# АНАЛИЗ РЕЗУЛЬТАТОВ ТЕСТИРОВАНИЯ В ПРОЦЕССЕ RASCH MEASUREMENT

Владимир Ким
Уссурийский государственный педагогический институт
vskim@mail.ru

Опубликовано в ж. Педагогические измерения №4, 2005 г.

Аннотация

В статье рассмотрены общие вопросы анализа результатов тестирования на основе одномерной модели Раша[1]. Приведены эмпирические данные по тесту «Механика» для средних общеобразовательных учреждений. Показано, что экспериментальные данные удовлетворительно согласуются с моделью Item Response Theory (IRT).

Ключевые слова: тест, теория Rasch Measurement, педагогические измерения.

Педагогический тест как средство измерения учебных достижений может дать достоверный результат только в случае его корректного применения. Корректность применения теста – это многоаспектное понятие, включающее в себя вопросы конструирования и дизайна теста, вопросы разработки и применения тестов и, разумеется, интерпретации результатов тестирования. В данной работе основное внимание уделено вопросам корректности интерпретации результатов педагогического тестирования, проводимого на основе модели Г.Раша. Анализ результатов обычно проводится на основе классической теории тестов или на основе Item Response Theory.

После выполнения работ по созданию теста и сбора данных на репрезентативной выборке испытуемых, производится интерпретация

[1] Rasch G. Probabilistic Models for Some Intelligence and Attainment Tests. Copenhagen, 1960, Danish Institute of Educational Research. (Expanded edition, Chicago, 1980, The University of Chicago Press).

результатов. Этот этап принципиально отличается от технологии, принятой, скажем в экспериментальной физике. Там экспериментальные данные пытаются описывать с помощью той или иной теории. Если теоретическая зависимость между исследуемыми величинами не соответствует наблюдаемой в эксперименте, то делается вывод, что теория недостаточно развита и требует дальнейшей разработки. В теории педагогических измерений может применяться иной подход. Если в физике законы природы не зависят от исследователя, то тесты в немалой степени зависят от его воли. Это принципиально важный момент.

IRT в настоящий момент является общепризнанной теорией. В качестве латентных параметров модели выступают как характеристики тестируемых, так и самого теста. Ю.Нейман и В.Хлебников делают вывод, что «…уникальность моделей семейства Г.Раша состоит в том, что они задают определенный механизм преобразования формальных наблюдений за исходом событий в объективные измерения на метрической шкале латентных стимулов этих событий». Это очень важно, так как недостаточно глубокое осознание этого факта, может приводить к тому, что положения педагогических измерений могут критически восприниматься специалистами в области точных наук[2]. Таким образом, несоответствие эмпирических данных модели Раша означает, что, например, имеются неточности в формулировке заданий, были нарушения в процедуре тестирования и т.д.

Как отмечает В.С.Аванесов, в литературе можно встретить немало критики по поводу неприменимости модели Раша к множеству «тестов», и поэтому ведется поиск других моделей, более адекватных полученным результатам. Но здесь есть один очень важный вопрос. В теории Г.Раша никогда не ставилась задача адекватного описания данных. Напротив, это пример другой философии измерения - model based measurement, где утверждается противоположное – не модель должна соответствовать эмпирическим данным, а данные должны соответствовать модели. Об этом можно спорить, но в соответствии с

философией Rasch шкалу (педагогический тест) образуют только те задания, которые отвечают данной модели измерения. Все остальные в тест не включаются[3].

Итак, при анализе результатов тестирования, необходимо проверить соответствие эмпирических данных модели Раша. Согласно Ф. Бейкеру[4] для этого всех N тестируемых, выполняющих К - заданий теста распределяют по шкале θ (ability) по своим диапазонам уровня подготовленности. Испытуемые делятся на J групп вдоль шкалы θ так, чтобы все тестируемые внутри данной группы имели одинаковый уровень подготовленности $\theta_j$. Всего внутри группы с номером j окажутся $m_j$ тестируемых, где j принимает значения из интервала j = 1,2,3,…,J.

В пределах каждой группы $r_j$ тестируемых отвечают правильно на данное задание теста. Таким образом, для уровня подготовленности (уровня знаний) равного $\theta_j$ вероятность правильного ответа на данное задание равна

$$p(\theta_j) = \frac{r_j}{m_j}$$

Величина $p(\theta_j)$ является эмпирическим значением вероятности правильного ответа на данное задание. На рисунке 1 показаны данные из работы F. Baker (2001).

На следующем этапе проверяется, насколько хорошо эмпирические данные описываются IRT-моделью. Результат сравнения показан на рисунке 2.

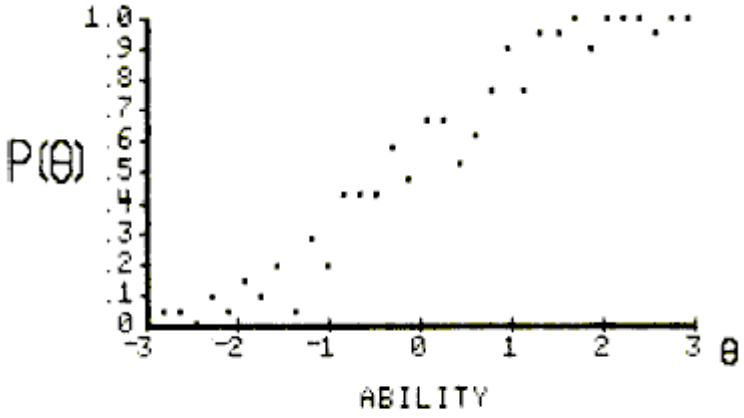
Из рисунка 2 видно, что наблюдается хорошее согласие эмпирических данных с IRT. В целом задача разработчика тестов состоит в том, чтобы

---

[3]Аванесов В.С. Применение тестовых форм в Rasch Measurement. См. статью в номере ПИ №4, 2005г.

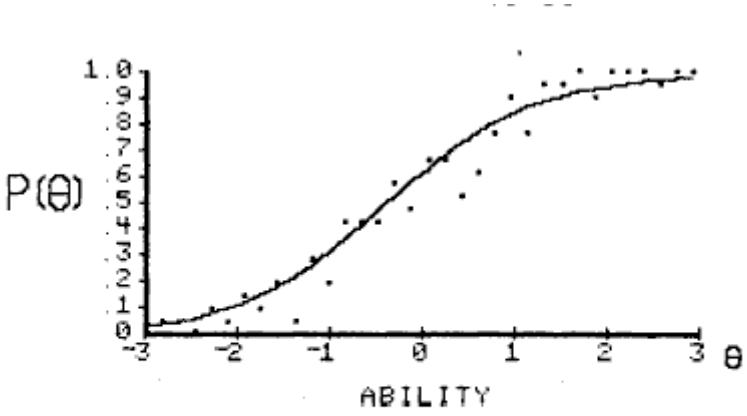
[4] Baker F.B. The Basics of Item Response Theory. –ERIC, 2001. -172 p

разработать такие тестовые задания и так осуществить процедуру тестирования, чтобы получить результаты, аналогичные тем, что показаны на рисунке 2.




Observed proportion of correct response as a function of ability

Рис. 1




Item characteristic curve fitted to observed proportions of correct response

Рис. 2

Основные результаты работы.

В данной работе приведены результаты тестирования учащихся средних общеобразовательных учреждений, по теме «Механика», учебной дисциплины «Физика». Тест содержал 30 заданий с выбором одного правильного ответа[5]. Всего было протестировано 60 испытуемых, по итогам была составлена матрица размером 60 х 30. После упорядочения матрицы по строкам и столбцам по стандартной процедуре были рассчитаны логистические кривые по одномерной модели Раша.

Для модели Раша вероятность успеха в j-м задании равна

$$P_j = \frac{1}{1+e^{-d(\theta-\beta_j)}}$$

где $d$ – фактор шкалирования, равный 1,702.

На рисунке 3 приведены результаты расчетов для всех 30 заданий теста.

Экспериментальные значения $P_j$, полученные по методике[4], приведены на рисунках 4-7. Экспериментальные данные показаны выборочно для четырех заданий различного уровня трудности – 3, 8, 20, 30 задания.

[5] Аванесов В.С. Форма тестовых заданий. М. Центр тестирования, 2005.

Рис.4

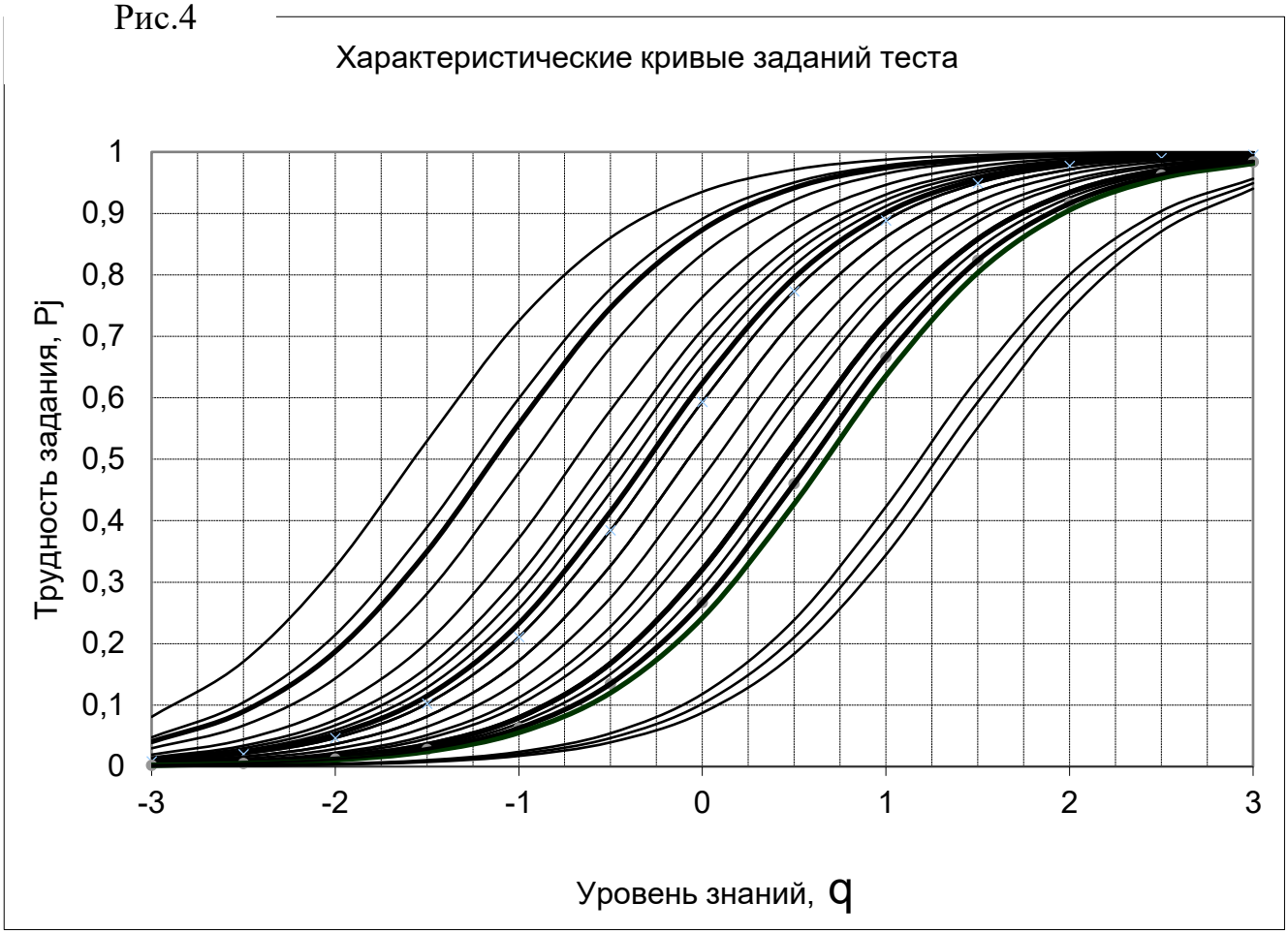
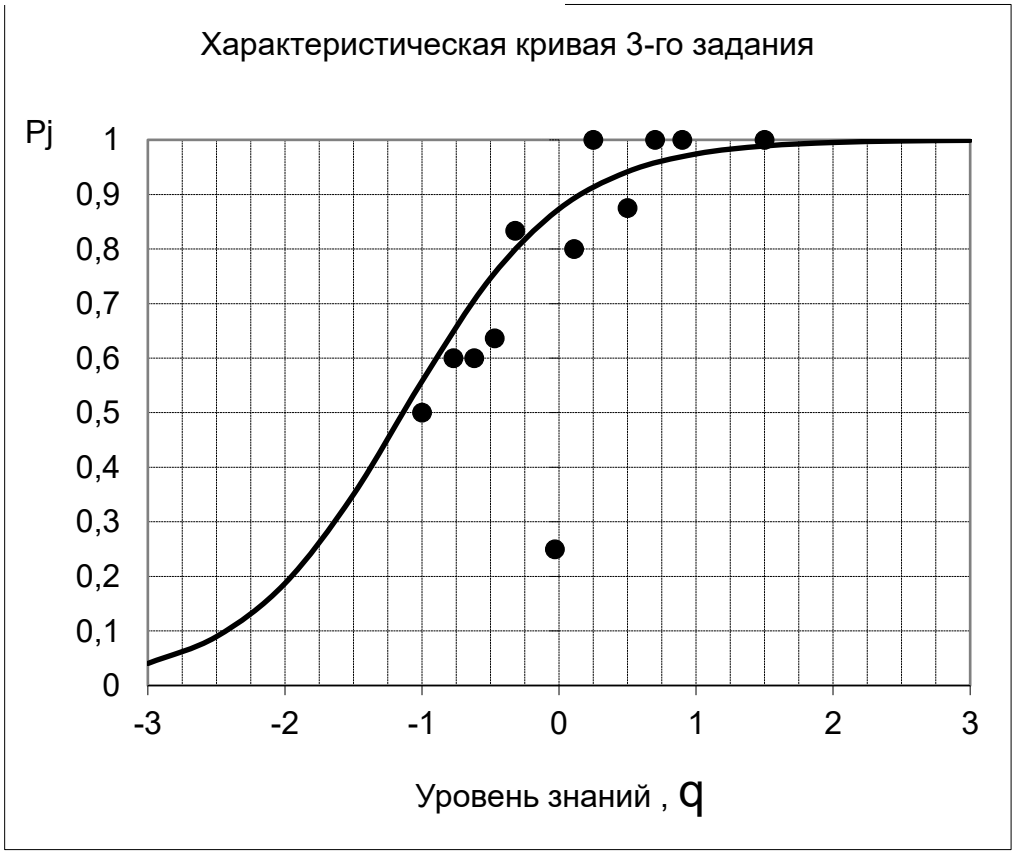

Характеристические кривые заданий теста
Трудность задания, Pj
1
0,9
0,8
0,7
0,6
0,5
0,4
0,3
0,2
0,1
0
-3
-2
-1
0
1
2
3
Уровень знаний, q

Рис. 3


Характеристическая кривая 3-го задания
Pj
1
0,9
0,8
0,7
0,6
0,5
0,4
0,3
0,2
0,1
0
-3
-2
-1
0
1
2
3
Уровень знаний , q

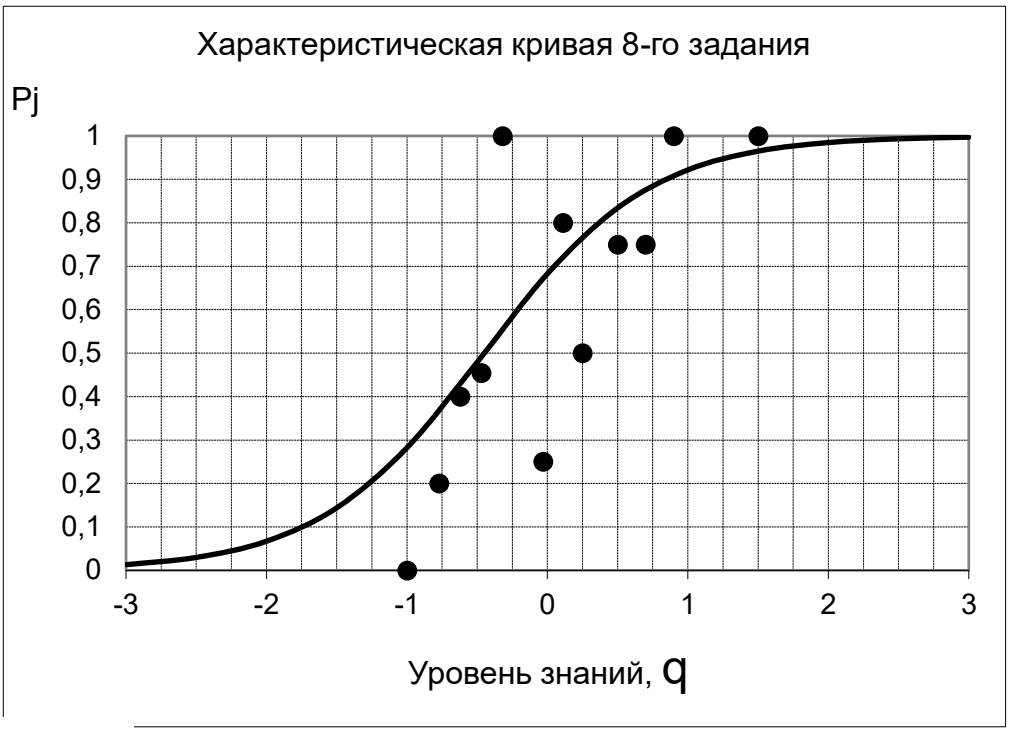

Характеристическая кривая 8-го задания
Pj
1
0,9
0,8
0,7
0,6
0,5
0,4
0,3
0,2
0,1
0
-3
-2
-1
0
1
2
3
Уровень знаний, q

Рис.5

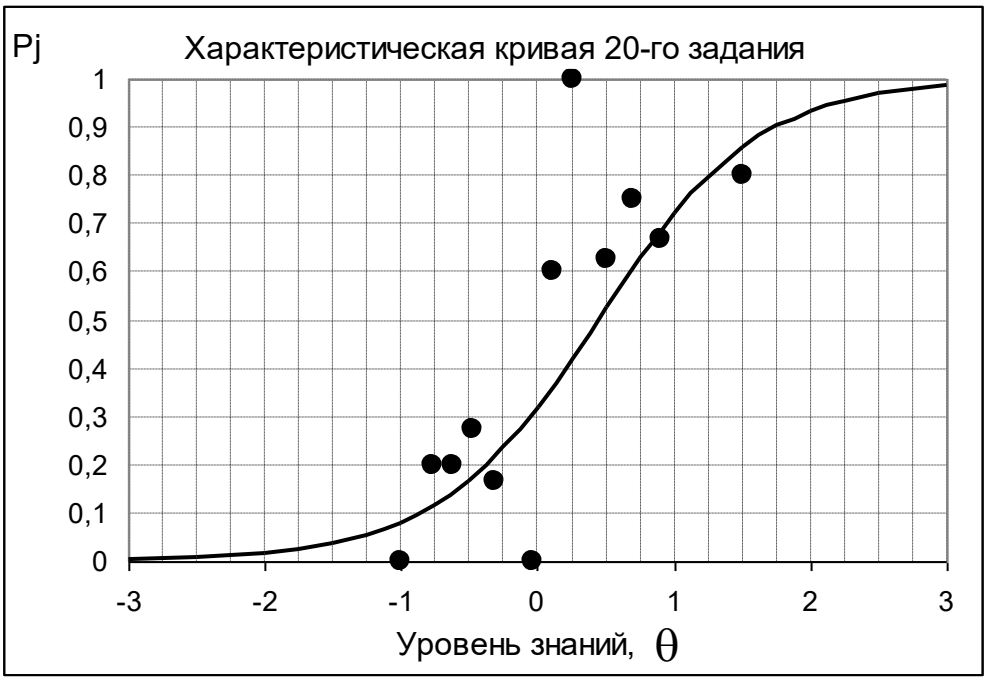
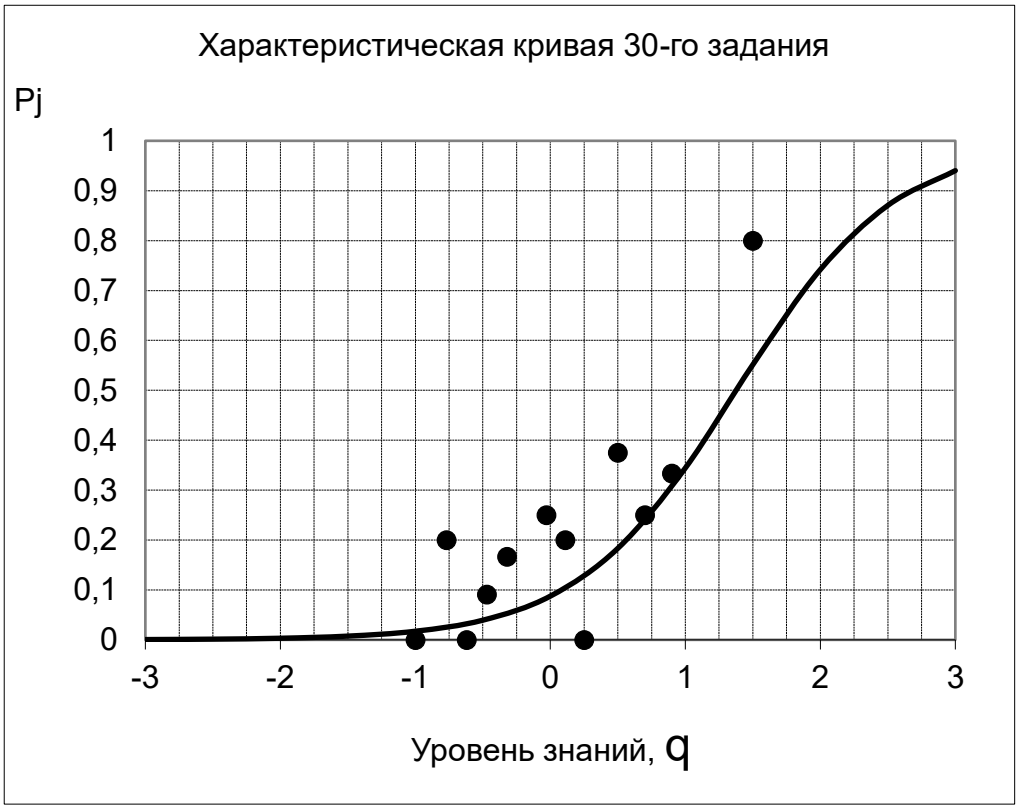

Pj
Характеристическая кривая 20-го задания
1
0,9
0,8
0,7
0,6
0,5
0,4
0,3
0,2
0,1
0
-3
-2
-1
0
1
2
3
Уровень знаний, θ

Рис.6

Рис.7




Обсуждение полученных результатов.

По результатам тестирования сразу можно получить матрицу, анализируя которую можно избавить тест от некоторых неподходящих заданий[6]. Дальнейшие расчеты возможны в трёх вариантах:

Однопараметрическая (одномерная) логистическая модель, или модель Раша[7];

Двухпараметрическая логистическая модель (2PL) Бирнбаума;

Трехпараметрическая логистическая модель (3PL) Бирнбаума.

Как известно, модели 2PL и 3PL предлагались для лучшего согласования теории с наблюдаемыми эмпирическими данными. Если считать, что согласования следует добиваться не видоизменением теории, а получением других эмпирических данных, то можно остановиться на модели Раша. Иными

---

[6] . Аванесов В.С. Основы научной организации педагогического контроля в высшей школе. М., 1989. -167 с.

[7] Rasch G. Probabilistic Models for Some Intelligence and Attainment Tests. Copenhagen, 1960, Danish Institute of Educational Research. (Expanded edition, Chicago, 1980, The University of Chicago Press).

словами, если экспериментальные данные не соответствуют модели Раша, то необходимо переработать тестовые задания и повторно провести эксперимент, добиваясь лучшего согласия с теорией, как указывалось выше.

Следуя такой парадигме, в данной работе все построения проводились по модели Г.Раша.

Из рисунка 3 видно, что задания теста по шкале уровня знаний θ перекрывают диапазон примерно от -3,5 до +3,5 логитов. Графики показаны последовательно слева- направо от 1-го (самого легкого) до 30-го - самого трудного задания. Характеристические кривые некоторых заданий, а именно 3 и 4; 10 и 11; 13, 14 и 15; 19, 20 и 21; 23 и 24; 25, 26 и 27 перекрываются. В связи с этим 4, 11, 14, 15, 19, 21, 24, 26, 27 задания могут быть удалены из теста без ущерба его измерительным свойствам.

На семействе логистических кривых тестовых заданий отчетливо наблюдается явная недостаточность отдельных заданий. Наличие «провалов» в монотонной последовательности характеристических кривых указывает на необходимость дополнительной оптимизации теста путем добавления новых тестовых заданий или переработки имеющихся. Переработкой тестовых заданий необходимо добиться появления добавочных характеристических кривых в интервале от -1,5 до - 0,5 и от +0,7 до +1,2 логита (на уровне $P_j = 0,5$).

Экспериментальные данные для $P_j$ имеют примерно одинаковое согласие с моделью Раша, которое можно считать удовлетворительным. Приведенные на рисунках 4-7 характеристические кривые некоторых заданий иллюстрируют это. При анализе вся совокупность тестируемых разбивалась на 12 групп (J=12).

Экспериментальные точки для характеристической кривой 3-го задания группируются в области от -1 до +1 логита для $P_j$ - от 0,5 до 1,0. Это относительно легкое задание и экспериментальные точки приблизительно соответствуют верхнему участку характеристической кривой. Задания 8 и 20 находятся примерно в средней области тестовых заданий (рисунок 3) и соответствуют заданиям средней сложности. Экспериментальные точки в этом случае группируются вблизи линейной области характеристических кривых $P_8$ и $P_{20}$

Задание №30 самое трудное и экспериментальные точки в основном сосредоточены вблизи нижнего загиба характеристической кривой $P_{30}$.

Для проверки гипотезы $H_o$ на соответствие полученных эмпирических данных одномерной модели IRT для всех заданий теста проводилось вычисление критерия $\chi^2$ согласно[8]

$$\chi^2 = \sum_{j=1}^{J} m_j \frac{(p(\theta_j) - P(\theta_j))^2}{P(\theta_j)Q(\theta_j)}$$

Расчетное значение критерия $\chi^2$ оказалось в пределах от 7 до 15 для различных заданий теста.

Таким образом, несмотря на довольно заметную вариацию данных, что вероятнее всего обусловлено недостаточной репрезентативностью выборки (60 испытуемых), можно констатировать удовлетворительное согласие экспериментальных результатов с Rasch Measurement.

[8] Baker F.B. The Basics of Item Response Theory. –ERIC, 2001. -172 p