

М е т о д о л о г и я

Применение компьютерной программы RUMM для создания качественных и объективированных методов педагогической диагностики и педагогических измерений

Галина Смирнова,
smirnova_g_i@mail.ru

Показаны метрические возможности компьютерной программы RUMM-2020 и RUMM-2030 для разработки методов педагогической диагностики и педагогических измерений. Для этого исследования были использованы две матрицы результатов тестирования. Первая матрица имеет размер 13×10 и является известным примером учебной матрицы В.С. Аванесова¹.

Вторая матрица представлена им же. Она взята из реальной практики проверки тестовых свойств заданий в тестовой форме, предложенных привлечёнными педагогами-методистами по русскому языку. Эта матрица имеет размер 52×50, где 52 — это число испытуемых, а 50 — число заданий. Матрица использована для эмпирической проверки гипотезы о пригодности представленных заданий при пробном тестировании абитуриентов по русскому языку, в Централизованном тестировании 1998 г. В процессе эмпирической апробации около трети предполагаемых заданий оказались непригодными для включения в тест Централизованного тестирования, а потому были удалены из предварительного набора заданий проектируемого теста.

¹ Аванесов В.С. Композиция тестовых заданий. М., 1998. <http://viperson.ru/articles/kompozitsiya-testovyh-zadaniy>

В данной статье вторая матрица использовалась для иллюстрации алгоритма разработки педагогического теста на основе модели Г. Раша, посредством программы RUMM-2020 (Rasch Unidimensional Measurement Models).

Анализ качества заданий и теста

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

Экстремальность заданий в тестовой форме

Экстремальными называются задания, на которые все испытуемые данной выборки отвечали одинаково — либо правильно, либо неправильно. Соответственно, они очень лёгкие или очень трудные. Такие задания не считаются тестовыми, потому что обладают нулевой дифференцирующей способностью.

Уровень трудности таких заданий выходит за пределы возможности проведения педагогических измерений на данной выборке испытуемых, а потому результаты, полученные по таким заданиям, исключаются из числа пригодных для включения в тест.

Экстремальные задания не включаются в тест при проведении измерений в данной выборке испытуемых. Но допускается возможность их применения в других выборках, если там они окажутся в числе заданий, дифференцирующих испытуемых.

В исследуемых матрицах экстремальных заданий не оказалось.

Совместимость заданий в тесте

Совместимость заданий является необходимым условием использования модели Раша для измерения латентных качеств посредством ка-

чественного педагогического теста — в частности, для измерения уровня подготовленности испытуемых. 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 13×10 *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 высокая. 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.

Как видно из табл. 1, *совместимость всех заданий матрицы результатов тестирования* постепенно

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

возрастала с нулевого значения на первом этапе до значения 0,222138 на последнем этапе. Данный положительный результат достигнут за счет удаления из матрицы тестирования некачественных заданий, значение Хи-квадрат которых не достигло критического, равного 0,05.

Анализ мер совместимости для каждого задания показал, что все задания в первом наборе учебного

примера матрицы 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 эмпирически полученной матрицы, на каждом из этапов исследования, отражен в табл. 2.

Таблица 2

Значение 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 ЭТАП	
25	0,033286
ШЕСТОЙ ЭТАП	
–	–

На последнем этапе анализ *совместимости для каждого задания* показал, что все отобранные для теста задания соответствуют требованиям модели Раша.

Графические образы заданий²

Графики заданий, исключенных 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. Приведем пример графика, типичного для данной группы (рис. 1).

Из рис. 1 видно, что данное тестовое задание можно отнести к заданиям с т.н. «*неупорядоченными ответами*». Испытуемые с низким и высоким уровнем подготовленности имеют сравнительно большую вероятность правильно ответить на это задание, чем испытуемые, имею-

щие средний уровень подготовленности. Такое задание «не вписывается» к требованиям к тестовым заданиям, что подтверждено данными табл. 2. Это задание нарушает основное положение (assumption) двух основных теорий — Rasch Measurement и 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 логита до $+4$ логита³. 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 (10001);
- название тестового задания, здесь названия задания выбраны по

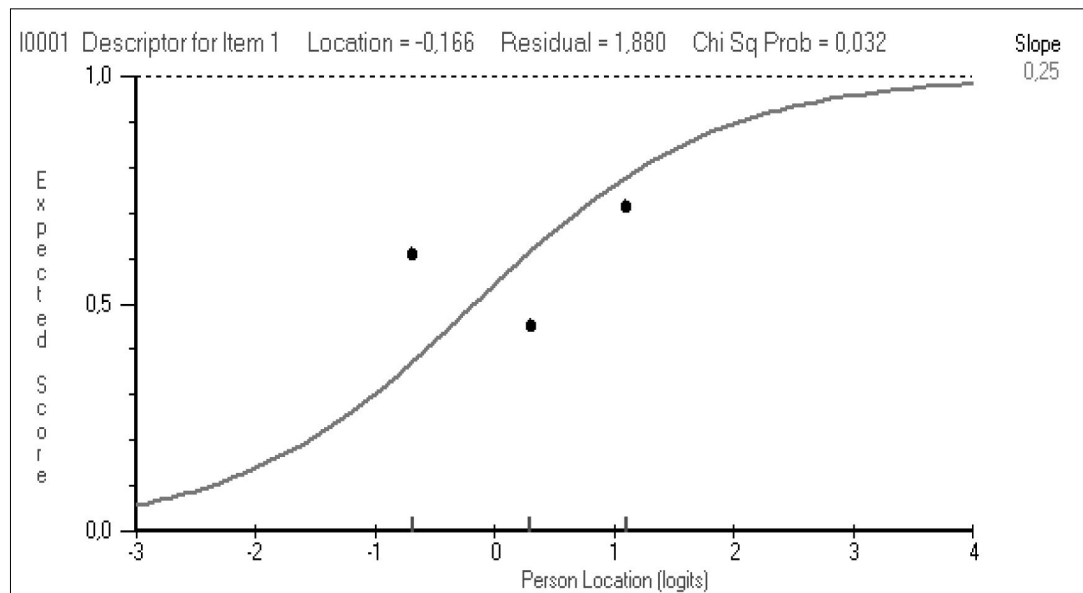


Рис. 1. График задания № 1

² В переводах на русский язык зарубежных статей и книг эти графики нередко называются характеристическими образами заданий (Прим. ред.).

³ В данном случае логит — это стандартная единица измерения уровня трудности задания. В Rasch Measurement используется и другой логит — стандартная единица измерения уровня подготовленности испытуемого.

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. Приведем примеры графиков, типичных для данной группы (рис. 2 и 3).

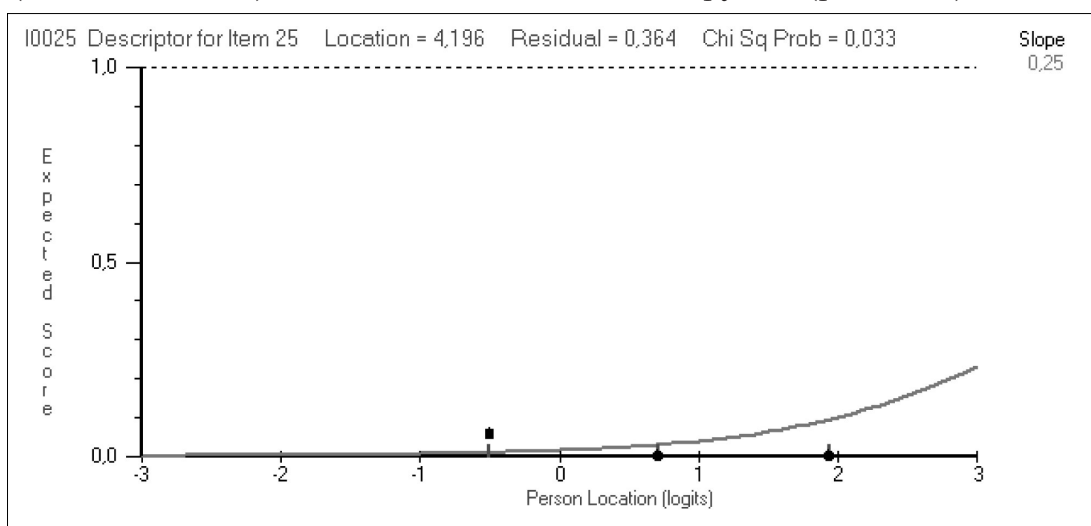


Рис. 2. График задания № 25

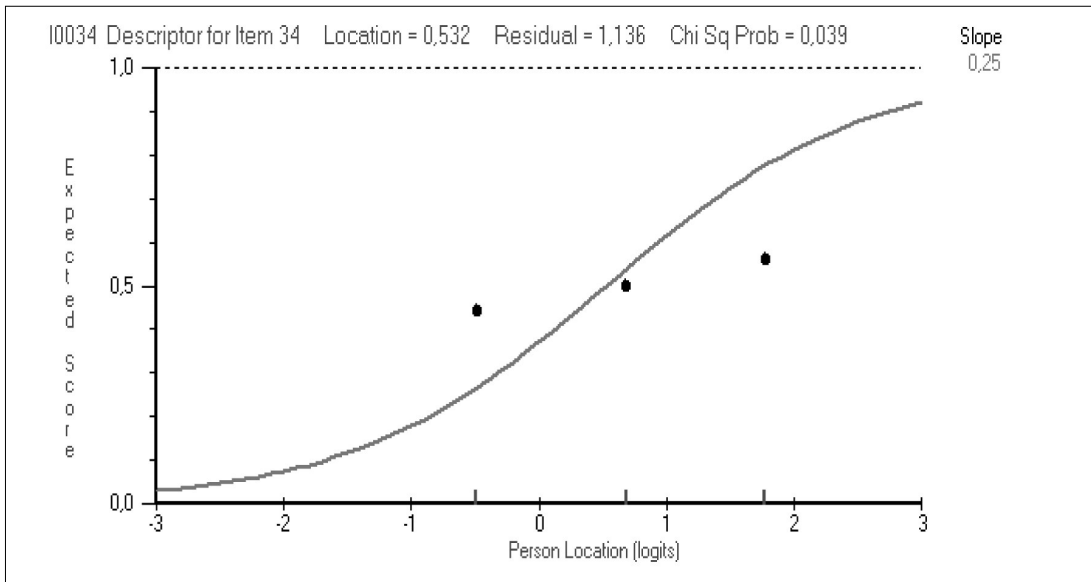


Рис. 3. График задания № 34

Из рис. 2 и 3 видно, что данные задания относятся к заданиям, которые не обладают способностью дифференцировать испытуемых по уровню их подготовленности. Ис-

пытующие с низким уровнем подготовленности имеют такую же, примерно, вероятность ответить правильно на данные задания, как и испытуемые с высоким уровнем под-

готовленности. Такие задания исключаются из состава проектируемого теста.

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

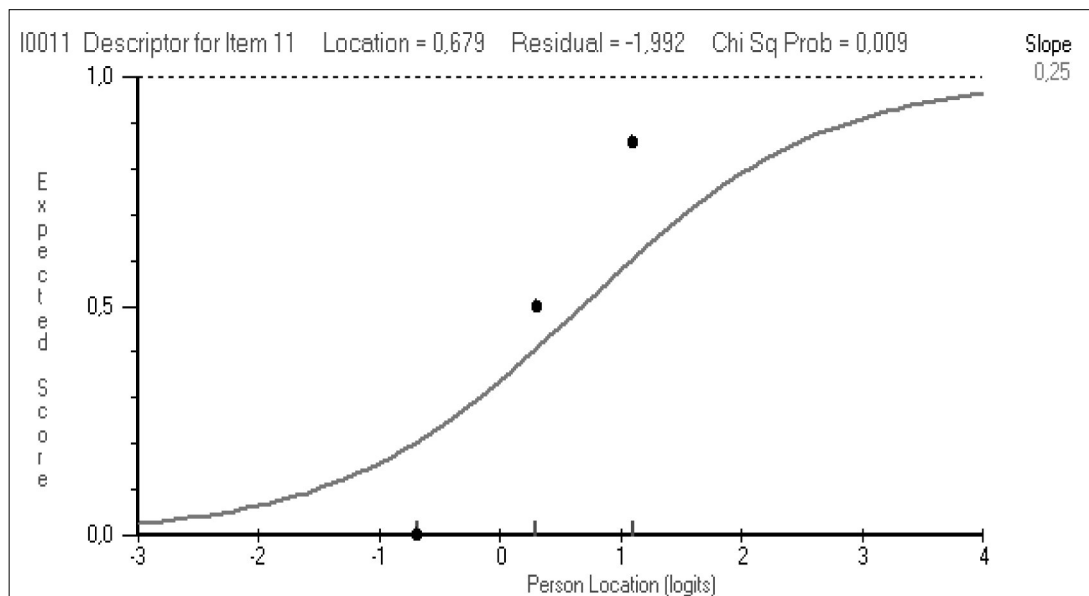


Рис. 4. График задания № 11

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

не отвечают соответственно модели, а испытуемые с высоким уровнем подготовленности — практически все отвечают верно.

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

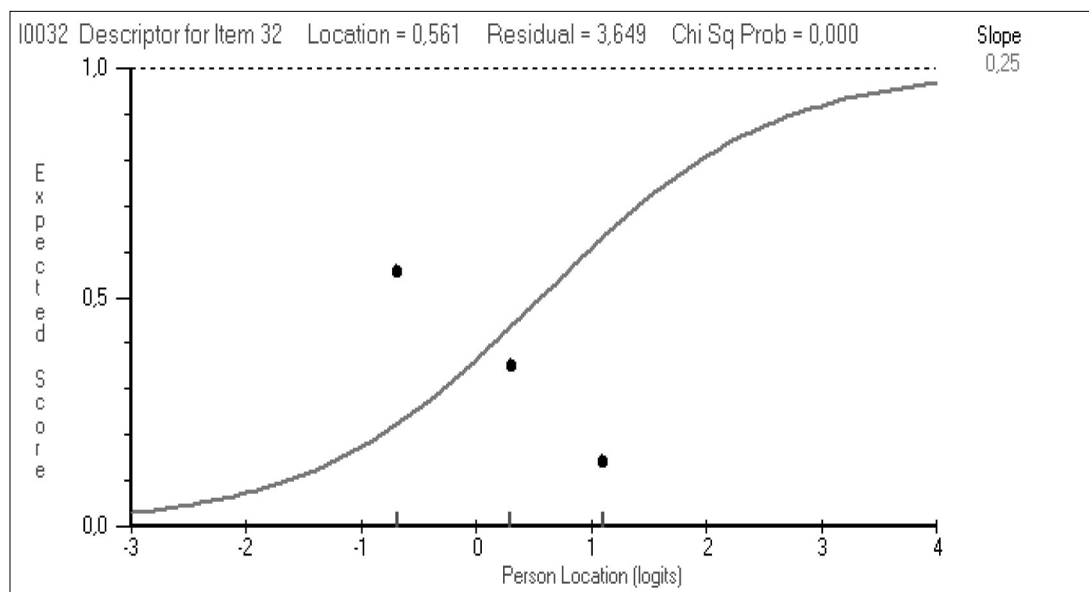


Рис. 5. График задания № 32

Согласно рис. 5, данное задание характеризуется как задание с т.н. «обратной дифференцирующей способностью». Испытуемые с низким и средним уровнем знаний имеют большую вероятность правильно ответить на это задание, чем испытуемые с высоким уровнем знаний. Такое задание исключается из совокупности гипотетически пригодных для измерения заданий, как алогичное.

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

Дистракторный анализ заданий

Одно из свойств качественного теста с выбором одного или нескольких правильных ответов — наличие качественных дистракторов. Дистрактором в задании называется ответ неверный, но правдоподобный, 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 дистрактор, то встает вопрос о целесообразности использования такого дистрактора.

Программа RUMM дает возможность увидеть работу дистракторов наглядно.

Равномерность возрастания меры трудности задания

Мера трудности 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 логита. Если это условие не выполняется, то тест не является эффективным измерительным инструментом. Это объясняется тем, что испытуемые внутри расширенного диапазона не дифференцируются. Согласно рис. 6 и 7, критерий распределения заданий по трудности для первого и второго наборов данных соблюдается, то есть трудность соседних заданий не превышает порог в 0,5 логита. Таким образом, оба отобранных посредством программы RUMM набора тестовых заданий могут являться качественным измерительным инструментом.

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 (рис. 8 и 9) данный критерий не соблюдается. Было получено следующее соответствие между уровнями подготовленности испытуемых и трудностью заданий (рис. 6 и 7).

Следовательно, подбор заданий по вариации уровня трудности заданий или состав выборки по вариации уровня подготовленности испытуемых нуждается в улучшении. Для небольших матриц это естественно, для больших матриц

означает наличие возможных дефектов.

Дефекты несоответствия уровня трудности заданий и уровня подготовленности испытуемых обнаруживаются и во второй матрице эмпирических данных.

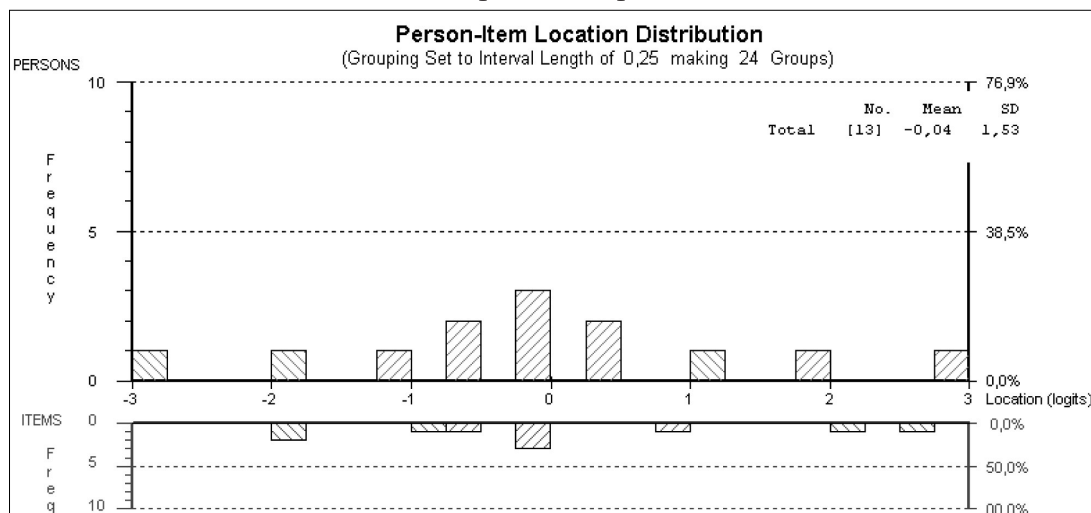


Рис. 6. Соответствие между уровнями подготовленности испытуемых и трудностью заданий в учебной матрице, с малым числом испытуемых и заданий

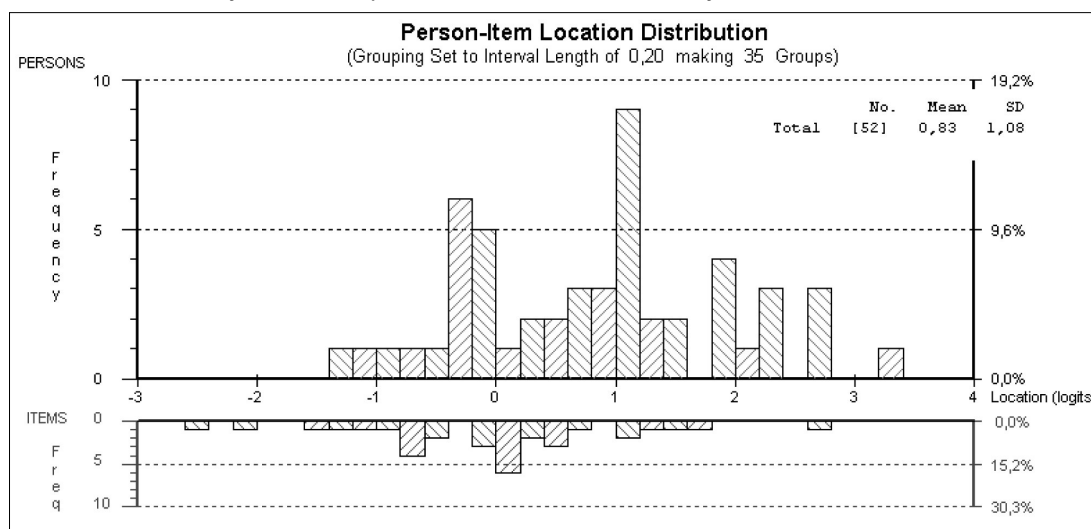


Рис. 7. Соответствие между уровнями подготовленности испытуемых и трудностью заданий второго набора данных

Данный рисунок является гистограммой. По оси абсцисс откладывается значение латентной переменной, по оси ординат — число индивидов (вверху) и заданий (внизу). Причем с левой стороны указаны абсолютные значения для индивидов/заданий, а с правой — относи-

тельные 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 ряда заданий. Этот критерий анализа можно отнести к числу важных и уникальных в зарубежных теориях педагогической диагностики и педагогических измерений.

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) группы 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 уровнем знаний. Гистограмма распределения меры трудности заданий (нижняя часть рисунка) должна быть близка к равномерному закону распределения: мера трудности 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*⁴. В этом определении курсивом выделены основные термины, позволяющие отграничить педагогические измерения от прочих методов — научных, псевдонаучных и ненаучных⁵. Задания теста должны быть равномерно распределены по шкале логитов. Это означает, что в общем разработанный набор заданий позволяет, для начала, качественно оценивать уровень знаний, но нуждается в дальнейшем улучшении.

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.

Достаточность вариации и размаха заданий по уровню их трудности

Данный критерий должен иметь значение больше трех логитов. Если размах не достигает данного значения, то необходимо добавлять в тест задания всех уровней трудности, и особенно — лёгкие и трудные.

Согласно рис. 6 и 7, 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, раз-

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

⁵ Аванесов В.С. Метрическая система Георга Раша — Rasch Measurement (RM)//ПИ №2. 2010. С. 57–80.

мах варьирования заданий составляет 4,8 (5,4) логита. Данные значения удовлетворяют выдвинутому критерию, а потому наборы тестовых заданий не следует усложнять.

Программа RUMM позволяет расшифровать значения верхних столбцов гистограммы. Для первого набора заданий эти данные представлены на рис. 8.

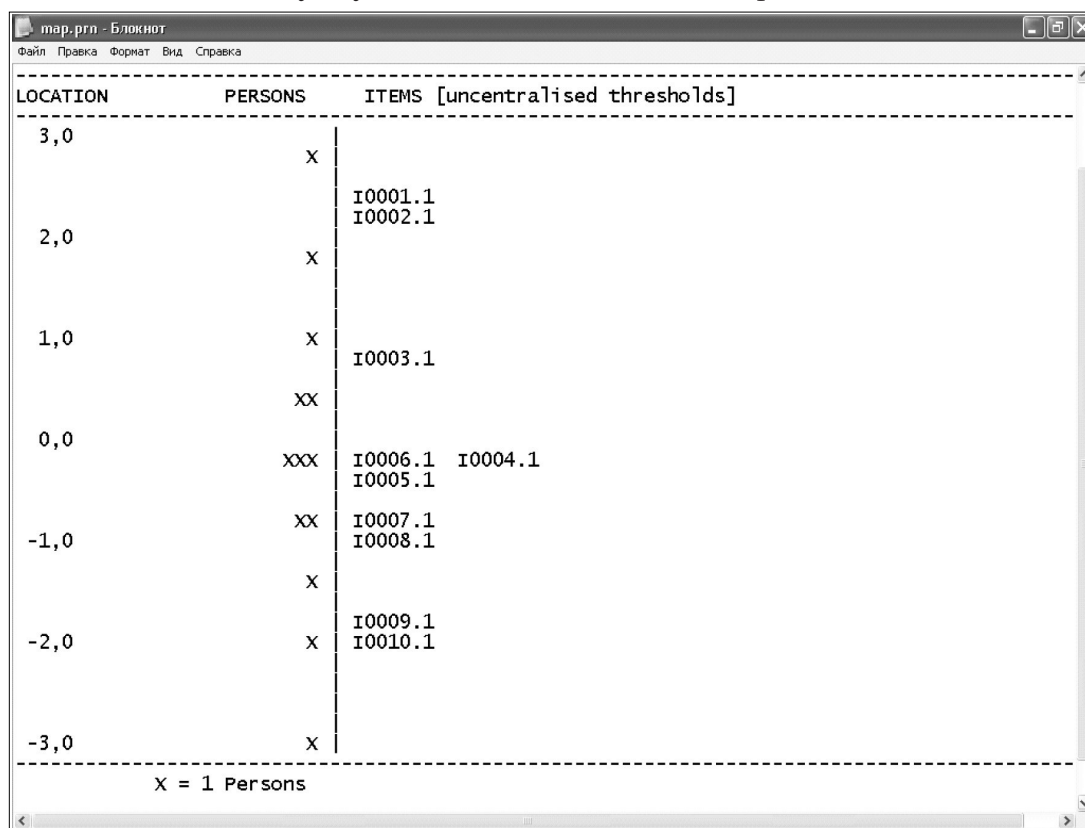


Рис. 8. Значение верхних столбцов гистограммы для первой матрицы

Из рис. 8 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.

Для второго набора тестовых заданий значения верхних столбцов гистограммы представлены на рис. 9.

Из рис. 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.

Определение уровня подготовленности испытуемых

Среди испытуемых легко выделяются те, кто обладает высоким, средним и низким уровнем подготовленности. Эти данные позволяют скорректировать процесс преподавания для конкретных испытуемых, применив, тем самым, индивидуальный подход в обучении.

Приведенные критерии являются важным средством повышения эффективности и качества тестов. Учет этих критериев позволяет разрабатывать педагогические тесты, которые достаточно объективно измеряют

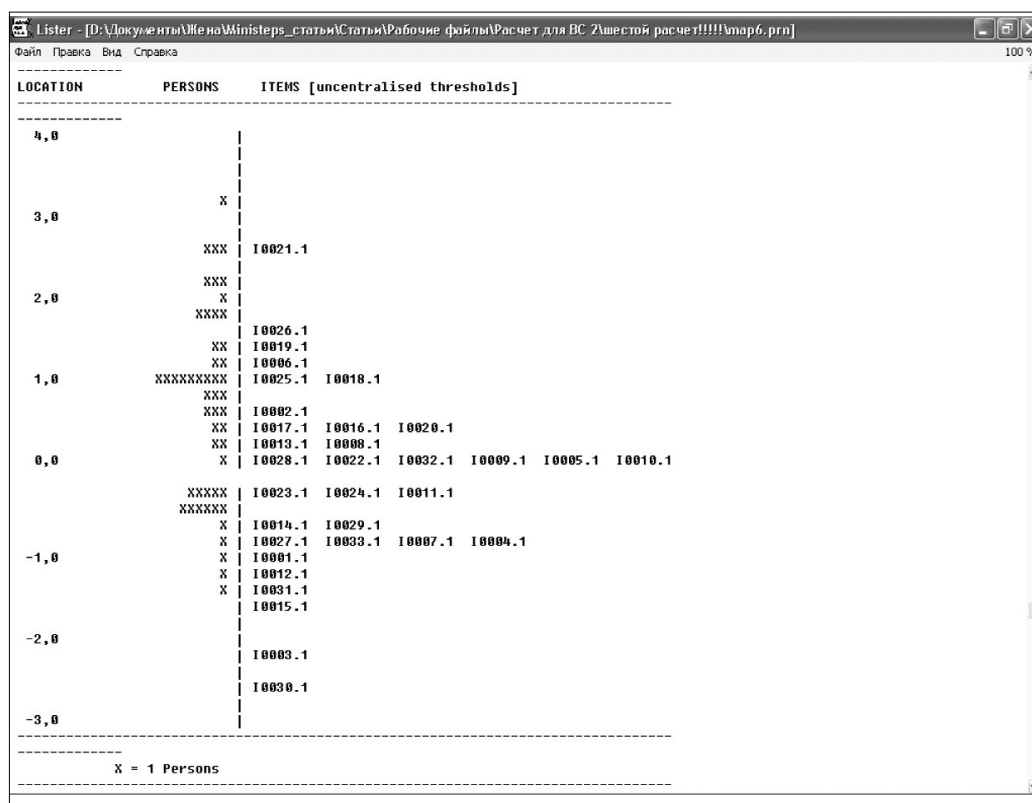


Рис. 9. Значение верхних столбцов гистограммы для второй матрицы

уровень подготовленности испытуемых и меру трудности заданий.

Вывод

Программа RUMM-2020 и RUMM-2030 позволяет оценить уровень подготовленности испытуемых и

пригодность многих, но далеко не всех заданий, рекомендуемых педагогами-предметниками для создания теста. Применение такой программы позволяет повысить качество разрабатываемых тестов, что является сверхактуальным для российской практики разработки диагностических тестов.