Галина Ивановна Смирнова, *Славянский-на-Кубани государственный педагогический институт*

ПРИМЕНЕНИЕ ПРОГРАММЫ RUMM-2020 ДЛЯ РАЗРАБОТКИ ПЕДАГОГИЧЕСКОГО ТЕСТА

В статье приведён анализ двух матриц результатов тестирования. Первая имеет размер 13x10 и является известным примером учебной матрицы В.С. Аванесова. Вторая матрица размером 52x50, представлена им же, по результатам пробной проверки заданий в тестовой форме по русскому языку, разработанных для Централизованного тестирования в РФ, в 1998 году. Матрица использована для отражения алгоритма разработки педагогического теста. Обе матрицы анализировались на основе модели Раша посредством программы RUMM-2020 (Rasch Unidimentional Measurement Models).

Ключевые слова: модель Раша, экстремальные задания, совместимость тестовых заданий, дистракторный анализ заданий, равномерность возрастания меры трудности задания, соответствие меры трудности разрабатываемого теста уровню подготовленности испытуемых, достаточность вариации и размаха заданий по уровню трудности.

Анализ качества заданий и теста

Математико-статистическое обоснование меры пригодности каждого отдельного задания и теста в целом проходило по следующим критериям:

1. Экстремальные задания. Экстремальными называются задания, на которые все испытуемые отвечали либо правильно, либо неправильно. Такие задания в тест не включаются, они считаются не тестовыми, потому что невозможно точно определить меру их трудности. Считается, что уровень трудности таких заданий выходит за пределы проводимых измерений на данной выборке испытуемых, а потому результаты, полученные по таким заданиям, необходимо исключить из матрицы данных тестирования. В исследуемых матрицах экстремальных заданий не оказалось.

2. Совместимость заданий. Совместимость заданий — необходимое условие использования модели Раша для измерения латентных качеств — в частности, уровня

подготовленности. Совместимость заданий проверяется на разных уровнях:

- для всей матрицы результатов тестирования;
- для каждого задания (по результатам тестирования для задания);
- для каждого испытуемого (по результатам тестирования для испытуемого);
- для каждого элемента матрицы тестирования.

Во всех четырёх случаях совместимость определяется на основе критерия хи-квадрат — результаты тестирования сопоставляются с ожидаемыми значениями на основе модели Раша. Критерий хи-квадрат Пирсона — наиболее простой критерий проверки значимости связи между заданиями. Критерий Пирсона основывается на том, что в таблице ожидаемые частоты при гипотезе «между переменными нет зависимости» можно вычислить непосредственно.

В первой матрице тестовых заданий размером 13x10 *совместимость всех заданий матрицы результатов тестирования* на основе критерия хи-квадрат составила 0,789050. Полученные значения говорят о том, что степень соответствия данных тестирования модели Раша достаточно высока. Чем ближе данное значение к единице, тем выше степень соответствия.

Для анализа второй матрицы тестовых результатов было проведено шесть этапов, на каждом из которых проверялось соответствие модели Раша как всей матрицы, так и каждого задания. Поэтапное значение совместимости всех заданий матрицы результатов тестирования отражено в табл. 1.

Таблица 1

Значение Хи-квадрат для всех заданий матрицы на каждом из этапов исследования

Этап	Значение Хи-квадрат для всей матрицы
1	0,000
2	0,081064
3	0,079946
4	0,068934
5	0,14492
6	0,222138

Как видно из табл. 1, *совместимость всех заданий матрицы результатов тестирования* постепенно возрастала с нулевого значения на первом этапе до значения 0,222138 на последнем этапе. Данный положительный результат достигнут за счёт удаления из матрицы тестирования некачественных заданий, значение которых не достигало критического, равного 0,05.

Анализ мер *совместимости для каждого задания* показал, что все задания в первом наборе соответствуют модели Раша, так как экспериментальное значение статистики хи-квадрат для них соответствует табличному значению, при критическом уровне 0,05. Анализ *совместимости для каждого задания* второй матрицы, на каждом из этапов исследования, отражён в табл. 2.

Таблица 2

Значения хи-квадрат для заданий, не соответствующих модели Раша

№ задания	Значение Хи-квадрат
ПЕРВЫЙ ЭТАП	
1	0,032165
10	0,036121
11	0,08717

12	0,013648
13	0,008812
14	0,01339
32	0,000002
33	0,003148
38	0,007126
40	0,005486
46	0,012621
ВТОРОЙ ЭТАП	
27	0,047780
29	0,011943
ТРЕТИЙ ЭТАП	
8	0,044027
ЧЕТВЁРТЫЙ ЭТАП	
33	0,038807
34	0,039249
ПЯТЫЙ ЭПАП	
25	0,033286
ШЕСТОЙ ЭТАП	
-	-

Анализ *совместимости для каждого задания* на последнем этапе показал, что все задания соответствуют модели Раша.

Графики заданий, исключённых из матрицы тестирования и описанных в табл.2, можно объединить в четыре группы (типа). Первую, самую многочисленную, группу составляют задания № 1, 8, 10, 33 (1-й этап), 33 (4-й этап), 40 и 46. Приведём пример графика, типичного для данной группы (рис.1).

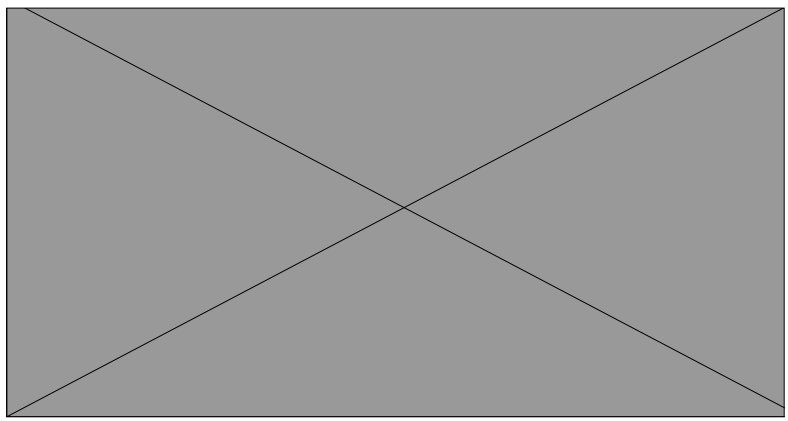

Рис. 1. **График задания № 1**

Из рис. 1 видно, что данное тестовое задание можно отнести к заданиям с *«неупорядоченными ответами»*. Испытуемые с низким и высоким уровнями подготовленности имеют большую вероятность правильно ответить на это задание, чем испытуемые, имеющие средний уровень подготовленности. Такое задание «не вписывается» к требованиям к тестовым заданиям, что подтверждено данными табл. 2. Оно нарушает основное положение (assumption) Item Response Theory — чем выше уровень подготовленности испытуемых, тем выше должна быть вероятность правильных ответов. Здесь наблюдается явное нарушение этого положения.

По оси абсцисс отложены значения латентной переменной «уровень знаний по изучаемой дисциплине». В данном случае латентная переменная варьирует от -3 логит до +4 логит. По оси ординат откладывается ожидаемый ответ индивида. Ожидаемый результат (Expected Score) варьирует от 0 до 1. В верхней части рисунка расположена следующая информация:

— код тестового задания (I0001);

— название тестового задания, здесь названия задания выбраны по умолчанию. В данном случае это (Descriptor for Item 1);

— трудность тестового задания (Location = - 0,166);

— суммарное отклонение ответов индивидов на данное задание от ожидаемых на основе модели Раша (Residual = = 1,880);

— степень соответствия данных модели Раша (Chi Sq Prob = 0,032);

— наклон графика (Slope = 0,25).

Вторую группу составляют задания № 25, 27, 29, 34 и 38. Приведём примеры графиков, типичных для данной группы (рис. 2 и 3).

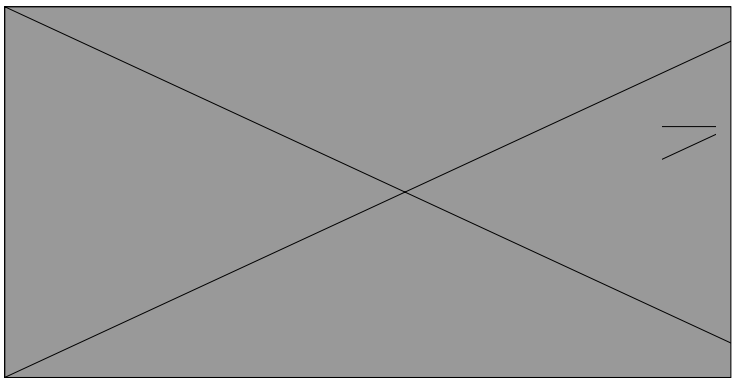

Рис. 2.
График задания № 25

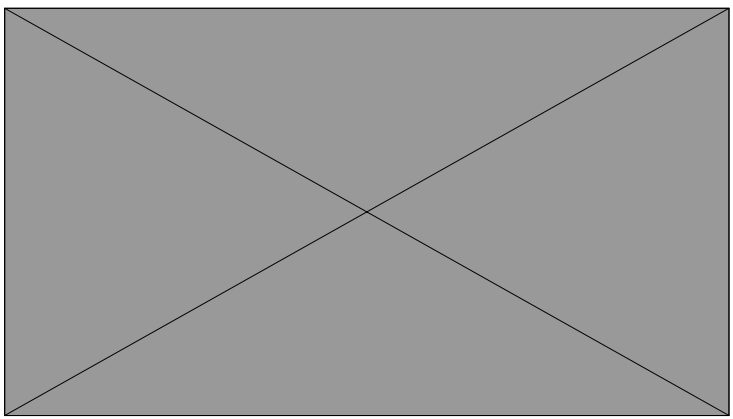

Рис. 3.
График задания № 34

Из рис. 2 и 3 видно, что данные задания относятся к заданиям, которые *не обладают способностью дифференцировать испытуемых по уровню их подготовленности.* Испытуемые с низким уровнем подготовленности имеют такую же, примерно, вероятность ответить правильно на данные задания, как и испытуемые с высоким уровнем подготовленности. Такие задания исключаются из состава проектируемого теста.

В третью группу вошли задания № 11, 12, 13 и 14. Приведём пример графика, типичного для данной группы (рис.4).

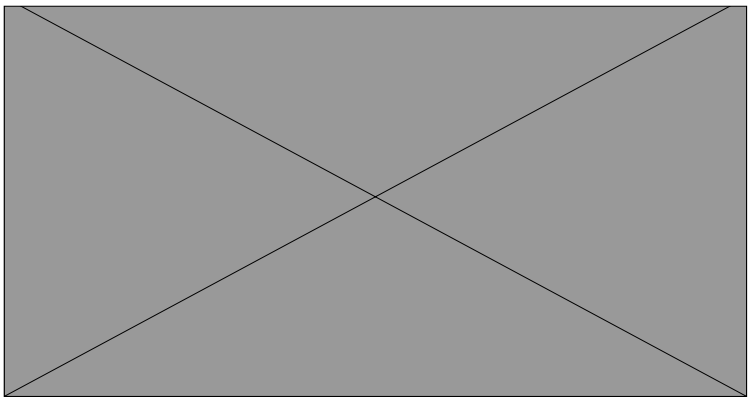

Рис. 4.
График задания № 11

Из рис. 4 видно, что данное задание можно отнести к заданиям со *«сверхвысокой дифференцирующей способностью»*. Испытуемые с низким уровнем подготовленности вообще не отвечают на данное задание, испытуемые со средним уровнем — отвечают соответственно модели, а испытуемые с высоким уровнем подготовленности — практически все отвечают верно.

И последняя группа включает в себя задание № 32. Приведём его график (рис. 5).

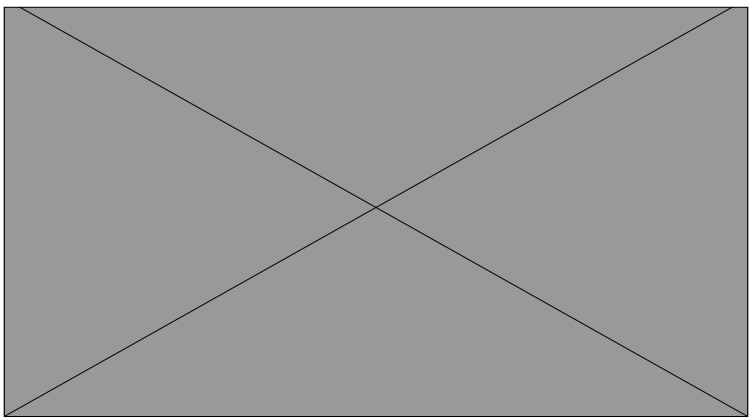

Рис. 5.
График задания № 32

Согласно рис. 5, данное задание характеризуется как задание с «*обратной дифференцирующей способностью*». Испытуемые с низким и средним уровнями знаний имеют большую вероятность правильно ответить на это задание, чем испытуемые с высоким уровнем знаний. Такое задание необходимо исключить.

Таким образом, дальше во втором наборе тестовых заданий мы будем исследовать данные, состоящие из 33 тестовых заданий, так как в течение шести этапов было исключено 17 заданий.

3. Дистракторный анализ заданий. Дистрактором называется неверный, но правдоподобный ответ, отвлекающий незнающих испытуемых от правильного ответа. То есть дистрактор — это готовый вариант ответа на задание в тестовой форме, похожий на правильный ответ, но таковым не являющийся. «Слабыми» в заданиях с выбором одного или нескольких правильных ответов называют дистракторы, которые выбирают очень мало испытуемых, «сильными» — те дистракторы, которые выбирают многие.

Хорошие дистракторы похожи на правильный ответ, они позволяют проверить уровень знаний испытуемого. Если ни один из испытуемых не выбирает какой-то определённый дистрактор, то встаёт вопрос о целесообразности его использования.

Программа RUMM даёт возможность увидеть работу дистракторов наглядно. Для первого набора тестовых заданий работа дистракторов представлена на рисунке 6.

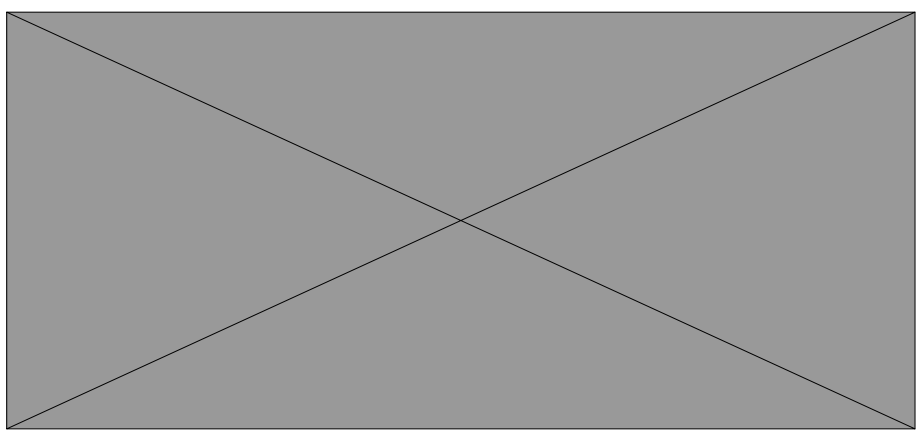

Рис. 6. **Дистракторы для первой матрицы**

По рис. 6 можно определить, какие из заданий имеют правильный / неправильный профиль. В анализируемом наборе тестовых заданий на задания № 4, 5, 6, 7, 8, 9 и 10 большинство тестируемых ответило правильно. А на задания № 1, 2 и 3 — больше неправильных ответов. Следовательно, в данных заданиях следует обратить внимание на формулировку дистракторов.

Для второго набора тестовых заданий работа дистракторов представлена на рисунке 7.

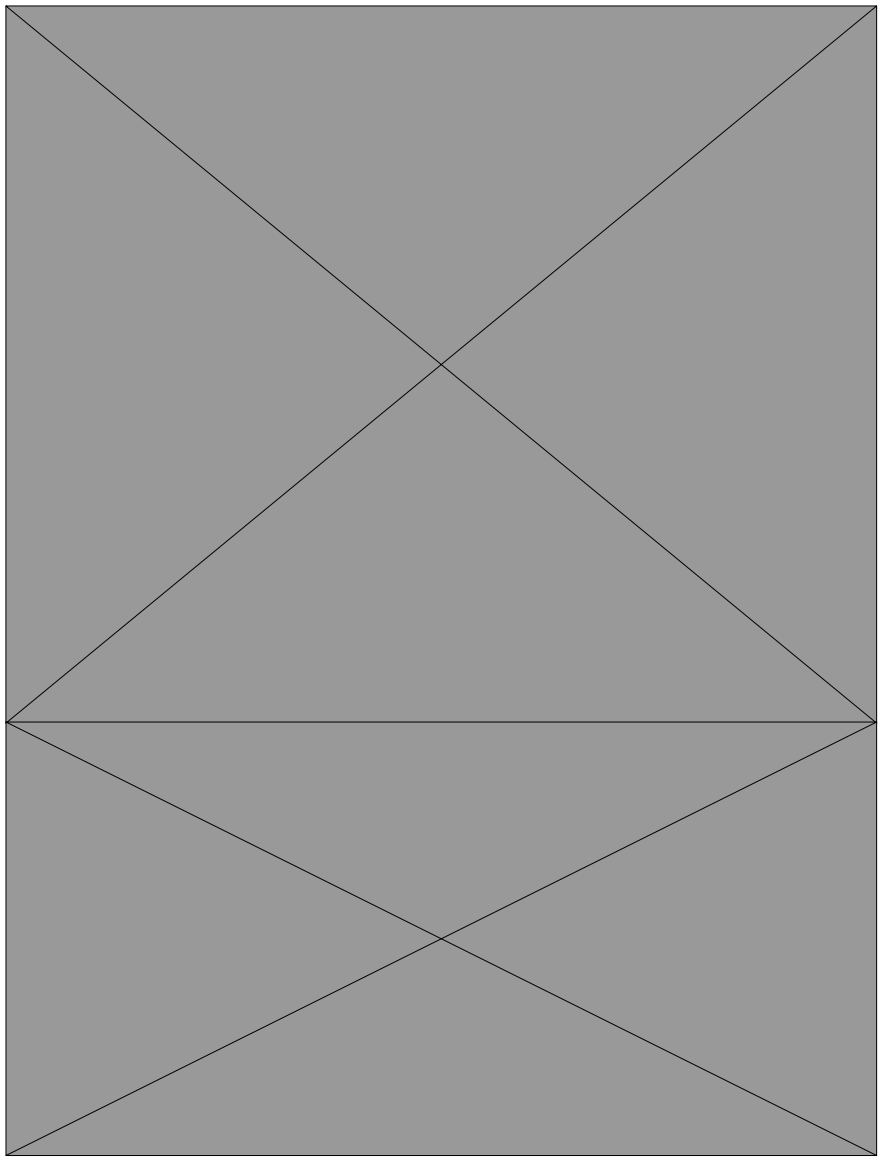

Рис. 7. **Дистракторы для второй матрицы**

По рис. 7 видно, что на задания № 3, 12, 15, 30 и 31 большинство испытуемых ответили правильно. А на задания № 2, 5, 6, 8, 18, 19, 21, 25 и 26 оказалось много неправиль-

ных ответов. Аналогично, в данных заданиях следует обратить внимание на формулировку дистракторов.

4. Равномерность возрастания меры трудности задания. Трудность соседних заданий в тесте не должна отличаться более чем на 0,5 логита. Значение 0,5 выбрано на том основании, что в хорошем тесте ошибка измерений уровня знаний испытуемых находится в пределах 0,25 логитов. Если это условие не выполняется, то тест не является эффективным измерительным инструментом. Это объясняется тем, что испытуемые внутри расширенного диапазона не дифференцируются.

Согласно рис. 8 и 9, критерий распределения заданий по трудности для первого и второго наборов данных соблюдается, то есть трудность соседних заданий не превышает порог в 0,5 логитов. Таким образом, оба набора тестовых заданий могут являться эффективным измерительным инструментом.

5. Соответствие меры трудности разрабатываемого теста уровню подготовленности испытуемых. Средняя трудность заданий не должна отличаться от среднего уровня подготовленности испытуемых более чем на 0,5 логита. Напомним, что при анализе результатов тестирования на основе модели Раша уровень подготовленности испытуемых и трудность тестовых заданий измеряются на одной и той же интервальной шкале. Поэтому если средняя трудность заданий отличается от среднего уровня подготовленности испытуемых более чем на 0,5 логита, то это означает, что уровень подготовленности некоторых испытуемых (в нижней или верхней части шкалы) плохо дифференцируется.

Для исследуемых наборов тестовых заданий (рисунки 8 и 9) данный критерий не соблюдается. Было получено следующее соответствие между уровнями подготовленности испытуемых и трудностью заданий (рисунках 8 и 9).

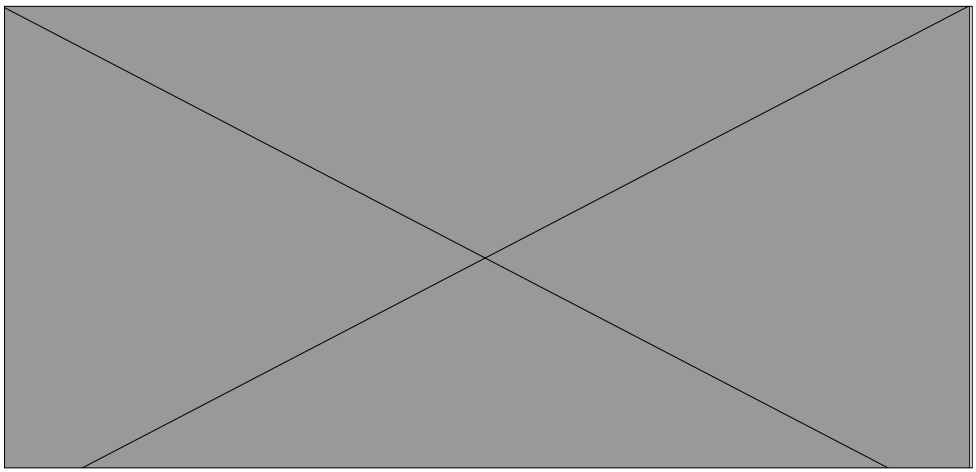

Рис. 8. **Соответствие между уровнями подготовленности испытуемых и трудностью заданий первого набора данных**

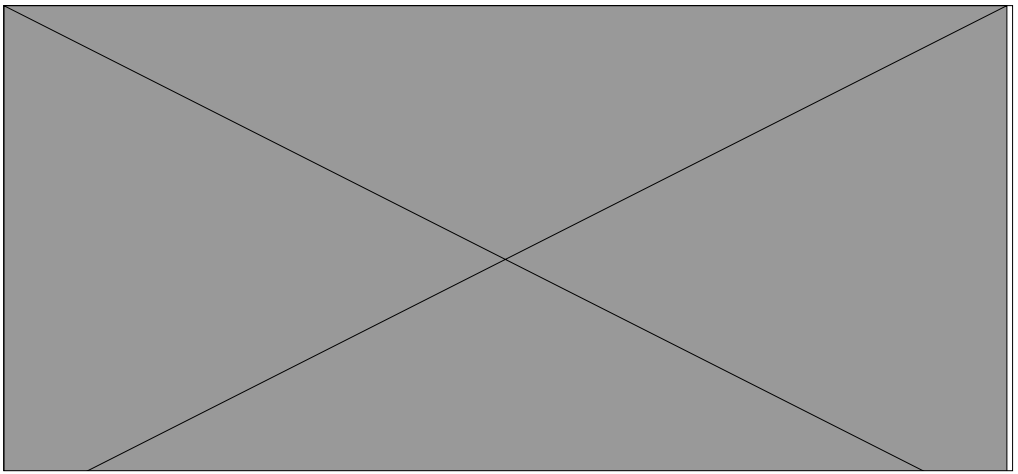

Рис. 9. **Соответствие между уровнями подготовленности испытуемых и трудностью заданий второго набора данных**

Данный рисунок является гистограммой. По оси абсцисс откладывается значение латентной переменной, по оси ординат — число индивидов (вверху) и заданий (внизу). Причём с левой стороны указаны абсолютные значения для индивидов/заданий, а с правой — относительные значения в процентах от общего числа индивидов/заданий.

В верхней части гистограммы первой матрицы указано, что число испытуемых в ней равно 13 (во второй матрице — 52). Средний уровень подготовленности испытуемых первой матрицы на 0,04 логита ниже среднего уровня трудности заданий. Уровень трудности заданий в процессе шкалирования выбирается равным нулю, что делается для удобства интерпретации. Для второй матрицы средний уровень подготовленности испытуемых на 0,83 логита оказался выше среднего уровня трудности заданий. Это означает, что в анализируемых данных обнаружено явное несоответствие уровня подготовленности и трудности заданий. В этой ситуации уровень подготовленности некоторых испытуемых дифференцируется недостаточно. В первой матрице необходимо уменьшить трудность некоторых заданий, а во второй матрице – увеличить трудность заданий. Этот критерий анализа уникален.

Среднеквадратическое отклонение для оценок уровней знаний в первом случае равно 1,53 (во втором — 1,08). Кроме того, в верхней части гистограммы указано, что цена деления на оси абсцисс равна 0,25 (0,20) логита, в результате в интервале от –3 до +3 логит рассматривается 24 (35) групп с различными значениями латентной переменной.

В идеальном случае гистограмма распределения ответов испытуемых (верхняя часть рисунка) должна быть близка к нормальному закону распределения: относительно небольшое число испытуемых с низким и высоким уровнем знаний и относительно

много со средним уровнями знаний. Гистограмма распределения меры трудности заданий (нижняя часть рисунка) должна быть близка к равномерному закону распределения: трудность заданий должна возрастать не скачкообразно, а равномерно.

Здесь уместно напомнить определение *качественного* теста, состоящего из *системы* заданий равномерно возрастающей трудности и позволяющего получать педагогически целесообразные результаты, отвечающие критериям *надёжности, валидности, объективности и эффективности*[1]. В этом определении курсивом выделены основные термины, позволяющие отграничить педагогические измерения от прочих методов — научных, псевдонаучных и ненаучных[2]. Задания теста должны быть равномерно распределены по шкале логитов. Это означает, что разработанный набор заданий позволяет качественно оценивать уровень знаний.

В первом наборе тестовых заданий уровни знаний испытуемых распределены по нормальному закону распределения. Уровни трудности заданий распределены по шкале логитов примерно равномерно. Во втором случае этот критерий не соблюдается. На гистограмме видны явные резкие скачки столбцов, что снова говорит о разнице уровней знаний испытуемых и трудности заданий.

6. Достаточность вариации и размаха заданий по уровню их трудности. Данный критерий должен иметь значение больше трёх логитов. Если размах не достигает данного значения, то необходимо добавлять в тест задания всех уровней трудности.

Согласно рис. 8 и 9, диапазон значений первого набора тестовых заданий находится на отрезке от –2 до +2,8 (второго от –2,6 до +2,8) по шкале логитов, следовательно, размах варьирования заданий составляет 4,8 (5,4) логита. Данные значения удовлетворяют выдвинутому критерию; наборы тестовых заданий не следует усложнять.

Программа RUMM позволяет расшифровать значения верхних столбцов гистограммы. Для первого набора заданий эти данные представлены на рис. 10.

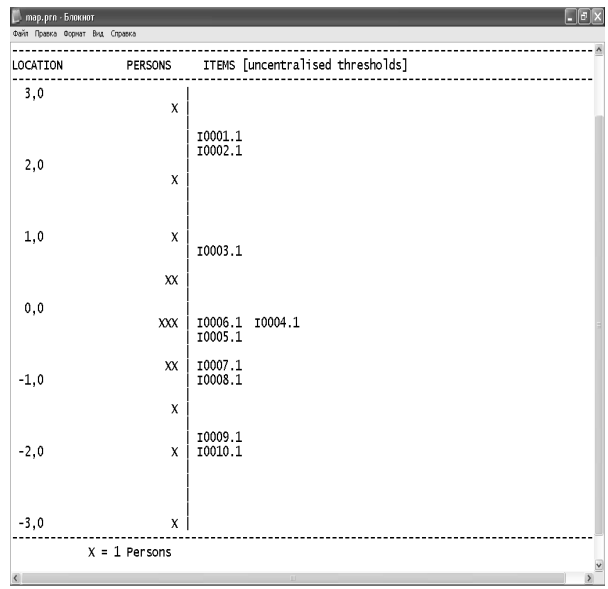

Рис. 10. **Значения верхних столбцов гистограммы для первой матрицы**

---

[1] *Аванесов В.С.* Понятие и методы математической теории педагогических измерений (Item Response Theory): статья третья. Педагогические измерения. № 4. 2009. С. 5.

[2] *Аванесов В.С.* Метрическая система Георга Раша — Rasch Measurement (RM)//ПИ № 2. 2010. С. 57–80.

Из рисунке 10 видно, что на задания № 4 и № 6 ответили правильно больше испытуемых, чем на задания № 9 и 10. Следовательно, первая группа заданий обладает меньшей трудностью, нежели вторая. Данные рисунка 10 подтверждают выводы, сделанные по рисунку 6.

Для второго набора тестовых заданий значения верхних столбцов гистограммы представлены на рисунке 11.

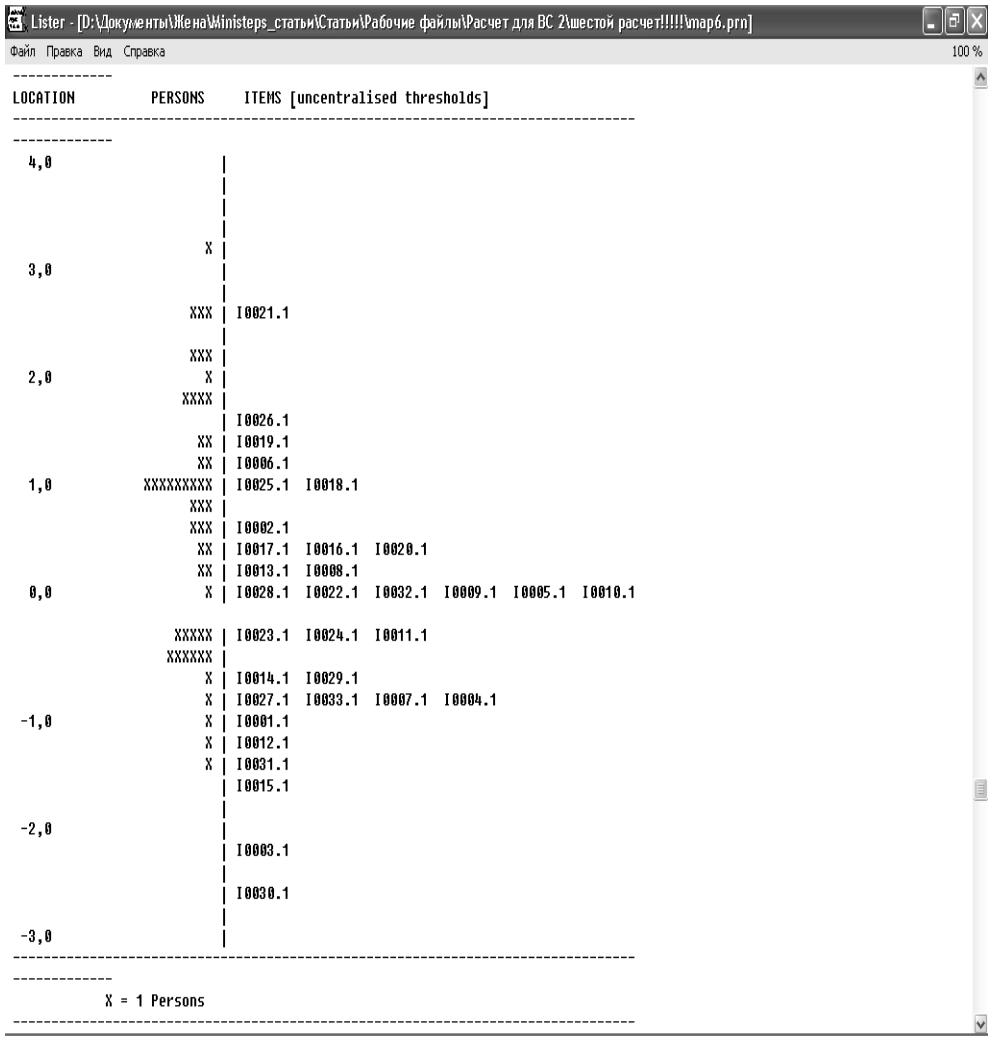

Рис. 11. **Значения верхних столбцов гистограммы для второй матрицы**

Из рисунка 8 видно, что на задания № 28, 22, 32, 9, 5 и 10 ответили правильно большее число респондентов, чем на задания № 21, 26, 19, 6, 2, 1, 12, 31, 15, 3 и 30. Следовательно, первая группа заданий обладает меньшей трудностью, нежели вторая. Данные рисунка 11 подтверждают выводы, сделанные по рисунку 7.

7. *Определение уровня подготовленности испытуемых.* Программа RUMM-2020 позволяет оценить уровень подготовленности испытуемых. Среди испытуемых легко выделяются те, кто обладает высоким, средним и низким уровнями подготовленности. Эти данные позволяют скорректировать процесс преподавания для конкретных испытуемых, применив тем самым индивидуальный подход в обучении.

Приведённые критерии являются важным средством повышения эффективности и качества тестов. Учёт этих критериев позволяет разрабатывать педагогические тесты, которые достаточно объективно измеряют уровень подготовленности испытуемых и меру трудности заданий.