ПЕДАГОГИЧЕСКИЕ ИЗМЕРЕНИЯ ПО МОДЕЛИ ГЕОРГА РАША

Вадим Аванесов

В СТАТЬЕ РАССМАТРИВАЮТСЯ ВОПРОСЫ СОЗДАНИЯ И ПРИМЕНЕНИЯ СИСТЕМЫ ПСИХОЛОГО-ПЕДАГОГИЧЕСКИХ ИЗМЕРЕНИЙ, НАЗВАННОЙ НА ЗАПАДЕ В ЧЕСТЬ ЕЁ ОСНОВАТЕЛЯ, ДАТСКОГО МАТЕМАТИКА ГЕОРГА РАША. ДАЁТСЯ АНАЛИЗ ЦЕЛЕЙ И ЗАДАЧ RM, ИЗУЧЕНЫ ПРИЧИНЫ ОТСТАВАНИЯ ИССЛЕДОВАНИЙ RM В РОССИИ ОТ ДРУГИХ СТРАН. СФОРМУЛИРОВАНА ПОДХОДЯЩАЯ ТЕРМИНОЛОГИЯ RM НА РУССКОМ ЯЗЫКЕ.

В процессе педагогических измерений акцент часто делается не только на результатах испытуемых, но и на необходимости определения устойчивой меры трудности заданий. Устойчивость здесь понимается как независимость или малая зависимость значений параметра трудности заданий от выборки испытуемых. Полезно напомнить, что если в рамках статистической теории педагогических измерений задания даются хорошо подготовленным испытуемым, то мера их трудности становится низкой; если же даются слабым испытуемым, то те же самые задания считаются лёгкими.

Различия в мерах трудности заданий в зависимости от уровня подготовленности испытуемых породили словосочетание «sample-dependant item characteristics». Симметрично можно говорить и о зависимости параметров уровня подготовленности испытуемых от уровня трудности заданий теста (item-depended person characteristics). Система RM и математическая теория измерений IRT[1] возникли из стремления преодолеть отмеченные зависимости.

Изначально выделяются два взаимосвязанных объекта измерений — уровни трудности заданий и уровни подготовленности испытуемых. В RM эти объекты участвуют одновременно в рамках одного общего исследования. Поэтому такое измерение часто называют совместно проводимым (joint measurement).

Определение главных понятий

В технологическом смысле Rasch Measurement можно определить как процесс и метрическую систему трансформации результатов тестирования в педагогические измерения. В этом определении главными стали словосочетания «процесс измерения», «трансформация» и «метрическая система». Отсюда вытекает, что, во-первых, RM — это больше чем математическая модель, во-вторых, RM — явление процессуальное и, в-третьих, явление системное, требующее системного подхода и системного анализа результатов. Результатом RM являются измерения со свойствами интервальной шкалы.

Педагогическое измерение в данной статье понимается как *процесс* определения меры *интересующего латентного свойства личности* испытуемого на *интервальной шкале* посредством *качественного* теста, состоящего из *системы* заданий равномерно возрастающей трудности, позволяющего получать педагогически целесообразные результаты, отвечающие критериям *надёжности, валидности, объективности и эффективности*. В этом определении курсивом выделены основные термины, позволяющие отграничить педагогические измерения от прочих методов — научных, псевдонаучных и ненаучных[2].

[1] **Вадим Аванесов.** Понятия и методы математической теории педагогических измерений (Item Response Theory, IRT). Статья третья.

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

Самое короткое и узкое определение RM — это метод трансформации исходных тестовых результатов в интервальную шкалу натуральных логарифмов. В этом определении главное — процесс трансформации исходных тестовых баллов в шкалу натуральных логарифмов, после чего, собственно, и появляется измерение. До процесса логарифмического преобразования исходные баллы испытуемых не рассматриваются как измерения[3].

Системный подход к RM позволил по-новому определить и понятие «педагогический тест». В новой формулировке он определяется так: это система заданий равномерно возрастающей трудности, позволяющая качественно оценить структуру и измерить уровень подготовленности испытуемых. Все задания педагогического теста должны иметь варианты замены, позволяющие защитить тест от рассекречивания или от списывания. В одной аудитории каждый испытуемый получает свой, отличающийся от других вариантов теста, но сопоставимый с другими вариантами по содержанию и по мере трудности[4].

Всё, что разработано Г. Рашем, сделано на языке математики, а потому не имеет конкретной привязки к педагогике или психологии, равно как и к измерению какого-либо одного свойства личности. Уже одно это свидетельствует об общности и оригинальности его теории. Это обстоятельство не воспринималось должным образом современниками Г. Раша.

И только спустя двадцать семь лет, после выхода его книги в США на английском языке, а также после появления в США первого последователя его теории, Бенджамена Райта[5], стало понятным, что RM — оригинальная научная система измерений, своеобразный подход к вопросам разработки тестов. Своеобразие этой системы в том, что она состоит не только из ряда теорий и научных положений, но и включает технологию разработки тестов, а также компьютерные программы для сопряжённого вычисления меры трудности заданий и уровня подготовленности испытуемых.

Постановка проблемы

RM популярна сегодня во всём мире, причём в различных сферах. Это касается не только педагогики и психологии, но и социологии, медицины. Преимущество RM заключается в следующем:

• этот метод обеспечивает получение валидных результатов посредством применения статистик адекватности (fit statistics), диагностической информации, карты (Person-item map) сравнения уровня трудности заданий с уровнем подготовленности испытуемых;

• даёт информацию о надёжности измерений посредством расчёта стандартных ошибок измерений, оценок параметров заданий и параметров подготовленности испытуемых на одной шкале;

• даёт возможность оценить параметры уровня подготовленности испытуемых независимо от уровня трудности заданий в имеющейся выборке заданий;

• оценивает параметры уровня трудности заданий независимо от уровня подготовленности выборки испытуемых;

• представляет параметры испытуемых и заданий на одной общей линейной шкале, что помогает критериально-ориентированной и нормативно-ориентированной интерпретации данных;

• ставит в фокус исследования отдельные задания и результат отдельных испытуемых, в отличие от статистической теории измерений (СТТ), где исследователь имеет дело с обобщённой статистикой свойств заданий и испытуемых;

• даёт возможность уравнивания баллов испытуемых, полученных на параллельных вариантах заданий, измеряющих одно и то же интересующее свойство[6].

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

[4] В КИМах ЕГЭ это требование параллельности заданий теста не выполняется. И это одна из важных причин, почему они не являются педагогическими измерениями. См. http://viperson.ru/wind.php?ID=563869&soch=1

[5] В 1964 г. B.D. Wright специально поехал в Данию познакомиться с Г. Рашем и его работами. См.: Review of cooperation between B D Wright and G Rasch. Rasch Measurement Transactions, 1988, 2:2 p. 19. http://www.rasch.org/rmt/rmt22c.htm

Практика применения RM насчитывает десятки тысяч исследований, проводившихся в разных странах в течение полувека и опубликованных на многих языках мира. Это свидетельствует об актуальности проблемы RM.

О. Г. Деменчёнок считает, что модель Раша можно рассматривать как научную гипотезу, основанную на следующих предположениях:

1) мера уровня подготовленности любого испытуемого t_i (т. е. количественная характеристика уровня подготовленности испытуемого по определённому множеству заданий теста) не должна зависеть от уровня трудности тестовых заданий $t_i \in (0;\infty)$;

2) вероятность правильного ответа испытуемого *Pi* зависит только от уровня подготовленности испытуемого и от уровня трудности тестового задания $b \in (0;\infty)$ (т. е. количественной характеристики тестового задания, не зависящей от выборки испытуемых и отражённой на определённой шкале по конкретному разделу определённой области знания) или $P = f(t, b)$[7].

Для построения шкалы измерений оказалось удобным выражать уровень подготовленности t и уровень трудности b на шкале логарифмов: $\theta = \ln(t)$, $\beta = \ln(b)$, где θ и β — значения уровней подготовленности и трудности, измеряемые в логарифмическом масштабе. В соответствии с принятой терминологией и нотацией далее под уровнями подготовленности и трудности будем понимать $\theta \in (-\infty;\infty)$ и $\beta \in (-\infty;\infty)$.

Автор этой статьи рассматривал проблему взаимосвязи форм тестовых заданий и требований модели Раша и сделал неожиданный для многих практиков вывод о непригодности повсеместно используемых сейчас заданий с выбором одного правильного ответа из 2–5 предлагаемых на выбор вариантов для применения в системе RM[8] с целью получения качественных измерений. Непригодность вытекает из-за неизбежности угадывания правильного ответа теми испытуемыми, которые подготовлены недостаточно. В итоге появляются ошибки измерения, снижающие качество педагогических измерений. Вместо критикуемых заданий с выбором одного правильного ответа были предложены задания с выбором нескольких правильных ответов, где вероятность угадывания правильных ответов со стороны неподготовленных испытуемых очень низка, порядка 0,001–0,010, то есть практически ничтожная.

Сам Г. Раш задания с выбором одного правильного ответа в своей работе не использовал, потому что в его время и в его окружении они не применялись. Он мыслил больше математически, чем технологически, предпочитал иметь дело с заданиями открытой формы, где угадывание исключено, не заботясь при этом о трудностях сбора данных посредством заданий такой формы в массовых исследованиях.

В наше время задания открытой формы автор этой статьи рекомендует применять только в двух случаях: для проверки правильности написания трудных слов и только в качестве пробного этапа разработки заданий с выбором одного или нескольких правильных ответов, для поиска подходящих дистракторов. В обоих случаях тестирование необходимо проводить на компьютерах, чтобы полностью исключить ручной труд и возможность намеренной или невольной фальсификации при сканировании данных.

B. D. Wright & J. M. Linacre определили RM как процесс сравнения результатов испытуемых на шкале натуральных логарифмов[9]. Математическую сторону и саму теорию Г. Раша успешно развивал D. Andrich[10]. Этими авторами было разработано несколько компьютерных программ, позволяющих проводить необходимые вычисления параметров заданий и испытуемых, а также давать компьютерную оценку пригодности данных для используемой модели.

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

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

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

9 **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 ;

10 **Andrich D.** Rasch Models for Measurement. In Series: Quantitative Applications in the Social Sciences. Sage University Paper. # 68.

Главное условие качества проведения RM — соответствие данных модели измерения. Если данные соответствуют модели, то в результате процесса измерения они представляются на интервальной шкале. При этом шкала RM устойчива к потере некоторых исходных данных.

RM является методом объективированного шкалирования данных. Иногда пишут «объективного» измерения, но сам Г. Раш эту лексику не поддерживал. Он считал нужным добавлять при этом слова «специфически объективного измерения», имея в виду неравенство понятия «объективное» в философии и того понятия объективности измерений, которое он мог достигать математическими методами. Автор этой статьи данную ситуацию выражает термином «объективированные» измерения[11].

Смысл специфически объективного измерения заключается в том, что сравнение мер трудности двух любых заданий теста рассматривается независимо от той или иной группы испытуемых. Симметричным образом, результаты сравнения любых двух взятых испытуемых не могут зависеть от систем заданий, образующих тест.

Метрическая система Г. Раша применима к исследованию любого интересующего свойства личности (если таковое существует устойчиво и наблюдаемо посредством системы эмпирических индикаторов), будь то знание, интеллект, социальные и психологические установки, отношение к чему-либо и пр.

Основу данной статьи составляют исходные понятия и идеи, которые привели к созданию нового, личностно-центрированного метода педагогических измерений. Хотя это открытие было сделано Г. Рашем в начале 50-х годов XX века, в литературе оно датируется обычно 1960 годом, моментом выхода из печати первого издания его главной книги[12]. Двадцать лет спустя она была издана в США[13]. Это и послужило главным толчком к признанию метрической системы Г. Раша в международном масштабе.

Основная проблема RM — это проведение качественных измерений. Оно возможно только тогда, когда есть чётко выраженная концепция измеряемого свойства (конструкт), подобрано нужное содержание теста, сформулированы задания в наиболее подходящей для данного содержания (вида знаний) тестовой форме.

Главная формула RM

С самого начала пятидесятых годов Г. Раша привлекли к оценке интеллектуальных способностей призывников датской армии. Он сразу же обратил внимание на основной недостаток применявшихся тогда психологами тестов, в которых были только задания с выбором одного правильного ответа. У этих заданий есть существенный дефект — высокая вероятность угадывания правильного ответа теми испытуемыми, кто не обладает интеллектуальными способностями. Поэтому первое, что он сделал, — решительно отказался от таких заданий и перешёл к применению заданий открытой формы[14]. Другие формы заданий в то время не были достаточно известны.

В основу разработки своей системы измерения Г. Раш положил метафору противоборства испытуемого с тестовым заданием. Если испытуемый имеет более чем достаточную подготовку для решения очередного задания, то он станет вероятным победителем противоборства, а потому получит, скорее всего, победный балл. Если подготовка недостаточна, побеждает, условно говоря, задание, и испытуемый получает 0 баллов.

Далее Г. Раш стал искать вероятностную математическую модель-функцию, позволяющую корректно описать свою метафору противоборства. После ряда проб он

[11] **Аванесов В.С.** Тесты в социологическом исследовании. М.: Наука, 1982 г.

[12] **Rasch G.** (1960). Probabilistic Models for Some Intelligence and Attainment Tests. Danish Institute for Educational Research, Copenhagen.

[13] Op. cit., reprinted by University of Chicago Press, 1980.

[14] **Rasch G.** (1960). Probabilistic Models for Some Intelligence and Attainment Tests. Danish Institute for Educational Research, Copenhagen. Reprinted by University of Chicago Press, 1980. То, что увидел он в начале пятидесятых годов прошлого века, не хотели видеть в России конца XX-го и не хотят видеть сейчас, в начале XXI-го века. В практике тестирования по-прежнему применяются преимущественно задания с выбором только одного правильного ответа.

остановился на формуле, представленной здесь в более удобной нотации Ф. Лорда.

$$P_j\{x_{ij} = 1 \mid \beta_j\} = \frac{\exp(\theta - \beta_j)}{1 + \exp(\theta - \beta_j)} \qquad (1)$$

В наше время эту формулу часто записывают в строку:

$$P_j(\theta) = \{x_{ij} = 1 \mid \beta_j\} = \exp(\theta - \beta_j)/(1 + \exp(\theta - \beta_j)$$

С этой модели начался подлинный успех Г. Раша, заметно усилился прогресс в разработке педагогических измерений и зарубежных образовательных систем.

Основные цели RM

Любое измерение начинается с общественно одобряемых целей и задач. Задания становятся операциональным определением измеряемого свойства тогда, когда их содержание и формы соответствуют открыто объявленным целям тестирования.

Г. Раш представлял главную цель измерения интересующего свойства личности на латентной переменной величине как относительно точное *позиционирование* каждого испытуемого на основе тестового балла. Этот балл получается на переменной величине, отображающей в количественном виде измеряемое свойство. Обычно принимается простая логика: чем выше тестовый балл, тем больше выражено у испытуемого интересующее свойство личности. Для определения переменной величины необходимы задания, подходящие для измерения данного свойства, а также значения исходных баллов испытуемых.

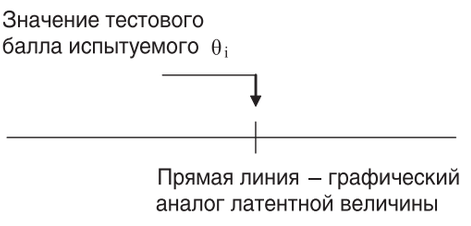


Рис. 1. Пример позиционирования испытуемого с номером i на латентной переменной величине «Уровень подготовленности испытуемых»

В наглядной форме эта идея была хорошо представлена в классической работе Б. Райта и М. Стоуна[15]. Рис. 1 из упомянутой книги даёт пример геометризации понятия «переменная величина» и результата измерения как точки на числовой оси.

Вторая цель RM — провести шкалирование уровня трудности заданий. Соответственно, возникла идея позиционирования заданий на латентной переменной величине «Трудность заданий». Реализация этой цели представлена на рис. 2.

Из всех известных моделей педагогических измерений модель Раша считается самой простой. Она требует информации о значениях только двух параметров: уровня подготовленности испытуемых и уровня трудности заданий.

Общие задачи RM

Помимо целей в RM можно выделить, по меньшей мере, два класса задач: общие и специфические.

Общие задачи педагогических измерений возникают при разработке теста по любой теории педагогических измерений, и в том числе по RM. *Специфические* задачи присущи преимущественно для RM.

Общие задачи педагогических измерений, используемые в RM, предшествуют специфическим. А потому без решения первых нет и качественного решения вторых.

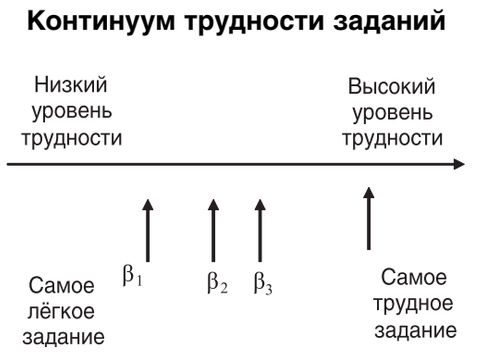


Рис. 2. Пример позиционирования четырёх заданий теста по уровню их трудности на латентной переменной величине «Трудность заданий теста»

[15] **Wright B.D., Stone M.H. Best Test Design.** Chicago. MESA Press, 1979.

1. *Композиция тестовых заданий.* В истории RM эта задача не ставилась в том виде, в каком ставится здесь. В этом легко убедиться, заглянув в работы зарубежных классиков. В их трудах RM — проблема математико-статистическая, программно-вычислительная и технологическая. Но значит ли это, что задача композиции тестовых заданий не существенна для RM?

Композиция тестовых заданий существенна для RM в той мере, в какой само это измерение зависит от качества формулирования заданий. Если содержание задания выражено некорректно или неясно либо представлено в неподходящей форме, то математическая теория не сможет сделать такие дефектные задания достойными. Но это и не её предмет. Вопросы содержания и формы заданий являются предметом педагогической теории измерений.

На сегодняшний день традиции RM таковы, что вопросы композиции считаются хотя и важными, но внешними для этой теории. С этим можно согласиться только при условии признания взаимосвязи педагогической теории и RM. На данный момент нет, однако, ни международного признания педагогической теории измерений, ни идеи взаимосвязи этой тории с RM. Существенное место в педагогической теории уделено вопросам композиции тестовых заданий[16].

Композиция определяется как форма деятельности педагога-творца, стремящегося создать задания, отвечающие требованиям современных образовательных технологий и педагогических измерений. Само слово «композиция» означает произведение, структуру, состав, а также соединение и взаимное расположение частей целого. Применительно к нашему предмету, целым является тест, частью целого — тестовое задание. Самое главное в композиции — умелое соединение формы и содержания заданий. Качественная композиция тестовых заданий — условие необходимое, но не достаточное для успешного проведения RM.

В науке и искусстве композицией называют состав и расположение частей целого, удовлетворяющих следующим условиям:
• ни одна часть целого не может быть изъята или заменена без ущерба для целого;
• части не могут меняться местами без ущерба для целого;
• ни один новый элемент не может быть присоединён к целому без ущерба для целого[17].

Успех в композиции заданий, как и в создании произведений искусства, зависит не только от оригинальности содержания, но и от мастерского владения формой. Успешная композиция может обладать свойствами эстетичности, эффективности, устойчивости и полезности.

Композиция тестовых заданий рассматривается не только как форма деятельности, но и как результат, получаемый в правильно организованном тестовом процессе. Цель композиции — создание таких заданий, которые можно было бы включить в тест, использовать в автоматизированных системах контроля и самоконтроля знаний, а также для организации самостоятельной работы учащихся.

В учебном процессе основная цель композиции — создание новых заданий, помогающих студентам (школьникам) обучаться и развиваться с использованием образовательных технологий.

Главная причина некачественности большинства имеющихся тестов коренится в игнорировании требований композиции тестовых заданий.

2. *Эмпирическая апробация заданий проектируемого теста*. Далее проводится апробация этих заданий на достаточной выборке испытуемых. Тест представляет собой результат умелого соединения теоретических концепций интересующего свойства личности и эмпирической проверки качества заданий. Эмпирическая апробация заданий проводится на типичной выборке испытуемых, очень похожей на так называемую целевую группу (target group). Это множество испытуемых, для которых разрабатывается тест. По итогам апробации строится матрица тестовых результатов, представленная в табл. 1.

[16] **Аванесов В.С.** Композиция тестовых заданий. 3 изд. М.: Центр тестирования. 2002.

[17] Проблемы композиции. Сб. науч. тр. М.: 1999. Под общ. ред. В.В. Ванслова. М.: НИИ Акад. художеств.

В этой матрице проводится редактирование, в процессе которого удаляются так называемые «экстремальные» задания и экстремальные испытуемые. В RM задания называются экстремальными в двух случаях: когда нет ни одного правильного ответа и когда, наоборот, все ответы правильные. Аналогично, из матрицы удаляются баллы так называемых экстремальных испытуемых, не имеющих ни одного правильного ответа, и испытуемых, ответивших на все задания правильно. Тем самым признаётся, что данным тестом уровень подготовленности экстремальных испытуемых определить невозможно.

3. *Дистракторный анализ заданий.* Дистракторами называют неправильные, но правдоподобные ответы в заданиях с выбором одного или нескольких правильных ответов. Дистракторный анализ проводится как в рамках общих методов педагогических измерений, так и в рамках RM. Без проведения дистракторного анализа тестов не бывает.

Общий дистракторный анализ сводится обычно к расчёту процентов выбора испытуемыми каждого ответа в каждом задании. В итоге появляются три группы дистракторов.

Первая группа — это те дистракторы, которые никто не выбирает или выбирают очень редко. Такой результат означает неудачу разработчика заданий, так как дистрактор не привлёк к себе внимания слабо подготовленных испытуемых. Дистрактор, который не оказался таковым фактически, удаляется из задания как не соответствующий требованиям композиции тестового задания. Вместо неудачного дистрактора подбирают другой. И снова потребуется эмпирическая проверка и проведение процентного анализа. Нижней границей приемлемости дистрактора можно считать 5%. Дистрактор, привлекающий к себе менее пяти процентов ответов неподготовленных испытуемых, считается неудачным.

Вторая группа — так называемые «работающие» дистракторы. Каждый из них привлекает внимание испытуемых, успешно отвлекает слабо подготовленных от правильного ответа.

Третья группа — это чрезмерно привлекательные дистракторы. Так, в задании

1. К. МАРКС РОДИЛСЯ В ГОРОДЕ
1) Трир
2) Берлин
3) Мюнхен
4) Карлмарксштадт
5) Франкфурт-на-Майне

испытуемые, не знающие правильный ответ, нередко выбирают четвёртый ответ, предполагая, что именно в честь данного события в бывшей ГДР и был назван город Карлмарксштадт.

Дистракторный анализ проводится в RM, а также в математической теории педагогических измерений (МТИ). Методика проведения такого анализ изложена в статье автора[18]. Обе теории — RM и МТИ — рассматриваются отдельно ввиду наличия у них существенных различий, несмотря на совпадающую формулу 1. Другие авторы, преимущественно российские, считают, что это одна теория. Нередко «объединённую» таким образом «теорию» называют «современной». Такое название в этой статье не поддерживается.

Специфические задачи RM

Другую часть задач можно назвать *специфическими* для RM.

1. *Расчёт вероятности правильного ответа испытуемых на задание теста.* Это задача вычислительного толка. Она решается посредством компьютерных программ при разработке тестов по системе RM и при применении математической теории педагогических измерений[19]. Вероятность правильного ответа каждого испытуемого на каждое задание в любой заданной точке θ можно определить посредством формулы 1. По итогам вычислений для каждой точки уровня подготовленности строится график задания теста.

[18] **Аванесов В.С.** Проблема эффективности педагогических измерений // Педагогические измерения. 2008. № 4. С. 3–24.

[19] Автор этой статьи считает эти две теории различающимися, в то время как другие авторы считают, что это одна теория. См. напр. Acton, G.S. What is Good About Rasch Measurement? Он пишет на стр. 902: «Rasch model is a one-parameter logistic model within item response theory». Rasch Measurement Transactions, 16, 902–903.

В модели Г. Раша принимается, что вероятность правильного ответа испытуемого на задание теста зависит только от двух показателей — уровня подготовленности испытуемого и уровня трудности задания. Чем больше разность $\theta_i - \beta_j$, тем больше вероятность правильного ответа испытуемого с номером i на задание с номером j. Эта закономерность графически выражена в работе B.D. Wright and M.D. Stone[20] (см. рис. 3).

Если испытуемый знает больше, чем того требует задание, значение разности больше, а потому большей чем 0,5 становится и вероятность правильного ответа, что видно из соответствующего графика на рис. 3. При любых значениях θ_i и β_j значения вероятности правильного ответа испытуемых с различной подготовкой на задания различного уровня трудности остаются в пределах от нуля до единицы, что достигается удачной структурой формулы 1.

Численный пример расчёта вероятности правильного ответа читатель найдёт в статье B.D. Wright[21]. Этот пример приводится также и в статье автора[22].

2. *Трансформация результатов тестирования.* RM — это метод трансформации данных тестирования. Процесс трансформации тестовых результатов делится на две части и проходит в два этапа. Первая часть процесса на английском языке называется Item calibration. На русский язык иногда это словосочетание переводят как «калибровка» или «калибрование» заданий. Автор статьи предлагает другой вариант: *шкалирование заданий по уровню их трудности*. Результатом процесса трансформации исходных баллов тестирования являются шкала исходных значений трудности заданий проектируемого теста. Эти значения представлены в строке $\ln q_j/p_j$ табл. 1.

Вторая часть процесса трансформации данных — это получаемая в RM шкала исходного уровня подготовленности испытуемых. Обе части процесса вычисления автор статьи называет измерением уровня подготовленности испытуемых и шкалированием заданий по уровню их трудности.

Здесь главное — трансформация исходных тестовых баллов в шкалу натуральных логарифмов, после чего, собственно, и появляется измерение. До начала процесса логарифмического преобразования исходные баллы тестирования не рассматриваются как результаты измерения[23].

[20] **Wright B.D. and Stone M.D.** Best Test Design. Chicago. MESA Press. 1979.

[21] **Wright B.D.** Solving measurement problems with the Rasch model. Journal of Educational Measurement 14 (2) pp. 97–116, Summer 1977. http://www.rasch.org/memo42.htm .

[22] **Аванесов В.С.** Проблема объективности педагогических измерений // Педагогические измерения. 2008. № 3. С. 3–40.

[23] См. подробнее на эту тему: **Аванесов В.С.** Являются ли КИМы ЕГЭ методом педагогических измерений? // Педагогические из-ния. 2009. № 1. С. 3–26.

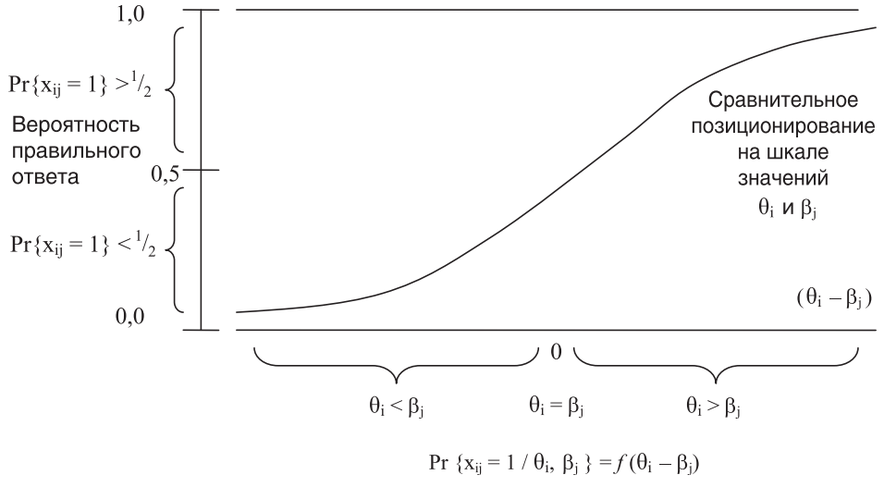


Рис. 3. График зависимости вероятности правильного ответа от разности между уровнем подготовленности испытуемых и уровнем трудности заданий

В методе Г. Раша исходные значения тестовых баллов трансформируются в исходные же логиты уровня подготовленности испытуемых. Учебный пример такого рода трансформации результатов представлен справа и внизу табл. 1.

Г. Раш отошёл от упрощённых оценок так называемого «уровня усвоения учебного материала», которые часто применяются при мониторинге в российских школах. Это процент правильных ответов на задания. Процент получается умножением долей правильных ответов испытуемых столбца p_i в правой стороне табл. 1 на сто. Появляется процентная мера усвоения каждого испытуемого (здесь не представлена).

Вместо этой меры Г. Раш предложил в правой стороне таблицы брать для испытуемых отношение $\ln p_i/q_i$, а в нижней её части — отношение $\ln q_j/p_j$. Первое отношение можно назвать логарифмической оценкой исходного уровня подготовленности (θ_i), второе — логарифмической оценкой исходной меры трудности задания β_j.

Тем самым Г. Раш сделал решающий шаг. Он ввёл общую логарифмическую меру измерения уровня подготовленности и уровня трудности задания, названную им, соответственно, логитом уровня подготовленности испытуемых и логитом трудности заданий.

Значения исходных логитов представлены в табл. 1.

Далее проводится второй этап шкалирования значений уровня трудности заданий и уровня подготовленности испытуемых: стандартизуются шкалы исходных логитов сопоставимыми значениями средних арифметических и стандартных отклонений. Только в этом случае возникает полная соизмеримость значений обеих переменных величин — уровня подготовленности испытуемых и уровня трудности заданий.

Таблица 1

Учебный пример таблицы тестовых результатов

№	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	Y_i	p_i	q_i	p_i/q_i	$\ln p_i/q_i$
1.	1	1	1	0	1	1	1	1	1	1	9	.90	.10	9	2.20
2.	1	1	0	1	1	1	1	1	1	0	8	.80	.20	4	1.39
3.	1	1	1	1	0	1	1	0	1	0	7	.70	.30	2.33	.85
4.	1	1	1	1	0	1	0	1	0	0	6	.60	.40	1.50	.40
5.	1	1	1	1	1	1	0	0	0	0	6	.60	.40	1.50	.40
6.	1	1	1	1	0	0	1	0	0	0	5	.50	.50	1.00	0
7.	1	1	0	1	1	0	1	0	0	0	5	.50	.50	1.00	0
8.	1	1	1	1	1	0	0	0	0	0	5	.50	.50	1.00	0
9.	1	0	1	0	1	1	0	0	0	0	4	.40	.60	.66	-.42
10.	0	1	1	0	0	0	0	1	0	1	4	.40	.60	.66	-.42
11.	1	1	1	0	0	0	0	0	0	0	3	.30	.70	.43	-.84
12.	1	1	0	0	0	0	0	0	0	0	2	.20	.80	.25	-1.39
13.	1	0	0	0	0	0	0	0	0	0	1	.10	.90	.11	-2.21
R_j	12	11	9	7	6	6	5	4	3	2	65				
W_j	1	2	4	6	7	7	8	9	10	11					
p_j	.923	.846	.692	.538	.462	.462	.385	.308	.231	.154	5				
q_j	.077	.154	.308	.462	.538	.538	.615	.692	.769	.846					
p_jq_j	.071	.130	.213	.248	.248	.248	.236	.213	.178	.130					
q_j/p_j	.083	.182	.445	.859	1.164	1.164	1.597	2.246	3.329	5.493					
$\ln q_j/p_j$	-2.489	-1.704	-.810	-.152	.152	.152	.468	.809	1.202	1.703					

В этой матрице рассчитывают:
• p_j — долю правильных ответов испытуемого i по всем заданиям теста;
• q_i — долю неправильных ответов того же испытуемого i по всем заданиям теста;
• p_i/q_i — потенциал подготовленности испытуемого i;
• ln p_i/q_i G. Rasch называет логитом подготовленности[24].
• ln q_j/p_j им же названа логитом трудности задания.

3. *Равномерность возрастания меры трудности заданий*. Решение этой задачи находится в соответствии с данным выше определением теста как системы заданий *равномерно* возрастающей трудности. Раньше этот уточняющий момент в определении теста не делался. В итоге задания некоторых так называемых «тестов» подбирались иногда с заметными «провалами» между заданиями, что сильно ухудшало метрические свойства метода — заметно снижалась точность измерений и дифференцирующая способность тестовых результатов. Можно с сожалением отметить, что ряд российских практиков и авторов этот критерий либо не признают, либо обходят стороной как несущественный. Например, вместо понятия «система заданий» используют словосочетание «совокупность» или «множество заданий», как будто между ними нет разницы.

B. D. Wright и M. D. Stone обратили внимание на важный системный фактор распределения заданий теста по уровню трудности. В педагогических измерениях по модели Г. Раша графики заданий теста отличаются только значениями проекций точек перегиба функций на ось абсцисс; чем труднее задание, тем правее располагается график относительно оси абсцисс. Трудность рядом стоящих заданий теста не должна отличаться более чем на 0,5 логита[25]. Иначе на шкале образуются провалы. Расстояние в 0,5 логита — это довольно либеральное требование. Лучше, когда расстояние между заданиями бывает не более чем 0,25 логита трудности. Это требование можно назвать условием достаточной плотности расположения числа заданий на шкале.

Обоснование вывода о равномерности расположения заданий теста, а следовательно, и пригодности предлагаемой системы заданий для измерения уровня подготовленности испытуемых, на данной переменной величине нуждается в эмпирических фактах. В качестве таких фактов в RM используется построение на одной плоскости графиков всех заданий теста. Для заданий учебной матрицы табл. 1 графики представлены на рис. 4.

[24] **Rasch,G.** On General Laws and the Meaning of Measurement in Psychology /In Proceedings of the Fourth Berkley Symposium on Mathematical Statistics and Probability. Berkley: Univ. of California Press, 1961; **Rasch, G.** On Specific Objectivity: An Attempt of Formalizing the Request for Generality and Validity of Scientific Statements / Danish Yearbook of Philosophy. 1977, v. 14, p. 58–94, Munksgaard, Copenhagen; **Rasch, G.** Probabilistic Models for Some Intelligence and Attainment Tests. With a Foreword and Afteword by B.D. Wright. The Univ. of Chicago Press. Chicago & London, 1980.

[25] Исходное значение логита трудности задания находится из выражения ln q_j/p_j, где q_j является долей неправильных ответов испытуемых на задании теста под номером j, а p_j — это доля правильных ответов испытуемых на то же самое задание под номером j.

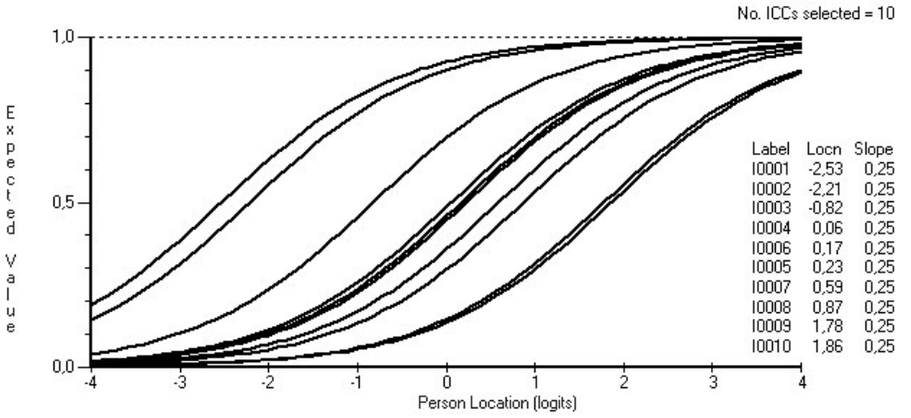


Рис. 4. Графики всех заданий, построенных по данным учебной матрицы табл. 1

Из рисунка 4 видно, что для достижения качественных измерений в учебном тесте табл. 1 не хватает заданий соответствующего уровня трудности между вторым и третьим, третьим и четвёртым, восьмым и девятым заданиями. Графики всех заданий имеют одну и ту же крутизну, это означает, что их дифференцирующая способность принимается равной. Хотя при использовании других моделей выявляются существенные отличия по крутизне заданий, в RM, тем не менее, значение параметра крутизны каждого задания принимается равным единице. Естественно поставить вопрос — почему в RM вводится столь странная унификация заданий по уровню их дифференцирующей способности?

Г. Раш полагал, что только в таком случае вероятность правильного ответа испытуемого будет зависеть только от значения θ и от меры трудности задания и не будет зависеть от других свойств заданий и от других факторов. С этим утверждением мало кто соглашался, но результат превзошёл ожидания. Модель оказалась работоспособной.

4. *Соответствие тестового задания модели измерения.* На рис. 5 представлен график первого, наиболее лёгкого задания учебной матрицы табл. 1. Видно вполне приемлемое совпадение теоретических и эмпирических точек; это доли правильных ответов слабой, средней и сильной части группы испытуемых. Об этом же свидетельствует и низкое значение отклонений эмпирических точек от графика (residual = –0,002).

В классической (статистической) теории педагогических измерений это задание было бы однозначно отбраковано по критерию очень низкой корреляции ответов испытуемых на это задание с суммой баллов проектируемого теста ($r_{1\,t'}$= 0,132).

Таблица 2

Коэффициенты корреляции ответов на задания учебного теста табл. 1 с суммой баллов

Номера заданий	Значения коэф. корр.
1	0,132
2	0,488
3	0,305
4	0,495
5	0,495
6	0,707
7	0,652
8	0,534
9	0,752
10	0,293

Теперь полезно посмотреть на пример плохого соответствия задания № 3 учебной матрицы табл. 1 требованиям модели Г. Раша. Соотношение эмпирических точек и графика задания на рис. 6. показывает, что это задание не годится ни для оценки испытуемых низкого уровня подготовленности, ни для оценки испытуемых высокого уровня подготовленности. Слабо подготовленные испытуемые отвечает на него лучше, чем прогнозирует вероятностная модель, а хорошо подготовленные отвечают хуже, чем прогнозируется

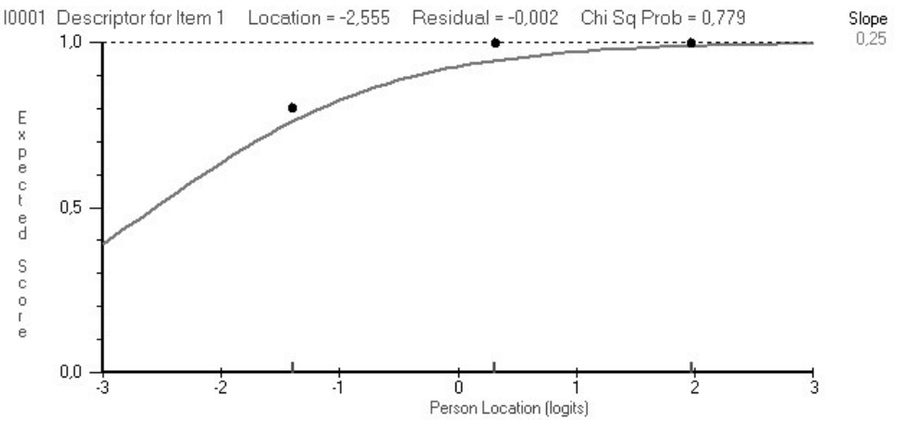


Рис. 5. График задания, совместимого с моделью Г. Раша

по модели. Это задание, скорее всего, имеет дефект в своей композиции; его правильно понимают только испытуемые среднего уровня подготовленности.

О неадекватности задания свидетельствует относительно большое значение отклонений точек от графика (residual = 0,918). Поэтому его нельзя отнести к числу соответствующих модели Г. Раша, даже если по минимальному значению критерия пригодности (хи-квадрат) оно считается подходящим. Такое задание может только испортить качественный тест.

Полезно заметить, что хотя коэффициент корреляции этого задания выше ($r_{3\,t}$ = 0,305), чем у первого, его соответствие модели Г. Раша оказалось ниже. В классической теории педагогических измерений это задание могло бы пройти в число тестовых, если бы там использовался обычный порог значений r > 0,300.

5. *Совместимость тестовых заданий.* Понятие «совместимость тестовых заданий» выражает идею возможности создать тест из совместимых между собой заданий. Наиболее часто применяемым показателем совместимости отдельного задания и общей совместимости всех заданий, образующих тест как систему заданий возрастающей трудности, является значение хи-квадрат, которое для случая учебной матрицы в табл. 1 равно 0,789. Чем больше значение хи-квадрат, делённое на число так называемых «степеней свободы», тем лучше совместимость.

В данном случае совместимость по установившейся практике считается более чем удовлетворительной. *Хорошая* совместимость появляется тогда, когда нет проблемных заданий. Совместимость становится *отличной*, если все задания проектируемого теста задания не только свободны от дефектов, но и наилучшим образом соответствуют требованиям модели Г. Раша.

6. *Соответствие меры трудности разрабатываемого теста уровню подготовленности учащихся.* Для проведения качественного педагогического измерения уровень трудности теста должен соответствовать уровню подготовленности испытуемых. Эта простая истина была известна ещё в статистической теории педагогических и психологических измерений. Её можно теперь увидеть посредством применения компьютерных программ по вычислению и наложению двух гистограмм — результатов испытуемых и мер трудности заданий. Вверху располагается гистограмма результатов испытуемых, внизу — гистограмма распределения заданий — от лёгкого к трудному.

7. *Достаточность вариации и размаха заданий по уровню их трудности.* В тесте должны быть задания равномерно возрастающей трудности. Это правило позволяет обеспечить варьирование заданий по уровню трудности. Разность между значением самого трудного и самого лёгкого задания называется размахом. В RM в качестве нормы принимаются пределы вариации значений трудности заданий в логитах от –3 до +3. Соответственно, приемлемая мера размаха равна шести логитам.

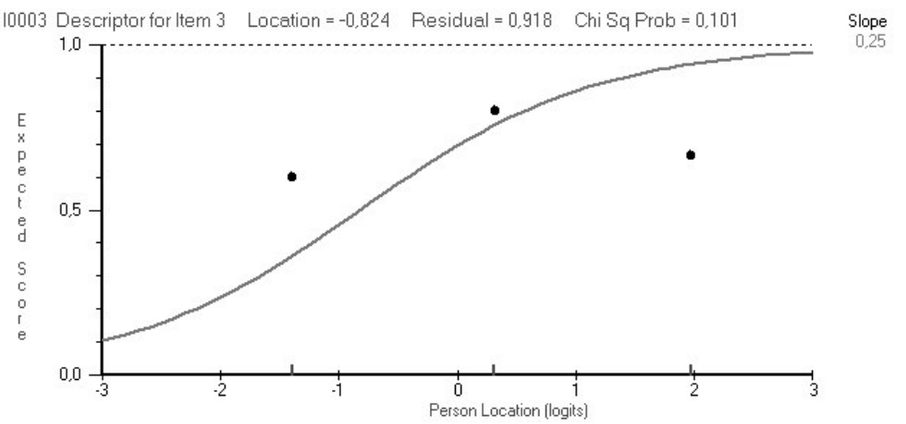
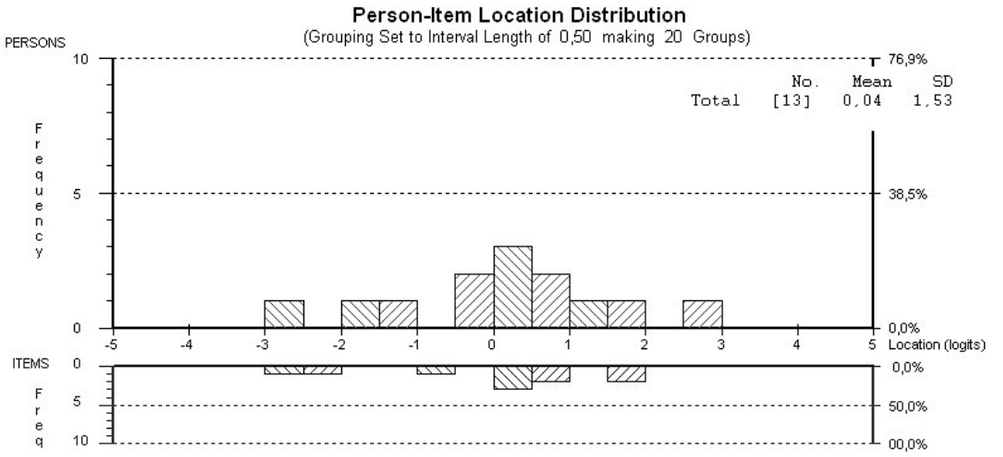


Рис. 6. График задания № 3, не соответствующего модели Г. Раша



Парадокс RM

Из общих соображений известно, что любая жёстко ограниченная система неизбежно порождает ряд противоречий. Одним из таких ограничений является требование метрической системы Г. Раша: все задания теста должны иметь одинаковую крутизну своих графиков, несмотря на фактические различия по их дифференцирующей способности. Отмеченное ограничение даёт начало парадоксу, который полезно назвать именем Г. Раша: чем большей, после некоторого уровня, дифференцирующей способностью обладает задание, тем больше оно противоречит системной идее RM. Следовательно, возрастает риск удаления из теста самых лучших его заданий!

Хорошо известно, что в одном тесте нет и не может быть одинаковых заданий: они все отличаются хотя бы по одной из характеристик, среди которых наиболее главная для теста как формальной системы — мера трудности заданий. Нет метрического смысла иметь в тесте два и более заданий одинакового уровня трудности.

Посмотрим пример задания № 7 на рис. 7. С точки зрения классической теории педагогических измерений, это задание обладает относительно высокой дифференцирующей способностью. Об этом свидетельствует значение коэффициента корреляции ответов испытуемых на это задание с суммой баллов по всему проектируемому тесту ($r_{7\,t}$ = 0,651, см. табл. 2.). Слабо

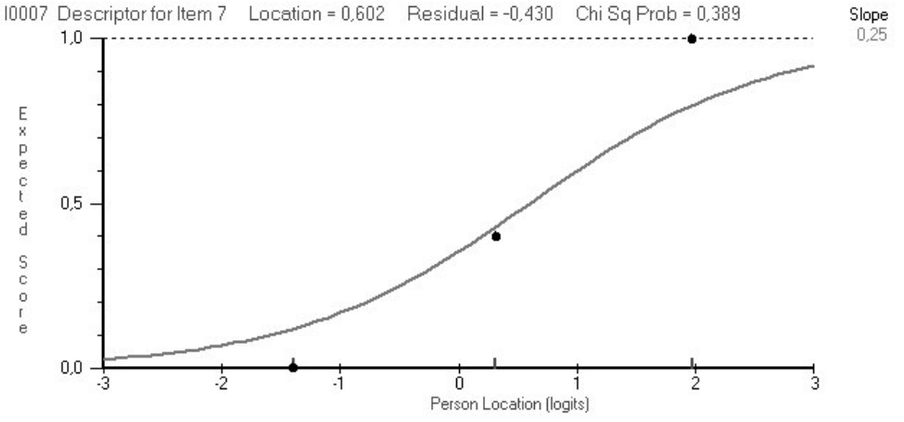


Рис. 7. График седьмого задания

подготовленные испытуемые отвечают хуже, чем это прогнозируется моделью Г. Раша, а хорошо подготовленные — лучше.

Если смотреть на такое задание с точки зрения требования одинаковой крутизны для всех графиков заданий теста, то получается парадокс: чем лучше задание, тем оно хуже с точки зрения требования RM[26].

И действительно, применение, например, двухпараметрической и трёхпараметрической модели МТИ позволило бы добиться лучшего совмещения эмпирических и теоретических (прогностических, по модели) точек.

Ещё один пример парадокса даёт задание № 9 на рис. 8. Оно имеет наибольшую дифференцирующую способность, если об этом судить по значению коэффициента корреляции ответов на задание с суммой баллов испытуемых ($r_{9\,t} = 0{,}752$). То же подтверждает расчёт коэффициента крутизны графика этого задания в математической теории педагогических измерений (МТИ[27], здесь не приводится). На задание № 9 правильно ответы дают только отлично и хорошо подготовленные испытуемые.

Из-за отмеченного парадокса и это задание придётся удалить из проектируемого теста. Здесь имеет место явление, называемое по-английски Overfit (на русский язык можно перевести примерно так: задание настолько хорошее, что в это верится с трудом).

Уровни RM

В RM можно выделить практику, теорию, методику, технологию и методологию.

Главные направления развития теории — это формирование собственного языка RM, разработка моделей, проверка пригодности заданий по статистическим критериям, разработка вычислительных методов RM.

Методика RM касается вопросов алгоритмизации применения различных методов в процессе педагогического измерения.

Методология RM имеет своим предметом развитие соответствующей теории и преобразование (повышение эффективности) практики педагогических измерений.

Причины отставания России в вопросах применения RM

В России педагогические измерения по модели датского математика Г. Раша не получили заметного распространения, хотя их значение и роль за последние десятилетия выросли во всём мире. В настоящее время

[26] My best items don't fit! Rasch Measurement Transactions, 2004, 18:3, p. 992. См. также G. Masters (1988). «Item discrimination: when more is worse», Journal of Educational Measurement, 25:1, 15–29, and www.rasch.org/rmt/rmt72f.htm — RMT 7:2, 289.

[27] **Аванесов В.С.** Истоки и основные понятия математической теории педагогических измерений (Item Response Theory) // Педагогические измерения. 2007. № 3. С. 3–36.

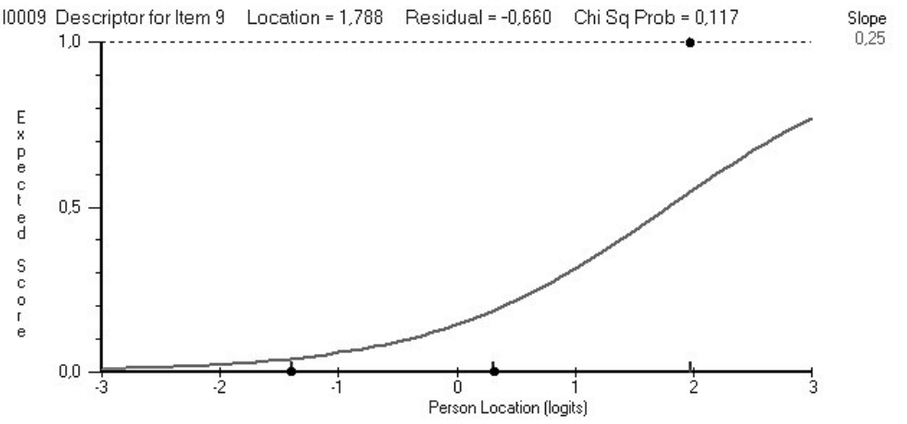


Рис. 8. Пример графика слишком хорошего задания, обладающего дифференцирующей способностью более высокой, чем это требуется по модели Г. Раша

RM применяется для качественного шкалирования интересующих объектов и показателей состояния этих объектов в таких сферах, как образование, медицина, социология, психология и др.

Можно выделить три причины неадекватности педагогических измерений, основанных на модели Г. Раша, требованиям времени.

Первая и главная причина — это сверхизбыточное государственное вмешательство в сферу, сопряжённую с педагогическими измерениями. Уже два десятилетия в России вместо педагогических измерений государством и его органами активно навязываются некачественные методы и затратные бюрократические схемы, так называемые ЕГЭ, КИМы, АПИМы, ОСОКО, уводящие научно-педагогическую общественность в сторону от разработки подлинных научных методов педагогических измерений.

Вторая причина — слабая информированность относительно сути и возможностей (RM). Результаты неинформированности проявляются в малом количестве случаев применения RM и, одновременно, в заметных масштабах ухудшения качества образования в стране. Связь качества образования с RM может показаться особо спорной и даже непонятной. Но если представить качество образования как одно из следствий недостатков используемых в практике учебных заданий, то становится понятным, что умелое управление собственной учебной деятельностью учащихся и студентов невозможно без качественной и объективной информации о сравнительной мере трудности каждого задания и о возможности задания быть включённым в тест.

Третья причина слабого применения RM в России — это отсутствие приемлемого педагогического языка RM, необходимых изданий, в том числе доступных большому числу начинающих исследователей. Тексты по этой проблеме написаны в основном математиками для математиков. Там используется язык теории вероятности и статистики, не очень понятный педагогам иных специальностей, неадекватные переводы иностранной лексики вроде «характеристических кривых заданий» (item characteristic curves).

На этом фоне предложенные автором данной статьи педагогическая теория педагогических измерений[28] и язык этой теории[29] в России остаются не востребованными. Но всё может существенно измениться с началом действительной модернизации образовательной деятельности. Однако когда начнётся подлинная модернизация российского образования, понимаемая как приведение к современности, — неведомо пока никому. ❑

[28] **Аванесов В.С.** Основы педагогической теории измерений // Педагогические измерения. 2004. №1. С. 15–21.

[29] **Аванесов В.С.** Определение исходных понятий теории педагогических измерений // Педагогические измерения. 2005. № 2. С. 6–24;
Аванесов В.С. Язык теории педагогических измерений // Педагогические измерения. 2009. № 2. С. 29–60.