ИСТОКИ И ОСНОВНЫЕ ПОНЯТИЯ МАТЕМАТИЧЕСКОЙ ТЕОРИИ ИЗМЕРЕНИЙ (ITEM RESPONSE THEORY)

Статья вторая

Вадим Аванесов
testolog@mail.ru

В первой статье[1] название Item Response Theory (IRT) было переведено на русский язык как «математическая теория педагогических измерений». Но эта теория применяется не только в педагогических, она широко используется также в психологических, социологических, медицинских измерениях. Для того, чтобы снять возникшую неточность, во второй статье название было поправлено. Оно потеряло неспецифическое слово «педагогических» и, таким образом, стало короче. Теперь IRT переводится на русский язык как математическая теория измерений (МТИ).
Во второй статье исследуются истоки МТИ, углублены и расширены ранее сформулированные понятия, представлены определения других основных понятий МТИ (IRT).

Спорный научный статус МТИ (IRT)

В первой статье уже отмечалось, что на русском языке IRT (МТИ) часто характеризуется как «современная» теория педагогических измерений. Этим допускаются сразу три ошибки.

1

Аванесов В.С. Item Response Theory: Основные понятия и положения. ПИ № 2, 2007 г. С. 3–28.

Во-первых, другие теории как бы автоматически переводятся в разряд несовременных, что, конечно же, порождает множество ошибок и отрицательных последствий в теории и в практике.

Во-вторых, IRT — теория вовсе не педагогическая, а математико-статистическая. Терминология IRT, язык и научный аппарат совершенно точно не входят в состав педагогической науки. Известные зарубежные авторы R.K. Hambleton & R.W. Jones[2], E.H. Haertal[3], W.M. Yen[4] и многие другие, если не все, также считают IRT не педагогической теорией. Они называют её статистической теорией. И действительно, IRT опирается на язык математики, статистики, в ней широко представлены знания по теории вероятностей, математической статистике и вычислительным методам математики. Всё это, вместе взятое, относится к прикладной математике и к математической статистике.

В современных науках немало таких теорий. Они используются для формального обоснования качества педагогических, психологических, социологических и прочих измерений. Учитывая общность языка математики, теории вероятности и математической статистики, а также факт широкого применения IRT в общественных науках, в данной статье в качестве основного названия на русском языке принято новое название IRT — «математическая теория измерений (МТИ)».

В-третьих, ещё одна ошибка заключается в том, что современными сейчас считаются[5] уже не одна, а все три основных теории, применяемые в педагогических и психологических измерениях: это статистическая (классическая) теория, МТИ (IRT) и расширенная статистическая теория (РСТ). Последняя по-английски называется Generalizability Theory (G-Theory)[6].

Вопрос понимания научного статус МТИ (IRT) усугубился попытками внедрения положений этой теории в систему российского образования, как это ни странно, не через науку, а через практику централизованного тестирования и ЕГЭ. А эта практика всегда держалась от педагогической науки очень далёко. И пока нет никаких признаков или попыток изменения этой нездоровой ситуации.

МТИ (IRT) совершенно определённо не занимается вопросами понятийного аппарата педагогических измерений, содержания и формы тестовых заданий. И не может этим заниматься, так как это собственный предмет другой ранее сформулированной теории — педагогической теории измере-

2 *Hambleton R.K., Jones R.W.* An NCME Instructional Module on COMPARISON OF CLASSICAL TEST THEORY AND ITEM RESPONSE THEORY AND THEIR APPLICATIONS TO TEST DEVELOPMENT. http://www.ncme.org/pubs/items.cfm

3 *Haertel, Edward H.* Problems in measurement: Item Response Theory. http://www.stanford.edu/group/mapss/syllabi/Educ353A-Psych249A.doc

4 Item response theory (IRT) is a family of statistical procedures for analyzing and describing test performance Yen W.M. (1992). Item response theory, In Alkin M. C. (Eds.), Encyclopedia of educational research. (vol. 2, pp. 657–667). NY: Maxwell Macmillan International.

5 *Sudweeks Richard.* Modern test theory is a composite of classical true-score theory, item response theory, and

ний (ПТИ)[7], включающей главные вопросы обеспечения качества педагогических измерений. И поскольку в МТИ (IRT) нет никакого педагогического содержания, её было бы правильнее рассматривать как теорию формальную, а не содержательную. Быть может, МТИ (IRT) когда-нибудь интегрируется в систему расширенного педагогического знания, но эффективно это может произойти через объединение с педагогической теорией измерений (ПТИ). Последняя намного ближе к педагогике, чем МТИ (IRT).

Педагогическая теория измерений является содержательно-педагогической, и только отчасти формальной педагогической теорией, имеющей своим предметом понятийный аппарат, форму и содержание тестовых заданий, а также вопросы разработки педагогических тестов, педагогической оценки уровня и структуры подготовленности испытуемых, проведения массового тестирования, сравнения и интерпретации результатов. Хотя объект ПТИ и МТИ совпадает — а это тестовый процесс — каждая из теорий имеет свой предмет, свой понятийный аппарат, свои методы обоснования качества педагогических измерений. И ни одна теория не в состоянии заменить другую.

Математическая теория измерений включает в себя также совокупность методов, позволяющих получить количественные оценки вероятности правильного ответа испытуемых на задания различного уровня трудности, уровня подготовленности испытуемых, меры трудности и различающей способности заданий, а также другие характеристики. Есть, правда, даже авторы, которые считают её просто теорией шкалирования данных. Тогда её можно определить как теорию и методику шкалирования уровней подготовленности испытуемых, а также уровней трудности и различающей способности тестовых заданий. Как и всякая развитая формальная теория, МТИ (IRT) имеет свою собственную систему понятий, опирающуюся главным образом на язык теории вероятностей, математики, вычислительных методов и статистики. В общем, на язык математики.

В фокусе исследования МТИ — тестовое задание и математическая функция, которая позволяет выразить вероятность правильного ответа испытуемых в зависимости от уровня их подготовленности. МТИ позволяет исследовать метрические свойства тестовых заданий, оценить их формальные свойства, пригодность для включения в тест, эффектив-

generalizability theory. http://education.byu.edu/ipt/courses/752.doc

6

Хои К. Суен, Пуи Ва Лей. Методологический анализ теорий педагогических измерений. Пер. с англ. // Педагогические измерения, № 1, 2007, С. 3–20.

7

Аванесов В.С. Основы педагогической теории измерений // Педагогические измерения, № 1, 2004, С. 15–21.

ность и качество тестовых заданий. Для этого используется вычисление стандартной ошибки измерения, значения хи-квадрат и определяется уровень достоверности получаемых выборочных статистик.

Вероятность правильного и неправильного ответа в МТИ рассматривается как функция от уровня подготовленности испытуемых и как функция от параметров заданий. В МТИ даётся решение и обратной задачи — определения меры правдоподобности оценок уровня подготовленности испытуемых как функции от теоретической вероятности наблюдаемых эмпирических результатов тестирования и от тех же параметров заданий. Наиболее правдоподобные значения принимаются в качестве оценок истинных значений (параметров) подготовленности испытуемых.

Для того, чтобы МТИ проявлялась как эффективная теория, она должна быть эмпирически верифицируемой, во всех своих многочисленных приложениях. Дело в том, что сама МТИ нередко используется в качестве методической гипотезы, в порядке апробации методов данной теории к результатам разрабатываемого или уже применяемого теста. Разумеется, речь может идти о гипотезе применимости МТИ для оценки тех или иных тестовых заданий.

Главная цель, сфера и главный смысл применения МТИ — научное исследование качества тестовых заданий. По большому счёту, к работе с нетестовыми заданиями МТИ недостаточно применима, а то и вовсе не применима. Хотя попытки применения делались много раз, в том числе и для ЕГЭ[8]. Но надежды математиков создать с помощью МТИ тесты из нетестовых заданий являются абсолютно тщетными. Это очень похоже на средневековую алхимию. Применение же МТИ для оценки качества нетестовых заданий порождает существенные системные противоречия в процессе измерения, особенно при интерпретации данных. Тесты можно создать только из тестовых заданий. Обходных путей в этом деле нет и не предвидится!

Известно, что попытки решения проблем одной науки средствами другой науки в принципе обречены на неудачу. Наблюдаемое во всём мире и особенно в России увлечение исключительно математической стороной обоснования качества тестов, в ущерб педагогической стороне — явление не новое и не исключительное. На ситуацию выхолащивания представителями одной науки содержания другой науки в своё время обратил внимание Гегель. Попытки математиков

8

Энтузиасты возлагали большие надежды на применение двухпараметрической модели IRT для улучшения качества т.н. КИМов. Однако при отсутствии должного внимания к формам заданий, общественного контроля за содержанием КИМов, при слабой проработанности понятийного аппарата и пренебрежительном отношении к тщательной разработке концептуально-теоретических вопросов научной структуры КИМов, а также при отсутствии научного проекта самого ЕГЭ улучшение невозможно. Не случайно ЕГЭ вызывает завышенную трату бюджетных средств, путаницу в умах учёных и народное недовольство. Особенно в вопросах перевода необъективных и ненадёжных баллов испытуемых из одной шкалы в другую. Недавно было заявлено, что в минобразования работают над вопросами перевода данных из шкалы в шкалу. Похоже, что ждать придётся ещё семь лет, потому что эта задача — не для чиновников.

задавать тон в философии он называл «варварским педантизмом или педантичным варварством, представленные во всей широте и со всей обстоятельностью, которые должны были привести к тому, чтобы геометрический метод лишился всякого доверия[9]».

Не случайно, например, в научной психологии, успешно применяющей математический аппарат уже примерно сто пятьдесят лет, было понято, что математика не может претендовать на полное и действительное решение проблем, принадлежащих психологии и другим наукам[10]. Есть надежда, что безусловное признание IRT в качестве формальной теории поможет снять необоснованные претензии некоторых представителей математико-вычислительного направления в нынешнем практическом тестировании на решение чуть ли не всех содержательных и общих теоретических задач педагогических измерений.

В нынешней практике педагогических измерений вопрос соотношения теорий становится одним из запутанных, и даже спекулятивных. В первой статье уже упоминались ранее дававшиеся эпитеты в адрес IRT: «научная и современная», чуть ли не единственная теория, способная решить все проблемы педагогических измерений; выходило, что все остальные теории отсталые и не нужные. На самом деле это не так. В педагогических измерениях требуется осмысленная концепция измерения, обоснование надёжности и валидности получаемых результатов, с точным указанием цели, для которой эти результаты адекватны[11]. Культурного педагогического измерения с амбивалентными целями, какие преследуются сейчас в ЕГЭ, не бывает.

Истоки МТИ

Гегель был прав, утверждая, что без истории нет теории. У каждой подлинно научной теории обычно бывает несколько источников и направлений развития. В случае с МТИ (IRT) часть таковых выделяется сравнительно легко. В зарубежной литературе можно насчитать десяток концептуальных источников и примерно два десятка фамилий авторов, внесших наибольший вклад в её развитие[12].

В этом разделе статьи внимание читателей обращается на семь источников.

Первый источник — это идея латентных (скрытых от непосредственного наблюдения) качеств личности. История возникновения таких идей прослеживается, начиная с трудов Платона. И хотя качество личности непосредственно

9

Гегель Г.В.Ф. Соч.: В 14 т. Т.XI. С. 364.

10

Психология и математика. М.: Наука, 1976.

11

Отсутствие валидности результатов – один из главных системных пороков т.н. КИМов, что вытекает из совокупности противоречивых целей и задач ЕГЭ. Подробнее см.: *Аванесов В.С.* Единый государственный экзамен в фокусе научного исследования // Педагогические измерения № 1, 2006 г. Единым госэкзаменом преследуется сразу несколько целей, а это путь к обнулению валидности. Что может быть следствием отсутствия действительно научной концепции ЕГЭ, если под таковой понимать форму организации научного знания, дающего исходное целостное представление о сущности структуре, функциях, целях и тенденциях развития, формах, методах и результатах измерения. Такой концепции не было, и нет до сих пор.

не наблюдаемо и не измеряемо, оно себя проявляет в идее понятийных и эмпирических индикаторов. В тестовой технологии положительный или отрицательный ответ испытуемого на каждое задание теста рассматривается как индикатор наличия или отсутствия у него интересующего латентного качества.

Как уже упоминалось в первой статье, почти любое интересующее латентное качество личности имеет общее название «ability». Дословный перевод как «способность» вызывает ошибки понимания. На русском языке, применительно к педагогическим измерениям, этому понятию лучше поставить в соответствие словосочетание «уровень подготовленности испытуемых».

Примеры распространённости идеи латентных качеств можно найти также в поэзии[13] и в художественной литературе[14].

Второй источник становления IRT — идея построения графических образов отдельных заданий теста на основе эмпирических данных (A. Binet & N. Simon[15], 1916, M.W. Richardson[16], 1936). Эти авторы первыми реально увидели, как могут выглядеть графики различных заданий, построенные на идее подбора функции для получаемых точек на плоскости.

Третий источник — это труды классика американской психометрики Л. Гутмана. В его представлении задания должны располагаться на той же числовой оси, на которой определяется уровень подготовленности испытуемых. Это стимулировало поиск такого методам шкалирования, который позволяет получить одну общую шкалу измерения как для интересующего свойства заданий, так и испытуемых. Такую шкалу позже получил Г. Раш на основе разработанной им теории. Этим открылась возможность численно сравнивать ранее несравнимые свойства личности и различных вещей. Некоторые несравнимые ранее понятия стали теперь сравнимыми.

В тестовой технологии начала XX века большое распространение получило простое решающее правило: всякий испытуемый за правильное выполнение задание получал один балл, за неправильное выполнение — ноль. После чего каждое задание теста стало исполнять роль очередного порога (threshold) возрастающей трудности, которые испытуемый старался преодолеть в процессе тестирования: чем больше правильных ответов, тем лучше. Если располагать задания по принципу возрастающей трудности, то это способствовало появлению высокого числа

12

History of item response theory (up to 1982). http://www.uic.edu/classes/ot/ot540/history.html

13

Милый друг, иль ты не видишь,
Что всё видимое нами,
Только отблеск, только тени
От не зримого очами.
В.С. Соловьев.
Русские поэты. Антология русской поэзии в 6 т. Москва: Детская лит-ра, 1996.
См. также http://www.litera.ru/stixiya/authors/solovev/milyj-drug-il.html

14

Самое важное – это то, что невидимо. «Главное — невидимо глазу», — повторил Маленький Принц ...
А. Сент-Экзюпери.
Маленький принц.
http://malenkiyprinc.narod.ru/
То, что полностью контролируемо, никогда не бывает вполне реальным, То, что реально, никогда не бывает вполне контролируемо.
В. Набоков.
Надо хотя бы иногда делать видимым то, что обычно невидимо. *Роуэн Уильямс, архиепископ Кентерберийский.*

баллов у хорошо подготовленных испытуемых. При ответах иногда возникали такие вектор-строки баллов испытуемых, в которых все нули следовали за всеми единицами. Такую вектор-строку можно назвать правильным профилем подготовленности испытуемого.

Редко у кого из испытуемых бывают правильные профили. Те профили, где наблюдаются одна или несколько инверсий, логично назвать неправильными профилями подготовленности личности. Иногда неправильные ответы даются на сравнительно лёгкие задания, а правильные — на трудные задания. Причин такого положения может быть несколько. Первая причина — это попытка угадывания правильного ответа в трудном задании, в случае использования заданий с выбором одного правильного ответа из числа предлагаемых на выбор.

В случае применения пяти ответов в каждом задании, из которых один правильный, а остальные неправильные, вероятность угадывания равна $1/5$. Это означает, что примерно пятая часть ответов на все задания теста может быть угадана. Вторая причина — отсутствие у учащихся системных знаний. Эти и другие причины приводят к тому, что большая часть профилей оказываются неправильными. Здесь, и таким образом, в частности, проявляют себя ошибки педагогического измерения.

В теории Л. Гутмана вероятность успешного выполнения для тех испытуемых, кто в состоянии выполнить задание, равна единице. Для тех, кто не в состоянии это сделать, вероятность равна нулю. На рис. 1 представлен графический образ задания среднего уровня трудности, которое безошибочно различает тех, кто знает, от тех, кто не знает. Это пример идеально функционирующего педагогического задания.

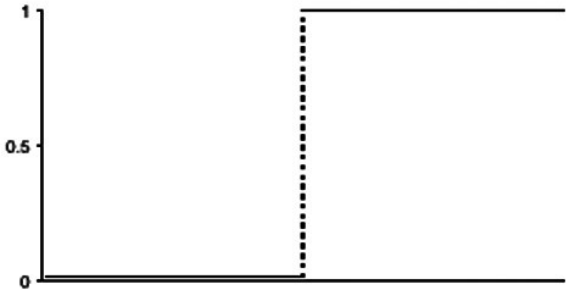

Рис. 1. Графический образ задания, вытекающий из теории Гутмана

Полезно выделить три важных условия получения идеальных заданий. Первое — содержание задания должно пониматься всеми испытуемыми, независимо от уровня

15 *Binet A., Simon T.H.* The Development of Intelligence in Young Children. Vineland, NJ: The Training School, 1916.

16 *Richardson M.W.* Notes on the Rationale of Item Analysis // Psychometrika, 1936, 1: 169–176.

их подготовленности. Иначе говоря, каждый должен понимать, о чём задание. Второе условие — форма заданий и инструкции к ней должны быть знакомыми. Третье условие идеальности задания — оно должно хорошо различать испытуемых ниже и выше точки трудности задания на оси абсцисс. Такие задания особенно привлекательны в случае, когда их много, и все они — непременно возрастающей трудности. Тогда на числовой оси можно располагать десятки заданий возрастающей трудности, вследствие чего возникает эффект идеального измерительного устройства, которое распределяет каждого испытуемого на числовой оси, в зависимости от числа правильно выполненных заданий[17] теста.

Результатом такого рода упорядочения испытуемых (по числу баллов) и расположения заданий по мере трудности получается своеобразная матрица тестовых результатов, элементы которой располагаются подобно прямоугольному треугольнику. На рис. 2 все правильные ответы располагаются в левом верхнем углу, все неправильные — в правом нижнем углу. Если при этом все профили испытуемых оказывались правильными, то такой вариант расположения результатов Л. Гутман называл шкалограммой.

1	1	1
1	1	0
1	0	0
0	0	0

Рис. 2. Пример т.н. шкалограммы

Позже требования Л. Гутмана к заданиям были названы детерминистической моделью измерения. У этой модели есть некоторые привлекательные свойства и одно, по меньшей мере, ограничение: все профили подготовленности должны быть правильными, что нереалистично. Правильность профиля педагогически можно истолковать так: испытуемый знает то, что знает (а это правильные ответы на сравнительно лёгкие задания) и не знает то, чего не знает — это неправильные ответы на трудные задания. Но для такого расположения испытуемых в матрице нужно было иметь и идеально упорядоченную, по мере возрастающей трудности, систему заданий.

Задания должны были поддаваться правильному решению со стороны тех, кто подготовлен, и не поддаваться правильному решению для неподготовленных испытуемых. Каждое задание

17 Получалось нечто похожее на русскую поговорку: всяк сверчок знай свой шесток!

должно чётко различать испытуемых на своём уровне трудности. Чем меньше ошибок, тем выше различающая способность задания. Но эту идеальную картинку отвергает противоречивая практика образования, которая редко у кого порождает стройную систему знаний. Суть этой главной проблемы образовательных систем и образовательных учреждений кратко и выразительно отметил великий русский поэт А.С. Пушкин: «Мы все учились понемногу, чему-нибудь и как-нибудь!»

Сам Гутман понимал идеалистичность своей теории, как понимал и её полезность для всякого, кто собирается создавать качественный тест. Нужно было искать более гибкие формы оценки качества заданий, допускающие возможности небольшого отклонения от образцовой шкалограммы. В каждом тестировании число неправильных профилей испытуемых и заданий было заметно больше числа правильных профилей.

В попытках уйти от идеализма и жёсткого формализма своей теории у Л. Гутмана возникла идея перехода к вероятностной логике оценки качества заданий. В каждом задании вероятность правильного ответа испытуемых должна расти по мере повышения уровня подготовленности испытуемых. Так Л. Гутман пришёл к идее M. Richardson строить для каждого задания теста график роста вероятности правильного ответа в зависимости от роста уровня подготовленности испытуемых. Чем выше подготовка, тем больше эмпирическая и теоретическая вероятность правильного ответа. Свой график Л. Гутман представлял в виде гладкой и непрерывно возрастающей функции. Такой график он назвал, вслед за П. Лазарсфельдом, trace line of the test item[18]. Пример графического образа задания представлен на рис. 3.

Четвёртый источник развития IRT — результаты исследования D.N. Lawley, из Эдинбургского университета[19]. В 1943 году он опубликовал работу, показывающую, что некоторым понятиям статистической (классической) теории тестов[20] можно поставить в соответствие другие формальные показатели, вычислять их по новым формулам и тем самым делать оценки точнее.

В поисках точек локализации мер трудности заданий на числовой оси авторы согласились считать в качестве показателя меры трудности заданий проекцию точки перегиба функции-графика задания на числовую ось. В однопараметрической модели именно в этой точке вероятность правильного ответа на задание равнялась $^1/_2$. То есть этот вопрос был решён конвенционально.

18

Cit in:
Van der Linden W.J. Trace lines in item response theory. Rasch Measurement Transactions, 1993, 7: 3 p. 308.

19

Lawley D.N.
On Problems Connected with Item Selection and Test Construction // Proceedings of the Royal Society of Edinburgh. Section A Mathematical and Physical Sciences. 43 v. LXI, 1943, part III, p. 273–287.

20

Так тогда называлась теория измерений.

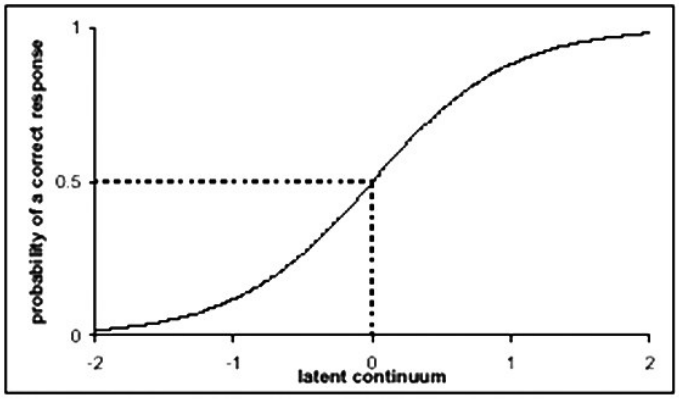

Рис. 3. Вероятность правильного ответа плавно возрастает по мере роста уровня подготовленности испытуемых

Пятым источником IRT можно назвать исследования по возможности совместного шкалирования данных, имеющих разные исходные шкалы, в одной общей шкале: уровня подготовленности испытуемых и параметров заданий[21]. Примерно так можно перевести на русский язык и идею, которая по-английски имеет название conjoint measurement (G. Rasch 1958). У истоков этой идеи стоял Л. Гутман, реализовали же её удачным образом Г. Раш и другие авторы[22].

Сам Г. Раш считал, что он совершил открытие в области психолого-педагогических измерений, но этот его образ мысли вызывал усмешку у большинства коллег и слушателей. На его лекциях в США в аудитории оставался только один слушатель — Б. Райт. Спустя много лет Б. Райт сумел донести до учёных США, а затем и всего мира, действительный смысл открытия в области психолого-педагогических измерений, которое сделал Г. Раш.

Шестой источник развития IRT — работы F.M. Lord[23], A. Birnbaum[24] и многих их коллег, усилиями которых IRT приобрела современный облик. Главной заслугой Ф. Лорда можно считать разработку двухпараметрической модели педагогических и психологических измерений, методов оценки параметров заданий и участие в создании эффективной компьютерной программы «Logist», для одновременного расчёта параметров заданий и уровня подготовленности испытуемых. У него хватило научной объективности оценить большую практичность моделей А. Бирнбаума[25].

21

Например, в одной шкале можно сравнить меру уровня подготовленности испытуемых и меру трудности заданий. На языке логике можно сказать, что такое измерение помогает превращать несравнимые понятия в сравнимые.

22

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. 199 p. ;
Luce R.D., Tukey J.W. Simultaneous conjoint measurement. Journal of Mathematical Psychology. 1964; 1: 1–27;
Perline R., Wright B.D., Wainer H. The Rasch model as additive conjoint measurement. Applied Psychological Measurement. 1979; 3: 237–256.

23

Lord F.M. Applications of item response theory to practical test problems. Hillsdale, NJ: Erlbaum. 1980.

Ф. Лорд одним из первых исследовал соотношение между баллом испытуемого на латентной переменной величине, называемой «ability» (в нашем случае, уровень подготовленности), наблюдаемым (исходным) тестовым баллом испытуемого и между истинным тестовым баллом испытуемого. Он установил, что все эти баллы должны иметь различное истолкование. Балл испытуемого на латентной переменной основан на идее абстрактно-истинного балла личности, независимого от теста, который применяется с целью педагогического измерения[26]. Наблюдаемый исходный тестовый балл испытуемого, естественно, зависит от применяемого теста (test dependent). Зависит от используемого теста и истинный (теоретический) тестовый балл, который определяется на основе получаемого при тестировании исходного тестового балла[27].

Публикация А. Бирнбаума была своеобразным ответом учёных США, безусловных лидеров в тестовых технологиях, на неожиданный научный прорыв датчанина Г. Раша, предложившего в конце пятидесятых годов три свои модели психолого-педагогического измерения. Одна из моделей в США была замечена больше других и стала называться там однопараметрической моделью измерения. По логике текстов американских математиков-психометриков выходило, что работа Г. Раша — лишь частный случай IRT, одна из трёх математических моделей измерения. Однако такая интерпретация не поддерживается рядом исследователей, равно как и автором этой статьи.

И, наконец, в качестве седьмого источника развития и внедрения IRT в практику хотелось бы упомянуть классическое пособие Ф. Бейкера[28]. По математической теории педагогических измерений (IRT) написано много хороших книг[29], но для начального изучения этой теории ничего лучшего по краткости, систематичности, доступности и культуре изложения пока нет. Поэтому начинать изучение IRT лучше всего с книги и компьютерных программ Ф. Бейкера.

Классификация основных понятий педагогических измерений

К настоящему времени научная лексика МТИ (IRT) насчитывает порядка тысячи терминов, помогающих проведению измерений во многих странах мира. Термином обычно называют слово или устойчивое словосочетание, которому приписывают определённое научное или специальное понятие.

24 *Birnbaum A.* «Some latent trait models and their use in inferring an examinee's ability.» Part 5 in F.M. Lord and M.R. Novick. Statistical Theories of Mental Test Scores. Reading, MA: Addison-Wesley, 1968.

25 *Lord F.M., Novick M.* Statistical Theories of Mental Test Scores. Reading, MA: Addison – Wesley, 1968.

26 Он писал: «Ability scores are more fundamental, because they are test independent». *Lord F.M.* The relation of test score to the trait underlying the test // Educational and Psychological Measurement, 1953, 13, 517–548.

27 *Lord F.M.* 1953. Там же.

28 *Baker F.B.* The Basics of Item Response Theory. 2 ed. Hieneman, Portsmouth, New Hempshire, 2001.

29 *Hambleton R.K., Swaminathan H.* Item response theory: principles and applications. Boston: Kluwer-Nijhoff Publishers. 1985.

Процесс внедрения IRT в российскую практику педагогических измерений тормозится из-за того, что заметная часть терминов и понятийный аппарат нуждаются в системной разработке. С целью хотя бы некоторого упорядочения положения дел, в настоящей статье выделены четыре группы основных понятий:

Первую группу естественно образовать из понятий педагогики и теории педагогических измерений. Это:

- испытуемые (тестируемые);
- исходный тестовый балл испытуемых;
- педагогическое задание, задание в тестовой форме, тестовое задание;
- свойства испытуемых и заданий;
- тестовый процесс;
- процесс педагогических измерений и различных видов педагогического оценивания;
- педагогический тест.

Ко второй группе можно отнести основные математико-статистические понятия, распространённые преимущественно в статистической (классической) теории педагогических измерений[30]. Это:

- мера трудности задания (item difficulty);
- тестовое задание и его статистические свойства;
- различающая способность задания (item discrimination);
- истинный тестовый балл испытуемого, определяемый в статистической (классической) теории тестов (True Test Score, обозначается T_i).

В классической теории есть два сильно различающихся варианта концептуализации истинного тестового балла. Первый называется specific true score. Это означает: истинный тестовый балл можно получить как средний балл из всех параллельных вариантов теста. Второй вариант истинного тестового балла испытуемого называется generic true score, который можно получить как средний балл тестов, образованных выборками заданий из одной и той же генеральной совокупности заданий;

- одномерность (unidimensionality);
- надёжность и валидность тестовых результатов;
- стандартная ошибка измерения.

Некоторые российские авторы считают, что классическая статистическая теорию устарела и после появления IRT стала не нужной. Кроме этого заблуждения, наблюдаются также случаи увлечения одним методом, одной теорией, например, только теорией Раша или Бирнбаума. Но такого рода ограниченный подход приводит к ошибкам интерпретации данных, и в конечном итоге может превратиться в признак возможного непрофессионализма. При сравнении практической

30

Gulliksen H. и др. Theory of Mental Tests. N–Y. Wiley. 1950. 486 p.

полезности теорий ведущие западные авторы считают классическую теорию обязательной, незаменимой[31]. И это абсолютно верное мнение.

Третью группу образуют понятия собственно IRT. Это:

- уровень подготовленности испытуемого, определяемый на латентной переменной величине (θ_i). Признание качества латентным означает, что в отличие от элементарных оценок и некоторых простых физических измерений процесс научно-педагогического измерения требует теоретизации. Куда входят проверка логической правильности имени измеряемого качества, определения ведущего понятия и выделяемого предмета, а также системы индикаторов, понятийных и эмпирических индикаторов, указывающих на наличие или отсутствие включаемых в понятие признаков интересующего качества[32];
- математические функции — модели измерения, определяющие вероятность правильного ответа на задание, в зависимости от уровня подготовленности испытуемых. Вероятность правильного ответа на задание зависит не только от уровня подготовленности испытуемого, но и от избранной модели измерения, каждая из которых может иметь один, два или три параметра[33].
- график функции, который в IRT ставится в соответствие каждому заданию. На английском языке эти функции называются Item Characteristic Curves. На русский язык их обычно переводят дословно и неблагозвучно: «характеристические кривые заданий» В данной статье такой перевод не поддерживается. Сумма функций заданий даёт понятие и график теста в целом. По-английски такой график традиционно называется Test Characteristic Curve.
- показатель уровня трудности заданий в IRT (item difficulty parameter); обозначается символом β_j, где подстрочный символ j представляет номер задания ($j = 1, 2....k$), k — число заданий проектируемого теста.
- показатель уровня различающей[34] способности задания (item discrimination parameter). Различающей способностью задания называется его способность (свойство) дифференцировать испытуемых по уровню подготовленности[35]. Чем выше такая способность задания, тем лучше деление испытуемых на подготовленных и на не подготовленных. Уровень такой способности обозначается символом a_j.
- вероятность угадать ответ в случае, если все испытуемые не знают правильного ответа (item guessing parameter)[36]. Потенциальная возможность угадать правильный ответ обозначается символом c_j.
- истинный балл испытуемого i, определяемого по формуле

31

Classical theory, which may be briefly described as an application of linear least squares estimation and analysis of variance along the lines of traditional error analysis, remains indispensable ..., but it cannot do the whole job either in theory or practice. r. Darrell Bock and Robert Wood. Test theory // Annual Review of Psychology. 1971, pp. 165–224.

32

Об этом процессе подробнее см., например, на стр. 87–105 книги *Аванесов В.С.* Тесты в социологическом исследовании. М.: Наука , 1982. 199 с.

33

Описание всех трёх моделей смотрите, например, в предыдущей статье автора: *Аванесов В.С.* Item Response Theory: Основные понятия и положения. ПИ № 2, 2007, С. 3–28.

34

На русский язык этот параметр нередко переводят как уровень дискриминантности (или дискриминативности) задания. На самом деле ничего дискриминационного в этом параметре нет.

Lawly. Обозначается двумя символами TS_i, с подстрочной буквой i, где i — номер испытуемого; i принимает значения от 1 до N — где N — число испытуемых;

- информационная функция задания и теста;
- свойство локальной независимости ответов испытуемых одинакового уровня подготовленности на задания теста (Local Independence);
- принцип инвариантности (независимости) оценок уровня подготовленности испытуемых от уровня трудности заданий теста, а также независимости значений параметров заданий от уровня подготовленности тестируемых групп испытуемых;
- единица измерения, которая называется логит;
- шкалирование значений параметров испытуемых и параметров заданий в единой стандартизованной шкале натуральных логарифмов (test calibration).
- параметры функции.
- метод χ^2-квадрат для проверки соответствия заданий избранной модели измерения;
- уровень значимости выборочных статистик χ^2-квадрат.

К понятиям этой же группы можно отнести сотни понятий математики и вычислительных методов, (максимального правдоподобия, Байесовские и другие), используемые при статистическом моделировании, определении параметров испытуемых и заданий;

Четвёртую группу понятий целесообразно соотнести со специально-философскими, общенаучными и социально-политическими аспектами педагогических измерений. Это понятия:

- личность испытуемого;
- объективность педагогических измерений;
- качество педагогических измерений, критерии качества;
- понятийные и эмпирические индикаторы признаков понятия.
- справедливость оценивания, понятие «право испытуемых на объективную и справедливую оценку уровня их подготовленности». Часть приведённых понятий были уже определены в ряде предыдущих работ автора[37]. В этой статье основное внимание уделяется понятиям собственно IRT.

Многие определения IRT были ранее сформулированы вне педагогического языка. В этой работе делается попытка дать определения основных понятий этой теории, согласованной с системой определений педагогических измерений, сформулированной ранее[38]. Часть понятий была определена в предыдущих статьях автора.

Задача педагогического измерения обычно формулируется как определение тестового балла и места испытуемого на

35 В предыдущих своих работах автор этой статьи использовал понятие «дифференцирующая способность задания». Но это название в западной литературе имеет несколько иной смысл. Поэтому здесь принято новое название «различающая способность задания».

36 Английские варианты понятий приведены здесь, во-первых, для создания возможностей построения параллельных вариантов лексики – русской и английской, необходимых для точного понимания сути теории, и, во-вторых, для лучшего понимания различий между переводом слова и смыслом научного понятия.

37 *Аванесов В.С.* Основы педагогической теории измерений // Педагогические измерения, № 1, 2004 г. С. 15–21 и другие статьи автора, в №№ 1–2, 2004 г. и в №№ 1–4, 2005 г., №№ 1–4, 2006 г. и №№ 1–2, 2007 г.

38 Там же.

числовой шкале уровня подготовленности. К этому добавляют и задачу определения параметров заданий. Что правильно с технической точки зрения. Но с педагогической точки зрения задача измерения стоит шире и глубже. Вопрос может ставиться об измерении интересующих свойств личности, а также содержательных и формальных свойств педагогических заданий. Поэтому самым первым следует признать понятие «личность испытуемого (тестируемого)»

Личность испытуемого

В нашей работе уже было определено, что в тестовом процессе испытуемые — это граждане, выражающие добровольное желание объективно определить уровень своей подготовленности и на этой основе решать вопросы своего социального и профессионального самоопределения. Если принцип добровольности не выполняется, то испытуемые превращаются в подопытных лиц, что нарушает их права, записанные в ст. 21 Конституции РФ. Главная разница между испытуемыми и подопытными лицами заключается именно в признаке добровольности участия в экспериментах (опытах)[39].

Все испытуемые имеют право на объективное измерение уровня их подготовленности, на своевременное получение объективной информации о собственных результатах, о результатах конкурирующих с ними других испытуемых, а также право на высокое качество измерений их знаний, умений, навыков и компетенций. Объективность возникает как следствие интеграции методов обоснования надёжности и валидности тестовых результатов[40].

Объективность обеспечивается также такими моделями измерения, которые позволяют оценить уровень подготовленности испытуемого независимо от выборки заданий, доставшейся испытуемому в виде теста, и общей научной организацией тестового процесса. Объективность может быть обеспечена только совместной координированной деятельностью профессиональных, общественных и государственных органов управления образовательной деятельностью.

Уровень подготовленности личности

Второе основное понятие педагогической и математической теорий измерений — это уровень подготовленности личности. Оно уже рассматривалось в первой статье. Здесь остаётся только углубить, расширить и уточнить его.

39

Поскольку ЕГЭ нарушает принцип добровольности, то эта форма объективно проявляет себя как форма насилия. См. напр.: *Аванесов В.С.* ЕГЭ как насилие // Независимая газета. 20.10.2006. http://www.ng.ru/politics/2006-10-20/3_kartblansh.html

40

В теории педагогических измерений утверждается, что надёжность – это свойство тестовых результатов (D.A. Frisbie. NCME Instructional module on Reliability of scores from teacher-made tests. p. 55). После семи лет экспериментирования с ЕГЭ было заявлено (Г.В. Ковалёва «Результаты ЕГЭ в 2006 году» // Школьные технологии. № 6 2006 г., с. 146), что этим же свойством обладают и КИМы ЕГЭ. Там написано: «...По сравнению с 2005 г. повысилось качество КИМ. Средняя надёжность (коэффициент альфа, Кронбах) КИМ по всем предметам находится в пределах от 0,85 до 0,93. Средняя дифференцирующая способность заданий по всем предметам,

кроме математики, колеблется от 30 до 60% (соответственно в 2005 г. 25–57%), причём для заданий с выбором ответа она составляет 30–51% (25–51), для заданий с кратким ответом — 36–56% (31–55) и для заданий с развёрнутым ответом — 25–60% (21–57)».

Между тем, приведенные проценты обладают такой же нулевой научной информативностью, как и среднебольничная температура. Упоминание процентов о «средней дифференцирующей способности заданий» противоречит реальной конструкции КИМов (вспомните сильно заниженную по трудности часть «А» и запредельную трудность части «С»), что, естественно, наводит на мысль о надуманности заявленных цифр, и укрепляет во мнении, что КИМы ЕГЭ действительно имеют неустранимые дефекты.

КИМы не допускают возможности получения качественных и объективных результатов педагогического измерения в принципе, из-за того, что в них континуум уровня подготовленности, например, по математике, разрывается на три сильно различающихся от

Известны пять основных оценок результатов тестирования испытуемых, не считая производных. Для предупреждения неизбежной в таких случаях путаницы каждой оценке даётся в соответствие своя символика.

1. Первая оценка — это исходное эмпирическое значение тестового балла испытуемого, или короче, исходный тестовый балл испытуемого (Raw Test Score or Test Score). Обозначается символом X_i или Y_i Исходный тестовый балл испытуемого получается элементарным сложением баллов, полученных каждым испытуемым по итогам выполнения всех заданий теста.

2. Вторая оценка — это истинный балл испытуемого в варианте классической теории тестов (так она называлась раньше)[41]. Истинный балл в этом варианте классической теории обозначается символом T_i. Он определяется как математическое ожидание результатов по любому параллельному варианту теста, на которые мог бы, предположительно, ответить испытуемый. Значение T_i вычисляется как интервальная оценка посредством построения доверительного интервала вокруг выборочного исходного тестового балла испытуемого *i*. Это делается по формуле $T = X_i \pm ts_e$, где t означает меру достоверности выборочных статистик по Стьюденту, а s_e — стандартная ошибка измерения. Для 5% уровня риска ошибки t принимается равным 1,96. Значение se определяется по формуле $s_e = s_x\sqrt{1-r_{xx}}$, где s_x равно стандартному отклонению исходных тестовых баллов испытуемых, а r_{xx} означает надёжность результатов, определяемую коррелированием результатов испытуемых в двух параллельных вариантах теста.

3. Истинный тестовый балл испытуемого (True Score), по версии D.N. Lawely, — это теоретически вычисляемый тестовый балл каждого испытуемого. Обозначается TS_i и вычисляется как сумма вероятностей правильных ответов испытуемого[42]. Для вычисления значений TS_i у каждого испытуемого он предложил формулу

$$TS_i = \sum_{j=1}^{k} P_j(\theta),$$

которая выражает смысл суммирования вероятностей правильного ответа каждого испытуемого на все задания теста, если известны параметры каждого задания.

4. Четвёртая оценка — это теоретически истинный тестовый балл испытуемого, значение параметра θ_i, не зависимого от выборки испытуемых и выборки заданий. Здесь уместно вновь обратиться к трактовке Ф. Лорда. Вопрос точности определения θ_i связан с уровнем дифференцированной стан-

дартной ошибки измерения для каждого измеряемого уровня. А это, помимо прочего, всегда производное от объёма выборок. Позже будет отмечено, что качество оценок, получаемых при применении IRT, сильно зависит от объёма выборок. И это одно из существенных ограничений, накладываемых на её применение в практике.

5. В IRT уровень подготовленности испытуемых вычисляется на латентной переменной, как обратная задача по отношению к расчёту вероятности правильного ответа. Знание эмпирических вероятностей правильного ответа и знание параметров заданий позволяет получить оценки значения $\hat{\theta}_i$ — уровней подготовленности испытуемых, где подстрочный символ i относится к испытуемому под номером i, и может принимать значения 1, 2, ... N.

Важно избавиться от распространённого мифа, что получаемые в единичном опыте измерения посредством IRT значения уровня подготовленности испытуемых $\hat{\theta}_i$ — это и есть истинное значение параметра подготовленности испытуемого. На самом деле, это очередная оценка значения латентного параметра. Если дать испытуемому другой вариант теста, то нередко изменяется профиль баллов испытуемого, другой становится ошибка измерения, а значит, меняется и оценка $\hat{\theta}_i$.

Вообще, получаемые значения $\hat{\theta}_i$ зависят от профиля тестовых баллов испытуемого, а он редко когда воспроизводится в точности у каждого испытуемого в параллельных вариантах теста. Таким образом, оценки значений $\hat{\theta}_i$ в каждом случае измерения не равны точно истинным значениям уровня подготовленности испытуемых (параметра), а всего лишь являются тоже оценками, меняющимися от измерения к измерению. Мерой вариации оценок $\hat{\theta}_i$ относительно параметра является стандартная ошибка измерения.

Математическая функция заданий и теста

Заданию как педагогическому феномену в нашем журнале уже были посвящены две большие статьи[43] автора. В этих статьях была поставлена задача создания теории педагогических заданий.

В МТИ (IRT) творческим образом соединяются два феномена разных наук: задание — со стороны педагогической науки и функция — со стороны математики. Всё началось с так называемой логистической функции, где аргументом является показатель степени

$$f(x)=\frac{e^x}{1+e^x}.$$

резка, сочетаются субъективные методы оценивания с объективными. Не случайно нет ни одного нормального научного отчёта с эмпирическими данными, где было бы показана фактическая надёжность и валидность результатов ЕГЭ.
В случае полномасштабного введения ЕГЭ в 2009 году эти оценки могут стать предметом массовых судебных разбирательств. Граждане России имеют право на объективную оценку уровня их подготовленности. К счастью, сам Гособрнадзор избегает делать оргвыводы о качестве образовательной деятельности директоров, учителей, школ и территорий на основе оценок ЕГЭ.

41 Теперь это weak true score theory.

42 Алгоритм расчёта вероятностей правильного ответа в зависимости от уровня подготовленности испытуемых представлен в статье автора: ПИ, № 2, 2007. стр. 20.

43 *Аванесов В.С.* Основы теории педагогических заданий. ПИ, №№ 2 и 3, 2006 г.

Поскольку в педагогических измерениях аргумент функции выражает не только число, но также и выражения, как буквенные, так и числовые, то принята другая, более удобная форма этой записи. Например, в функции вида

$$P_j(\theta) = \{x_{ij} = 1 | \beta_j\} = \cdot = \frac{\exp(\theta - \beta_j)}{1 + \exp(\theta - \beta_j)},$$

вместо x появилась разность значений двух параметров, $(\theta - \beta_j)$. Эта разность позволяет связать уровень подготовленности испытуемых (θ_i) с вероятностью правильного ответа на задание уровня трудности β_j, где j — номер задания. Многие зарубежные авторы считают т.н. логистическую функцию самым существенным элементом IRT.

График такой функции на английском языке имеет устаревшее название Item Characteristic Curve (ICC). Как отмечают W. Van der Linden и R. Hambllетone[45], это название в 1946 году ввёл Tuker[46]. Уже в 1950 году P. Lazarsfeld дал графику такой функции другое название[47]. На русском языке ICC нередко переводятся, как говорится, «в лоб» — в виде «характеристические кривые заданий», что нельзя признать удовлетворительным. График функции — это линия на плоскости, отображающая уровни подготовленности испытуемых в значения вероятностей правильного ответа на задание известного уровня трудности. В нашем случае график полезно рассматривать как образ тестового задания, представленный в системе прямоугольных координат на плоскости.

Хотя в практике обычно рассматривают один график, в МТИ (IRT) полезно принимать во внимание два графика. На рис. 4 оба графика показывают значения вероятностей правильного ответа (красная линия) и неправильного ответа (чёрная линия) на одно и тоже задание. Значения вероятностей рассчитаны по однопараметрической модели, в зависимости от уровня подготовленности испытуемых. При изображении вероятности неправильного ответа по оси ординат откладывают значения $Q(\theta_i)$, равные $1 - P(\theta_i)$. Поскольку точки графика $Q(\theta_i)$ легко находятся из разности $1 - P(\theta_i)$, то основное внимание исследователи обычно обращают на графики функции $P(\theta_i)$.

Часто бывает полезным построить график не отдельного задания, а теста в целом. На рис. 5 чёрные линии — это графики заданий, красная линия – график «теста», построенного всего лишь по пяти заданиям.

Когда математики говорят о функциях, то считается полезным указать на некоторые их свойства.

45 *Van der Linden, Wim and Hambleton R.K.* Handbook of Modern Item Response Theory. 1997. Springer-Verlag, New-York Inc. p. 5.

46 *Tuker L.R.* Maximum validity of a test with equivalent items. Psychometrika, 11, 1946, 1–13.

47 Подтверждение этого факта см. в работе Van der Linden, W.J. Trace lines in item response theory. Rasch Measurement Transactions, 1993, 7: 3 p. 308.

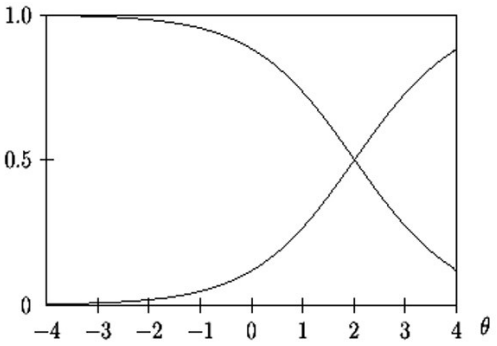

Рис. 4. Графики зависимости вероятности правильного ответа (красная линия) и неправильного ответа (чёрная линия) испытуемых от уровня подготовленности

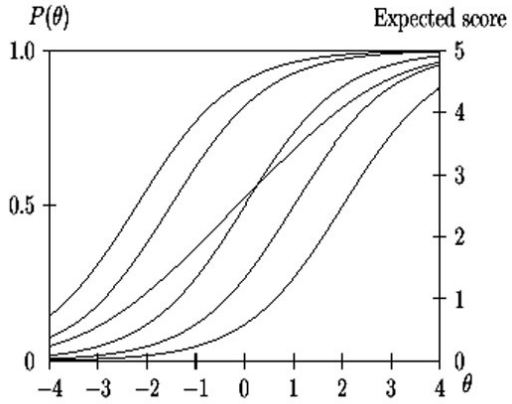

Рис. 5. Графики функций заданий и теста

1. Функция задания монотонная, возрастающая на оси q. Чем выше значение испытуемого на латентной переменной, тем выше вероятность правильного ответа.

2. Вероятность правильного ответа принимает значения между нулём и единицей.

3. График этой функции — непрерывная линия.

Параметры функций

В математике параметрами функции обычно называются числа, которые в процессе вычисления значений зависимой переменной по значениям независимой переменной остаются постоянными. Простой пример функции $y = 3 + 2x$, где x и y переменные величины, а коэффициенты 3 и 2 могут быть названы параметрами линейной функции. В данном случае параметры функции — это два числа, константы.

Функции бывают разными, их число может быть неограниченно. Для того, чтобы число функций в IRT как-то минимизировать, было принято решение не фиксировать каждый раз в формуле значения параметров, а мыслить их в качестве переменных величин, так называемых переменных параметров функций. Так что в IRT слово «параметр» имеет другой смысл. Это качественно фиксированная, но переменная в количественном отношении латентная величина, различная для каждого задания

и испытуемого. Параметры в IRT — это числа, которые используются в каждом вычислительном процессе. Их называют параметрами только потому, что они заранее неизвестны, за исключением одного. И их значения предстоит определить. Иначе говоря, надо определить значения параметров заданий и параметров испытуемых.

Обычно выделяется четыре параметра:

- уровень трудности задания (βj) , где j — номер задания;
- параметр крутизны графика функции задания. Обозначается a_j;
- уровень подготовленности испытуемых (θ_i), i — номер испытуемого;
- мера возможного угадывания правильного ответа на различные задания теста (c_j). Чем меньше число ответов к заданию, тем выше возможность угадывания правильного ответа.

Из этих четырёх только значение cj определяется явно и может быть известно заранее, исходя из числа ответов к каждому заданию. Кроме того, в процессе поиска подходящего ответа проявляется не только удачливость, но и способности понимать, рассуждать, логически обосновывать своё решение. Вероятно, поэтому Ф. Лорд называл c_j, параметром псевдоугадывания. Автор этой статьи склонен называть c_j псевдопараметром. Ещё один, пятый параметр γ (читается гамма) представляет больше теоретический, чем практический интерес. Он плохо интерпретируем, но иногда бывает полезен. Формула с этим параметром представлена, в частности, в диссертационной работе автора данной статьи[48].

Главным из трёх параметров заданий можно назвать параметр трудности. На рис. 2 представлены проекции точки перегиба функции на ось абсцисс и на ось ординат. Поскольку в IRT уровень подготовленности испытуемых измеряется в одной и той же стандартной шкале логитов, как и уровень трудности заданий, проекция этой точки перегиба функции на ось абсцисс даёт меру трудности заданий. На рис. 6 красная стрелка показывает, что представленное, для примера, задание имеет уровень трудности плюс один логит. В теории Раша проекция точки перегиба функции на ось ординат в точности равна точке $^1/_2$ на шкале вероятности правильного ответа. Там угадывание не допускается. Полезно эту же мысль выразить несколько иначе. Трудность задания определяется в такой точке графика, чтобы проекция этой точки на ось тета в точности совпадала с проекций на ось ординат, там, где вероятность правильного ответа на задание равнялось бы $^1/_2$.

48

Аванесов В.С. Методологические и теоретические основы тестового педагогического контроля. Дис. ... д-ра пед. наук. С-Пб. Госуниверситет, 1994 г.

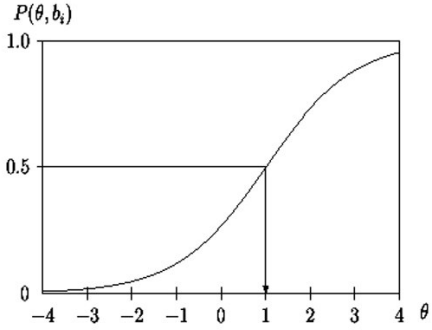

Рис. 6. Определение значения параметра трудности задания

Чем труднее задание, тем правее располагается его график. Точнее, проекция точки перегиба функции на ось абсцисс более трудного задания располагается правее. Этим объясняется другое английское название параметра трудности задания — location parameter. Графическая иллюстрация сравнительной меры трудности двух заданий представлена на рис. 7.

На рис. 8 представлена система заданий возрастающей трудности. В Интернете можно найти такой пример графиков пяти заданий, с дидактической надписью на английском языке: «Rasch curves never cross», что означает, что в однопараметрической модели измерения графики заданий никогда не пересекаются. Это и есть одно из системных условий качественного педагогического измерения, присущих исключительно модели Г. Раша. При пересекающихся графиках заданий возрастает число ошибок измерения, а следова-

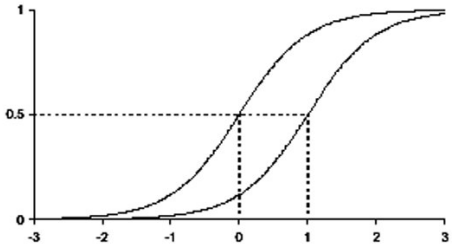

Рис. 7. Различие графиков двух заданий по уровню трудности

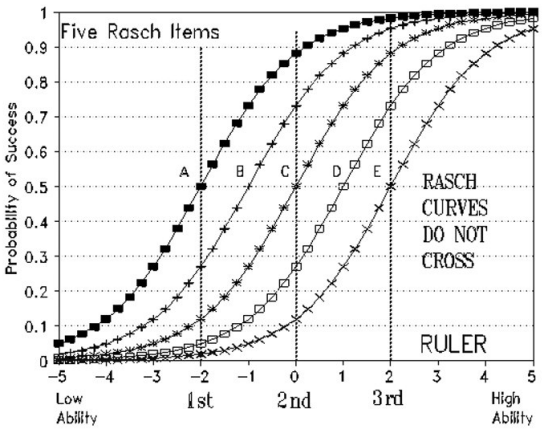

Рис. 8. Графики пяти заданий возрастающей трудности, с одинаковым значением параметра крутизны

тельно, ухудшается и качество теста. Это обстоятельство сильно снижает ценность двух- и трёхпараметрических моделей IRT для создания теста как системы заданий возрастающей трудности, с непересекающимися графиками.

Графический образ задания меняется в зависимости от значений параметров a_j и c_j. На рис. 9 задание, расположенное справа, труднее и имеет сравнительно большую различающую способность.

Влияние параметра a_j на графики двух заданий разного уровня трудности представлено на рис. 9. Этот параметр называется по-английски item discrimination parameter. Чем больше значения a_j, тем круче выглядит график задания. Соответственно, «крутыми» иногда называют и задания.

На рис. 10 представлены задания одинакового, среднего уровня трудности, но имеющие неодинаковую меру различающей способности.

Можно думать, что идея применения параметра a_j возникла из-за желания расшить возможности исследователей при подборе более подходяще-

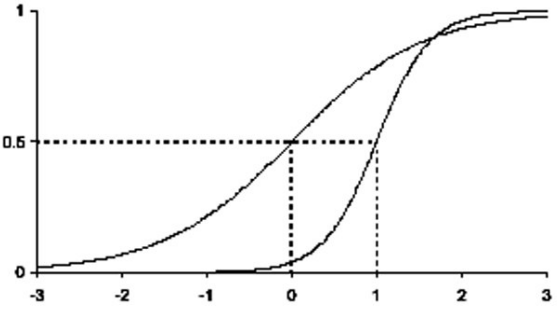

Рис. 9. Второе задание труднее и имеет более высокий уровень различающей способности

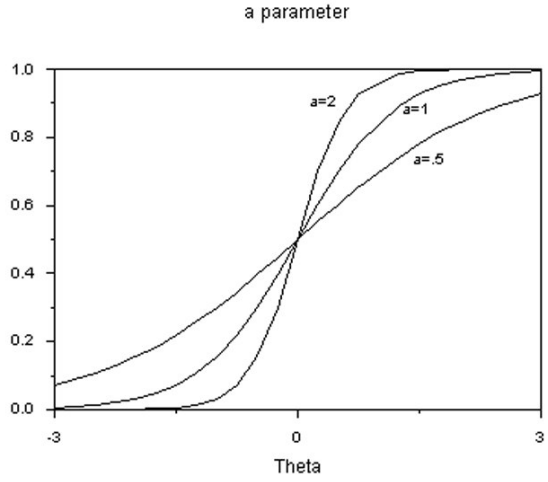

Рис. 10. Изменение графиков заданий в зависимости от значений параметра a_j

го графика для эмпирических точек. Сторонники IRT указывают на преимущества двухпараметрической модели перед однопараметрической, опираясь именно на этот тезис. Вопрос, однако, в том, что идея индивидуального подбора графиков для имеющихся данных находится в противоречии с идеей измерения по теории Раша: там, наоборот, данные должны соответствовать требованиям измеряемой модели. И это очень плодотворная идея с точки зрения создания теста как системы заданий равномерно возрастающей трудности. В этом пункте идеи Л. Гутмана и Г. Раша полностью совпали.

Влияние параметра c_j на изменение графика функции $P(\theta_i)$ представлено на рис. 11. Чисто теоретически идея применения параметра c_j в трёхпараметрической модели измерения кажется привлекательной. Появляется возможность делать коррекцию на угадывание правильного ответа в случае неподготовленности испытуемых. Испытуемые с низкой подготовкой имеют возможность угадать правильный ответ с вероятностью, зависящей от качества задания. Абстрактно c_j может принимать значения от нуля до единицы. Практические пределы изменения значения c_j — от нуля до 0,5, в случае задания с двумя ответами, при качественном дистракторе. При некачественных дистракторах вероятность угадывания резко возрастает. Отсюда важная роль педагогического анализа не только содержания задания, но и содержания каждого ответа.

Г. Раш возражал против введения параметра c_j и использования при тестировании заданий, где есть вероятность угадывания правильного ответа. Вот почему для своего первого теста он просил коллег-психологов собрать данные с

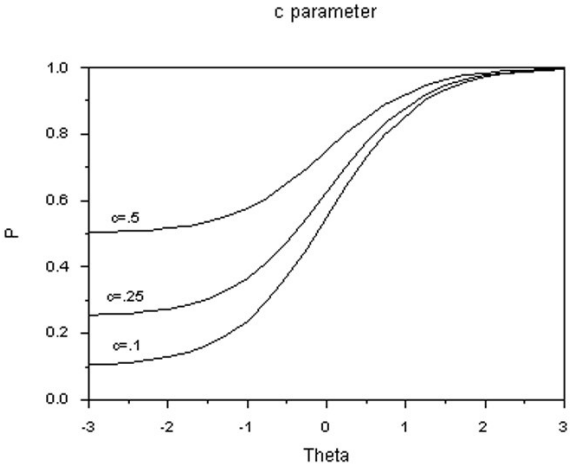

Рис. 11. Влияние значения параметра c_j на изменение графика задания

применением заданий открытой формы, где вероятность угадывания считается нулевой. Не одобрял он также использование в психолого-педагогических измерениях и параметра a_j, справедливо полагая, что из заданий, имеющих графики различной крутизны, качественный тест не создать. Но к его мнению не прислушались.

Изменение значения параметра c_j влияет на крутизну графика задания. А это означает, что меняются значения параметров a_j и β_j . В случае применения трёхпараметрической модели измерения мера трудности задания равна проекции на ось абсцисс точки графика

$$P(\theta) = 0{,}5 \cdot \left(c_j + \left(1 - c_j\right)\right) = \frac{c_j + \left(1 - c_j\right)}{2}.$$

Нижним пределом значения функции становится не ноль, а значение c_j.

Не все функции IRT можно признать логистическими. Трёхпараметрическая модель измерения логистической не считается. Потому что не все свойства логистической функции на неё распространяются.

Свойство одномерности и локальной независимости

Одномерность (unidimensionality) — это свойство отдельных заданий, включаемых в тест, и свойство теста как системы заданий. Суть этого свойства можно понимать так, что все задания в тесте измеряют преимущественно одно и тоже интересующее свойство личности. Если методами факторного (регрессионного) анализа элиминировать вариацию, определяемую данным свойством, то корреляции между заданиями должны становиться близкими нулю.

В МТИ это свойство означает, что все задания подобраны так, что они имеют в своей основе только один латентный фактор. Хотя так бывает очень редко, ценность идеи одномерности от этого нисколько не снижается. Это тот идеал, к которому стремятся. Наличие этого свойства является решающим при оценке содержательной валидности тестовых результатов. Отсутствие одномерности, а следовательно, и валидности результатов теста делает всю работу общественно бесполезной.

Старые тестологи сравнивали критерии валидности и надёжности тестовых результатов с определением времени по часам. Часы, имеющие ненадёжный ход, не могут показывать точное время. Но и надёжные (точно работающие) часы, будучи поставленными на неправильное время, тоже оказываются непригодными для ответа на главный во-

прос — который час? Часы, имеющие ненадёжный ход и поставленные неизвестно на какой час, полностью теряют возможность определить время. Так и тесты.

Если понятие «одномерность» относится к языку факторного анализа, то понятие локальная независимость относится к языку теории вероятности. Смысл этого термина «локальная независимость» надо понимать так: Для испытуемых одного одинакового уровня подготовленности (отсюда слово «локальная») вероятность правильного ответа на одно задание не зависит от вероятности правильного ответа на любое другое задание теста.

Эти два понятия связаны. Как показал F. Lord, если свойство одномерности заданий подтверждается, то проявляет себя и свойство локальной независимости заданий (Lord, 1980).

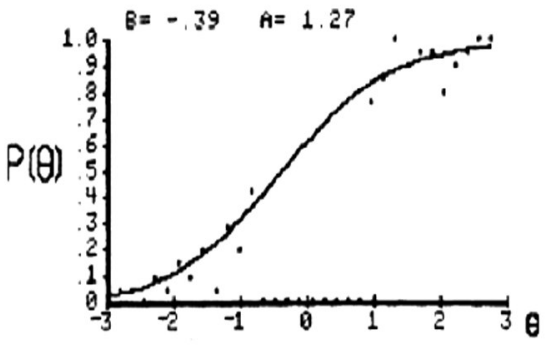

Рис. 12. Свойство инвариантности параметров функции, независимых от уровня подготовленности испытуемых

Инвариантность значений параметров

В рамках классической теории педагогических измерений оценки уровня трудности заданий зависят от уровня подготовленности группы. Чем лучше подготовлена тестируемая группа, тем выше доля (процент) правильных ответов на задание, тем легче оказывается задание. И наоборот, в слабой группе испытуемых процент выполнения заданий заметно ниже. По этому поводу сложилась даже своеобразная терминология. Оценки трудности заданий, оцениваемые в рамках классической статистической теории, называют зависимыми от уровня подготовленности испытуемых каждой группы. По-английски такие оценки называют group-dependent.

В IRT значения параметров заданий принципиально (теоретически) независимы от уровня подготовленности групп тестируемых. Это очень полезное свойство вытекает из свойства функции задания. График рис. 12 показывает распределение вероятностей правильных ответов в двух группах испытуе-

мых — сильной, точки расположены в правой части графика, и слабой, точки которой расположены в левой части графика. Функция одна, параметры функции задания те же самые для обеих групп. Точки в окрестности графика указывают на пригодность данной функции задания для определения вероятности правильного ответа в любой из этих групп.

Результаты слабой и сильной группы испытуемых располагаются вокруг соответствующих частей одного и того же графика функции задания.

Какую бы группу испытуемых мы не взяли для определения параметров заданий, свойства функции заданий проявляются одинаковым образом. Это и есть свойство независимости параметров задания. На английском языке это называется the item invariance of an examinee's ability. В данном случае инвариантность означает практическую неизменность значений параметров заданий от уровня подготовленности испытуемых. Фактически некоторые флуктуации в зависимости от групп есть, но они не столь существенны по сравнению с оценками трудности заданий, оцениваемыми в статистической теории тестов. Принцип остаётся в силе.

Симметрично независимы и оценки параметра уровня подготовленности испытуемых от уровня трудности заданий. Иначе говоря, значение уровня подготовленности испытуемых можно определять в разных по уровню подготовленности группах (the group invariance of an item's parameters).

Таким образом, в IRT имеют место два вида независимых (инвариантных) значений: параметры заданий независимы от параметров подготовленности испытуемых, а параметры подготовленности испытуемых независимы от параметров заданий. Это и есть одно из главных открытий Г. Раша, на основе которого стала развиваться IRT. Измерения по теории Г. Раша и по IRT дали мощный толчок становлению западных образовательных технологий.

Информационная функция

Информационная функция и методы её расчёта являются важной частью научного аппарата IRT.

Это понятие и метод вычисления ввёл A. Birnbaum[49].

$$I(\theta)=\sum_{j=1}^{k}\frac{\left(P_j'\right)^2}{P_jQ_j},$$

где $I(\theta)$ означает информационную функцию от латентной переменной величины θ;
$\left(P_j'\right)^2$ – квадрат значения производной функции в интересующей точке θ;

49

В книге: *Lord F.M., Novick M.* Statistical Theories of Mental Test Scores. Addison-Wesley Publ. Co. Reading, Mass. 1968. 560 pp.

Q_j — вероятность неправильного ответа на то же задание j, в той же точке θ;
k — число заданий теста.

Статус информационной функции трактуется по-разному. Одни авторы считают, что это понятие близко к понятию «надёжность тестовых результатов», другие — что оно имеет отношение и к валидности, третьи — к тому и другому. Автор данной статьи связывает эту функцию, кроме того, также с обоснованием эффективности теста и тестовых заданий[50] при проведении педагогических измерений. И нельзя сказать, что кто-то неправ, потому что информационная функция позволяет интерпретировать результаты с точки зрения всех перечисленных критериев.

Например, если вопрос касается надёжности тестовых результатов, то информационная функция показывает дифференцированную надёжность измерения каждого уровня подготовленности испытуемых. Не случайно в IRT полагают, что информационная функция указывает на локальную надёжность результатов (local reliability). При такой интерпретации информационной называют функцию, которая позволяет оценить меру точности измерения каждым отдельным заданием или тестом в целом.

Информационная функция свидетельствует о количестве информации, которое даёт каждое задание для измерения уровня подготовленности каждого испытуемого. Это количество зависит от значений близости оценок уровня подготовленности испытуемых и трудности задания. Чем ближе эти значения, тем более информативно (эффективно) задание для измерения уровня подготовленности испытуемых именно такого уровня подготовленности.

Для каждой модели измерения используется своя информационная функция. Для данных, представленных в дихотомической шкале (1/0) для модели Раша максимальное значение информационной функции равно 0,25. Это значение получается для заданий, имеющих среднюю меру трудности ($p = q = 0{,}5$). Тогда произведение $pq = 0{,}25$. В классической теории тестов этот показатель назывался дисперсией тестовых баллов по заданию, оцениваемому дихотомически. Доказательство максимума дисперсии, а равно и информационной функции, уже приводилось в наших работах. Если для данных модели Раша графики задания и информационной функции расположить на одной плоскости (рис. 13), то чёрная линия — это график задания, синяя линия — график информационной функции задания. На рис. 13 красная линия указывает на проекции точки перегиба функции на ось абсцисс и ось ординат.

50

Аванесов В.С. Методологические и теоретические основы тестового педагогического контроля. Дис. ... д-ра пед. наук. С-Пб. Госуниверситет, 1994 г.

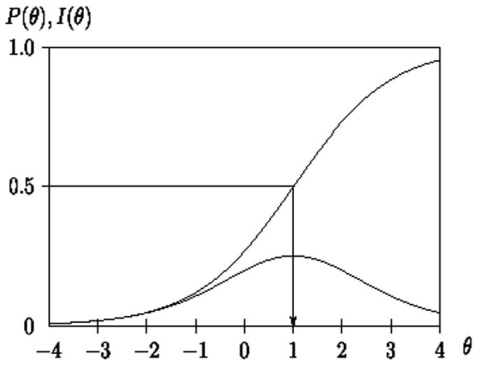

Рис. 13. Графики функции задания и информационной функции для модели Г. Раша

Максимум информации задание даёт для измерения уровня подготовленности испытуемых, у которых этот уровень в точности равен уровню трудности задания. Чем больше отличаются значения этих двух показателей, тем задание менее информативно с точки зрения эффективности измерения.

Поскольку информационная функция является функцией от латентной переменной, то смысл введения этой переменной требует краткого разъяснения. θ представляет собой форму одномерной реализации идеи существования ненаблюдаемой переменной, детерминирующей, как фактор, результаты испытуемых на наблюдаемой переменной, получаемой элементарным сложением баллов. Результаты теста всегда содержат в себе ошибки измерения, затрудняющие оценку значения тестового балла на латентной переменной. Поскольку латентная переменная появляется в результате концептуализации, она всегда остаётся гипотетической переменной величиной, относительно которой с большей или меньшей точностью оцениваются истинные результаты испытуемых, получаемые на основе эмпирических данных.

Необходимую для оценки информативности производную функции Q_j, например, по теории G. Rasch получают, дифференцируя выражение

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

где x_{ij} = 1, если ответ любого испытуемого (i) на j-е задание правильный; θ — уровень знаний, латентная переменная; β_j — уровень трудности j-го задания теста, измеряемой на латентном континууме.

Вероятность неправильного ответа на задание j, обозначаемая (Q_j) и равная, как принято в теории вероятностей, 1 – P, выражается так:

$$Q_j\left\{x_{ij}=0\,|\,\beta_j\right\}=1-\frac{\exp\left(\theta-\beta_j\right)}{1+\exp\left(\theta-\beta_j\right)}.$$

Элементарные преобразования правой части последней формулы посредством приведения разности к общему зна-

менателю $1+\exp(\theta-\beta_j)$ позволяют выразить её в более удобном виде

$$Q_j = \frac{1}{1+\exp(\theta-\beta_j)}.$$

Симметрично возникла и модель, описывающая вероятность правильного ответа студентов с уровнем знаний θ_i на задания различного уровня трудности

$$P_i\{x_{ij}=1|\theta\} = \frac{\exp(\theta-\beta_j)}{1+\exp(\theta-\beta_j)}.$$

Значения θ_i и β_j могут быть аппроксимированы из матрицы эмпирических данных.

Производная функции получается посредством дифференцирования дроби

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

$$= \frac{\left(e^{D(\theta-\beta_j)}\right)'\left(1+e^{D(\theta-\beta_j)}\right) - \left(e^{D(\theta-\beta_j)}\right)\left(1+e^{D(\theta-\beta_j)}\right)'}{\left(1+e^{D(\theta-\beta_j)}\right)^2} =$$

$$= \frac{\left(De^{D(\theta-\beta_j)} + De^{2(\theta-\beta_j)} - De^{2(\theta-\beta_j)}\right)}{\left(1+e^{D(\theta-\beta_j)}\right)^2}.$$

После уничтожения подобных членов последнее выражение становится равным

$$\frac{De^{D(\theta-\beta_j)}}{1+De^{D(\theta-\beta_j)}} \cdot \frac{1}{De^{D(\theta-\beta_j)}} = DP_jQ_j.$$

Для варианта модели с добавлением константы D производная равна DP_jQ_j.

Для модели G. Rasch, подставляя в формуле (3.31) вместо $(P_j')^2$ равное ему выражение $D^2P^2{}_jQ^2{}_j$ и сокращая полученное выражение на P_jQ_j, после выведения константы D за знак суммы получаем

$$I(\theta) = D^2\sum_{j=1}^{k} P_jQ_j.$$

Например, информационная функция теста, состоящего из k числа заданий, для модели G. Rasch, вычисляется посредством формулы

$$I(\theta_i) = D^2(P_1Q_1 + P_1Q_1 + ... + P_kQ_k).$$

Поскольку значение константы заранее известно, то вопрос расчёта информативности теста в избранной точке θ_i сводится, таким образом, к нахождению суммы произведений вероятности правильного ответа на вероятность неправильного ответа в каждом задании. По справедливому утверждению F.M. Lord, информационная функция теста указывает на меру эффективности измерения на каждом уровне континуума знаний.

Интерпретация графика информационной функции столь же проста, сколь и эффективна для разработчика теста: чем больше значение $I(\theta_i)$, тем лучше тест измеряет. Максимум информации при изме-

рении знаний испытуемых получается в той точке, где $I(\theta_i)$ принимает максимальное значение. В таких случаях можно говорить, что тест разработан для измерения знаний студентов с уровнем, где $I(\theta_i)$ принимает значение максимума. Там, где значение $I(\theta_i)$ минимально, можно определённо говорить о неэффективности теста для измерения знания у студентов с соответствующей подготовкой.

Использование информационной функции приводит, по сути, к оценке дифференцированной точности измерения, чего не было в статистической (классической) теории тестов. Там ошибка измерения принималась для всех испытуемых одинаковой. Если в качестве ошибок измерения se брать значения $1/I(\theta_i)$ в точке θ_i, то ошибка измерения у разных испытуемых станет отличаться, в зависимости от полученного значения θ_i и от значений $I(\theta_i)$.

Этот, казалось бы, на первый взгляд, математико-измерительный нюанс на самом деле отражает нечто большее, связанное с научным статусом тестов. Критики тестов интуитивно осознавали невозможность точного измерения знаний испытуемых различного уровня подготовленности с помощью одного и того же теста. Это одна из причин того, что в практике стремились обычно создавать тесты, рассчитанные на измерение подготовленности испытуемых самого многочисленного, среднего уровня. Естественно, что при такой ориентации теста знания у сильных и слабых испытуемых измеряются с меньшей точностью.

Информационная функция в IRT считается не только для отдельного задания, но и для теста в целом. Для этого надо суммировать количество информации, даваемое в интересующей точке θ_i каждым заданием. Определение информационной функции теста для каждой интересующей точки оси тета даёт возможность целенаправленной разработки такого теста, который позволяет решать задачи, например, профотбора. Если надо отобрать половину лучших, то точность измерения повышают на уровне средней подготовленности, где и применяется данное решающее правило. В таких случаях в тест добавляют задания среднего уровня трудности. Но тогда очень плохо измеряется подготовка сильных и слабых испытуемых. Пример решения такого рода представлен на рис. 14.

Если нужно, чтобы тест был эффективен на всём интересующем диапазоне подготовленности испытуемых, то задания теста стараются подобрать равномерно возрастающей трудности. Учебный пример

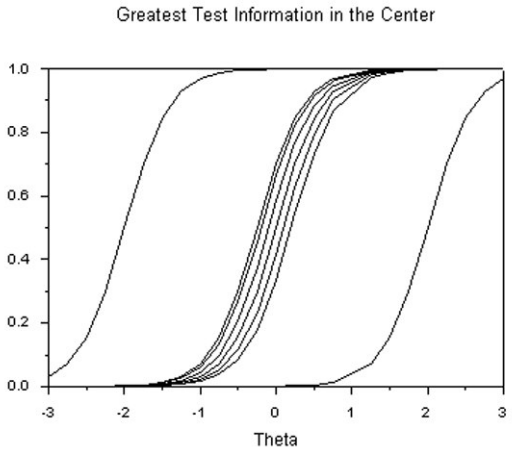

Рис. 14. Измерение точнее для средне подготовленных испытуемых

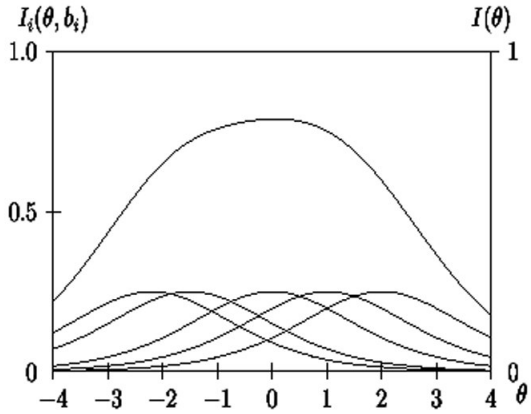

Рис. 15. Информационные функции каждого задания (синий цвет) и «теста» в целом (красный цвет)

такого рода равномерного подбора заданий «теста», состоящего всего из пяти заданий представлен на рис. 15.

Если нужно отобрать 10% лучших, то в тесте лёгкие задания не нужны. Их информационная ценность близка к нулю, в то время как включение трудных и очень трудных заданий при решении такой задачи повышает точность принимаемых кадровых решений и улучшает качество профотбора в целом.

Последним и очень важным в IRT понятием, которое получает здесь своё определение, является то, что по-английски называют test calibration. В текстах на русском языке это словосочетание переводится дословно как калибровка заданий или теста, что автору этой работы ни о чём хорошем не говорит. Потому что этим словосочетанием операцией определяется не калибр теста, а совсем другое.

Ключ к интерпретации словосочетания test calibration лежит в правильном понимании сущности теста. Автору этой статьи уже не раз приходилось обращать внимание на то, что в тест – это не просто набор или некое множество заданий, но также резуль-

таты тестирования, а также интерпретация результатов тестирования.

Таким образом, понятие «test calibration» на русском языке надо понимать как совместное шкалирование значений уровней подготовленности испытуемых, уровней трудности заданий и параметров крутизны заданий. Поскольку в модели Г. Раша фактор крутизны является константой, то остаются шкалирование уровня подготовленности испытуемых и уровней трудности заданий.

Преимущества и ограничения МТИ (IRT)

По сравнению с другими теориями измерений МТИ (IRT) имеет некоторые преимущества и ограничения. R.E. Schumacker[51] выделяет следующие преимущества IRT:

- оценки параметров заданий, определяемые с помощью IRT, не зависят от выборок испытуемых;
- оценки параметров испытуемых не зависят от уровня трудности теста в целом и от уровня трудности отдельных заданий;
- посредством IRT легче добиться соответствия уровня трудности заданий уровню подготовленности испытуемых, что повышает качество измерений;
- уровень подготовленности испытуемых и уровень трудности заданий определяется на одной и той же стандартной логарифмической шкале.
- надёжность тестовых результатов можно определить без применения параллельных вариантов теста.

Последнее из указанных преимуществ МТИ (IRT) открывает, казалось бы, возможность математического решения трудной педагогической задачи — создания параллельных вариантов теста. Однако история, теория и особенно практика педагогических измерений убеждают в необходимости разрабатывать параллельные варианты теста. Без чего невозможно качественно проводить массовое тестирование.

К перечисленным выше преимуществам иногда добавляют, что в МТИ (IRT) нет необходимости предполагать нормальность распределения тестовых результатов. Но для специалистов по педагогическим измерениям и это преимущество сомнительного толка. Так как отклонения исходных тестовых баллов от нормального распределения всегда связаны с нарушением баланса подбора заданий различного уровня трудности, а также несоответствием уровня трудности заданий уровню подготовленности испытуемых.

Кроме того, к положительным сторонам МТИ (IRT) относят философское свойство

51 *Schumacker Randall E.* ITEM RESPONSE THEORY. http://www.appliedmeasurementassociates.com/White%20Papers/ITEM%20RESPONSE%20THEORY.pdf

фальсифицируемости этой теории. То есть всегда есть возможность убедиться в степени пригодности или непригодности любой применяемой модели для собранных данных в случае несоответствия искать другую модель. Но и это преимущество оборачивается недостатком, если смотреть на IRT с позиции теории измерения по Г. Рашу. В теории Г. Раша заложена противоположная философия: данные должны подходить под используемую там математическую модель. И только тогда можно получить качественные измерения на интервальной шкале.

В числе ограничений этот же автор указывает, что применение IRT:

- основано на более строгих предположениях, чем применение вариантов классической статистической теории;
- требует математической подготовки, а потому эта теория плохо понимается педагогической общественностью и социумом, в котором эта теория применяется.
- требует больших выборок испытуемых для обоснования оценок параметров.

Критики МТИ (IRT) указывают также на поразительную корреляцию тестовых баллов испытуемых, полученных методами IRT (θ_i) и методами классической статистической теории измерений (X_i). Значения коэффициента корреляции между значениями θ_i и X_i на одном и том же множестве испытуемых равняются 0.99 и даже выше[52]. Тогда ставится вопрос — зачем применять IRT, если сравнительно простые методы приводят к столь сходным результатам[53]?

Простого ответа на этот вопрос нет. Здесь многое зависит от целей, критериев качества, уровня допустимых затрат и научности массовых исследований, программной оснащённости, от уровня некомпетентности и особенно от фактора Public Relation (PR). МТИ (IRT) хорошо использовать для экспертизы теста и тестовых заданий, для оценки соответствия уровня трудности задания уровню подготовленности испытуемых. Посредством МТИ (IRT) хорошо решаются локальные задачи определения ошибки измерения испытуемых определённого уровня подготовленности. МТИ (IRT) позволяет находить локальные коэффициенты надёжности тестовых результатов испытуемых интересующего уровня подготовленности на основе значений информационной функции.

Вообще, МТИ (IRT) — это теория, используемая для научного анализа качества тестовых заданий и теста в целом. Массовое применение этой теории для разработки тестов требует специального обуче-

52

Fan X. Item response theory and classical test theory: An empirical comparison of their item/person statistics. Educational and Psychological Measurement. 1998; 58(3): 357–381.

53

Anderson John O. Does complex analysis (IRT) pay any dividends in achievement testing? http://www.educ.uvic.ca/epls/faculty/anderson/documents/Paper2.pdf

Если, например, при обсуждении недостатков ЕГЭ отвечают, что Вам не понять, какая сложная математика используется для перевода данных из одной шкалы в другую и как, в конце концов, определяются итоговые баллы госэкзаменуемых, то это и есть один из неприемлемых случаев массового применения IRT.

ния тех, кто уже владеет началами педагогической и статистической теорий измерений. Широкомасштабное применение МТИ (IRT), задуманное из благих целей, при недостатке образования может привести к профанации и этой теории. Так что и в этом деле требуется определённая сдержанность и культура[54].

Что касается обобщённых или интегральных характеристик тестовых результатов, таких, как надёжность и валидность тестовых результатов, то здесь значительная роль отводится классической статистической теории измерений. МТИ (IRT) бесполезна для решения ряда задач педагогической теории измерений: при разработке тестовых форм, содержания тестовых заданий и теста. Потому что, как уже отмечалось, IRT — формальная педагогическая теория.