

УДК 37.091.3:004.9

Смирнова Г.И.

ПРИМЕНЕНИЕ КОМПЬЮТЕРНОЙ ПРОГРАММЫ RUMM ДЛЯ РАЗРАБОТКИ ПЕДАГОГИЧЕСКОГО ТЕСТА

Ключевые слова: уровень подготовленности испытуемого, уровень сложности задания, модель Раша, латентная переменная, педагогический тест.

Педагогический тест – репрезентативная система параллельных заданий возрастающей трудности, специфической формы, определенного содержания, система, создаваемая с целью объективной оценки структуры и уровня подготовленности обучаемых [1]. Для разработки педагогического теста необходимо преодолеть три этапа. Первый начинается с разработки заданий в тестовой форме. Вторым этапом – статистическая проверка заданий с целью обоснования их в качестве тестовых. И третий – разработка теста.

В данной статье описан второй этап разработки педагогического теста.

Вначале необходимо дать определение понятия «латентная переменная». Латентные переменные величины – это положительные и отрицательные качества личности, не поддающиеся непосредственному измерению. Примерами являются «подготовленность студентов», «знание учебной дисциплины», «способность понимать», «интеллектуальное развитие» и многое другое. Это означает, что практически все измеряемые свойства личности рассматриваются как латентные.

Тесты анализируются на основе модели Раша. Модель Раша – простая логистическая модель [3]. Модель Раша превращает измерения, сделанные в дихотомических и порядковых шкалах, в линейные измерения, в результате качественные данные анализируются с помощью количественных методов. Модель Раша позволяет использовать широкий спектр статистических процедур для анализа результатов измерений. Оценка трудности тестовых заданий не зависит от выборки испытуемых, на которых была получена [2].

В качестве примера, на котором будут рассмотрены параметры оценки качества заданий педагогического теста, возьмем латентную переменную

«уровень подготовленности студентов по дисциплине “Современные средства оценивания результатов обучения”». Данная переменная оценивалась в группе студентов факультета педагогики и методики начального образования Славянского-на-Кубани государственного педагогического института (СГПИ). Исследование проходило по следующим критериям:

1. Экстремальные задания. Выявление и исключение из теста экстремальных заданий, трудность которых нельзя измерить при использовании модели Раша. Так, если все испытуемые правильно ответили на некоторое тестовое задание, то невозможно определить трудность этого задания – трудность такого задания выходит за пределы измерений. Такое задание необходимо исключить из матрицы результатов тестирования. Аналогичным образом необходимо поступить и с теми заданиями, на которые никто из испытуемых не дал правильного ответа. В исследуемом наборе тестовых заданий таких заданий нет.

2. Совместимость тестовых заданий. Совместимость тестовых заданий является необходимым условием использования модели Раша для измерения латентных переменных – в частности, измерения уровня подготовленности. Совместимость тестовых заданий проверяется на разных уровнях: для всей матрицы результатов тестирования; для каждого тестового задания (по результатам тестирования для задания); для каждого испытуемого (по результатам тестирования для испытуемого); для каждого элемента матрицы тестирования.

Во всех четырех случаях совместимость определяется на основе критерия хи-квадрат – результаты тестирования сопоставляются с ожидаемыми значениями на основе модели Раша.

Критерий хи-квадрат Пирсона – это наиболее простой критерий проверки значимости связи между двумя категоризованными переменными. Критерий Пирсона основывается на том, что в таблице ожидаемые частоты при гипотезе «между переменными нет зависимости» можно вычислить непосредственно [6].

В исследуемом наборе тестовых заданий *совместимость всей матрицы результатов тестирования* на основе критерия хи-квадрат составляет 0,239529. Полученные значения говорят о том, что степень соответствия данных тестирования модели Раша средняя. Чем ближе данное значение к единице, тем выше степень соответствия.

Анализ *совместимости для каждого тестового задания* показал, что задания № 11, 18 и 26 в исследуемом наборе тестовых заданий не соответствуют модели Раша, так как экспериментальные значения статистики хи-квадрат для них не соответствуют табличному и составляют 0,037040, 0,031945 и 0,012470 соответственно при критическом значении 0,05. Для примера приведем формулировку одного тестового задания, которое не соответствует модели Раша.

Тестовое задание № 26

Установите соответствие:

Свойство знаний:	Фраза:
1) рефлексивность;	а) если я знаю, что некто знает нечто, то из этого следует, что я знаю это нечто;
2) транзитивность;	б) если я знаю кого-то, то это не значит, что он меня знает;
3) антисимметричность.	в) я не только знаю нечто, но и знаю, что я это знаю.

Ответы: 1 __, 2 __, 3 __.

(правильный ответ: 1 – в, 2 – а, 3 – б).

Рассмотрим график тестового задания № 26 (рис. 1).

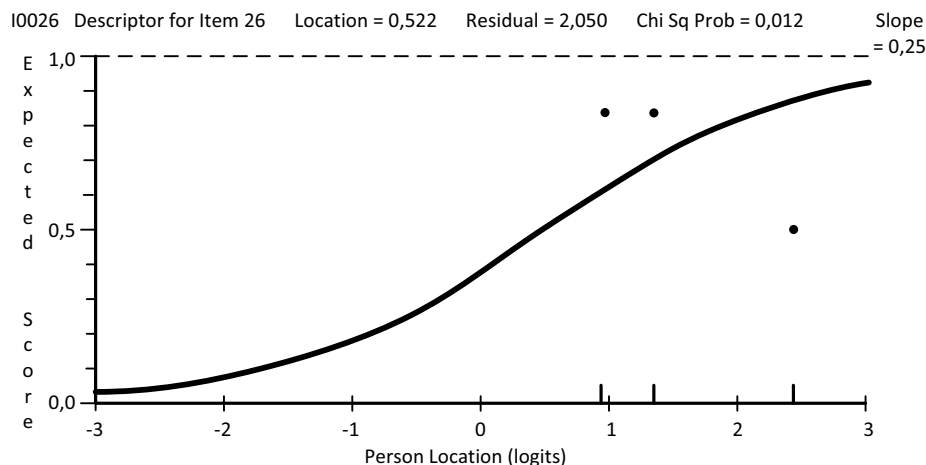


Рис. 1. График тестового задания № 26

По оси абсцисс отложены значения латентной переменной «уровень знаний по изучаемой дисциплине». В данном случае латентная переменная варьирует от -3 логита до $+3$ логита. 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 (I0026);
- название тестового задания, здесь названия задания выбраны по умолчанию. В данном случае это Descriptor for Item 26;
- трудность тестового задания (Location = 0,522);
- суммарное отклонение ответов индивидов на данное задание от ожидаемых на основе модели Раша (Residual = 2,050);
- степень соответствия данных модели Раша (Chi Sq Prob = 0,012);
- наклон кривой (Slope = 0,25).

Логит – единица измерения уровней подготовленности (знаний) участников тестирования и трудностей тестовых заданий в рамках логистических моделей тестирования [4].

Из рис. 1 видно, что данное тестовое задание можно отнести к заданиям с

«обратной дифференцирующей способностью». Такое задание должно быть переформулировано.

3. Анализ качества дистракторов. Дистрактор – это ложная, отвлекающая альтернатива среди перечня возможных ответов на вопрос тестового задания. То есть дистрактор – это готовый вариант ответа на закрытое тестовое задание, похожий на правильный ответ, но таковым не являющийся. «Слабыми» в тестах с выбором ответа называют дистракторы, которые по статистике выбирают очень мало испытуемых, «сильными» – те, которые выбирают многие [5].

Хорошие дистракторы похожи на правильный ответ, они позволяют проверить уровень знаний испытуемого. Если ни один из испытуемых не выбирает некоторый дистрактор, то встает вопрос о целесообразности использования такого дистрактора.

Программа RUMM дает возможность увидеть работу дистракторов наглядно. Для исследуемого набора тестовых заданий работа дистракторов представлена на рис. 2.

По рис. 2 можно определить, какие из заданий имеют правильный/неправильный профиль. В анализируемом наборе тестовых заданий на задания

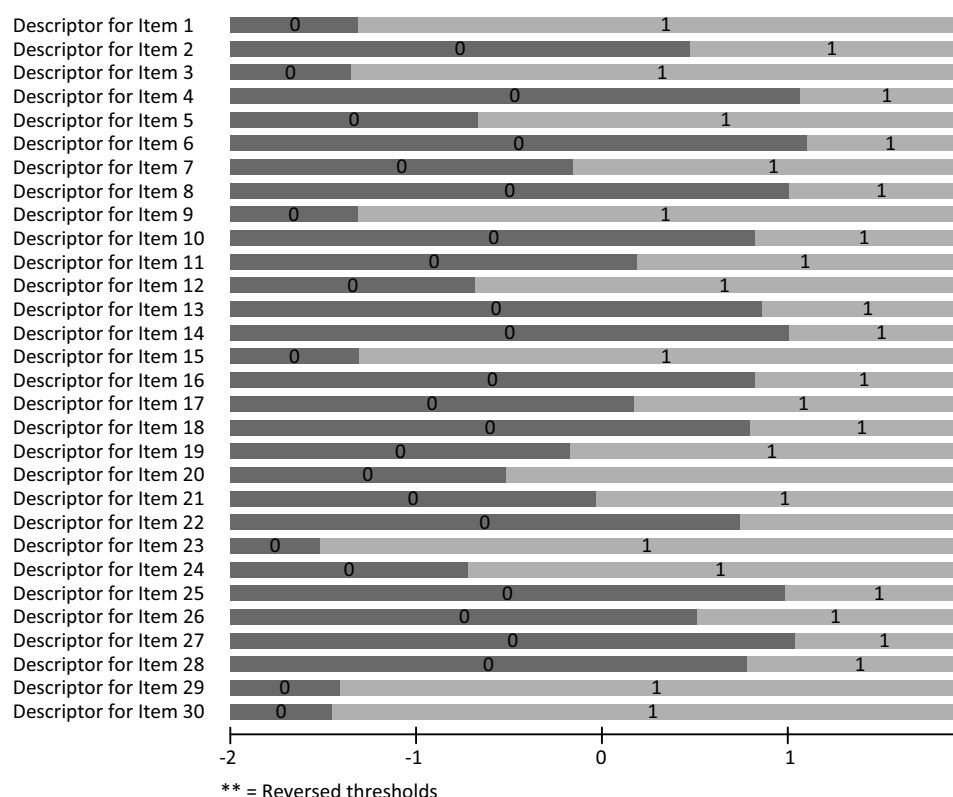


Рис. 2. Работа дистракторов

№ 1, 3, 9, 15, 23, 29 и 30 большинство тестируемых ответило правильно. А на задания № 4, 6, 8, 13, 14, 25 и 28 – больше неправильных ответов. Следовательно, в данных заданиях следует обратить внимание на формулировку дистракторов.

4. Равномерность распределения заданий по трудности. 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 логита. Если это условие не выполняется, то тест не является эффективным измерительным инструментом. Это объясняется тем, что испытуемые внутри такого диапазона не дифференцируются.

Согласно рис. 3, критерий распределения заданий по трудности соблюдается, т.е. трудность соседних заданий данного набора тестовых заданий не превышает порог в 0,5 логита. Таким образом, анализируемый набор тестовых заданий может являться эффективным измерительным инструментом.

5. Соответствие трудности 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

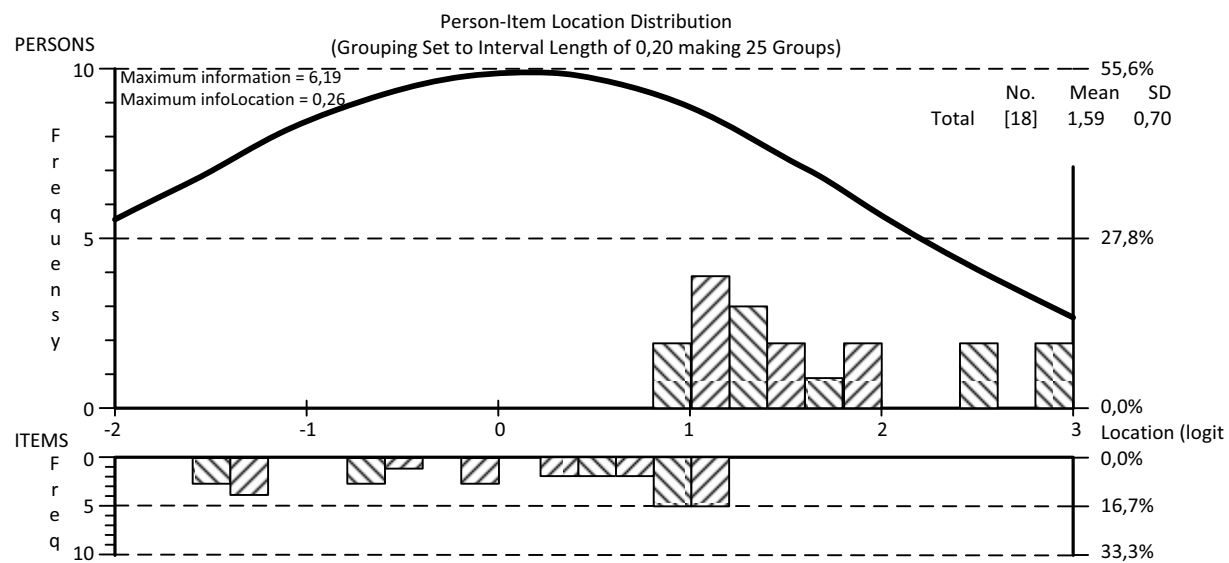


Рис. 3. Соответствие между уровнем знаний студентов и трудностью заданий

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 исследуемого набора тестовых заданий (рис. 3) данный критерий не соблюдается.

Было получено следующее соответствие между уровнем знаний студентов и трудностью заданий (рис. 3).

Данный рисунок является гистограммой. По оси абсцисс откладывается значение латентной переменной («уровень подготовленности испытуемых по учебной дисциплине «Современные средства оценивания результатов обучения»» и «уровень трудности задания»), по оси ординат – число индивидов (вверху) и заданий (внизу). Причем 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 указано, что число индивидов равно 18, средний уровень знаний индивидов на 1,59 логита выше среднего уровня сложности заданий, который для удобства интерпретации выбран равным 0. Это означает, что в анализируемых данных обнаружено явное несоответствие уровня подготовленности и сложности заданий. То есть уровень подготовленности некоторых испытуемых плохо дифференцируется, так как необходимо увеличить сложность некоторых тестовых заданий. Этот критерий анализа уникален.

Среднеквадратическое отклонение для оценок уровней знаний равно 0,70. Кроме того, в верхней части гистограммы указано, что цена деления по оси абсцисс равна 0,20 логита, в результате в интервале от –2 до +3 логитов рассматривается 25 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 студентов с низким и высоким уровнем знаний и относительно много со средним уровнем знаний. Гистограмма распределения трудностей заданий теста (нижняя часть рисунка) должна быть близка к равномерному закону распределения: трудности заданий должны быть равномерно распределены по шкале логитов – это означает, что разработанный набор тестовых заданий позволяет эффективно оценивать уровень знаний.

В данном наборе тестовых заданий уровни знаний испытуемых распределены по нормальному закону распределения. Уровни трудности заданий распределены по шкале логитов примерно равномерно.

6. Диапазон варьирования заданий должен иметь значение больше трех логитов. Если диапазон не достигает данного значения, то необходимо добавлять в тест задания всех уровней сложности. Согласно рис. 3, диапазон значений данного набора тестовых заданий находится на отрезке от –1,8 до +1,2 по шкале логитов, следовательно, диапазон варьирования заданий составляет 3 логита. Данное значение удовлетворяет выдвинутому критерию, данный набор тестовых заданий не следует усложнять.

Программа RUMM позволяет расшифровать значения верхних столбцов гистограммы. Для исследуемого набора тестовых заданий эти данные представлены на рис. 4.

Из рис. 4 видно, что на такие тестовые задания, как 8, 14, 27, 4 и 6 ответило большее число респондентов, чем на вопросы № 23, 30 и 29.

LOCATION	PERSONS	ITEMS [uncentralised thresholds]				
3,0	XX					
	XX					
2,0	XX					
	X					
	XX					
	XXX					
1,0	XXXX	I0008.1	I0014.1	I0027.1	I0004.1	I0006.1
	XX	I0018.1	I0016.1	I0010.1	I0013.1	I0025.1
		I0022.1	I0028.1			
		I0002.1	I0026.1			
		I0011.1	I0017.1			
0,0						
		I0019.1	I0007.1	I0021.1		
		I0020.1				
		I0024.1	I0012.1	I0005.1		
-1,0						
		I0003.1	I0001.1	I0015.1	I0009.1	
		I0023.1	I0030.1	I0029.1		
-2,0						

X = 1 Persons

Рис. 4. Значения верхних столбцов гистограммы

Следовательно, первая группа заданий обладает меньшей трудностью, нежели вторая. Данные рис. 4 подтверждают выводы, сделанные по рис. 3.

7. Также программа RUMM позволяет оценить уровень подготовленности испытуемых. Оказалось, что среди испытуемых легко выделить тех, кто обладает высоким, средним и низким уровнем подготовленности. Эти данные позволяют скорректировать процесс преподавания для конкретных студентов, применив тем самым индивидуальный подход в обучении.

Приведенные критерии являются важным средством повышения эффективности и качества тестов. Учет этих критериев позволяет разрабатывать тесты, которые объективно и точно измеряют уровень подготовленности испытуемых. Для анализируемого набора тестовых заданий было выявлено, что

необходимо переработать следующие тестовые задания: № 11, 18 и 26.

Так как некоторые тестовые задания требуют улучшения формулировки, данный набор тестовых заданий был апробирован еще несколько раз, до тех пор, пока все задания не стали соответствовать выдвинутым критериям. Для этого было проведено еще два этапа тестирования в различных группах студентов СГПИ. В результате были переформулированы тестовые задания № 13, 22 и 24.

После завершения данных этапов работы набор тестовых заданий перешел в разряд педагогических тестов. Результаты показали, что данный педагогический тест является эффективным измерительным инструментом, он может быть использован для измерения латентной переменной «уровень подготовленности испытуемых по учебной

дисциплине “Современные средства оценивания результатов обучения”». Кроме того, программа RUMM позволяет проводить не только оценку уровня подготовленности и сложности задания в отдельности, но и показывает несоответствие уровня подготовленности уровню трудности, если таковое имеется. Этот критерий анализа теста уникален.

Данная методика анализа теста может применяться для измерения латентной переменной «уровень подготовленности испытуемых» по любой учебной дисциплине.

Литература

1. Аванесов, В.С. Форма тестовых заданий: учеб. пособие для учителей школ, лицеев, преподавателей вузов и колледжей / В.С. Аванесов. 2-е изд., перераб. и расшир. М.: Центр тестирования, 2005.
2. Анисимова, Т.С. Измерение латентных переменных в образовании / Т.С. Анисимова. М.: Исследовательский центр проблем качества подготовки специалистов, 2004.
3. Качество высшего образования. Тезаурус / В.С. Аванесов [и др.]. Славянск-на-Кубани: Изд. центр СГПИ, 2005.
4. <http://ru.science.wikia.com>.
5. <http://www.ht.bitnet.ru>.
6. <http://www.statsoft.ru>.