МЕТРИЧЕСКАЯ СИСТЕМА ГЕОРГА РАША – RASCH MEASUREMENT (RM).

СТАТЬЯ ВТОРАЯ

Вадим Аванесов
testolog@mail.ru

В статье продолжается[1] исследование ключевых вопросов развития и применения метрической системы Г.Раша (RM). Рассмотрены вопросы сущности и содержания RM, метрические свойства тестов и заданий, вопросы проектирования качественных тестов на основе теории и технологии RM. В начале статьи даётся краткий обзор публикаций журнала ПИ по метрической системе Г. Раша.

Ключевые слова: Rasch Measurement (RM), основные понятия, критерии качества тестов, отвечающих требованиям RM.

If you want to do measurement,
you have to do Rasch.
R.J. Mead[2]

Проблемная ситуация и проблема исследования

Актуальность исследования. К настоящему времени практика применения RM насчитывает тысячи, в разных странах мира. Ещё бо-

1 Первые две статьи: *Аванесов В.С.* Применение тестовых форм в Rasch Measurement //Педагогические измерения. № 4. 2005. С. 3–20; *Аванесов В.С.* Метрическая система Георга Раша // Педагогические измерения.

2 *Mead R.J.* A Rasch primer: the measurement theory of Georg Rasch. Psychometrics services research memorandum 2008–001. Maple Grove, MN: Data Recognition Corporation. 2008.

лее обширен мир теоретических исследований и публикаций по данному направлению науки, что является свидетельством актуальности RM. Но вот парадокс: в российских учебниках педагогики Георг Раш и его метрическая система даже не упоминаются. Точно такая же ситуация была раньше и с тестами. Вряд ли случайно, что RM, будучи одной из теорий и технологий разработки качественных тестов, повторяет в России трудную судьбу тестов. Причины такого положения лежат, скорее всего, в плоскости управления образованием и наукой в стране. Этот вопрос, очевидно, заслуживает отдельного анализа.

Хотя за прошедшие сорок лет ситуацию с тестами удалось чуть-чуть поправить, появилось, однако, новое препятствие. В России появился т.н. единый государственный экзамен (ЕГЭ), а вслед за этим экзаменом вновь возникла и усилилась критика «тестов» ЕГЭ. К счастью, вся эта критика, в сущности, относится не к тестам, а к т.н. «контрольно-измерительным материалам» (КИМам) ЕГЭ, много лет выдававшихся официозом за тесты, и даже за педагогические измерения. Но, как было недавно показано[3], никаких тестов и измерительных материалов в этих КИМах нет. А потому пришло время их называть правильнее — контрольными материалами ЕГЭ, без добавления слов «измерительные» или «тесты». Пора образумиться, там в действительности, нет ни тестов, ни измерений.

Литература по RM на русском языке тоже находится в неопределённой ситуации. В ней есть немало текстов, которые трудно читать и понимать даже заметно образованному человеку, из-за неудобоваримой лексики. Такое бывает нередко, когда лица одной профессии пишут своим языком для представителей другой профессии. Особенно часто это случается с математиками, которые активно подключаются к исследованиям не только в педагогике, но практически и по всем научным дисциплинам. Как известно, в наше время без математики не обходится ни одна настоящая наука.

Тем более это относится к педагогическим измерениям, где всё пронизано математикой, программами и вычислениями. На этом языке RM определяется (в приблизительном переводе) как метод получения объективных, фундаментальных, аддитивных измерений, пригодность и качество которых подтверждается статистическими методами определения значений стандартных ошибок результатов эмпирических данных[4]. Нормальный перевод такого рода определений невозможен без предварительного

3

Аванесов В.С. Проблема развития педагогических измерений //Педагогические измерения № 2. 2011. С. 3–35.

4

Rasch analysis is a method for obtaining objective, fundamental, additive measures (qualified by standard errors and quality-control fit statistics) from stochastic observations of ordered category responses. Linacre J.M. Rasch analysis and Winsteps. http://www.winsteps.com/winman/.

истолкования всех терминов. Иначе говоря, этот тот случай, когда нужен не перевод, а развёрнутая интерпретация.

Есть публикации на русском языке, по поводу которых могут возникать вопросы. В них RM предстаёт как «теория измерения латентных переменных» или как «теория измерения латентных переменных на основе модели Раша»[5]. Из чего видна ключевая роль словосочетания «латентная переменная». Но RM — это теория измерения определённых латентных свойств личности, заданий и иных объектов, а не просто неименованных абстрактных латентных переменных. В педагогике латентные переменные величины возникают вследствие концептуализации явлений, операционализации понятий, применения метрических операций и математических моделей к изучаемому свойству.

Например, вероятностная модель Г. Раша позволяет прогнозировать результат испытуемого под номером i на то или иное задание, под номером j, если известны так называемые значения параметров уровня подготовленности испытуемого и уровня трудности задания. Можно использовать также двух- и трёхпараметрические модели определения вероятности правильных ответов испытуемых на задания переменной трудности. Что делается в рамках другой, математической теории педагогических измерений (Item response Theory, IRT)[6]. При этом исследователи также опираются на понятия латентных переменных величин, как и в RM. Что даёт основания многим авторам считать, что RM и IRT — это одна и та же теория. Но это не так. Это разные теории, и результаты их применения различны, хотя проблема одна и та же — педагогические измерения.

Среди недостатков, имеющихся в российской и зарубежной литературе, можно отметить часто встречающиеся попытки противопоставить классическую (статистическую) теорию тестов теориям RM и IRT. При этом последние теории нередко называют современными, а классическую теорию пытаются выставить в качестве устаревшей.

Пример с логической наукой позволит лучше понять ошибочность подобных текстов. В логике много веков торжествовала дедуктивная логика (силлогистика). Затем возникла средневековая схоластическая логика, модальная логика и усовершенствованная логика Аристотеля. В XVI–XIX веках пришло время индуктивной логики, вслед за которой появилась математическая логика, неклассическая, конструктивная и др. логики. И никому не пришло в голову объявить самую

5 Неприемлемые определения RM, опубликованные за пределами публикаций журнала ПИ, в других источниках, в этой статье не рассматриваются.

6 *Аванесов В.С.* Понятия и методы математической теории педагогических измерений (Item Response Theory). Статья третья. // Педагогические измерения. № 4. 2009. С. 3–28.

первую логическую теорию — силлогистику устаревшей и ненужной.

Научный прогресс, как правило, не отменяет научные теории, а обнаруживает лишь ограничения на их применение в новых условиях, недостаточность прежних теорий для объяснения усложняющихся познавательных объектов, а потому ранее сформулированные теории часто дополняют (но не отменяют полностью) другими теориями.

Применительно к педагогическим измерениям можно сказать, что многие положения классической (статистической) теории тестов — по надёжности и валидности тестовых результатов, параллельности заданий теста, по форме и содержанию тестовых заданий, по статистическим характеристикам теста — широко используются и сейчас. Без них просто невозможно создать качественный тест. Вот почему суждения об устаревшей классической теории тестов можно отнести к числу неконструктивных.

При анализе проблемной ситуации уместно писать не только о том, что написано, но и о том, чего нет, но должно быть. Нет учебника по RM на русском языке, нет качественных переводов зарубежной научной литературы по данной проблеме. Мало качественных и доступных для педагогов публикаций по методике RM, нет пособий по применению компьютерных программ по RM. Трудно найти авторов, способных хорошим русским языком изложить ключевые вопросы RM.

И всё-таки что-то делается. К числу немногих исключений можно отнести статьи по RM, опубликованные в журнале «Педагогические измерения» (см. обзор далее).

Проблему настоящей статьи образуют исторические предпосылки появления RM, исследование сущности такого измерения, метрические свойства тестов и заданий, вопросы проектирования тестов на основе теории и технологии RM, а также развития педагогического языка RM. Вопросы языка в этой статье затрагиваются не специально, а по ходу изложения текста статьи.

Некоторые из перечисленных вопросов проблемы исследуются сейчас в России, но системной и эффективной работу в данном направлении признать невозможно. Нет, соответственно, и весомых достижений в этой сфере. К исключениям можно отнести лишь публикации статей по проблеме RM в журнале «Педагогические измерения[7]» и эпизодические курсы для узкого круга слушателей.

В двух предыдущих работах автора этой статьи[8,9] использование английских слов Rasch Measurement (RM) в названиях

7 *Аванесов В.С.* Обзор публикаций журнала «Педагогические измерения» по Rash Measurement обзор публикаций //Педагогические измерения. № 3. 2011.

8 *Аванесов В.С.* Применение тестовых форм в Rasch Measurement // Педагогические измерения. № 4. 2005. С. 3–20.

9 *Аванесов В.С.* Метрическая система Георга Раша (Rasch Measurement, RM) // Педагогические измерения. № 2. 2010. С. 57–80.

статей объяснялось отсутствием подходящих терминов на русском языке. В 2010 году RM получило новое название — метрическая система Георга Раша[10].

Обзор публикаций журнала «Педагогические измерения» по Rasch Measurement

Примечательно, что к вопросам применения и развития метрической системы RM журнал «Педагогические измерения» обратился с самого первого номера, вышедшего в свет в конце 2004 года. С той поры в данном направлении российскими авторами был пройден путь от отдельного опыта использования компьютерной программы RUMM-2010 до уровня теоретических исследований и развития вопросов методологии RM. И работа по развитию теории и методологии RM продолжается журналом.

Сейчас можно даже сказать, что для авторов российского научно-методического журнала «Педагогические измерения» RM стало ведущей исследовательской проблемой. Гораздо раньше RM получил возрастающую поддержку ведущего зарубежного специализированного научного журнала «Психометрика»[11].

Как отмечали зарубежные авторы[12] нашего журнала, RM стало популярным во всём мире и в различных сферах. Эта популярность касается не только сфер педагогики и психологии, но и социологии, медицины. В статье этих авторов были перечислены преимущества RM, по сравнению с классической (статистической) теорией тестов и с математической теорией педагогических измерений (Item response Theory, IRT).

Все публикации журнала «Педагогические измерения» по RM полезно разделить на три группы, в зависимости от привычного в науке деления на уровни исследований. К первому уровню можно отнести опыты применения технологии RM для решения практических задач и проведения эмпирических исследований. Ко второму уровню можно отнести статьи, в которых ставятся и решаются вопросы теории и технологии RM. К третьему уровню можно отнести статьи по методологии RM. Очевидно, что в науке нужны и важны исследования каждого уровня.

1. *Применение программ.* Первой, по времени, была опубликована статья, в которой специальная компьютерная программа RUMM-2010, технологически реализующая возможности RM, была использована для получения информации о сравнительном уровне развития

10 Там же.

11 The Rasch model is a special case of additive conjoint measurement, a form of fundamental measurement... A fit of the Rasch model implies that the cancellation axiom will be satisfied... It then follows that items and persons are measured on an interval scale with a common unit.» *Brogden H.E.* 1977. The Rasch model, the law of comparative judgement and additive conjoint measurement. Psychometrika. 42, p. 633.

12 *Smith Everett V.Jr., Karen M. Conrad, Karen Chang, Jo Piazza.* Введение в Rasch Measurement. Перевод с англ. //Педагогические измерения. № 1. 2006. С. 65–81.

образования в областях и регионах РФ. То, что Москва и Санкт-Петербург имеют наиболее высокий уровень развития сферы образования, являлся ожидаемым результатом. Несколько неожиданным и необъяснённым оказался третий ранг Чукотского автономного округа[13].

Этот и другие результаты применения программы по RM требовали интерпретации полученного вывода, в смысле зависимости результатов от количества и качества используемых показателей. Иначе говоря, здесь напрашивалось обсуждение вопроса соотношения фактов и возможных артефактов, вызванных несовершенством используемых сейчас счётных показателей. Это вопрос валидности используемой системы показателей. Но данный аспект редко когда затрагивается. Между тем вопрос соотношения получаемых данных и их педагогически обоснованной интерпретации — один из главных и трудных в педагогических измерениях. Потому что все содержательные выводы зависят от используемых показателей, от меры их пригодности для решаемой проблемы[14].

Попытка разработки социологической анкеты, с применением технологии RM, для исследования вредных привычек у молодёжи была предпринята в исследовании А.А. Маслака[15].

Затем последовала публикация по вопросам применения компьютерной программы RUMM-2010, для анализа качества заданий с выбором одного правильного ответа[16]. Было показано, что эта программа, основанная на технологии RM, позволяет провести более углублённый анализ метрических свойств теста и каждого задания. Анализ качества заданий в той статье проводился по следующим, как там было написано, аспектам RM:

- выявление и исключение из экстремальных заданий;
- совместимость заданий;
- соответствие уровня трудности разрабатываемого теста уровню подготовленности студентов;
- равномерность распределения заданий по трудности;
- диапазон варьирования трудности тестовых заданий;
- соответствие тестового задания модели измерения;
- оценка качества дистракторов.

По сути, здесь перечислены основные задачи, решаемые разработчиком теста посредством компьютерной программы.

Публикация статей по применению компьютерной программы RUMM-2010 была продолжена работами В.С. Кима. В первой его статье была сделана удачная попытка анализа результатов тестирования[17]. Во второй его статье эта же программа целенаправленно использовалась для

13
Анисимова, Т.С., Маслак, А.А., Седых С.И. Измерение уровня развития сферы образования в регионах России. Педагогические измерения. № 1. 2004. С. 97–128.

14
Аванесов В.С. Проблема качества педагогических измерений //Педагогические измерения. № 2. 2004. С. 3–27.

15
Маслак А.А. Оценка статистической взаимосвязи между склонностью старшеклассников к курению и условиями их жизни и учёбы // Педагогические измерения, № 2. 2005. С. 101–119.

16
Анисимова, Т.С., Маслак, А.А. Осипов С.А. Анализ качества заданий с выбором одного правильного ответа. Педагогические измерения. № 3. 2005.

17
Ким В.С. Анализ результатов тестирования в Rasch Measurement. Педагогические измерения. № 4. 2005. С. 39–45.

проверки качества заданий теста[18]. В третьей статье, методической направленности, был представлен алгоритм вычислений, используемый в программе RUMM-2020[19], что представляет особую методическую ценность. В четвёртой статье была сделана удачная попытка применить эту программу для углублённого научного анализа качества метода социологического исследования[20].

Алгоритм обработки эмпирических данных посредством программы RUMM-2010 изложила Г.И. Смирнова, сделав это доступным для педагогов языком[21]. Она же выявила различия в двух версиях программы RUMM[22].

В журнале «Педагогические измерения» была сделана попытка ввести в научный оборот российских исследователей ещё одну компьютерную программу — Winsteps. На эту тему написали статьи Г.И. Смирнова, с А. Смирновым[23], а также С.И. Янченко[24].

Полученные выше научные результаты применения программ RUMM (в модификациях 2010 и 2020) и WINSTEPS (Ministep-бесплатной версией) для решения проблем развития практики педагогических измерений в России можно считать полезными. Именно этот успешный опыт возбудил в стране некоторый интерес к вопросам следующего уровня.

2. Теория и технология RM. Начало теоретическому процессу разработки вопросов RM в России положила обзорная статья О.В. Михеева. Хотя RM там было уделено очень небольшое внимание (в общем ряду с другими метрическими системами), это была полезная работа для становления общей культуры педагогических измерений в стране[25].

Пожалуй, первым опубликованным в журнале чисто теоретическим исследованием оказалась статья С.И. Янченко. Она сделала интересную попытку соединить возможности математической модели Г. Раша с актуальной задачей изучения уровня и структуры подготовленности испытуемых[26]. Полученные выводы были оригинальны, приведённый материал обладал заметным потенциалом развития. Однако развитие этого направления теории, к сожалению, не последовало.

После этого было опубликовано редкое, для нашего времени, теоретико-экспериментальное исследование точности оценок параметров модели Раша на основе алгоритма PROX. Выбор этого алгоритма не случаен: он самый простой и доступный, особенно для педагогов-нематематиков, желающих практически освоить методику RM. Проведённый авторами имитационный эксперимент позволил получить интересные выводы,

18

Ким В.С. Анализ тестовых заданий в модели G. Rasch// Педагогические измерения. № 1. 2008. С. 49–58.

19

Ким В.С. Обработка результатов тестирования компьютерной программой RUMM-2020 //Педагогические измерения. № 4. 2008. С. 53–69. Эта же статья была напечатана из-за ошибки редакции повторно в № 1, 2009. К счастью, статья оказалась очень полезной и нужной.

20

Ким В.С. Использование компьютерной программы RUMM-2020 в социологических исследованиях// Педагогические измерения. № 2. 2009. С. 61–75.

21

Смирнова Г.И. Алгоритм обработки матриц результатов тестирования с оценкой 0-1-2 и более с помощью программы RUMM-2010 // Педагогические измерения. № 4. 2007. С. 86–90.

совпадающие с теорией и важные для практики измерения латентных переменных величин.

Во-первых, оказалось, что минимальные и максимальные уровни подготовленности[30] оцениваются менее точно, чем уровни подготовленности, расположенные в середине интервала варьирования измеряемой латентной переменной (±0,2–0,4 логита). Во-вторых, точность измерения латентной переменной значимо зависит от распределения измеряемой величины. Чем меньше дисперсия, тем выше точность измерения. Вместе с тем, точность измерения измеряемой латентной переменной величины оказалась независимой от вида её распределения[31].

Г.И. Смирнова в двух номерах журнала сделала попытку создания первого варианта тезауруса RM на русском языке[32]. Она же провела теоретико-эмпирический анализ метрических свойств заданий проектируемого теста[33].

Математические основы RM обстоятельно изложил в теоретической статье О.Г. Деменчёнок[34]. Ему удалось установить преемственность и математически показать научную связь модели Раша со сформулированным ранее Л.Л. Терстоуном и другими классиками требованием независимости (инвариантности) средств измерения от объекта измерения[35].

Интересен вывод О.Г. Деменчёнка относительно спорного соотношения данных модели и наоборот. По мнению J.M. Linacre[36], исправлять нужно не модель Раша, а исходные данные[37]. Подтверждая этот тезис, О.Г. Деменчёнок усиливает его следующим выводом: «Никакая модель не в состоянии корректно устранить все искажения исходных данных. Поэтому актуальность задачи получения пригодных исходных данных высока и не зависит от выбранной математической модели[38]. С этим можно полностью согласиться.

В недавних номерах журнала «Педагогические измерения» проявил себя двумя новыми теоретическими исследованиями Ю.Н. Каргин. В первой его работе была представлена т.н. альтернативная однопараметрическая модель педагогических измерений. Ключевая идея этой его статьи — это возможность измерения уровня подготовленности испытуемых и уровня трудности заданий в шкале более высокого уровня — шкале *отношений*[39].

Во второй работе этот же автор представил новый аналитический метод решения основной задачи педагогических измерений — измерение уровня

22
Смирнова Г.И. Различия в программах RUMM-2010 и RUMM-2020 //Педагогические измерения. № 3. 2007. С. 69–77.

23
Смирнова Г.И., Смирнов А. Начало работы с программой MINISTEP // Педагогические измерения. №3. 2006. С. 106–113.

24
Янченко С.И. Начало работы в WINSTEPS с данными статистического пакета SPSS // Педагогические измерения. №3. 2006. С. 115–118.

25
Михеев О.В. Математические модели педагогического измерения // Педагогические измерения. № 2. 2004. С. 75–88.

26
Янченко С.И. Оценка уровня и структуры знаний испытуемых // Педагогические измерения. № 3. 2005. С. 38–64.

подготовленности испытуемых и уровня трудности тестовых заданий, а также разработал алгоритм решения этой задачи[40], чем внёс свой вклад в развитие технологии RM.

3. Методология RM. В статье В.С. Аванесова был дан развёрнутый анализ целей и задач RM, изложены начала педагогически адекватной терминологии на русском языке, названы причины отставания в России исследований по RM от других стран мира. Сама система измерения Г. Раша была названа *метрической*. В этой системе изначально выделяются два взаимосвязанных объекта измерений – уровни трудности заданий и уровни подготовленности испытуемых, которые участвуют одновременно в процессе измерения посредством одной и той же корректируемой шкалы. Поэтому такое измерение часто называют совместно проводимым (joint или conjoint measurement)[41]. Коррекция делается с учётом средней арифметической и значения дисперсии каждой шкалы. В итоге получается шкала с общей единицей измерения — логит.

Здесь главное — метод трансформации исходных тестовых баллов в шкалу натуральных логарифмов (логитов), после чего, собственно, и появляется измерение. До начала процесса логарифмического преобразования исходные баллы тестирования не рассматриваются как результаты измерения[42]. Это одно из существенных требований к педагогическим измерениям, которое актуализировалось к концу XX–началу XXI века. К сожалению, на это требование качественного измерения российский официоз взирает молча и отстранённо. Не принимает и не отвергает. Как будто это их не касается. Следствием чего мы получили и имеем неметрический и некачественный ЕГЭ.

Rasch Measurement — это не просто однопараметрическая модель более общей, казалось бы, математической теории измерений (Item Response Theory), как это обычно считается, а другая, более высокая культура измерения. В RM вместо привычного поиска математической модели для наилучшего описания данных требовалось, наоборот, чтобы данные соответствовали математической модели процесса тестирования и прогнозирования результатов. Иначе, по мысли Г. Раша, измерение не получится. Не всем это требование нравилось. Но со временем, увидев преимущества, часть западных, главным образом американских авторов притерпелась к этому.

Ранее, в другой статье В.С. Аванесова, была исследована связь между RM и формой тестовых заданий. Широко используемые сейчас задания с

27

Это абстрактная стандартизованная единица измерения любых научно обоснованных признаков, интересующих исследователя. В нашем случае — это уровень подготовленности испытуемых и уровень трудности заданий теста. По-английски произносится как «лоджит», с ударением на первом слоге. В традициях произношения слов с этим корнем на русском языке — использование буквы «г».

28

Анисимова Т.С., Маслак А.А., Осипов С.А., Хмара И.А. Исследование точности оценивания параметров модели Раша на основе алгоритма PROX. Педагогические измерения. № 2. 2005. С. 80–100.

29

Смирнова Г.И. Разработка тезауруса педагогических измерений Г. Раша. Педагогические измерения. № 3. 2005. С. 83–86; *Смирнова Г.И.* Разработка тезауруса педагогических измерений Г. Раша. Педагогические измерения. № 4. 2005. С. 62–64.

выбором одного правильного ответа не подходят, в принципе, для применения в метрической системе Г. Раша, поскольку в них всегда присутствует возможность угадывания правильного ответа теми, кто недостаточно подготовлен. А эта возможность в математической модели Г. Раша не предусмотрена. Все, кто использует технологию RM для заданий с выбором одного правильного ответа, как-то об этом забывают либо игнорируют это обстоятельство. И затем удивляются, когда довольно много заданий оказываются неподходящими для включения в тест. Главная причина неподтверждения метрических свойств заданий с выбором одного правильного ответа — это именно высокая вероятность угадывания. Сам Г. Раш пошёл по пути использования заданий открытой формы, но они недостаточно технологичны.

Вместо тех и других заданий автором данной статьи были предложены задания с выбором нескольких правильных ответов, в которых вероятность угадывания правильного ответа довольно близка к нулю. К тому же они технологичны. Это и создаёт подходящие формальные условия применения заданий с выбором нескольких правильных ответов в метрической системе RM.

Проведённый обзор показывает, что за время, прошедшее с начала выпуска российского научно-методического журнала «Педагогические измерения», были получены некоторые результаты. Они могли бы быть более весомыми, если бы ресурсы государства не тратились на бессмысленные контрольные материалы ЕГЭ, а направлялись бы на развитие культуры педагогических измерений, в том числе и RM. Но этого пока нет. И когда случится отказ федерального органа управления образованием от некачественных контрольных материалов — неведомо никому.

Исторические предпосылки появления RM

Возражение западных классиков против имевшейся тогда практики тестирования сводилось к зависимости получаемых тестовых баллов от субъективного подбора заданий по уровню трудности. Чем легче оказывались, в целом, задания, тем выше оказывались тестовые баллы испытуемых. И наоборот. Подбор трудных и очень трудных заданий приводил к уменьшению общего числа баллов испытуемых.

Ещё в 1926 году E.L.Thorndike выражал недовольство тем, что сложение и иные операции с тестовыми баллами не обоснованы с

30 *Смирнова Г.И.* Анализ качества заданий педагогического теста по учебной дисциплине «Математика и информатика» // Педагогические измерения. № 4. 2006. С. 86–100.

31 *Деменчёнок О.Г.* Математические основы Rasch Measurement // Педагогические измерения. № 1. 2010.

32 *Thurstone L.L.* Attitudes can be measured// American Journal of Sociology. Vol. 33. January. 1928. 529–544 pp.

33 Читается по-русски «Линека», с ударением на первом слоге.

34 *Linacre J.M.* The Rasch Model cannot be «Disproved»! // Rasch Measurement Transactions. 1996. 10: 3. p. 512–514.

35 *Деменчёнок О.Г.* Математические основы Rasch Measurement // Педагогические измерения. № 1. 2010. С. 3–21.

точки зрения требования теории измерений[43]. Иначе говоря, счёт — это не измерение[44]. Нетестовые методы, основанные исключительно на подсчёте баллов, метрических свойств не имеют[45].

От подбора заданий менялись и статистические распределения тестовых результатов. Они становились асимметричными, в левую или правую сторону, в зависимости от того или иного подбора заданий теста. Соответственно менялись и качественные характеристики теста.

Самый эффективный метод трансформации счётных данных в измерения открыл именно Г. Раш. С этой точки зрения, одним из самых коротких и точных определений RM — это метод трансформации счётных тестовых баллов в измерения.

Одним из источников поисков Г. Раша была научная деятельность Л.Л. Терстоуна, который по проблеме психологических измерений написал более двух десятков статей в двадцатые и тридцатые годы XX века. Л.Л. Терстоуну мы обязаны чёткой формулировкой требований к психолого-педагогическим измерениям:

- Шкала измерений должна быть линейной, что позволяет использование арифметических операций:
- Параметры трудности заданий не должны зависеть от параметров уровня подготовленности испытуемых, и наоборот, параметры испытуемых не должны зависеть от параметров заданий.
- Те и другие параметры должны быть независимы от выборок испытуемых и от выборок заданий.
- Неполнота данных не должна становиться препятствием для шкалирования.
- Методы шкалирования должны быть достаточно простыми.

Вторым источником явилась публикация D.A. Walker[46] и работы по шкалограммному анализу Л.Л. Гутмана, основанные на упорядочении заданий теста по возрастающей трудности. Из такого порядка вытекало требование: если испытуемый с правильно сформированной структурой знаний успешно отвечает на задание под номером j, то он должен был бы правильно отвечать и на все предыдущие, сравнительно лёгкие задания теста[47].

Проблему зависимости баллов испытуемых от подбора заданий по уровню трудности пытались решить многие выдающиеся учёные, но только Г. Рашу удалось первым создать эффективный метод преодоления отмеченной зависимости. Новаторский аспект подхода Г. Раша к психолого-педагогическим измерениям был замечен не сразу и не многими, а лишь самыми

36 *Каргин Ю.Н.* Построение альтернативной модели педагогических измерений по системе Г. Раша //Педагогические измерения. 2010. № 4. С. 62–71.

37 *Каргин Ю.Н.* Аналитический метод решения основной задачи педагогических измерений // Педагогические измерения. 2011. № 2. С. 54–76.

38 *Аванесов В.С.* Метрическая система Георга РАША — Rasch Measurement (RM). Педагогические измерения. № 2. 2010. С. 57–80.

39 См. подробнее на эту тему: *Аванесов В.С.* Являются ли КИМы ЕГЭ методом педагогических измерений? Педагогические измерения. № 1. 2009. С. 3–26.

40 *Thorndike E. L. et al.* The Measurement of Intelligence. New York: Columbia University, Teachers College. 1926.

проницательными исследователями. В числе первых оказалась J. Loevinger. Фрагмент текста из её работы заслуживает цитирования[48].

Признание его метрических идей длилось долго, более двадцати лет. Примерно столько же лет ушло на разработку более сотни методов, компьютерных программ и подпрограмм для автоматизации вычислительного процесса обоснования качества результатов RM. И хотя многое уже сделано, этот процесс продолжается.

Вторая проблема, которую Г. Раш успешно решил, — это трансформация счётных данных (исходных тестовых баллов) в педагогические измерения. Этот успех позволил некоторым авторам определить RM как метод трансформации счётных тестовых баллов в педагогические измерения. Это было, конечно, одностороннее определение, оно высвечивало только вторую сторону достижения Г. Раша.

Сущность RM

Rasch Measurement — это методология, теория и технология, а также методика измерения уровня трудности тестовых заданий и уровня подготовленности испытуемых. Итогом применения RM становятся качественные задания и тесты, обладающие метрическими свойствами, и педагогические измерения интересующего свойства испытуемых.

В нашем определении RM говорится о теории, технологии и о системе методов RM. В логике данного определения формальные модели процесса тестирования — лишь их часть. В RM математические модели используются для прогнозирования вероятности правильного или неправильного ответа испытуемого на задания возрастающей трудности, для трансформации получаемых при тестировании исходных баллов в линейную интервальную шкалу, а также и для проверки приемлемости профилей исходных баллов испытуемых и заданий[49], что важно для оценки степени пригодности профилей при построении двух шкал — уровня подготовленности испытуемых и уровня трудности заданий.

В определении названы два предмета RM — уровень подготовленности испытуемых и мера трудности каждого задания теста. Для психологии и социологии объектами измерения являются интересующие признаки испытуемых (респондентов) и мера трудности тестовых заданий соответствующего содержания. Система RM нацелена на проверку метрических свойств каждого задания и тес-

41 Обычно исходные тестовые баллы являются ничем иным, как следствием подсчёта числа правильных ответов испытуемых на задания теста, Это только начало измерений, но ещё не сами измерения. См. перевод статьи: Б.Д.Райт и Дж.М.Линека. Различия между исходными тестовыми баллами и измерениями. Сокр. перевод Г.И.Смирновой и А.В.Смирнова // Педагогические измерения. №2. 2006. С. 83–86.

42 Differences between scores and measures. *Wright B.D., Linacre J.M.* Rasch Measurement Transactions. 1989. 3:3 p. 63.

43 *Walker D.A.* Answer pattern and score scatter in tests and examinations. British j. of Psychology. 1931. 22. 73–86.

44 We shall call a set of items of common content a scale if [and only if] a person with a higher rank than another person is just as high or higher on every item than the other person. Guttman L.L. 1950.

та в целом. При отсутствии таковых программно-технологическая часть системы информирует об этом.

В литературе на русском языке RM нередко определяют как измерение по модели Раша. Но исходная модель Раша — это формула не измерения, а расчёта вероятности правильного ответа испытуемого, имеющего уровень подготовленности $и_i$, на задание с уровнем трудности $в_j$. Измерение возникает позже, вследствие трансформации отношения долей правильных ответов к долям неправильным ответов в логарифмическую шкалу.

Своеобразное, можно сказать, графическое истолкование RM дали ведущие зарубежные исследователи RM: это позиционирование испытуемых и заданий на континууме[50], представляющем все мыслимые значения интересующей латентной переменной величины[51]. Пример позиционирования на основе измерения уровня подготовленности испытуемых и уровня трудности заданий читатель может видеть на рис. 1[52].

Каждая латентная величина формируется системой заданий общего тематического содержания, соотносимых с исследуемым признаком. Геометрически латентная величина представляется прямой линией.

B.D. Wright & J.M. Linacre определили RM как процесс сравнения результатов испытуемых на шкале натуральных логарифмов[53]. Не все разработчики это готовы признать, особенно авторы сомнительных, по качеству, тестов и тестоподобных изделий. И действительно, не все тестовые баллы есть результаты измерения. Баллы некоторых, если не большинства, тестов — это обычно результаты элементарного подсчёта количества правильно выполненных заданий.

Придавать большее значение модели (части), чем методу (целому), неразумно, а потому обычное сведение понятия RM к математической модели — слишком заметное упрощение того действительного процесса обновления педагогических измерений, начало которому положил Г. Раш. Он стал основоположником нового, более каче-

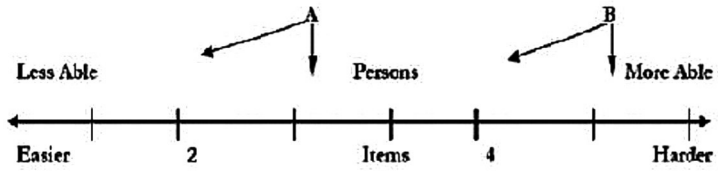

Рис. 1. Позиционирование испытуемых и заданий

The basis for scalogram analyis. In Stouffer et al. Measurement and Prediction. The American Soldier, vol. IV. N-Y, p. 62. New York: Wiley.

45

«Rasch (1960) has devised a truly new approach to psychometric problems. He makes use of none of the classica1 psychometrics, but rather applies algebra anew to a probabilistic model. The probability that a person will answer an item correctly is assumed to be the product of an ability parameter pertaining only to the person and a difficulty parameter pertaining only to the item. Beyond specifying one person as the standard of ability or one item as the standard of difficulty, the ability assigned to an individua1 is independent of that of other members of the group and of the particular items with which he is tested; similarly for the item difficulty». Цитата из статьи: Loevinger, Jane. Person and population as psychometric concepts. Psychological Review, 1965, 72, 143–155.

ственного процесса психолого-педагогического измерения на основе сравнительного анализа исходных результатов тестирования и их преобразования в логарифмическую шкалу.

С математической точки зрения цель применения RM — найти значения так называемого параметра уровня подготовленности каждого испытуемого и параметра уровня трудности каждого задания. Итогом применения RM становятся два множества — параметров испытуемых(θ_i), где i — номер испытуемого, и параметров меры трудности каждого задания (β_j), где j — номер задания. Всего вычисляется $n + m$ число параметров, где n — число испытуемых, m — число заданий.

И именно этот сравнительный анализ привёл Г. Раша к поразительным результатам в смысле успешного преодоления зависимости статистических характеристик заданий от характеристик испытуемых, и наоборот. Что открыло путь для новой, более объективной и эффективной технологии проектирования и разработки теста, а также для более качественного обоснования качества каждого задания и теста в целом с системных позиций.

Самое главное в RM — это то, что создаваемый при этой технологии тест гарантированно становится качественным средством педагогического измерения. Все прочие известные методы и технологии разработки тестов такой гарантии не давали и не дают. Тем самым теорией RM был открыт новый путь к эффективной технологии разработки тестов более высокого качества.

Преимущество вытекает из специфических особенностей модели G. Rasch. Получаемые с её помощью оценки знаний, в силу относительной независимости от конкретного подбора того или иного задания, приобретают характер достаточно объективированных результатов, что также положительно отражается на качестве оценок, используемых в педагогическом контроле. Эта идея превратилась в новое для науки положение о т.н. специфической объективности, как основе для получения справедливых оценок, не зависимых ни от конкретного набора заданий, ни от подбора групп испытуемых. Не случайно измерения по Г. Рашу в западной литературе называют model based measurement.

Ещё одно преимущество связано с возможностями получения интервальной шкалы. Rasch Measurement имеет все достаточные признаки фундаментальной теории. Эта теория имеет сравнительно простую аксиоматику, сводящуюся к простым утверждениям: инте-

45

Профиль испытуемого — это последовательность баллов испытуемого, получаемых им при ответах на каждое задание теста. В матрице тестовых результатов такой профиль представляется в виде вектора-строки. Профиль задания — это последовательность баллов, получаемых испытуемыми в каждом задании. Такой профиль в матрице тестовых результатов представляется в виде вектора-столбца.

47

Протяжённость в пространстве, изображаемая прямой линией, не имеющей ни начала, ни конца, на которой наглядно позиционируют интересующие объекты.

48

A measure is a location on a line. Measurement is the process of constructing lines and locating individuals on lines. Wright, D. N. and M. H. Stone (1979)...Best Test Design. Chicago: MESA Press.

ресующее свойство личности существует в латентном состоянии, оно устойчиво и потому измеряемо с некоторой погрешностью. Вероятность правильного ответа испытуемого зависит от соотношения уровня его подготовленности и от уровня трудности задания.

Теория Г. Раша оказалась непротиворечивой, эффективной, имеющей подтверждения в тысячах практических приложений. Выдвинутый им т.н. принцип separability of estimates позволил освободиться от неустойчивых статистик на выборках заданий и выборках испытуемых.

Следующим преимуществом рассматриваемой модели является сравнительная устойчивость рассчитываемых значений уровня знаний и трудности задания. Это позволяет утверждать: однопараметрические модели удачно оценивают интересующие качества личности, недоступные непосредственному измерению.

Наконец, можно говорить и о методологии Rasch Measurement, которая основана на философии объективного познания, имеет собственный метод т.н. фундаментального измерения, с собственной единицей измерения, с уже упоминавшимися свойствами интервальной шкалы. Эта методология отвечает всем требованиям, выдвигавшимся ведущими учёными к психологическим и педагогическим измерениям:

1. Линейность, что допускает возможность применения арифметических свойств и операций.
2. Параметры заданий и испытуемых не должны быть взаимно зависимы.
3. Метод измерения должен быть сравнительно лёгким.
4. Одномерность измеряемого свойства.
5. Монотонность отображения свойства в числовую шкалу[54].

Можно также говорить и о процессе, и о культуре Rasch Measurement. Процесс состоит из этапов, культура включает в себя философские, теоретические и метрические основы измерения. Rasch Measurement имеет известные преимущества построения линейной интервальной шкалы, с достаточной статистикой, а также эффективные математические обоснования.

RM основано на идее т.н. локальной независимости результатов, которая формулируется как аксиома: для испытуемых одинакового уровня подготовленности вероятность правильного ответа на одно задание не должна зависеть от вероятности правильного ответа на любое другое задание теста. Обычно эта идея формулируется как аксиома, не требующая ни доказательств, ни споров. Кому не нравится, может эту аксиому не принимать во внима-

49

Snider P.D. Exploring the relationship between Individualism-Collectivism Scale (ICS) and attitudes toward Counseling among ethnic Chinese, Australian, and American University Students. Murdoc Univercity, 2003. p.106. http://researchrepository.murdoch.edu.au/319/2/02Whole.pdf

50

Wright B.D., Linacre J.M. A measurement is the quantification of a specifically defined comparison. Rasch model derived from objectivity. Rasch Measurement Transactions 1:1 p. 5. 1987.4; A measure is a location on a line. Measurement is the process of constructing lines and locating individuals on lines. Wright *B.D., Stone M.H.* Best test Design: Rasch Measurement. Chicago: MESA Press. 1979.

51

D. Andrich. Relatioship Between the Thurstone and Rasch. (Approaches to Item Scaling Appl. Psychol. Measurement V2. 1978, pp. 451–462.

ние. Но тогда отрезается возможность опираться на теорему умножения вероятностей независимых событий, используемую при разработке настоящих тестов, обладающих метрическими свойствами.

Тестовый процесс в истолковании Георга Раша

Тестовый процесс в RM представляется метафорой единоборства испытуемого с заданием. При этом предполагается, что испытуемый обладает некоторым уровнем подготовленности, а задание обладает некоторым уровнем трудности. Неподготовленный испытуемый и абсолютно лёгкое задание не участвуют в тестовом процессе; в RM их относят к экстремальным явлениям, мешающим исследованию параметров личности и заданий. Исход единоборства испытуемого с заданием можно априори предположить зависимым от количественного соотношения этих двух свойств.

На уровне здравого смысла предполагалось, что если ответ на задание правильный, то испытуемый подготовлен на уровне трудности данного задания. Если ответ неправильный, то не подготовлен. Случаи угадывания, списывания нарушают правильность этого предположения. Далее предполагалось, что если у испытуемого много правильных ответов, то он, вероятно, подготовлен лучше тех, у кого меньше правильных ответов. Формально это предположение выражается понятием «сумма исходных баллов испытуемого».

В RM вероятность правильного ответа испытуемых зависит от двух показателей — от уровня их подготовленности θ_i и от меры трудности заданий β_j.

Для упрощения метафоры тестового процесса Г. Раш предположил, что вероятность успеха испытуемого должна зависеть только от разности параметров[55]. На математическом языке это выглядит так:

$$P_{ij} = f(\theta_i - \beta_j).$$

И действительно, в RM вероятность правильного ответа испытуемых зависит от разности двух показателей — от уровня их подготовленности θ_i и от меры трудности заданий β_j. Чем больше значение разности, тем больше вероятность правильного ответа испытуемого i на задание j.

Это было очевидной редукцией мультифакторного процесса взаимодействия испытуемого с заданием, и это действительно является некоторым огрублением действительности. Редукция оказалась весьма удачной, сравнительно простой и эффективной. Примерно такая же логи-

52 It is the difference $(\theta_i - \beta_j)$ which governs the probability of a right answer. Wright B.D. Solving measurement problems with the Rasch model. J. of Educational Measurement 14 (2) pp. 97–116, Summer 1977.

53 *Деменчёнок О.Г.* Подбор параметров модели педагогических измерений // Педагогические измерения. № 1. 2008. С. 27–49.

54 Это таблица 1.4.1 из книги: *Wright, D. N. and M. H. Stone* (1979). Best Test Design. Chicago: MESA Press.

55 *Birnbaum A.* Some latent trait models and their use in inferring an examinee's ability. In F.M. Lord and M.R. Novick, Statistical theories of mental test scores (pp. 397–545). Reading, Mass.: Addison-Wesley. 1968.

ка заключена в работе по уменьшению числа показателей в регрессионном анализе. Тогда редуцированная система показателей, если она умело произведена, оказывается эффективнее, чем нередуцированная система.

К 1958 году у G. Rasch возникла идея выразить вероятность правильного ответа на задание j посредством так называемой логистической функции вида. Функцию, описывающая зависимость вероятности правильного ответа испытуемого от уровня подготовленности испытуемого и от меры трудности задания, обычно записывается таким образом:

$$P_j\left\{X_{ij}=1|\beta_j\right\}=\frac{\exp(\theta-\beta_i)}{1+\exp(\theta-\beta_i)},$$

где $X_{ij} = 1$, если ответ любого испытуемого (i) на любое задание под номером j правильный; θ – уровень знаний, значение на латентной переменной величине; β_j — уровень трудности j-го задания теста.

В результате модель Г. Раша стала называться однопараметрической, в которой вероятность правильного ответа стала зависеть только от разности двух параметров.

Эту функцию можно записать и другими способами, например, в строку:

$$P_j\left\{X_{ij}=1|\beta_j\right\}=\frac{\exp(\theta-\beta_j)}{1+\exp(\theta-\beta_j)},$$

где $X_{ij} = 1$, если ответ любого испытуемого на j-е задание правильный.

Ещё один способ записи:

$$P_j(\theta)\left\{X_{ij}=1|\beta_j\right\}=\frac{1}{1+e^{-L}}$$ или

совсем короче $$P_j(\theta)=\frac{1}{1+e^{-L}}.$$

$$P_j(\theta)=\frac{1}{1+e(-L)},$$

где L представляет разность параметров испытуемых i и заданий j.

Если берётся конкретное задание под номером j, то разность записывается с индексом соответствующего номера β_j. Если берётся конкретный испытуемый i с присущим ему уровнем подготовленности, то разность пишется $(\theta_i - \beta)$.

Эта формула даёт возможность непосредственно сопоставить любое задание с любым испытуемым и на основе такого сопоставления вычислить вероятность получения правильного ответа. На основе такого сопоставления ЭВМ подбирает очередное задание в системах адаптивного обучения и контроля знаний. Если вероятность низкая, то подбирается задание полегче; если высокая, то потруднее. Общий принцип подбора заданий — в районе 50% вероятности получения правильного ответа.

Значения θ_i и β_j находятся на самых первых этапах разра-

ботки теста. Самый простой алгоритм расчёта PROX уже освещался в нашем журнале[56]. Знание параметров позволяет перейти к прогнозированию вероятности правильного ответа каждого испытуемого на каждое задание.

Возьмём пример, где значение параметров a и b для задания принимается для начала неизменным, равным 1,0. Это случай очень низкого уровня подготовленности испытуемого со значением $\theta_i = -3{,}0$.

Первый шаг: $L = (\theta - \beta_j)$.

Подставляя эти данные, получаем $L = (-3{,}0 - 1{,}0) = -4{,}0$,

$e - L = 2{,}71828 - (-4{,}0) = 54{,}59801$.

Второй шаг. Находится значение знаменателя формулы (1)

1 + 54, 59801 = 55,59801.

Третий шаг. Находится вероятность правильного ответа при использованных данных:

$$P_j(\theta) = \frac{1}{1+e^{-L}} = \frac{1}{55{,}59801} = 0{,}018315.$$

Интерпретация полученного результата для испытуемых очень низкого уровня подготовленности, равного –3,0 логита, вероятность правильного ответа на задание уровня трудности 1,0 логит равна 0,018315. Из чего видно, что правильный ответ малоподготовленного испытуемого на трудное задание весьма маловероятен. В это трудно поверить.

Теперь пришло время посмотреть, как меняется вероятность правильного ответа на одно и то же задание, с уровнем трудности b = 1,0, для испытуемых, имеющих различный уровень подготовленности. Для этого достаточно провести небольшой вычислительный эксперимент, в котором надо последовательно (с шагом +1) брать разные уровни подготовленности и полученные данные свести в табл. 1 результатов вычислительного эксперимента по определению вероятности правильного ответа испытуемых различного уровня подготовленности на задание с параметрами трудности b = 1,0.

График функции при b = 1 представлен на рис. 2.

Интересно проследить, как меняется вероятность правильного ответа при изменении значений обоих параметров. Этот результат представлен в табл. 2.

В последнем столбце табл. 2 представлены значения вероятности правильного ответа[54].

А. Бирнбаум[55] разработал двух- и трёхпараметрическую модель определения вероятности правильного ответа испытуемого, в зависимости не только от уровня трудности задания, но и от уровня его дифференцирующей способности и от потенциальной вероятности угадать правильный ответ в данном задании[56].

56

Эти формулы приводятся в работе: *Аванесов В.С.* Применение тестовых форм в Rasch Measurement // Педагогические измерения. № 4. 2005. С. 3–20.

Таблица 1

Значения вероятности правильного ответа испытуемых в зависимости от уровня их подготовленности

Уровень подготовленности испытуемых, θ_i	$L = j(\theta - \beta_j)$	e^{-L}	$1+e^{-L}$	$P_j(\theta)$
–3,0	(–3 – 1) = –4	54,59	55,598	0,018
–2,0	(–2 –1) = –3	20,08	21,086	0,047
–1,0	(–1–1) = –2	7, 389	8,389	0,166
0	(0 – 1) = –1,0	2, 718	3, 716	0,269
1,0	(1 – 1) = 0	1	2	0,500
2,0	(2 – 1) = 1,0	0,368	1,368	0,731
3,0	(3 – 1) = 2	0,135	1,135	0,881

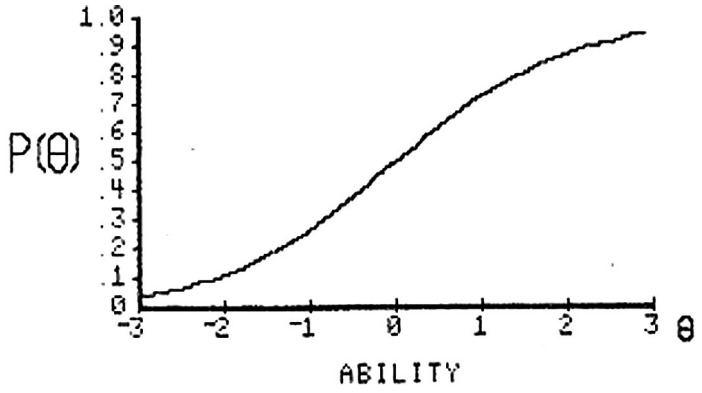

Рис. 2

Это было сделано в рамках математической теории педагогических измерений (Item Response Theory, IRT). Применение заданий по математическим моделям А. Бирнбаума улучшило метрический анализ свойств заданий, но ухудшило условия разработки теста как системы заданий возрастающей трудности. Главной проблемой стало пересечение графиков заданий, а значит и появление дополнительных ошибок измерений

Можно сказать так: IRT позволяет глубже разобраться в метрических достоинствах каждого отдельного задания, что полезно при решении вопроса о его включении в состав теста. A RM позволяет создать лучшую систему заданий возрастающей трудности, отвечающей критериям качества педагогических измерений.

Таблица 2

Зависимость вероятности правильности правильного ответа от разности параметров

Значения параметра испытуемых	Значения параметра трудности заданий	Разность параметров	Значения r	
			Числителя	Вероятности
θ_i	β_j	$\theta_i - \beta_j$	$\exp(\theta_i - \beta_j)$	P_{ij}
5	0	5	148	0,99
4	0	4	54,6	0,98
3	0	3	20,1	0,95
2	0	2	7,39	0,88
1	0	1	2,72	0,73
0	0	0	1,00	0,50
0	1	–1	0,368	0,27
0	2	–2	0,135	0,12
0	3	–3	0,050	0,05
0	4	–4	0,018	0,02
0	5	–5	0,007	0,01

Главные идеи RM

Как и всякое учение, RM характеризуется рядом идей. Можно назвать пять основных идей, реализация которых и позволила появиться метрической системе RM. Вспомним, что создавая тесты в рамках классической теории, разработчик всё время стоял перед проблемой зависимости тестовых результатов испытуемых от уровня трудности используемых заданий. Включение лёгких заданий в тест приводило к высоким тестовым баллам испытуемых. И наоборот, включение в тест заметного числа трудных заданий снижало тестовые баллы большинства испытуемых. Кроме того, баллы и место отдельного испытуемого при тестировании сильно зависят от состава тестируемой группы. В сильной группе его результаты окажутся низкими, в слабой группе — высокими[57].

Ответы на каждое задание теста априорно рассматриваются независимыми от ответов на любое другое задание теста. Впрочем, допустимость этой аксиомы для имеющихся данных при разработке качественного теста всегда проверяется. Если аксиома нарушается, тест с такими заданиями не будет обладать метрическими свойствами.

В RM не должны быть взаимно зависимы параметры за-

57

Как справедливо писал о недостатках традиционного тестирования B.D.Wright, «My ability depended not only on which items I took but on who I was and the company I kept»! Источник: Sample-free Test Calibration and Person Measurement. http://www.rasch.org/memo1.htm

даний и испытуемых. Иначе теряются метрические свойства теста. С этим свойством идеального теста никто не спорил. Все были «за», но реально преодолел зависимость параметров испытуемых и заданий, исключительно математическим средствами, только Г. Раш. Г. Раш изначально поставил задачу разработки такого метода измерения, где отмеченная зависимость не должна себя проявлять (parameter separation).

1. Мера трудности каждого задания должна иметь устойчивое значение (т.н. параметр трудности задания).

2. Вторая идея RM формулировалась симметрично первой идее, но она касалась испытуемых. Уровень подготовленности испытуемых не должен меняться в зависимости от уровня трудности заданий, включаемых в тест. Здесь использовался тот же эффект.

3. Все задания гомогенного теста должны измерять только одно интересующее свойство личности. Это идея одномерности, что проявляется в стремлении измерять, в одном метрическом процессе только одно интересующее свойство личности. Этот подход развивается в русле идей классиков психометрики.

4. Вероятность правильного ответа более подготовленного испытуемого должна быть выше вероятности правильного ответа менее подготовленного испытуемого.

5. Вероятность правильного ответа испытуемого на лёгкое задание должна быть выше вероятности правильного ответа на трудное задание.

RM опирается также на вероятностную идею т.н. локальной независимости результатов, которая в некоторых работах формулируется как аксиома: для испытуемых одинакового уровня подготовленности вероятность правильного ответа на одно задание не должна зависеть от вероятности правильного ответа на любое другое задание теста.

Метрические свойства тестов и заданий

Из нового определения RM вытекает, что критерий наличия или отсутствия метрических свойств позволяет отделить настоящие тесты от псевдотестов, и от всего того, что сейчас в России выдаётся за средства педагогических измерений. Важно отметить — те тесты, которые делаются по технологии RM, имеют метрические свойства. Это следствие самой технологии разработки тестов с метрическими свойствами. Проблема демаркация педагогических измерений от множества тестоподобной и иной некачественной продук-

ции не случайно сформулирована как главное направление развития педагогических измерений в России[58].

Отсюда становится понятно, что методология, теория, технология и методы несводимы к одной лишь математической модели. В логике данного определения формальные модели процесса тестирования — лишь их часть. В RM математические модели используются для прогнозирования вероятности правильного или неправильного ответа испытуемого на задания возрастающей трудности, для трансформации получаемых при тестировании исходных баллов в линейную интервальную шкалу, а также и для проверки приемлемости *профилей исходных баллов испытуемых и заданий*. Что важно для оценки степени пригодности профилей при построении сразу двух шкал — уровня подготовленности испытуемых и уровня трудности заданий.

Эффект проверки метрических свойств каждого задания в RM можно сравнить с появлением рентгена для диагностики органов человека.

Это стало возможным за счёт более качественного анализа метрических свойств каждого отдельного задания.

Посмотрим примеры графиков заданий, не обладающие по разным причинам метрическими свойствами.

На рис. 3 приведён график задания, которое не способно объективно оценить вероятность правильного ответа у хорошо подготовленных испытуемых[59]. Для них данное задание оказалось труднее, чем прогнозирует модель. Результаты выполнения этого задания оказались хуже прогнозируемого уровня по модели Г. Раша. График указывает на то, что с этим заданием что-то не так. Причины такого дефекта задания могут скрываться в словесной формулировке задания, которое могло быть непонятным или неправдоподобным именно для этих испытуемых. Точный диагноз дефекта задания может дать только экспертиза его содержания, а также формы этого задания.

На рис. 4 приведён график, который указывает на неспособность задания объективно оценить вероятность правильного ответа у испытуемых среднего уровня подготовленности. Средний арифметический результат в этой группе оказывается заметно ниже ожидаемого[60].

Причины такого дефекта задания могут скрываться в словесной формулировке задания: оно оказалось непонятным или неправдоподобным именно для испытуемых со средней подготовкой. Точный диагноз дефекта задания опять может

58 *Аванесов В.С.* Проблема развития педагогических измерений// Педагогические измерения. № 2. 2011. С. 3–35.

59 *Аванесов В.С.* Метрическая система Георга Раша. – Rasch Measurement // Педагогические измерения. № 2. 2010. С. 57–81.

60 *Аванесов В.С.* Метрическая система Георга Раша. – Rasch Measurement // Педагогические измерения. № 2. 2010. С. 57–81.

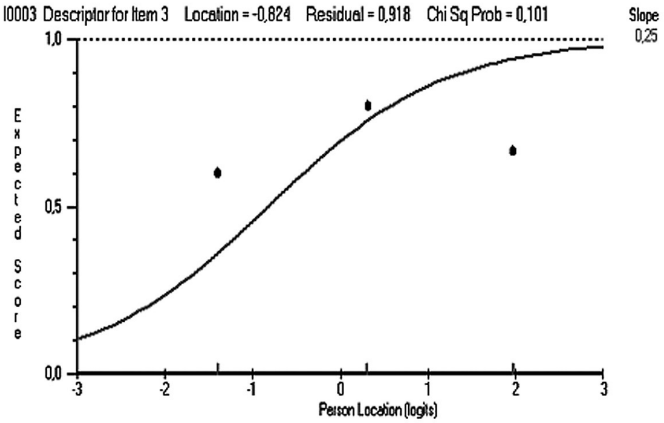

Рис. 3

дать только экспертиза его формы и содержания.

Два следующих графика дают примеры заданий, не имеющих метрических свойств. Это видно из значений средних арифметических всех трёх групп испытуемых. Испытуемые с низким уровнем подготовленности имеют большую вероятность правильно ответить на это задание, чем испытуемые, имеющие средний и высокий уровень подготовленности[61].

На рис. 6 приводится пример графика задания, которое должно было бы лучше фиксировать различия в группах испытуемых. На деле же оно занижает прогноз успешности у слабых испытуемых и завышает

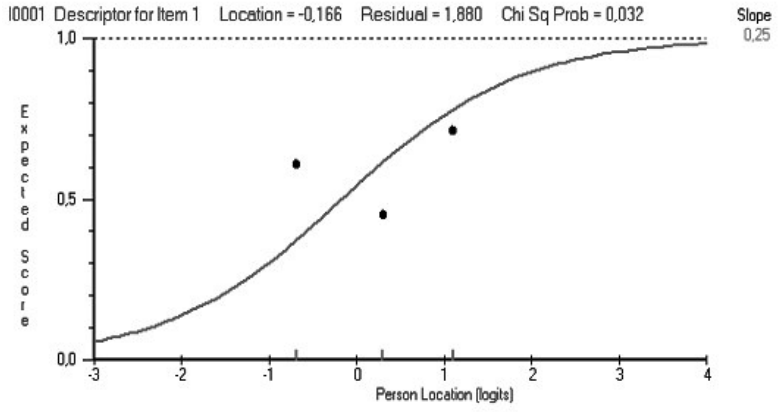

Рис. 4

Смирнова Г.И. Применение программы RUMM-2020 для разработки педагогического теста // Педагогические измерения. № 3. 2010. С. 20–32.

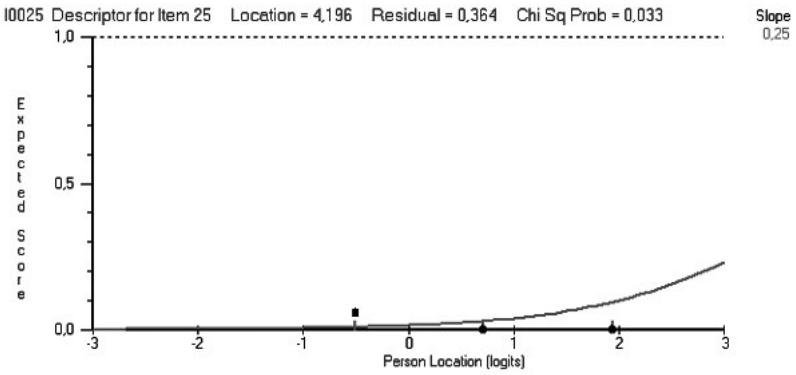

Рис. 5

средний балл у хорошо подготовленных испытуемых[62].

Только после появления метрической системы Rasch Measurement стало понятно, что можно создать сколько угодно тестов как обладающих, так и не обладающих метрическими свойствами. В последнем случае понятие «тест» вымывается или, иначе говоря, выхолащивается. А посему в определение теста возникла необходимость обязательного включения идеи теста как средства педагогического измерения. Только тогда появляется основание и возможность отделить настоящие тесты от прочих форм и методов, выдаваемых за педагогические измерения[63].

Вряд ли случайно, что большинство классиков западной психометрики ранее пользовались понятием «теория тестов», а не понятием «теория педаго-

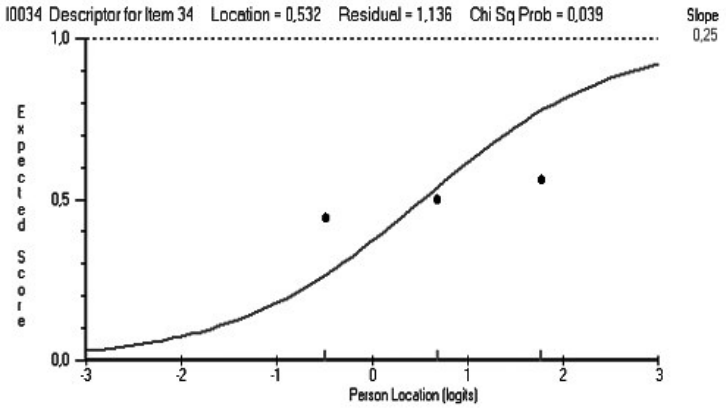

Рис. 6

62 Там же.

63 Как это случилось в России с т.н. КИМами ЕГЭ, производимыми Федеральным институтом педагогических измерений.

гических измерений»[64]. Вероятно, многие из них понимали, что те тесты ещё не отвечали требованиям измерений, проводимых на линейной интервальной шкале. Отсюда становится понятной необходимость различать такие понятия, как «*тест*» и «*тестирование*» от понятия «*педагогические измерения*» и «*педагогическое оценивание*».

Появление метрической системы Rasch Measurement актуализировало идею постоянной проверки тестов и заданий на обладание метрическими свойствами.

Метрическими свойствами обладает не каждый тест, а только такой, который соответствует *критериям качества педагогических измерений*. Остальные тесты, различные формы и методы только оценивают[65]. В этой связи полезно напомнить, что и сейчас все так называемые «тесты» и «тестовые баллы», упоминаемые в отчётах ФИПИ, в действительности не являются педагогическими измерениями. И как отмечалось в предыдущей статье[66], совсем не являются педагогическими измерениями результаты применения т.н. «КИМов ЕГЭ». Это лишь оценки, к тому же некачественные.

Метрические свойства теста в целом и тестовых заданий — тоже сравнительно новые понятия, нуждающиеся в определениях. Метрические свойства не возникают сами по себе, из одного только факта разработки теста. Нужно организовать ещё два процесса шкалирования получаемых тестовых баллов испытуемых и уровней трудности заданий, а также наладить процесс обоснования качества каждого задания теста и результата каждого испытуемого. И только после этого у теста возникают метрические свойства. Наглядным выражением наличия метрических свойств теста является рис. 7, выражающий связь между уровнем подготовленности испытуемых (ось абсцисс, в шкале логитов) и суммой получаемых исходных тестовых баллов[67] (ось ординат). На рис. 7 представлен пример такой связи. Она, очевидно, нелинейна, монотонна, непрерывна. Большему уровню подготовленности соответствуют и сравнительно большие значения исходных тестовых баллов испытуемых.

Метрическими свойствами обладает только такой тест, который соответствует *критериям качества педагогических измерений*. Четыре основных критерия уже рассматривались в журнале: это надёжность, валидность и объективность тестовых результатов вместе с критерием эффективности теста и тестовых заданий. Остальные методы, не отвечающие названным критериям, только оценивают испытуемых. Здесь возникают вопросы принципиальных

64

Gulliksen. H. The theory of Mental Test Scores. N-Y., Wiley. 1950.

65

На этот случай полезно использовать другое английское понятие «assessment». В действительности, КИМы ЕГЭ дают некоторые оценки, иногда очень грубые, как это случилось, например, с оценками ЕГЭ по математике 2010 г. для наиболее подготовленных абитуриентов вузов. Абсолютное большинство баллов испытуемых там оказалось ниже среднего арифметического исходной шкалы. Заданий запредельного уровня трудности было немного, и их выполнило слишком мало испытуемых. А потому такой же запредельно неприемлемой и недопустимой оказалась и погрешность измерения. См. обоснование данного вывода в статье: Аванесов В.С. Ошибочные цели – плачевные результаты. Педагогические измерения. № 4. 2010.

66

Аванесов В.С. Проблема развития педагогических измерений. Педагогические измерения. № 2. 2011.

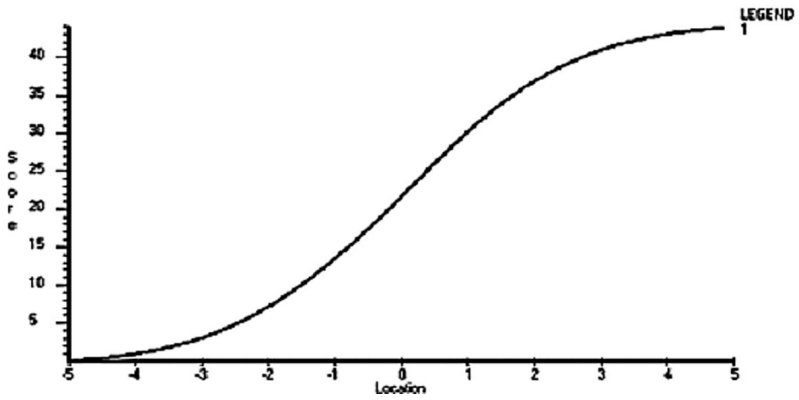

Рис. 7

различий оценок (оценивания) от педагогических измерений. Но это предмет другой статьи.

Проектирование тестов в RM

Классический учебник по RM, написанный последователями Г. Раша, называется «Best Test Design»[68]. Элементарный перевод этого словосочетания — «Проектирование наилучшего теста», что сводило бы всю проблематику к двум основным вопросам: что такое наилучший тест? и как его надо проектировать?

Уже в одном только названии классического учебника появились сравнительно новые понятия, такие как «наилучший тест», «проектирование наилучшего теста», разработка (создание, проектирование, конструирование) наилучшего теста. Результаты именно такого теста непременно становятся педагогическими измерениями. Эти понятия в научной литературе на русском языке, насколько известно, никогда не обсуждались и не вводились в уже опубликованные тексты по педагогическим измерениям. Возможно, что раз не было понятий, то не было и профессионального движения в сторону именно таких тестов, которые непременно обладают метрическими свойствами.

Потенциальные оппоненты могут возразить, что движения в сторону лучших тестов не было хотя бы потому, что само это понятие вряд ли можно считать научно обоснованным. Оно

67 http://en.wikipedia.org/wiki/Rasch_model

68 *Wright B.D., Stone M.H.* Best test Design: Rasch Measurement. Chicago: MESA Press. 1979.

действительно содержит в себе элемент метафоры. В науке метафоры — не редкость. Но каждый, кто их использует, должен отдавать себе отчёт о границах применимости метафоры и об опасностях упрощения сути исследуемого объекта[69].

Хотя сказать, что в словосочетании «best test» нет науки, было бы неверно. Идеальные объекты были всегда самыми привлекательными для разработки критериев качества и эффективности. Да и все сторонники этого направления уверены, что лучше метрической системы Г. Раша нет ничего, и что только эта система измерения способна породить лучшие тесты. Автор данной статьи разделяет такую точку зрения. Но это совсем не значит, что в RM критиковать нечего.

Например, название «best test» звучит немного метафорично, вызывающе, а потому данной статье предлагается рассматривать его скромнее, менее метафорично, более эмпирично. На ум приходит выражение «качественный тест», с присущими ему метрическими свойствами. Именно эти свойства можно подтвердить или опровергнуть, теоретически и эмпирически, по имеющимся в педагогических измерениях критериям качества тестов. По критериям традиционным и по критериям, специально сформулированным в технологии RM.

В процессе проектирования теста важно выделить целевую группу, в которой предполагается использование теста. Это перевод английского словосочетания target group.

Здесь принимается во внимание, в первую очередь, уровень подготовленности испытуемых. Если уровень высокий, то и задания теста подбираются нелёгкими. Иначе результаты теста окажутся невалидными для поставленной цели.

Кроме того, принимается во внимание уровень вариации интересующего свойства у испытуемых. В гомогенной группе вариация меньше. Соответственно, и задания теста будут меньше варьировать по уровню трудности.

Напротив, если предстоит тестировать гетерогенную группу, где представлены испытуемые с заметно различающимся уровнем подготовленности, то и задания теста должны больше варьировать по уровню трудности. Увеличивается и число необходимых заданий, иначе точность измерений становится недостаточной.

В лучшем тесте средний арифметический балл испытуемых равен среднему числу используемых заданий, коэффициенты асимметрии и эксцесса не отклоняются от значений для стандартной кривой нормального распределения результатов. Хорошо, если значе-

69

В России больше распространены два других мифа. Первый о контрольных материалах ЕГЭ, которые якобы обладают измерительными свойства0и. Никто и никогда эти свойства не видел. Второй миф — о Едином государственном экзамене, который якобы может объективно и с одинаково высоким качеством измерить уровень подготовленности как ординарных выпускников школ, так и продвинутых абитуриентов хороших вузов. Нет никаких данных, подкрепляющих этот официозный миф. А те аргументы и факты, которые стали известны, обнуляют оба эти мифа на корню. См. Аванесов В.С. Проблема развития педагогических измерений //Педагогические измерения. №2. 2011.

ния средней арифметической, моды и медианы совпадают. Это признак точной нацеленности общего уровня трудности теста на уровень подготовленности испытуемых.

В тесте нужны задания, измеряющие только интересующее содержание учебной дисциплины. Это правило основывается на принципе гомогенности содержания теста.

Задания теста должны быть равномерно возрастающей трудности. Это вытекает из определения теста и принципа соответствия уровня трудности заданий уровню подготовленности испытуемых. В RM равномерность распределения заданий проверяется и является одним из методологических требований.

Задания, имеющие сходные параметры, избыточны для сбалансированного теста. А потому такие задания в тест не включаются.

В тесте желательно иметь достаточно большой размах значений параметра трудности, в пределах, по размаху, не менее $-3 < \beta < +3$.

Лучше, если значения a_j принимаются одинаковыми. Реально они имеют заметно различающиеся значения. Но в модели Г. Раша с общим значением параметра крутизны заданий системные свойства теста появляются лучшим образом.

И наконец, при проектировании теста закладываются возможности получения достаточных значений уровня надёжности и валидности тестовых результатов в целевой группе. Информация обо всех этих показателях должна обязательно публиковаться в отчёте по разработке теста.

При использовании технологии RM в обязательном порядке приводятся данные по решению каждой из упомянутых в обзоре задач. В самом отчёте рекомендуется указывать номера заданий и тексты экстремальных заданий. Специалисты–предметники анализируют причины возникновения эффекта экстремальности. Обычные причины – задания по трудности не соответствуют уровню подготовленности испытуемых. Экстремальные задания – это те, на которые либо отвечают правильно все испытуемые, либо отвечают неправильно.

Технология RM позволяет определить соответствие уровня трудности разрабатываемого теста уровню подготовленности студентов. Эффект несоответствия (или недостаточного соответствия) виден на гистограммах (рис. 8).

По гистограмме распределения исходных баллов испытуемых можно видеть, что в тесте не хватает средних, трудных и очень трудных заданий. Из-за этого качество измерения оказывается неприемлемым. Например, для наиболее подготовлен-

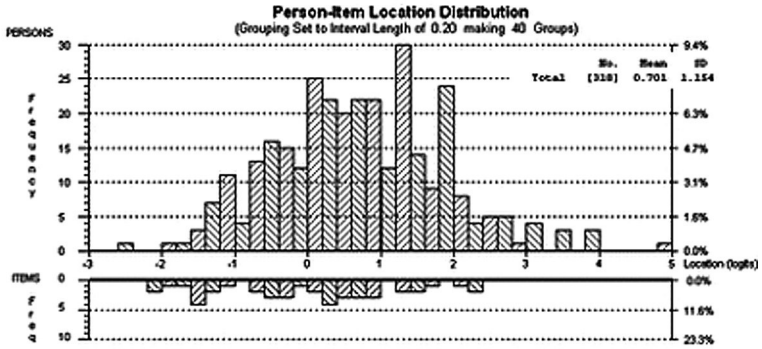

Рис. 8. Совмещённые гистограммы распределения исходных тестовых баллов испытуемых и мер трудности заданий (то и другое — в логитах). По оси ординат представлены соответствующие частоты

ных испытуемых в тесте (рис. 8) не оказалось достаточного числа заданий повышенной трудности.

Достаточность числа заданий теста – это критерий формальный и концептуальный одновременно. Анализ критериев качества RM – предмет отдельного исследования и новой журнальной публикации.