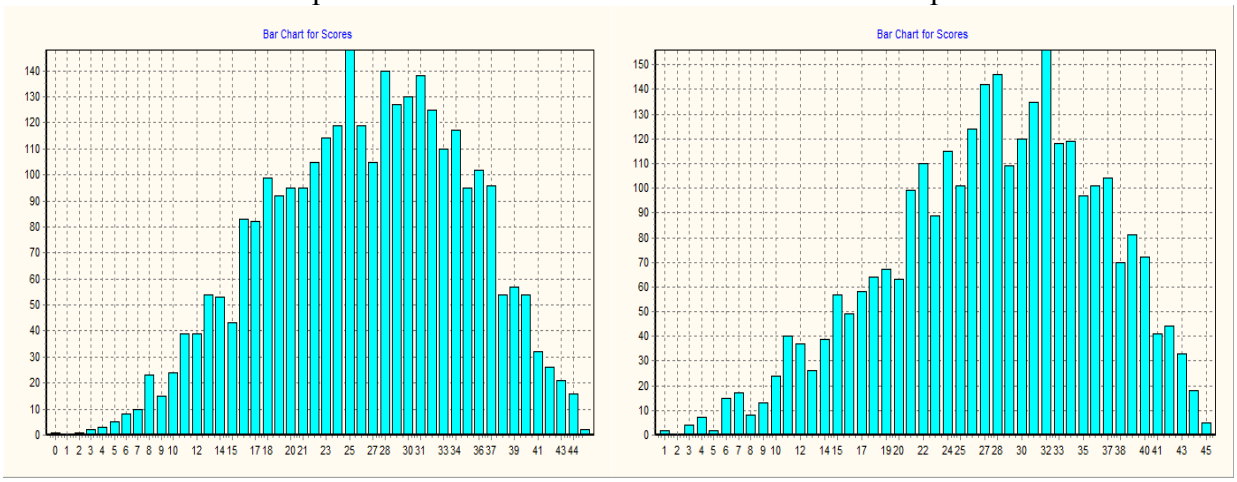

Рис. 2.1. Гистограммы распределения первичных баллов участников тестирования

Из анализа таблицы следует, что статистические показатели двух вариантов достаточно близки. Показатели надежности имеют достаточно высокие значения – не ниже 0,9 по обоим вариантам. Средние коэффициенты решаемости и дискриминативности близки к оптимальным значениям.

В табл. 2.2 представлена детальная информация по заданиям **варианта 1**. При классическом анализе основными характеристиками задания являются его трудность и дискриминативность. Трудность задания (коэффициент решаемости) определяется как доля испытуемых выборки, выполнивших задание правильно. Чем выше этот показатель, тем задание легче.

Дискриминативность характеризует дифференцирующую способность задания, т.е. его способность различать испытуемых с различным уровнем подготовки. В отчете задействованы два коэффициента дискриминативности: классический показатель дискриминативности (разность между коэффициентами решаемости задания двумя группами испытуемых - 27 % лучших (испытуемых, имеющих высокие баллы) и 27 % худших (испытуемых, имеющих низкие баллы)) и скорректированный коэффициент точечной бисериальной корреляции (коэффициент корреляции между баллом по заданию и общим баллом по тесту после удаления результатов по этому заданию). В качестве критического значения коэффициентов дискриминативности выбрано значение 0,2.

Из анализа табл. 2.2 можно заключить, что все задания анализируемого теста хорошо функционируют. Четыре задания первого уровня оказались очень легкими для данного контингента испытуемых – коэффициент решаемости выше 0,9, что означает, что эти задания выполняют более 90% испытуемых (эти задания в таблице выделены зеленым цветом). Эти же четыре задания обладают низкой дискриминативностью (задания с низкой дискриминативностью выделены в таблице розовым цветом). Именно экстремальный уровень легкости этих заданий обуславливает их низкий уровень дифференцирующей способности: задание, которое выполнили почти все, не обладает дифференцирующей силой.

Все остальные задания теста демонстрирует хорошие статистические показатели. Заметим, что внутри каждого блока из трех заданий 1-го, 2-го и 3-го уровней наблюдается иерархия по трудности.

Табл. 2.2 Статистические показатели функционирования заданий (математика, вариант 1)

№ задания и его код		Коэффициент. решаемости	Показатель дискриминативности	Скорректир. точ-бис. корреляция
1	M-C-01-1-1	0.93	0.13	0.21
2	M-C-01-1-2	0.63	0.62	0.49
3	M-C-01-1-3	0.17	0.37	0.38
4	M-C-03-1-1	0.98	0.06	0.15
5	M-C-03-1-2	0.71	0.47	0.39
6	M-C-03-1-3	0.29	0.37	0.30
7	M-M-02-1-1	0.88	0.31	0.39
8	M-M-02-1-2	0.73	0.51	0.43
9	M-M-02-1-3	0.33	0.53	0.42
10	M-M-03-1-1	0.87	0.28	0.32
11	M-M-03-1-2	0.66	0.63	0.50
12	M-M-03-1-3	0.45	0.60	0.44
13	M-M-06-1-1	0.74	0.49	0.44
14	M-M-06-1-2	0.49	0.68	0.51
15	M-M-06-1-3	0.43	0.54	0.40
16	M-M-11-1-1	0.89	0.20	0.27
17	M-M-11-1-2	0.56	0.72	0.54
18	M-M-11-1-3	0.21	0.42	0.39
19	M-R-02-1-1	0.73	0.60	0.51
20	M-R-02-1-2	0.52	0.67	0.50
21	M-R-02-1-3	0.36	0.51	0.40
22	M-R-05-1-1	0.97	0.05	0.14
23	M-R-05-1-2	0.68	0.49	0.39
24	M-R-05-1-3	0.16	0.34	0.35
25	M-G-01-1-1	0.82	0.27	0.28
26	M-G-01-1-2	0.49	0.56	0.41
27	M-G-01-1-3	0.20	0.32	0.29
28	M-D-03-1-1	0.81	0.28	0.28
29	M-D-03-1-2	0.54	0.61	0.45
30	M-D-03-1-3	0.31	0.54	0.45
31	M-D-05-1-1	0.81	0.40	0.42
32	M-D-05-1-2	0.70	0.50	0.40
33	M-D-05-1-3	0.25	0.42	0.35
34	M-D-08-1-1	0.93	0.15	0.28
35	M-D-08-1-2	0.66	0.54	0.44
36	M-D-08-1-3	0.19	0.26	0.25
37	M-R-03-1-1	0.87	0.28	0.35
38	M-R-03-1-2	0.62	0.47	0.35
39	M-R-03-1-3	0.35	0.49	0.38
40	M-C-05-1-1	0.82	0.35	0.35
41	M-C-05-1-2	0.59	0.61	0.46

42	М-С-05-1-3	0.40	0.72	0.54
43	М-М-08-1-1	0.88	0.29	0.37
44	М-М-08-1-2	0.46	0.71	0.53
45	М-М-08-1-3	0.35	0.67	0.52

На рисунках ниже показатели из табл. 2.1 представлены графически.

На рис.2.2 представлено распределение коэффициентов решаемости заданий анализируемого теста в виде столбчатой диаграммы. На оси абсцисс расположены коды заданий (в порядке следования в тесте), по оси ординат – коэффициенты решаемости. Задания 1-го уровня выделены синим цветом, задания 2-го уровня – красным, и задания 3-го уровня – зеленым. На диаграмме хорошо видна иерархия по трудности внутри каждого блока. На рис.2.3 также показано распределение коэффициентов решаемости заданий, но в виде графика.

На рис. 2.4-2.5 показаны распределения коэффициентов дискриминативности заданий теста. И наконец, на рис.2.6 показано совместное распределение коэффициентов решаемости и показателей дискриминативности заданий.

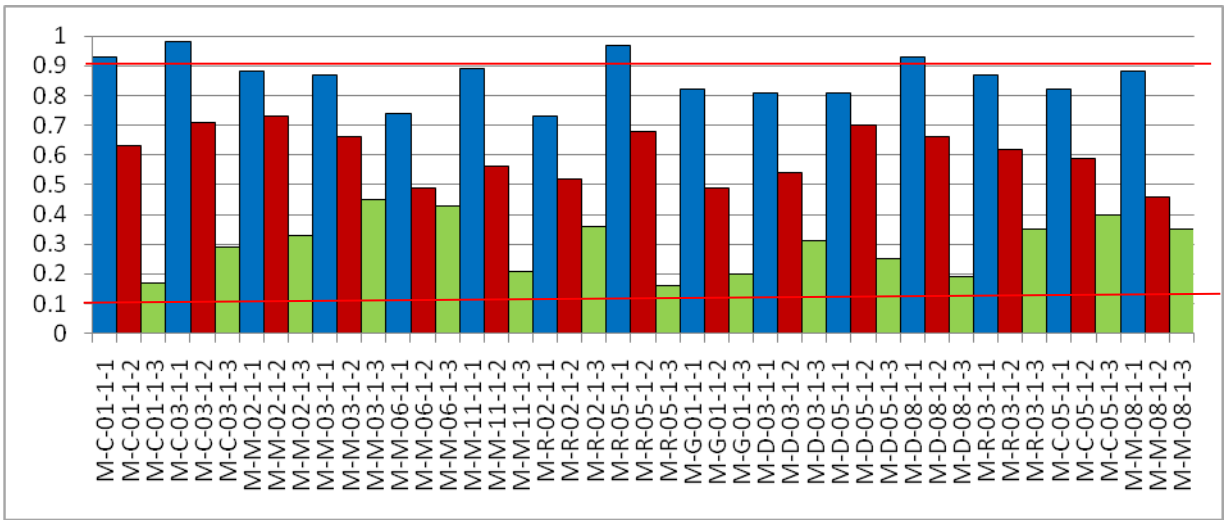

Рис.2.2. Диаграмма распределения коэффициентов решаемости (математика, вариант 1)

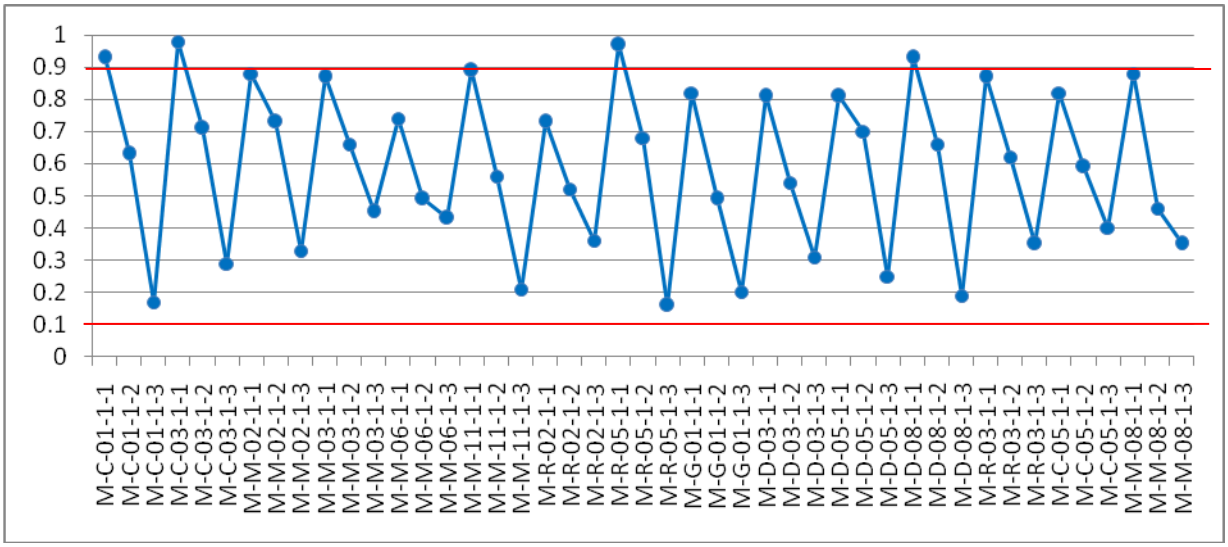

Рис. 2.3. График распределения коэффициентов решаемости (математика, вариант 1)

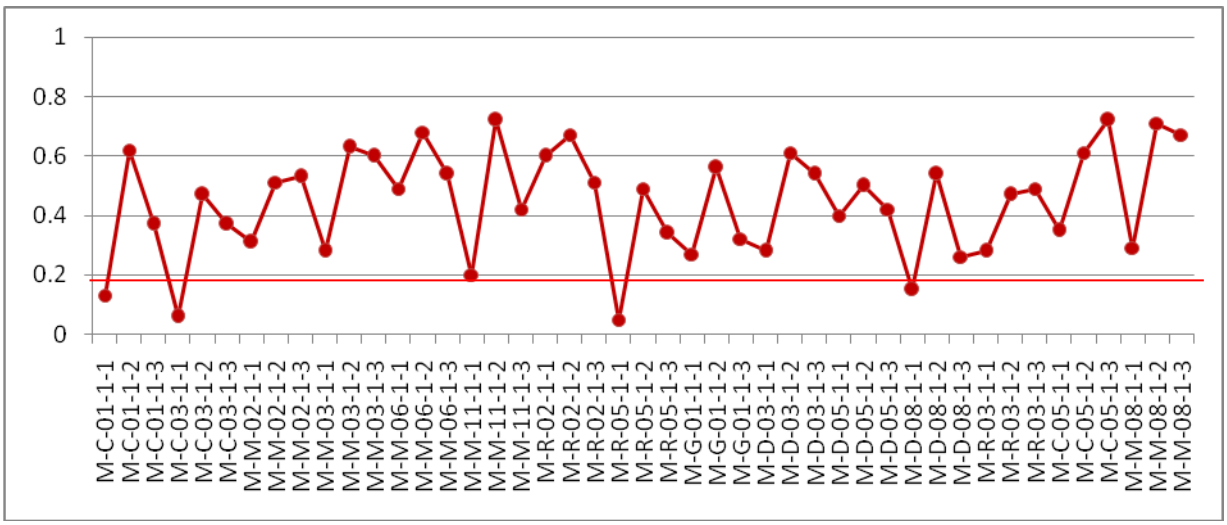

Рис.2.4. График распределения показателей дискриминативности заданий (математика, вариант 1)

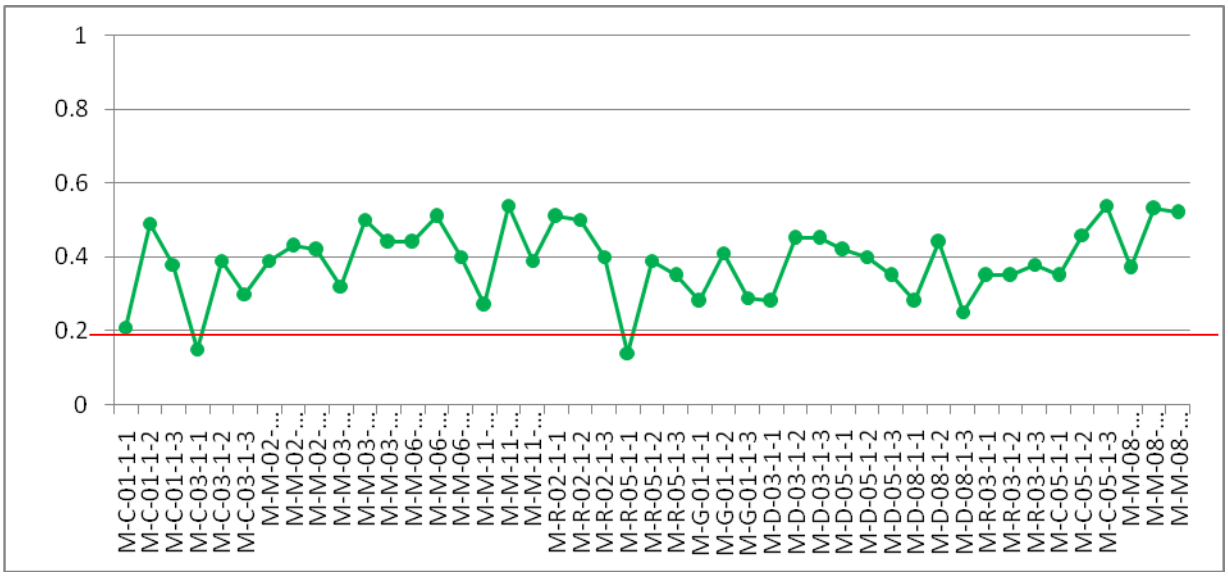

Рис.2.5. График распределения коэффициентов точечно-бисериальной корреляции (математика, вариант 1)

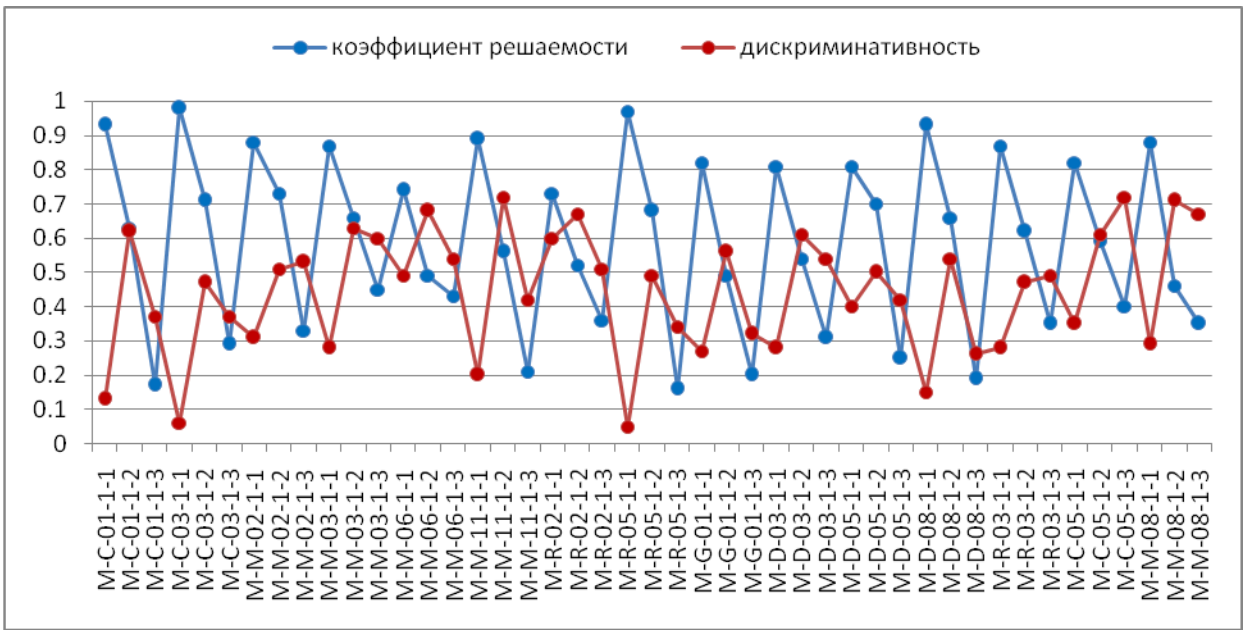

Рис.2.6. Совместное распределение коэффициентов решаемости и показателей дискриминативности (математика, вариант 1)

На рисунках хорошо видно, что практически все задания имеют хорошие показатели – оптимальную трудность и высокую дискриминативность. Только 4 экстремально легких задания 1-го уровня имеют показатели дискриминативности ниже нормы.

Дополнительно был проведен **анализ функционирования дистракторов** в заданиях закрытой формы (с выбором ответа из нескольких предложенных). В данном тесте таких заданий было 7.

Для анализа дистракторов, во-первых, было проанализировано распределение ответов испытуемых по всем ответным альтернативам заданий закрытого типа с целью определения неработающих дистракторов. Во-вторых, были вычислены коэффициенты корреляции между дистракторами и общим баллом по тесту (точечно-бисериальный коэффициент корреляции). Дистрактор функционирует правильно, если испытуемые с высоким уровнем подготовки его не выбирают в качестве правильного ответа. В этом случае коэффициент корреляции будет отрицательным (желательно, чтобы он был меньше -0,2). И, наоборот, коэффициент корреляции для правильного варианта ответа должен быть положительным (коэффициент корреляции для правильного ответа совпадает с коэффициентом дискриминативности заданий, рассмотренным ранее).

Серьезных проблем в функционировании дистракторов закрытых заданий анализируемого теста диагностировано не было.

В таблицах 2.3 и 2.4 показаны примеры анализа дистракторов для двух заданий. Правильный вариант ответа помечен звездочкой и выделен цветом. Для задания М-М-03-1-2 все показатели находятся в пределах нормы: 66 % участников тестирования выполнили задание правильно. Высокий положительный коэффициент корреляции говорит о том, что это сильные учащиеся, имеющие высокие баллы по тесту. Остальные учащиеся, не выполнившие задание, распределились по трем дистракторам. Все дистракторы при этом работают – каждый выбрало не менее 5% испытуемых. Отрицательные коэффициенты корреляции говорят о том, что это слабые учащиеся, имеющие низкие баллы по тесту.

Табл.2.3. Анализ дистракторов задания М-М-03-1-2

Задание М-М-03-1-2

Варианты ответов	**А**	**Б***	**В**	**Г**
Распределение ответов испытуемых	11%	66%	7%	6%
Коэффициент корреляции	-0,22	0,54	-0,29	-0,23

Таблица 2.4. Анализ дистракторов задания М-М-03-1-3

Задание М-М-03-1-3					
Варианты ответов	**a + b**	**a * b**	**a + 2b**	**2a + b***	**2a + 2b**
Распределение ответов испытуемых	11%	9%	9%	44%	22%
Коэффициент корреляции	-0,20	-0,31	-0,10	0,49	-0,05

Задание М-М-03-1-3 более трудное: его выполнили только 44% испытуемых и 5% его пропустили. Высокий положительный коэффициент корреляции у правильного варианта ответа говорит о том, что это сильные учащиеся, имеющие высокие баллы по тесту. Остальные учащиеся, не выполнившие задание, распределились по четырем дистракторам. Все дистракторы при этом работают – каждый выбрало не менее 5% испытуемых. Отрицательные коэффициенты корреляции говорят о том, что это слабые учащиеся, имеющие низкие баллы по тесту. Однако, в данном задании дистрактор 2a + 2b является предпочтительным – его выбрало 22% испытуемых. При этом его корреляция с общим баллом незначительна, что означает, что этот дистрактор является привлекательным как для слабых испытуемых, так и для испытуемых с хорошей подготовкой. Такая ситуация бывает характерна для заданий, содержащих подвох, который не замечают сильные учащиеся.

Таким образом, анализ заданий варианта 1 по математике в рамках классической теории тестирования показал, что все задания имеют хорошие показатели.

Более тщательный анализ теста был проведен **в рамках современной теории тестирования IRT.** Ниже приведены основные результаты этого анализа.

В качестве модели тестирования используется одномерная дихотомическая модель Раша. В соответствии с этой моделью каждое задание теста характеризуется одним параметром – трудностью, каждый участник тестирования характеризуется одним параметром – уровнем подготовленности. Оценки всех параметров – и участников тестирования и заданий теста - находятся на общей метрической шкале (единица измерения на этой шкале называется логит) и снабжены характеристиками точности оценивания. Начало этой шкалы не определено и может быть выбрано произвольно. С целью фиксирования начала отсчета среднее значение оценок трудности всех заданий положено равным 0.

В первую очередь было проведено исследование размерности теста. Для этого применялся факторный анализ нормированных уклонений ответа испытуемого на задание от его математического ожидания согласно модели [1]. Было показано, что **тест может быть признан существенно одномерным**. Это означает, что тест измеряет единственную латентную характеристику испытуемых – математическую компетентность участников тестирования.

Далее было проанализировано согласие экспериментальных данных тестирования с используемой моделью измерения. В табл. 2.5 представлены статистические данные по заданиям теста. Задания расположены в порядке их следования в тесте.

Таблица 2.5. Статистические показатели заданий теста

Номер задания	Тип задания	Оценка трудности	Ошибка измерения	Коэффициент корреляции	Статистики согласия	
					Взвешенная	Простая
1	M-C-01-1-1	-2.36	0.07	0.24	1.06	1.32
2	M-C-01-1-2	0.01	0.04	0.51	0.94	0.90
3	M-C-01-1-3	2.73	0.06	0.45	0.95	0.87
4	M-C-03-1-1	-3.80	0.13	0.15	1.03	1.09
5	M-C-03-1-2	-0.58	0.05	0.40	1.05	1.10
6	M-C-03-1-3	1.92	0.05	0.36	1.14	1.24
7	M-M-02-1-1	-1.84	0.06	0.38	0.93	0.79
8	M-M-02-1-2	-0.61	0.05	0.46	0.97	0.91
9	M-M-02-1-3	1.36	0.05	0.48	1.00	1.02
10	M-M-03-1-1	-1.68	0.06	0.34	1.02	0.96
11	M-M-03-1-2	-0.15	0.05	0.52	0.93	0.85
12	M-M-03-1-3	1.01	0.04	0.49	1.00	1.00
13	M-M-06-1-1	-0.62	0.05	0.47	0.94	0.91
14	M-M-06-1-2	0.84	0.04	0.55	0.90	0.87
15	M-M-06-1-3	1.01	0.04	0.43	1.07	1.09
16	M-M-11-1-1	-1.87	0.06	0.29	1.04	1.28
17	M-M-11-1-2	0.43	0.04	0.57	0.86	0.80
18	M-M-11-1-3	2.41	0.05	0.45	0.98	0.93
19	M-R-02-1-1	-0.52	0.05	0.54	0.86	0.76
20	M-R-02-1-2	0.65	0.04	0.55	0.91	0.86
21	M-R-02-1-3	1.31	0.05	0.44	1.05	1.14
22	M-R-05-1-1	-4.17	0.16	0.10	1.03	1.41
23	M-R-05-1-2	-0.30	0.05	0.42	1.05	1.04
24	M-R-05-1-3	2.74	0.06	0.42	0.98	1.07
25	M-G-01-1-1	-1.34	0.06	0.29	1.11	1.43
26	M-G-01-1-2	0.84	0.04	0.46	1.03	1.07
27	M-G-01-1-3	2.56	0.05	0.34	1.08	1.53
28	M-D-03-1-1	-1.14	0.05	0.31	1.10	1.46
29	M-D-03-1-2	0.47	0.04	0.48	1.01	1.01
30	M-D-03-1-3	1.54	0.05	0.51	0.96	1.00
31	M-D-05-1-1	-1.20	0.05	0.43	0.93	0.96
32	M-D-05-1-2	-0.76	0.05	0.40	1.06	1.04
33	M-D-05-1-3	1.79	0.05	0.43	1.05	1.28
34	M-D-08-1-1	-2.85	0.09	0.26	0.95	1.53
35	M-D-08-1-2	-0.19	0.05	0.46	1.00	0.97
36	M-D-08-1-3	2.63	0.05	0.31	1.11	1.52
37	M-R-03-1-1	-1.80	0.06	0.34	0.97	1.32
38	M-R-03-1-2	0.01	0.04	0.39	1.12	1.15
39	M-R-03-1-3	1.40	0.05	0.41	1.08	1.18
40	M-C-05-1-1	-1.28	0.06	0.36	1.03	1.10
41	M-C-05-1-2	0.26	0.04	0.49	0.97	1.00
42	M-C-05-1-3	0.84	0.05	0.58	0.86	0.82

43	М-М-08-1-1	-1.96	0.07	0.36	0.94	0.90
44	М-М-08-1-2	0.87	0.04	0.57	0.88	0.83
45	М-М-08-1-3	1.38	0.05	0.57	0.87	0.81

Из анализа таблицы следует, что в тесте нет заданий, которые не были достигнуты испытуемыми (т.е. не были выполнены ввиду недостатка времени). В столбцах 3 и 4 таблицы представлены оценки трудности заданий и соответствующие ошибки измерения (в логитах). В следующем столбце приведен коэффициент корреляции между баллами по заданию и уровнем подготовленности испытуемых. Этот показатель интепретируется так же, как коэффициенты дискриминативности в классической теории тестирования. Только два задания имеют коэффициент корреляции ниже значения 0,2 (задания № 4 и 22), это задания первого уровня, они имеют пониженный уровень трудности.

В последних двух столбцах табл. 2.5 приведены значения статистик согласия, характеризующих согласие данных с используемой моделью измерения. В данном отчете используются две версии статистики согласия – простая и взвешенная. Простая статистика согласия более чувствительна к экстремально неожиданным ответам, когда сильный испытуемый неожиданно неправильно отвечает на легкое задание, или наоборот, слабый испытуемый неожиданно правильно отвечает на трудное задание. Взвешенная версия статистики согласия позволяет уменьшить влияние экстремально неожиданных ответов. Таким образом, эти две статистики относятся к различным частям распределения уровней подготовленности участников тестирования. В литературе продуктивным для измерений называется интервал значений статистик согласия (0,5; 1,5), а в случае широкомасштабных тестирований с важными последствиями – интервал (0,8; 1,2) [2]. Именно последний интервал и рассматривался в данном исследовании в качестве допустимого для статистик согласия.

Из анализа табл. 2.5 следует, что значения взвешенной статистики согласия находятся в норме для всех заданий теста. У ряда заданий значение простой статистики согласия превышает правое критическое значение 1,2 (такие задания выделены в таблице розовым цветом). Это означает, что есть испытуемые, давшие неожиданные ответы на эти задания.

Таким образом, практически все задания теста находятся в удовлетворительном согласии с используемой моделью измерения. Задания, имеющие значение простой статистики согласия выше критического, будут рассмотрены ниже. Но сначала приведем пример хорошего задания, функционирующего в идеальном согласии с моделью – задания №2. Это задание 2-го уровня, средней трудности. На рис. 2.7 приведена характеристическая кривая этого задания, представляющая собой вероятность правильного ответа на задание в зависимости от уровня подготовленности испытуемых (красная линия). Крестиками на рисунке обозначены точки эмпирического распределения ответов испытуемых на данное задание. Они представляют собой средний балл по данному заданию по группам испытуемых (вся выборка была поделена на 10 частей в зависимости от балла по тесту). Также на рисунке показаны границы 95%-ного доверительного интервала для точек эмпирического распределения.

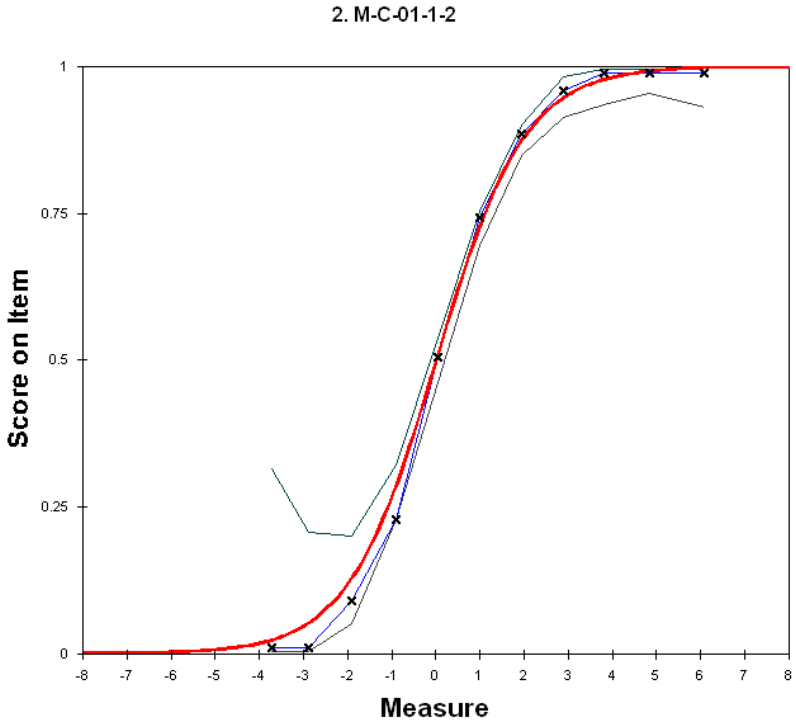

Рис.2.7. Характеристическая кривая задания М-С-01-1-2

Для данного задания мы видим, что все точки эмпирического распределения не выходят за границы доверительного интервала, что говорит о том, что различия между модельными ожиданиями и эмпирическими данными не являются значимыми, т.е. о хорошем согласии ответов испытуемых на данное задание с используемой моделью. Это же справедливо для абсолютного большинства заданий теста.

Для сравнения рассмотрим более подробно одно из проблемных заданий, например, задание № 36, значение простой статистики согласия которого равно 1,52 – одно из самых высоких среди всех заданий. Характеристическая кривая этого задания приведена на рис.2.8. Это задание очень трудное, 3-го уровня. Только 19% участников тестирования выполнили его правильно. При этом, как видно из рисунка, есть испытуемые, которые выполнили это задание правильно, хотя вероятность этого была очень низка. Однако таких испытуемых немного, и в целом задание находится в удовлетворительном согласии с моделью.

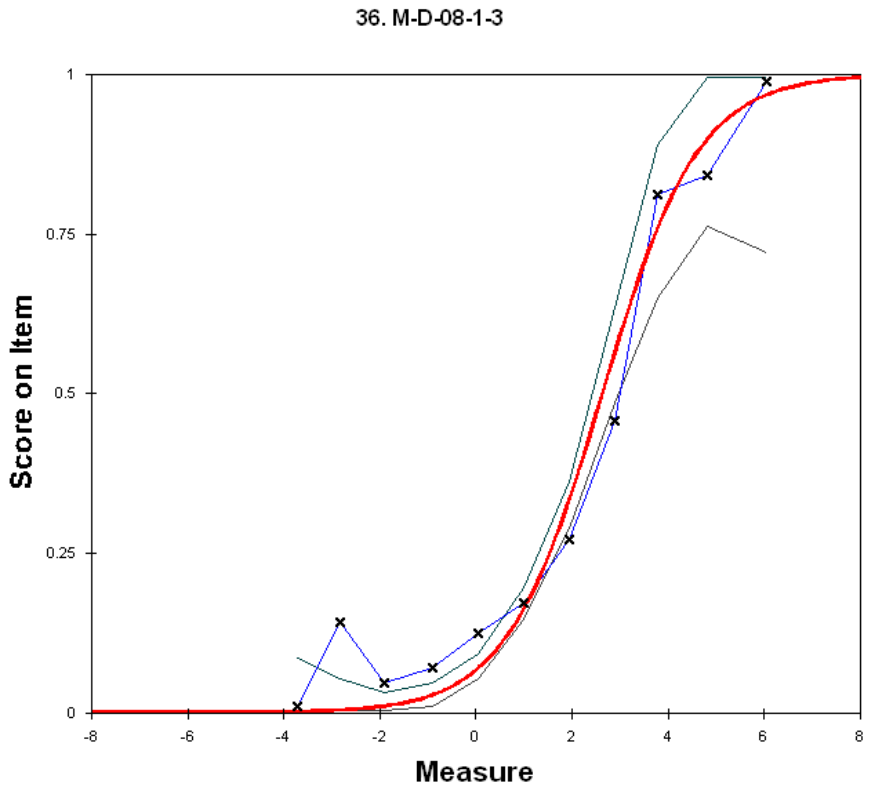

Рис.2.8. Характеристическая кривая задания M-D-08-1-3

На рис. 2.9 представлена карта переменных, на которой показано распределение испытуемых и заданий относительно друг друга на общей метрической шкале. Слева на рисунке находится шкала логитов. На карте испытуемые представлены слева, задания – справа. Более трудные задания и более подготовленные испытуемые находятся в верхней части карты, более легкие задания и менее подготовленные испытуемые находятся в нижней части карты. Карта позволяет проанализировать совместное распределение заданий относительно данного контингента испытуемых и диагностировать проблемы теста.

```
        PERSON - MAP - TASKS
            <more>|<rare>
5             .  +
             .#  |
                 |
                 |
              .  |
              .  |
4            .#  +
              .  |
              .  |
             .#  |
              . T|T
             .#  |
3          .###  +
              .  |
           .###  |  M-C-01-1-3  M-D-08-1-3  M-R-05-1-3
           .###  |  M-G-01-1-3
            .##  |  M-M-11-1-3
         .###### S|
2         ######  +  M-C-03-1-3
        .#######  |  M-D-05-1-3
        .#######  |S
       .########  |  M-D-03-1-3
        .#######  |  M-M-02-1-3  M-M-08-1-3  M-R-02-1-3  M-R-03-1-3
   .############  |
1      #########  +  M-M-03-1-3  M-M-06-1-3
```

Задания 3-го

```
        .######## M|  M-C-05-1-3  M-G-01-1-2  M-M-06-1-2  M-M-08-1-2
         .#######  |  M-R-02-1-2
        .########  |  M-D-03-1-2  M-M-11-1-2
      .##########  |  M-C-05-1-2
          .######  |
 0         .#####  +M M-C-01-1-2  M-R-03-1-2
          .######  |  M-D-08-1-2  M-M-03-1-2
        .########  |  M-R-05-1-2
           .##### S|  M-C-03-1-2  M-R-02-1-1
            .####  |  M-M-02-1-2  M-M-06-1-1
            .####  |  M-D-05-1-2
-1            .##  +
             .###  |  M-D-03-1-1  M-D-05-1-1
              .##  |  M-C-05-1-1  M-G-01-1-1
              .##  |
               .#  |S M-M-03-1-1
               .# T|  M-M-02-1-1  M-M-11-1-1  M-R-03-1-1
-2              .  +  M-M-08-1-1
                .  |
                .  |  M-C-01-1-1
                .  |
                .  |
                   |  M-D-08-1-1
-3              .  +
                .  |
                   |T
                .  |
                .  |
                   |  M-C-03-1-1
-4                 +
                   |  M-R-05-1-1
                   |
                   |
                   |
                   |
-5              .  +
            <less>|<frequ>
```

Задания 2-го

Задания 1-го

Рис. 2.9. Карта переменных (математика, вариант 1)

Как можно заметить, распределение мер испытуемых близко к нормальному. Выборка испытуемых смещена вверх относительно выборки заданий, что означает, что тест оказался довольно легким для данного контингента. Стоит обратить внимание на довольно большой размах мер испытуемых. Некоторые задания 1-го уровня оказались очень легкими, в то время, как в верхней части карты не хватает трудных заданий 3-го уровня для испытуемых с высоким уровнем подготовленности.

Все задания 1-го уровня находятся в нижней части карты, далее следуют задания 2-го уровня, и далее – задания 3-го уровня, самые трудные в тесте.

В целом можно рекомендовать усложнить некоторые задания 1-го и 3-го уровней (находящиеся в нижней и верхней частях карты), что позволит более центрировать тест относительно контингента испытуемых и тем самым снизить ошибку измерения слабых и сильных испытуемых.

Аналогичный анализ был проведен **для варианта 2 по математике** (данные анализа в отчете не приводятся): все показатели в норме, все выводы аналогичны.

В заключение было проведено **выравнивание параметров** (мер испытуемых и трудностей заданий) и перенос всех параметров на единую шкалу. Для проведения процедуры выравнивания в оба варианта теста были включены общие задания в количестве 6 штук (2 общих блока, задания 7-9 и 22-24), что составляет 14% от общего числа заданий. В табл.2.6 приведены характеристики общих заданий в рамках классической теории тестирования.

Табл. 2.6. Характеристики общих заданий в двух вариантах теста по математике

Номер задания в тесте	Код задания	**Вариант 1**		**Вариант 2**	
		Коэффи-циент решаемости	Показатель дискримина-тивности	Коэффи-циент решаемости	Показатель дискримина-тивности
7	M-M-02-1-1	0.88	0.31	0.87	0.32
8	M-M-02-1-2	0.73	0.51	0.76	0.51
9	M-M-02-1-3	0.33	0.53	0.35	0.55
22	M-R-05-1-1	0.97	0.05	0.97	0.06
23	M-R-05-1-3	0.68	0.49	0.66	0.50
24	M-R-05-1-2	0.16	0.34	0.13	0.26

Показатели общих заданий в двух вариантах очень близки и удовлетворяют критериям качества, а также демонстрируют хорошее согласие с моделью измерения. Исключение составляет задание №22, которое ранее отмечалось, как задание экстремальной легкости – в обоих вариантах его выполнили 97 % учащихся. Вследствие этого, это задание обладает очень низкой дискриминативностью.

Для создания общей шкалы был выбран метод отдельной калибровки с фиксацией общих параметров и отражением всех параметров на общей шкале [3]. В качестве единой шкалы была выбрана шкала варианта 1.

На рис.2.10 показаны характеристические кривые двух вариантов теста по математике. Мы видим, что характеристические кривые практически совпадают, что говорит об эквивалентности мер испытуемых, полученных в обоих вариантах, т.е. об успешности процедуры выравнивания. Погрешность выравнивания может быть оценена статистически, однако в данном случае в этом нет необходимости.

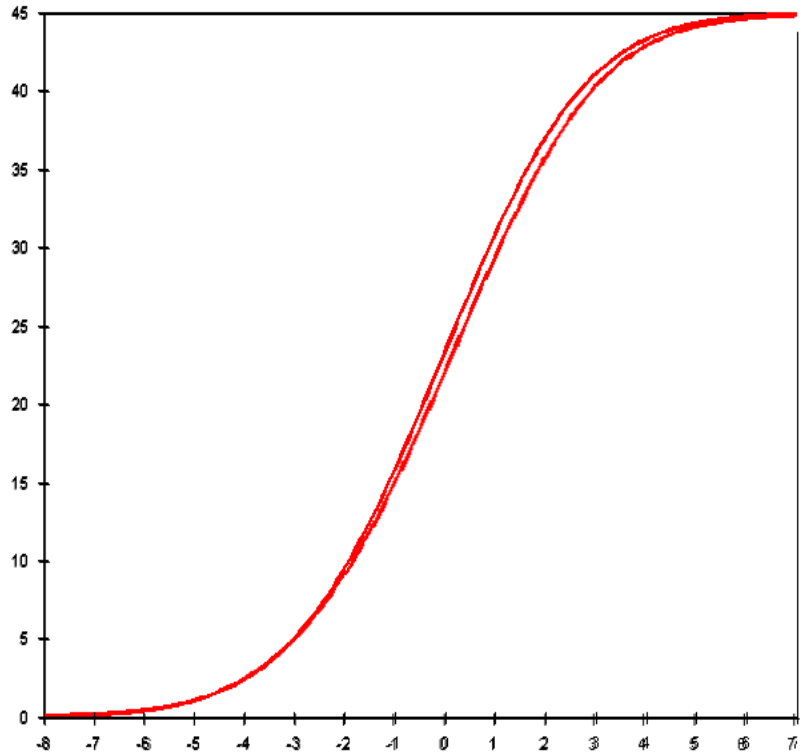

Рис. 2.10. Характеристические кривые двух вариантов теста по математике после процедуры выравнивания

Таким образом, анализ теста по математике в рамках классической и современной теорий тестирования позволяет сделать вывод, что **тест представляет собой достаточно качественный измерительный инструмент и может использоваться для оценивания предметных образовательных результатов по математике.**

2.2. Результаты статистического анализа данных апробации теста по русскому языку

В апробации были использованы два варианта теста по русскому языку. Тест состоял из 45 заданий, объединенных в 15 блоков. В табл. 2.7 представлены суммарные статистические показатели по двум вариантам теста **в рамках классической теории тестирования**, на рис.2.11 - гистограммы распределения первичных баллов участников тестирования.

Табл. 2.7. Суммарные результаты теста по русскому языку

	1 вариант	2 вариант
Число испытуемых	2980	2967
Средний балл	23,35	23,13
Стандартное отклонение	8,74	8,70
Коэффициент асимметрии	-0,14	-0,19
Коэффициент эксцесса	-0,68	-0,69
Средний коэффициент решаемости	0,52	0,51
Средний индекс дискриминативности	0,46	0,47
Средний скор. коэф. точ.-бис. корреляции	0,39	0,39
Показатель надежности (KR20)	0,90	0,90
Стандартная ошибка измерения	2,70	2,70

1 вариант 2 вариант

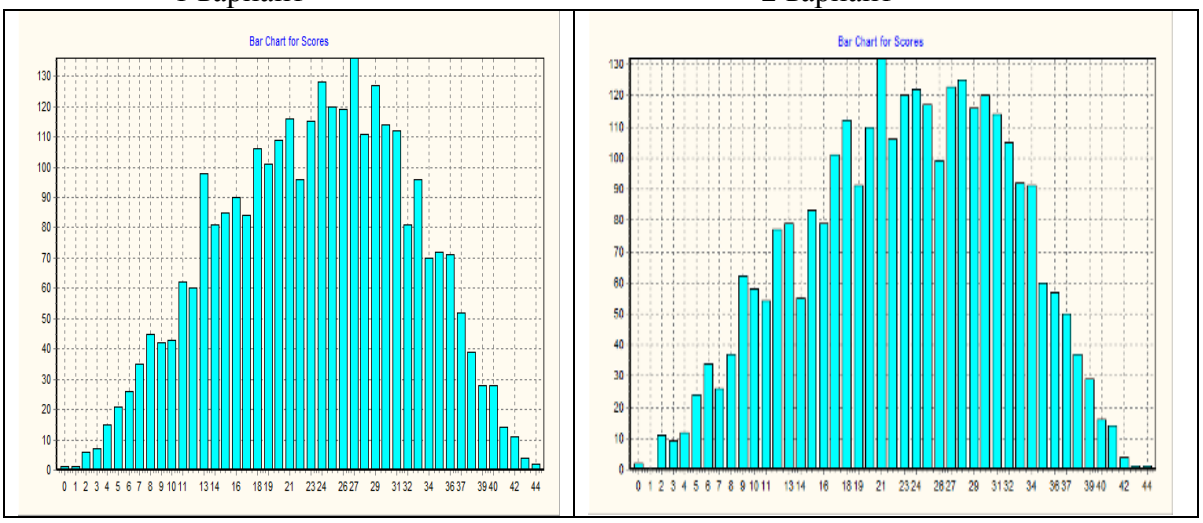

Рис. 2.11. Гистограммы распределения первичных баллов участников тестирования

Из анализа таблицы следует, что статистические показатели двух вариантов очень близки. Дополнительно была проверена гипотеза о параллельности вариантов (методом хи-квадрат), и было показано, что варианты могут быть признаны параллельными. Показатели надежности имеют достаточно высокие значения – 0,9 по обоим вариантам.

Средние коэффициенты решаемости и дискриминативности близки к оптимальным значениям.

В табл. 2.8 представлена детальная информация по заданиям **варианта 1**. При классическом анализе основными характеристиками задания являются его трудность и дискриминативность. Трудность задания (коэффициент решаемости) определяется как доля испытуемых выборки, выполнивших задание правильно. Чем выше этот показатель, тем задание легче.

Дискриминативность характеризует дифференцирующую способность задания, т.е. его способность различать испытуемых с различным уровнем подготовки. В отчете задействованы два коэффициента дискриминативности: классический показатель дискриминативности (разность между коэффициентами решаемости задания двумя группами испытуемых - 27 % лучших (испытуемых, имеющих высокие баллы) и 27 % худших (испытуемых, имеющих низкие баллы)) и скорректированный коэффициент точечной бисериальной корреляции (коэффициент корреляции между баллом по заданию и общим баллом по тесту после удаления результатов по этому заданию). В качестве критического значения коэффициентов дискриминативности выбрано значение 0,2.

Из анализа таблицы 2.8 можно заключить, что все задания анализируемого теста хорошо функционируют. Одно задание первого уровня оказалось очень легким для данного контингента испытуемых (коэффициент решаемости 0,96, что означает, что это задание выполнилит более 96% испытуемых) и 1 задание оказалось очень трудным (коэффициент решаемости 0,07, что означает, что это задание выполнили 7% испытуемых), эти задания в таблице выделены зеленым цветом. Эти же два задания обладают низкой дискриминативностью (задания с низкой дискриминативностью выделены в таблице розовым цветом). Именно экстремальный уровень легкости и трудности этих заданий обуславливает их низкий уровень дифференцирующей способности.

Все остальные задания теста демонстрирует хорошие статистические показатели. Заметим, что внутри каждого блока из трех заданий 1-го, 2-го и 3-го уровней наблюдается иерархия по трудности. Исключение составляет блок заданий 14-16, в котором задание 3-го уровня оказалось слегка легче задания 2-го уровня (эти задания выделены в таблоице синим цветом).

Табл. 2.8. Статистические показатели функционирования заданий (русский язык, вариант 1)

№		Коэф. решаемости	Индекс дискриминативности	Скорректир. точ-бис. корр.
1	L-O-1-01-1-1	0.85	0.33	0.39
2	L-O-1-01-1-2	0.80	0.33	0.32
3	L-O-1-01-1-3	0.36	0.56	0.44
4	L-O-2-01-1-1	0.96	0.10	0.23
5	L-O-2-01-2-2	0.68	0.53	0.43
6	L-O-2-01-1-3	0.36	0.48	0.38
7	L-M-3-01-1-1	0.75	0.39	0.33
8	L-M-3-01-1-2	0.39	0.53	0.41
9	L-M-3-01-1-3	0.28	0.49	0.41
10	L-M-2-01-1-1	0.74	0.35	0.29
11	L-M-2-01-1-2	0.68	0.56	0.48
12	L-M-2-01-1-3	0.24	0.37	0.33

13	L-M-1-02-1-1	0.80	0.37	0.36
14	L-M-1-02-1-2	0.33	0.53	0.42
15	L-M-1-02-1-3	0.43	0.56	0.43
16	L-F-1-02-1-1	0.82	0.36	0.39
17	L-F-1-02-1-2	0.42	0.58	0.45
18	L-F-1-02-1-3	0.32	0.60	0.48
19	L-F-2-02-1-1	0.79	0.46	0.45
20	L-F-2-02-1-2	0.61	0.33	0.25
21	L-F-2-02-1-3	0.35	0.35	0.27
22	L-F-2-01-1-1	0.85	0.31	0.35
23	L-F-2-01-1-2	0.50	0.6	0.46
24	L-F-2-01-1-3	0.28	0.55	0.47
25	L-F-2-03-1-1	0.67	0.59	0.5
26	L-F-2-03-1-2	0.57	0.76	0.58
27	L-F-2-03-1-3	0.21	0.38	0.37
28	L-L-3-01-1-1	0.88	0.28	0.37
29	L-L-3-01-1-2	0.65	0.5	0.4
30	L-L-4-01-1-3	0.29	0.42	0.35
31	L-L-2-01-1-1	0.56	0.58	0.45
32	L-L-2-01-1-2	0.38	0.48	0.35
33	L-L-2-01-1-3	0.23	0.32	0.29
34	L-L-1-01-1-1	0.63	0.45	0.35
35	L- L-1-01-1-2	0.30	0.44	0.36
36	L-L-1-01-1-3	0.07	0.11	0.19
37	L-L-5-01-1-1	0.75	0.46	0.41
38	L-L-5-01-1-2	0.38	0.56	0.43
39	L-L-5-01-1-3	0.20	0.41	0.39
40	L-S-2-01-1-1	0.77	0.47	0.45
41	L-S-2-01-1-2	0.54	0.64	0.5
42	L-S-2-01-1-3	0.26	0.53	0.46
43	L-S-1-01-1-1	0.75	0.49	0.46
44	L-S-1-01-1-2	0.34	0.47	0.37
45	L-S-1-01-1-3	0.33	0.61	0.5

На рисунках ниже показатели из табл. 2.8 представлены графически.

На рис.2.12 представлено распределение коэффициентов решаемости заданий анализируемого теста в виде столбчатой диаграммы. На оси абсцисс расположены коды заданий (в порядке следования в тесте), по оси ординат – коэффициенты решаемости. Задания 1-го уровня выделены синим цветом, задания 2-го уровня – красным, и задания 3-го уровня – зеленым. На диаграмме хорошо видна иерархия по трудности внутри каждого блока. На рис.2.13 также показано распределение коэффициентов решаемости заданий, но в виде графика.

На рис. 2.14-2.15 показаны распределения коэффициентов дискриминативности заданий теста. И наконец, на рис. 2.16 показано совместное распределение коэффициентов решаемости и показателей дискриминативности заданий.

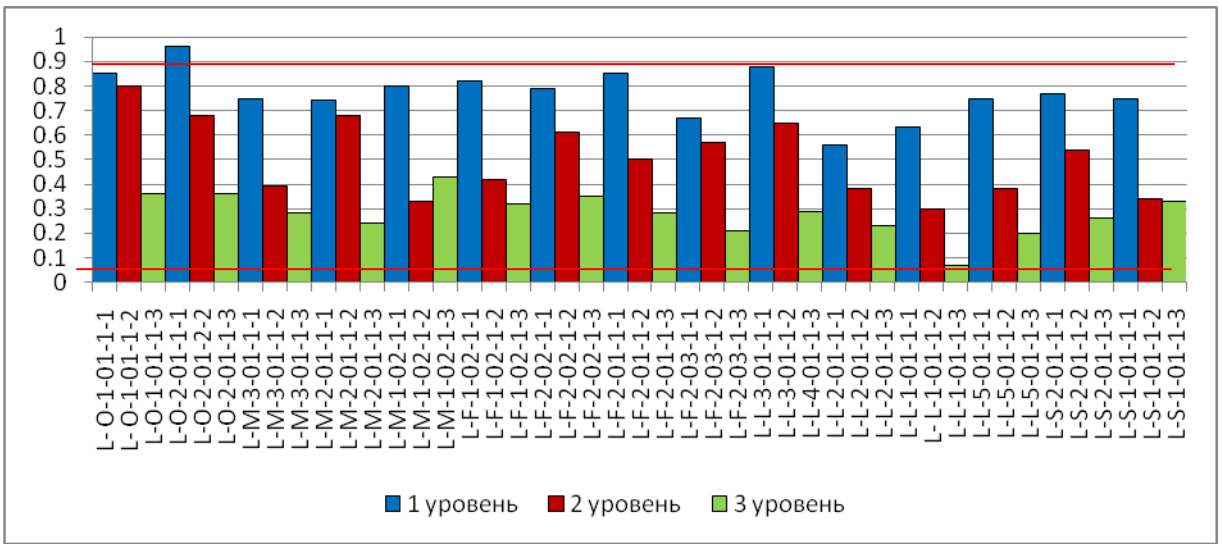

Рис.2.12. Диаграмма распределения коэффициентов решаемости (русский язык, вариант 1)

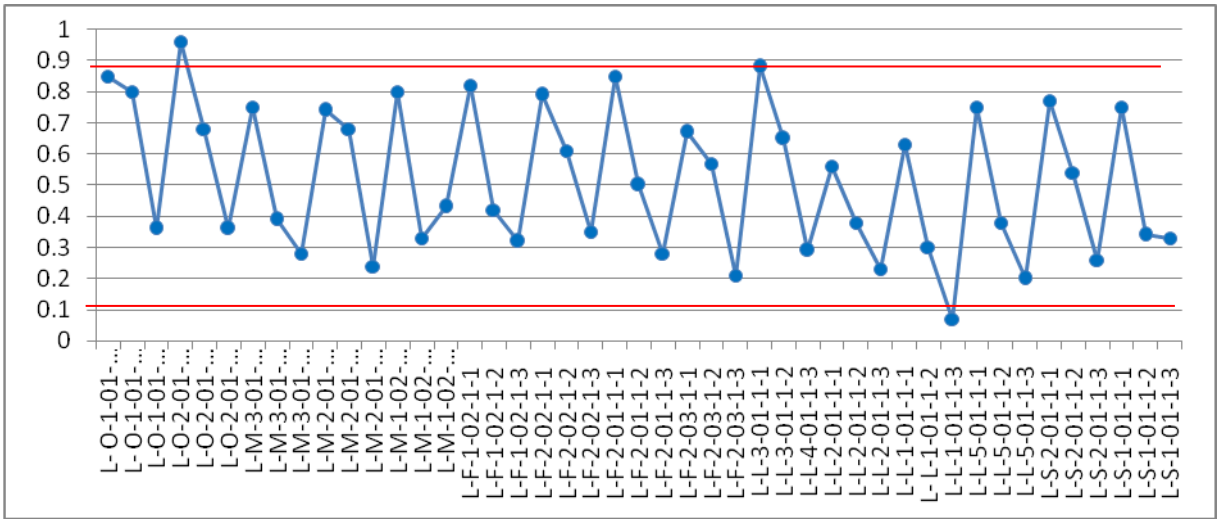

Рис. 2.13. График распределения коэффициентов решаемости (русский язык, вариант 1)

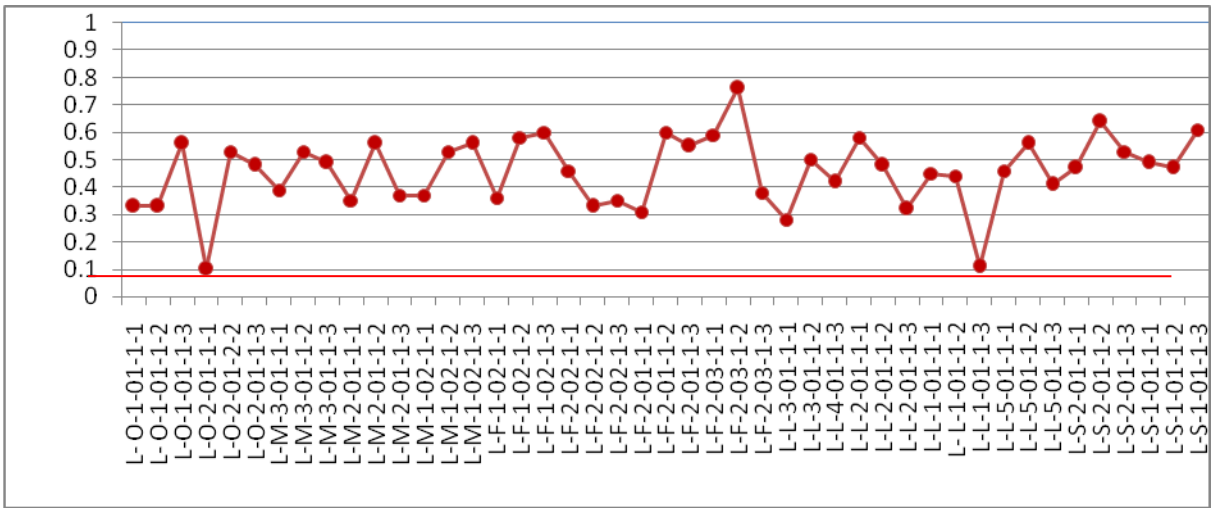

Рис.2.14. График распределения показателей дискриминативности заданий (русский язык, вариант 1)

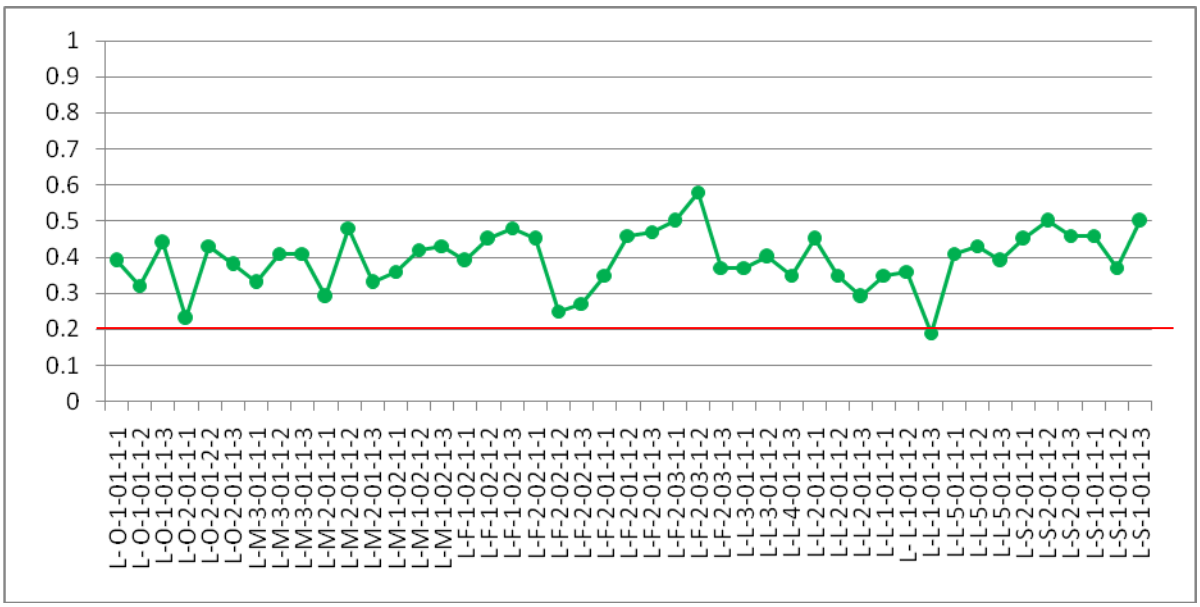

Рис.2.15. График распределения коэффициентов точечно-бисериальной корреляции (русский язык, вариант 1)

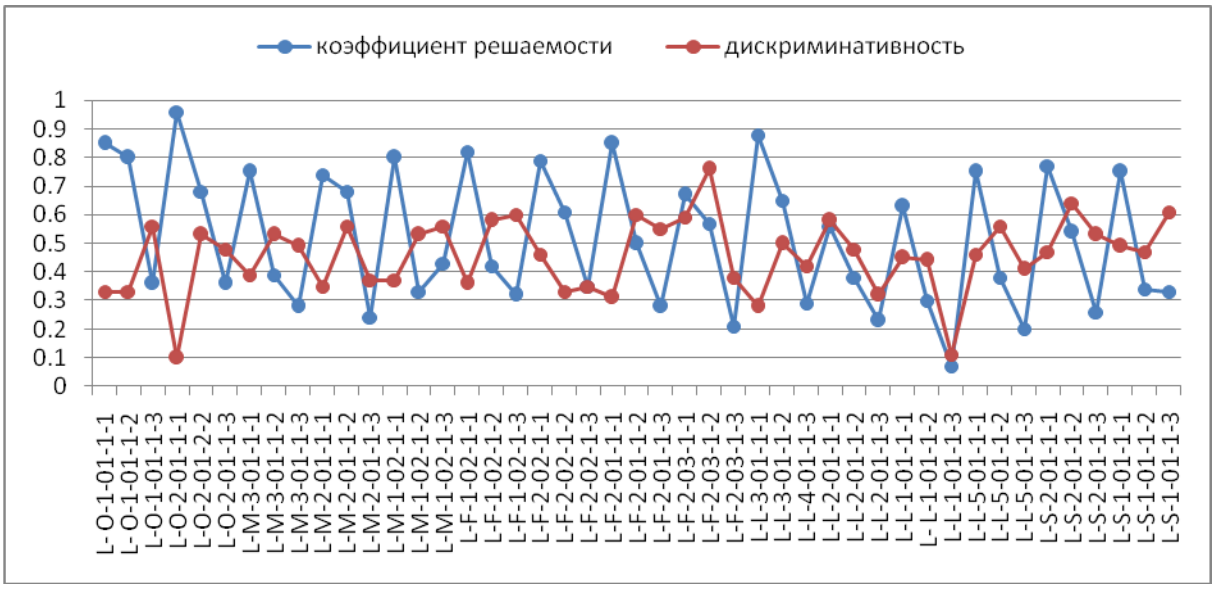

Рис.2.16. Совместное распределение коэффициентов решаемости и показателей дискриминативности (русский язык, вариант 1)

На рисунках хорошо видно, что практически все задания имеют хорошие показатели – оптимальную трудность и высокую дискриминативность. Только 2 задания имеют показатели дискриминативности ниже нормы – самое легкое задание и самое трудное задание теста, уже отмеченные выше.

Дополнительно был проведен **анализ функционирования дистракторов** в заданиях закрытой формы (с выбором ответа из нескольких предложенных). В данном тесте таких заданий было 14, причем некоторые предполагали выбор более одного правильного ответа.

Для анализа дистракторов, во-первых, было проанализировано распределение ответов испытуемых по всем ответным альтернативам заданий закрытого типа с целью определения неработающих дистракторов. Во-вторых, были вычислены коэффициенты корреляции между дистракторами и общим баллом по тесту (точечно-бисериальный коэффициент корреляции). Дистрактор функционирует правильно, если испытуемые с

высоким уровнем подготовки его не выбирают в качестве правильного ответа. В этом случае коэффициент корреляции будет отрицательным (желательно, чтобы он был меньше -0,2). И, наоборот, коэффициент корреляции для правильного варианта ответа должен быть положительным (коэффициент корреляции для правильного ответа совпадает с коэффициентом дискриминативности заданий, рассмотренным ранее).

Серьезных проблем в функционировании дистракторов закрытых заданий анализируемого теста диагностировано не было.

В табл. 2.8 и 2.9 показаны примеры анализа дистракторов для двух заданий. Правильный вариант ответа помечен звездочкой и выделен цветом. Для задания L-M-3-01-1-1 все показатели находятся в пределах нормы: 75 % участников тестирования выполнили задание правильно. Высокий положительный коэффициент корреляции говорит о том, что это сильные учащиеся, имеющие высокие баллы по тесту. Остальные учащиеся, не выполнившие задание, распределились по трем дистракторам (3% участников это задание пропустили). Все дистракторы при этом работают – каждый выбрало не менее 5% испытуемых. Отрицательные коэффициенты корреляции говорят о том, что это слабые учащиеся, имеющие низкие баллы по тесту.

Таблица 2.8. Анализ дистракторов задания L-M-3-01-1-1

Задание L-M-3-01-1-1				
Варианты ответов	**А**	**Б**	**В***	**Г**
Распределение ответов испытуемых	8%	7%	75%	7%
Коэффициент корреляции	-0,15	-0,15	0,33	-0,19

Таблица 2.9. Анализ дистракторов задания L- F-2-02-1-2

Задание L- F-2-02-1-2				
Варианты ответов	**А**	**Б**	**В***	**Г**
Распределение ответов испытуемых	17%	8%	61%	12%
Коэффициент корреляции	-0,09	-0,29	0,25	0,01

Задание L-F-2-02-1-2 более трудное: его выполнили 61% испытуемых и 2% его пропустили. Учащиеся, не выполнившие задание, распределились по трем дистракторам, при этом все дистракторы работают – каждый выбрало не менее 5% испытуемых. Дистракторы А и Г являются более привлекательными для учащихся, и их корреляция с общим баллом близка к нулю, что означает, что эти дистракторы являются привлекательными, как для слабых испытуемых, так и для испытуемых с хорошей подготовкой. Такая ситуация бывает характерна для заданий, содержащих подвох, который не замечают сильные учащиеся.

Таким образом, анализ заданий варианта 1 по русскому языку в рамках классической теории тестирования показал, что все задания имеют хорошие показатели.

Более тщательный анализ теста был проведен **в рамках современной теории тестирования IRT.** Ниже приведены основные результаты этого анализа.

В качестве модели тестирования используется одномерная дихотомическая модель Раша. В соответствии с этой моделью каждое задание теста характеризуется одним параметром – трудностью, каждый участник тестирования характеризуется одним параметром – уровнем подготовленности. Оценки всех параметров – и участников

тестирования и заданий теста - находятся на общей метрической шкале (единица измерения на этой шкале называется логит) и снабжены характеристиками точности оценивания. Начало этой шкалы не определено и может быть выбрано произвольно. С целью фиксирования начала отсчета среднее значение оценок трудности всех заданий положено равным 0.

В первую очередь было проведено исследование размерности теста. Для этого применялся факторный анализ нормированных уклонений ответа испытуемого на задание от его математического ожидания согласно модели [1]. Было показано, что **тест может быть признан существенно одномерным**. Это означает, что тест измеряет единственную латентную характеристику испытуемых – языковую компетентность участников тестирования.

Далее было проанализировано согласие экспериментальных данных тестирования с используемой моделью измерения. В табл. 2.10 представлены статистические данные по заданиям теста. Задания расположены в порядке их следования в тесте.

Табл. 2.10. Статистические показатели заданий теста

Номер задания	Тип задания	Оценка трудности	Ошибка измерения	Коэффициент корреляции	Статистики согласия	
					Взвешенная	Простая
1	L-O-1-01-1-1	-1.93	0.06	0.42	0.94	0.82
2	L-O-1-01-1-2	-1.57	0.05	0.37	1.05	1.07
3	L-O-1-01-1-3	0.97	0.04	0.47	0.97	0.94
4	L-O-2-01-1-1	-3.54	0.10	0.28	0.95	0.76
5	L-O-2-01-2-2	-0.72	0.05	0.47	0.97	0.96
6	L-O-2-01-1-3	0.98	0.04	0.42	1.03	1.05
7	L-M-3-01-1-1	-1.19	0.05	0.37	1.09	1.07
8	L-M-3-01-1-2	0.87	0.04	0.46	0.99	0.97
9	L-M-3-01-1-3	1.31	0.05	0.44	0.98	0.97
10	L-M-2-01-1-1	-1.05	0.05	0.33	1.13	1.22
11	L-M-2-01-1-2	-0.71	0.05	0.51	0.91	0.87
12	L-M-2-01-1-3	1.70	0.05	0.37	1.05	1.07
13	L-M-1-02-1-1	-1.66	0.05	0.40	0.99	0.96
14	L-M-1-02-1-2	1.21	0.04	0.46	0.96	0.92
15	L-M-1-02-1-3	0.40	0.04	0.45	1.01	0.99
16	L-F-1-02-1-1	-1.70	0.06	0.42	0.96	0.91
17	L-F-1-02-1-2	0.72	0.04	0.49	0.95	0.94
18	L-F-1-02-1-3	1.17	0.04	0.53	0.88	0.82
19	L-F-2-02-1-1	-1.44	0.05	0.48	0.92	0.79
20	L-F-2-02-1-2	-0.34	0.04	0.31	1.19	1.38
21	L-F-2-02-1-3	0.70	0.05	0.27	1.24	1.33
22	L-F-2-01-1-1	-2.11	0.06	0.38	0.97	0.92
23	L-F-2-01-1-2	0.20	0.04	0.49	0.96	0.93
24	L-F-2-01-1-3	1.49	0.05	0.51	0.88	0.84
25	L-F-2-03-1-1	-0.95	0.05	0.50	0.91	0.81
26	L-F-2-03-1-2	-0.07	0.04	0.60	0.82	0.76
27	L-F-2-03-1-3	1.85	0.05	0.40	0.98	1.00
28	L-L-3-01-1-1	-2.59	0.07	0.36	0.93	0.85
29	L-L-3-01-1-2	-0.59	0.04	0.42	1.04	1.06
30	L-L-4-01-1-3	1.26	0.05	0.37	1.08	1.25

31	L-L-2-01-1-1	-0.11	0.04	0.47	0.99	0.98
32	L-L-2-01-1-2	0.84	0.04	0.37	1.10	1.16
33	L-L-2-01-1-3	1.54	0.05	0.32	1.11	1.23
34	L-L-1-01-1-1	-0.89	0.05	0.33	1.14	1.20
35	L-L-1-01-1-2	1.31	0.05	0.40	1.04	1.02
36	L-L-1-01-1-3	3.29	0.08	0.22	1.04	1.26
37	L-L-5-01-1-1	-1.40	0.05	0.39	1.03	1.01
38	L-L-5-01-1-2	0.85	0.04	0.46	0.98	0.97
39	L-L-5-01-1-3	1.81	0.05	0.42	0.97	0.89
40	L-S-2-01-1-1	-1.56	0.05	0.44	0.95	0.83
41	L-S-2-01-1-2	-0.11	0.04	0.49	0.96	0.94
42	L-S-2-01-1-3	1.32	0.05	0.48	0.93	0.90
43	L-S-1-01-1-1	-1.64	0.06	0.39	0.98	0.97
44	L-S-1-01-1-2	1.05	0.04	0.38	1.07	1.09
45	L-S-1-01-1-3	0.99	0.05	0.51	0.92	0.87

Из анализа таблицы следует, что в тесте нет заданий, которые не были достигнуты испытуемыми (т.е. не были выполнены ввиду недостатка времени). В столбцах 3 и 4 таблицы представлены оценки трудности заданий и соответствующие ошибки измерения (в логитах). В следующем столбце приведен коэффициент корреляции между баллами по заданию и уровнем подготовленности испытуемых. Этот показатель интерпретируется так же, как коэффициенты дискриминативности в классической теории тестирования. Только два задания имеют коэффициент корреляции ниже значения 0,2 (задания № 4 и 22), это задания первого уровня, они имеют пониженный уровень трудности.

В последних двух столбцах таблицы 4.9 приведены значения статистик согласия, характеризующих согласие данных с используемой моделью измерения. В данном отчете используются две версии статистики согласия – простая и взвешенная. Простая статистика согласия более чувствительна к экстремально неожиданным ответам, когда сильный испытуемый неожиданно неправильно отвечает на легкое задание, или наоборот, слабый испытуемый неожиданно правильно отвечает на трудное задание. Взвешенная версия статистики согласия позволяет уменьшить влияние экстремально неожиданных ответов. Таким образом, эти две статистики относятся к различным частям распределения уровней подготовленности участников тестирования. В литературе продуктивным для измерений называется интервал значений статистик согласия (0,5; 1,5), а в случае широкомасштабных тестирований с важными последствиями – интервал (0,8; 1,2) [2]. Именно последний интервал и рассматривался в данном исследовании в качестве допустимого для статистик согласия.

Из анализа таблицы 2.10 следует, что значения взвешенной статистики согласия находятся в норме для всех заданий теста, кроме одного. У нескольких заданий значение простой статистики согласия превышает правое критическое значение 1,2 (такие задания выделены в таблице розовым цветом). Это означает, что есть испытуемые, давшие неожиданные ответы на эти задания.

Таким образом, практически все задания теста находятся в удовлетворительном согласии с используемой моделью измерения. Задания, имеющие значение простой статистики согласия выше критического, будут рассмотрены ниже. Но сначала приведем пример хорошего задания, функционирующего в идеальном согласии с моделью – задания №3. Это задание 3-го уровня, трудности выше среднего. На рис. 2.17 приведена характеристическая кривая этого задания, представляющая собой вероятность правильного ответа на задание в зависимости от уровня подготовленности испытуемых (красная линия). Крестиками на рисунке обозначены точки эмпирического распределения ответов испытуемых на данное задание. Они представляют собой средний балл по данному заданию по группам испытуемых (вся выборка была поделена на 10 частей в

зависимости от балла по тесту). Также на рисунке показаны границы 95%-ного доверительного интервала для точек эмпирического распределения.

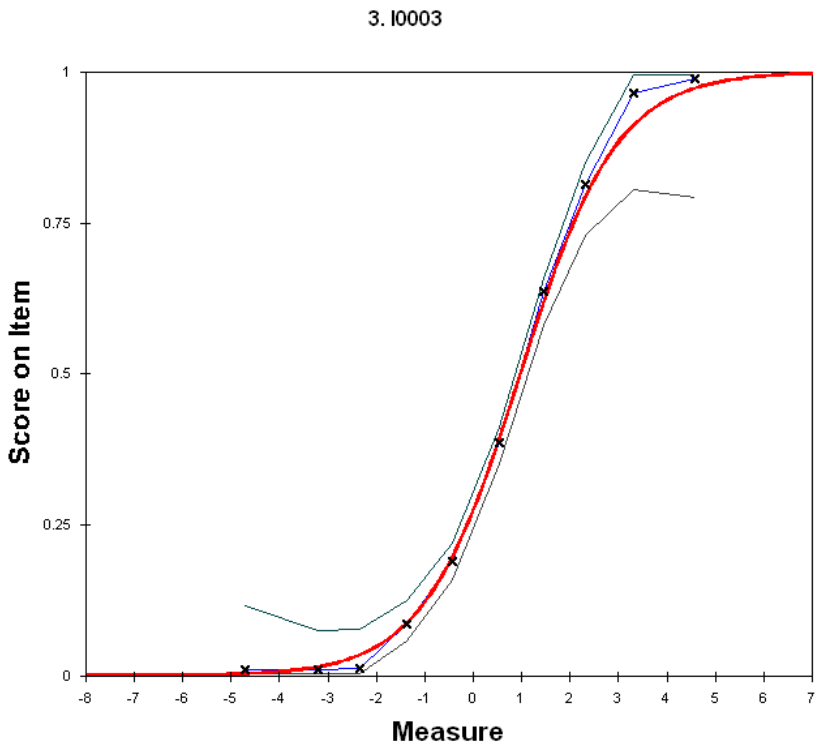

Рис.2.17. Характеристическая кривая задания L-O-1-01-1-3

Для данного задания мы видим, что все точки эмпирического распределения не выходят за границы доверительного интервала, что говорит о том, что различия между модельными ожиданиями и эмпирическими данными не являются значимыми, т.е. о хорошем согласии ответов испытуемых на данное задание с используемой моделью. Это же справедливо для абсолютного большинства заданий теста.

Для сравнения рассмотрим более подробно одно из проблемных заданий, например, задание № 10, значение простой статистики согласия которого равно 1,22. Характеристическая кривая этого задания приведена на рис.2.18. Это задание 1-го уровня, довольно легкое – 74% испытуемых выполнили это задание правильно. При этом, как видно из рисунка, есть слабые испытуемые, которые выполнили это задание правильно, хотя вероятность этого была очень низка. Это задание с выбором одного правильного ответа из 4-х предложенных, поэтому можно предположить, что имело место случайное гадание. Чтобы проверить это предположение, приведем в табл. 2.11 анализ дистаркторов по этому заданию.

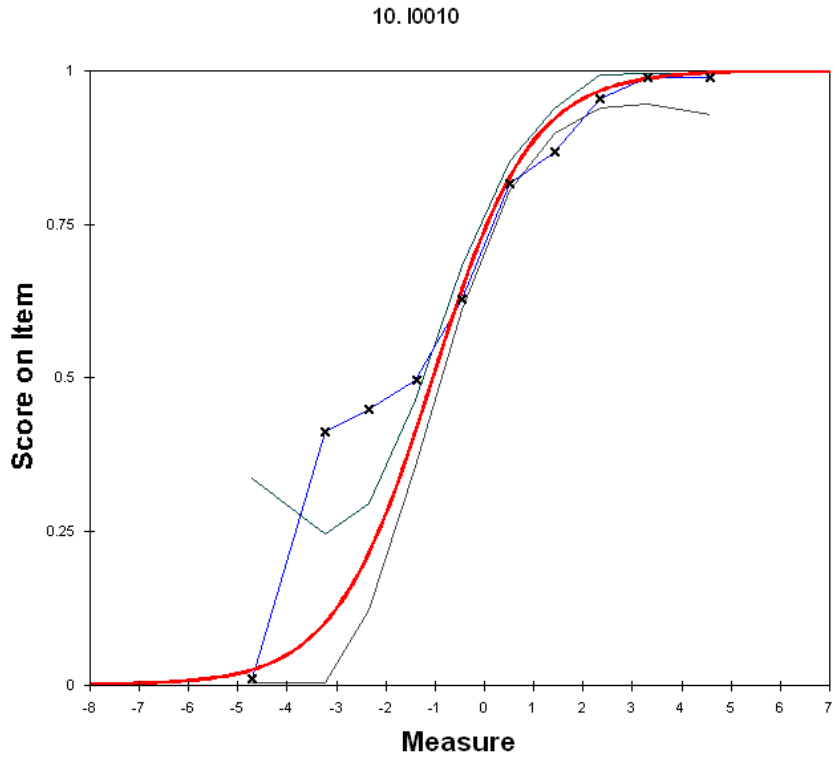

Рис. 2.18. Характеристическая кривая задания L-M-2-01-1-1

Табл. 2.11. Анализ дистракторов задания L-M-2-01-1-1

Задание L-M-2-01-1-1				
Варианты ответов	**А**	**Б***	**В**	**Г**
Распределение ответов испытуемых	10%	74%	2%	12%
Коэффициент корреляции	-0,02	0,29	-0,1	-0,28

Мы видим, что дистрактор В не работает – его выбирают только 2% испытуемых. Довольно высокая положительная корерляция правильного ответа голворит нам, что его выбирали сильные учащиеся. Испытуемые, которые не знали правильного ответа, гадали между двумя вариантами – А и Г. Поэтому вероятность отгадывания правильного овтета слабыми учащимися в этом задании была достаточно высока, что и нашло отражение на рис. 2.19 и в статистиках данного задания.

```
          PERSON - MAP - TASKS
              <more>|<rare>
5                   +
                    |
                .   |
                .   |
                    |
                    |
4                   +
                .   |
                .   |
                .   |
                .   |  L-L-1-01-1-3
                .   |
3               .   +
                .   |T
               .# T|
```

```
                .#  |
               .##  |
                .#  |
 2            .###  +
             .####  |  L-F-2-03-1-3  L-L-5-01-1-3
            .#####  |  L-|-2-01-1-3
       .########## S|S L-F-2-01-1-3  L-L-2-01-1-3
           .######  |  L-L-1-01-1-2  L-L-4-01-1-3  L-S-2-01-1-3  L-|-3-01-1-3
          .#######  |  L-F-1-02-1-3  L-M-1-02-1-2
 1       .########  +  L-S-1-01-1-2  L-S-1-01-1-3  L-+-1-01-1-3  L-+-2-01-1-3
         .########  |  L-L-2-01-1-2  L-L-5-01-1-2  L-M-3-01-1-2
     .############  |  L-F-1-02-1-2  L-F-2-02-1-3
        .#########  |
         .######## M|  L-M-1-02-1-3
     .############  |  L-F-2-01-1-2
 0       .#######  +M L-F-2-03-1-2
         .########  |  L-L-2-01-1-1  L-S-2-01-1-2
          .#######  |  L-F-2-02-1-2
            .#####  |
           .######  |  L-L-3-01-1-2  L-|-2-01-1-2  L-+-2-01-2-2
            .##### S|  L-L-1-01-1-1
-1          .#####  +  L-F-2-03-1-1  L-|-2-01-1-1
            .#####  |  L-M-3-01-1-1
              .###  |  L-L-5-01-1-1
               .##  |S L-F-2-02-1-1  L-S-2-01-1-1  L-+-1-01-1-2
               .##  |  L-F-1-02-1-1  L-M-1-02-1-1  L-S-1-01-1-1
                .#  |
-2               .#  +  L-+-1-01-1-1
                 .# T|  L-F-2-01-1-1
                 .#  |
                  #  |
                  .  |  L-L-3-01-1-1
                  .  |T
-3                .  +
                  .  |
                  .  |
                  .  |  L-+-2-01-1-1
                  .  |
                     |
-4                   +                          EACH "#" IS 17.
                     |                         EACH "." IS 1 TO 16
                     |
                     |
                  .  |
                     |
-5                   +
              <less>|<frequ>
```

Рис. 2.19. Карта переменных (русский язык, вариант 1)

На рис. 2.19 представлена карта переменных, на которой показано распределение испытуемых и заданий относительно друг друга на общей метрической шкале. Слева на рисунке находится шкала логитов. На карте испытуемые представлены слева, задания – справа. Более трудные задания и более подготовленные испытуемые находятся в верхней части карты, более легкие задания и менее подготовленные испытуемые находятся в нижней части карты. Карта позволяет проанализировать совместное распределение заданий относительно данного контингента испытуемых и диагностировать проблемы теста.

Как можно заметить, распределение мер испытуемых близко к нормальному. Выборка испытуемых смещена немного вверх относительно выборки заданий, что означает, что тест оказался довольно легким для данного контингента. Стоит обратить внимание на довольно большой размах мер испытуемых. Некоторые задания 1-го уровня

оказались очень легкими, в то время, как в верхней части карты не хватает трудных заданий 3-го уровня для испытуемых с высоким уровнем подготовленности.

Все задания 1-го уровня находятся в нижней части карты, далее следуют задания 2-го уровня, и далее – задания 3-го уровня, самые трудные в тесте.

В целом можно рекомендовать усложнить некоторые задания 1-го и 3-го уровней (находящиеся в нижней и верхней частях карты), что позволит более центрировать тест относительно контингента испытуемых и тем самым снизить ошибку измерения сильных испытуемых.

Аналогичный анализ был проведен **для варианта 2 по русскому языку** (данные анализа в отчете не приводятся): все показатели в норме, все выводы аналогичны.

В заключение было проведено **выравнивание параметров** (мер испытуемых и трудностей заданий) и перенос всех параметров на единую шкалу. Для проведения процедуры выравнивания в оба варианта теста были включены общие задания в количестве 6 штук (2 общих блока, задания 4-6 и 34-36), что составляет 14% от общего числа заданий. В табл.2.12 приведены характеристики общих заданий в рамках классической теории тестирования.

Таблица 2.12. Характеристики общих заданий в двух вариантах теста по русскому языку

Номер задания в тесте	Код задания	**Вариант 1**		**Вариант 2**	
		Коэффициент решаемости	Показатель дискриминативности	Коэффициент решаемости	Показатель дискриминативности
7	L-O-2-01-1-1	0.96	0.10	0.96	0.10
8	L-O-2-01-2-2	0.68	0.53	0.70	0.52
9	L-O-2-01-1-3	0.36	0.48	0.33	0.50
22	L-L-1-01-1-1	0.63	0.45	0.57	0.51
23	L-L-1-01-1-2	0.30	0.44	0.29	0.43
24	L-L-1-01-1-3	0.07	0.11	0.07	0.12

Показатели общих заданий в двух вариантах очень близки и удовлетворяют критериям качества, а также демонстрируют хорошее согласие с моделью измерения. Исключение составляют задания №7 и №24, которые ранее отмечались, как самое легкое и самое трудное задания теста. Однако психометрические характеристики этих заданий в рамках IRT, в частности, согалсие с моделью, удовлетворительные.

Для создания общей шкалы был выбран метод отдельной калибровки с фиксацией общих параметров и отражением всех параметров на общей шкале [3]. В качестве единой шкалы была выбрана шкала варианта 1.

На рис.2.20 показаны характеристические кривые двух вариантов теста по математике. Мы видим, что характеристические кривые практически совпадают, что говорит об эквивалентности мер испытуемых, полученных в обоих вариантах, т.е. об успешности процедуры выравнивания. Погрешность выравнивания может быть оценена статистически, однако в данном случае в этом нет необходимости.

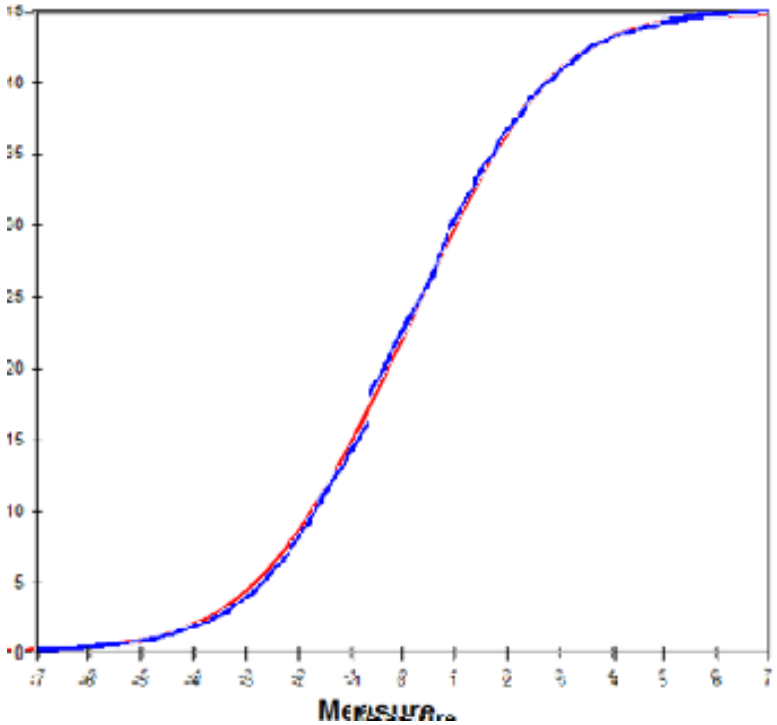

Рис. 2.20. Характеристические кривые двух вариантов теста по русскому языку после процедуры выравнивания

Таким образом, анализ теста по русскому языку в рамках классической и современной теорий тестирования позволяет сделать вывод, что **тест представляет собой достаточно качественный измерительный инструмент и может использоваться для оценивания предметных образовательных результатов по русскому языку.**

Литература

1. *Smith Jr. E. V.* Detecting and Evaluating the Impact of Multidimensionality using Item Fit Statistics and Principal Component Analysis of Residuals // Journal of Applied Measurement. – 2002,Vol. 3, №2, p.205-231
2. *Smith R.M.* Applications of Rasch Measurement. – Chicago: Mesa Press, 1992
3. Карданова Е.Ю., Нейман Ю.М. Проблема выравнивания в современной теории тестирования. *Вопросы тестирования в образовании*, 2003, № 8, с. 21-40.

3. Обоснование валидности SAM

Разработка образовательного теста включает валидизацию, т.е. оценку его адекватности предмету и цели тестирования. Как правило, имеются в виду три основные показателя валидности: валидность по содержанию, конструктная (концептуальная) валидность и критериальная валидность (текущая и прогностическая). Каждый из названных показателей отвечает на свой вопрос. Так, показатель валидности по содержанию характеризует меру адекватности тестовых материалов проверяемой компетенции. Показатель конструктной валидности устанавливает меру согласованности результатов тестирования с психологической моделью тестируемой способности. Наконец, показатель критериальной валидности фиксирует меру согласованности результатов тестирования с известными объективными критериями.

Отдельно стоит рассмотреть валидизацию установленных для интерпретации результатов SAM пороговых оценок. Это исследование описано в главе «Оценивание участников тестирования».

3.1. Валидность по содержанию

Очевидно, что ни один тест не может вбирать в себя содержание всей учебной дисциплины. Содержательная валидность теста характеризует то, насколько оптимально отобрано содержание. Исследование валидности теста по содержанию имеет целью исследование свойств выборки заданий, которые включены в тест, с точки зрения их адекватности проверяемой компетенции. Оценить валидность по содержанию какой-либо единой количественной мерой невозможно. Главным методом исследования валидности теста по содержанию является экспертное оценивание содержания каждого задания в отдельности и содержания теста в целом специалистами, т.е. проведение внешней экспертизы по содержанию.

Принципы отбора содержания тестов SAM изложены и обоснованы в спецификациях тестов и представлены подробно в [1].

Тест по математике включает основные разделы математики, представленные в программах для начальной школы.

При отборе содержания использовался Федеральный государственный стандарт начального общего образования (Приказ Минобрнауки России «Об утверждении и введении в действие федерального государственного образовательного стандарта начального общего образования» от 06.10.2009 г. № 373).

Тематическое содержание теста разделено на пять разделов: Числа и вычисления, Измерение величин, Закономерности, Зависимости, Элементы геометрии. Содержательная основа теста может быть представлена в виде матрицы (таблица 5.1), в которую включены:

- разделы предметного содержания (5 разделов);
- математические средства (понятия, представления, принципы, правила, формулы, схемы и проч.), овладение которыми лежит в основе математической компетентности.

Табл. 3.1. Содержание теста по математике

Разделы содержания	Средства ориентировки математических действий
Числа и вычисления	– последовательность натуральных чисел – числовая прямая – позиционный принцип – свойства арифметических действий – порядок действий
Измерение величин	– отношение между числом, величиной и единицей – отношение «целого и частей» – формула площади прямоугольника
Закономерности	– «индукционный шаг» – повторяемость (периодичность)
Зависимости	– отношения между однородными величинами (равенство,

	неравенство, кратности, разностное, «целого и частей») – прямая пропорциональная зависимость между величинами – производные величины: скорость, производительность труда и др. – соотношения между единицами
Элементы геометрии	– форма и другие свойства фигур (основные виды геометрических фигур) – пространственные отношения между фигурами – симметрия

Структурной единицей теста является блок из трех заданий (1-го, 2-го и 3-го уровня), соответствующий одному разделу предметного содержания. В табл. 3.2 представлено примерное соотношение заданий по различным разделам.

Табл. 3.2. Примерная представленность разделов в тестовой тетради

Разделы	**Количество блоков**	**Количество задач**
Числа и вычисления	4	12
Измерение величин	5	15
Закономерности	2	6
Зависимости между величинами	2	6
Элементы геометрии	2	6
ВСЕГО	15	45

Тест по русскому языку включает основные разделы грамматики русского языка, представленные в программах для начальной школы.

При отборе содержания использовались следующие нормативные документы:

- Федеральный государственный стандарт начального общего образования (Приказ Минобрнауки России «Об утверждении и введении в действие федерального государственного образовательного стандарта начального общего образования» от 06.10.2009 г. № 373).
- Примерная основная образовательная программа начального общего образования, рекомендованная к использованию образовательными учреждениями решением Координационного совета при Департаменте общего образования Минобрнауки России по вопросам организации введения ФГОС (протокол заседания Координационного совета №1 от 27-28 июля 2010 г.).

Предмет тестирования соответствует ФГОС (см. Предметные результаты освоения программы: ФГОС, с. 10-11) в следующей части:

- овладение первоначальными представлениями о нормах русского литературного языка (орфоэпических, лексических, грамматических) и правилах речевого этикета; умение ориентироваться в целях, задачах, средствах и условиях общения, выбирать адекватные языковые средства для успешного решения коммуникативных задач;
- овладение учебными действиями с языковыми единицами и умение использовать знания для решения познавательных, практических и коммуникативных задач.

Предметное содержание теста по русскому языку разделено на две области: «Слово и его значение и написание», «Высказывание и его оформление в письменной речи», что

соответствует двум аспектам речевой деятельности (номинативному и коммуникативному).

Первая содержательная область предполагает освоение слова прежде всего с точки зрения соотношения формы и значения, что означает овладение языковыми механизмами формирования и выражения понятий и представлений. Задачи этой части теста построены на материале, соответствующем таким разделам учебного предмета «Русский язык», как «Фонетика», «Состав слова», «Морфология» и «Лексика».

Вторую содержательную область определяет действие последовательного развертывания различных содержаний в языке и речи. Таким образом, данная область охватывает виды синтагматических связей между словами, словосочетаниями, частями предложения, т.е. связана с освоением синтаксических средств языка. Задачи этой части теста построены на материале раздела «Синтаксис и пунктуация».

Содержание теста по русскому языку может быть представлено в виде матрицы (табл. 3.3), в которую включены:

- области предметного содержания;
- языковые средства (понятия, представления, правила, схемы...), овладение которыми лежит в основе языковой компетентности.

Табл. 3.3. Содержание теста по русскому языку

Области и разделы содержания	**Средства ориентировки в языковом материале (понятия, принципы, схемы)**
Слово, его значение и написание	
Звучание и написание (фонетика, графика и орфография)	– отношение «звук — буква» – принципы русского письма – орфографические правила – принцип слогоделения
Состав слова (морфемика и словообразование)	– отношение «форма — значение» – отношение производности
Значение слова (лексика)	– семантические отношения (синонимические, антонимические, родо-видовые) – многозначность – модели переноса значения
Формы слова (морфология)	– принципы выделения частей речи – парадигматические отношения (типы словоизменения: склонение, спряжение) – грамматические категории (род, число, падеж, лицо, время)
Высказывание и его оформление в письменной речи	
Предложение (синтаксис и пунктуация)	– синтаксические отношения: сочинение, подчинение – грамматическая основа предложения – принципы русской пунктуации

Структурной единицей теста является блок из трех заданий (1-го, 2-го и 3-го уровня), соответствующий одному разделу предметного содержания. В табл. 3.4 представлено соотношение заданий по различным разделам содержания и уровням.

Табл. 3.4. Примерная представленность разделов в тестовой тетради

Разделы	**Количество блоков**	**Количество задач**
Фонетика, графика и орфография	2	6
Морфемика и словообразование	3	9
Морфология	4	12
Лексика	4	12
Синтаксис и пунктуация	2	6
ВСЕГО	15	45

Все задания тестов по математике и русскому языку прошли внутреннюю тестологическую экспертизу: проверку правильности формулировок и адекватности формы содержанию задания.

Разработка и апробация тестовых заданий включала также консультации с учителями начальной школы и преподавателями-предметниками по содержанию теста.

На заключительном этапе проведено экспертное оценивание содержания обоих тестов внешними экспертами – предметниками. Экспертные заключения представлены в приложениях.

3.2. Конструктная валидность

Согласно Анастази «конструктная валидность теста показывает, насколько его результаты могут рассматриваться в качестве меры некоего теоретического конструкта» [2]. Для обоснования валидности конструкта формулируются гипотезы о связях результатов тестирования с психологической моделью тестируемой способности.

В модели SAM основным конструктом, подлежащим верификации, является трехуровневая таксономия освоения учебного содержания, вписанная в возрастной контекст. Этот конструкт связан как минимум с двумя гипотетическими утверждениями, требующими проверки:

1. Задания трех уровней, относящиеся к одному блоку и отвечающие теоретически заданным критериям трех уровней, должны обнаруживать соответствующую иерархию по трудности.

2. В начальной школе при хорошем обучении предметный материал усваивается преимущественно на втором (рефлексивном) уровне, т.е. на уровне понимания. Освоение этого материала на третьем (функциональном) уровне происходит в рамках основной школы.

Дополнительный способ валидизации конструкта, состоит в сравнении трудностей заданий, предсказанных разработчиками, с эмпирической трудностью этих же заданий..

Ниже изложены результаты исследований по обоснованию валидности конструкта.

Проверка первой гипотезы

Апробация тестов SAM дала положительный ответ по первому пункту. На диаграмме рис.3.1 представлены коэффициенты решаемости заданий разных уровней теста по математике (данные по задачам 1-го уровня представлены синими столбиками, 2-го уровня – красными и 3-го уровня – зелеными). В каждом из блоков задачи обнаруживают правильную иерархию по трудности: коэффициент решаемости уменьшается от первого уровня к третьему.

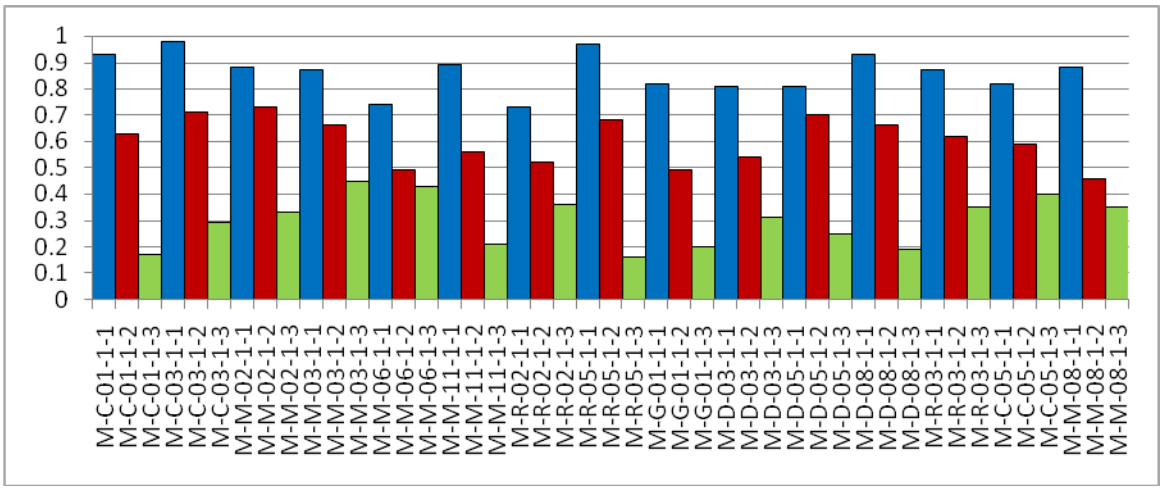

Рис.3.1. Распределение коэффициентов решаемости (математика, вариант 1)

Дополнительно в табл. 3.5 представлены коэффициенты решаемости заданий в зависимости от уровня. В целом задания 1-го уровня легче заданий 2-го уровня, а задания 2-го уровня легче заданий 3-го уровня. Это же подтверждается и в разрезе отдельных содержательных областей (см. табл. 3.6, в которой представлены средние значения коэффициентов решаемости заданий в зависимости от содержательной области).

Таким образом, задачи одного блока, отвечающие теоретически заданным критериям трех уровней, обнаруживают соответствующую иерархию по трудности, т.е. отображают логику функционального генеза в соответствии с принятой таксономией, что свидетельствует в пользу валидности этого конструкта.

Табл. 3.5. Коэффициенты решаемости тестовых заданий в зависимости от их уровня

	Количество заданий		Уровень трудности		
		Среднее значение	Стандарт-ное отклонение	Миним. значение	Максим. значение
Задания 1-го уровня	15	0,86	0,07	0,73	0,98
Задания 2-го уровня	15	0,60	0,09	0,46	0,73
Задания 3-го уровня	15	0,30	0,09	0,16	0,45
Всего	45	0,59	0,25	0,16	0,98

Табл. 3.6. Средние значения коэффициентов решаемости заданий в зависимости от содержательной области

Содержательная область	Весь тест	1-й уровень	2-й уровень	3-й уровень
Числа и вычисления	0,61	0,91	0,64	0,29
Измерение величин	0,59	0,85	0,58	0,35
Закономерности	0,59	0,86	0,61	0,29
Зависимости	0,58	0,85	0,64	0,25
Элементы геометрии	0,51	0,82	0,49	0,20

Проверка второй гипотезы

Рамки проекта не давали возможности исчерпывающим образом проверить вторую гипотезу. Поэтому проведенное исследование вынужденно носило пробный, предварительный характер. Учитывая, что в нашем распоряжении был всего год, мы выбрали систему срезов, которая: а) покрывала все классы от 4-х до 11-х; б) показывала реальный шаг годичного развития четырех возрастных групп учащихся.

В соответствии с задачей исследования - выявление нормальной картины распределения предметных образовательных результатов - в качестве объекта тестирования были взяты параллельные классы двух хороших школ, представляющие относительно благополучный вариант культурного развития детей.

Исследование проводилось в 2011-2012 гг. В 2011 г. тесты по математике и русскому языку были предложены учащимся 4-х возрастных групп – 4-х, 6-х, 8-х и 10-х классов (тестировались по две параллели в каждой школе). Через год, в 2012 г. те же тесты были предложены тем же учащимся, которые в тот момент были учащимися 5-х, 7-х, 9-х и 11-х классов. Тестирование проходило весной, в конце учебного года.

Результаты анализа данных тестирования по математике

Первый срез по математике выполняли 396 учащихся, второй - 412 учащихся. Всего учеников, которые выполняли оба среза, – 374. В табл. 3.7 представлены данные о распределении учащихся по параллелям.

Табл. 3.7. Распределение учащихся по параллелям (математика)

	1 срез	2 срез	оба среза
4 – 5 класс	104	102	93
6 – 7 класс	103	108	99
8 – 9 класс	104	111	100
10 – 11 класс	85	91	82
ВСЕГО	396	412	374

Для интерпретации результатов участников тестирования SAM на основе трехуровневой модели тестирования были установлены пороговые оценки, позволяющие поделить всех участников на 4 группы по ступеням достижений. Процедура установления границ будет описана в главе 7. Всего выделено 4 категории достижений, которые отвечают следующим содержательным критериям:

Ниже первой ступени – не освоен даже первый уровень;

Первая ступень – освоен только первый уровень;

Вторая ступень – освоен второй уровень;

Третья ступень – освоен третий уровень.

В табл. 3.8 показано распределение участников тестирования разных возрастов по ступеням достижений в зависимости от класса. (В таблице отсутствует нулевая ступень, т.к. число участников тестирования, находящихся на этой ступени было ничтожно мало. Это объясняется довольно сильной выборкой, что отмечалось ранее). Из таблицы видно, что наименьшие изменения в доле учащихся, продемонстрировавших определенный уровень достижений, наблюдается от 10-го к 11-му классу. В данных классах различия в долях по каждому уровню составляют не более 2%. К этому времени результаты становятся устойчивыми: большинство участников находятся на 3-ей ступени. Наибольшие изменения в доле учащихся, находящихся на различных ступенях, наблюдаются от 6-го к 7-му и от 7-го к 8-му классам, т.е. в основной школе. Здесь различия по второму и третьему уровню составляют 8% – 16%. Это подтверждает нашу гипотезу, что освоение предметного материала на функциональном уровне может и должно происходить в рамках основной школы.

Табл. 3.8. Распределение учащихся разных классов по уровням достижений (математика)

	1 ступень	2 ступень	3 ступень
4 класс	16%	64%	18%
5 класс	10%	60%	30%
6 класс	7%	55%	38%
7 класс	4%	43%	53%
8 класс	1%	29%	70%
9 класс	2%	24%	74%
10 класс	1%	17%	82%
11 класс	1%	15%	84%

К концу начальной школы (4-ый класс) в общеобразовательной школе предметный материал усваивается на уровне понимания – большинство детей (64%) находятся на 2-ой ступени. Третий уровень только начинает формироваться – на 3-ей ступени находятся только 18% детей (и это сильные классы). Это подтверждает нашу гипотезу, что в норме к концу начальной школы материал может и должен усваиваться на рефлексивном уровне, т.е. на уровне понимания.

На рис.3.2 данные таблицы 3.8 показаны графически. Можно заметить, что доля учащихся, демонстрирующих второй уровень достижений, неизменно падает с увеличением школьного возраста. Доля учащихся, демонстрирующих третий уровень, растет.

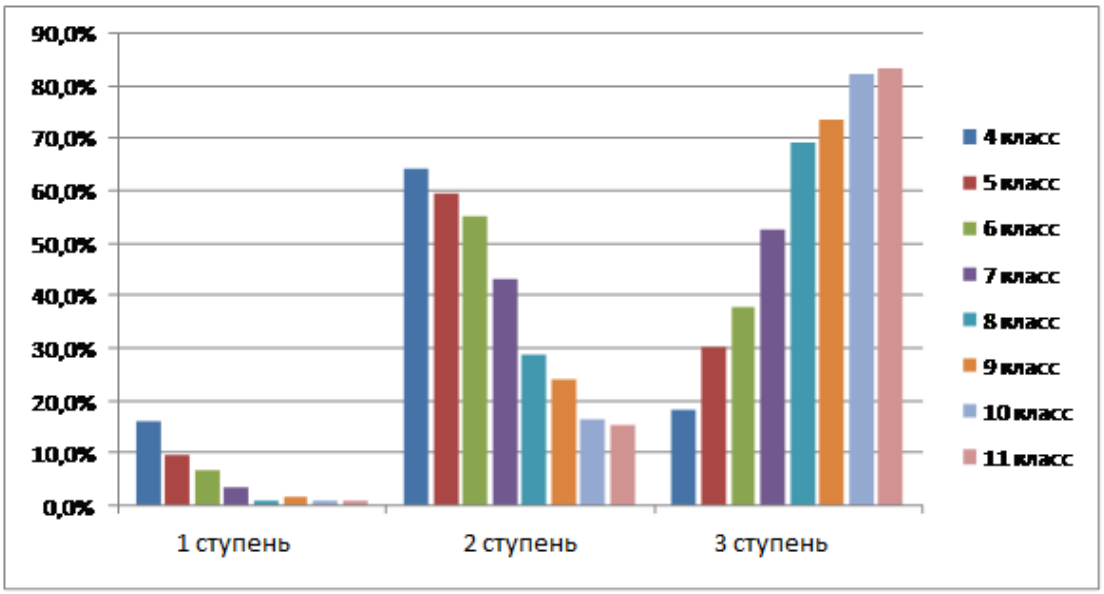

Рис.3.2. Распределение участников тестирования по ступеням достижений в зависимости от класса (математика)

На рис. 3.3 показано соотношение учащихся различных классов, достигших разных уровней в освоении математики. На диаграмме хорошо видна возрастная динамика присвоения учебного содержания. К концу начальной школы (4-ый класс) вторая ступень является доминирующей. Начиная с 8-го класса (окончание основной школы) доминирует 3-ий уровень. Это обеспечивает поддержку теоретической модели SAM, т.е. свидетельствует в пользу ее валидности.

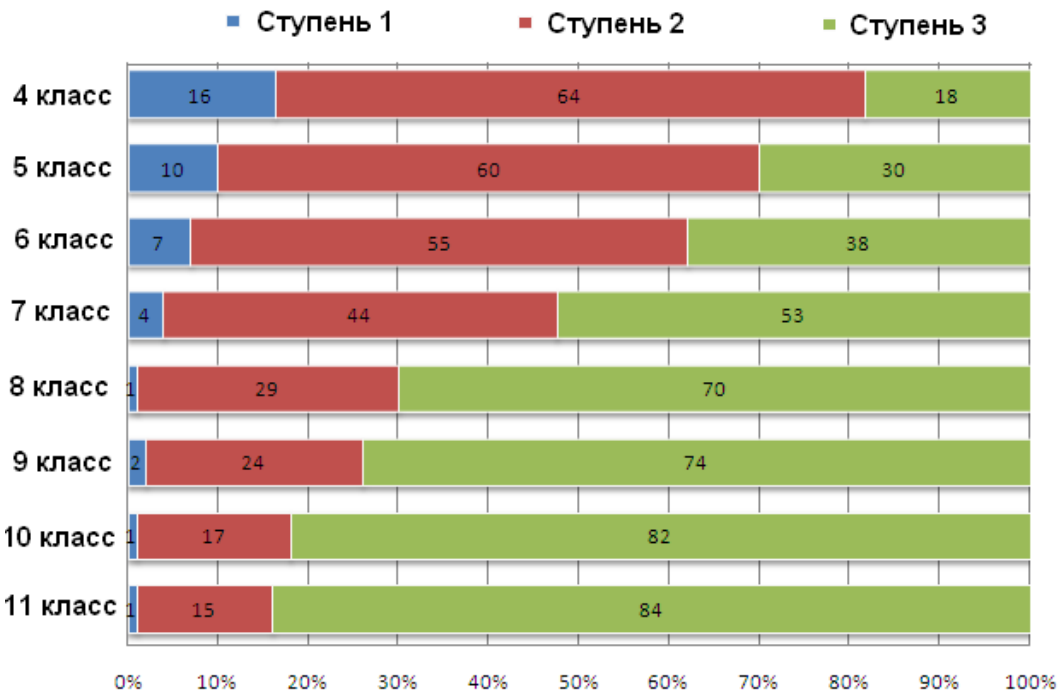

Рис.3.3. Соотношение учащихся различных классов, достигших разных уровней освоения математики

Отметим, что результаты проведенного исследования могут быть использованы и шире, а именно для оценивания индивидуального прогресса учащихся с течением времени. Оценить индивидуальный прогресс позволяет тот факт, что одни и те же дети были протестированы дважды, с разницей в один год. Прогресс учащихся можно отследить по двум показателям – тестовому баллу и ступени, на которой ученик находится. Повышение или понижение ступени ученика является следствием повышения или понижения тестового балла этого ученика. Однако изменение тестового балла не всегда ведет к изменению ступени.

Именно повышение ступени ученика свидетельствует о его индивидуальном прогрессе в освоении предмета. Изменение тестового балла будет демонстрировать, есть ли предпосылки к повышению уровня или ученик «стоит на месте».

В табл. 3.9 представлены данные о доле учеников двух параллелей – 4-5 классов и 6-7 классов, продемонстрировавших определенное сочетание ступеней на первом и втором срезах. Из таблицы мы видим, что 9% учащихся 5-го класса понизили свой уровень по сравнению со своим же результатом в 4-м классе: 2% перешли со 2-ой ступени на 1-ую и 7% - с 3-ей на 2-ую. 29% учащихся 5-ых классов повысили свой уровень: 10% перешли с 1-ой ступени на 2-ую; 1% - с 1-ой на 3-ю и 18% учащихся перешли со 2-ой ступени на 3-ю. И наконец, 61% участников тестирования сохранили свою ступень: 4% - 1-ую, 44% - 2-ую и 13% - 3-ю. Аналогично могут быть проинтерпретированы данные по параллели 6-7 классов. Здесь, наибольшее число участников – 23% - перешли со 2-ой ступени на 3-ю.

Табл. 3.9. Процент учеников, продемонстрировавших определенное сочетание ступеней на первом и втором срезах (математика)

Сочетания ступеней 1 срез/ 2 срез	4-5 класс	6-7 класс
1/ 0	0	0
1/ 1	4	2
1/ 2	10	4
1/ 3	1	0
2/ 0	0	0
2/ 1	2	1
2/ 2	44	31
2/ 3	18	23
3/ 0	0	0
3/ 1	0	0
3/ 2	7	9
3/ 3	13	29

Отметим, что у учащихся других параллелей – 8-9 и 10-11 классов существенного изменения ступени не происходит, т.к. в этих параллелях доминирующее число участников находятся на 3-ей ступени.

Результаты анализа данных тестирования по русскому языку

Первый срез по русскому языку выполняли 382 учащихся. Второй срез по русскому языку выполняли 409 учащихся. Всего учеников, которые выполняли оба среза, – 363. В табл. 3.10 представлены данные о распределении учащихся по параллелям.

Табл. 3.10. Распределение учащихся по параллелям (русский язык)

	1 срез	2 срез	оба среза
4 – 5 класс	102	102	92
6 – 7 класс	102	107	99
8 – 9 класс	98	111	95
10 – 11 класс	80	89	77
ВСЕГО	382	409	363

В табл. 3.11 показано распределение участников тестирования разных возрастов по ступеням достижений в зависимости от класса. Из таблицы видно, что наименьшие изменения в доле учащихся, продемонстрировавших определенный уровень достижений, наблюдается от 10-го к 11-му классу. В данных классах различия в долях по каждому уровню составляют не более 2%. К этому времени результаты становятся устойчивыми: большинство участников находятся на 3-ей ступени. Наибольшие изменения в доле учащихся, находящихся на различных ступенях, наблюдаются от 6-го к 7-му и от 7-го к 8-му классам, т.е. в основной школе. Здесь различия по второму и третьему уровню составляют 8% – 16%. Это подтверждает нашу гипотезу, что освоение предметного материала на функциональном уровне может и должно происходить в рамках основной школы.

Табл. 3.11. Распределение учащихся разных классов по уровням достижений (русский язык)

	1 ступень	2 ступень	3 ступень	0 ступень
4 класс	37%	52%	10%	1%
5 класс	24%	51%	20%	5%
6 класс	21%	55%	22%	2%
7 класс	13%	54%	31%	2%
8 класс	16%	40%	43%	1%
9 класс	7%	41%	52%	0%
10 класс	1%	24%	75%	0%
11 класс	3%	26%	71%	0%

К концу начальной школы (4-ый класс) в общеобразовательной школе предметный материал усваивается на уровне понимания – большинство детей (52%) находятся на 2-ой ступени. Третий уровень только начинает формироваться – на 3-ей ступени находятся только 10% детей (и это сильные классы). Это подтверждает нашу гипотезу, что к концу начальной школы материал усваивается на рефлексивном уровне, т.е. на уровне понимания.

На рис.3.4 данные таблицы 3.11 показаны графически. Можно заметить, что доля учащихся, находящихся на 1-ой ступени (т.е. демонстрирующих первый уровень

достижений), падает с увеличением школьного возраста, а доля учащихся, находящихся на 3-ей ступени (т.е. демонстрирующих третий уровень), растет от четвертого к десятому классу (с незначительными колебаниями в районе 7-8 класса для 1-ой ступени и в районе 10-11 класса для 3-ей ступени). Что касается 2-ой ступени, то доля учащихся, находящихся на этой ступени (т.е. демонстрирующих второй уровень достижений) является относительно устойчивой для 4–7 классов (на уровне 50% - 55%), а затем начинает снижаться: до уровня 40% в 8 – 9 классах и 25% в 10 – 11 классах.

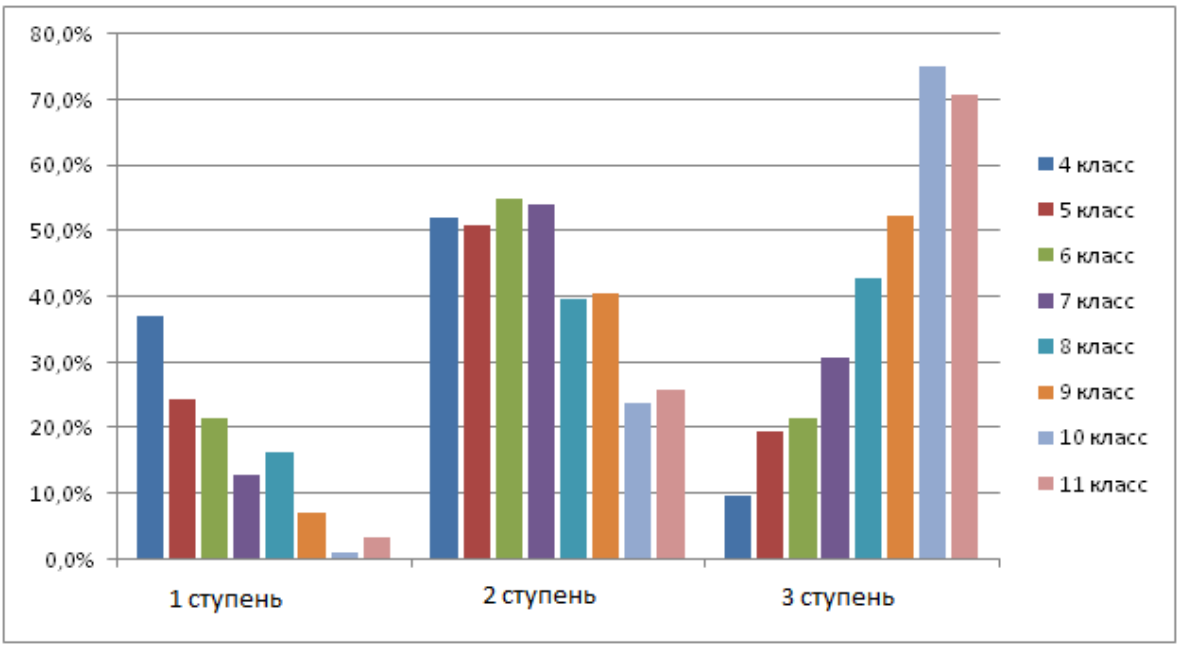

Рис.3.4. Распределение участников тестирования по ступеням достижений в зависимости от класса (русский язык)

Интересно заметить, что наименьшие изменения в доле учащихся, продемонстрировавших определенный уровень образовательных результатов, наблюдаются от 5-го к 6-му классу и от 10-го к 11-му классу. В данных классах различия в долях по каждому уровню составляют не более 4,2%. Наибольшие изменения в доле учащихся наблюдаются:

- по первому уровню – от четвертого к пятому классу (составляет 13%);
- по второму уровню – от седьмого к восьмому классу (14,5%), от девятого к десятому классу (17%);
- по третьему уровню – от девятого к десятому классу (23%).

На рис. 3.5 показано соотношение учащихся различных классов, достигших разных уровней в освоении предмета «русский язык». На диаграмме хорошо видна возрастная динамика образовательных результатов. К концу начальной школы (4-ый класс) вторая ступень является доминирующей. Начиная с 9-го класса (окончание основной школы) доминирует 3-ий уровень. Это обеспечивает поддержку теоретической модели SAM, т.е. свидетельствует в пользу ее валидности.

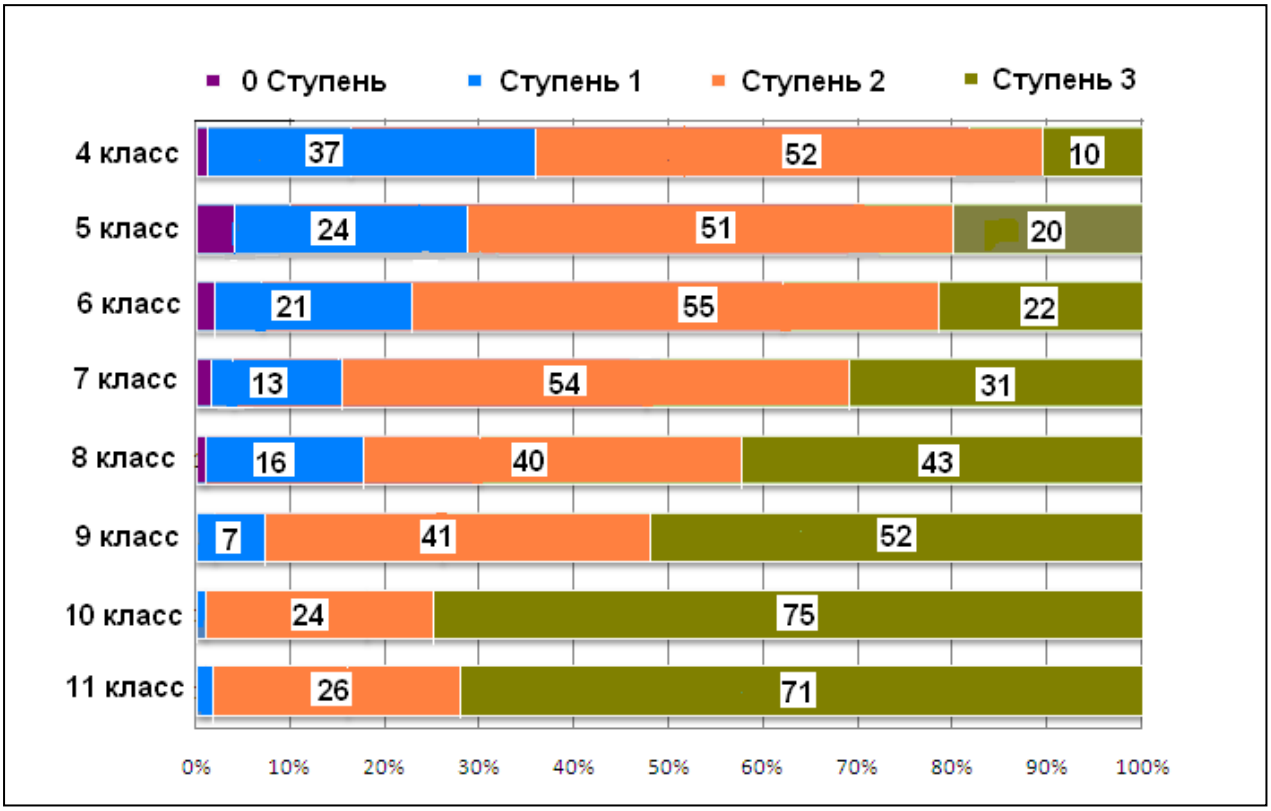

Рис.3.5. Соотношение учащихся различных классов, достигших разных уровней освоения предмета «русский язык»

Отметим, что также как по математике, результаты проведенного исследования по русскому языку могут быть использованы и шире, а именно для оценивания индивидуального прогресса учащихся с течением времени. Оценить индивидуальный прогресс позволяет тот факт, что одни и те же дети были протестированы дважды, с разницей в один год. Прогресс учащихся можно отследить по двум показателям – тестовому баллу и ступени, на которой ученик находится. Повышение или понижение ступени ученика является следствием повышения или понижения тестового балла этого ученика. Однако изменение тестового балла не всегда ведет к изменению ступени.

Именно повышение ступени ученика свидетельствует о его индивидуальном прогрессе в освоении предмета. Изменение тестового балла будет демонстрировать, есть ли предпосылки к повышению уровня или ученик «стоит на месте».

В табл. 3.12 представлены данные о доле учеников двух параллелей – 4-5 классов и 6-7 классов, продемонстрировавших определенное сочетание ступеней на первом и втором срезах. Из таблицы мы видим, что 9% учащихся 5-го класса понизили свой уровень по сравнению со своим же результатом в 4-м классе: 2% перешли с 1-ой ступени на 0-ую; 3% перешли со 2-ой ступени на 1-ую и 4% - с 3-ей на 2-ую. 27% учащихся 5-ых классов повысили свой уровень: 13% перешли с 1-ой ступени на 2-ую и 14% учащихся перешли со 2-ой ступени на 3-ю. И наконец, 63% участников тестирования сохранили свою ступень: 20% - 1-ую, 36% - 2-ую и 7% - 3-ю. Аналогично могут быть проинтерпретированы данные по парралели 6-7 классов. Здесь, наибольшее число участников – 34% - сохранили 2-ую ступень и 17% - перешли со 2-ой ступени на 3-ю.

Табл. 3.12. Процент учеников, продемонстрировавших определенное сочетание ступеней на первом и втором срезах

Сочетания ступеней 1 срез/ 2 срез	4-5 класс	6-7 класс
1/ 0	2	0
1/ 1	20	10
1/ 2	13	11
1/ 3	0	0
2/ 0	0	0
2/ 1	3	3
2/ 2	36	34
2/ 3	14	17
3/ 0	0	0
3/ 1	0	0
3/ 2	4	6
3/ 3	7	16

Отметим, что у учащихся других параллелей – 8-9 и 10-11 классов существенного изменения ступени не происходит, т.к. в этих параллелях доминирующее число участников находятся на 3-ей ступени.

Сравнение трудностей заданий, предсказанных разработчиками, с их эмпирической трудностью

При разработке каждого тестового задания, авторы должны представлять себе когнитивную основу действия по его выполнению, и на этом основании иметь возможность прогнозировать сравнительный уровень трудности разных заданий для потенциальной выборки участников тестирования. Таким образом, если прогнозы разработчиков теста близки к эмпирическим значениям, то можно с определенной вероятностью заключить, что их представления о действиях испытуемых соответствуют действительности.

Для проведения соответствующего исследования был выбран тест SAM по математике. Разработчикам было предложено оценить предполагаемую трудность заданий до их апробации. Был оценен коэффициент согласованности среди экспертов: он составил 0.831. Затем, после проведения апробации этих заданий был оценен коэффициент согласованности между ожидаемыми и эмпирическими трудностями заданий по всему тесту и для каждого субтеста с заданиями первого, второго и третьего уровней в отдельности. В табл. 3.13 представлены коэффициенты корреляции.

Таблица 5.13. Коэффициенты корреляции между ожидаемыми и эмпирическими трудностями заданий

Задания 1-го уровня	0,882
Задания 2-го уровня	0,973
Задания 3-го уровня	0,928
Весь тест	0,931

На рис. 3.6 показано отношение между ожиданиями экспертов и эмпирическими показателями трудности заданий для всего теста. Эти результаты показывают достаточно высокий уровень согласованности между экспертами, а также между их ожиданиями и эмпирическими результатами, что является еще одним свидетельством валидности конструкта.

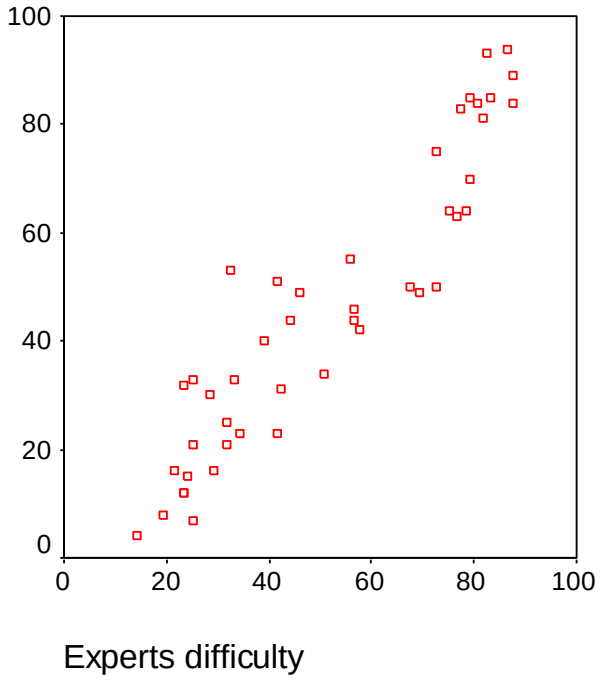

Рис. 3.6. Корреляционная связь между экспертно оцененным и эмпирическим значением трудности

3.3. Критериальная валидность

Как отмечалось выше, критериальная валидность предполагает для установления обоснованности результатов тестирования и их интерпретации сопоставление результатов с неким внешним критерием, связанным с измеряемым конструктом. Критериальная валидность подразделяется на прогностическую и текущую. Прогностическая валидность отвечает на вопрос о том, насколько хорошо тест предсказывает баллы по критерию в будущем. Текущая критериальная валидность отвечает на вопрос о связи результатов тестирования с критерием в настоящее время.

Ниже приведены результаты исследования критериальной валидности SAM.

Исследование прогностической валидности SAM

Исследование проводилось на базе данных апробации тестов SAM в одном из регионов РФ весной 2011 г. Общий объем выборки составил 941 человек из 12 школ. Выборка составлялась как репрезентативная со стратификацией по двум основаниям: тип школы (общеобразовательная школа / гимназия, лицей) и местоположение школы (город / село). Тестирование проходило в конце 4-го класса, т.е. по окончании начальной школы.

Результаты тестирования SAM переводились на 1000-балльную шкалу со средним значением около 500 и стандартным отклонением 50 (процедура шкалирования будет подробно описана в главе 7). Дополнительно, как отмечалось выше, для интерпретации результатов участников тестирования SAM на основе трехуровневой модели тестирования были установлены пороговые оценки, позволяющие поделить всех участников на 4 группы по ступеням достижений. На рис.3.7 показано распределение участников тестирования по ступеням достижений по математике.

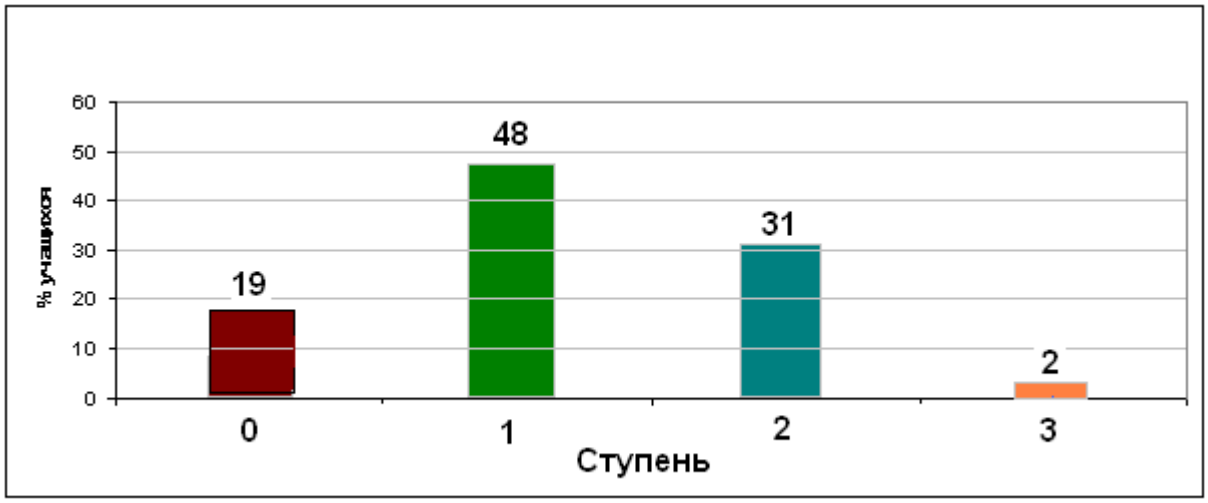

Рис.3.7. Распределение участников тестирования по ступеням достижений (математика)

Мы видим, что в целом по региону 19% участников тестирования находятся на 0-ой ступени (не освоен даже 1-ый уровень); 48% - на 1-ой (освоен только 1-ый уровень); 31% - на 2-ой (освоен 2-ой уровень) и только 2% - на 3-ей ступени (освоен 3-ий уровень).

На рис.3.8 показано распределение участников по ступеням достижений в зависимости от школы (коды школ указаны слева, по вертикальной оси). Школы упорядочены по убыванию среднего тестового балла. Из графика видно, что число детей, находящихся на 0-ой ступени колеблется от 8 до 48% в зависимости от школы, а число детей, находящихся на 3-ей ступени, не превышает 7%. Во всех школах доминирует 1-я ступень.

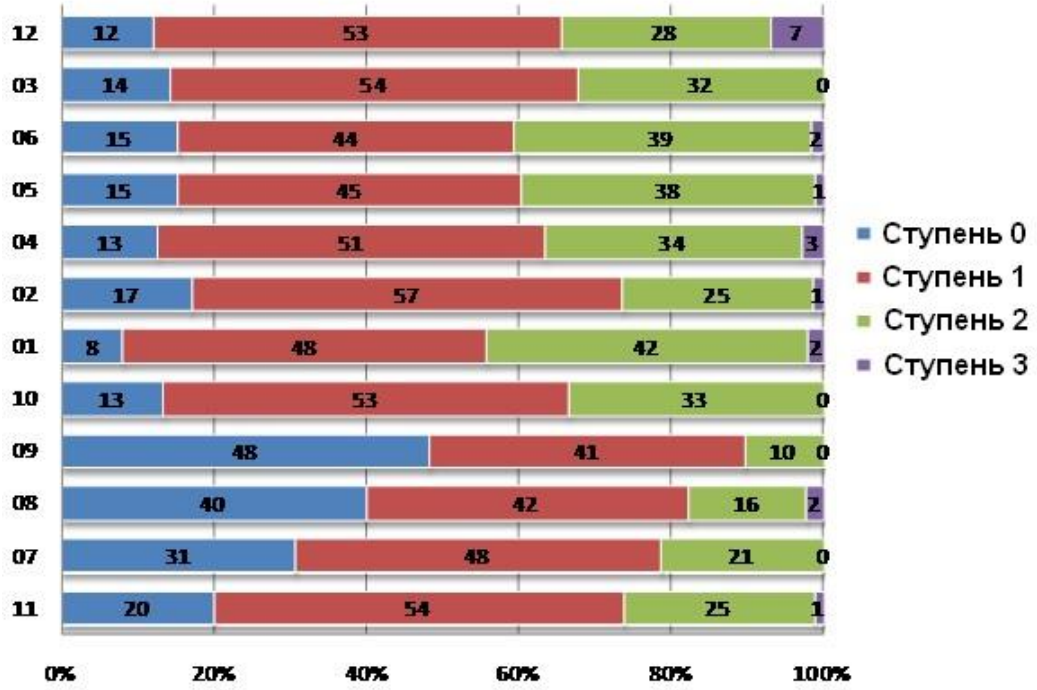

Рис.3.8. Распределение участников тестирования по ступеням достижений в зависимости от школы (математика)

Для проведения исследования прогностической валидности теста SAM по математике были собраны оценки по математике этих же учащихся в 5-м классе (т.е. через год после проведения тестирования). Конечно, школьные оценки являются достаточно субъективными и шкала оценок обладает очень низкой различительной способностью. Однако, другого универсального критерия, показывающего успехи детей в математике в 5-м классе, не существует. Оценки удалось собрать по 649 учащимся, что составляет 67% от общей выборки тестирования. Поэтому для проведения исследования были отобраны результаты только тех учащихся, по которым известны оценки по математике за 5-ый класс. Из отобранных учащихся 110 (17%) находятся на 0-ой ступени; 311 (48%) - на 1-ой; 219 (34%) - на 2-ой (освоен 2-ой уровень) и 10 (1%) - на 3-ей ступени. Следует отметить, что распределение по ступеням отобранных учащихся очень близко к распределению в целом по региону, представленному на рис.3.7.

Из 649 учащихся 1 человек имел оценку 2. Этот учащийся был удален из дальнейшего анализа. Остальные учащиеся распределились по оценкам следующим образом: 218 (34%) имеют оценку 3; 275 (42%) - оценку 4 и 155 (24%) - оценку 5.

В табл. 3.14 показано распределение оценок учащихся, находящихся на различных ступенях, на рис.3.9 это же распределение показано графически. Мы видим, что все учащиеся, отнесенные к 3-ей ступени достижений по математике по результатам теста SAM, имеют оценку 5. Учащиеся, отнесенные ко 2-ой ступени, распределились, в основном, между оценками 4 и 5. Из учащихся, отнесенных к 1-ой ступени, половина имеет оценку 4 и немногим более трети – оценку 3. И, наконец, для учащихся, отнесенных к 0-ой ступени доминирующей оценкой является оценка 3.

Табл. 3.14. Распределение оценок учащихся, находящихся на различных ступенях, % (математика)

Ступень	Оценка		
	3	**4**	**5**
0 Ступень	82	17	1
1 Ступень	35	49	16
2 Ступень	9	48	43
3 Ступень			100

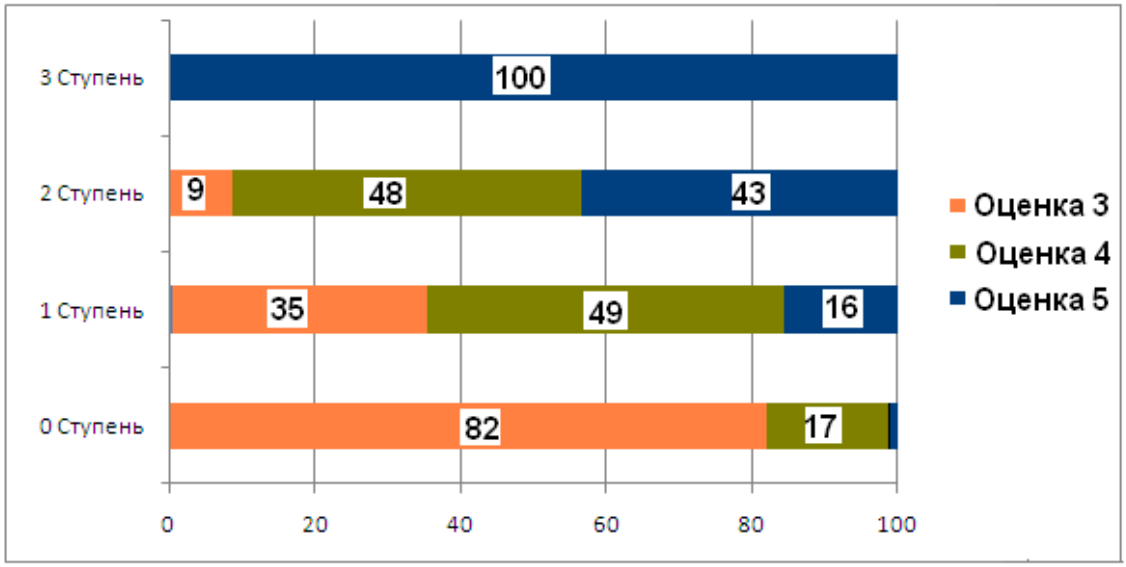

Рис. 3.9. Распределение оценок учащихся в зависимости от ступени достижений (математика)

В табл. 3.15 показано распределение по ступеням учащихся, имеющих различные оценки, на рис.3.10 это же распределение показано графически. Мы видим, что из учащихся, имеющих оценку 5, 62% были отнесены ко 2-ой ступени; из учащихся, имеющих оценку 4, 55% были отнесены к 1-ой ступени и 38% - ко 2-ой. И наконец, из учащихся, имеющих оценку 3, 41% был отнесен к 0-ой ступени и 50% - к 1-ой ступени. Никто их тех, кто был отнесен к 3-ей ступени, не имеет оценку ниже 5.

Табл. 3.15. Распределение по ступеням учащихся, имеющих различные оценки, % (математика)

Оценка	Ступень			
	0-ая	**1-ая**	**2-ая**	**3-я**
3	41	50	9	
4	7	55	38	
5		31	62	7

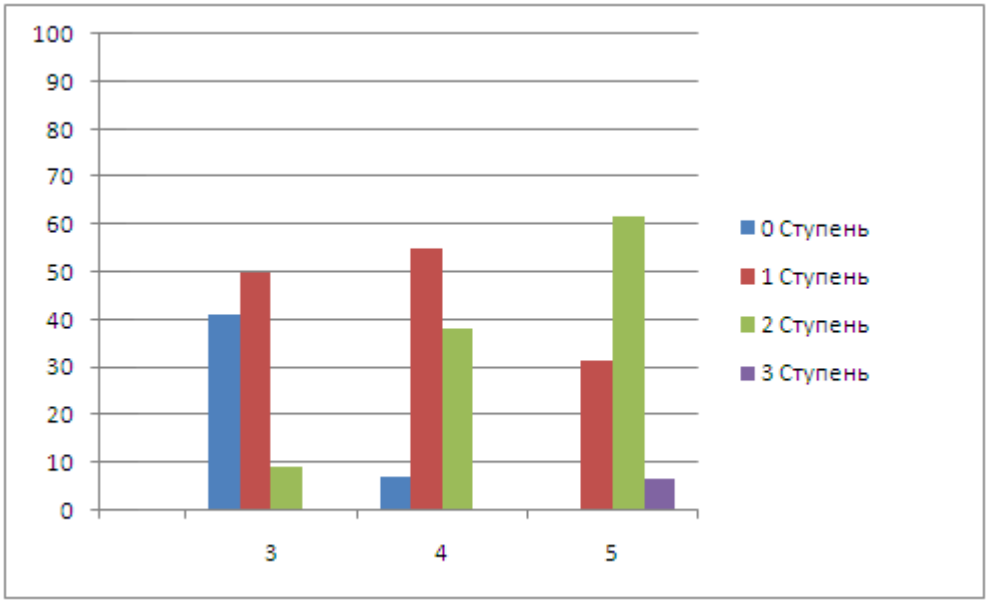

Рис. 3.10. Распределение учащихся по ступеням достижений в зависимости от оценки (математика)

Дополнительно были посчитаны коэффициенты корреляции между тестовым баллом учащихся и их школьными оценками, а также между ступенью, к которой был отнесен учащийся, и его школьной оценкой. Коэффициенты корреляции соответственно равны 0,6 и 0,56. Эти показатели являются довольно высокими, свидетельствующими в пользу прогностической валидности теста SAM по математике.

Аналогичное исследование было проведено для теста SAM по русскому языку. В целом ситуация аналогична. На рис. 3.11 показано распределение участников тестирования по ступеням достижений по русскому языку в целом по региону. Отметим, что из всей выборки к 3-ей ступени был отнесен только 1 человек. Этот учащийся был исключен из дальнейшего анализа, и поэтому 3-ий уровень далее представлен не будет.

Мы видим, что в целом по региону 15% участников тестирования находятся на 0-ой ступени (не освоен даже 1-ый уровень); 64% - на 1-ой (освоен только 1-ый уровень); 21% - на 2-ой (освоен 2-ой уровень).

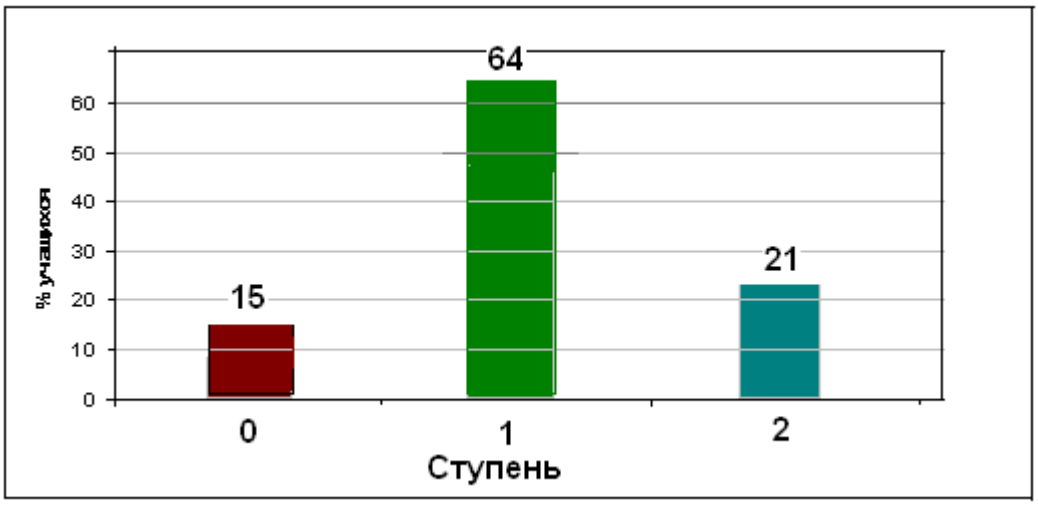

Рис.3.11. Распределение участников тестирования по ступеням достижений (русский язык)

Для проведения исследования прогностической валидности теста SAM по русскому языку были собраны оценки по русскому языку этих же учащихся в 5-м классе (т.е. через год после проведения тестирования). Выборка учащихся, у которых оказались доступны оценки, составила 638 учащихся, что составляет 67% от общей выборки тестирования. Поэтому для проведения исследования были отобраны результаты только тех учащихся, по которым известны оценки по русскому языку за 5-ый класс. Из отобранных учащихся 159 (25%) находятся на 0-ой ступени; 348 (55%) - на 1-ой; 130 (20%) - на 2-ой (освоен 2-ой уровень) и 1 учащийся – на 3-ей ступени (из анализа удален). Следует отметить, что распределение по ступеням отобранных учащихся отличается от распределения в целом по региону, представленному на рис. 3.13.

Из 638 учащихся ни один человек не имел оценку 2. Учащиеся распределились по оценкам следующим образом: 228 (36%) имеют оценку 3; 316 (50%) - оценку 4 и 94 (14%) - оценку 5.

В табл. 3.16 показано распределение оценок учащихся, находящихся на различных ступенях, на рис.3.12 это же распределение показано графически. Мы видим, что 82% учащихся, отнесенных ко 0-ой ступени достижений по русскому языку по результатам теста SAM, имеют оценку 3. Из учащихся, отнесенных к 1-ой ступени, большинство (62%) имеют оценку 4. Учащиеся, отнесенные ко 2-ой ступени, распределились, в основном, между оценками 4 и 5.

Таблица 3.16. Распределение оценок учащихся, находящихся на различных ступенях, % (русский язык)

Ступень	Оценка		
	3	**4**	**5**
0 Ступень	82	17	1
1 Ступень	28	62	10
2 Ступень	7	49	43

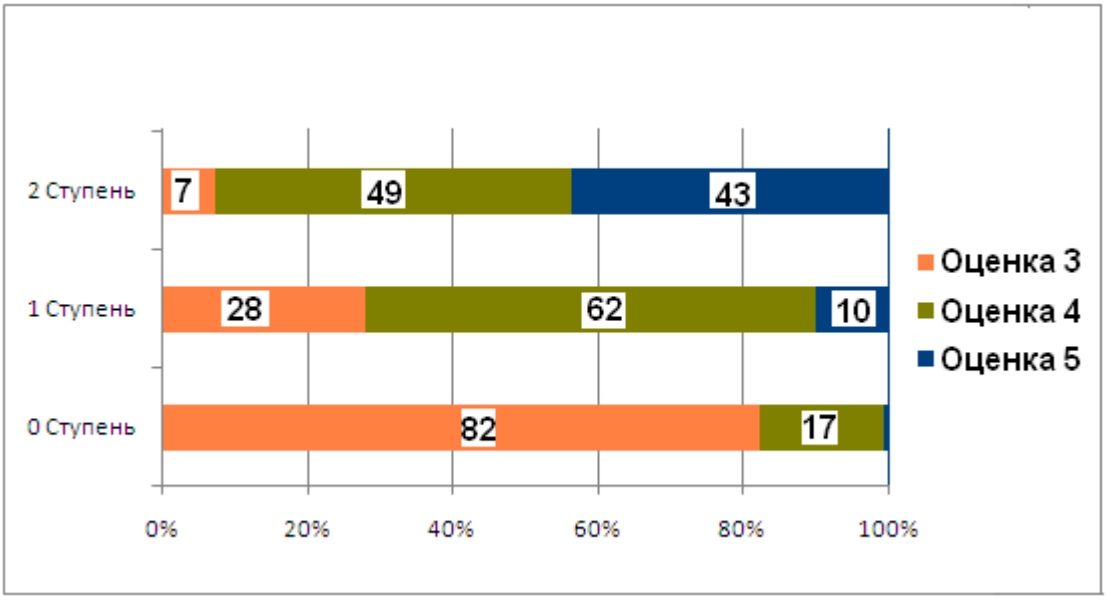

Рис. 3.12. Распределение оценок учащихся в зависимости от ступени достижений (русский язык)

В табл. 3.17 показано распределение по ступеням учащихся, имеющих различные оценки, на рис.3.13 это же распределение показано графически. Мы видим, что из учащихся, имеющих оценку 5, 61% были отнесены ко 2-ой ступени; из учащихся, имеющих оценку 4, 70% были отнесены к 1-ой ступени и 21% - ко 2-ой. И наконец, из учащихся, имеющих оценку 3, 55% был отнесен к 0-ой ступени и 41% - к 1-ой ступени.

Табл. 3.17. Распределение по ступеням учащихся, имеющих различные оценки, % (русский язык)

Оценка	Ступень		
	0-ая	**1-ая**	**2-ая**
3	55	41	4
4	9	70	21
5	1	37	61

Дополнительно были посчитаны коэффициенты корреляции между тестовым баллом учащихся по русскому языку и их школьными оценками, а также между ступенью, к которой был отнесен учащийся, и его школьной оценкой. Коэффициенты корреляции соответственно равны 0,64 и 0,58. Эти показатели близки к соответствующим показателям для теста по математике. Как отмечалось выше, они являются довольно высокими, свидетельствующими в пользу прогностической валидности теста SAM по русскому языку.

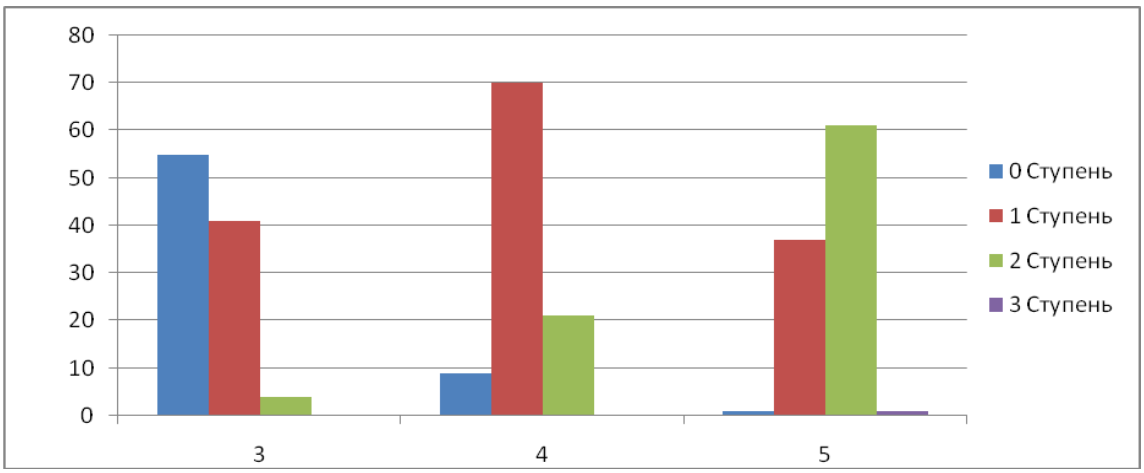

Рис. 3.13. Распределение учащихся по ступеням достижений в зависимости от оценки (русский язык)

Таким образом, проведенное исследование позволяет сделать вывод, что тесты SAM по математике и русскому языку обладают достаточно высокой прогностической валидностью.

Исследование текущей валидности SAM

Исследование проводилось на базе данных апробации тестов SAM в одной из областей РФ весной 2012 г. Общий объем выборки составил 4406 человек по математике и 4385 по русскому языку. Особенностью данной апробации являлось то, что были протестированы практически все учащиеся 4-го класса начальной школы данного региона. Текущая валидность исследуется с помощью внешнего критерия, по которому информация собирается одновременно с экспериментами по проверяемой методике. В качестве критерия для исследования текущей валидности были выбраны оценки по предмету, которые ожидались у учащихся по окончании начальной школы. Эти оценки были собраны путем опроса учителей во время тестирования. Оценки удалось собрать у 3955 человек по математике, 3893 человек по русскому языку. В табл. 3.18 представлено распределение оценок по математике и русскому языку.

Табл. 3.18. Распределение ожидаемых оценок учащихся, %

Ожидаемая оценка за год по предмету	Математика	Русский язык
3	39	41
4	52	51
5	9	8

В табл. 3.19 представлено распределение участников тестирования по ступеням достижений. Мы видим, что по математике в целом по региону 2% участников тестирования находятся на 0-ой ступени (не освоен даже 1-ый уровень); 27% - на 1-ой (освоен только 1-ый уровень); 54% - на 2-ой (освоен 2-ой уровень) и 17% - на 3-ей (освоен и 3-ий уровень). По русскому языку распределение по ступеням имеет вид: 4%, 39%, 39% и 12% - соответственно для 0-ой, 1-ой, 2-ой и 3-ей ступеней.

Табл. 3.19. Распределение участников тестирования по ступеням достижений, %

Ступень	Математика	Русский язык
0	2	4
1	27	39
2	54	39
3	17	12

В табл. 3.20 показано распределение оценок по математике учащихся, находящихся на различных ступенях, на рис. 3.14 это же распределение показано графически. Мы видим, что 89% учащихся, отнесенных ко 0-ой ступени достижений по математике по результатам теста SAM, имеют оценку 3. Из учащихся, отнесенных к 1-ой ступени, 64% имеют оценку 3 и 34% - оценку 4. Большинство учащиеся, отнесенных ко 2-ой ступени (59%), имеют оценку 4. То же справедливо и для учащихся, находящихся на 3-ей ступени, однако 25% этих учащихся имеют оценку 5.

Табл. 3.20. Распределение оценок учащихся, находящихся на различных ступенях, % (математика)

Ступень	Оценка		
	3	**4**	**5**
0 Ступень	89	9	2
1 Ступень	64	34	2
2 Ступень	33	59	8
3 Ступень	13	62	25

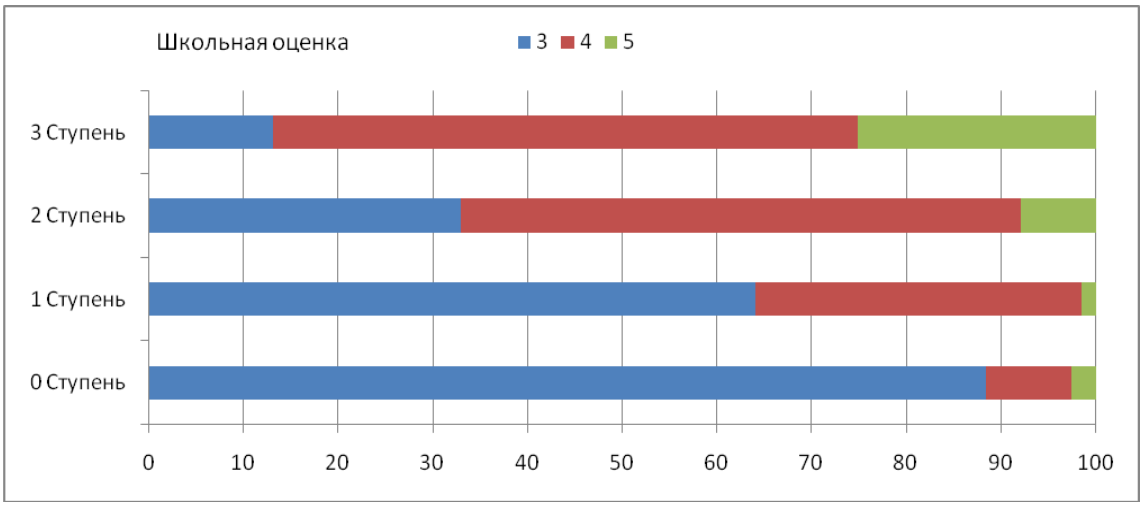

Рис. 3.14. Распределение оценок учащихся в зависимости от ступени достижений (математика)

В табл. 3.21 показано распределение по ступеням учащихся, имеющих различные оценки, на рис.3.15 это же распределение показано графически. Мы видим, что из учащихся, имеющих оценку 5, 47% были отнесены ко 2-ой ступени и 48% - к 3-ей; из учащихся, имеющих оценку 4, 61% были отнесены ко 2-ой ступени и 21% - к 3-ой. И наконец, из учащихся, имеющих оценку 3, 4% были отнесены к 0-ой ступени, 44% - к 1-ой и 46% - ко 2-ой ступени.

Табл. 3.21. Распределение по ступеням учащихся, имеющих различные оценки, % (математика)

Оценка		Ступень		
	0-ая	**1-ая**	**2-ая**	**3-я**
3	4	44	46	6
4	0	18	61	21
5	1	4	47	48

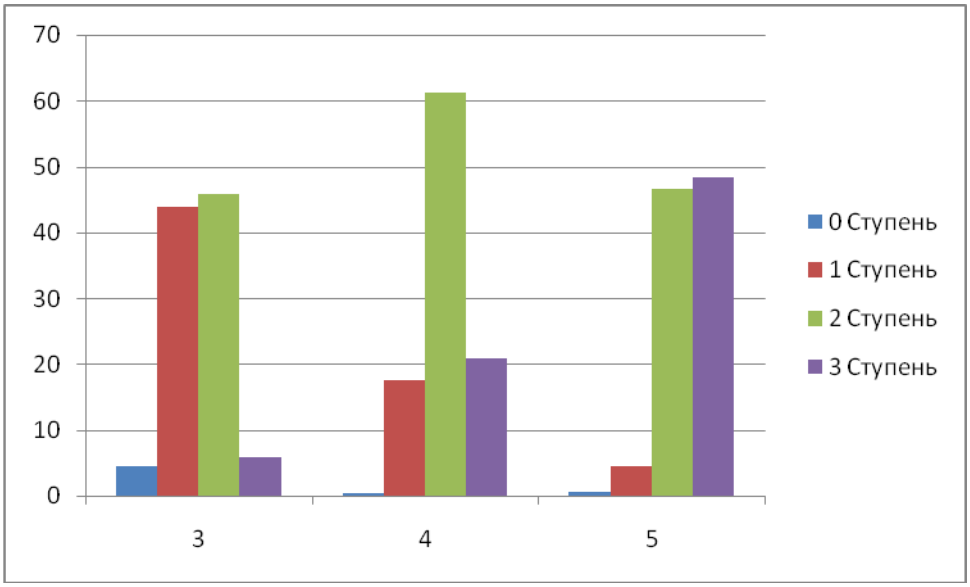

Рис. 3.15. Распределение учащихся по ступеням достижений в зависимости от оценки (математика)

Дополнительно были посчитаны коэффициенты корреляции между тестовым баллом учащихся по математике и ожидаемыми школьными оценками, а также между ступенью, к которой был отнесен учащийся, и его школьной оценкой. Коэффициенты корреляции соответственно равны 0,46 и 0,41. Эти показатели ниже соответствующих показателей для прогностической валидности, что может быть объяснено тем, что учителя опасались ставить высокие ожидаемые оценки учащимся. В пользу этого аргумента свидетельствует и тот факт, что процент ожидаемых 5 не превышает 10% (табл. 3.18), что, на наш взгляд, не соответствует реальному числу учащихся, имеющих 5 по математике и русскому языку по окончании начальной школы.

Далее представлены результаты аналогичного исследования по русскому языку. В табл. 3.22 показано распределение оценок по русскому языку учащихся, находящихся на

различных ступенях, на рис.3.16 это же распределение показано графически. Мы видим, что 84% учащихся, отнесенных ко 0-ой ступени достижений по русскому языку по результатам теста SAM, имеют оценку 3. Из учащихся, отнесенных к 1-ой ступени, 53% имеют оценку 3 и 45% - оценку 4. Большинство учащиеся, отнесенных ко 2-ой ступени (65%), имеют оценку 4. То же справедливо и для учащихся, находящихся на 3-ей ступени, однако 28% этих учащихся имеют оценку 5.

Табл. 3.22. Распределение оценок учащихся, находящихся на различных ступенях, % (русский язык)

Ступень	Оценка		
	3	**4**	**5**
0 Ступень	84	15	1
1 Ступень	53	45	2
2 Ступень	27	65	9
3 Ступень	11	61	28

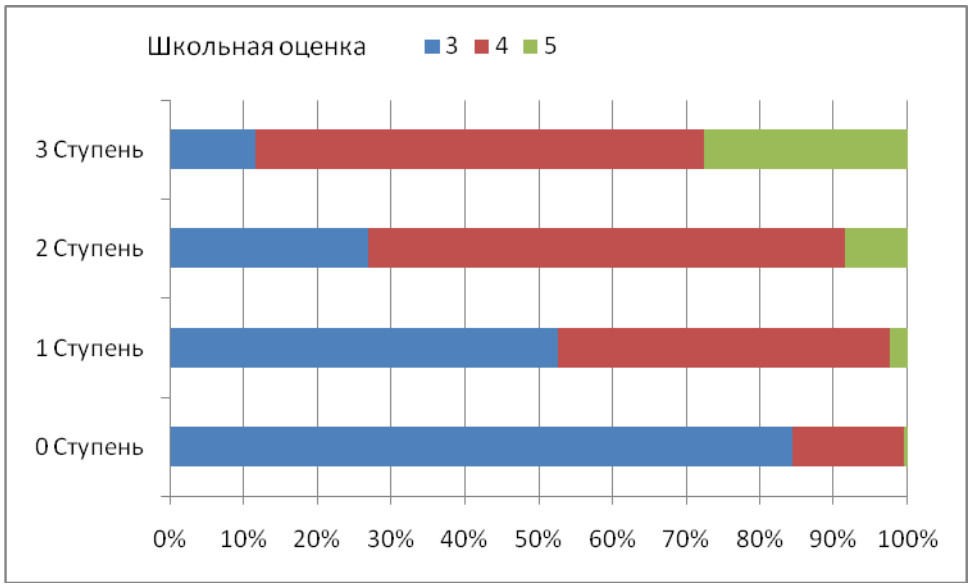

Рис. 3.16. Распределение оценок учащихся в зависимости от ступени достижений (русский язык)

В табл. 3.23 показано распределение по ступеням учащихся, имеющих различные оценки, на рис.3.17 это же распределение показано графически. Мы видим, что из учащихся, имеющих оценку 5, 43% были отнесены ко 2-ой ступени и 44% - к 3-ей; из учащихся, имеющих оценку 4, 49% были отнесены ко 2-ой ступени и 14% - к 3-ой. И наконец, из учащихся, имеющих оценку 3, 21% были отнесены к 0-ой ступени, 51% - к 1-ой и 25% - ко 2-ой ступени.

Табл. 3.23. Распределение по ступеням учащихся, имеющих различные оценки, % (русский язык)

Оценка		Ступень		
	0-ая	**1-ая**	**2-ая**	**3-я**
3	21	51	25	3
4	3	34	49	14
5	1	12	43	44

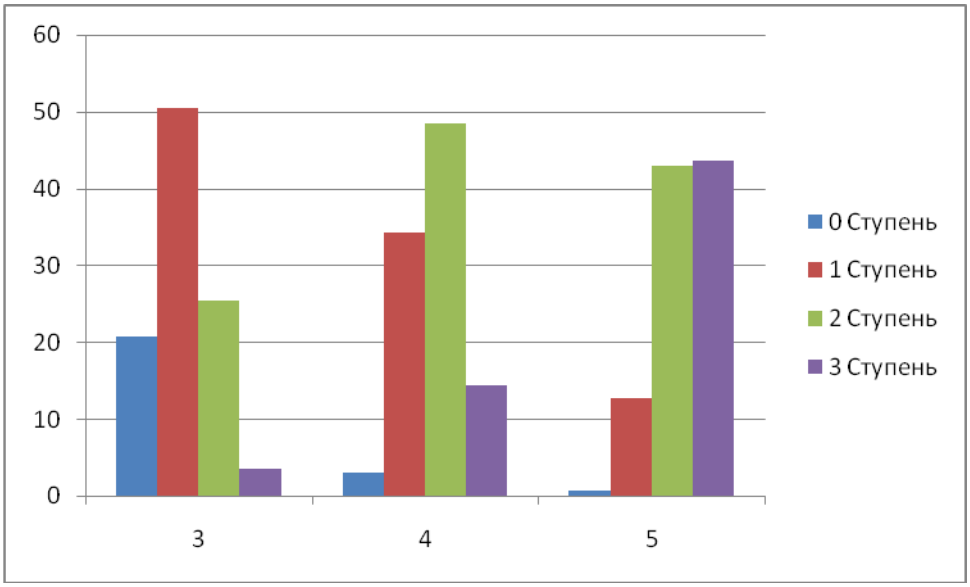

Рис. 3.17. Распределение учащихся по ступеням достижений в зависимости от оценки (русский язык)

Дополнительно были посчитаны коэффициенты корреляции между тестовым баллом учащихся по русскому языку и ожидаемыми школьными оценками, а также между ступенью, к которой был отнесен учащийся, и его школьной оценкой. Коэффициенты корреляции соответственно равны 0,48 и 0,45. Эти показатели близки к аналогичным показателям по математике.

Литература

1. SAM (School Achievement Monitoring): Инструмент мониторинга учебных достижений школьников // *под ред. Нежнова П.Г., Кардановой Е.Ю.*, 2011, 104 с.
2. *Анастази А.*, Урбина С. Психологическое тестирование. СПб.: Питер, 2003. 674 с.

4. Анализ функционирования заданий по отношению к различным группам испытуемых (DIF анализ)

Задание демонстрирует DIF (Differential Item Functioning), если участники тестирования с одинаковым уровнем подготовленности, принадлежащие к различным

группам, имеют различные шансы выполнить задание правильно. Другими словами, задание по-разному функционирует для различных групп тестируемых, и представители одной из групп могут быть несправедливо оценены. Например, пусть задание в тесте по математике содержит слова, непонятные или плохо знакомые определенной группе испытуемых. Тогда вероятность успешного выполнения задания участниками из этой группы будет ниже, чем у других участников, даже при идентичном уровне математической подготовки. Если в тесте таких заданий несколько, то это повлияет на тестовые баллы участников из данной группы: они окажутся существенно ниже. DIF анализ имеет целью выявление заданий, демонстрирующих различное функционирование по отношению к различным группам испытуемых, и степени влияния этого явления на оценки участников тестирования.

Применительно к SAM основаниями для выделения различных групп участников тестирования могут быть пол участника, регион проживания (различные субъекты Российской Федерации), страна проживания (при использовании тестов в РФ и странах СНГ), язык тестирования (при переводе тестов на другие языки), форма тестирования (бланковая или компьютерная). В данном отчете представлены результаты анализа различного функционирования заданий тестов SAM по отношению к группам участников разного пола (мальчики / девочки).

Для моделирования тестов SAM используется дихотомическая модель Раша. В условиях этой модели оценки параметров заданий должны быть инвариантны относительно групп испытуемых, по результатам тестирования которых эти оценки получены (в предположении, что задание находится в хорошем согласии с моделью) [1]. Если задание демонстрирует DIF по отношению к некоторым группам участников, то свойство инвариантности выполнено не будет: оценки трудности этого задания, полученные по результатам тестирования этих групп, будут статистически значимо различаться. На этом основан один из методов обнаружения DIF (который мы применяем в данной работе) - t-тест с отдельными калибровками для каждой группы (separate calibration t-test) [2]. По каждому заданию отдельно рассчитывается статистика:

$$t_i = \frac{\delta_i^{(1)} - \delta_i^{(2)}}{\sqrt{(\sigma_i^{(1)})^2 + (\sigma_i^{(2)})^2}},$$

где $\delta_i^{(1)}$ - оценка трудности i-го задания теста, полученная по результатам выполнения теста первой группой испытуемых, $\delta_i^{(2)}$ - оценка трудности i-го задания теста, полученная по результатам выполнения теста второй группой испытуемых; $\sigma_i^{(1)}$, $\sigma_i^{(2)}$ - среднеквадратические ошибки измерения $\delta_i^{(1)}$ и $\delta_i^{(2)}$. Статистика t_i имеет асимптотически нормальное распределение с нулевым математическим ожиданием и единичной дисперсией. Значения этой статистики, по модулю большие 2 (т.е. $|t_i| > 2$), указывают на слишком большие отклонения оценок трудностей заданий, т.е. на присутствие DIF.

Для сравнения трудностей заданий необходимо, чтобы трудности заданий, полученные в результате двух калибровок, находились на одной шкале. Для установления единой шкалы в данной работе использовался метод «постоянных заданий» (Constant Items) [3], в котором набор «DIF-свободных» заданий служит якорным звеном для установления общей шкалы, и все другие задания исследуются на DIF. Для отбора таких заданий использовались как нестатистические процедуры (содержательный анализ

заданий), так и статистические (сравнение характеристик заданий для всей совокупности данных и для каждой из групп в отдельности).

Следует отметить, что при больших объемах выборки статистика *t* (равно, как и любая другая) может показывать статистически значимые значения даже при незначительном DIF в задании. Поэтому при анализе заданий на DIF устанавливают минимальное различие в трудности задания у двух групп испытуемых, как некоторый порог, определяющий его практическую значимость. В качестве такого порога мы рассматриваем количество DIF, равное 0.5 логита [4].

Дополнительно к *t*-статистике в данной работе применялся один из самых известных методов исследования DIF в классической теории тестирования - метод Mantel-Haenszel (MH), который подразумевает прямое сравнение шансов выполнить задание правильно испытуемыми одного и того же уровня подготовки из различных групп [5]. Статистика MH в условиях справедливости нулевой гипотезы от отсутствии DIF, имеет хи-квадрат распределение с одной степенью свободы.

Реальный тест никогда не является идеальным и всегда в некоторой степени содержит DIF. Необходимо определить, является ли величина DIF относительно небольшой, и тест можно считать инвариантным в практическом смысле (т.е. не искажающим оценки испытуемых из различных групп), или задания, содержащие DIF, вносят существенные искажения в измерение испытуемых. С этой целью в данной работе сравнивались характеристические кривые теста для отдельных групп испытуемых[1]. Предварительно тесты были выровнены методом фиксации параметров заданий, не содержащих DIF. Если характеристические кривые практически совпадают, то наличие заданий с DIF не оказывает существенного влияния на тестовый балл участника тестирования, и результаты можно признать справедливыми относительно различных групп участников. В противном случае, если кривые не совпадают, наличие заданий с DIF оказывает влияние на оценки участников и должно быть дополнительно исследовано.

Таким образом, в данной работе для анализа заданий тестов SAM на наличие DIF по гендерному признаку применялась процедура, состоящая из следующих этапов:

1) установление единой шкалы для двух групп;

2) выявление заданий, подозрительных на наличие DIF, с применением метода «t-тест с отдельными калибровками для каждой группы» и метода Mantel-Haenszel (MH);

3) выявление заданий, демонстрирующих DIF, на основе результатов шага 2 и с учетом порога практической значимости (минимальный порог в разнице трудностей заданий для различных групп участников принят равным 0.5 логита);

4) определение значимости DIF на уровне теста: сравнение характеристических кривых теста для двух групп участников.

Для проведения DIF-анализа использовалась программа Winsteps [6].

4.1. DIF анализ заданий теста по математике

В табл. 4.1 приведены данные о количестве участников тестирования каждого пола и их средних баллах для теста по математике, вариант 1. Данные по анализу варианта 2 аналогичны и в отчете не приведены. На рис.4.1 представлены гистограммы

---

[1] Характеристическая кривая теста представляет собой график функции математического ожидания балла за тест в зависимости от уровня подготовленности учащегося.

распределений тестовых баллов участников из анализируемых групп: розовый цвет – девочки, голубой – мальчики.

Табл. 4.1. Показатели теста для обоих полов (математика, вариант1)

Показатель	Женский пол	Мужской пол
Выборка	1471	1546
Наблюдаемый балл: среднее (SD)	26.7 (8.4)	26.2 (8.3)
Оценка уровня способностей: среднее (SD)	0.69 (1.32)	0.61 (1.11)

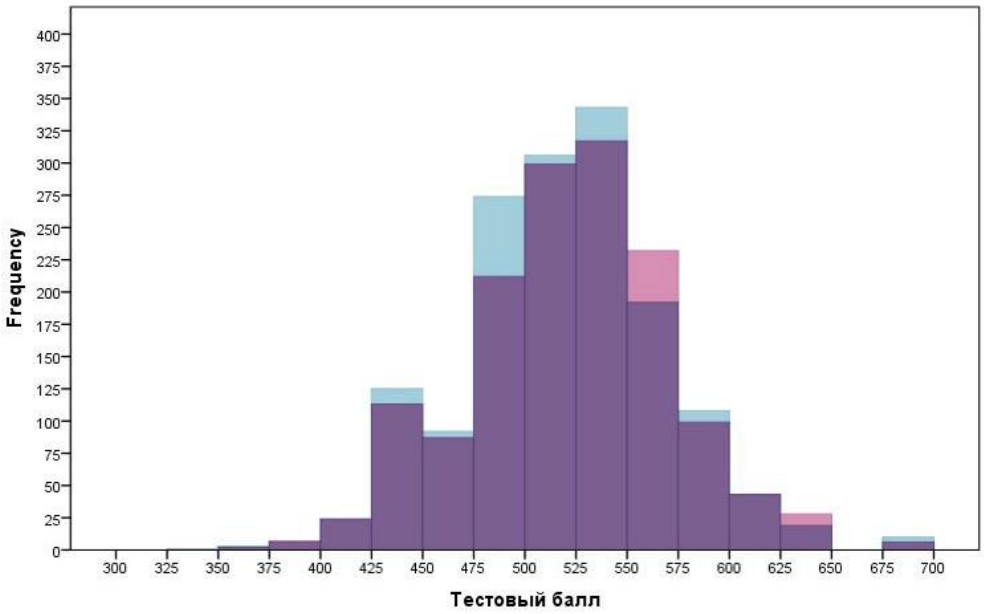

Рис. 4.1. Распределение тестовых баллов участников тестирования анализируемых групп (математика, вариант 1)

Таким образом, результаты девочек и мальчиков в целом очень близки: девочки выполнили тест чуть лучше мальчиков, но эти различия вряд ли значимы (значимость различий будет рассмотрена в главе 5).

Результаты DIF-анализа заданий теста SAM приведены в табл. 4.2. Задания расположены в порядке их следования в тесте. В столбцах 2 и 3 таблицы представлены оценки трудности заданий и соответствующие ошибки измерения (все данные приведены в логитах), полученные по результатам тестирования девочек. В следующих двух столбцах – те же данные, полученные по результатам тестирования мальчиков. В столбцах 6 и 7 приведена разница в оценках трудности заданий для двух групп участников и соответствующая ошибка измерения. В следующем столбце приведены значения *t*-статистики. Для этой статистики допустимыми являются значения из промежутка (-2,+2). И наконец, в последних двух столбцах приведены значения статистики Mantel-Haenzel (MH) и ее уровня значимости (вероятности наблюдения таких данных в предположении, что задание не демонстрирует DIF). Минимальный порог этой вероятности принимается равным 0.05. Значения DIF-статистик, превышающие свои критические значения, выделены в таблице розовым цветом.

Табл. 4.2. Результаты DIF-анализа (математика, вариант1)

№	Девочки		Мальчики		Разница в тр-ти	S.E. разницы	t-статистика	Mantel-Hanzel	
	Трудн.	Ош.изм	Трудн.	Ош.изм.				статистика	Prob.
1	-2.58	0.11	-2.34	0.10	-0.25	0.15	-1.66	3.47	0.06
2	-0.1	0.06	-0.01	0.06	-0.09	0.09	-1.05	1.44	0.23
3	2.37	0.07	2.37	0.07	0	0.1	0	0.1	0.75
4	-3.37	0.15	-3.57	0.16	0.2	0.22	0.93	0.33	0.56
5	-0.4	0.06	-0.55	0.06	0.15	0.09	1.7	1.85	0.17
6	1.57	0.06	1.62	0.06	-0.05	0.09	-0.55	0.26	0.61
7	-1.7	0.08	-1.78	0.08	0.08	0.12	0.67	0.11	0.74
8	-0.65	0.07	-0.65	0.06	0	0.09	0	0.02	0.89
9	1.22	0.06	1.16	0.06	0.06	0.08	0.75	0.94	0.33
10	-1.48	0.08	-1.56	0.08	0.08	0.11	0.75	0.33	0.57
11	-0.17	0.06	-0.12	0.06	-0.05	0.09	-0.58	0.05	0.82
12	0.79	0.06	0.79	0.06	0	0.08	0	0.06	0.81
13	-0.5	0.07	-0.65	0.06	0.16	0.09	1.72	3.32	0.07
14	0.67	0.06	0.62	0.06	0.05	0.08	0.56	1.03	0.31
15	0.81	0.06	0.81	0.06	0	0.08	0	0.52	0.47
16	-1.54	0.08	-1.6	0.08	0.06	0.11	0.49	0	0.95
17	0.43	0.06	0.36	0.06	0.07	0.08	0.8	2.11	0.15
18	1.81	0.06	1.92	0.06	-0.11	0.09	-1.18	1.11	0.29
19	-0.52	0.07	-0.42	0.06	-0.09	0.09	-1.04	1.03	0.31
20	0.45	0.06	0.64	0.06	-0.19	0.08	-2.24	4.47	0.03
21	1.02	0.06	1.06	0.06	-0.04	0.08	-0.54	0.46	0.5
22	-2.58	0.11	-2.46	0.1	-0.13	0.15	-0.83	0.92	0.34
23	-0.19	0.06	-0.07	0.06	-0.12	0.09	-1.42	1.92	0.17
24	2.1	0.07	2.19	0.07	-0.09	0.1	-0.9	1.43	0.23
25	-0.94	0.07	-1.09	0.07	0.15	0.1	1.48	0.97	0.33
26	0.71	0.06	0.71	0.06	0	0.08	0	0.17	0.68
27	1.91	0.06	1.79	0.06	0.12	0.09	1.31	1.14	0.29
28	-0.95	0.07	-1.13	0.07	0.18	0.1	1.82	2.16	0.14
29	0.51	0.06	0.24	0.06	0.27	0.08	3.19	12.07	0
30	1.26	0.06	1.21	0.06	0.04	0.08	0.5	0.82	0.37
31	-1.01	0.07	-1.11	0.07	0.1	0.1	1.01	0.33	0.57
32	-0.5	0.07	-0.44	0.06	-0.06	0.09	-0.68	0.2	0.66
33	1.51	0.06	1.51	0.06	0	0.09	0	0	0.98
34	-1.93	0.09	-1.86	0.08	-0.08	0.12	-0.61	1.58	0.21
35	-0.03	0.06	-0.2	0.06	0.18	0.09	2.06	3.66	0.06
36	1.93	0.07	1.93	0.07	0	0.09	0	0.04	0.85
37	-1.68	0.08	-1.26	0.07	-0.42	0.11	-3.79	16.19	0
38	0	0.06	0	0.06	0	0.09	0	0.07	0.8
39	1.05	0.06	1.18	0.06	-0.13	0.09	-1.55	2.88	0.09
40	-1.09	0.07	-1.09	0.07	0	0.1	0	0	0.97
41	0.07	0.06	0.28	0.06	-0.21	0.09	-2.45	6.67	0.01
42	0.7	0.06	0.96	0.06	-0.27	0.09	-3.1	8.37	0
43	-1.24	0.08	-1.6	0.08	0.37	0.11	3.3	11.5	0

44	0.91	0.06	0.8	0.06	0.11	0.09	1.28	3	0.08
45	1.38	0.07	1.25	0.07	0.14	0.09	1.43	3.22	0.07

Всего выделено 7 заданий, значения хотя бы одной статистики которых превышают критические: 5 из них функционируют в пользу девочек и 2 – в пользу мальчиков. Зеленым цветом выделены значения разности трудностей заданий для групп девочек и мальчиков. Ни одно из этих значений не превышает принятый порог практической значимости (0.5 логитов), что говорит о незначительном расхождении в трудности заданий для анализируемых групп учащихся. Незначительное превышение критических значений может быть объяснено большим объемом выборки и случайными факторами. Это подтверждается и содержательным анализом самих заданий, отмеченных как функционирующие в пользу мальчиков или девочек

В подтверждение сказанного на рис.4.2. представлены распределения трудностей заданий отдельно для выборок мальчиков и девочек.

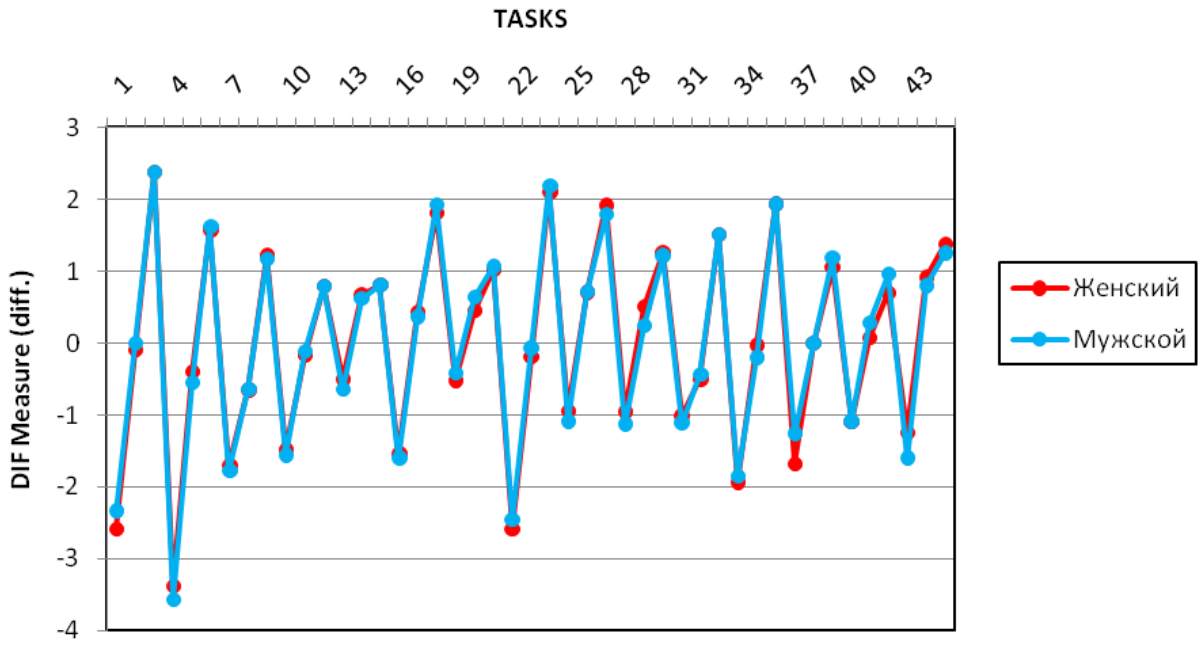

Рис. 4.2. График распределения трудностей заданий для различных групп учащихся (математика, вариант 1)

На рис. 4.3 приведена характеристическая кривая задания 37, у которого значение *t*-статистики самое большое по модулю (-3.79), равно как и самое большое значение разности между трудностями задания для двух групп (-0.42). Это задание функционирует в пользу девочек. Таким образом, это задание самое плохое в смысле демонстрации DIF. Характеристическая кривая задания представляет собой вероятность (в соответствии с используемой моделью) правильного ответа на задание в зависимости от уровня подготовленности испытуемых (зеленая линия на рисунке). Крестиками на рисунке обозначены точки эмпирического распределения ответов испытуемых на данное задание: красными крестиками помечены ответы мальчиков, синими – девочек. Они представляют собой средний балл по данному заданию по группам испытуемых с различным уровнем подготовки.

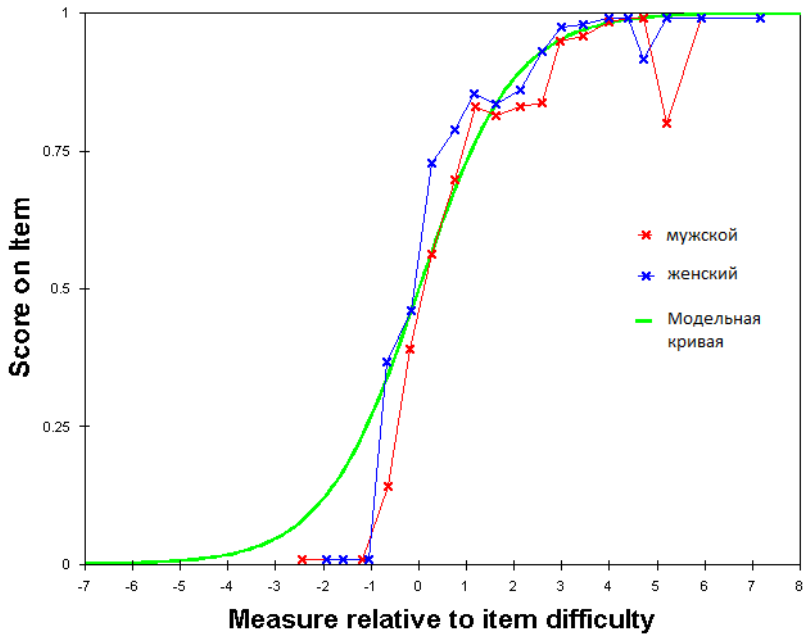

Рис.4.3. Характеристическая кривая и точки эмпирического распределения для двух групп участников для задания 37 (математика, вариант 1)

Для сравнения на рис. 4.4 приведена характеристическая кривая задания 12, у которого оценки трудности для двух групп учащихся совпадают и, таким образом, это задание полностью свободно от DIF.

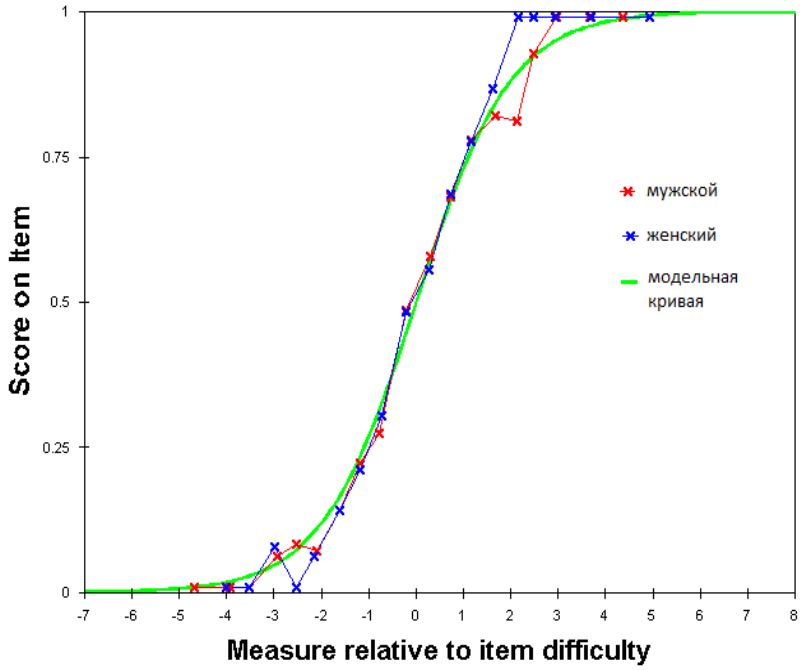

Рис.4.4. Характеристическая кривая и точки эмпирического распределения для двух групп участников для задания 12 (математика, вариант 1)

Мы видим, что незначительные различия в эмпирической трудности заданий наблюдаются всегда. Однако, как показали результаты исследования, в данном случае эти различия не являются значимыми в практическом смысле.

В подтверждение того, что небольшие различия в трудности заданий не оказывают влияния на результаты тестирования представителей обеих анализируемых групп, на рис. 4.5 показаны характеристические кривые теста для обеих групп (синяя линия – для мальчиков, красная – для девочек). В качестве якорных заданий между двумя выборками были выбраны задания 3, 8, 12, 15, 26, 31, в которых по результатам предварительного исследования не было обнаружено DIF. Характеристические кривые теста для мальчиков и девочек практически совпадают, что означает, что наличие DIF не оказывает влияния на тестовый балл испытуемого, а, следовательно, на его ступень.

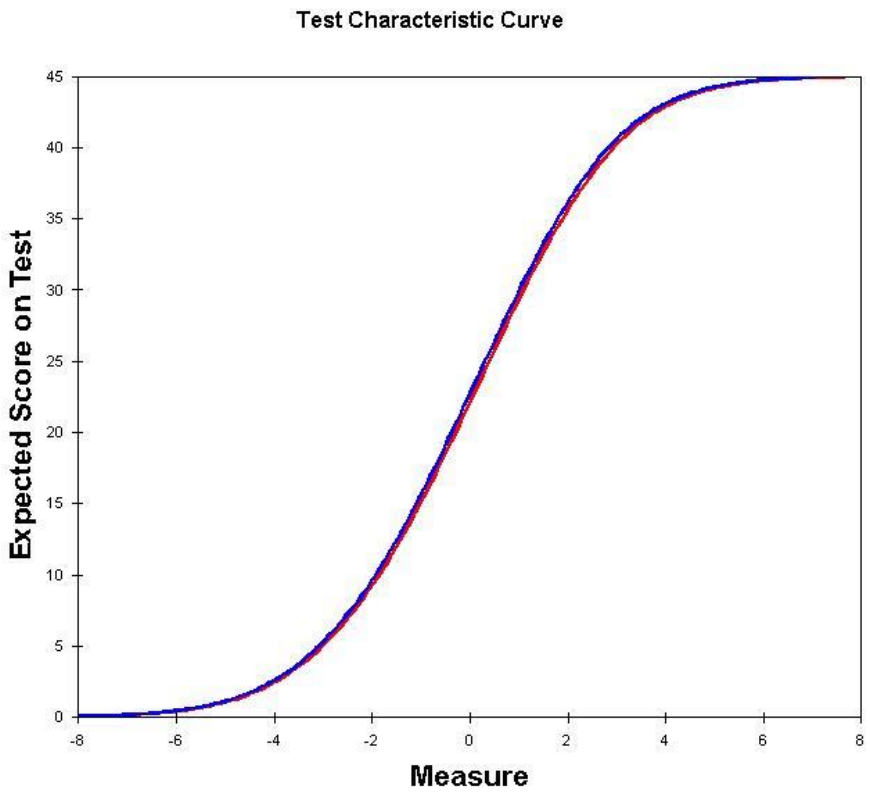

Рис. 4.5. Характеристические кривые теста для двух анализируемых групп учащихся (математика, вариант 1)

Таким образом, результаты анализа заданий теста по математике однозначно говорят, что задания не демонстрируют различного функционирования по отношению к представителям разного пола.

4.2. DIF анализ заданий теста по русскому языку

В табл. 4.3 приведены данные о количестве участников тестирования каждого пола и их средних баллах для теста по русскому языку, вариант 1. Данные по анализу варианта 2 аналогичны и в отчете не приведены. На рис.4.6 представлены гистограммы распределений тестовых баллов участников из анализируемых групп: розовый цвет – девочки, голубой – мальчики.

Таблица 4.3. Показатели теста для обоих полов (русский язык, вариант1)

Показатель	Женский пол	Мужской пол
Выборка	1440	1537
Наблюдаемый первичный балл: среднее (SD)	24.7 (8.5)	22.1 (8.8)
Оценка уровня способностей: среднее (SD)	0.34 (1.22)	-0.04 (1.26)

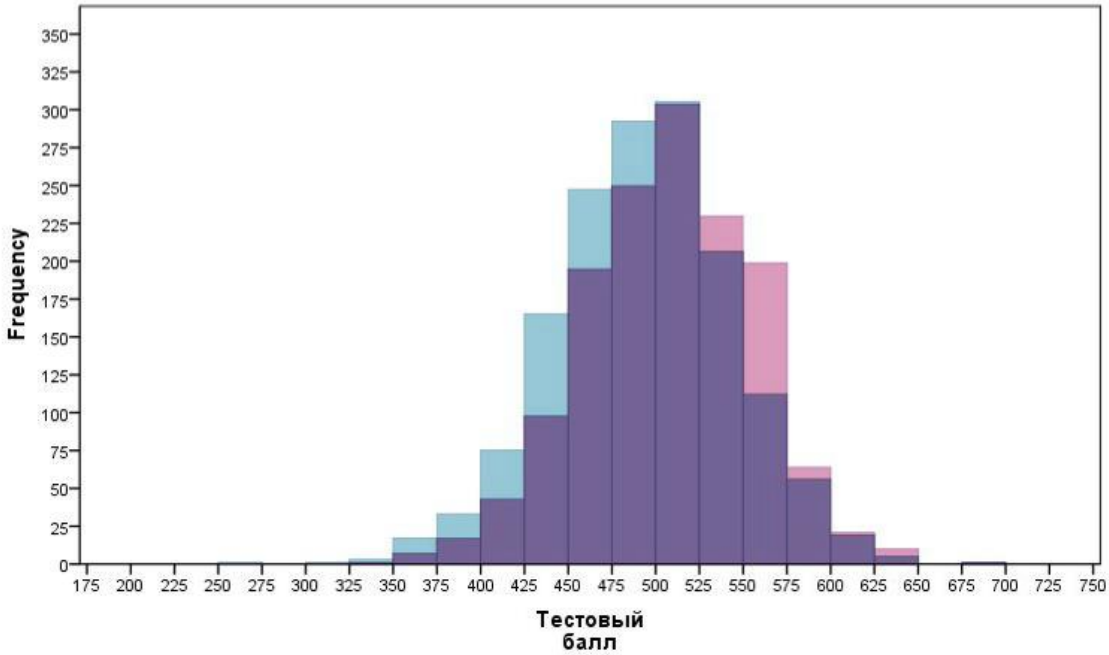

Рис. 4.6. Распределение тестовых баллов участников тестирования анализируемых групп (русский язык, вариант 1)

Таким образом, девочки демонстрируют несколько лучшие результаты по русскому языку по сравнению с мальчиками (значимость различий будет рассмотрена в главе 5).

Результаты DIF-анализа заданий теста SAM приведены в табл. 4.4. Задания расположены в порядке их следования в тесте. В столбцах 2 и 3 таблицы представлены оценки трудности заданий и соответствующие ошибки измерения (все данные приведены в логитах), полученные по результатам тестирования девочек. В следующих двух столбцах – те же данные, полученные по результатам тестирования мальчиков. В столбцах 6 и 7 приведена разница в оценках трудности заданий для двух групп участников и соответствующая ошибка измерения. В следующем столбце приведены значения *t*-статистики. Для этой статистики допустимыми являются значения из промежутка (-2,+2). И наконец, в последних двух столбцах приведены значения статистики Mantel-Haenzel (MH) и ее уровня значимости (вероятности наблюдения таких данных в предположении, что задание не демонстрирует DIF). Минимальный порог этой вероятности принимается равным 0.05. Значения DIF-статистик, превышающие свои критические значения, выделены в таблице розовым цветом.

Табл. 4.4. Результаты DIF-анализа (русский язык, вариант1)

№	Девочки		Мальчики		Разница в тр-ти	S.E. разницы	t-статистика	Метод Mantel-Hanzel	
	Трудн.	Ош.изм	Трудн.	Ош.изм.				Ст-ка хи-квадрат	Prob.
1	-2,26	0,09	-2,01	0,07	-0.25	0.12	-2.14	3.93	0.05
2	-1,74	0,08	-1,68	0,07	-0.06	0.11	-0.56	0.87	0.35
3	0,86	0,06	0,94	0,06	-0.08	0.09	-0.9	0.33	0.57
4	-3,79	0,16	-3,51	0,12	-0.28	0.2	-1.45	1.92	0.17
5	-0,81	0,07	-0,81	0,06	0	0.09	0	0.22	0.64
6	0,92	0,06	0,92	0,06	0	0.09	0	0.32	0.57
7	-1,27	0,07	-1,34	0,07	0.07	0.1	0.68	0.00	0.99
8	0,84	0,06	0,65	0,06	0.19	0.09	2.21	4.34	0.04
9	1,33	0,06	1,39	0,07	-0.06	0.09	-0.67	0.23	0.63
10	-1,04	0,07	-1,33	0,07	0.29	0.09	3.07	3.71	0.05
11	-0,87	0,07	-0,76	0,06	-0.11	0.09	-1.22	0.46	0.50
12	1,64	0,07	1,57	0,07	0.07	0.09	0.78	0.11	0.74
13	-1,7	0,08	-1,59	0,07	-0.11	0.1	-1.03	1.36	0.24
14	1,01	0,06	1,14	0,06	-0.14	0.09	-1.55	1.90	0.17
15	0,57	0,06	0,48	0,06	0.08	0.08	1	1.27	0.26
16	-1,88	0,08	-1,67	0,07	-0.21	0.11	-1.98	2.73	0.10
17	0,65	0,06	0,56	0,06	0.09	0.08	1.05	1.52	0.22
18	1,06	0,06	1,16	0,06	-0.1	0.09	-1.08	0.51	0.48
19	-1,69	0,08	-1,49	0,07	-0.2	0.1	-1.95	1.57	0.21
20	-0,32	0,06	-0,6	0,06	0.27	0.09	3.19	1.79	0.18
21	0,96	0,06	0,96	0,06	0	0.09	0	1.17	0.28
22	-2,31	0,09	-1,95	0,07	-0.36	0.12	-3.07	9.91	0.002
23	0,15	0,06	0,15	0,06	0	0.08	0	0.59	0.44
24	1,42	0,06	1,35	0,07	0.07	0.09	0.73	2.25	0.13
25	-0,71	0,06	-0,84	0,06	0.13	0.09	1.46	4.82	0.03
26	-0,16	0,06	-0,24	0,06	0.08	0.08	0.89	5.76	0.02
27	1,87	0,07	1,83	0,07	0.04	0.1	0.43	0.10	0.75
28	-2,49	0,1	-2,23	0,08	-0.26	0.13	-2.06	3.06	0.08
29	-0,56	0,06	-0,8	0,06	0.24	0.09	2.73	5.84	0.02
30	1,36	0,06	1,24	0,06	0.12	0.09	1.32	0.41	0.52
31	-0,14	0,06	-0,14	0,06	0	0.08	0	0.05	0.82
32	0,91	0,06	0,63	0,06	0.27	0.09	3.19	7.11	0.01
33	1,68	0,07	1,68	0,07	0	0.1	0	0.92	0.34
34	-0,56	0,06	-0,56	0,06	0	0.09	0	0.42	0.52
35	1,24	0,06	1,24	0,06	0	0.09	0	0.15	0.70
36	3,47	0,11	3,17	0,11	0.3	0.15	1.96	2.26	0.13
37	-1,4	0,07	-1,18	0,06	-0.23	0.1	-2.33	5.44	0.02
38	0,77	0,06	0,82	0,06	-0.05	0.09	-0.57	0.23	0.63
39	1,8	0,07	2,01	0,07	-0.21	0.1	-2.07	3.59	0.06
40	-1,53	0,08	-1,3	0,06	-0.23	0.1	-2.3	2.67	0.10
41	-0,08	0,06	-0,01	0,06	-0.07	0.08	-0.87	0.07	0.80
42	1,47	0,06	1,55	0,07	-0.08	0.09	-0.84	0.02	0.88

43	-1,33	0,07	-1,21	0,06	-0.12	0.1	-1.26	0.67	0.42
44	1,06	0,06	0,97	0,06	0.09	0.09	1.02	0.47	0.49
45	1,05	0,06	1,05	0,06	0	0.09	0	1.53	0.22

Всего выделено 11 заданий, значения хотя бы одной статистики которых превышают критические: 6 из них функционируют в пользу девочек и 5 – в пользу мальчиков. Зеленым цветом выделены значения разности трудностей заданий для групп девочек и мальчиков. Ни одно из этих значений не превышает принятый порог практической значимости (0.5 логитов), что говорит о незначительном расхождении в трудности заданий для анализируемых групп учащихся. Незначительное превышение критических значений может быть объяснено большим объемом выборки и случайными факторами. Это подтверждается и содержательным анализом самих заданий, отмеченных как функционирующие в пользу мальчиков или девочек

В подтверждение сказанного на рис.4.7. представлены распределения трудностей заданий отдельно для выборок мальчиков и девочек.

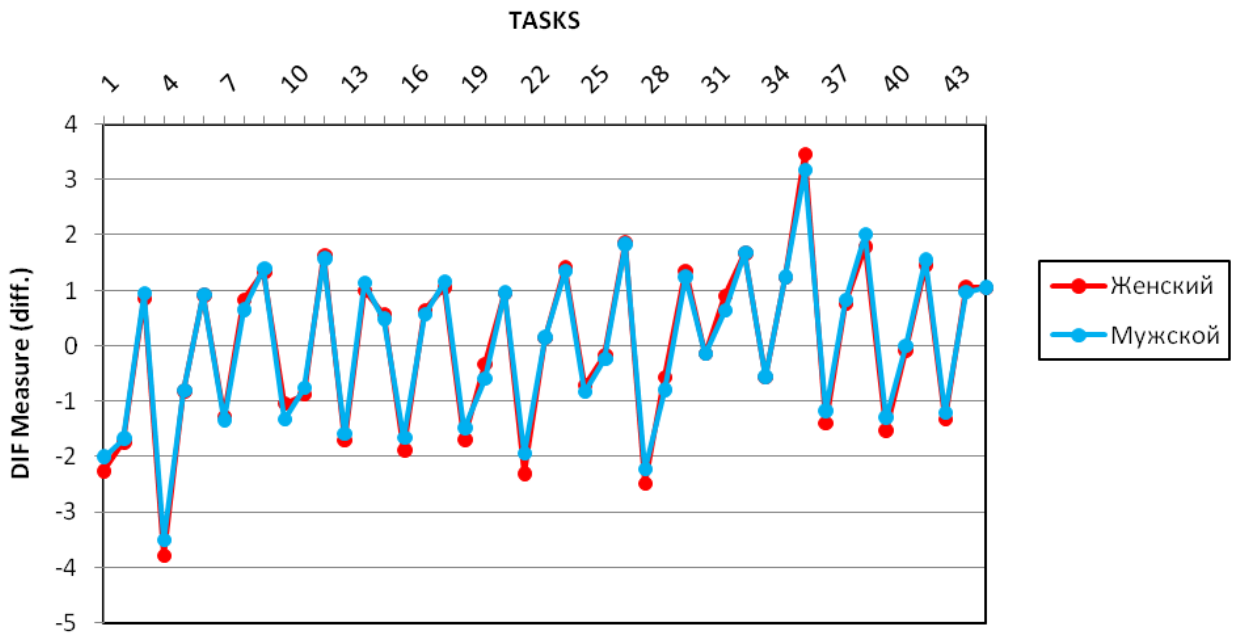

Рис. 4.7. График распределения трудностей заданий для различных групп учащихся (русский язык, вариант 1)

На рис. 4.8 приведена характеристическая кривая задания 32, у которого значение *t*-статистики самое большое по модулю (3.19), равно как и самое большое значение разности между трудностями задания для двух групп (0.27). Это задание функционирует в пользу мальчиков. Таким образом, это задание самое плохое в смысле демонстрации DIF. Характеристическая кривая задания представляет собой вероятность (в соответствии с используемой моделью) правильного ответа на задание в зависимости от уровня подготовленности испытуемых (зеленая линия на рисунке). Крестиками на рисунке обозначены точки эмпирического распределения ответов испытуемых на данное задание: красными крестиками помечены ответы мальчиков, синими – девочек. Они представляют собой средний балл по данному заданию по группам испытуемых с различным уровнем подготовки.

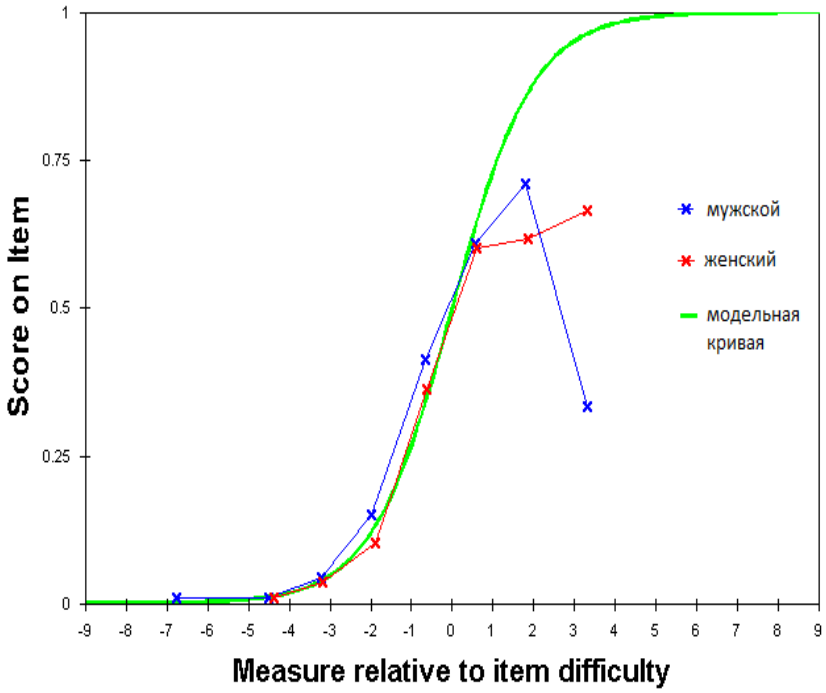

Рис.4.8. Характеристическая кривая и точки эмпирического распределения для двух групп участников для задания 32 (русский язык, вариант 1)

Для сравнения на рис. 4.9 приведена характеристическая кривая задания 5, у которого оценки трудности для двух групп учащихся совпадают и, таким образом, это задание полностью свободно от DIF.

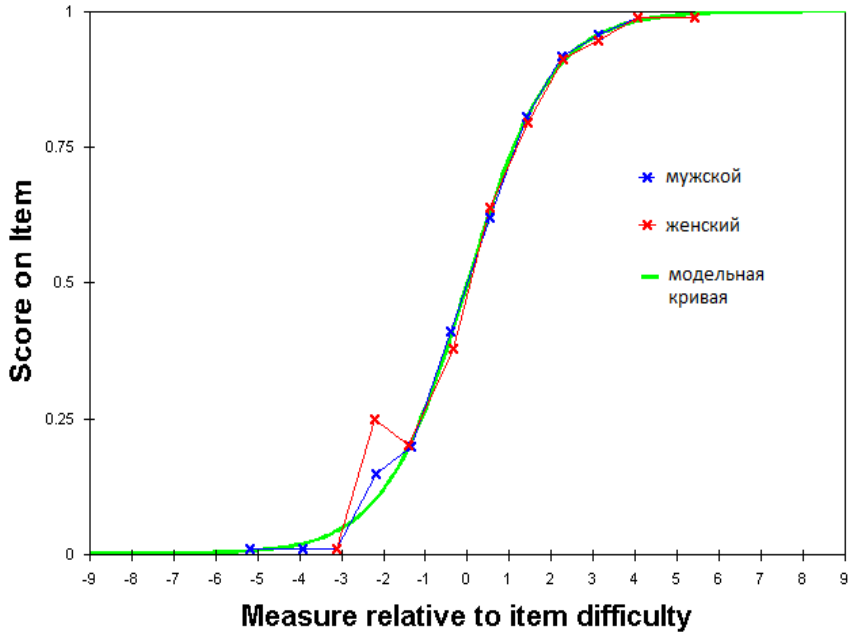

Рис.4.9. Характеристическая кривая и точки эмпирического распределения для двух групп участников для задания 5 (русский язык, вариант 1)

Мы видим, что незначительные различия в эмпирической трудности заданий наблюдаются всегда. Более того, у задания 32 наблюдается выброс в части высоких

баллов, когда учащиеся с высоким уровнем способности, выполнили задание хуже, чем от низ ожидалось. Причем выброс наблюдается и у девочек, и у мальчиков, т.е. независимо от пола участника. Однако таких учащихся очень мало (в противном случае задание не находилось бы в хорошем согласии с моделью измерения, и это было бы зафиксировано ранее, в главе 2). Таким образом, это задание необходимо проанализировать на предмет чистоты его формулировки, но с точки зрения DIF оно функционирует удовлетворительно - как показали результаты исследования, различия в трудности не являются значимыми в практическом смысле.

В подтверждение того, что небольшие различия в трудности заданий не оказывают влияния на результаты тестирования представителей обеих анализируемых групп, на рис. 4.10 показаны характеристические кривые теста для обеих групп (синяя линия – для мальчиков, красная – для девочек). В качестве якорных заданий между двумя выборками были выбраны задания 12, 13, 17, 35, 41, 42, в которых по результатам предварительного исследования не было обнаружено DIF. Характеристические кривые теста для мальчиков и девочек практически совпадают, что означает, что наличие DIF не оказывает влияния на тестовый балл испытуемого, а, следовательно, на его ступень.

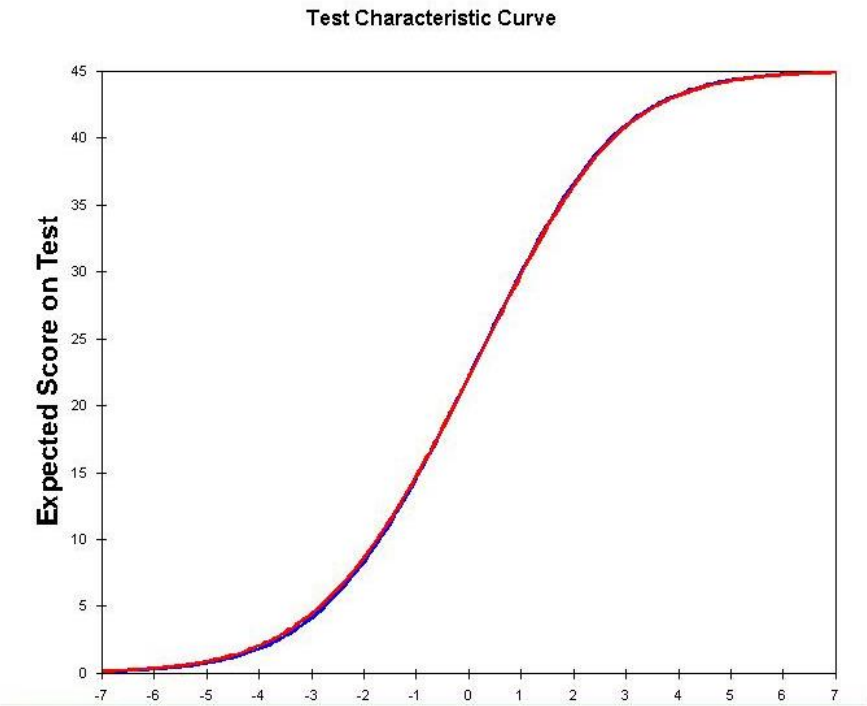

Рис. 4.10. Характеристические кривые теста для двух анализируемых групп учащихся (русский язык, вариант 1)

Таким образом, результаты анализа заданий теста по русскому позволяют сделать вывод, что задания не демонстрируют различного функционирования по отношению к представителям разного пола.

Литература

1. Smith, R. M., and Suh, K. K. (2003). Rasch fit statistics as a test of the invariance of item parameter estimates. Journal of Applied Measurement, 4, 153-163.

2. Smith. R. M. (2004). Detecting item bias with the Rasch model. Journal of Applied Measurement, 5(4), 430-449
3. Wang, W.-C. (2004). Effects of anchor item methods on the detection of differential item functioning within the family of Rasch models. Journal of Experimental Education, 72, 221-261.
4. Wang, W.-C. (2008). Assessment of differential item functioning. Journal of Applied Measurement, 9(4), 387-408
5. Dorans N.J. (1989). Two New Approaches to Assessing Differential Item Functioning: Standardization and the Mantel-Haenszel Method. Applied Measurement in Education, 2(3), 217-233.
6. Linacre J. M. (2011). A User's Guide to WINSTEPS. Program Manual 3.71.0. (http://www.winsteps.com/a/winsteps.pdf).

5. Оценивание участников тестирования SAM

Оценивание участников тестирования происходит с использованием двух подходов - нормативно- и критериально-ориентированного, совмещенных в соответствии с современной теорией тестирования.

В рамках нормативно-ориентированного подхода каждому участнику тестирования в результате математической обработки результатов (в качестве модели тестирования используется дихотомическая модель Раша) присваивается интегральный тестовый балл. Тестовые баллы всех участников тестирования находятся на единой метрической шкале, независимо от времени прохождения теста и конкретного набора выполненных заданий. Для сообщения тестовых баллов участникам тестирования и заинтересованным лицам используется 1000-балльная шкала со средним около 500 и стандартным отклонением 50. В будущем все оценки, полученные в результате практического применения инструмента SAM, будут помещаться на эту же шкалу, что даст возможность сравнивать предметные образовательные результаты каждого участника во времени, т.е. осуществлять мониторинг достижений.

Для реализации критериально-ориентированного подхода разработан ступенчатый вариант шкалы образовательных результатов, основанный на интегральных баллах участников тестирования и пороговых значениях, делящих всех участников на группы, соответствующие различным качественным ступеням. Это обеспечивает возможность качественной оценки предметных образовательных результатов через указание ведущего типа ориентировки в решении задач разных уровней. Методы построения единой шкалы и установления пороговых оценок подробно описаны в [1].

Разработчиками инструмента было принято, что уровень можно рассматривать как достигнутый, если по крайней мере 50% заданий данного уровня выполнены верно. (Разумеется, речь идет о вероятностных оценках.) Всего выделено 4 категории достижений, которые отвечают следующим содержательным критериям:

Ниже первой ступени – не освоен даже первый уровень: учащийся выполняет менее 50% заданий 1-го уровня.

Первая ступень – освоен первый уровень: учащийся выполняет не менее 50% заданий 1-го уровня.

Вторая ступень – освоен второй уровень: учащийся выполняет не менее 50% заданий 2-го уровня.

Третья ступень – освоен третий уровень: учащийся выполняет не менее 50% заданий 3-го уровня.

Для математики имеем следующие пороговые оценки: переход на 1-ую ступень – 430 баллов; на 2-ую ступень 2 – 500 баллов; на 3-ю ступень – 570 баллов. Пороговая оценка определяет нижнюю границу соответствующей ступени. Если тестовый балл участника «перешагнул» этот порог, соответствующая ступень рассматривается как достигнутая. Это означает, что с вероятностью выше 50% данный участник выполнит более 50% заданий данного уровня. Все участники, результат тестирования которых находится ниже данной границы, считаются не достигшими этой ступени, равно как и всех последующих. Интерпретация ступеней на примере теста по математике показана на рис.5.1.

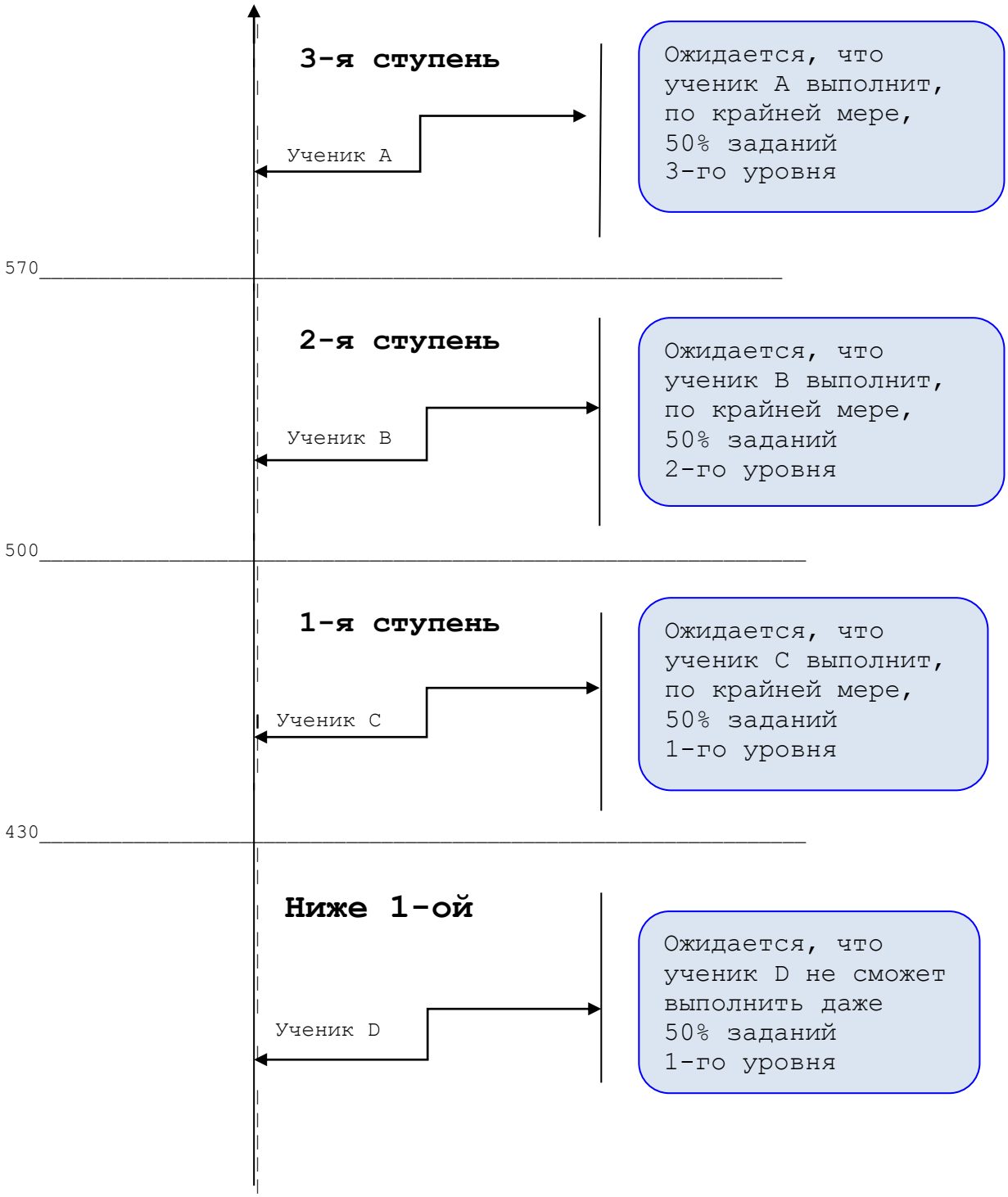

Рис. 05.1. Категоризация шкалы образовательных результатов по математике

Таким образом, каждому участнику тестирования, помимо интегрального тестового балла определяется ступень, которой он достиг. Если учащийся не достиг даже 1-ой ступени, он выполняет менее 50% заданий 1-го уровня. При этом вероятность выполнения им заданий 2-го и 3-го уровней практически равна нулю. Если учащийся достиг 1-ой ступени (но не достиг вышестоящих ступеней), то он выполняет не менее 50% заданий 1-го уровня. При этом он выполняет менее 50% заданий 2-го уровня, а вероятность выполнения заданий 3-го уровня совсем незначительна. Далее, если учащийся достиг 2-ой ступени (но не достиг 3-ей), то он выполняет не менее 50% заданий 2-го уровня. При этом он может выполнить большинство заданий 1-го уровня, но менее 50% заданий 3-го уровня. И, наконец, если учащийся достиг 3-ей ступени, он выполняет не менее 50% заданий 3-го уровня. При этом он почти наверняка выполнит любое задание 1-го уровня и большинство заданий 2-го уровня. Приведенное деление участников тестирования по ступеням достижений позволяет интерпретировать результаты в терминах уровневой модели оценивания, т.е. дать интегральному баллу содержательную интерпретацию.

Рис. 5.2 представляет собой карту переменных одного из вариантов теста по математике (структура карты переменных описана в разделе 2).

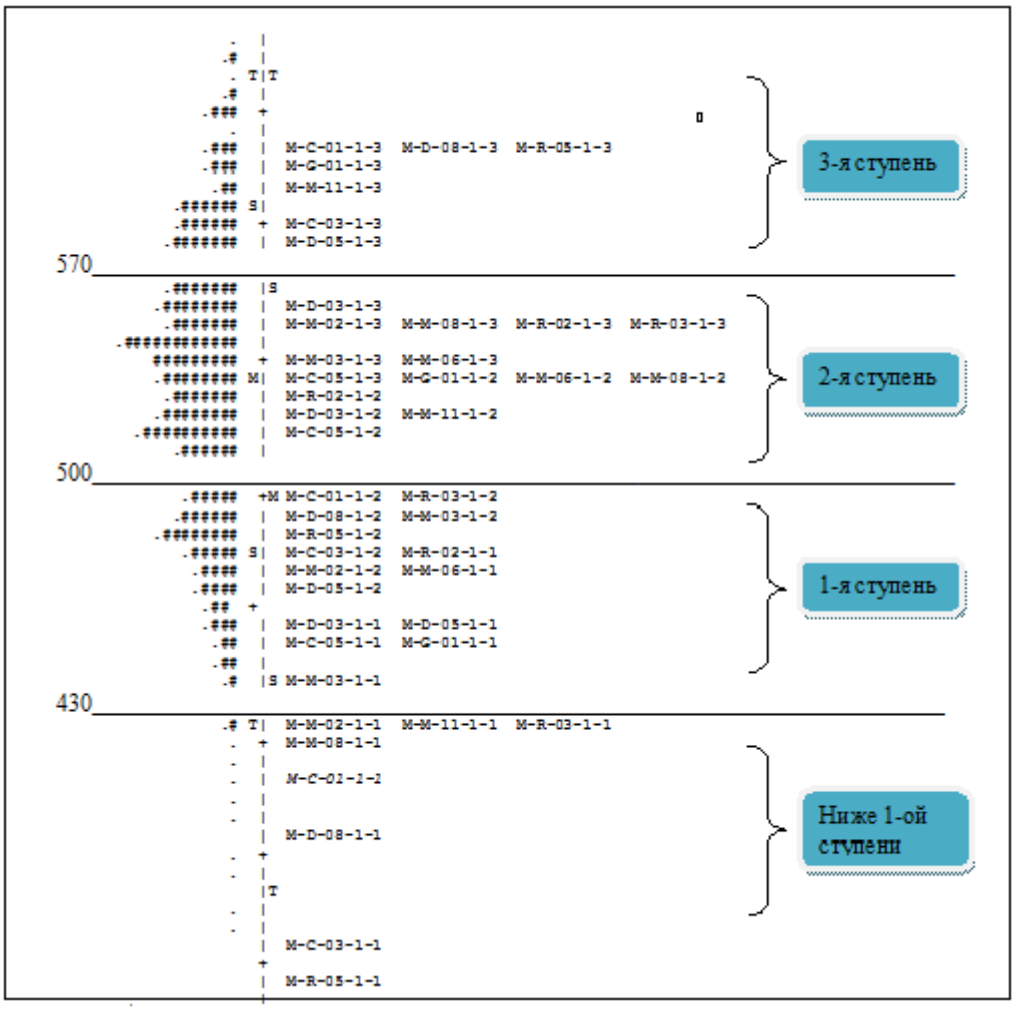

Рис. 5.2. Карта переменных (математика, вариант 1)

На карте горизонтальными линиями показаны границы между ступенями, соответствующие пороговым оценкам. Таким образом, большинство участников тестирования находятся на 1-ой и 2-ой ступенях, при доминировании 2-ой. Это соответствует ожиданиям разработчиков модели оценивания, согласно которым к концу начальной школы учебное содержание усваивается на 2-ом (рефлексивном) уровне.

Аналогичным образом оцениваются участники тестирования по русскому языку.

Установление пороговых оценок и деление участников тестирования на группы позволяет дать интегральному баллу качественную интерпретацию. Для подтверждения обоснованности выводов о степени подготовки участников тестирования было проведено исследование по валидизации указанных пороговых оценок. Оно включало следующие направления:

- Проверка соответствия используемой модели тестирования отдельно каждой группе участников (на каждой ступени достижений)
- Установление пороговых оценок другим методом и сравнение результатов (с этой целью пороговые оценки были установлены методом Ангоф, часто используемым в традиционном тестировании)
- Сравнение коэффициентов решаемости заданий, ожидаемых в соответствии с моделью, с эмпирическими коэффициентами решаемости этих заданий для испытуемых, находящихся на разных ступенях
- Оценка ошибки измерения пороговых оценок.

Результаты исследования представлены в [2]. Основной вывод исследования: пороговые оценки установлены валидно и могут использоваться для качественной интерпретации результатов тестирования.

Установленные пороговые оценки основаны на трехуровневой модели оценивания SAM и позволяют каждому участнику тестирования определить достигнутую им ступень достижений. Таких ступеней может быть достигнуто три – в соответствии с тремя уровнями модели. Однако степень достижения ступени может быть различной: учащийся может только-только достигнуть ступени, «зацепиться» за нее, а может уже ее освоить, т.е. стоять на ней уверенно. Напомним, что ступень считается достигнутой, если, по крайней мере, 50% заданий данного уровня выполнены правильно. Будем считать, что ступень не только достигнута, но и освоена, если правильно выполнены не менее 75% заданий данного уровня. Такая интерпретация результатов тестирования позволит выделить учащихся, уверенно стоящих на ступени и готовых к переходу на следующую ступень. Иллюстрация достижения и освоения ступеней показана на рис.5.3.

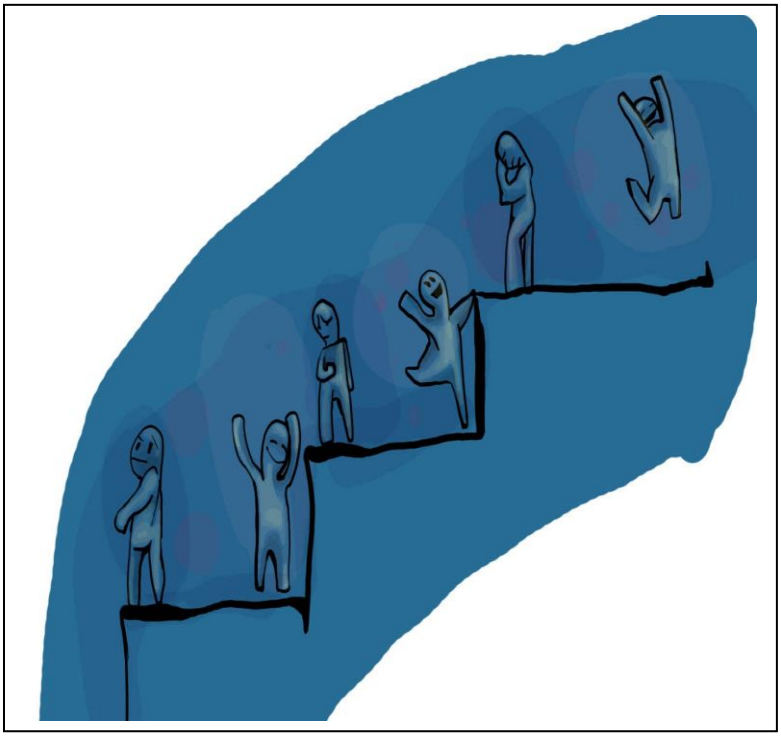

Рис. 5.3. Иллюстрация различной степени достижения ступеней

На рис.5.4 показано распределение участников апробации 2012 г. по математике (выборка около 6000 человек) по ступеням достижений: 2% не достигли 1-ой ступени, а половина всех участников находится на 2-ой ступени.

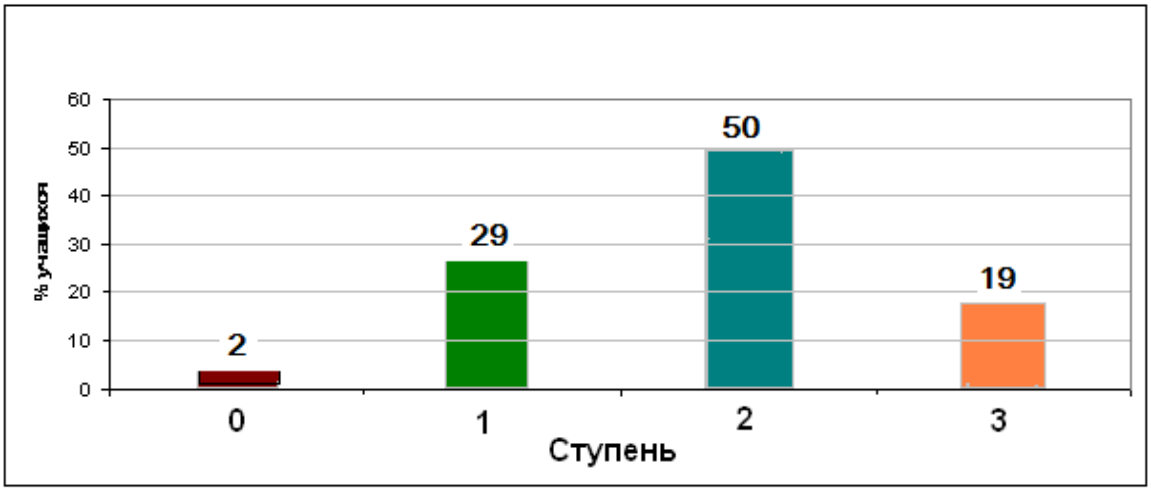

Рис. 5.4. Распределение участников тестирования по ступеням достижений (математика)

В табл. 5.1 приведены данные о распределении участников по ступеням достижений с учетом степени освоения ступени.

Табл. 5.1. Распределение участников тестирования по ступеням (математика)

Ступень	Всего человек	Процент	Степень освоения	Всего человек	Процент от общего	Процент от ступени
1	1725	29	достиг	665	11	39
			освоил	1060	18	61
2	2974	50	достиг	1864	32	63
			освоил	1110	18	37
3	1124	19	достиг	760	13	68
			освоил	364	6	32

Таким образом, мы видим, что из 2974 учащихся, находящихся на 2-ой ступени, освоили ее только 37%. Аналогично, из 1124 учащихся, находящихся на 3-ей ступени, освоили ее только 32%. Что касается 1-ой ступени, то большинство учащихся (61%), находящихся на ней, уже ее освоили.

На рис.5.5 распределение учащихся по ступеням достижений с учетом степени освоения ступени показано графически. Такая интерпретация результатов тестирования может быть полезна для более тонкой оценки достижений учащихся.

Рис. 5.5. Распределение участников тестирования по ступеням достижений с учетом меры освоения ступени (математика)

На рис.5.6 показано распределение участников апробации 2012 г. по русскому языку (выборка около 6000 человек) по ступеням достижений. Таким образом, 11% не достигли 1-ой ступени, и примерно поровну участников находятся на 1-ой и 2-ой ступенях.

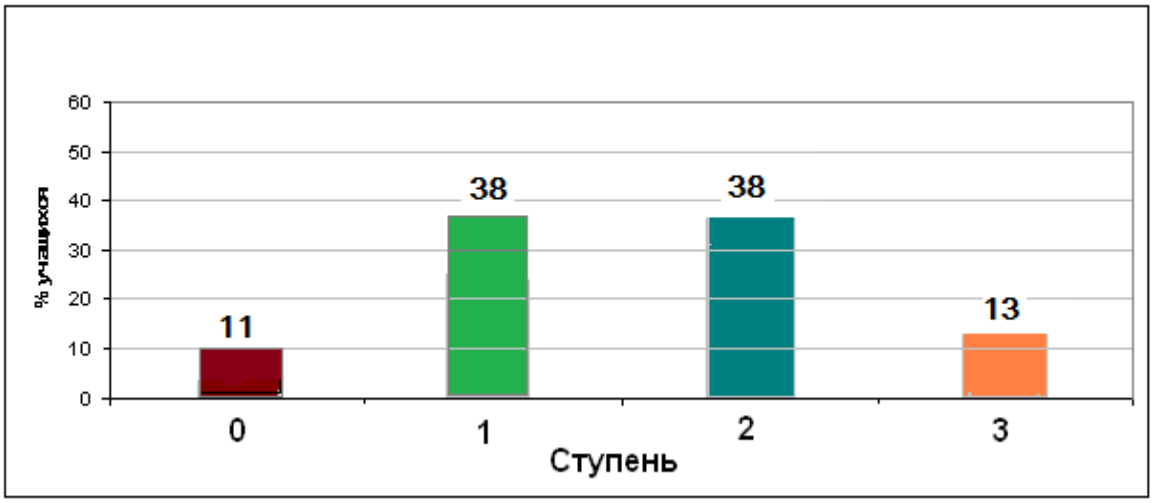

Рис. 5.6. Распределение участников тестирования по ступеням достижений (русский язык)

В табл. 5.2 приведены данные о распределении участников по ступеням достижений с учетом меры освоения ступени. В таблицу не включены 631 учащийся (11% от общего числа), не достигших 1-ой ступени.

Табл. 5.1. Распределение участников тестирования по ступеням (русский язык)

Ступень	Всего человек	Процент	Степень освоения	Всего человек	Процент от общего	Процент от ступени
1	2287	38	достиг	1278	21	56
			освоил	1013	17	44
2	2271	38	достиг	1898	33	84
			освоил	373	6	16
3	757	13	достиг	664	11	88
			освоил	93	2	12

Таким образом, мы видим, что в данном случае из 2271 учащихся, находящихся на 2-ой ступени, освоили ее (т.е. стоят уже уверенно) только 16% (значительно меньше, чем по математике). Остальные учащиеся, находящиеся на этой ступени, могут быть рассмотрены, как только-только ее достигнувшие. Аналогично, из 757 учащихся, находящихся на 3-ей ступени, освоили ее только 12%. Что касается 1-ой ступени, то на ней сейчас находятся 2287 учащихся (38%). Из них большинство учащихся (56%) ее еще не освоили.

На рис.5.7 распределение учащихся по ступеням достижений с учетом степени освоения ступени показано графически.

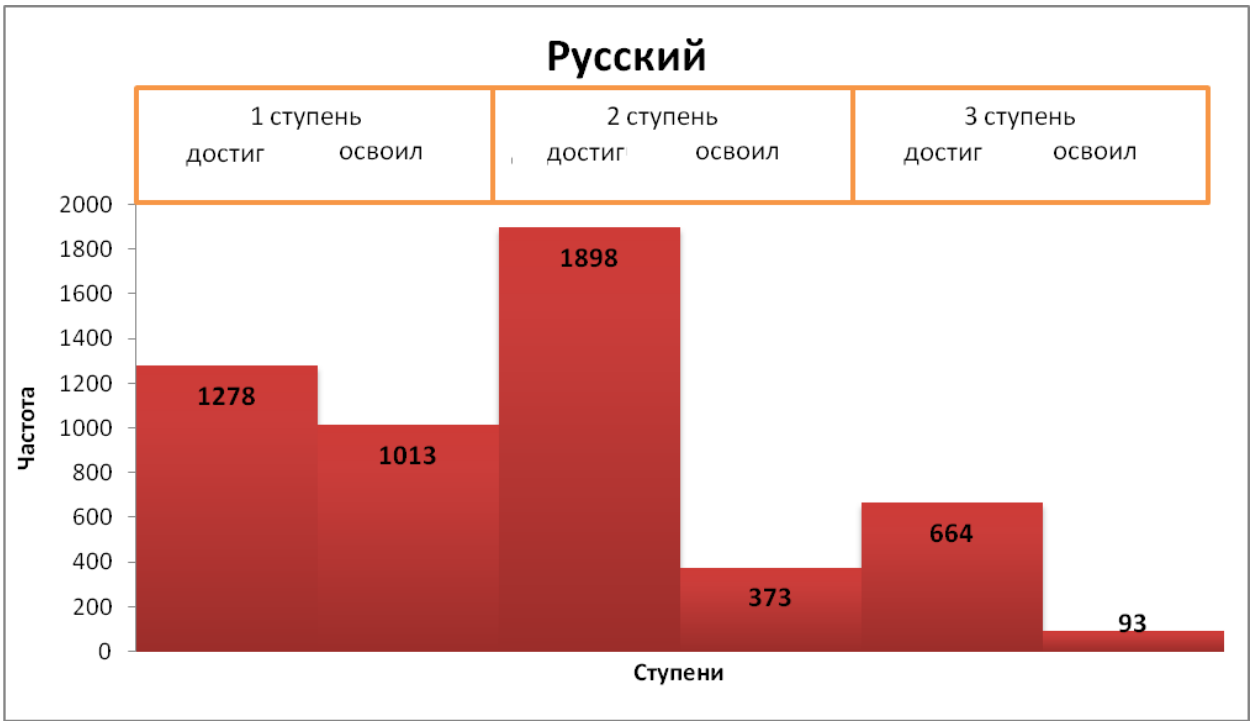

Рис. 5.7. Распределение участников тестирования по ступеням достижений с учетом меры освоения ступени (русский язык)

Литература

1. SAM (School Achievement Monitoring): Инструмент мониторинга учебных достижений школьников // *под ред. Нежнова П.Г., Кардановой Е.Ю.*, 2011, 104 с.

2. *Kardanova E., Gaponova N.* New technologies of assessment: School Achievements Monitoring Toolkit // The paper presented at EDULEARN12, Barcelona, Spain, 2012. http://library.iated.org/view/KARDANOVA2012NEW

6. Первичный анализ результатов тестирования

Результаты тестирования SAM могут быть представлены тремя ключевыми группами показателей: интегральные (первичные и тестовые) баллы, ступени образовательных результатов и трехмерные профили.

Первичный балл – это сумма баллов по всем заданиям, выполненным участником тестирования. Тестовый балл – окончательный балл учащегося, полученный в результате математической обработки первичных результатов, имеющей целью получить оценки на единой метрической шкале, общей для всех участников независимо от времени прохождения теста и конкретного набора заданий, которые они выполняли. Для представления результатов тестирования по данному показателю используются 1000-балльные шкалы, полученные для каждого предметного теста с помощью специальных исследований на базовой выборке учащихся.

Помимо интегрального тестового балла каждому участнику тестирования определяется ступень, которой он достиг. Всего выделено 4 категории достижений: ниже первой ступени (учащийся не освоил предмет даже на 1-м уровне), первая ступень (освоен 1-ый уровень), вторая (освоен 2-ой уровень) и третья (освоен 3-ий уровень). Деление участников тестирования по ступеням позволяет интерпретировать результаты в терминах уровневой модели оценивания, т.е. дать интегральному баллу содержательную интерпретацию.

Дополнительно тест SAM позволяет получить структурную характеристику оцениваемого образовательного результата: его трехмерный профиль. Профиль строится по первичным (процентным) баллам, полученным по каждому уровню, и показывает, какой процент материала усвоен на каждом из трех уровней, то есть из чего складывается тот или иной интегральный результат, разведенный по трем субшкалам. Именно профили и их изменения в ходе мониторинга, наиболее непосредственно и выразительно отображают процесс присвоения предметного содержания.

Первичный анализ результатов тестирования был проведен на базе данных апробации тестов SAM в одном из регионов РФ весной 2012 г. Особенностью данной апробации являлось то, что были протестированы практически все учащиеся 4-го (выпускного) класса начальной школы данного региона. В табл. 6.1 представлены общие сведения о данном тестировании.

Табл. 6.1. Общие сведения о тестировании

	Количество учащихся	Количество школ	Количество классов	Количество населенных пунктов
Математика	4639	192	293	136
Русский язык	4649	192	293	136

В табл. 6.2. приведены данные о распределении учащихся региона по ступеням образовательных результатов в зависимости от предмета тестирования. На рис.6.1 это же распределение показано графически. Мы видим, что в среднем результаты по математике несколько лучше результатов по русскому языку. По математике 2-ая ступень доминирует (более 50% учащихся ее достигли), по русскому языку – примерно поровну детей на 1-ой и 2-ой ступенях (около 40%). Число детей на 3-ей ступени невелико – 12% по русскому и 18% - по математике.

Табл. 6.2. Распределение учащихся по ступеням достижений

		Ниже 1 ступени	1 ступень	2 ступень	3 ступень	всего
Математика	Число	106	1305	2390	838	4639
	Процент	2%	28%	52%	18%	100%
Русский язык	Число	494	1801	1791	563	4649
	Процент	11%	39%	38%	12%	100%

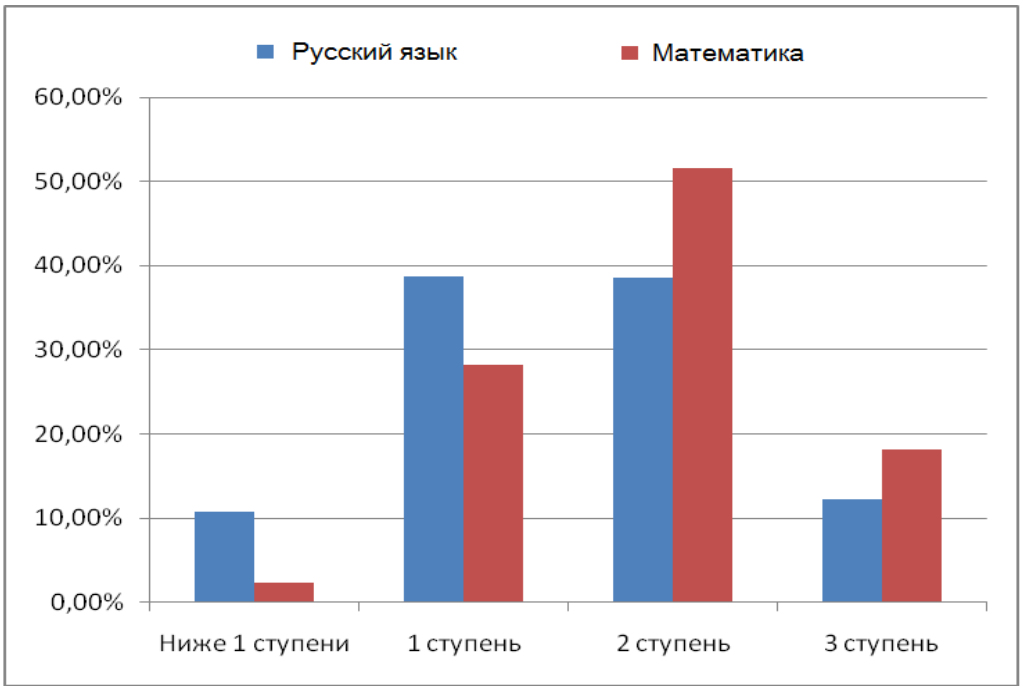

Рис. 6.1. Распределение учащихся по ступеням достижений в зависимости от предмета

В табл. 6.3 приведены средние проценты решенных задач разного уровня в зависимости от предмета. На рис.6.2 эта информация представлена графически.

Табл. 6.3. Средний процент решенных задач разных уровней

	Задачи 1-го уровня	Задачи 2-го уровня	Задачи 3-го уровня
Математика	86	61	32
Русский язык	76	49	28

Мы видим, что результаты по математике выше результатов по русскому языку по всем уровням. Однако для заданий 3-го уровня это преимущество не значительно, тогда как для заданий 1-го и 2-го уровней оно существенно. В целом, выпускники начальной школы хорошо справляются с заданиями 1-го уровня (материал формально освоен), но только половина материала освоена на 2-ом, рефлексивном, уровне (с пониманием), а на функциональном уровне освоено около трети.

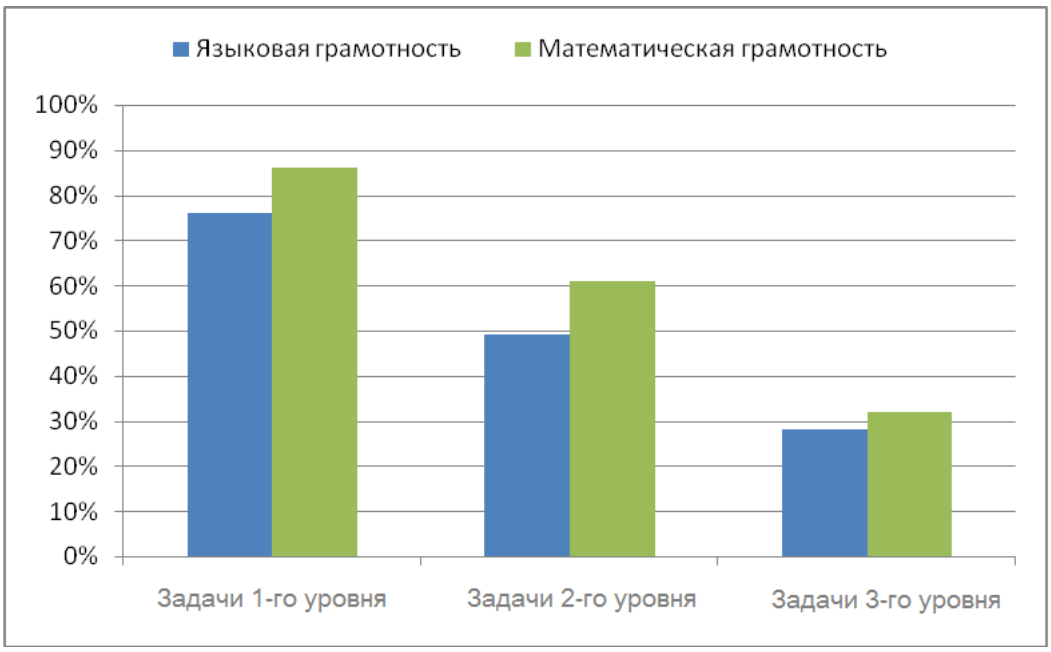

Рис. 6.2. Процент решенных задач разного уровня в зависимости от предмета

С целью сравнительной оценки учебных достижений учащихся можно использовать нормы выполнения теста. Они представляют собой набор показателей, которые устанавливаются эмпирически по результатам выполнения теста четко определенной выборкой испытуемых. Эта выборка должна быть репрезентативной по отношению к генеральной совокупности лиц, для которых предназначен тест, а также быть достаточно большой по объему. В данном случае в тестировании принимали участие все выпускники начальной школы региона, поэтому могут быть установлены нормы выполнения теста для данного региона (региональные нормы).

Нормы могут быть установлены на индивидуальном и на групповом уровнях. На индивидуальном уровне могут быть рассмотрены следующие показатели:

- Среднестатическая индивидуальная норма – средний показатель и стандартное отклонение выполнения теста учащимися выборки (табл. 6.4).
- Процентильные нормы – показатели, основанные на проценте участников тестирования, выполнивших тест на определенном уровне.

Табл. 6.4. Среднестатистические индивидуальные нормы

	Среднее значение	Стандартное отклонение
Математика	522	49
Русский язык	499	50

В качестве процентильных норм могут быть предложены, например, показатель выполнения теста, соответствующий 90-му процентилю (что означает, что 90% учащихся выполнили тест хуже) и 10-му процентилю (90% учащихся выполнили тест лучше) – табл. 6.5.

Табл. 6.5. Процентильные индивидуальные нормы

	10-ый процентиль	90-ый процентиль
Математика	459	581
Русский язык	433	557

Таким образом, если тестовый балл участника тестирования по математике выше 581, то он находится в группе 10% самых лучших; если балл участника меньше 459, то он находится в группе 10% самых худших. Аналогично интерпретируются процентильные нормы по русскому языку. Отметим, что нормы (и среднестатистические, и процентильные) по математике выше соответствующих показателей по русскому языку, что означает, что результаты по математике выше результатов по русскому в данной выборке учащихся.

Дополнительно, помимо среднестатистических данных на индивидуальном уровне можно выделить усредненные результаты по школам. На этом уровне могут быть рассмотрены следующие показатели:

- Среднестатистическая групповая норма – средний по школам показатель выполнения теста и его стандартное отклонение (таблица 6.6).
- Социокультурная норма – средний показатель группы школ-лидеров (табл. 6.7).

Табл. 6.6. Среднестатистические групповые нормы

	Среднее значение	Стандартное отклонение
Математика	517	34
Русский язык	499	36

Группа лидеров – это группа школ (например, 25% от общего числа), имеющих самые высокие показатели. Лидирующая группа школ выступает в качестве ориентира для всего образовательного сообщества, то есть задает социокультурную норму, как реалистичную среднестатистическую норму «завтрашнего дня» и дополнительное основание оценки тестовых результатов.

Табл. 6.7. Социокультурные нормы

	Тестовый балл
Математика	561
Русский язык	543

Эти нормы являются чисто статистическими, но они помогают интерпретировать результаты тестирования с точки зрения их сравнимости между собой.

6.1. Анализ результатов тестирования по математике

На рис.6.3 показана гистограмма распределения тестовых баллов участников тестирования по математике. Среднее значение баллов равно 522, стандартное отклонение 41. В целом распределение тестовых баллов близко к нормальному.

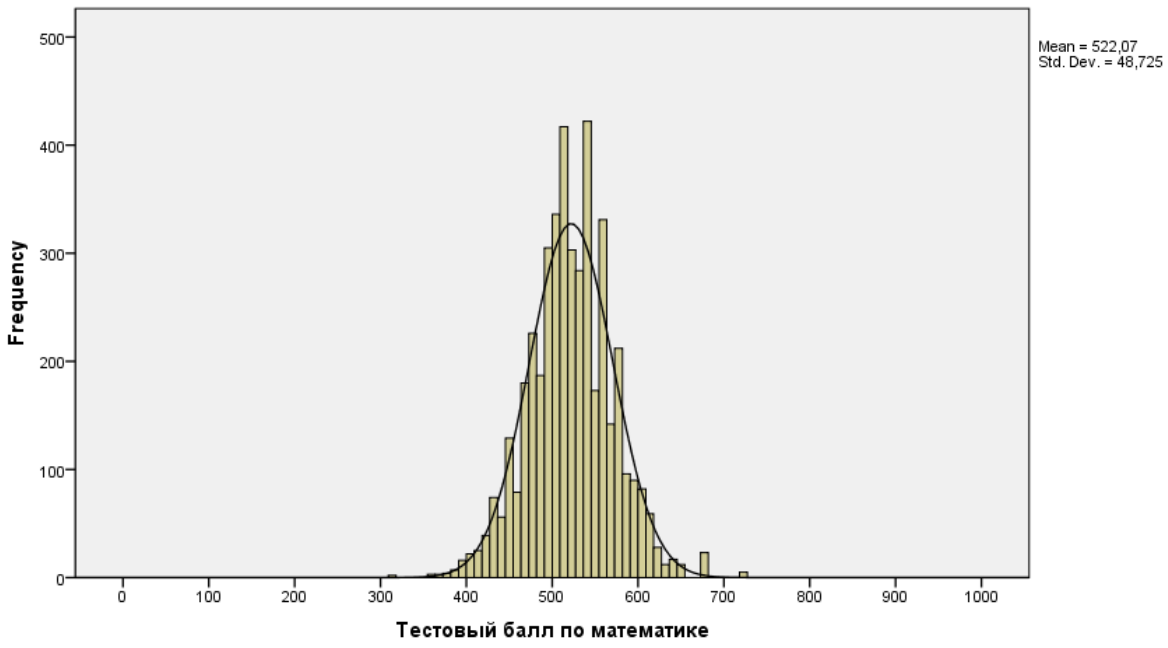

Рис.6.3. Гистограмма распределения тестовых баллов участников тестирования (математика)

На рис. 6.4 представлено распределение участников тестирования по математике по ступеням достижений.

Рис. 6.4. Распределение участников тестирования по ступеням достижений (математика)

На рис. 6.5 показан профиль образовательных результатов по математике для данной выборки учащихся (средние проценты решенных задач в зависимости от уровня).

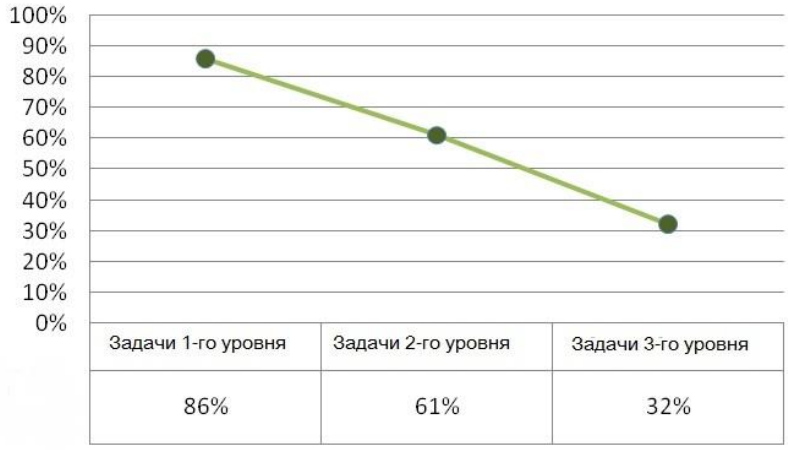

Рис. 6.5. Профиль по математике для данной выборки учащихся

Профиль прочитывается относительно просто. Если принять, что тест в нужной пропорции покрывает основные разделы учебного содержания, то шкала первого уровня сообщает, что основная часть этого содержания формально освоена (более 80% задач 1-го уровня выполнены); вторая шкала показывает, что рефлексивно (с пониманием), усвоено около 60%; и последняя шкала информирует, что функционально усвоено около трети.

На рис.6.6 показано распределение участников тестирования по ступеням достижений в различных школах данного региона (на рисунке представлены 20 школ из 192, отобранных случайным образом). По горизонтальной оси отложен процент учащихся, находящихся на каждой ступени, а по вертикальной оси указаны школы. Для каждой школы в скобках указан средний тестовый балл учащихся данной школы. Школы упорядочены по возрастанию среднего значения общих тестовых баллов.

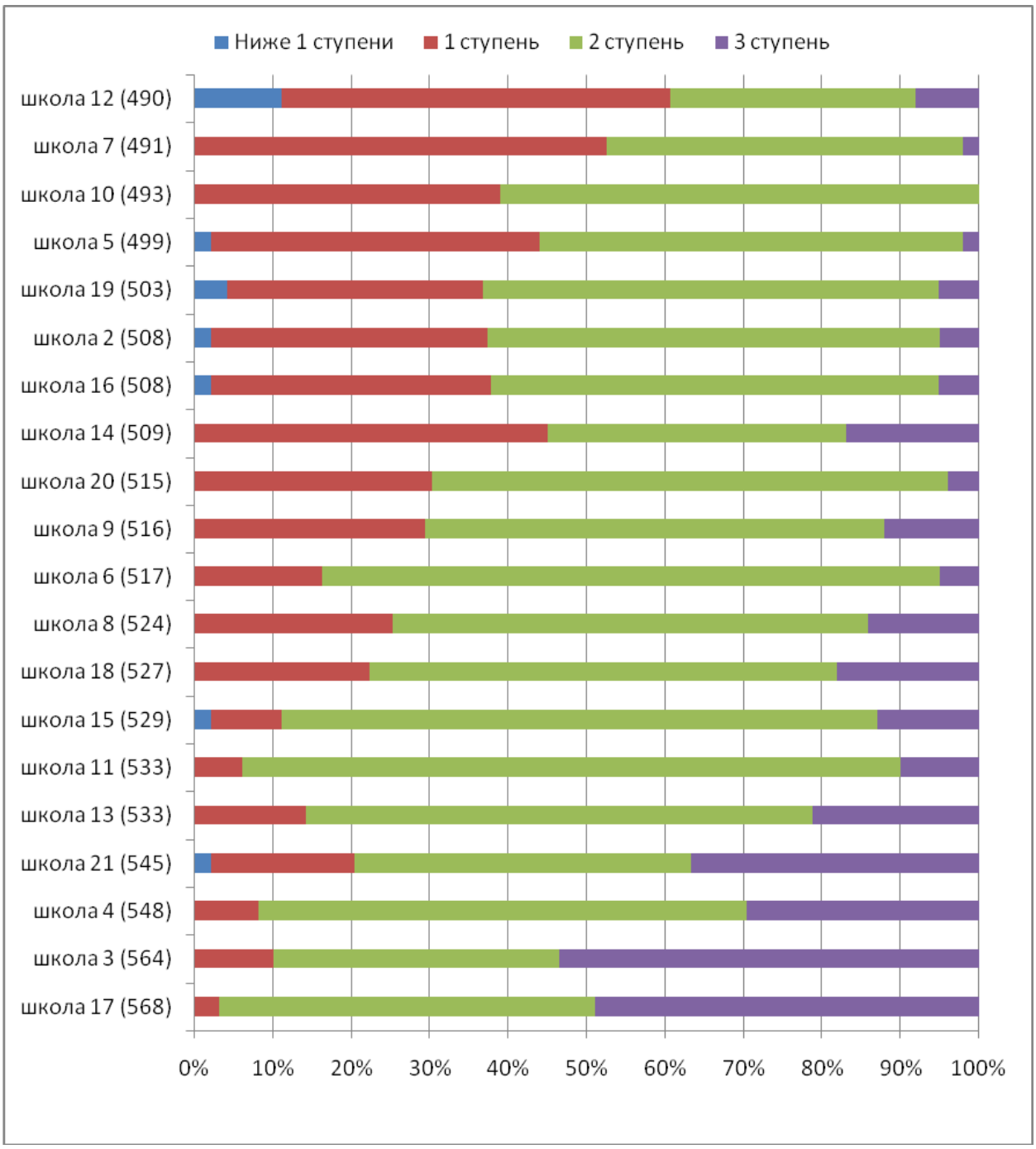

Рис. 06.6. Распределение учащихся разных школ по ступеням достижений (математика)

Таким образом, мы видим, что количество детей на разных ступенях сильно варьируется в зависимости от школы. Так, в верхних школах доминирует первая ступень (красный цвет), тогда как в нижних – третья (фиолетовый цвет). Однако в большинстве школ доминирует вторая ступень (зеленый цвет), что означает, что содержание усвоено на рефлексивном уровне (уровне понимания). Число участников, не достигших 1-ой ступени, не велико даже в школах с низкими результатами.

Интересно сравнить распределение учащихся по ступеням достижений в различных классах одной школы. На рис.6.7 показано такое распределение для одной из школ данного региона, средние показатели которой ниже средних по региону.

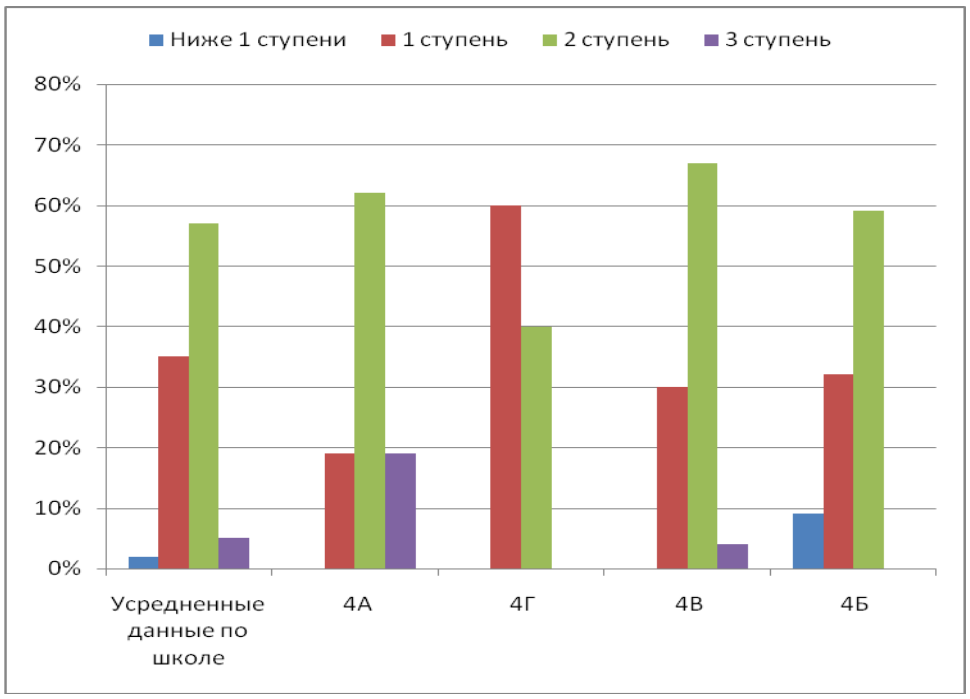

Рис. 6.7. Распределение по ступеням достижений учащихся различных классов одной школы (математика)

Сравнивая распределения по ступеням достижений в различных классах данной школы, мы видим, что в классах 4А, 4Б и 4В доминирует 2-ая ступень, а в классе 4Г - 1-я ступень (что соответствует первому, формальному уровню освоения предметного материала). Самым сильным является класс 4А – почти 20% детей достигли 3-ей ступени и менее 20% находятся еще на 1-ой ступени. За ним идет класс 4В – 30% детей находятся на 1-ой ступени и более 65% - на 2-ой. Функциональный уровень овладения содержанием в классе 4В практически отсутствует - на 3-ей ступени находятся не более 5% учащихся. В классе 4Б почти 10% детей не достигли 1-ой ступени – не освоен даже 1-ый уровень, а 30% детей освоили материал на формальном уровне (без понимания). Знание этих соотношений весьма существенно для выбора стратегии работы с тем или иным классом. Еще важнее знание причин такой ситуации, для чего требуются дополнительные исследования.

В рамках данного отчета представлены первичные результаты анализа связи результатов теста по математике с различными факторами – полом участника, вариантом, который он выполнял, типом учебного заведения, местоположением школы и др. Более глабокое изучение факторов, оказывающих влияние на результаты обучения в начальной школе, выходит за рамки данной работы.

Ниже будут перечислены основные результаты исследования.

1) Результат SAM по математике статистически значимо зависит от пола участника: девочки выполняют тест лучше, чем мальчики.

В табл. 6.8 показано соотношение мальчиков и девочек в данной выборке, а в табл. 6.9 – распределение участников разного пола по ступеням достижений. На рис.6.8 это же распределение показано графически.

Табл. 6.8. Соотношение мальчиков и девочек (математика)

Пол	**Кол-во**	**%**
Женский	2078	47,2
Мужской	2328	52,8

Табл. 6.9. Распределение мальчиков и девочек по ступеням достижений (математика)

			Ступень по математике				Всего
			Ниже первой	1	2	3	
пол	Женский	Кол-во	45	523	1100	408	2076
		%	2,2%	25,2%	53,0%	19,7%	100,0%
	Мужской	Кол-во	58	664	1233	370	2325
		%	2,5%	28,6%	53,0%	15,9%	100,0%

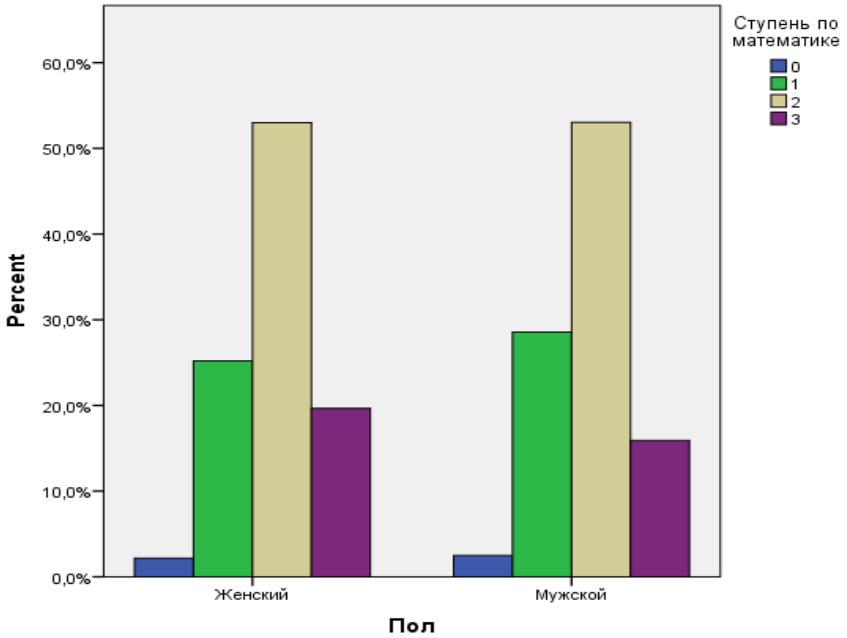

Рис. 6.8. Распределение участников тестирования по ступеням достижений в зависимости от пола (математика)

Статистически было показано, что связь между результатом тестирования по математике и полом участника есть (хоть и очень слабая): девочки выполняют тест несколько лучше, чем мальчики.

2) Результат SAM по математике статистически значимо зависит от типа учебного заведения: учащиеся гимназий выполняют тест лучше, чем учащиеся общеобразовательных школ. В табл. 6.10 показано соотношение учащихся учебных заведений различного типа в данной выборке, а в табл. 6.11 – распределение участников разного пола по ступеням достижений. На рис.6.9 это же распределение показано графически.

Табл. 6.10. Соотношение учащихся учебных заведений различного типа (математика)

Тип школы	**Кол-во**	**Процент**
Общеобразовательная	3754	85,2
Гимназия	647	14,7

Табл. 6.11. Распределение участников тестирования по ступеням достижений в зависимости от типа учебного заведения (математика)

Тип учебного заведения * ступень по математике Crosstabulation

		Ступень по математике				Всего
		Ниже первой	Первая	Вторая	Третья	
Общеобразовательная	Кол-во	98	1082	1986	588	3754
	%	2,6%	28,8%	52,9%	15,7%	100,0%
Гимназия	Кол-во	5	105	347	190	647
	%	,8%	16,2%	53,6%	29,4%	100,0%

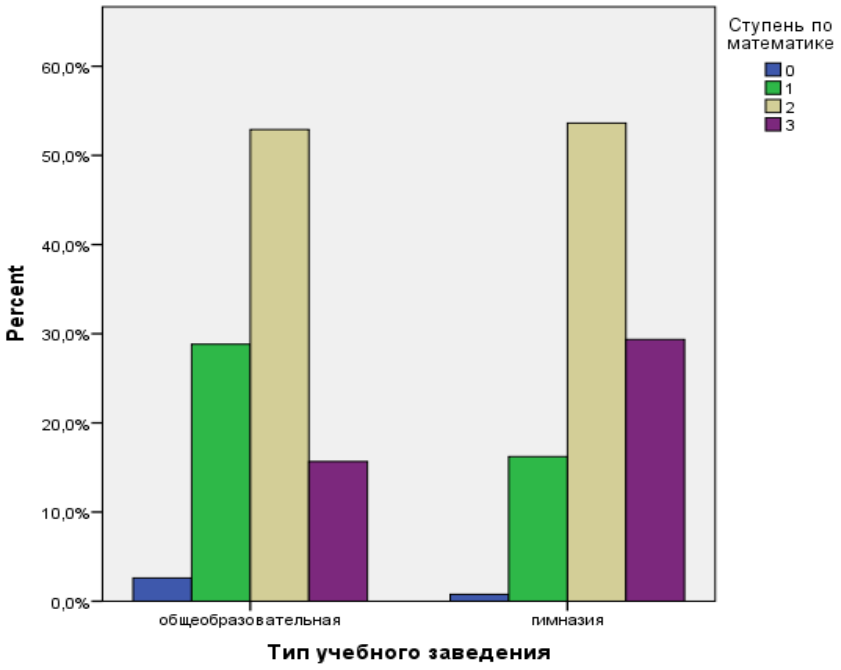

Рис. 6.9. Распределение участников тестирования по ступеням достижений в зависимости от типа учебного заведения (математика)

Статистически было показано, что связь между результатом тестирования по математике (тестовым баллом и ступенью достижений) и типом учебного заведения есть (но слабая): учащиеся из специализированных школ (гимназий) выполняют тест несколько лучше, чем учащиеся из общеобразовательных школ.

3) Результат SAM по математике статистически значимо зависит от типа населенного пункта: учащиеся городов и поселков выполняют тест лучше, чем учащиеся сельской местности

В табл. 6.12 показано частотное распределение учащихся из населенных пунктов разного типа, а в табл. 6.13 – распределение этих участников по ступеням достижений. На рис.6.10 показана зависимость результатов тестирования (тестового балла) от типа населенного пункта.

Табл. 6.12. Частотное распределение учащихся из населенных пунктов разного типа (математика):

		Частота	Процент
Valid	Город	3161	71,7
	Поселок и поселок городского типа	618	14,0
	Село и деревня	606	13,8
	Total	4385	99,5
Missing	System	21	,5
Total		4406	100,0

Таблица 6.13. Распределение *учащихся из населенных пунктов разного типа по ступеням достижений(математика)*:

Тип населенного пункта		Ступень по математике				Total
		Ниже первой	Первая	Вторая	Третья	
Город	Кол-во	58	754	1725	624	3161
	%	1,8%	23,9%	54,6%	19,7%	100,0%
Поселок и поселок городского типа	Кол-во	16	188	312	102	618
	%	2,6%	30,4%	50,5%	16,5%	100,0%
Село и деревня	Кол-во	28	237	290	51	606
	%	4,6%	39,1%	47,9%	8,4%	100,0%

Статистически было показано, что связь между результатом тестирования по математике (тестовым баллом и ступенью достижений) и типом населенного пункта есть: учащиеся из городов выполняют тест лучше учащихся из поселков, а учащиеся из поселков – лучше, чем учащиеся из сельской местности.

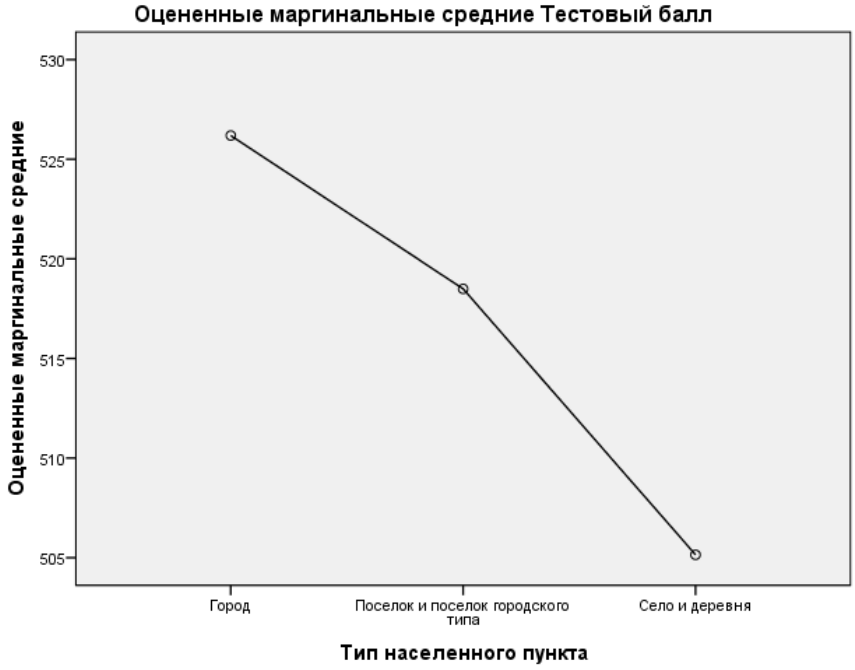

Рис. 6.10. Зависимость результатов тестирования от типа населенного пункта (математика)

Для более точного выяснения факторов, оказывающих влияние на результаты обучения по математике в начальной школе, требуется дополнительное исследование.

6.2. Анализ результатов тестирования по русскому языку

На рис.6.11 показана гистограмма распределения тестовых баллов участников тестирования по русскому языку.

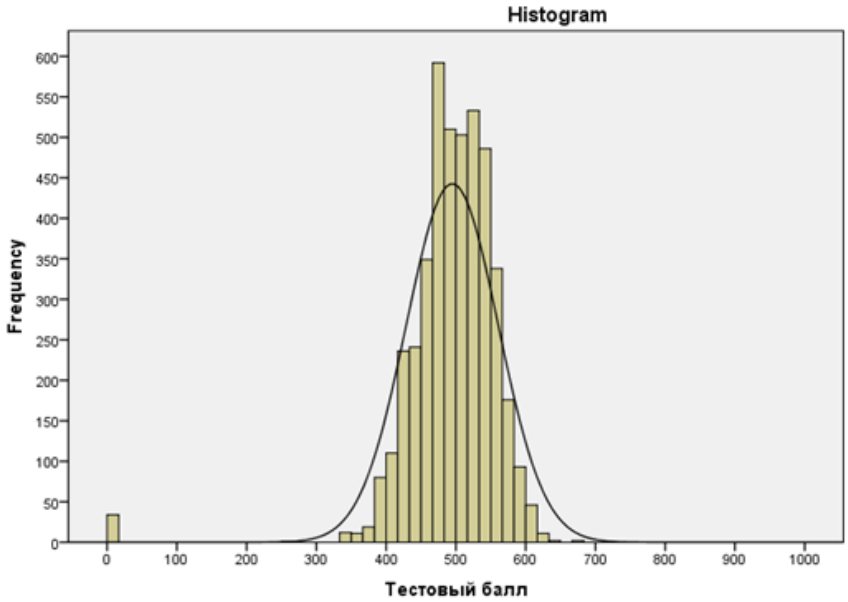

Рис.6.11. Гистограмма распределения тестовых баллов участников тестирования (русский язык)

На рис. 6.12 представлено распределение участников тестирования по русскому языку по ступеням достижений.

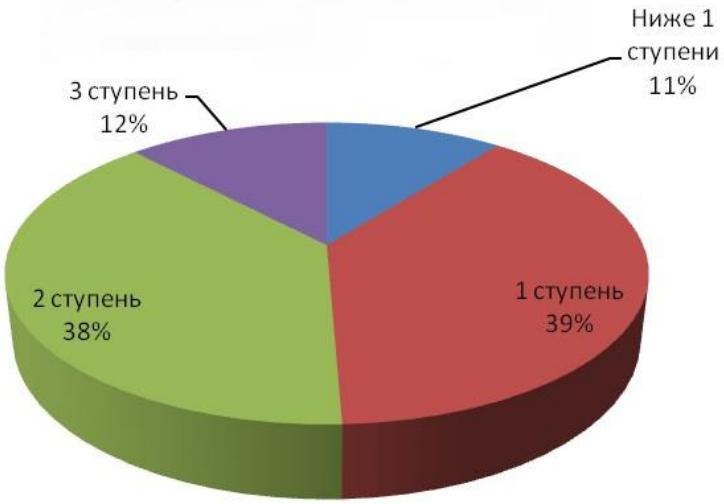

Рис. 6.12. Распределение участников тестирования по ступеням достижений (русский язык)

На рис. 6.13 показан профиль успешности по русскому языку для данной выборки учащихся (средние проценты решенных задач в зависимости от уровня).

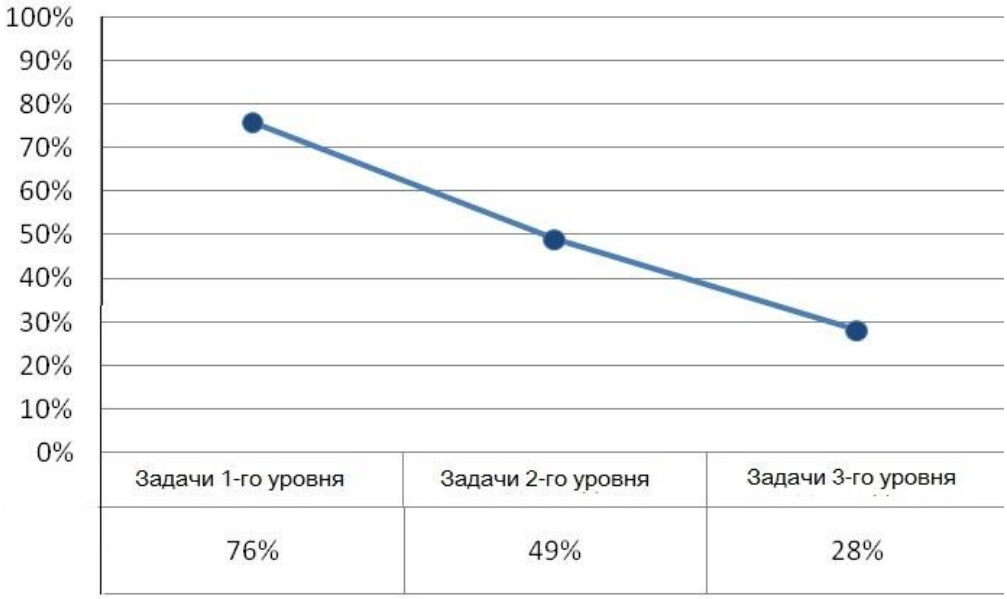

Рис. 6.13. Профиль по русскому языку для данной выборки учащихся

Если принять, что тест в нужной пропорции покрывает основные разделы учебного содержания, то шкала первого уровня сообщает, что основная часть этого содержания формально освоена (почти 80% задач 1-го уровня выполнены); вторая шкала показывает, что рефлексивно (с пониманием), усвоено только 50%; и последняя шкала информирует, что функционально усвоено менее трети.

На рис.6.14 показано распределение участников тестирования по ступеням достижений в различных школах данного региона (на рисунке представлены 20 школ из 192 - те же, что и по математике на рис.6.6). По горизонтальной оси отложен процент учащихся, находящихся на каждой ступени, а по вертикальной оси указаны школы. Для

каждой школы в скобках указан средний тестовый балл учащихся данной школы. Школы упорядочены по возрастанию среднего значения общих тестовых баллов.

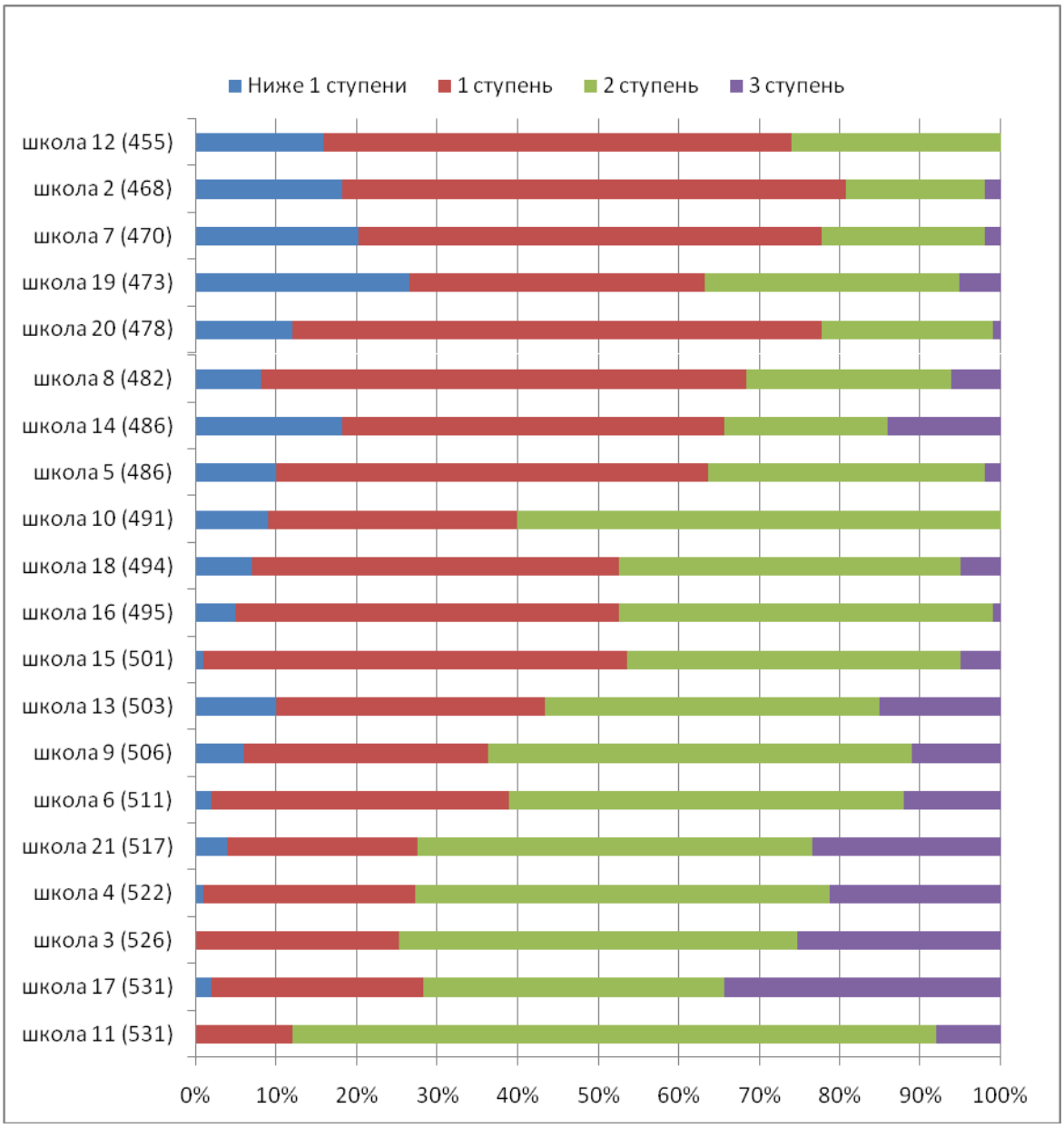

Рис. 06.14. Распределение учащихся разных школ по ступеням достижений (русский язык)

Таким образом, мы видим, что количество детей на разных ступенях сильно варьируется в зависимости от школы. Так, в верхних школах доминирует первая ступень (красный цвет) и относительно большое число детей находится ниже 1-ой ступени (голубой цвет), 3-я ступень в этих школах почти не представлена. В нижних школах доминирует 2-я ступень (зеленый цвет) и уже активно представлена 3-я ступень (фиолетовый цвет). В целом результаты по русскому языку хуже результатов этих же школ по математике.

Интересно сравнить распределение учащихся по ступеням достижений в различных классах одной школы. На рис.6.15 показано такое распределение для одной из школ данного региона, средние показатели которой ниже средних по региону. (Мы взяли ту же школу, что и на рис.6.7).

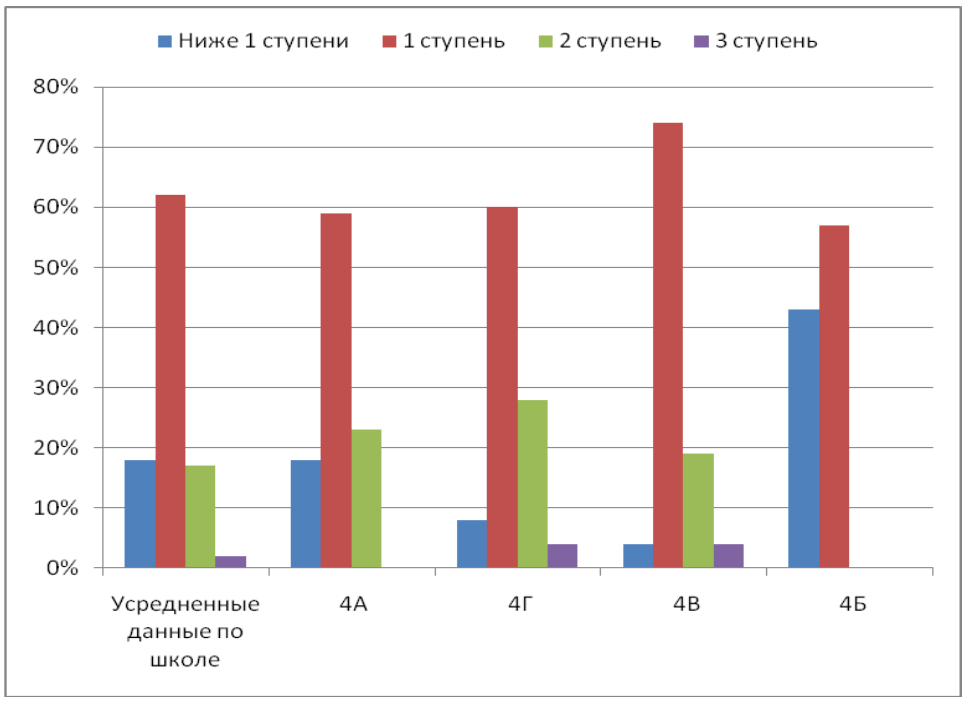

Рис. 6.15. Распределение по ступеням достижений учащихся различных классов одной школы (русский язык)

Сравнивая распределения по ступеням достижений в различных классах данной школы, мы видим, что во всех классах доминирует 1-я ступень. Класс 4Б демонстрирует самые плохие результаты: более 40% детей не достигли 1-ой ступени – не освоен даже 1-ый уровень, а остальные учащиеся данного класса освоили материал на формальном уровне (без понимания). В этом классе нет детей на более высоких ступенях. Самым сильным сейчас является класс 4Г: около 30% детей освоили материал на 2-м уровне (уровне понимания). Однако даже в этом классе почти 10% детей не достигли 1-ой ступени.

Для сравнения рассмотрим еще одну школу (рис. 6.16). В этой школе классы еще более различаются по качеству подготовки. Самым сильным является класс 4А – более 60% детей достигли 2-ей ступени и почти 10% - 3-ей. В противоположность ему в классе 4В 55% детей не достигли 1-ой ступени, а остальные 45% находятся на 1-ой ступени. Рефлексивный и функциональный уровни овладения содержанием в классе 4В полностью отсутствуют. Класс 4Б является очень разнородным по своей подготовке – в нем 20% детей не достигли 1-ой ступени, а 10% - достигли 3-ей. Остальные поделились почти поровну между 1-ой и 2-ой ступенями.

Знание этих соотношений весьма существенно для выбора стратегии работы с тем или иным классом. Еще важнее знание причин такой ситуации, для чего требуются дополнительные исследования.

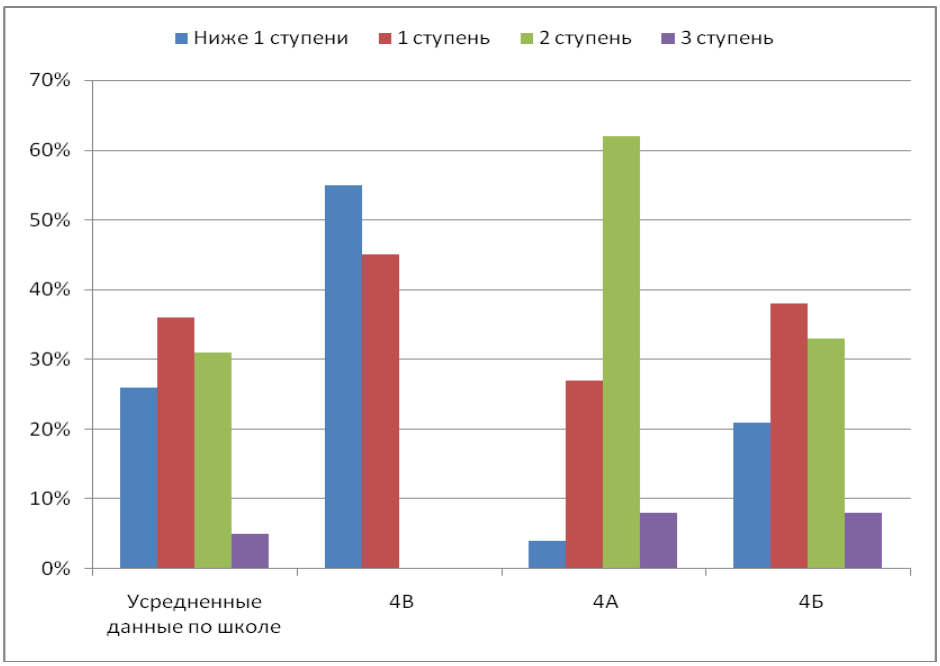

Рис. 6.16. Распределение по ступеням достижений учащихся различных классов школы 2 (русский язык)

В рамках данного отчета представлены первичные результаты анализа связи результатов теста по русскому языку с различными факторами – полом участника, вариантом, который он выполнял, типом учебного заведения, местоположением школы и др. Более глабокое изучение факторов, оказывающих влияние на результаты обучения в начальной школе, выходит за рамки данной работы.

Ниже будут перечислены основные результаты исследования.

1) Результат SAM по русскому языку статистически значимо зависит от пола участника: девочки выполняют тест лучше, чем мальчики.

В табл. 6.14 показано соотношение мальчиков и девочек в данной выборке, а в табл. 6.15 – распределение участников разного пола по ступеням достижений. На рис.6.17 это же распределение показано графически.

Табл. 6.14. Соотношение мальчиков и девочек (русский язык)

Пол	**Кол-во**	**%**
Женский	2075	47,1
Мужской	2326	52,9

Табл. 6.15. Распределение мальчиков и девочек по ступеням достижений (русский язык)

			Ступень по математике				Всего
			Ниже первой	1	2	3	
пол	Женский	Кол-во	161	736	849	329	2075
		%	7,8%	35,5%	40,9%	15,9%	100,0%
	Мужской	Кол-во	314	988	811	213	2326
		%	13,5%	42,5%	34,9%	9,2%	100,0%

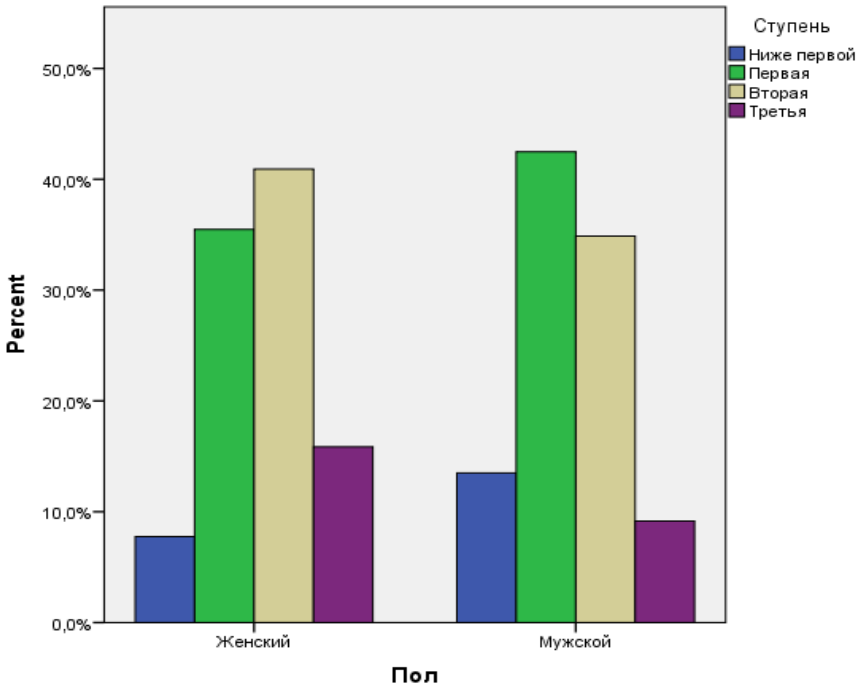

Рис. 6.17. Распределение участников тестирования по ступеням достижений в зависимости от пола (русский язык)

Статистически было показано, что связь между результатом тестирования по русскому языку и полом участника есть (хоть и слабая): девочки выполняют тест несколько лучше, чем мальчики.

2) Результат SAM по русскому языку статистически значимо зависит от типа учебного заведения: учащиеся гимназий выполняют тест лучше, чем учащиеся общеобразовательных школ.

В табл. 6.16 показано соотношение учащихся учебных заведений различного типа в данной выборке, а в табл. 6.17 – распределение участников разного пола по ступеням достижений. На рис.6.18 это же распределение показано графически.

Табл. 6.16. Соотношение учащихся учебных заведений различного типа (русский язык)

Тип школы	**Кол-во**	**Процент**
Общеобразовательная	3754	85,3
Гимназия	647	14,7

Табл. 6.17. Распределение участников тестирования по ступеням достижений в зависимости от типа учебного заведения (русский язык)

		Ступень по русскому языку				Всего
		Ниже первой	Первая	Вторая	Третья	
Общеобразовательная	Кол-во	444	1489	1406	415	3754
	%	11,8%	39,7%	37,5%	11,1%	100,0%
Гимназия	Кол-во	31	235	254	127	647
	%	4,8%	36,3%	39,3%	19,6%	100,0%

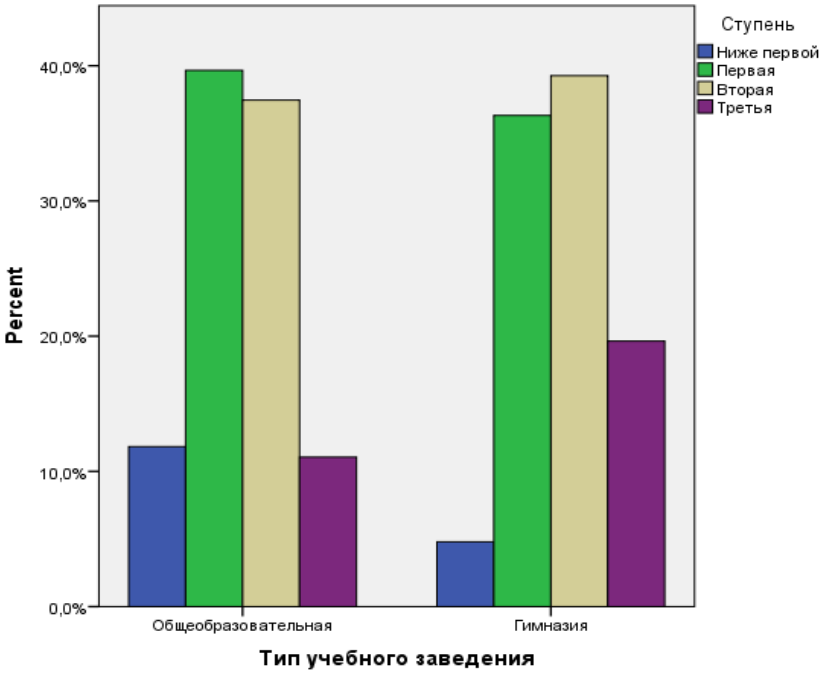

Рис. 6.18. Распределение участников тестирования по ступеням достижений в зависимости от типа учебного заведения (русский язык)

Статистически было показано, что связь между результатом тестирования по русскому языку (тестовым баллом и ступенью достижений) и типом учебного заведения есть (но очень слабая): учащиеся из специализированных школ (гимназий) выполняют тест несколько лучше, чем учащиеся из общеобразовательных школ.

3) Результат SAM по русскому языку не зависит от типа населенного пункта.

В табл. 6.18 показано частотное распределение учащихся из населенных пунктов разного типа, а в табл. 6.19 – распределение этих участников по ступеням достижений. На рис.6.19 показана зависимость результатов тестирования (тестового балла) от типа населенного пункта..

Табл. 6.18. Частотное распределение учащихся из населенных пунктов разного типа (русский язык):

		Частота	Процент
Valid	Город	3161	71,8
	Поселок и поселок городского типа	618	14,0
	Село и деревня	606	13,8
	Total	4385	99,6
Missing	System	16	,4
Total		4401	100,0

Табл. 6.19. Распределение *учащихся из населенных пунктов разного типа по ступеням достижений(русский язык)*:

Тип населенного пункта		Ступень по русскому языку				Total
		Ниже первой	Первая	Вторая	Третья	
Город	Кол-во	281	1237	1228	415	3161
	%	8,9%	39,1%	38,8%	13,1%	100,0%
Поселок и поселок городского типа	Кол-во	91	254	208	65	618
	%	14,7%	41,1%	33,7%	10,5%	100,0%
Село и деревня	Кол-во	98	229	218	61	606
	%	16,2%	37,8%	36,0%	10,1%	100,0%

Статистически было показано, что связь между результатом тестирования по русскому языку (тестовым баллом и ступенью достижений) и типом населенного пункта отсутствует: результат ученика не связан значимо с типом населенного пункта. На рис.6.19 видно, что различия в средних баллах участников из различных населенных пунктов менее 10 баллов.

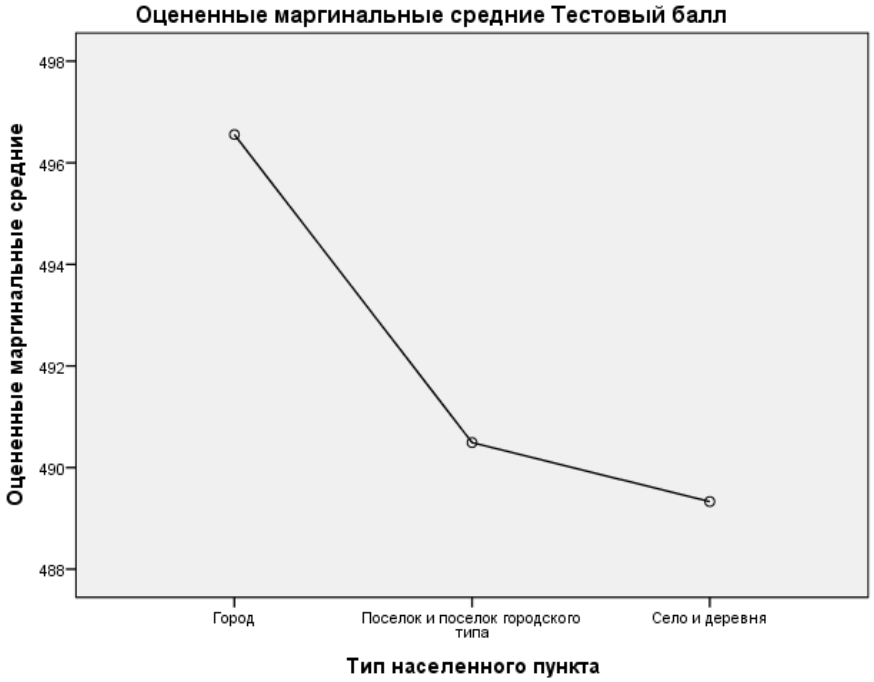

Рис. 6.19. Зависимость результатов тестирования от типа населенного пункта (русский язык)

4) В заключение на примере теста по русскому языку рассмотрим связь между результатом тестирования и вариантом теста. В данном регионе были задействованы два варианта.

В табл. 6.20 показано соотношение учащихся, выполнявших различные враианты в данной выборке, а в табл. 6.21 – распределение участников различных вариантов по ступеням достижений. На рис.6.20 это же распределение показано графически.

Табл. 6.20. Соотношение учащихся по варинтам (русский язык)

Вариант	**Кол-во**	**%**
Первый	2219	50,4
Второй	2182	49,6

Табл. 6.21. Распределение участников тестирования по ступеням достижений в зависимости от варианта (русский язык)

.

			Ступень по русскому языку				Всего
			Ниже первой	Первая	Вторая	Третья	
Вариант по Математике	1	Кол-во	231	871	832	285	2219
		%	10,4%	39,3%	37,5%	12,8%	100,0%
	2	Кол-во	244	853	828	257	2182
		%	11,2%	39,1%	37,9%	11,8%	100,0%

.

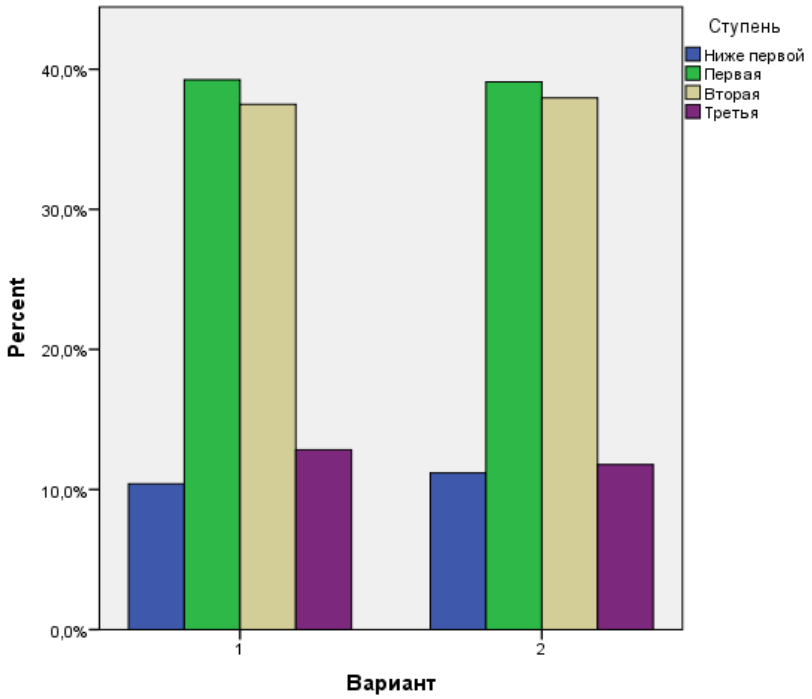

Рис. 6.20. Распределение участников тестирования по ступеням достижений в зависимости от варианта теста (русский язык)

Статистически было показано, что связь между результатом тестирования по русскому языку (тестовым баллом и ступенью достижений) и вариантом теста отсутствует: результат ученика не связан значимо с вариантом.

Для более точного выяснения факторов, оказывающих влияние на результаты обучения по математике в начальной школе, требуется дополнительное исследование.

Заключение

Представленные в данном техническом отчете результаты исследований основных характеристик инструментария SAM свидетельствуют о соответствии этих характеристик нормативным требованиям. Таким образом, SAM можно квалифицировать как надежный инструмент качественно-количественной оценки предметных образовательных результатов учащихся начальной школы в области математики и русского языка.

Полное описание инструментария, включающее теоретические основания, конструктивные особенности тестов, используемые математические модели, способы применения, основные результаты и формы их презентации, представлены в паспорте «SAM (Student Achievement's Monitoring) – инструмент мониторинга результатов обучения школьников». Опыт применения SAM обобщен в руководстве «Тесты SAM: практическое применение»